Published in final edited form as:
Nature. 2009 July 16; 460(7253): 352–358. doi:10.1038/nature08160.

The genome of the blood fluke *Schistosoma mansoni*

A full list of authors and affiliations appears at the end of the article.

Abstract

Schistosoma mansoni is responsible for the neglected tropical disease schistosomiasis that affects 210 million people in 76 countries. We report here analysis of the 363 megabase nuclear genome of the blood fluke. It encodes at least 11,809 genes, with an unusual intron size distribution, and novel families of micro-exon genes that undergo frequent alternate splicing. As the first sequenced flatworm, and a representative of the lophotrochozoa, it offers insights into early events in the evolution of the animals, including the development of a body pattern with bilateral symmetry, and the development of tissues into organs. Our analysis has been informed by the need to find new drug targets. The deficits in lipid metabolism that make schistosomes dependent on the host are revealed, while the identification of membrane receptors, ion channels and more than 300 proteases, provide new insights into the biology of the life cycle and novel targets. Bioinformatics approaches have identified metabolic chokepoints while a chemogenomic screen has pinpointed schistosome proteins for which existing drugs may be active. The information generated provides an invaluable resource for the research community to develop much needed new control tools for the treatment and eradication of this important and neglected disease.

Correspondence and requests for materials should be addressed to M.B. (mb4@sanger.ac.uk) and N.M.E.S (elsayed@umd.edu).
[a]The Broad Institute, 7 Cambridge Center, Cambridge, MA 02142, USA
[b]Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil
[c]Dept. Medicine, University of Pittsburgh, Pittsburgh, PA, USA
[d]Fios Genomics Ltd, ETTC, King's Buildings, Edinburgh, EH9 3JL, UK
[e]Biomedical Imaging Unit, School of Medicine, University of Southampton, Southampton, SO16 6YD, UK
[f]John Innes Centre, Norwich Research Park, Colney, Norwich, NR4 7UH, UK
[g]Leiden University Medical Centre, Parasitologie, Albinusdreef, 2333 ZA Leiden, The Netherlands
[h]Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD, USA
[i]Immunology/Microbiology, Rush University Medical Center, 1735 West Harrison Street, Chicago, IL 60612-3824, USA
[j]Department of Biochemistry, School of Medicine and Biomedical Research, State University of New York at Buffalo, Buffalo NY 14214 and Developmental Genomics Group, New York State Center of Excellence in Bioinformatics and Life Sciences, 701 Ellicott Street, Buffalo, NY 14203

Author Contributions: AI, AW, CMFL, DAJ, NMES and PTLV initiated the project; MQ constructed DNA libraries and JP and JR directed sequencing; YG and ZN assembled the genome sequence data; HH, PTLV, RP and YH produced the mapping data; ADe, ADj, ART, BH, DB, DL, GCC, JW, MAA, MAR, MSa, OW, PDA, RH, SLS and TE provided computational and bioinformatic support; ART, MAA and RH setup and maintained the genome database; CM, DB, GB, GCC and JG produced the gene finding training set; BH, MP and MSt trained genefinding software; AP, BB and BH annotated the genome data; AC, AZ, BAL, CRC, DLW, GO, GPD, JPO, LFA, MSa, MZ, PMV, RC, RDM, STM, TAD and WW contributed specific analysis topics presented in this manuscript; CHF and EG contributed to general project and sequencing management; BB, CHF, CRC and JP commented on the manuscript drafts; GCC performed data submission to Genbank; AW, MB, NMES, PTLV drafted and edited the paper; AW, DAJ and PTLV provided DNA resources for the sequencing; MB and NMES directed the project and assembled the manuscript

Author Information: The annotated genome sequence has been submitted to EMBL with the accession numbers FN357292-FN376313. All data are also available for browsing in the GeneDB database (http://www.genedb.org/genedb/smansoni/). The CHORI BAC clones used in this study are available from http://bacpac.chori.org/.

The authors have no competing financial interests.

Schistosomiasis is a Neglected Tropical Disease that ranks with malaria and tuberculosis as a major source of morbidity affecting approximately 210 million people in 76 countries of the world, despite of strenuous control efforts1. Is caused by blood flukes of the genus *Schistosoma* (Phylum Platyhelminthes), which exhibit dioecy and have complex life cycles comprising multiple morphologically distinct phenotypes in definitive human and intermediate snail hosts. *S. mansoni*, one of the three major human species, occurs across much of sub-Saharan Africa, parts of the Middle East, Brazil, Venezuela and some West Indian islands. The mature flukes dwell in the human portal vasculature, depositing eggs in the intestinal wall that either pass to the gut lumen and are voided in the faeces, or travel to the liver where they trigger immune-mediated granuloma formation and peri-portal fibrosis2. Approximately 280,000 deaths per annum are attributable to schistosomiasis in sub-Saharan Africa alone3. However, the disease is better known for its chronicity and debilitating morbidity4. A single drug, praziquantel, is almost exclusively used to treat the infection but this does not prevent reinfection and with the large-scale control programs in place, there is concern about the development of drug resistance. Indeed, resistance can be selected for in the laboratory and there are reports of increased drug tolerance in the field5.

In this study we present the sequence and analysis of the *S. mansoni* genome. Previous metazoan projects have been restricted to Deuterostomia (e.g. *Homo*, *Mus*, *Ciona*) and the ecdysozoan clade of the Protostomia (e.g. *Drosophila*, *Caenorhabditis*, *Brugia*). Together with the accompanying article on *S. japonicum*, we present the first descriptions of metazoan genomes from the lophotrochozoan clade. The genome reveals features that aid our understanding of the evolution of complex body plans. We have mined the genome to predict new drug targets, based on searches involving traditional areas for drug discovery, metabolic reconstruction, and bioinformatics screens that exploit shared pharmacology. It is hope that these, and other, targets will accelerate drug discovery, generating the much needed new treatments for control and eradication of schistosomiasis.

Genome structure and content

The nuclear genome sequence of *S. mansoni* was determined by whole genome shotgun and assembled into 5,745 scaffolds greater than 2 kb (Supplementary Table 1), totalling 363 megabases (Mb). Although 40% of the genome is repetitive, 50% percent is assembled into scaffolds of at least 824.5 kb. Furthermore, 43% of the genome assembly (distributed over 153 scaffolds) was unambiguously assigned to chromosomes (7 autosomal, plus ZW sex-determination pairs) using Fluorescence *In Situ* Hybridization (Fig. 1, Supplementary Fig. 1 and Supplementary Table 2).

We identified 72 families of both LTR and non-LTR transposons, comprising 15% and 5% of the genome, respectively, and containing 63 and 60 new families each (Supplementary Table 3). The LTR transposons are from the Gypsy/Ty3 and BEL clades while the non-LTR transposons are restricted to the RTE, CR1 and R2 clades. Two previously described non-LTR retrotransposon families from the RTE clade (SR2 and perere-3)6,7, appear to have undergone a burst of transposition events after divergence of *S. mansoni* and *S. japonicum* and contribute to an overall higher representation of non-LTR retrotransposons in *S. mansoni* (15%, *cf.* 8% in *S. japonicum*). A novel DNA transposon belonging to the Mu family was also found, which represents the first instance in a flatworm. The presence of target site duplications in some copies implies recent transposition and suggests that active copies may still exist in the genome. A lack of terminal inverted repeats –a feature of Mu family members – suggests a peculiar mechanism for recognition of this element by the transposition apparatus.

