

Genomic organization of the *Schistosoma mansoni* aspartic protease gene, a platyhelminth orthologue of mammalian lysosomal cathepsin D

Maria E. Morales^{a,b}, Bernd H. Kalinna^c, Oliver Heyers^c, Victoria H. Mann^a,
Alexandra Schulmeister^c, Claudia S. Copeland^{a,b}, Alex Loukas^d, Paul J. Brindley^{a,b,*}

^aDepartment of Tropical Medicine, School of Public Health and Tropical Medicine, Tulane University Health Sciences Center, New Orleans, LA, 70112, USA

^bInterdisciplinary Program in Molecular and Cellular Biology, School of Public Health and Tropical Medicine, Tulane University Health Sciences Center, New Orleans, LA, 70112, USA

^cDepartment of Molecular Parasitology, Institute for Biology, Humboldt University Berlin, 10115 Berlin, Germany

^dDivision of Infectious Diseases and Immunology, Queensland Institute of Medical Research, Brisbane, Queensland 4029, Australia

Received 15 March 2004; received in revised form 6 May 2004; accepted 17 May 2004

Received by D. Finnegan

Abstract

Schistosomes are considered the most important of the helminth parasites of humans in terms of morbidity and mortality. Schistosomes employ proteolytic enzymes to digest host hemoglobin from ingested human blood, including a cathepsin D-like, aspartic protease that is overexpressed in the gut of the adult female schistosome. Because of its key role in parasite nutrition, this enzyme represents a potential intervention target. To continue exploration of this potential, here we have determined the sequence, structure and genomic organization of the cathepsin D gene locus of *Schistosoma mansoni*. Using the cDNA encoding *S. mansoni* cathepsin D as a probe, we isolated several positive bacterial artificial chromosomes (BAC) from a BAC library that represents an ~ 8-fold coverage of the schistosome genome. Sequencing of BAC clone 25_J_24 revealed that the cathepsin D gene locus was ~ 13 kb in length, and included seven exons interrupted by six introns. The exons ranged in length from 49 to 294 bp, and the introns from 30 to 5025 bp. The genomic organization of schistosome cathepsin D was similar in sequence, structure and complexity to human cathepsin D, including to a greater or lesser extent the conservation of all six exon/intron boundaries of the schistosome gene. It was less similar to aspartic protease genes of the nematodes *Caenorhabditis elegans* and *Haemonchus contortus*, and dissimilar to those of plasmepsins from malarial parasites. Examination of the introns revealed the presence of endogenous mobile genetic elements including SR2, the ASL-associated retrotransposon, and the SINE-like element, SM α . Phylogenetically, schistosome cathepsin D appeared to be more closely related to mammalian cathepsin D than to other sub-families of eukaryotic aspartic proteases known from mammals. Taken together, these features indicated that schistosome cathepsin D is a platyhelminth orthologue of mammalian lysosomal cathepsin D.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Cathepsin D; Schistosome; Aspartic protease; Phylogeny; Exon; Intron; Orthologue; Parasite; Hemoglobin digestion; Retrotransposon

1. Introduction

Aspartic proteases are considered the most conserved of the protease groups, with eukaryotic aspartic proteases likely

Abbreviations: BAC, bacterial artificial chromosome; DIG, digoxigenin; EST, expressed sequence tag; LTR, long terminal repeat; ORF, open reading frame; SR2, schistosome retrotransposon 2; HSP70, heat shock protein 70 kDa; ASL, adenylysuccinate lyase.

* Corresponding author. Department of Tropical Medicine, SL-17, Tulane University, Health Sciences Center, 1430 Tulane Avenue, New Orleans, LA 70112, USA. Tel.: +1-504-988-4645; fax: +1-504-988-6686.

E-mail address: paul.brindley@tulane.edu (P.J. Brindley).

having arisen from an ancient gene duplication event that is now reflected by a structure composed of two homologous lobes and, at least in vertebrate homologues, by the presence of residues encoded by four exons on each of the lobes (Tang et al., 1978; Holm et al., 1984; Redecker et al., 1994). The active site of eukaryotic aspartic proteases is formed by a dyad of aspartic acid residues, one from each of the homologous lobes of the enzyme. The catalytic dyad residues coordinate the attacking water molecule; carboxyl groups of the Asp residues mediate a proton from the water molecule to the leaving nitrogen atom of the peptide substrate, initiating cleavage of the scissile bond (see Barrett et

al., 1998). According to the Merops classification (<http://www.merops.sanger.ac.uk/>), cathepsin D belongs to the Clan AA proteases (i.e., aspartic proteases), peptidase family A1 (pepsin family), and is numbered as peptidase A01.009. Cathepsin D is one of the eight discrete sub-families of the peptidase family A1 characterized from mammals, along with pepsin, gastricin, chymosin, renin, cathepsin E, napsin, and memapsin (= β -secretase). In human tissues, cathepsin D (EC 3.4.23.5) contributes to lysosomal proteolysis, processing of at least some antigens for the MHC class II system, and other functions (see Conner, 1998). Cathepsin D knockout mice develop normally during the first 2 weeks, stop thriving in the third week and die in a state of anorexia at about day 26 (Saftig et al., 1995).

Aspartic proteases occur also in invertebrates where, among other roles, they participate in digestive processes in the gut of nematodes (Longbottom et al., 1997; Tcherpanova et al., 2000; Williamson et al., 2002) and platyhelminths (Brindley et al., 2001; Brinkworth et al., 2001). With regard to parasites, aspartic proteases play key roles in the degradation of hemoglobin obtained from ingested or parasitized erythrocytes. In particular, aspartic proteases termed plasmepsins play a key role in hemoglobin proteolysis in the digestive vacuole of the intraerythrocytic stages of the malaria parasite, *Plasmodium falciparum*, and represent potential targets for the development of next generation anti-malarial drugs (e.g., see Banerjee et al., 2002). Aspartic proteases also play central roles in the degradation of ingested blood and hemoglobin in blood-feeding parasitic worms, and sites of cleavage within hemoglobins by both hookworm and schistosome aspartic proteases have been mapped (Brindley et al., 2001; Brinkworth et al., 2001; Williamson et al., 2002). The substrate specificity of these enzymes for mammalian hemoglobins is hypothesized to be associated with evolution of host species range and specificity (Williamson et al., 2002).

Focusing on the cathepsin D-like protease involved in digestion of hemoglobin within the gut of *Schistosoma mansoni* (Brindley et al., 2001), we report here the isolation of the gene locus from a bacterial artificial chromosome (BAC) library of the *S. mansoni* genome. The protease is encoded by seven exons interrupted by six introns spanning ~ 13 kb of the genome. Further, we have compared and contrasted the genomic structure of this invertebrate orthologue of human cathepsin D to human cathepsin D, to aspartic proteases of other helminths, and to plasmepsins from *P. falciparum*, that likewise are involved in digestion of human hemoglobin.

2. Materials and methods

2.1. Screening bacterial artificial chromosomes

Le Paslier et al. (2000) described the construction and characterization of a BAC library of the *S. mansoni* genome.

The library, constructed in the BAC plasmid vector, pBelo-Bac11, using gDNA from cercariae of a Puerto Rican strain of *S. mansoni* partially digested with *HindIII*, consists of ~ 21,000 clones, with a mean insert size of ~ 100 kb, providing a ~ 8-fold coverage of the schistosome genome. A cathepsin D specific probe was produced by PCR amplification using high-fidelity polymerase (Platinum *Taq*; Invitrogen) with an expression plasmid containing the cDNA encoding the proenzyme of *S. mansoni* cathepsin D (Brindley et al., 2001) (GenBank U60995) as a template, and 5'-GGGGTACCTGAAGTGGTTAGGATCCCTCT-3' (forward) and 5'-GGGGTACCAACTTCATCTGAAA-GAA-3' (reverse) primers, with restriction sites for *KpnI* added to facilitate cloning into plasmid vectors. The PCR-amplified cathepsin D gene probe of ~ 1.2 kb was isolated from an agarose gel, labeled with digoxigenin (DIG) using the DIG Labeling System from Roche (Indianapolis, IN), and used to screen the BAC library. Hybridizations of the probe with high density nylon filters containing ~ 21,000 BAC clones were carried out at 42 °C overnight in hybridization solution (DIG Labeling System; Roche), after which the filters were washed at high stringency at 65 °C for 30 min in 0.5 × SSC, 0.1% sodium dodecyl sulfate. Visualization of the DIG label was accomplished by incubation with an anti-DIG-alkaline phosphatase conjugated antibody (Roche) after which disodium 3-(4-methoxy-spiro(1,2-dioxetane-3,2'-(5'-chloro)tricyclo[3,3.1.1^{3,7}]decane)-4-yl) phenyl phosphate (CSPD) was used as the substrate for the development of the chemiluminescence. Hybridization was detected using X-ray film. Positive clones were cultured as described (Le Paslier et al., 2000), after which BAC plasmid DNAs were isolated and purified (Plasmid Mini Kit; Qiagen, Valencia, CA).

