Evolutionary histories of expanded peptidase families in *Schistosoma mansoni*

Larissa Lopes Silva¹,²,³, Marina Marcet-Houben⁴, Adhemar Zerlotini¹,², Toni Gabaldón⁴, Guilherme Oliveira¹,², Laila Alves Nahum¹,²/+ 

¹Grupo de Genômica e Biologia Computacional, Instituto de Pesquisas René Rachou, Instituto Nacional de Ciência e Tecnologia em Doenças Tropicais ²Centro de Excelência em Bioinformática-Fiocruz, Belo Horizonte, MG, Brasil ³Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil ⁴Bioinformatics and Genomics Programme, Centre de Regulació Genòmica, Universitat Pompeu Fabra, Barcelona, Spain

Schistosomiasis is one of the three main causative agents of human schistosomiasis, a major health problem with a vast socio-economic impact. Recent advances in the proteomic analysis of schistosomes have revealed that peptidases are the main virulence factors involved in the pathogenesis of this disease. In this context, evolutionary studies can be applied to identify peptidase families that have been expanded in genomes over time in response to different selection pressures. Using a phylogenomic approach, we searched for expanded endopeptidase families in the S. mansoni predicted proteome with the aim of contributing to the knowledge of such enzymes as potential therapeutic targets. We found three endopeptidase families that comprise leishmanolysins (metallopeptidase M8 family), cercarial elastases (serine peptidase SI family) and cathepsin D proteins (aspartic peptidase AI family). Our results suggest that the Schistosoma members of these families originated from successive gene duplication events in the parasite lineage after its diversification from other metazoans. Overall, critical residues are conserved among the duplicated genes/proteins. Furthermore, each protein family displays a distinct evolutionary history. Altogether, this work provides an evolutionary view of three S. mansoni peptidase families, which allows for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle.

Key words: phylogenomics - maximum likelihood analysis - homology prediction - functional annotation - proteases - paralogous families - parasite genomics

Schistosomiasis, which is caused by different species from the *Schistosoma* genus, remains one of the most prevalent tropical neglected diseases, affects 210 million people worldwide, and is responsible for at least 280,000 deaths every year (van der Werf et al. 2003, Steinmann et al. 2006, Han et al. 2009). *Schistosoma mansoni* is one of the three major species that infect humans and is the causative agent of intestinal and hepatic schistosomiasis mainly in Africa and South America (Han et al. 2009). Measures to control schistosomiasis rely almost entirely on praziquantel®, which is the only drug available for mass chemotherapy. Despite the effectiveness of this treatment, re-infection is common and drug-resistant parasites have been found in the laboratory and in the field, which demonstrate the urgent need to develop additional chemotherapeutic agents and effective vaccines (Liang et al. 2003, Pica-Mattoccia & Cioli 2004, Botros & Bennett 2007, Melman et al. 2009).

Over the past several years, advances in the molecular analysis of major parasites have identified some key factors involved in parasitic diseases and peptidases as one of the major factors of pathogenicity (McKerrow et al. 2006, Kasný et al. 2009). These enzymes have been implicated in processes that are crucial to the development and survival of helminth parasites, including digestion, invasion from host tissues, activation of inflammation and evasion of the host immune system (McKerrow et al. 2006, Kasný et al. 2009).

Peptidases (also termed proteases, proteinases or proteolytic enzymes) are hydrolytic enzymes that cleave peptide bonds in proteins. Endopeptidases cleave internal peptide bonds, whereas exopeptidases hydrolyse the amino terminus (aminopeptidases) or carboxy terminus (carboxypeptidases) of different proteins. Enzymatic specificity is determined based on the chemical groups responsible for catalysis in the peptide’s active site. Thus, peptidases are classified into one of the following classes: asparagine, aspartic, cysteine, glutamic, metallo, serine, threonine and unknown peptidases (Rawlings & Barrett 1993, Rawlings et al. 2010).

Asparagine peptidases are enzymes that have active sites composed of an aspartic acid and an asparagine, the latter being the PI residue, the amino acid or molecule, which can be found at a specific location in the cleavage site (Rawlings et al. 2010). In turn, aspartic peptidases have their catalytic centres formed by two aspartate residues that activate a water molecule that mediates the nucleophilic attack on the peptide bond (James 2004, Rawlings...
et al. 2010). In general, cysteine peptidases have cysteine and histidine residues forming their “catalytic dyad”. Meanwhile, other active site residues have been found. Glutamic peptidases have glutamic acid residues as their primary catalytic residues, which are probably the nucleophilic attack mediators involved in the catalysis (Fujinaga et al. 2004, Rawlings et al. 2010). In metalloendopeptidases, the catalytic mechanism usually involves a single catalytic zinc ion tetrahedrally coordinated by one glutamate and two histidine residues (Rawlings et al. 2010). Serine peptidases have serine residues at their active sites, which together with two other variable amino acids constitute the “catalytic triad” (Hedstrom 2002, Rawlings et al. 2010). Threonine peptidases have threonine residues as their nucleophiles during catalysis. For unknown peptidases, the active site residues have not yet been determined.

Evolutionary analyses have been applied to a broad range of studies, which include the identification of gene/protein families that have expanded in a specific lineage over evolutionary time and possibly indicate the existence of selective pressure (Irving et al. 2003, Sargeant et al. 2006, Nahum & Pereira 2008, Robinson et al. 2008, Wu et al. 2009, Huzurbazar et al. 2010). The availability of faster and more powerful computers combined with the development of automated pipelines has enabled the investigation of such evolutionary processes through the reconstruction of phylogenetic trees for the complete set of proteins encoded in a genome (known as phylome). The results obtained by this analysis provide a broad view of the evolution of an organism’s genome and proteome, which allows for a deeper understanding of genomic complexity and lineage-specific adaptations (Huerta-Cepas et al. 2007, 2010b).