We identified 11,809 putative genes encoding 13,197 transcripts. Considering genes that do not span a gap, the average gene size is 4.7 kb, typically with large introns (average is 1692 bp) and much smaller exons (average is 217 bp). Moreover, the introns display a strikingly skewed size distribution that has not been observed in other eukaryotes, whereby 5′ introns are smaller than 3′ introns (Fig. 2, Supplementary Information, Supplementary Table 5). In multi-exon genes the first few introns can be as small as 26 bp, whereas introns towards the 3′ end are typically kilobases in length (largest is 33.8 kb). The reason for this is unclear but suggests unusual transcriptional control. However, a survey of conserved transcription factor domains reveals *S. mansoni* to be broadly similar to other eukaryotes (Supplementary Information, Supplementary Fig. 2 and Supplementary Table 6). It is noteworthy that 43% of transcription factor families with schistosome representatives also contained vertebrate sequences, nearly twice the number that matched nematode worms, emphasising their evolutionary distance.

Micro-exon genes (MEGs)

At least 45 genes have an unusual micro-exon structure. Individual micro-exons have been described in other genomes, dispersed among numbers of normal exons8. However, *S. mansoni* is remarkable in containing micro-exon genes (MEGs) where micro-exons comprise 75% of the coding sequence, are flanked at the 5′ and 3′ extremes by conventional exons and have lengths that are multiples of three bases (from 6 to 36).

Other than having shared gene structure, no similarity could be detected between 14 MEG families (each with up to 23 members; Fig. 3 and Supplementary Table 7). Moreover, they displayed no similarity with annotated genes from outside *Schistosoma* spp, nor any identifiable motifs or functional domains. Comparisons between MEG family members and related proteins from *S. japonicum* suggest that some gene duplication events preceded the divergence of the two species. Almost all encode a signal peptide at the 5′ end and three have membrane anchors, so most are probably secreted. Examination of the large EST data set from across the life cycle reveals that genes from all MEG families are transcribed in the intramammalian stages of the life cycle, and the germ balls of daughter sporocysts that develop into infective cercariae, but probably not in miracidia that infect the snail intermediate host (Fig. 3).

Sequencing of transcripts from three MEG families revealed the occurrence of numerous alternative splice variants formed by exon skipping. In one of the families analyzed, all internal exons except those coding for the signal peptide were missing in at least one transcript sampled and a gene from a second family presented different transcripts with extended exons produced by the use of alternative splicing sites. These observations suggest a ‘pick and mix’ strategy is used to create protein variation

Evolution of the triploblast state, parasitism and distinct tissues

Schistosomes are the first Platyhelminthes to be fully sequenced and provide insights into the evolution of 'simple' animals. Using Treefam to make comparisons with the sea anemone *Nematostella vectensis*, a representative of the Radiata, we sought gene families restricted to, or expanded in the Bilateria (Supplementary Table 8). The advent of a third germ layer in flatworms is paralleled by the expansion of genes encoding cell adhesion molecules such as cadherins. Similarly, tissue-patterning developmental cues (e.g. Notch/ Delta) and histone-modifying enzymes (e.g. histone acetyltransferases) have proliferated. Some genes such as the Tetraspanins that encode membrane structural proteins have greatly proliferated in schistosomes suggesting a critical role in worm physiology/parasitism. The large array of paralogues for fucosyl and xylosyltransferases involved in the generation of novel glycans expressed at the host-parasite interface, may be important for subverting the immune system. The expansion of proteases in schistosomes also appears directly related to parasitism, as it includes families involved in host invasion (Invadolysins) and blood feeding (Cathepsins). Finally, G-Protein-coupled receptors show varying levels of contraction in schistosomes whereas several classes (e.g. peropsins) are greatly expanded in *Nematostella* implying functions associated with the free-living lifestyle.

Although schistosomes are acoelomate, they possess tissues approaching the sophistication of organs such as gut, nephridia, nerve and muscle, concerned with discrete physiological processes such as feeding, excretion and locomotion. However, as lophotrochozoans they are evolutionarily distant from the previously sequenced parasitic nematodes, *Brugia*9 and *Meloidogyne*10,11 (both ecdysozoans). Compartmentalisation of schistosome tissues and the formation of epithelial barriers are crucial for life in the hostile environment of the host bloodstream. Schistosomes possess the typical machinery of higher metazoa to interact with the cytoskeleton and control cell polarity (Supplementary Information, Supplementary Table 9), organise epithelia and denote tissue boundary lines.

S. mansoni posses a nervous system that includes an anterior brain and longitudinal nerve cords, which extend from the brain to run the length of the worm body. In addition, a variety of sensory structures (at least six types in the cercaria12) are able to transduce a wide range of stimuli that assist in host location, penetration and navigation through the vasculature. In common with more complex organisms, schistosomes possess the tools needed to mediate neurogenesis and control axon growth cones and migration of neural cells (Supplementary Information, Supplementary Table 9), supporting the ancient origins of neural complexity.

Insights into possible new drug targets

Historically, anti-schistosomiasis agents were identified by *in vivo* screening in animal models. The *S. mansoni* genome project makes a more target-based approach to drug discovery feasible and some promising leads have already emerged. These include a family of nuclear receptors 13 (Supplementary Information) and a redox enzyme, thioredoxin glutathione reductase, recently validated as a drug target14. The condensed redox biochemistry of *S. mansoni*, relative to its human host, may offer further drug development targets (Supplementary Information). In the context of drug discovery, we have explored other potential areas of vulnerability: lipid metabolism G-protein coupled receptors, ligand- and voltage-gated ion channels; kinases; proteases; and neuropeptides. We also undertook two bioinformatics-led approaches: metabolic reconstruction to identify chokepoints, and sequence searches for structures related to known drug targets.

Lipid metabolism

S. mansoni contains a full complement of genes required for most core metabolic processes, such as glycolysis, tricarboxylic acid cycle and the pentose phosphate pathway. However, schistosomes are incapable of *de novo* synthesis of sterols or free fatty acids and must utilise complex precursors from the host15. An extensive lipid carrying protein repertoire could be identified but, despite producing precursors for fatty acid synthesis, fatty acid synthase could not be identified. An inability to utilize isoprene products of the mevalonate pathway most likely accounts for the lack of sterol biosynthesis (Supplementary Table 11, Supplementary Information). The genes necessary for a complete beta oxidation pathway are present, and this usually inactive pathway might operate in reverse to perform syntheses16. Despite constituting 40% or more of the lipid content of adult worms15, triacylglycerol plays an uncertain role in the schistosome's life cycle; they are slow to turn over, do not contribute to the formation of other lipids15 and their use as an energy store is doubtful16. Nevertheless, *S. mansoni* possesses lipases capable of breaking down triacylglycerol, so they may have functions beyond preventing too high concentrations of intracellular fatty acids15. Pathways responsible for synthesizing the phospholipid components of membranes are well represented except that phosphatidylcholine must be derived from diacylglycerol17 and the parasite must depend on its host as a source of inositol.