2.2. Southern hybridizations

BAC plasmid clones were digested with restriction enzymes, separated through 0.8% agarose gel by electrophoresis, transferred to nylon (Zeta-Probe) (Bio-Rad, Hercules, CA) by capillary action, and cross-linked to the membrane by UV irradiation. Hybridization of the cathepsin D probe to the Southern blots and development of signals were accomplished as described in Section 2.1.

2.3. Restriction analysis of BAC clones; Nucleotide sequencing; Long PCR

Positive sub-clones of BAC clone 25_J_24 digested with *HindIII* were cloned into plasmid pNEB193 (New England Biolabs, Beverly, MA). The inserts of clones p1.5, p1.8, p1.30, pB.1 and pB.2 were sequenced, and these sequences analyzed by Blastn. Blast results for clones p1.5, p1.8, pB.1 and pB.2 indicated that these clones included the 5'-end of the gene and the first four exons of the schistosome cathepsin D gene locus. Clone p1.30 Blast results indicated the presence of the exon 7

and the 3'-UTR region of the genomic structure of cathepsin D. The region of BAC 25_J_24 not represented in these sub-clones, which turned out to be the central region of the gene locus (see below), was amplified by long PCR using a 5'-oligonucleotide primer based on the insert sequence of p1.5 and a 3'-primer based on the sequence of p1.30. Long PCR was performed in 50.0 µl reaction volumes containing 20 ng of BAC 25_J_24, 5.0 µl of 10 × reaction buffer, 0.2 mM of each dNTP, 1.5 mM MgCl₂, 1 U of Platinum *Taq* polymerase (Invitrogen, Carlsbad, CA), and oligonucleotide primers 5'-CTTAC-TATGCATACCTCGTG-3' (from clone p1.5, below) and 5'-CATAGGGATCTGTAAAGCTG-3' (from clone p1.30, below). The samples were heated at 94 °C for 1 min followed by 30 cycles of amplification at 94 °C 30 s, 50 °C for 30 s, 68 °C for 6 min, and followed by a final extension step at 68 °C for 10 min. Amplified products were sized by electrophoresis through agarose gels, isolated from the gels, and ligated into plasmid pCR-XR-TOPO (Invitrogen). An amplicon of ~ 14 kb was cloned and the recombinant plasmid named pJ24_LP. The nucleotide sequence of the insert of pJ24_LP was determined using the vector specific primers M13 forward (– 20) and M13 reverse, and insert specific primers, using BigDye Terminator (Applied Biosystems [ABI]) chemistry and ABI Prism 3100 sequencers.

2.4. Bioinformatics analyses

Contiguous sequences (Contigs) and multiple sequence alignments were assembled with assistance from the Jellyfish software package (LabVelocity, San Francisco, CA). Bioinformatics analysis of the gene locus using Blast and other on-line tools was employed to determine exon and intron structures and boundaries of the schistosome protease gene, upstream and downstream regulatory sequences, and identities of unrelated sequences residing in the introns of the gene. The positions of splice sites were predicted using the NetGene2 program, <http://www.cbs.dtu.dk/services/NetGene2/>, according to Hebsgaard et al. (1996). A structural alignment was created using the Swiss Model server <http://www.expasy.org/swissmod/SWISS-MODEL.html> First Approach mode comparing *S. mansoni* cathepsin D with aspartic proteases where the crystal structure had been determined. Model sequences included cardosin A from the flowers of *Cynara cardunculus* (pdb accession **1B5F**), pepsinogen from *Sus scrofa* (**2PSG**) and human cathepsin D (**ILYA**). Structural alignments and models were viewed using the Swiss pdb Viewer <http://www.expasy.org/spdbv/>. Secondary structural features predicted for the homology model of the three dimensional structure of *S. mansoni* cathepsin D were used to determine whether intron–exon boundaries occurred in structurally complex regions, α-helices or β-sheets, of the *S. mansoni* cathepsin D open reading frame. Predictions of the cleavage point of signal peptides

from propeptide domains of aspartic protease were accomplished with the signalP algorithm, www.cbs.dtu.dk/services/SignalP.

2.5. Phylogenetic analysis

To investigate the phylogenetic relationship of the *S. mansoni* cathepsin D like protease with other eukaryotic aspartic proteases, phylograms were generated from regions homologous to the conserved aspartyl protease domain, Pfam00026, Eukaryotic Aspartyl Protease, conserved among pepsins, cathepsins D and E, renins, and other eukaryotic proteases. These conserved sequences ranged in size from 275 to 370 amino acid residues. Regions of 40 discrete aspartic proteases aligning with Pfam00026 were extracted from the public databases. Alignments of these amino acid sequences and generation of a bootstrapped phylogenetic tree (Brocchieri, 2001) were accomplished using CLUSTALX (Thompson et al., 1997) and Njplot (Saitou and Nei, 1987) software. Positions containing gaps were excluded from the analysis. The tree was rooted by using nucellin, an aspartic protease from flowering plants (Chen and Foolad, 1997), as the outgroup. Branch length ratios were preserved upon transfer into PowerPoint for display. Sequences of aspartic proteases used in the phylogenetic analysis were obtained from the GenBank, EMBL and PIR databases, as follows: *S. mansoni*, U60995; *Schistosoma japonicum*, L41346; *Necator americanus*, CAC00543; *Ancylostoma ceylanium*, AAO22152; *Caenorhabditis elegans*, AAC02571; *Drosophila melanogaster*, AAF23824; *Anopheles gambiae*, XP_307784; *Aedes aegypti*, Q03168; *Silurus asotus*, AAM62283; *Gallus gallus*, Q05744; *Mus musculus*, NP_034113; *Rattus norvegicus*, NP_599161; *Homo sapiens*, AAP36305; *Bos taurus*, BAB21620; *Sus scrofa*, **KHPGD**; *Onchocerca volvulus*, AAD00524; *C. cardunculus*, CAB40134; *Brassica napus*, U55032; *Oryza sativa*, D32165; *R. norvegicus*, P08424; *H. sapiens*, NP_004842; *Blatella germanicum*, P54958; *P. falciparum*, P39898; *P. falciparum*, CAA15605; *Eimeria tenella*, AJ293830; *Trematomus bernachii*, CAA69878; *Xenopus laevis*, BAC57453; *H. sapiens*, NP_055039; *R. norvegicus*, NP_579818; *R. norvegicus*, NP_068521; *H. sapiens*, NP_055039; *B. taurus*, **CMBO**; *Brugia malayi*, BAC05688; *Strongyloides stercoralis*, AAD09345; *Haemonchus contortus*, AF079402; *C. elegans*, AAC02571; *Aspergillus awamori*, M34454; *Candida albicans*, AAA34369; *Saccharomyces cerevisiae*, P32329; *M. musculus*, NP_035922, and nucellin-like aspartic protease from *Arabidopsis thaliana*, NP_177872.

2.6. GenBank accession number

The genomic sequence of the *S. mansoni* cathepsin D-like aspartic protease gene locus has been assigned accession number AY309267.