In a previous study, we described the reconstruction of the *S. mansoni* phylome to improve gene/protein functional annotation and provide insights into parasite’s biology (phyromed.org). By applying an automated pipeline, we also identified lineage-specific gene duplications, which may have led to a potential diversification of several protein families that are relevant for host-parasite interactions, such as tetraspanins, fucosyltransferases and sperm-coating protein-like proteins. Here, we explore the *S. mansoni* phylome data to analyse three endopeptidase families that expanded in this lineage since its diversification from 15 other metazoan species with the aim of contributing to the available knowledge of parasite biology and host-parasite interactions from an evolutionary perspective. The members of these families include leishmanolysins (metalloendopeptidase M8 family), cercarial elastases (serine peptidase S1 family) and cathepsin D proteins (aspartic peptidase A1 family).

The present paper is centred on two main research questions: (i) Did any peptidase families expand in the *S. mansoni* genome/proteme and if so, which ones? (ii) What are the evolutionary histories of these peptidase families? To address these questions, we used a so-called species-overlap algorithm (Huerta-Cepas et al. 2007) to detect lineage-specific duplications that occurred during the evolution of the parasite’s genome. We also integrated information on sequence alignments, phylogenetic trees, protein architecture and the conservation of critical residues to characterise these proteins. Our results indicate that each peptidase family has a unique evolutionary history within/ across the analysed species. Furthermore, our data support the hypothesis that gene duplication events followed by divergence is the main mechanism shaping the evolution of *S. mansoni*-specific paralogous groups.

The analysis of the evolutionary histories of these three *S. mansoni* families is relevant to functional genomics, evolutionary biology, medicine and biotechnology, especially taking into account the importance of *S. mansoni* peptidases in the development of schistosomiasis and that they have been described as promising vaccine and drug targets (McKerrow et al. 2006, Abdulla et al. 2007, Kasný et al. 2009).

**MATERIALS AND METHODS**

**Organisms and sequence data** - The dataset of species selected for analysis includes eight invertebrates (*Nematostella vectensis*, *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *S. mansoni*, *Drosophila melanogaster*, *Anopheles gambiae*, *Bombbyx mori* and *Strongylocentrotus purpuratus*), one tunicate (*Ciona intestinalis*), one cephalochordate (*Branchiostoma floridae*), three vertebrates (*Danio rerio*, *Mus musculus* and *Homo sapiens*), three fungi (*Neurospora crassa*, *Saccharomyces cerevisiae* and *Ustilago maydis*) and one plant (*Arabidopsis thaliana*). Information on the selected taxa is provided as Supplementary data.

This dataset is particularly rich in metazoans (76% of the selected species) that cover important evolutionary innovations, for example, the origin of bilateral symmetry, the third germ layer, the development of organs, systems, complex patterns of communication and the emergence of the adaptive immune system, which makes it especially suitable for addressing the evolutionary innovations in *S. mansoni* in comparison with other metazoan species (phyromed.org).

The *S. mansoni* predicted proteome dataset was downloaded from SchistoDB version 2.0 (schistodb.net) (Zerlotini et al. 2009). Proteomes derived from the 16 fully sequenced genomes were downloaded from the Broad Institute *Ustilago maydis* Database, Ensembl, Integrg8, JGI Genome Projects, National Center for Biotechnology Information Genome Database and SilkDB, which can be collectively accessed through the Genomes OnLine Database (genomesonline.org).

**Endopeptidase protein families** - Peptidases are hydrolases that act on peptide bonds [Enzyme Commission (EC) 3.4]. Three endopeptidase families were selected and analysed in detail in the present work. They include the metalloendopeptidase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-) members and belong to three peptidase clans (MA, PA and AA, respectively), as described in the MEROPS database (Rawlings et al. 2010).

Information on enzymes was collected from the literature and database references and included in the Supplementary data. The EC numbers were collected from the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology database, which is available online (chem.qmul.ac.uk/iubmb/enzyme/).
Alignments and phylogenetic trees - Sequence alignments and phylogenetic trees of the endopeptidase families selected for analysis were retrieved from the S. mansoni phylome data, which were reconstructed through a comparative analysis among all proteins encoded by the parasite genome and their potential homologs in 16 other eukaryotic species (phylomedb.org) (Huerta-Cepas et al. 2011).

Briefly, the S. mansoni phylome was reconstructed using each protein encoded in the S. mansoni genome ("seed" proteins) and the potential homologs identified through similarity-based searches (Smith & Waterman 1981) against the dataset of selected proteome data described above. The groups of homologous sequences were aligned using MUSCLE v3.6 (Edgar 2004) and gap-rich columns were filtered using trimAl (Capella-Gutierrez et al. 2009). Phylogenetic analyses were performed using the neighbour-joining and maximum likelihood (ML) methods, as implemented in PhyML (Guindon & Gascuel 2003).

For the phylogenetic reconstruction of each "seed protein", we tested four different evolutionary models (JTT, WAG, BLOSUM62 and VT). In all cases, a discrete gamma-distribution model with four rate categories plus invariant positions was assumed with the gamma parameter and the fraction of invariant positions estimated from the data. Tree support values were computed using the approximate likelihood ratio test as implemented in PhyML (Guindon & Gascuel 2003, Anisimova & Gascuel 2006). The evolutionary model best fitting the data was determined by comparing the likelihood of the used models according to the Akaike Information Criterion (Akaike 1973). The resulting alignments, phylogenies and homology prediction can be accessed at PhylomeDB (phylomedb.org) (Huerta-Cepas et al. 2011) through protein sequence identifiers (e.g., UniProt: C4PZH6; SchistoDB: Smp_127030; PhylomeDB: Phy000V7EC_SCHMA).