G-protein coupled receptors, ligand gated ion channels and voltage gated ion channels

G-protein couple receptors (GPCRs), ligand-gated ion channels (LGICs) and voltage-gated ion channels (VGICs) are targets for 50% of all current pharmaceuticals18. At least 92 putative GPCR-encoding genes are present (Supplementary Table 12), the bulk (82) from the rhodopsin family. The largest groups are the alpha subfamily (30), which includes amine receptors and the beta subfamily (24), which contains neuropeptide and hormone receptors. The diversity of the former subfamily underlines the wide range of potential amine/ neurotransmitter reactivities of schistosomes but the tentative identities assigned need to be confirmed by functional studies, as has already been performed for a histamine receptor19. Schistosomes detect chemosensory cues but a large, unique clade of the mediating receptors was not found. However, the 26 “orphan” rhodopsin family GPCRs may include proteins with this role. Outside the large rhodopsin family, representatives from each of the smaller families of GPCRs, glutamate family (2), frizzled family (3), and the secretin/adhesion family (4) are present.

Each of the three major LGIC families, the Cys-loop family, Glutamate activated cation channels, and ATP-gated ion channels, are represented in the schistosome genome. Of the 13 Cys-Loop Family LGICs, nine encode nicotinic acetylcholine receptor subunits (Supplementary Fig. 4 and Supplementary Table 13). The remaining four anion channel subunits group amongst GABA, glycine and glutamate receptors but it is not possible to assign precise identities. The seven schistosome glutamate-activated cation channels comprise at least two sequences from each of the three common sub-groupings. The presence of a functional P2X receptor for ATP-mediated signalling in schistosomes was already known20, and the data here reveal at least four more.

VGICs generate and control membrane potential in excitable cells and are central to ionic homeostasis. There are examples of successful drugs targeting voltage-gated sodium, potassium and calcium channels21. Although voltage-gated sodium channels were not found, at least 41 members from each of the major of six transmembrane (6TM) and four transmembrane (4TM) families of potassium channel (Supplementary Table 14) are present. The 6TM voltage-gated potassium channel family (20 members) is the largest, including the well-characterized Kv1.1 channel found in nerve and muscle of adult schsitosomes22. Other classes of 6TM potassium channels include the KQT channels, large calcium-activated channels, small calcium-activated channels, and cyclic-nucleotide-gated groups. This last comprising 8 members is most often associated with signal transduction in primary olfactory and visual sensory cells (*C. elegans* has only 523). *S. mansoni* possesses six 4TM inward-rectifying TWIK-related postassium channels (*cf.* 46 in *C. elegans*). There are four alpha and two beta subunits of voltage-gated calcium channels in schistosomes and a beta subunit is implicated as a molecular target of the antischistosomal praziquantel24.

The kinome

Protein kinases are important regulators of many different cellular functions. Both they and their inhibitors have entered the drug development pipeline in recent years25 but few schistosome kinases have been characterized to date. The *S. mansoni* genome encodes 249 kinases, including 22 genes with alternative splicing (Supplementary Information). This corresponds to 1.9% of the total coding proteins in the genome, a figure comparable to that found in other species26 (Supplementary Fig. 6). *S. mansoni* possesses representatives of all of the main kinase groups (Supplementary Fig. 7), the largest of which is the CMGC (cyclin-dependent, MAP-, glycogen synthase kinase 3 and CK2-related kinases) group, in contrast to other analysed eukaryotic genomes. However, a single class (RCK) is absent from the CMGC family, a deficiency shared with yeast but not nematodes or mammals.

The least represented groups are the Casein Kinase (CK1) and Receptor Guanylate Cyclase (RGC) families with only 7 and 3 members, respectively, contrasting with *C. elegans* where CK1 is the largest group and RGC has 27 members. CK1 (and CMGC) group members that are expressed in sperm or during spermatogenesis in *C. elegans*, are missing in *S. mansoni*.

The degradome

Proteolytic enzymes (proteases), making up an organism's ‘degradome’27, operate in virtually every biological and pathological phenomenon28 and are proven drug targets in diverse biomedical contexts29,30. All five major classes of proteases (aspartic, cysteine, metallo-, serine and threonine) are represented as various clans (mechanistically related groups) in the parasite genome (Supplementary Table 17). The percentage distribution of the major clans is generally similar to that of the human host with some notable exceptions, mainly due to the expansion of constituent protease families in humans. Sixty-one of the 73 protease families found in humans are also in *S. mansoni* and 60 families are shared. With 335 sequences, proteases comprise 2.5% of the putative proteome (Supplementary Table 18), consistent with the proportion in other organisms (1-5%), but only one-third that in humans (945 sequences, if A2 family retrovirus and retrotransposon proteases are included).

The greatest difference between host and parasite is in the paucity of Family S1 chymotrypsin-like enzymes in the latter (22 vs. 135 human sequences). This reflects the evolution and diversification of Family S1 for complex and highly regulated proteolysis cascades in vertebrates and some invertebrates such as innate immunity, development, blood coagulation and complement activation31-33. From a therapeutic standpoint, the reduced complexity may prove valuable with fewer parasite proteases available for essential life-sustaining functions. For example, robust drug discovery programs are in place for Families S134 and C14 (caspases)35, upon which anti-schistosomal drug discovery could ‘piggy-back’36. It is also notable that a smaller number of schistosome protease families (e.g., C1, M8 and M13) have more members than the respective families in humans. C1 proteases are involved in nutrient digestion by the parasite, which contrasts with the S1 enzymes employed in the host. This disparity has already been exploited for a promising anti-schistosome therapy37. One protease family (C83) is apparently unique to *S. mansoni*.

Apart from the degradome, but involved in its modulation, 34 protease inhibitors were found (Supplementary Table 19). The majority of these are serine protease inhibitors belonging to Families I2 (Kunitz-type) and I4 (serpins). Two inhibitors of cysteine proteases (cystatins38,39) and two alpha-2-macroglobulin homologues (I39) were also identified, as were three Inhibitor of Apoptosis proteins (I32), one of which is highly expressed in adults, where it may function to regulate one or more of the four schistosome caspases.

Neuropeptides

Thirteen putative neuropeptides were identified (Supplementary Table 20), indicating that schistosomes may display much greater diversity than the two described previously. Apart from the neuropeptide Fs (NPFs), most are apparently restricted to the Platyhelminthes, their absence from humans making them a credible source of anthelmintic drug leads. The predicted product of npp-6 (AVRLMRLamide) resembles molluscan myomodulin, while the two NPP-13 peptides display 100% C-terminal identity with vertebrate neuropeptide-FF-like peptides (PQRFamides); neither of these has previously been reported in any non-vertebrate organism. The discovery of a second NPF (Sm-NPP-21b) additional to the known Sm-NPP-21a40 is reminiscent of the vertebrate neuropeptide Y (NPY) superfamily, and strengthens the argument that NPFs and NPYs have a common ancestry.