3. Results

3.1. *Schistosoma aspartic protease is encoded by a single gene*

When the BAC library of Le Paslier et al. (2000) of ~ 21,000 clones was screened with the labeled *S. mansoni* aspartic protease encoding cDNA U60995), six positive clones were identified; BAC clones 14_M_9, 23_M_1, 25_J_24, 29_M_15, 28_M_23 and 36_M_18 (not shown). Since this BAC library represents ~ 8-fold coverage of the schistosome genome (Le Paslier et al., 2000), this finding suggested that a single copy gene encoded this transcript and supports earlier findings using Southern hybridization analysis that a single gene encodes this schistosome aspartic protease (Becker et al., 1995, Wong et al., 1997).

3.2. *Cathepsin D gene locus is organized as seven exons interrupted by six introns spanning ~ 13 kb*

As a representative of the positive clones, and because there appeared to be just a single gene encoding the cathepsin D-like aspartic protease, we examined in detail one of the positive BACs, 25_J_24. In brief, five *Hind*III fragments of BAC 25_J_24 that cross-hybridized with the *S. mansoni* gene probe were sub-cloned into pNEB193, and named clones p1.5, p1.8, p1.30, pB.1 and pB.2. The size of inserts of these clones along with Blastn results obtained when they were used to query the GenBank nr

database indicated that p1.5 (insert size, ~ 4.5 kb), p1.8 (6 kb), pB.1 (4.5 kb) and pB.2 (4.5 kb) contained similar or related sequences representing the 5'-region of the cathepsin D gene locus regions upstream of the cathepsin D coding sequence. By contrast, p1.30 of ~ 7 kb represented the 3'-region of the gene along with downstream sequences. In addition, using long PCR, and primers based on the inserts of p1.5 and p1.30, a fragment of BAC 25_J_24 of ~ 14 kb was amplified and cloned into pCR-TOPO-XL as pJ24_LP.

From the nucleotide sequences of these cloned fragments, a contiguous sequence (contig) of 18,576 bp was assembled and assigned GenBank accession AY309267. The contig extended 2470 bp upstream of the translation start site and 3954 bp downstream of the translation stop site. By Blastn, this contig matched accession AJ318869, a 969-bp fragment of genomic DNA that spans exon 1 and the promoter region immediately upstream of the *S. mansoni* cathepsin D gene (to be reported elsewhere; Schulmeister and others, manuscript in preparation) and also matched the *S. mansoni* cDNA, U60995. By comparing the contig genomic sequence with the cDNA, and by predictions of splice sites, a gene structure of seven exons interrupted by six introns was determined for the *S. mansoni* cathepsin D-like aspartic protease (Fig. 1). The sizes of exons one to seven were 49, 142, 242, 203, 125, 243 and 294 bp, respectively (Table 1, Fig. 1). The introns ranged in size from 30 bp (intron 1) to 5025 bp (intron 6) (Table 1, Fig. 1). Notably, the exact 5'-end of the transcript encoding the cathepsin D had not been reported (Wong et

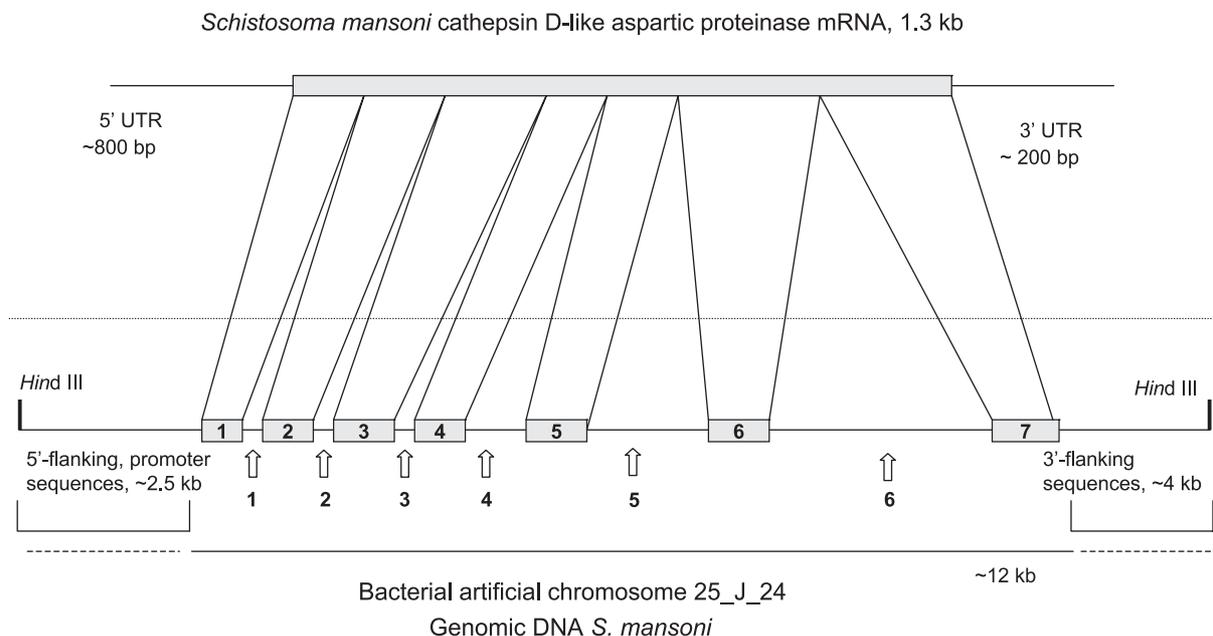


Fig. 1. Structure of the *S. mansoni* cathepsin D aspartic protease gene locus, as determined by nucleotide sequence analysis of bacterial artificial chromosome (BAC) clone no. 25_J_24 (Le Paslier et al., 2000). The top panel shows the size of the cDNA while the bottom panel shows a schematic of positions and relative sizes of the 7 exons and 6 introns (arrowed) that comprise the gene. Positions of the 5'- and 3'-*Hind* III at the boundaries of the genomic fragment are also shown, along with upstream and downstream flanking regions.

Table 1
Schistosoma mansoni cathepsin D like, aspartic protease gene locus sizes and identities of exons, introns and flanking regions

Structure	Nucleotide position in sequence of AY309267	Length (bp)	Identification	Blast match to accession number(s)
Upstream flanking region	1–1647	2470	Adenylosuccinate Lyase (ASL) <i>S. mansoni</i> repeat sequence <i>SR2</i> retrotransposon	AF448820 D87492
Promoter region	1648–2470	832	5' UTR	AJ318869
Exon 1	2471–2519	49	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 1	2520–2550	30	No matches	–
Exon 2	2551–2693	142	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 2	2694–2729	33	No matches	–
Exon 3	2728–2970	242	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 3	2971–3008	39	No matches	–
Exon 4	3009–3212	203	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 4	3213–4849	1636	No matches	–
Exon 5	4850–4975	125	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 5	4976–9060	4084	<i>SR2</i> subfamily A non-LTR retrotransposon <i>S. mansoni</i> HSP70 <i>SMα</i> SINE element (Sm7 satellite)	AF025672 L02415 AF036744
Exon 6	9061–9304	243	<i>S. mansoni</i> aspartic protease mRNA	U60995
Intron 6	9305–14,328	5025	<i>S. mansoni</i> ASL <i>S. mansoni</i> ASL long terminal repeat retrotransposon	AF448820 AF448821
Exon 7	14,329–14,623	294	<i>S. mansoni</i> aspartic protease mRNA	U60995
3' UTR and downstream flanking region	14,624–18,576	3954	No matches	–

The contiguous 18,576 bp of genomic sequence of the gene locus has been assigned Genbank accession AY309267.

al., 1997) U60995). The genomic sequence of exon 1 revealed that the deduced full-length preproenzyme was just two amino acids longer than partial sequence of Wong et al. (1997), with Met and Leu as first and second amino terminal residues (Fig. 2).