To integrate information from SchistoDB (Zerlotini et al. 2009) and PhylomeDB (Huerta-Cepas et al. 2011), we built a local relational database, named SchistoPhylomeSQL, which allowed us to extract and interpret the large amount of data in this work (Fig. 1). Access to this local database was implemented using DbVisualizer version 7.0.5 (dbvis.com). The SchistoPhylomeSQL database was the main resource for data mining in this work. In-house Perl scripts and Structured Query Language queries were used to parse data files during the database building and searching processes.

Paralogy and orthology relationships - To derive a complete catalogue of the paralogy and orthology relationships between S. mansoni proteins and those from other eukaryotic proteomes, we applied a "species-overlap" algorithm, as previously described (Huerta-Cepas et al. 2007). This algorithm uses the level of species overlap between the two daughter partitions of a given node to define it as a duplication or speciation event, which give rise to paralogs and orthologs, respectively. Once all the nodes have been classified, the algorithm establishes the paralogy and orthology relationships between the "seed protein" and other proteins included in the phylogenetic tree, according to the original definition of these terms (Fitch 1970, Gabaldón 2008).

Lineage-specific duplications - Using a python Environment for Tree Exploration (Huerta-Cepas et al. 2010a), we analysed the S. mansoni phylome data (phylomedb.org) to identify protein families that were specifically expanded in the S. mansoni lineage since its diversification from the other selected taxa (Supplementary data). The duplication events defined by the "species-overlap" algorithm that only comprised paralogs from S. mansoni were considered lineage-specific duplications. In cases where more than one phylogenetic tree contained the same paralogous proteins, by changing only the "seed" protein position, the data were filtered to obtain a non-redundant list of in-paralogs.

Protein architecture and critical residues - In this study, we used the Pfam database (Finn et al. 2010) to identify the presence and organisation of protein sequence domains as well as critical residues present in the three S. mansoni endopeptidase families. Pfam is a large and widely used database of protein domains families. This database contains multiple sequence alignments and profile hidden Markov models (profile HMMs) for each protein family. Pfam-A entries are derived from the underlying sequence database, which is termed Pfam-seq. This database is built from the most recent release of UniProtKB at a given time point (Finn et al. 2010, Apweiler et al. 2011). To predict active sites in new sequenc- es, Pfam uses the information available in UniProtKB for homologous proteins, whose catalytic residues have been experimentally characterized (Mistry et al. 2007). Based on Pfam information, the illustrations of the S. mansoni protein domain architectures were generated using DOG 2.0 (Ren et al. 2009).

Fig. 1: flowchart of the applied methodology. The Schistosoma mansoni proteome data was retrieved from SchistoDB and each protein was used as "seed" to reconstruct the S. mansoni phylome. The resulting alignments, phylogenies, and homology predictions are available at PhylomeDB. To integrate information from SchistoDB and PhylomeDB, we built the SchistoPhylomeSQL, a local relational database as the main resource for data mining in this work.
RESULTS

Comparative genomics has revealed a great deal of sequence and/or functional diversity within and across organisms with respect to gene/protein family size, composition and the relatedness of their members (Huerta-Cepas et al. 2007, 2010b, Nahum et al. 2009, Andrade et al. 2011, Avelar et al. 2011). The rationale underlying the present work is that lineage-specific duplications may reflect molecular biodiversity and that the adaptation of organisms to different environments may ultimately help to identify potential therapeutic targets against parasitic diseases.

Our previous work identified lineage-specific gene duplications that led to the diversification of several families in S. mansoni (phylomedb.org). Furthermore, recent advances in proteomic analyses of schistosomes have revealed that peptidases are one of the main virulence factors involved in the pathogenesis of schistosomiasis (McKerrow et al. 2006, Kasný et al. 2009). In this work, we performed a phylogenomic analysis to address the two main questions of (i) whether peptidase families are expanded in the S. mansoni proteome and (ii) whether they share similar evolutionary histories.

Endopeptidase family members are duplicated in S. mansoni - To investigate which peptidase families are expanded in the S. mansoni genome, we explored the parasite phylome data available at PhylomeDB (Huerta-Cepas et al. 2011). Phylogenetic analyses were performed using an automated pipeline and a complete list of the paralogy relationships among the S. mansoni proteins was retrieved using a “species-overlap” algorithm that identifies family members originated by lineage-specific duplication events (Huerta-Cepas et al. 2007).

Based on the functional annotation available from the SchistoDB (Zerlotini et al. 2009) and UniProt (Apweiler et al. 2011) databases, the results revealed that the most significant peptidase expansions in the S. mansoni proteome corresponded to endopeptidases such as leishmanolysins, cercarial elastases and cathepsin D proteins. These enzymes belong to three distinct endopeptidase families, metallopeptidase M8 family (EC 3.4.24.-), serine peptidase S1 family (EC 3.4.21.-) and aspartic peptidase A1 family (EC 3.4.23.-), as described in the MEROPS database (Rawlings et al. 2010) and represent promising targets for vaccine and drug development.

In total, we identified 12 leishmanolysins, 13 cercarial elastases (Supplementary data) and 11 cathepsin D proteins (Supplementary data) in the predicted S. mansoni proteome. These proteins vary in length and sequence composition, but they are highly conserved with respect to the presence of a conserved sequence domain, which is distinct for each protein family as defined by the Pfam database (see details below). Currently, no crystal structure has been obtained for the S. mansoni peptidases described here.