Metabolic chokepoints

A chokepoint analysis of metabolic pathways reconstructed from the *S. mansoni* genome was used to identify additional targets. A total of 607 enzymatic reactions could be placed in pathways and 120 of these enzymes were identified as chokepoints (Supplementary Table 21). The list of chokepoints includes many that are drug targets in other organisms as well as target reactions already characterized in *S. mansoni*, validating the approach (Supplementary

Information). The list also contains new candidate targets and comprises approximately 1% of the *S. mansoni* proteome.

Chemogenomics screening

In the context of Neglected Tropical Diseases, with constrained investment in drug discovery, 'piggy-backing'36 or 'drug-repositioning' strategies41 that re-use existing drugs, offer potential time savings and cost benefits. We adopted a two-fold strategy to find significant matches between proteins from the parasite and known 'druggable' protein targets of the human host and human-infective pathogens. Using conservative parameters of > 50% sequence identity over > 80% of the target, we first performed a similarity search against a database of targets curated from medicinal chemistry literature. This revealed 240 distinct *S. mansoni* transcripts with matches to targets against which there are high quality compounds (Supplementary Table 22). Given the need for short-course, oral therapies against schistosomiasis, this list was further reduced to 94 *S. mansoni* targets by filtering for potency and predicted bioavailability. A second search, against a database of the targets for human-directed drugs, revealed 66 significant matches with currently marketed pharmaceuticals (Supplementary Table 23), corresponding to 34 *S. mansoni* targets (26, after representing multicopy genes as a single instance; Table 1). For instance disulfiram, for controlling substance abuse, was highlighted as a potential anti-schistosomal drug; its anti-parasite properties have already been investigated42. Manual inspection of the list for compounds with side effects and toxicity can further refine choices, e.g, by eliminating the immunosuppressants, cyclosporin and rapamycin. The remaining known drugs could be directly tested in animal models, and either applied unmodified in anti-schistosomal therapy, or could serve as leads for further optimisation. Widening the search beyond the initial strict criteria would expand opportunities, e.g. Topoisomerase 1, is retrieved below our initial threshold, at 71% identity but only 58% overlap.

Conclusion

A century after Louis Sambon first named the species in 1907, the sequencing of the *S. mansoni* genome is a landmark event. The sequence provides the scientific community with multiple avenues to study this under-researched human pathogen and will drive future evolutionary, genetic and functional genomic research. Not least, given that just one drug is widely available to treat schistosomiasis, the genome sequence, including the genome-mining analysis presented, offers the possibility that new drug candidates will be identified soon.

Methods Summary

Mixed sex cercariae from the Puerto Rico isolate of *S. mansoni*43, released from infected *Biomphalaria glabrata* snails, were placed in low-melting agarose plugs and genomic DNA prepared by standard methods. Approximately, six-fold coverage of the nuclear genome was obtained using a whole genome shotgun sequencing approach where libraries of different cloned insert sizes (in plasmid, fosmid and BAC vectors) were randomly sequenced by Sanger technology from either end. Sequence reads were assembled and scaffolds were

FISH mapped to individual chromosomes where possible (Supplementary Table 2). The output of several gene prediction algorithms, trained using 409 manually curated gene structures, were integrated into a single set of gene predictions (v4), which were used for subsequent analyses. Data were accessed via GeneDB (http://www.genedb.org) and Artemis was used for subsequent manual annotation and curation of a further 958 genes during subsequent analyses (as described previously44).

Full methods and all associated references are available in the online version of the paper at www.nature.com/nature.

Further details for additional methods used in this study are provided in Supplementary Information.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Matthew Berriman[1], Brian J. Haas[2,a], Philip T. LoVerde[3], R. Alan Wilson[4], Gary P. Dillon[4], Gustavo C. Cerqueira[5,6,7], Susan T. Mashiyama[8,9], Bissan Al-Lazikani[10], Luiza F. Andrade[11], Peter D. Ashton[4], Martin A. Aslett[1], Daniella C. Bartholomeu[2,b], Gaelle Blandin[2], Conor R. Caffrey[8], Avril Coghlan[12], Richard Coulson[13], Tim A. Day[14], Art Delcher[6], Ricardo DeMarco[4,15,16], Appoliniare Djikeng[2], Tina Eyre[1], John A. Gamble[1], Elodie Ghedin[2,c], Yong Gu[1], Christiane Hertz-Fowler[1], Hirohisha Hirai[17], Yuriko Hirai[17], Robin Houston[1], Alasdair Ivens[1,d], David A. Johnston[18,e], Daniela Lacerda[2,b], Camila D. Macedo[5,7], Paul McVeigh[14], Zemin Ning[1], Guilherme Oliveira[11], John P. Overington[13], Julian Parkhill[1], Mihaela Pertea[6], Raymond J. Pierce[19], Anna V. Protasio[1], Michael A. Quail[1], Marie-Adèle Rajandream[1], Jane Rogers[1,f], Mohammed Sajid[8,g], Steven L. Salzberg[6,7], Mario Stanke[20], Adrian R. Tivey[1], Owen White[2,h], David L. Williams[21,i], Jennifer Wortman[2,h], Wenjie Wu[3,j], Mostafa Zamanian[14], Adhemar Zerlotini[11], Claire M. Fraser-Liggett[2,h], Barclay G. Barrell[1], and Najib M. El-Sayed[2,5,6,7]

Affiliations

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, UK [2]The Institute for Genomic Research/The J. Craig Venter Institute, 9712 Medical Center Dr., Rockville, MD 20850, USA [3]Departments of Biochemistry and Pathology, Mail Code 7760, University of Texas, Health Science Center, San Antonio, TX 78229-3900, USA [4]Department of Biology, University of York, PO Box 373, York, YO10 5YW, UK [5]Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA [6]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA [7]Maryland Pathogen Research Institute, University of Maryland, College Park, MD 20742, USA [8]Sandler Center for Basic Research in Parasitic Diseases, California Institute for Quantitative Biomedical Research (QB3), Byers Hall, 1700 4th Street,

University of California, San Francisco, CA 94158-2330, USA [9]Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research (QB3), Byers Hall, 1700 4th Street, University of California, San Francisco, CA 94158-2330, USA [10]Cancer Research UK Centre for Cancer Therapeutics, The Institute of Cancer Research, Haddow Laboratories, 15 Cotswold Road, Belmont, Sutton, Surrey SM2 5NG, UK [11]Centro de Pesquisas René Rachou (CPqRR) – FIOCRUZ, Av Augusto de Lima 1715, Belo Horizonte, MG 30190002, Brazil [12]Department of Microbiology, University College Cork, Ireland [13]European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK [14]Department of Biomedical Sciences, Iowa State University, Ames, IA 50011, USA [15]Instituto de Química, Universidade de São Paulo, Brazil [16]Instituto de Física de São Carlos, Universidade de São Paulo, Brazil [17]Primate Research Institute, Kyoto University, Inuyama, Aichi 484 – 8506, Japan [18]Biomedical Parasitology Division, The Natural History Museum, London, SW7 5BD, UK [19]Inserm, U 547, Université Lille 2, Institut Pasteur de Lille, IFR 142, Lille, France [20]Institut für Mikrobiologie und Genetik, Abteilung Bioinformatik, Universität Göttingen, Goldschmidtstraβe 1, Göttingen, 37077, Germany [21]Department of Biological Sciences, Illinois State University, Normal, IL 61790-4120, USA