3.3. Splice sites; identities of intron sequences

The splice donor and acceptor sites of all six introns in the *S. mansoni* cathepsin D gene conformed to the established consensus GT at the 5'-end of the intron and AG at the 3'-end (Senapathy et al., 1990) (Table 2). The ~ 17 kb of non-coding region of the *S. mansoni* cathepsin D-like aspartic protease gene was analyzed by blastn and blastx, which revealed that several endogenous retrotransposable elements have come to reside in the introns and flanking sequences of this gene. These mobile elements included the *SR2* non-long terminal repeat (LTR) retrotransposon (Drew et al., 1999) and the *SMα* SINE like element (Ferbeyre et al., 1998) within intron 5, and the adenylosuccinate lyase (ASL)-associated LTR retrotransposon (Foult et al., 2002) within intron 6 and also within the upstream flanking regions of the gene. In addition, fragments of heat shock protein 70 kDa (HSP70) and ASL coding sequences were also evident within the intron sequences (introns 5 and 6), a situation that likely had resulted from movement of these sequences associated with the mobilization of retrotransposons (Brindley et al., 2003).

3.4. Conservation of intron structure between schistosome and human cathepsins D

The aspartic protease genes of vertebrates all share a conserved 9-exon gene structure, including two homologous clusters of four exons (Holm et al., 1984). In human cathepsin D, which is representative and characteristic of vertebrate aspartic proteases (Redecker et al., 1994), exons 2–5 are homologous to exons 6–9. The active site DTG motif is located in exon 3 and in exon 7. By contrast, *S. mansoni* cathepsin D has a seven-exon structure, with the active site DTGs in exons 3 and 6 (Fig. 2). When the exon structure and the exon/intron boundaries of *S. mansoni* cathepsin D were compared with those of human cathepsin D, remarkable similarities were evident in spite of the reduced number of exons in the schistosome gene. Specifically, all of the six exon/intron boundaries of schistosome cathepsin D were conserved to a greater or lesser degree in the human enzyme. As illustrated in Fig. 2, red (dark) stars highlight the three exon/intron boundaries that are absolutely conserved while yellow (light) stars locate the three in similar, though not identical, positions. Specifically, the absolutely conserved exon/intron boundaries were at introns 1, 2 and 4, and those with similar boundaries at introns 3, 5 and 6 (schistosome gene numbering).

To reiterate, allowing for the inclusion of the two additional introns in the human cathepsin D genes, i.e., introns 3 and 7 (marked by blue circles in Fig. 2), the exon/

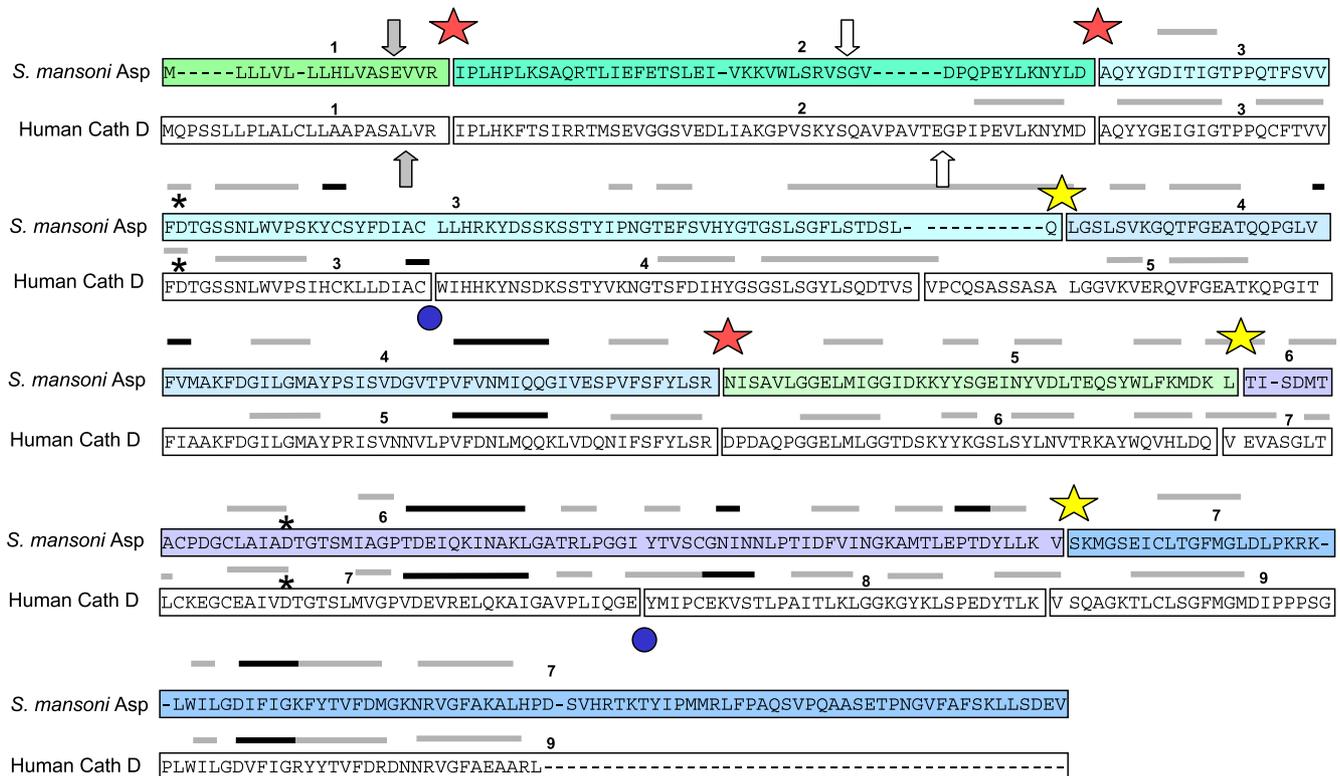


Fig. 2. Comparison of the exon/intron structures of *S. mansoni* cathepsin D (*S. mansoni* Asp) and human cathepsin D (Human cath D), using the system of Jean et al. (2001a). The two enzymes were aligned for maximal homology and the sequence corresponding to an exon for each protein was separated and boxed. Exon numbering is shown above the boxes. The position of the active site dyad of DTG motifs is indicated with asterisks (*). Red (dark) stars indicate where intron positions are conserved between the two enzymes, while yellow (light) stars indicate where introns occur in similar, though not exactly the same, positions. Blue circles indicate positions of introns in the human cathepsin D gene that are absent from the schistosome cathepsin D gene. The secondary structure of the schistosome enzyme, as predicted by Swiss Model First Approach mode, is indicated above the primary amino acid sequence, where gray bars indicate beta sheets and black bars indicate alpha helices. In like fashion, the secondary structure of human cathepsin D is indicated above the amino acid sequence. The position of cleavage of signal peptides is indicated with gray arrows, and the position of cleavage of the propeptides from mature enzymes indicated with white arrows.

intron boundaries are more or less exactly conserved at all of the 6 exon/intron boundaries in the by *S. mansoni* cathepsin D gene. The obvious difference in structure between these two genes was the presence of two additional exons, exons 4 and 8, in the human cathepsin D. Exons 4 and 8 in the human gene did not have obvious orthologues in the schistosome gene (Fig. 2). As mentioned above, exons 2–5 and exons 6–9 are homologous to one another, and reflect a primordial gene duplication event during the evolution of eukaryotic aspartic proteases. In *S. mansoni*

cathepsin D, exons 2–4 were homologous to exons 5–7 (Fig. 2).