Leishmanolysins (also called invadolysin) is a major surface peptidase member of the metallopeptidase M8 family. Leishmanolysins are believed to share the same mechanism used by the other zinc metalloproteinases, such as thermolysin. The conserved glutamate residue in the catalytic site acts in conjunction with a zinc ion to deprotonate and activate a water molecule. In turn, the activated water molecule acts as a nucleophile to attack the carbonyl of the peptide bond of a variety of substrates (Macdonald et al. 1995, Schlagenhauf et al. 1998). In Leishmania, these proteins are involved in different types of processes, such as the inhibition or perturbations of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). These proteins may have similar activities in schistosomes. Indeed, the S. mansoni protein, SmPepM8 (Smp_090100), is the second most abundant constituent in cercarial secretions, which provides insight on how it may contribute to tissue invasion by schistosomes and suggests this protein as a potential anti-parasitic target (Curwen et al. 2006, Fitzpatrick et al. 2009).

The catalytic triad of serine, histidine and aspartate residues is conserved in members of the serine protease family (Wilmouth et al. 2001, Hajjar et al. 2010). In elastases, this triad and an essential water molecule are involved in the catalysis. The peptide to be cleaved is bound noncovalently in the enzyme near the catalytic triad. In the first reaction step, the hydroxyl of the serine residue performs a nucleophilic attack on the substrate amide bond to form an ester. The amino terminus of the substrate is then covalently bound to the enzyme. The histidine residue abstracts a proton from a water molecule, which then attaches to the ester carbon to give rise to an oxyanion intermediate. Cercarial elastases play a key role in the penetration by the cercariae of mammalian skin to initiate infection and recent studies have revealed that these peptidases are also employed by the schistosomes to overcome or evade the host immune response (Salter et al. 2002, Aslam et al. 2008).

Cathepsin D is a member of the aspartic protease family. The active site of cathepsin D contains two aspartate residues, which perform an acid-base catalysis. This enzymatic mechanism involves the deprotonation of water by an ionised aspartate residue. This water molecule attacks the peptide carbonyl and there is a simultaneous protonation of the carbonyl oxygen by the other aspartate residue (e.g., Northrop 2001). Schistosome cathepsin D is involved in haemoglobin digestion, a process that provides the parasite with its main source of amino acid nutrients and that is essential for its development, growth and reproduction (Brindley et al. 2001, Caffrey et al. 2004, Delcroix et al. 2006). Given the essential function of cathepsin D in parasite nutrition and the ability of recombinant forms to cleave human immunoglobulin G, this protein is considered a potential target for novel anti-parasitic interventions (Verity et al. 2001, Morales et al. 2008).

The phylogenetic relationships of each endopeptidase family (Figs 2-4) are shown with protein sequences represented by identifiers in PhylomeDB (phylomedb.org) (Huerta-Cepas et al. 2011), UniProt (uniprot.org) (Apweiler et al. 2011) and/or SchistoDB (schistodb.net) (Zerlotini et al. 2009). In each phylogenetic tree, the S. mansoni endopeptidases form a well-supported clade of closely related proteins.

Together, the analysis of the S. mansoni proteome through an evolutionary approach identified endopeptidase family members that arose by gene duplication after
the divergence of this parasite from the other eukaryotic species studied in this work. These lineage-specific duplications are related to the parasite’s biology and evolution.

**Leishmanolysins (metallopeptidase M8 family)** - Our pipeline identified 12 *S. mansoni* leishmanolysins (Supplementary data). Proteins Smp_171330 and Smp_171340 are located in the same genomic region of Smp_090100 and Smp_090110, respectively, and could not be retrieved from the UniProt (Apweiler et al. 2011) and GeneDB (genedb.org) databases, which suggests that these genes were incorrectly annotated and probably deleted from these databases. Similar findings were obtained in two previous studies (Berriman et al. 2009, Bos et al. 2009).

To reconstruct the evolutionary history of *S. mansoni* leishmanolysins and their homologs in selected taxa, we performed a sequence alignment of 32 protein sequences identified as potential homologs by our pipeline. The trimmed alignment contained 1,822 sites, which cover most of the conserved protein domain identified in these proteins.

By analysing the phylogenetic tree (Fig. 2), it is possible to demonstrate that *S. mansoni* leishmanolysins have homologs in most species analysed in the present work, with the exception of *C. intestinalis* (tunicata) and fungi. However, this result does not completely discard the presence of homologous proteins in other organisms because they may be very divergent from the others in the database and therefore be missed by the pipeline search. The same is true for the other protein families mentioned in this paper.

Based on the information available in the literature and curated databases, three leishmanolysin homologs have been experimentally confirmed in *D. melanogaster*, *M. musculus* and *H. sapiens* and their function is related to the coordination of mitotic progression and cell migration (for details see Supplementary data). Although predicted functions or experimental evidence are not yet available, the metallopeptidase M8 family is also expanded in the sea anemone (*N. vectensis*) and sea urchin (*S. purpuratus*). The metallopeptidase M8 family also has more paralogs in the schistosomes (12 proteins).

---

**Fig. 2**: phylogenetic relationships of the *S. mansoni* leishmanolysins and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
compared to *H. sapiens* (4 proteins), which is in contrast with what is normally observed for the human peptidase families (Berriman et al. 2009).

A conserved protein domain (Pfam: PF01457), which characterises members of the metallopeptidase M8 family, was identified in all *S. mansoni* proteins analysed here (Fig. 5). Length variation and conservation of active sites were also observed. According to the Pfam profile HMMs, truncated domains were identified in all proteins, which possibly reflects the presence of different protein isoforms, as has been described elsewhere (Floris et al. 2008). The truncated domains could also indicate that parts of the sequences are missing at the N-terminal, C-terminal regions, or both due to annotation issues.

The data also reveals that the protein domain is duplicated in Smp_167090, Smp_167120 and Smp_135530. Seven *S. mansoni* proteins (Smp_090100, Smp_090110, Smp_127030, Smp_135530, Smp_153930, Smp_167090 and Smp_173070) were identified as active due to the presence of expected active site residues and metal ligand sites in the correct positions based on alignments with reference sequences, as previously described (Berriman et al. 2009).