Acknowledgments

This genome sequencing and annotation work was funded by the Wellcome Trust [grant number WT085775/Z/08/Z] and NIH-NIAID grant AI48828 to NES. We thank Neil D. Rawlings of the MEROPS database team at the Wellcome Trust Sanger Institute for his invaluable help, Dr. Jean C. Illes for discussions on polarity complexes, and Fransisco Prosdocimi and Maria Rosa Domingo Sananes for early discussions and analyses in the project. FISH chromosome mappings were partially supported by Oyama Health Foundation (H.H.), JSPS (13557021) (H.H.), 21st century COE and global COE of MEXT. Additional support by The Sandler Foundation (C.R.C., M.S.), NIH R01 GM60595 (P.C.B.), NIH-Fogarty 5D43TW006580 (P.T.L.), NIH-Fogarty 5D43TW007012-03, NIH Grant AI054711-01A2 (R.A.W. and G.P.D.), FAPEMIG REDE-281/05 (G.O.), the PhRMA Foundation (Postdoctoral Fellowship in Informatics to S.T.M.), The Burroughs Wellcome Fund (P.T.L.) and The UNICEF/UNDP/World bank/WHO Special program for research and training in tropical diseases (TDR) (P.T.L.). R.D.M. was a recipient of CAPES and FAPESP fellowships.

References

1. Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J. Schistosomiasis and water resources development: systematic review, meta-analysis, and estimates of people at risk. Lancet Infect Dis. 2006; 6:411–425. [PubMed: 16790382]
2. Gryseels B, Polman K, Clerinx J, Kestens L. Human schistosomiasis. Lancet. 2006; 368:1106–1118. [PubMed: 16997665]
3. van der Werf MJ, et al. Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. Acta Trop. 2003; 86:125–139. [PubMed: 12745133]
4. King CH, Dickman K, Tisch DJ. Reassessment of the cost of chronic helmintic infection: a meta-analysis of disability-related outcomes in endemic schistosomiasis. Lancet. 2005; 365:1561–1569. [PubMed: 15866310]
5. Doenhoff MJ, Pica-Mattoccia L. Praziquantel for the treatment of schistosomiasis: its use for control in areas with endemic disease and prospects for drug resistance. Expert review of anti-infective therapy. 2006; 4:199–210. [PubMed: 16597202]

6. Drew AC, Minchella DJ, King LT, Rollinson D, Brindley PJ. SR2 elements, non-long terminal repeat retrotransposons of the RTE-1 lineage from the human blood fluke Schistosoma mansoni. Mol Biol Evol. 1999; 16:1256–1269. [PubMed: 10486981]

7. DeMarco R, Machado AA, Bisson-Filho AW, Verjovski-Almeida S. Identification of 18 new transcribed retrotransposons in Schistosoma mansoni. Biochem Biophys Res Commun. 2005; 333:230–240. [PubMed: 15939396]

8. Volfovsky N, Haas BJ, Salzberg SL. Computational discovery of internal micro-exons. Genome Res. 2003; 13:1216–1221. [PubMed: 12799353]

9. Ghedin E, et al. Draft genome of the filarial nematode parasite Brugia malayi. Science. 2007; 317:1756–1760. [PubMed: 17885136]

10. Abad P, et al. Genome sequence of the metazoan plant-parasitic nematode Meloidogyne incognita. Nature biotechnology. 2008; 26:909–915.

11. Opperman CH, et al. Sequence and genetic map of Meloidogyne hapla: A compact nematode genome for plant parasitism. Proc Natl Acad Sci U S A. 2008; 105:14802–14807. [PubMed: 18809916]

12. Dorsey CH, Cousin CE, Lewis FA, Stirewalt MA. Ultrastructure of the *Schistosoma mansoni* cercaria. Micron. 2002; 33:279–323. [PubMed: 11742750]

13. Wu W, Niles EG, Hirai H, LoVerde PT. Evolution of a novel subfamily of nuclear receptors with members that each contain two DNA binding domains. BMC Evol Biol. 2007; 7:27. [PubMed: 17319953]

14. Sayed AA, et al. Identification of oxadiazoles as new drug leads for the control of schistosomiasis. Nat Med. 2008; 14:407–412. [PubMed: 18345010]

15. Brouwers JF, Smeenk IM, van Golde LM, Tielens AG. The incorporation, modification and turnover of fatty acids in adult Schistosoma mansoni. Mol Biochem Parasitol. 1997; 88:175–185. [PubMed: 9274878]

16. Barrett, J. Biochemistry of Parasitic Helminths. Macmillan Publishers; London: 1981.

17. de Kroon AI. Metabolism of phosphatidylcholine and its implications for lipid acyl chain composition in Saccharomyces cerevisiae. Biochim Biophys Acta. 2007; 1771:343–352. [PubMed: 17010666]

18. Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there? Nat Rev Drug Discov. 2006; 5:993–996. [PubMed: 17139284]

19. Hamdan FF, et al. A novel Schistosoma mansoni G protein-coupled receptor is responsive to histamine. Mol Biochem Parasitol. 2002; 119:75–86. [PubMed: 11755188]

20. Agboh KC, Webb TE, Evans RJ, Ennion SJ. Functional characterization of a P2X receptor from Schistosoma mansoni. J Biol Chem. 2004; 279:41650–41657. [PubMed: 15292267]

21. Kaczorowski GJ, McManus OB, Priest BT, Garcia ML. Ion channels as drug targets: the next GPCRs. The Journal of general physiology. 2008; 131:399–405. [PubMed: 18411331]

22. Kim E, Day TA, Bennett JL, Pax RA. Cloning and functional expression of a Shaker-related voltage-gated potassium channel gene from Schistosoma mansoni (Trematoda: Digenea). Parasitology. 1995; 110(Pt 2):171–180. [PubMed: 7533899]

23. Salkoff L, et al. Potassium channels in C. elegans. WormBook. 2005:1–15. [PubMed: 18050399]

24. Jeziorski MC, Greenberg RM. Voltage-gated calcium channel subunits from platyhelminths: potential role in praziquantel action. Int J Parasitol. 2006; 36:625–632. [PubMed: 16545816]

25. Boyle SN, Koleske AJ. Dissecting kinase signaling pathways. Drug Discov Today. 2007; 12:717–724. [PubMed: 17826684]

26. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science. 2002; 298:1912–1934. [PubMed: 12471243]