The signal peptides are encoded by exon 1 in both the schistosome and human enzymes, and the propeptide of each cathepsin, which is released to generate the mature active protease, is encoded by parts of exons 1 and 2 in similar fashion in both proteases (Fig. 2). Based on the molecular model of *S. mansoni* cathepsin D (not shown; available on request) constructed using the Swiss Model First Approach mode, we identified secondary structures

Table 2

Intron- and exon boundaries, splice acceptor and splice donor sites in the gene encoding *Schistosoma mansoni* cathepsin D-like aspartic protease

Exon	Exon length (bp)	Donor		Intron	Size	Acceptor	
		Exon	Intron			Intron	Exon
1	49	GGTTAG	gtaatttacag	1	30	agtaagacctag	GATCCC
2	143	CTTGAT	gtatgattttt	2	33	ttctgtctctag	GCTCAA
3	243	CTTCAG	gttggttatta	3	39	atatttaaacag	TTGGGC
4	204	CAGCAG	gtagtagtagt	4	1636	tttataatttag	GAATAT
5	126	GGACAA	gtaagctaaaa	5	4084	tccttcaaacag	GCTGAC
6	244	TTGAAG	gtaatttcggt	6	5025	attcgataacag	GTATCT
7	296						

including α -helices and β -sheets. The positions of these are indicated on Fig. 2. The topographical locations of the introns within the coding region of mature *S. mansoni* cathepsin D were mapped to β -sheets for introns 3 and 5 while the remaining introns were mapped to areas without defined secondary structures (loops) (Fig. 2). In like fashion, we inspected the secondary structure of human cathepsin D, *ILYA*. The topographical locations of the introns within the coding region of mature human cathepsin D were mapped to β -sheets for introns 4, 6, 7 and 8 while the remaining introns were in loops. In summary, both enzymes exhibited β -sheets at two homologous intron sites (see above), introns 3 and 5 in the *S. mansoni* locus and introns 4 and 6 in the human gene, and loops at introns 4 (schistosome) and 5 (human). The main difference in complex structure at intron insertion sites was at schistosome intron 6 (loop)/human intron 8 (β -sheet) (Fig. 2).

Next, we aligned the seven-exon/six-intron structure of *S. mansoni* cathepsin D with two nematode aspartic proteases,

AAC02571 from *C. elegans* and pepsinogen from *H. contortus* AF079402) to compare aspartic protease gene organization among several helminths (Fig. 3). *C. elegans* has at least six aspartic protease genes, with varying exon numbers (one to six) (Tcherepanova et al., 2000). AAC02571 of *C. elegans* has a seven-exon/six-intron structure like *S. mansoni* cathepsin D. *H. contortus* pepsinogen has nine exons (Longbottom et al., 1997; Jean et al., 2001a). As illustrated in Figs. 3 and 4, there was little or no apparent agreement in the exon/intron relationships among these helminth aspartic protease genes, although two intron insertion sites appeared to be located in similar, but not identical, sites among these enzymes (Fig. 3; yellow [light] stars). Overall, however, no obvious evolutionary relationships among them were apparent. Indeed, despite the probable closer evolutionary distance between trematodes and nematodes than between trematodes and vertebrates, the *S. mansoni* cathepsin D gene structure was more similar to human cathepsin D than to the two aspartic proteases of nematodes that we examined here.

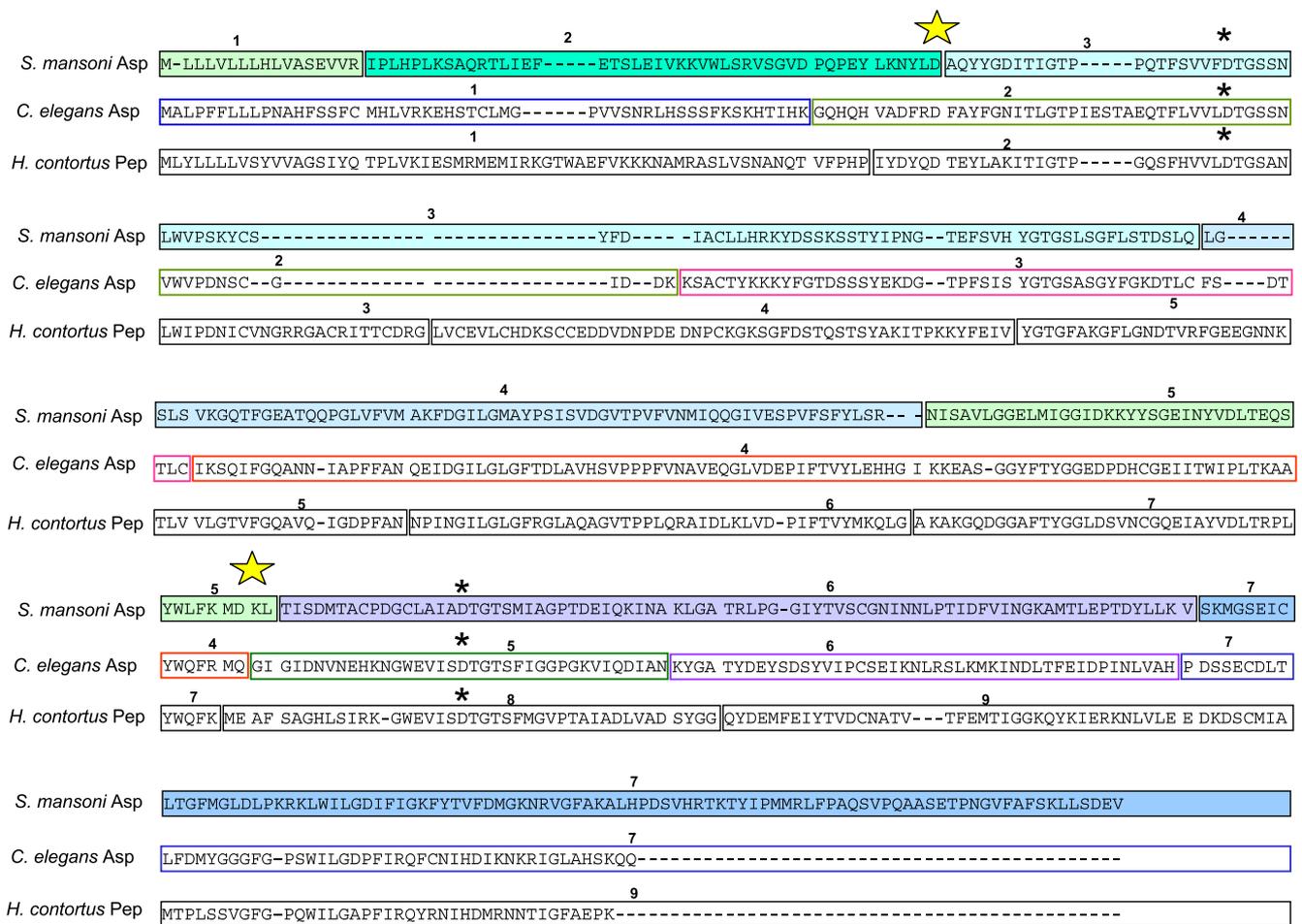


Fig. 3. Comparison of the exon/intron structures of *S. mansoni* cathepsin D AY309267), a *C. elegans* aspartic protease AAC02571), and pepsinogen Pep1 of *H. contortus* AF079402), using the system of Jean et al. (2001a). The three enzymes were aligned for maximal homology and the sequence corresponding to an exon for each protein was separated and boxed. Exon numbering is shown above the boxes. The position of the dual active site DTG motifs is indicated with asterisks (*). The position of the dual active site DTG motifs is indicated with asterisks (*). Yellow (light) stars indicate where introns occur in similar, though not exactly the same, positions.