**Cercarial elastases (serine peptidase SI family)** - Our analysis identified a total of 13 cercarial elastases encoded in the *S. mansoni* genome (Supplementary data). Similar results were obtained by Berriman et al. (2009). However, with TreeFam (Ruan et al. 2008), Berriman et al. (2009) also

---

**Fig. 3:** conserved protein domain architecture of the *Schistosoma mansoni* leishmanolysins. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF01457) is present in all proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and active sites are also shown.
identified the Smp_192850 protein, which is annotated as a hypothetical protein and only contains 69 amino acids. Two proteins, Smp_152560.2 and Smp_056680.2, are encoded in the same genomic location and could not be recovered in UniProt (Apweiler et al. 2011). Searches for the former protein in GeneDB (genedb.org) retrieved only the latter (Smp_056680), which indicated that the Smp_152560.2 gene was improperly annotated and thus was eliminated from both databases. In the original version of the S. mansoni genome, some sequences were interpreted as isoforms and different gene models were constructed. However, further studies indicated that these were actually mistakes in the genome assembly/annotation due to low sequence coverage. In the new version of the parasite genome, which is to be released by the Wellcome Trust Sanger Institute (sanger.ac.uk), many of these sequences have been collapsed.

Whole amino acid sequences from 35 proteins were aligned and filtered to remove gap-rich columns as previously described. The trimmed alignment contains 583 sites, which cover the conserved protein domain. The phylogenetic analysis of the S. mansoni elastases and their homologs in the other species included in this work was performed as already described. The parasite elastases form a well-supported monophyletic clade, which suggests that these proteins originated from a common ancestor by gene duplication events followed by divergence in the Schistosoma lineage.

In observing the resulting phylogeny (Fig. 3), it is possible to demonstrate that S. mansoni elastases have homologs in six of the 16 other species considered in this analysis (N. vectensis, D. melanogaster, An. gambiae, B. floridae, M. musculus and H. sapiens). The serine peptidase S1 family is also expanded in all of these species except for one, D. melanogaster. According to the information available in UniProt (Apweiler et al. 2011), seven homologs have been experimentally confirmed in D. melanogaster, M. musculus and H. sapiens, and their function is related to a digestive process and immune response (Supplementary data). It is believed that similar activities are performed by elastases in schistosomes (Salter et al. 2002, Aslam et al. 2008).

Fig. 4: phylogenetic relationships of the Schistosoma mansoni elastases and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
A conserved protein sequence domain (Pfam: PF00089), which is found in all characterised members of the serine peptidase S1 family, was identified in the *S. mansoni* elastases and ranges in length from 141-265 amino acids (Fig. 6). The catalytic triad of histidine, aspartate and serine residues is present in most of these proteins. Based on profile HMMs available in Pfam, truncated regions were assigned to all 12 of these elastases, perhaps reflecting their degree of divergence in relation to other proteins in the database. Meanwhile, it is important to emphasise that protein databases do not cover all of the existing diversity in nature.

Together, these results indicate that the correct number of cercarial elastases encoded in the *S. mansoni* genome is 12 and not 13 as described before. However, only Smp_006510, Smp_006520 and Smp_141450 were previously predicted as active proteins (Berriman et al. 2009). Smp_194800 has a much shorter domain compared to others. This difference could reflect either the presence of an elastase pseudogene in the parasite genome or that the sequence was incorrectly annotated due to an error in the gene model. Considering that the first-pass annotation of the *S. mansoni* genome was produced by a combination of gene-finding algorithms (Augustus,

---

Fig. 5: conserved protein domain architecture of the *Schistosoma mansoni* elastases. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00089) is present in all proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and the catalytic triad of histidine (H), aspartate (D), and serine (S) are also shown.
Fig. 6: phylogenetic relationships of the *Schistosoma mansoni* cathepsins D and their homologs in selected taxa. Analysis was performed with trimmed sequence alignment by using the maximum likelihood method as implemented in PhyML. Best fit model (WAG) and support values for each node were estimated by the Akaike Likelihood Ratio Test (aLRT). Sequence labels follow the PhylomeDB internal identifier. For details, see Supplementary data.
Twinscan and GlimmerHMM) (Berriman et al. 2009), this genome has not received extensive manual curation and therefore, many gene models will be refined in the future. Furthermore, EVidenceModeler (Haas et al. 2008) has also been used to incorporate expressed sequence tag (EST) evidence into the data.

Cathepsin D proteins (aspartic peptidase A1 family) - Our pipeline identified 11 S. mansoni cathepsin D proteins (Supplementary data) that were duplicated after the divergence of S. mansoni from the other metazoans analysed here. The evolutionary history of cathepsin D proteins was reconstructed from the sequence alignment of 111 protein sequences from S. mansoni and the selected taxa. The final trimmed alignment contained 1,676 sites, which covered most of the conserved protein domain (Pfam: PF00026). Two S. mansoni proteins corresponded to alternative splicing products (Smp_136830.2 and Smp_013040.2). Similar results were found by Berriman et al. (2009).

The phylogenetic tree indicates that the S. mansoni cathepsin D proteins have homologs in all but one species (S. purpuratus) analysed in this work (Fig. 4). The aspartic peptidase A1 family has also been expanded in 12 of the 15 species in which homologous proteins were identified (A. thaliana, U. maydis, N. crassa, C. elegans, D. briggae, D. melanogaster, C. intestinalis, B. floridiae, D. rerio, M. musculus and H. sapiens). The number of paralogous proteins ranges from two-17 and includes different aspartic peptidases, such as pepsins, renins, gastricsin and cathepsin D proteins. Based on the information available in the literature and curated databases, these homologous proteins are involved in digestion and protein degradation (Supplementary data). In schistosomes, cathepsin D proteins play an integral role in haemoglobin proteolysis (Caffrey et al. 2004, Delcroix et al. 2006).