27. Lopez-Otin C, Overall CM. Protease degradomics: a new challenge for proteomics. Nat Rev Mol Cell Biol. 2002; 3:509–519. [PubMed: 12094217]

28. Rawlings ND, Morton FR. The MEROPS batch BLAST: a tool to detect peptidases and their non-peptidase homologues in a genome. Biochimie. 2008; 90:243–259. [PubMed: 17980477]

29. Abbenante G, Fairlie DP. Protease inhibitors in the clinic. Med Chem. 2005; 1:71–104. [PubMed: 16789888]

30. Fear G, Komarnytsky S, Raskin I. Protease inhibitors and their peptidomimetic derivatives as potential drugs. Pharmacol Ther. 2007; 113:354–368. [PubMed: 17098288]

31. Page MJ, Di Cera E. Serine peptidases: classification, structure and function. Cell Mol Life Sci. 2008; 65:1220–1236. [PubMed: 18259688]

32. Krem MM, Di Cera E. Evolution of enzyme cascades from embryonic development to blood coagulation. Trends Biochem Sci. 2002; 27:67–74. [PubMed: 11852243]

33. Zou Z, Lopez DL, Kanost MR, Evans JD, Jiang H. Comparative analysis of serine protease-related genes in the honey bee genome: possible involvement in embryonic development and innate immunity. Insect Mol Biol. 2006; 15:603–614. [PubMed: 17069636]

34. Ieko M, et al. Factor Xa inhibitors: new anti-thrombotic agents and their characteristics. Front Biosci. 2006; 11:232–248. [PubMed: 16146728]

35. Okun I, Balakin KV, Tkachenko SE, Ivachtchenko AV. Caspase activity modulators as anticancer agents. Anti-Cancer Agents in Medicinal Chemistry. 2008; 8:322–341. [PubMed: 18393791]

36. Caffrey CR, Steverding D. Recent initiatives and strategies to developing new drugs for tropical parasitic diseases. Expert Opinion in Drug Discovery. 2008; 3:173–186.

37. Abdulla MH, Lim KC, Sajid M, McKerrow JH, Caffrey CR. Schistosomiasis mansoni: novel chemotherapy using a cysteine protease inhibitor. PLoS Med. 2007; 4:e14. [PubMed: 17214506]

38. Cao M, Chao H, Doughty BL. A cDNA from Schistosoma mansoni eggs sharing sequence features of mammalian cystatin. Mol Biochem Parasitol. 1993; 57:175–176. [PubMed: 8426612]

39. Morales FC, Furtado DR, Rumjanek FD. The N-terminus moiety of the cystatin SmCys from Schistosoma mansoni regulates its inhibitory activity in vitro and in vivo. Mol Biochem Parasitol. 2004; 134:65–73. [PubMed: 14747144]

40. Humphries JE, et al. Structure and bioactivity of neuropeptide F from the human parasites Schistosoma mansoni and Schistosoma japonicum. J Biol Chem. 2004; 279:39880–39885. [PubMed: 15229227]

41. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov. 2004; 3:673–683. [PubMed: 15286734]

42. Nash T, Rice WG. Efficacies of zinc-finger-active drugs against Giardia lamblia. Antimicrobial agents and chemotherapy. 1998; 42:1488–1492. [PubMed: 9624499]

43. Fletcher M, LoVerde PT, Woodruff DS. Genetic variation in Schistosoma mansoni: enzyme polymorphisms in populations from Africa, Southwest Asia, South America, and the West Indies. The American journal of tropical medicine and hygiene. 1981; 30:406–421. [PubMed: 7235131]