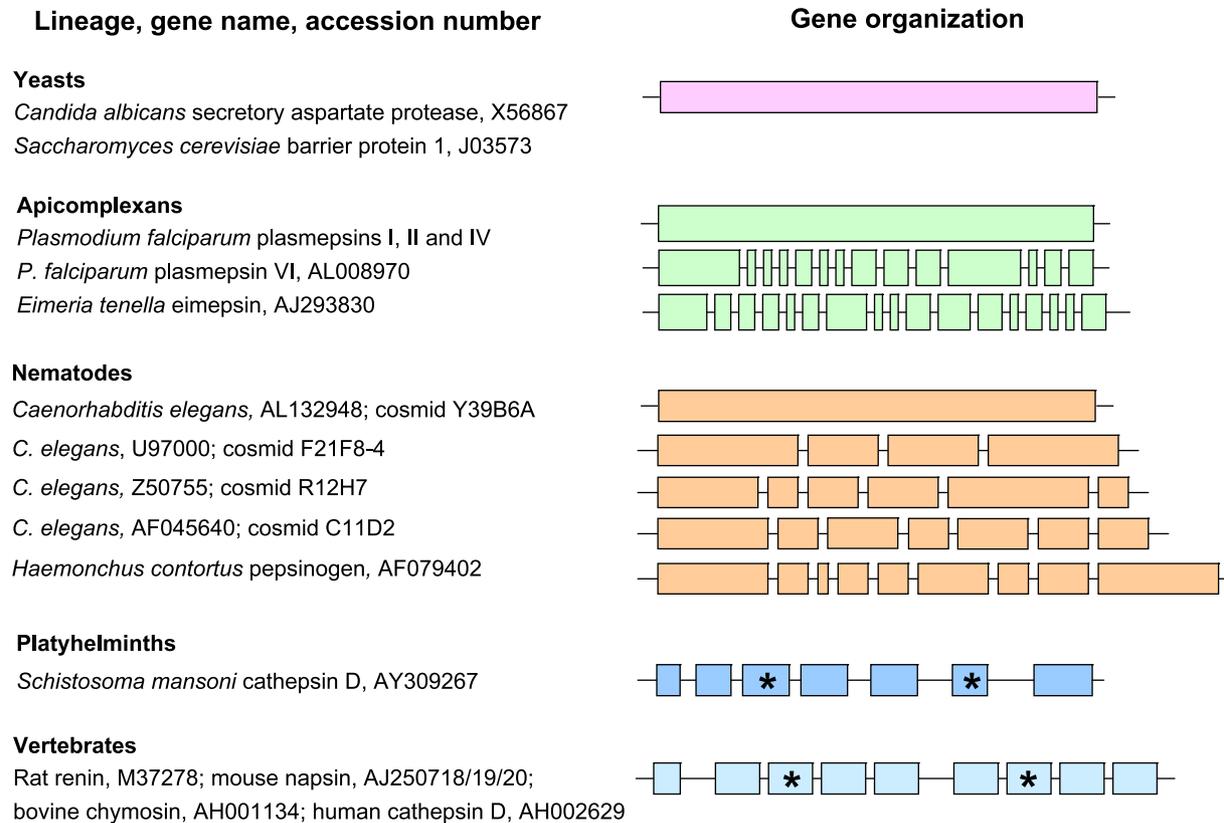


Fig. 4. Schematic to compare the structure and exon numbers of the aspartic proteinase family genes, including the cathepsin D gene of *S. mansoni* (Phylum Platyhelminthes) and genes from species of several other major taxa—yeasts, apicomplexans, nematodes, and vertebrates. Colored blocks represent exons, thin lines represent introns, and the asterisks on the schistosome and vertebrate blocks identify the exon that encodes the dyad of active site DTG residues. Database accession numbers are provided for the illustrated sequences. Additional information for this illustration was adapted from Jean et al. (2001a).

3.5. Phylogenetic relatedness of schistosome aspartic protease with mammalian cathepsin D

We constructed a phylogenetic tree from an alignment of the conserved regions of *S. mansoni* cathepsin D, with cathepsin D from other eukaryotes, other mammalian non-cathepsin D aspartic proteases, and from other informative aspartic proteases including plasmepsins, eimepsin, cardosin and nucellin. Signal peptides and carboxyl extensions and other elaborations were trimmed from the sequences to leave only the eukaryotic aspartic protease (Pfam00026) conserved domain. Nucellin served as the outgroup. *S. mansoni* cathepsin D occupied a clade with its congeneric orthologue from *S. japonicum* (Fig. 5). Together, these two *Schistosoma* enzymes clustered with a clade of enzymes from dipteran insects including cathepsin D from *D. melanogaster* and lysosomal aspartic protease from *A. aegypti* and with the large clade representing vertebrate cathepsin D. Of the eight major groupings of mammalian aspartic proteases, pepsin, gastrin, renin, chymosin, napsin, memapsin, and cathepsins D and E, schistosome cathepsin D clustered closest to mammalian cathepsin D. Other parasite aspartic proteases to which roles have been ascribed in the degradation of mammalian hemoglobin from parasitized or ingested red cells, including pepsinogen of *H. contortus*, the

barber's pole worm of sheep (Longbottom et al., 1997) and the malarial plasmepsins (see Banerjee et al., 2002) were located on branches more distant from the schistosome cathepsin D and distinct from the mammalian cathepsin D (Fig. 5). However, Na-APR-1, a cathepsin D-like protease from the human hookworm *N. americanus*, and its orthologue Ay-APR-1 from the zoonotic hookworm *A. ceylanicum*, enzymes that participate in hemoglobin digestion in the hookworm gut (Williamson et al., 2002), also clustered near the cathepsins D.

4. Discussion

4.1. Gene structure of schistosome aspartic protease

Schistosome cathepsin D is overexpressed within the gut of the adult female schistosome where it participates in digestion of hemoglobin released from ingested erythrocytes. While the presence of cathepsin D transcripts noted in other developmental stages indicates additional roles for this enzyme in the physiology of the schistosome (Verity et al., 1999; Hu et al., 2003), overexpression within the gut tissue of adult female schistosomes demonstrates that a primary function of this enzyme is in hemoglobin digestion (Brind-

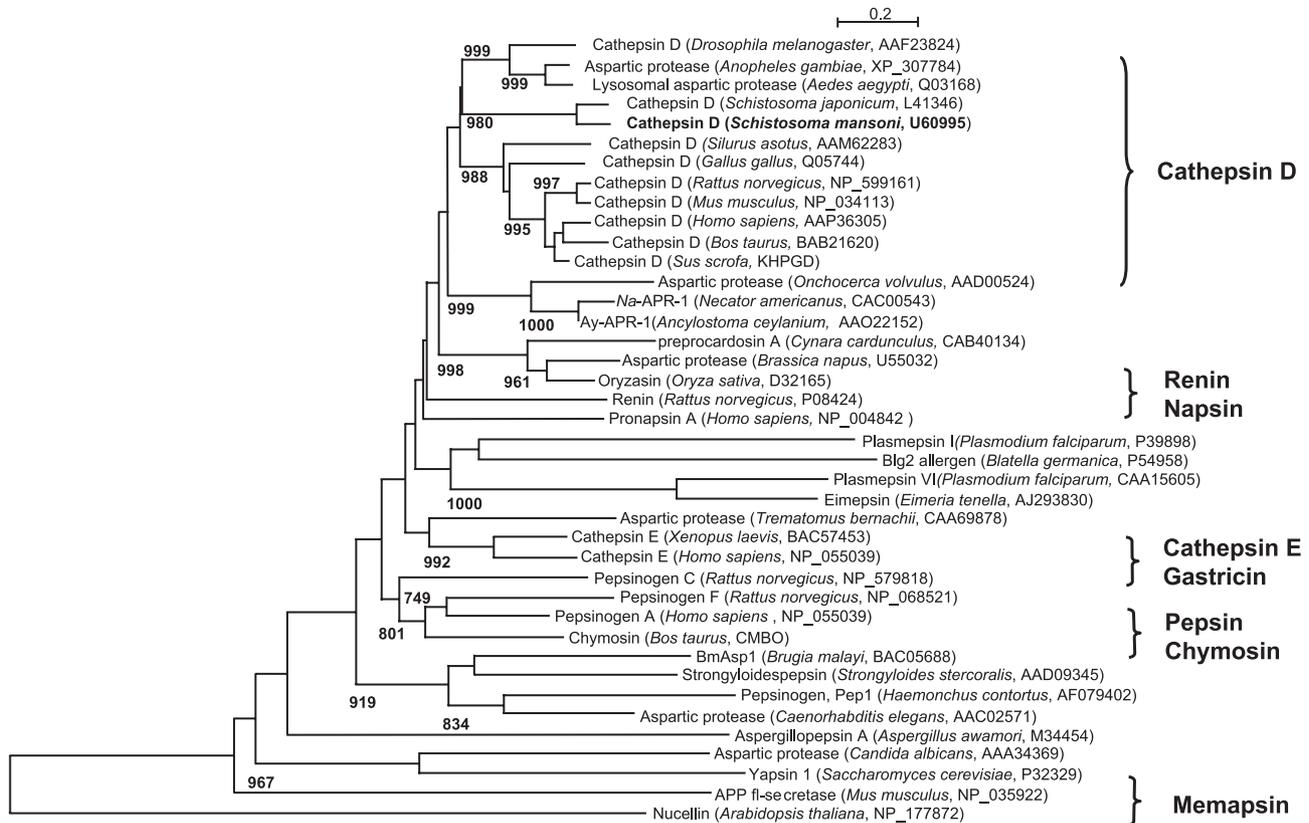


Fig. 5. Phylogenetic relationships of *S. mansoni* cathepsin D-like aspartic protease and other eukaryotic aspartic proteases. The bootstrapped tree was rooted using as the outgroup the phylogenetically distinct enzyme, nuclillin, an aspartic protease like protein involved in programmed cell death of cells of flowering plants (Chen and Foolad, 1997). The tree was constructed using CLUSTALX and NJplot, and subsequently imported into PowerPoint for annotation, where branch lengths were preserved. Bootstrap values for 1000 replicates are shown where values greater than 700 were obtained.