To predict the protein domain architecture of S. mansoni cathepsin D proteins, we applied the same methodology as previously described. The conserved domain (Pfam: PF00026), which has been found in all characterised aspartic peptidase A1 family members, was also identified in the S. mansoni proteins with sequence lengths ranging from 94-430 amino acids (Fig. 7). Active sites are also indicated. Based on the profile HMMs available in Pfam, truncated regions were observed in the N-terminal, C-terminal or both regions. The data also indicate that an additional short sequence domain (Pfam: PF07966), which is known as the A1 propeptide domain, is present at the N-terminal region of two S. mansoni proteins, Smp_013040.1 and Smp_013040.2. Smp_136840 has a much shorter domain compared to other proteins in the same family.

In a previous study, four S. mansoni cathepsin D proteins (Smp_013040.1, Smp_013040.2, Smp_136730 and Smp_136830.2) were identified as active proteins (Berriman et al. 2009), but the variation in the domain architecture and its implications in functional complexity were not investigated. One interesting study would be to analyse the functional properties of Smp_013040.1 and Smp_013040.2, which contain the A1 propeptide domain (PF07966).

DISCUSSION

We found that three endopeptidase families are expanded in the helminth parasite S. mansoni, which include members of the metallopeptidases (M8 family), serine peptidases (S1 family) and aspartic peptidases (A1 family). In this work, a comparative analysis of these three protein families revealed their distinct evolutionary histories and provided further information with respect to the sequence and functional features of the parasite family members.

Based on the S. mansoni genomic data, 335 peptidases were identified, which comprise 2.5% of the predicted proteome (Berriman et al. 2009). They include members of five major classes of peptidases (aspartic, cysteine, metallo, serine and threonine). Of the 61 peptidase families, 44 are expanded in this parasite and the number of paralogous proteins range from two-26.

Using a computational approach, Bos et al. (2009) analysed all putative peptidases encoded in the parasite’s genome in addition to using EST data, which is similar to work by Berriman et al. (2009). After removing redundant sequences, inactive homologs, likely pseudogenes and sequences smaller than 100 amino acids from the dataset, they identified a total of 255 peptidase sequences from the five catalytic classes.

Our results are not fully comparable to those obtained by Bos et al. (2009) with respect to elastases and cathepsin D proteins. However, it is worth noting that the phylogenetic analysis of the serine peptidase S1 family performed by these authors also indicated a well-supported clade of four S. mansoni elastases, which are corroborated by our findings. The other homologs with high similarities to the cercarial elastases were likely pseudogenes and, for this reason, they were excluded from the analysis by Bos et al. (2009).

Our results suggest that Schistosoma members of these endopeptidase families originated from successive gene duplication events in the parasite lineage after its diversification from the other metazoans analysed here. These results were corroborated by previous proteomic and phylogenetic analyses on Fasciola hepatica peptidases, which showed that the repertoire of virulence-associated cathepsin L proteins was established by a series of gene duplication events (Irving et al. 2003, Robinson et al. 2008). These studies also indicate that the gene duplications were followed by active site residue refinements, which interfere with the substrate specificity of the F. hepatica cathepsin L proteins. Whether the S. mansoni proteins share a similar refinement remains to be established.

Gene duplication followed by divergence is known to be the most predominant mechanism of molecular evolution and represents the main source of raw material for the generation of new genes and proteins through the processes of neo and sub-functionalisation (Ohno 1970, Conant & Wolfe 2008, Nahum & Pereira 2008, Hamilton et al. 2009). Although in some cases sequences have diverged to the extent that it is impossible to recognise homologous relationships, different proteins that arose by gene duplication may be related at distinct levels, such as sequence, structure, function or a combination of these features and can be grouped into families and superfamilies (Nahum & Pereira 2008).
Gene fusion, gene fission and domain shuffling were not observed as mechanisms shaping the evolution of the *S. mansoni* endopeptidase families analysed in this work. Whether gene fusion/fission also plays a role in the evolution of the *S. mansoni* genome will be a subject of a future work. Our previous study indicated that domain shuffling is one of the main evolutionary forces driving the sequence and functional diversification of the protein kinases of this parasite (Andrade et al. 2011, Avelar et al. 2011).

Peptidases have been implicated in various processes that are crucial to the development and survival of parasites, including host invasion, degradation of haemoglobin in blood feeding, immune evasion and activation of inflammation (McKerrow et al. 2006, Kasný et al. 2009).

Experimental work suggests that the SmPepM8 metallopeptidase (leishmanolysin) may contribute to tissue invasion by schistosome cercariae. This peptidase was the second most abundant protein released during the transformation of *S. mansoni* cercariae into schistosomula (Curwen et al. 2006). Leishmanolysins are a major surface peptidase member of the metallopeptidase M8 family, which in leishmaniasis are involved in different types of processes, such as the inhibition or perturbation of host cell interactions and the degradation of the extracellular matrix (Fitzpatrick et al. 2009). It is speculated that these proteins could perform similar activities in schistosomes during host-parasite interactions (Curwen et al. 2006, Fitzpatrick et al. 2009).

---

**Fig. 7:** conserved protein domain architecture of the *S. mansoni* cathepsin Ds. Protein identifiers were assigned in SchistoDB. The conserved protein domain according to Pfam (PF00026) is present in all proteins. The PF07966 additional N-terminal domain was identified in two proteins. Truncated regions (yellow block) are indicated. Sequence length, domain limits, and active sites are also shown.
Invasion of host skin is the initial event in establishing an infection in mammalian hosts. Considering the complexity of host skin barriers that the cercariae must go through during invasion, it has been suggested that multiple enzyme activities are required for this process (Salter et al. 2002). However, only one peptidase (cercarial elastase) has been identified as a major secretary product released during skin penetration (Knudsen et al. 2005, Hansell et al. 2008). These proteins may also be involved in eliminating the outer layer of the cercariae during transformation. Although cercarial elastases were named based on their ability to degrade insoluble elastin, numerous substrates for these enzymes have been identified, which include collagen, keratin and extracellular matrix proteins (Salter et al. 2002, McKerrow 2003, Knudsen et al. 2005).