44. Berriman M, Harris M. Annotation of parasite genomes. Methods Mol Biol. 2004; 270:17–44. [PubMed: 15153621]

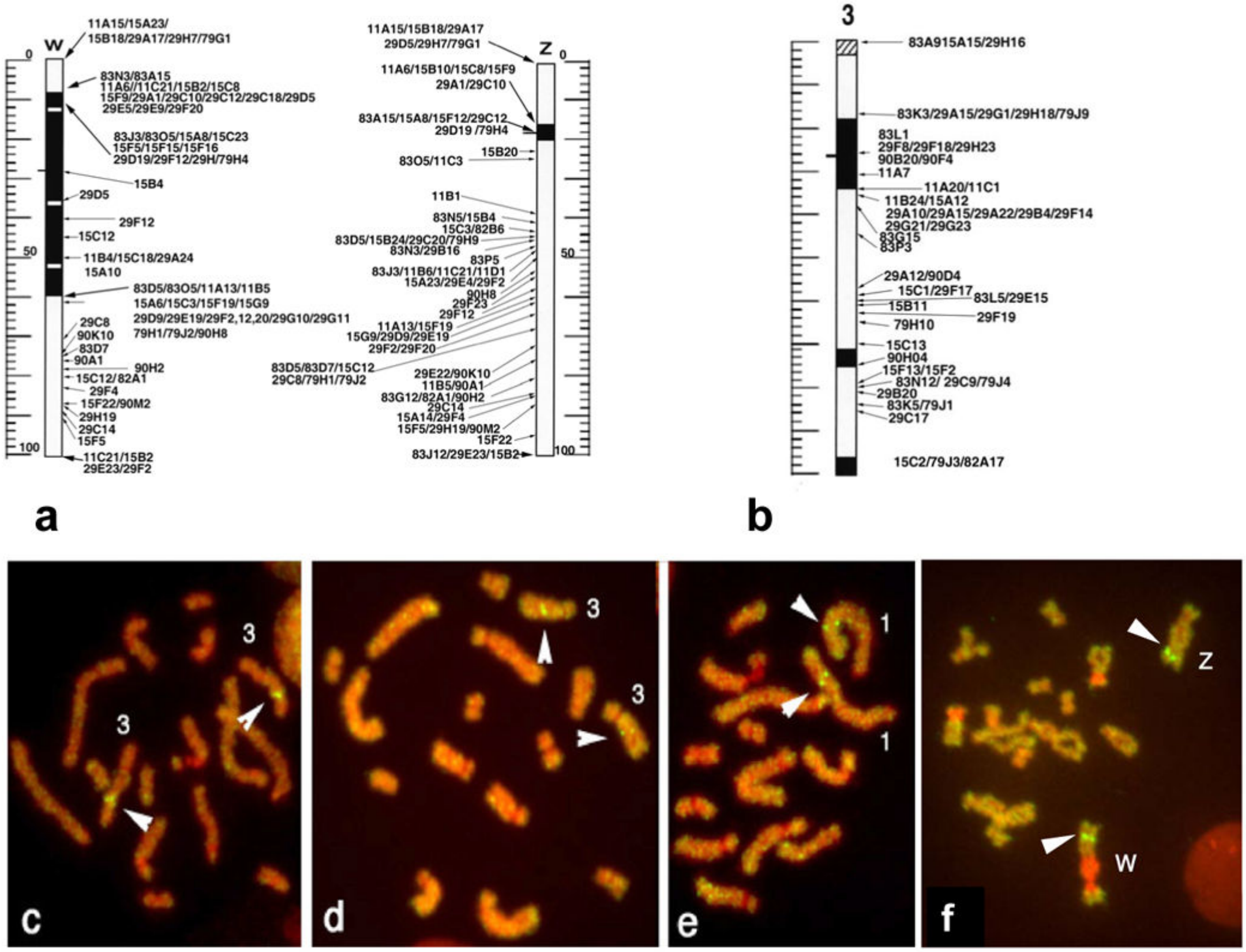

Figure 1. Physical map of Schistosoma mansoni

Idiogram of *S. mansoni* chromosomes W, Z (a) and 3 (b). *S. mansoni* BAC clones were mapped to the karyotype of *S. mansoni* by FISH. The solid black areas are heterochromatin and the open areas are euchromatin. The BAC clones are identified by BAC numbers. Panels c-f show chromosomes spreads with FISH mapped BACS. FISH mapped BACS are identified by arrow heads on labelled chromosomes. Idiograms for all *S. mansoni* chromosomes are included in the Supplementary Online Materials.

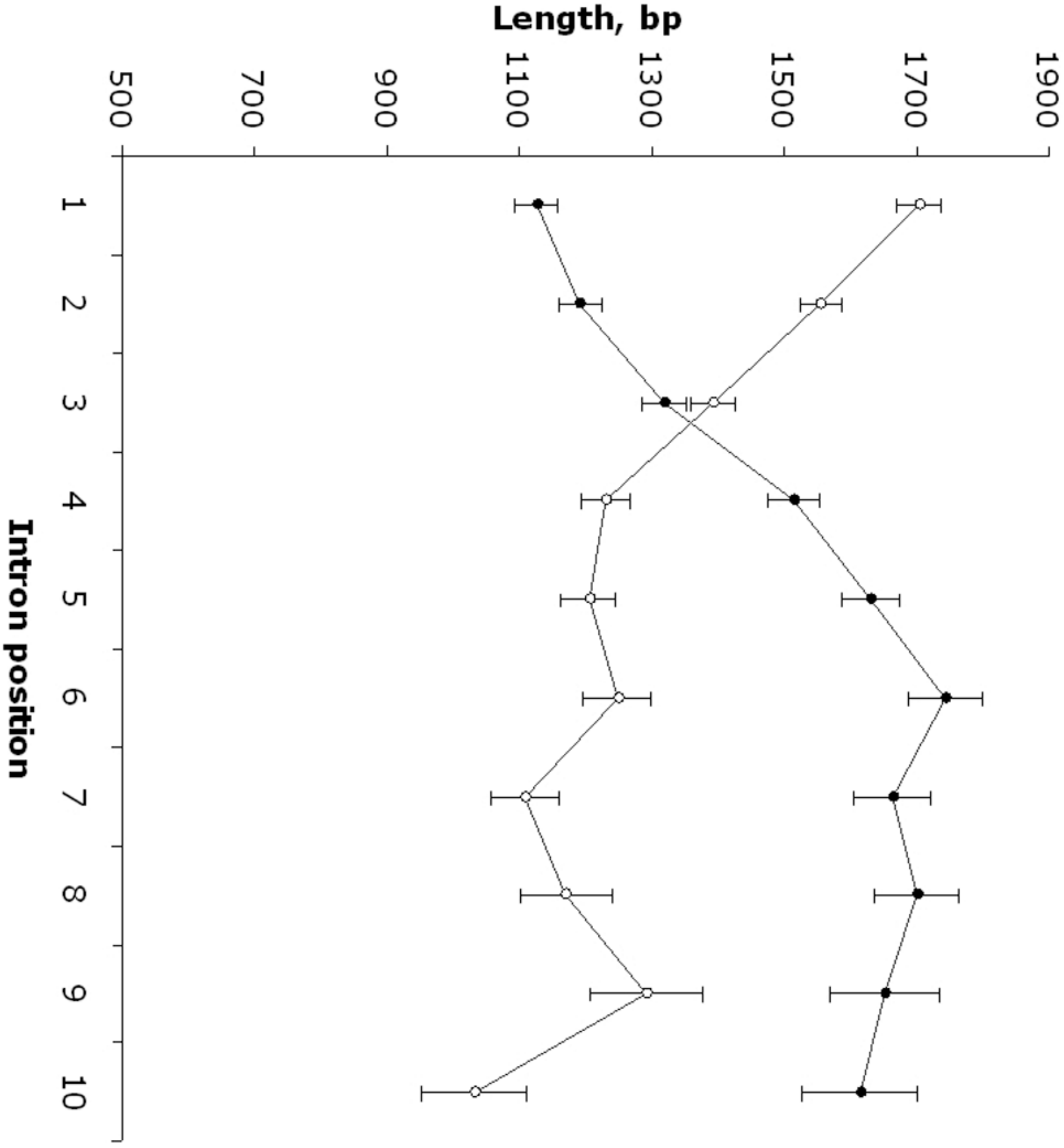

Figure 2. Intron size distribution

The length of introns varies according to their position in a transcript, counting from the 5′ end (solid circles) and the 3′ end (open circles). After approximately 5 introns, the length difference is no longer apparent due to the variation in the number of introns per transcript (See Supplementary Information).