ley et al., 2001). The locus of the *S. mansoni* cathepsin D gene was located on BAC 25_J_24, spanning ~ 13 kb of the chromosome, and was constituted of seven exons interrupted by six introns. Only six BAC clones were positive when the entire genomic library of ~ 21,000 BACs was screened with the *S. mansoni* cathepsin D gene probe. Since this genomic library represents an ~ 8-fold coverage of the genome (Le Paslier et al., 2000) (haploid genome size, 270 MB), the result suggests that this enzyme is encoded by a single gene, a finding that supports earlier findings using Southern hybridizations that had indicated the presence of just a single gene (Becker et al., 1995; Wong et al., 1997).

The seven-exon/six-intron structure of *S. mansoni* cathepsin D gene is of similar complexity to aspartic proteases of some other eukaryotes (Fig. 4) (see Jean et al., 2001a). Vertebrate aspartic proteases display a nine-exon/eight-intron structure that is widely conserved among species of vertebrates and indeed among the major groupings of vertebrate aspartic proteases. The nine exons include two homologous groups each of four exons that resulted from a primordial gene duplication event and which encode each of the two homologous lobes of the three-dimensional structure of these enzymes (Holm et al., 1984). By contrast, the exon/intron structure of eukaryotic aspartic proteases in

non-vertebrate taxa is more diverse. The genome of *C. elegans* includes at least six aspartic proteases, with a range of exon structures (Tcherepanova et al., 2000), and the malarial parasite *P. falciparum* has at least 10 plasmepsins with diverse exon structures. Jean et al. (2001a) have noted that plasmepsins I, II and IV, which are involved in hemoglobin degradation, have a single exon structure whereas plasmepsin VI and eimepsin from the related apicomplexan *E. tenella* have much more complex exon/intron structures with 15–18 exons (Jean et al., 2001a,b). Jean et al. (2001a) also have suggested that the single exon format might be linked to a role in hemoglobin digestion. If so, this hypothesis cannot be broadened to cover hemoglobin-digesting enzymes from metazoan parasites since, as reported here, schistosome cathepsin D has a more complex exon/intron structure, as does pepsinogen from *H. contortus*, and both these aspartic proteases participate in degradation of hemoglobin from erythrocytes ingested by these parasitic worms.

4.2. A platyhelminth orthologue of mammalian cathepsin D

A comparison of the gene structure revealed that schistosome cathepsin D was more similar to human cathepsin D than it was to aspartic proteases from *C. elegans* or *H.*

contortus (Figs. 2–4). In addition, the phylogenetic analysis suggested that schistosome cathepsin D was more closely related in primary amino acid sequence to mammalian cathepsin D than to enzymes of the other major categories of vertebrate aspartic proteases such as pepsin, chymosin or renin (Fig. 5). Indeed, the similarity in exon/intron structure of schistosome cathepsin D and human cathepsin, including the retention of at least four exon/intron boundaries, reflected the close phylogenetic relationship between these two enzymes. These two enzymes are 52% identical and 69% similar at the amino acid level, by pairwise alignment (not shown), they exhibit remarkably similar secondary (Fig. 2) and three-dimensional structure (Brinkworth et al., 2001), and they share some cross-reactive epitopes (Valdivieso et al., 2003). On the other hand, they differ in their fine substrate specificity (Brindley et al., 2001), by the presence in the schistosome orthologue of a long COOH terminal extension (Fig. 2), and by the secretion and deployment into the parasite gut where schistosome cathepsin D performs a key role in digestion of ingested host hemoglobin (Brindley et al., 2001). (Mammalian cathepsin D probably digests hemoglobin also, in the context of catabolism and recycling of senescent erythrocytes within the spleen and liver.) Taken together, the similarity in genomic organization, including retention of most of the conserved vertebrate aspartic protease intron boundaries, the sequence identity, and the phylogenetic relatedness, indicated that this *S. mansoni* aspartic protease gene is a platyhelminth orthologue of mammalian lysosomal cathepsin D.

4.3. Introns in schistosome cathepsin D gene locus targeted by mobile genetic sequences

The sizes of the six introns of the schistosome cathepsin D gene ranged from 30 to 5025 bp. Introns as small as the diminutive introns 1 (of 30 bp), 2 (33 bp) and 3 (39 bp) have been long recognized in schistosome genes (Craig et al., 1989), as have much longer (>10 kb) examples (Bentley et al., 2003). The larger introns 4 (4084 bp) and 5 (5025 bp) in the *S. mansoni* cathepsin D gene locus have been targeted by mobile genetic elements including SR2, a non-LTR retrotransposon, the ASL-associated LTR retrotransposon, and the SINE like element, SM α (Drew et al., 1999; Foulk et al., 2002; Ferbeyre et al., 1998). In like fashion, the ASL-associated retrotransposon was identified in the 5'-flanking region of the gene. Mobilization of these kinds of elements to introns, rather than exons, likely carries much less danger from deleterious mutations for the host genome. Additionally, host genomes, including the schistosome genome, have adapted to the presence of mobile genetic elements in proximity to and indeed within host genes and appear to have incorporated these sequences into regulatory elements and other structures (e.g., see Brindley et al., 2003). We speculate that the presence of these retrotransposable elements, which occur in high copy number in the schistosome genome (Brindley et al., 2003), may have influenced the

evolutionary divergence of the organization of this platyhelminth aspartic protease gene away from the vertebrate 9-exon paradigm, by facilitating unequal crossing over (homologous recombination).

Only a few complete genomic structures have been reported previously for any schistosome genes (e.g., Craig et al., 1989; Neumann et al., 1992; Bentley et al., 2003). Since information about species-specific gene structure is valuable for training algorithms to detect the presence of genes (see Aggarwal et al., 2003), and since international efforts are currently underway to determine the entire genome sequences of *S. mansoni* and *S. japonicum*, the present information on the genome structure of schistosome cathepsin D should be of value also in these genomics endeavours aiming to fully sequence and annotate these parasite genomes.

Acknowledgements

We thank Dr. Philip LoVerde for provision of the BAC library. This work was supported in part by Infrastructure Grant no. ID-IA-0037-02 from the Ellison Medical Foundation to PJB and others, and grant no. KA 866/2-1 of the Deutsche Forschungsgemeinschaft to B.H.K. P.J.B. is a recipient of a Burroughs Wellcome Fund Scholar Award in Molecular Parasitology.