Orthologous genes encoding elastase proteins were found in Schistosoma haematobium, Schistosoma japonicum and Schistosoma douthitti (Salter et al. 2002, Zhou et al. 2009). The expression of S. japonicum cercarial elastases was confirmed in both the sporocyst and cercarial stages and evidence that this peptidase is released by the parasite during the invasion of mammalian skin was obtained by anti-recombinant SjCE antibodies in infected mouse skin (Zhou et al. 2009). However, orthologous peptidases to S. mansoni cercarial elastases were not detected in the acatabular secretions of S. japonicum (Dvorák et al. 2008). Furthermore, the faster penetration by S. japonicum into the host skin may reflect the differential use of proteolytic enzymes in addition to those characterised in S. mansoni or even involve new peptidases not yet characterised (Chlichlia et al. 2005, He et al. 2005). Recent studies have also demonstrated that S. mansoni elastases are capable of cleaving IgE molecules from human, mouse and rat, indicating that the parasite may be able to overcome or evade the IgE response (Aslam et al. 2008). However, this subject remains controversial.

The biological complexity of S. mansoni is related to evolutionary innovations that took place before and after its diversification from other metazoa. Because duplicated genes are important substrates for improving an organism’s adaptation to its environment, understanding how members of protein families evolved may link evolutionary studies to parasite biology. In turn, this knowledge will provide insights into host-parasite relationships and accelerate the identification of novel vaccine and drug targets aimed at the treatment and eradication of schistosomiasis.

In conclusion, this paper provides an evolutionary view of three S. mansoni peptidase families, thus allowing for a deeper understanding of the genomic complexity and lineage-specific adaptations potentially related to the parasitic lifestyle. In the future, our results obtained using a systemic approach (proteome-wide analyses) may accelerate the understanding of schistosomiasis, its etiologic agents and host-parasite interactions and optimise the discovery of therapeutic targets for the development of new drugs and vaccines.

ACKNOWLEDGEMENTS

To the use of the computing resources of CRG (Spain) and CEBio-Fiocruz (Brazil), to Jaime Huerta-Cepas (CRG), Eric Aguilar and Francislion Silva (CEBio), for bioinformatics technical support, to Mariana de Oliveira (CEBio), for help with illustrations, and to the two anonymous reviewers, for the valuable suggestions that improved this paper.

REFERENCES

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

Akaike H 1973. Information theory and an extension of the maximum likelihood principle. In BN Petrov, F Csaki (eds.), Second International Symposium on Information Theory, Academiai Kiado, Budapest, p. 267-281.

Andrade LF, Nahum LA, Avelar LG, Silva LL, Zerlotini A, Ruiz JC, Oliveira G 2011. Eukaryotic protein kinases (ePKs) of the helminth parasite Schistosoma mansoni. BMC Genomics 12: 215.

Anisimova M, Gascuel O 2006. Approximate likelihood-ratio test for branches: a fast, accurate and powerful alternative. Syst Biol 55: 539-552.

Apweiler R, Martin MJ, O’Donovan C, Migrane M, Alam-Faruque Y, Antunes R, Barrell D, Bely B, Bingley M, Binns D, Bower L, Browne P, Chan WM, Dimmer E, Eberhardt R, Fazzini F, Fedotov A, Foulger R, Garavelli J, Castro LG, Huntley R, Jacobson J, Kleen M, Laiho K, Legge D, Lin Q, Liu W, Luo J, Orchard S, Patient S, Pichler K, Poggioi D, Pontikos N, Pruess M, Rosanoff S, Sawford T, Sehra H, Turner E, Corbett M, Donnelly M, van Rensburg P, Xeranios I, Bouguerel L, Aucinchloss A, Argoud-Puy G, Axelsen K, Bairoch A, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Bolland L, Boutet E, Quintaie SB, Breuza L, Bridge A, deCastro E, Coudert E, Cusin I, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gehan S, Ferro S, Gasteiger E, Gateau A, Gerritsen V, Gos A, Gruau-Gumowski N, Hinz U, Hulo C, Hulo N, James J, Jimenez S, Jungo F, Kapppler T, Keller G, Lara V, Lemercier P, Lieberherr D, Martin X, Masson P, Moinat M, Morgat A, Paesano S, Pedruzzl I, Pilbou S, Poux S, Pozzato M, Redaschi N, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staelhi S, Stanley E, Stutz A, Sundaram S, Tognolli M, Verbrugge L, Veehty AL, Wu CH, Arighi CN, Arminski L, Barker WC, Chen C, Chen Y, Dubey P, Huang H, Mazumder R, McGarvey P, Natale DA, Natarajan TG, Nchoutmboube J, Roberts M, Rong V, Roussel M, Scholz G, Steiner R, Strobel J, Suck J, Supek F, Tanaka N, Tekaya A, Santolii M, Verbrugge L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Barker WC, Chen C, Chen Y, Dubey P, Huang H, Mazumder R, McGarvey P, Natale DA, Natarajan TG, Ngchoumbouabe J, Roberts NV, Suzek BE, Ugocuchku U, Vinaikaya CR, Wang Q, Wang Y, Yeh LS, Zhang J 2011. Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res 39: D214-219.

Aslam A, Quinn P, McIntosh RS, Shi J, Ghumra A, McKerrow JH, Caffrey CR 2007. Schistosoma mansoni: proteases • Larissa Lopes Silva et al. 2011. Functional diversity of the Schistosoma mansoni tyrosine kinases. J Signal Transduc: 603290.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyaymy ST, Al-Lazikani B, Andrade LF, Ashton PD, Ashley MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eayre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pernet M, P egret RJ, Pratsio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Worman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM 2009. The genome of the blood fluke Schistosoma mansoni. Nature 460: 352-358.