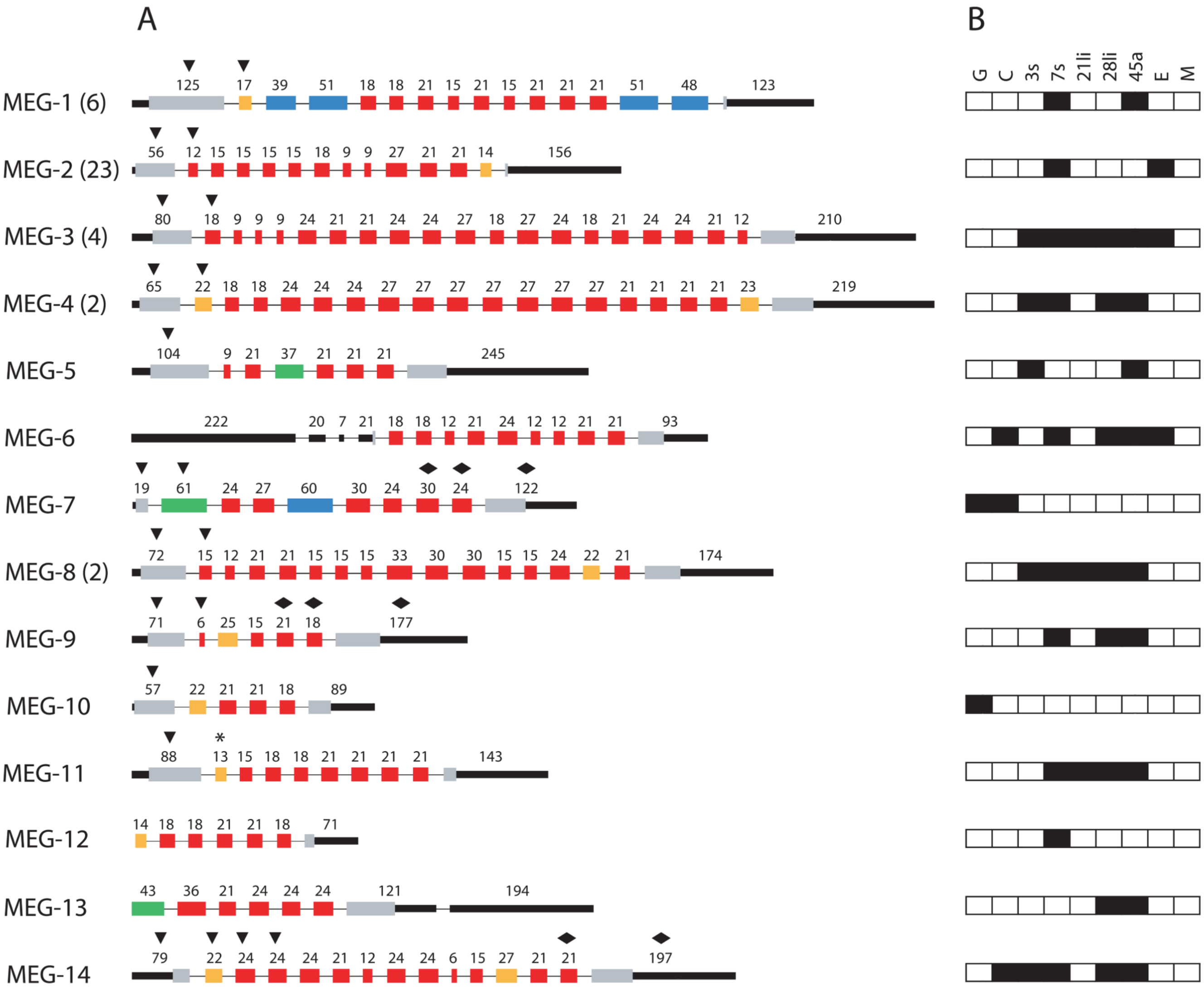

Figure 3. Schematic representation of gene structure from MEG family members

a, Structure of a representative member from each MEG family. Where multiple members were found, the total number detected is indicated in parentheses. Each box represents an exon drawn to scale and the number above it indicates the exon size in nucleotides. For illustrative purposes, the introns are shown with fixed length. Black triangles and diamonds indicate exons encoding predicted signal peptides and transmembrane helices, respectively. Other characteristics associated with exons are indicated by colour and grouped as follow: micro-exons having lengths of either multiples of 3 bp (red) or indivisible by 3 bp (orange); exons longer than 36 bp and having lengths of either multiples of 3 bp (blue) or indivisible by 3 bp (green); putative initiation and termination exons (grey); untranslated region (UTR) (black). Asterisk indicates exon deduced from transcript data, which did not match sequenced genome. MEG-12 and 13 structures were only partially predicted due to the lack of transcripts containing the 5′ end of these genes. b, RT-PCR or EST-based evidence of transcription (black box) for each family across different life cycle stages (G, C, E and M:

germball, cercaria, egg and miracidium; 3s and 7s: 3- and 7-day schistosomula; 21li and 28li: 21- and 28-day liver worms; 45a: 45-day adult worm pairs).

Table 1

List of *S. mansoni* genes that match a human gene with currently marketed pharmaceuticals

Gene ID	Protein description	Potential drugs
Smp_005210	histone deacetylase 1 (SmHDAC1)	Vorinostat[B]
Smp_009030	ribonucleoside-diphosphate reductase, alpha subunit, putative	Fludarabine Phosphate[B]
Smp_012930	inosine-5-monophosphate dehydrogenase, putative	Mycophenolate Mofetil[A], Mycophenolic Acid[A], Ribavirin[C]
Smp_015020	Na+/K+ ATPase alpha subunit (SNaK1)	Digoxin[B], Digitoxin[B], Acetyldigitoxin[C], Deslanoside[D],
Smp_016780[1]	tubulin alpha chain, putative	Vinblastine[A], Colchicine[A], Vincristine[A],
Smp_022960	aldehyde dehydrogenase, putative	Disulfiram[B]
Smp_026560	calmodulin, putative	Bepridil[C]
Smp_030730[1]	tubulin beta chain, putative	Colchicine[A], Vinblastine[A], Vincristine[A], Albendazole[B], Mebendazole[B], Paclitaxel[B], Thiabendazole[B], Vinorelbine[B], Docetaxel[C],
Smp_040130	cyclophilin (Smp17.7)	Cyclosporine[A]
Smp_040790	cyclophilin B	Cyclosporine[A]
Smp_044440	alcohol dehydrogenase, putative	Fomepizole[A]
Smp_048430	thioredoxin reductase (TGR)	Auranofin[B]
Smp_050390	aldehyde dehydrogenase, putative	Disulfiram[B]
Smp_053220	aldo-keto reductase, putative	Tolrestat[A]
Smp_055890	ribonucleoside-diphosphate reductase small chain, putative	Hydroxyurea[A], Gemcitabine[B]
Smp_065120	deoxyhypusine synthase, putative	Ciclopirox[B]
Smp_069160	cyclophilin, putative	Cyclosporine[A]
Smp_079230	immunophilin FK506 binding protein FKBP12, putative	Pimecrolimus[A], Temsirolimus[A], Sirolimus[B], Tacrolimus[B]
Smp_093280	histone deacetylase 3 (SmHDAC3)	Vorinostat[B]
Smp_094810	cyclophilin E	Cyclosporine[A]
Smp_121920	vesicular amine transporter, putative	Rauwolfia Serpentina[B], Reserpine[B], Deserpidine[C], Rescinnamine[C], Alseroxylon[C]
Smp_135460	bifunctional dihydrofolate reductase-thymidylate synthase, putative	Pemetrexed[B], Flucytosine[B], Floxuridine[B], Capecitabine[B], Fluorouracil[B]
Smp_136300	tyrosine kinase 5	Dasatinib[B]
Smp_147050	ATP synthase alpha subunit vacuolar, putative	Tiludronate[C]
Smp_171580	aromatic amino acid decarboxylase, putative	Carbidopa[B]
Smp_173280	cyclophilin, putative	Cyclosporine[A]

[1] There are multiple copies of tubulin (alpha: Smp_027920, Smp_090120, Smp_103140; beta: Smp_192110, Smp_079960, Smp_079970, Smp_078040, Smp_035760).

[A] The confidence in the assignment of the target to the efficacy of the drug in a human. Direct and clear evidence that this interaction is primarily responsible for the therapeutic action of the drug.

[B] The confidence in the assignment of the target to the efficacy of the drug in a human. Direct and clear evidence that this interaction is one mechanism for the drug, other targets/mechanisms may also exist.

[C] The confidence in the assignment of the target to the efficacy of the drug in a human. Indirect or inferred evidence of the association of the drug, target and therapeutic action, although the exact mechanism is still speculative.