References

- Aggarwal, G., Worthey, E.A., McDonagh, P.D., Myler, P.J., 2003. Importing statistical measures into Artemis enhances gene identification in *Leishmania* genome project. BMC Bioinformatics, 4/23.
- Banerjee, R., Liu, J., Beatty, W., Pelosof, L., Klemba, M., Goldberg, D.E., 2002. Four plasmepsins are active in the *Plasmodium falciparum* food vacuole, including a protease with an active-site histidine. Proc. Natl. Acad. Sci. U. S. A. 99, 990–995.
- Barrett, A.J., Rawlings, N.D., Woessner, J.F., 1998. Chapter 270, Introduction: aspartic peptidases and their clans. In: Barrett, A.J., Rawlings, N.D., Woessner, J.F. (Eds.), Handbook of Proteolytic Enzymes. Academic Press, San Diego, p. 801.
- Becker, M.M., Harrop, S.A., Dalton, J.P., Kalinna, B.H., McManus, D.P., Brindley, P.J., 1995. Cloning and characterization of the *Schistosoma japonicum* aspartic proteinase involved in hemoglobin degradation. J. Biol. Chem. 270, 24496–24501.
- Bentley, G.N., Jones, A.K., Agnew, A., 2003. Mapping and sequencing of acetylcholinesterase genes from the platyhelminth blood fluke *Schistosoma*. Gene 314, 103–112.
- Brindley, P.J., Kalinna, B.H., Wong, J.Y.M., Bogitsh, B.J., King, L.T., Smyth, D.J., Verity, C.K., Abbenante, G., Brinkworth, R.I., Fairlie, D.P., Smythe, M.L., Milburn, P.J., Bielefeldt-Ohmann, H., Zheng, Y., McManus, D.P., 2001. Proteolysis of human hemoglobin by schistosome cathepsin D. Mol. Biochem. Parasitol. 112, 103–112.
- Brindley, P.J., Laha, T., McManus, D.P., Loukas, A., 2003. Mobile genetic elements colonizing the genomes of metazoan parasites. Trends Parasitol. 19, 79–87.
- Brinkworth, R.I., Prociv, P., Loukas, A., Brindley, P.J., 2001. Hemoglobin-degrading, aspartic proteases of blood-feeding parasites: substrate specificity revealed by homology models. J. Biol. Chem. 276, 38844–38851.

- Brocchieri, L., 2001. Phylogenetic inferences from molecular sequences: review and critique. *Theor. Popul. Biol.* 59, 27–40.
- Chen, F., Foolad, M.R., 1997. Molecular organization of a gene in barley which encodes a protein similar to aspartic protease and its specific expression in nucellar cells during degeneration. *Plant Mol. Biol.* 35, 821–831.
- Conner, G.E., 1998. Cathepsin D. In: Barrett, A.J., Rawlings, N.D., Woessner, J.F. (Eds.), *Handbook of Proteolytic Enzymes*. Academic Press, San Diego, pp. 828–836.
- Craig III, S.P., Muralidhar, M.G., McKerrow, J.H., Wang, C.C., 1989. Evidence for a class of very small introns in the gene for hypoxanthine–guanine phosphoribosyltransferase in *Schistosoma mansoni*. *Nucleic Acids Res.* 17, 1635–1647.
- Drew, A.C., Minchella, D.J., King, L.T., Rollinson, D., Brindley, P.J., 1999. SR2, non-long terminal repeat retrotransposons of the RTE-1 lineage, from the human blood fluke *Schistosoma mansoni*. *Mol. Biol. Evol.* 16, 1256–1269.
- Ferbyre, G., Smith, J.M., Cedergren, R., 1998. Schistosome satellite DNA encodes active hammerhead ribozymes. *Mol. Cell. Biol.* 18, 3880–3888.
- Foulk, B.W., Pappas, G., Hirai, Y., Hirai, H., Williams, D.L., 2002. Adenylosuccinate lyase of *Schistosoma mansoni*: gene structure, mRNA expression, and analysis of the predicted peptide structure of potential chemotherapeutic target. *Int. J. Parasitol.* 32, 1487–1495.
- Hebsgaard, S.M., Korning, P.G., Tolstrup, N., Engelbrecht, J., Rouze, P., Brunak, S., 1996. Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information. *Nucleic Acids Res.* 24, 3439–3452.
- Holm, I., Ollo, R., Panthier, J.J., Rougeon, F., 1984. Evolution of aspartyl proteases by gene duplication: the mouse renin gene is organized by two homologous clusters of four exons. *EMBO J.* 3, 557–562.
- Hu, W., Yan, Q., Shen, D.-K., Liu, F., Xu, X.-R., Zhu, Z.-D., Wu, X.-W., Zhang, X., Wang, J.-J., Xu, X., Wang, Z., Huang, J., Wang, S.-Y., Wang, Z.-Q., Brindley, P.J., McManus, D.P., Xue, C.-L., Feng, F., Chen, Z., Han, Z.-G., 2003. Evolutionary and biomedical implications of a *Schistosoma japonicum* complementary DNA resource. *Nat. Genet.* 35, 139–147.
- Jean, L., Long, M., Young, J., Pery, P., Tomley, F., 2001a. Aspartyl proteinase genes from apicomplexan parasites: evidence for evolution of the gene structure. *Trends Parasitol.* 17, 491–498.
- Jean, L., Pery, P., Dunn, P., Bumstead, J., Billington, K., Ryan, R., Tomley, F., 2001b. Genomic organisation and developmentally regulated expression of an apicomplexan aspartyl proteinase. *Gene* 262, 129–136.
- Le Paslier, M.C., Pierce, R.J., Merlin, F., Hirai, H., Wu, W., Williams, D.L., Johnston, D., LoVerde, P.T., Le Paslier, D., 2000. Construction and characterization of a *Schistosoma mansoni* bacterial artificial chromosome library. *Genomics* 65, 87–94.
- Longbottom, D., Redmond, D.L., Russell, M., Liddell, S., Smith, W.D., Knox, D.P., 1997. Molecular cloning and characterisation of a putative aspartate proteinase associated with a gut membrane protein complex from adult *Haemonchus contortus*. *Mol. Biochem. Parasitol.* 88, 63–72.
- Neumann, S., Ziv, E., Lantner, F., Schechter, I., 1992. Cloning and sequencing of an hsp70 gene of *Schistosoma mansoni*. *Mol. Biochem. Parasitol.* 56, 357–360.
- Redecker, B., Heckendorf, B., Grosch, H.W., Mersmann, G., Hasilik, A., 1994. Molecular organization of the human cathepsin D gene. *DNA Cell Biol.* 10, 423–431.
- Saftig, P., Hetman, M., Schmahl, W., Weber, K., Heine, L., Mossmann, H., Koster, A., Hess, B., Evers, M., von Figura, K., et al., 1995. Mice deficient for the lysosomal proteinase cathepsin D exhibit progressive atrophy of the intestinal mucosa and profound destruction of lymphoid cells. *EMBO J.* 14, 3599–3608.
- Saitou, N., Nei, M., 1987. The Neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* 4, 406–425.
- Senapathy, P., Sharp, M.B., Harris, N.L., 1990. Splice junctions, branch point sites and exons: sequence statistics, identification, and applications to genome project. *Methods Enzymol.* 183, 252–278.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The Clustal X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 24, 4876–4882.
- Tang, J., James, M.N., Hsu, I.N., Jenkins, J.A., Blundell, T.L., 1978. Structural evidence for gene duplication in evolution of acid proteases. *Nature* 271, 618–621.
- Tcherepanova, I., Bhattacharyya, L., Rubin, C., Freedman, J., 2000. Aspartic proteases from the nematode *Caenorhabditis elegans*. Structural organization and developmental and cell-specific expression of *asp-1*. *J. Biol. Chem.* 275, 26359–26369.
- Valdivieso, E., Bermudez, H., Hoebeke, J., Noya, O., Cesari, I.M., 2003. Immunological similarity between *Schistosoma* and bovine cathepsin D. *FEBS Letts.* 89, 81–88.
- Verity, C.K., McManus, D.P., Brindley, P.J., 1999. Developmental expression of the cathepsin D aspartic protease of *Schistosoma japonicum*. *Int. J. Parasitol.* 29, 1819–1824.
- Williamson, A.L., Brindley, P.J., Abbenante, G., Prociv, P., Berry, C., Girdwood, K., Pritchard, D.I., Fairlie, D.P., Hotez, P.J., Dalton, J.P., Loukas, A., 2002. Cleavage of hemoglobin by hookworm cathepsin D aspartic proteases and its potential contribution to host specificity. *FASEB J.* 16, 1458–1460.
- Wong, J.Y.M., Harrop, S.A., Day, S.R., Brindley, P.J., 1997. Schistosomes express two forms of cathepsin D. *Biochim. Biophys. Acta* 1338, 156–160.