Avelar LG, Nahum LA, Andrade LF, Oliveira G 2011. Functional diversity of the Schistosoma mansoni tyrosine kinases. J Signal Transduc: 603290.

Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyaymy ST, Al-Lazikani B, Andrade LF, Ashton PD, Ashley MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eayre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pernet M, P egret RJ, Pratsio AV, Quail MA, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Worman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM 2009. The genome of the blood fluke Schistosoma mansoni. Nature 460: 352-358.

Bos DH, Mayfield C, Michellc DJ 2009. Analysis of regulatory protease sequences identified through bioinformatic data mining of the Schistosoma mansoni genome. BMC Genomics 10: 488.
Secor WE, Mkoji GM, Loker ES 2009. Reduced susceptibility to praziquantel among naturally occurring Kenyan isolates of *Schistosoma mansoni*. *PLoS Negl Trop Dis* 3: e504.

Mistry J, Bateman A, Finn RD 2007. Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* 8: 298.

Morales ME, Rinaldi G, Gobert GN, Kines KJ, Tort JF, Brindley PJ 2008. RNA interference of *Schistosoma mansoni* cathepsin D, the apical enzyme of the hemoglobin proteolysis cascade. *Mol Biochem Parasitol* 157: 160-168.

Nahum LA, Goswami S, Serres MH 2009. Protein families reflect the metabolic diversity of organisms and provide support for functional prediction. *Physiol Genomics* 38: 250-260.

Nahum LA, Pereira SL 2008. Phylogenomics, protein family evolution and the tree of life: an integrated approach between molecular evolution and computational intelligence. In TG Smolinski, MG Milanova, AE Hassanien (eds.), *Studies in computational intelligence (SCI)*, Springer-Verlag, Berlin, pp. 259-279.

Northrop DB 2001. Follow the protons: a low-barrier hydrogen bond unifies the mechanisms of the aspartic proteases. *Acc Chem Res* 34: 790-797.

Ohno S 1970. *Evolution by gene duplication*, 1st ed., Springer-Verlag, Heidelberg, 160 pp.

Pica-Mattoccia L, Cioli D 2004. Sex- and stage-related sensitivity of *Schistosoma mansoni* to *in vivo* and *in vitro* praziquantel treatment. *Int J Parasitol* 34: 527-533.

Rawlings ND, Barrett AJ 1993. Evolutionary families of peptidases. *Biochem J* 290: 205-218.

Rawlings ND, Barrett AJ, Bateman A 2010. MEROPS: the peptidase database. *Nucleic Acids Res* 38: D227-233.

Ren J, Wen L, Gao X, Jin C, Xue Y, Yao X 2009. DOG 1.0: illustrator of protein domain structures. *Cell Res* 19: 271-273.

Robinson MW, Tort JF, Lowther J, Donnelly SM, Wong E, Xu W, Stack CM, Padula M, Herbert B, Dalton JP 2008. Proteomics and phylogenetic analysis of the cathepsin L protease family of the helminth pathogen *Fasciola hepatica*: expansion of a repertoire of virulence-associated factors. *Mol Cell Proteomics* 7: 1111-1123.

Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R 2008. TreeFam: 2008 Update. *Nucleic Acids Res* 36: D735-740.

Salter JP, Choe Y, Albrecht H, Franklin C, Lim KC, Craik CS, McKeever JH 2002. Cercarial elastase is encoded by a functionally conserved gene family across multiple species of schistosomes. *J Biol Chem* 277: 24618-24624.

Sargeant TJ, Marti M, Caler E, Carlton JM, Simpson K, Speed TP, Cowman AF 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biol* 7: R12.

Schlagenhauf E, Eiges R, Metcalf P 1998. The crystal structure of the *Leishmania major* surface protease leishmanolysin (gp63). *Structure* 6: 1035-1046.

Smith TF, Waterman MS 1981. Identification of common molecular subsequences. *J Mol Biol* 147: 195-197.

Steinmann P, Keiser J, Bos R, Tanner M, Utzinger J 2006. Schistosomiasis and water resources development: systematic review, meta-analysis and estimates of people at risk. *Lancet Infect Dis* 6: 411-425.

van der Werf MJ, de Vlas SJ, Brooker S, Looman CW, Nagelkerke NJ, Habbema JD, Engels D 2003. Quantification of clinical morbidity associated with schistosome infection in sub-Saharan Africa. *Acta Trop* 86: 125-139.

Verity CK, McManus DP, Brindley PJ 2001. Vaccine efficacy of recombinant cathepsin D aspartic protease from *Schistosoma japonicum*. *Parasite Immunol* 23: 153-162.

Wilmouth RC, Edman K, Neutze R, Wright PA, Clifton IJ, Schneider TR, Schofield CJ, Hajdu J 2001. X-ray snapshots of serine protease catalysis reveal a tetrahedral intermediate. *Nat Struct Biol* 8: 689-694.

Wu DD, Wang GD, Irwin DM, Zhang YP 2009. A profound role for the expansion of trypsin-like serine protease family in the evolution of hematophagy in mosquito. *Mol Biol Evol* 26: 2333-2341.

Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G 2009. SchistoDB: a *Schistosoma mansoni* genome resource. *Nucleic Acids Res* 37: D579-582.

Zhou Y, Zheng H, Chen Y, Zhang L, Wang K 2009. The *Schistosoma japonicum* genome reveals features of host-parasite interplay. *Nature* 460: 345-351.
| Kingdom        | Phylum       | Subphylum    | Code  | Scientific name               | NCBI_taxid | Proteome_data                                                                                                                                 |
|---------------|--------------|--------------|-------|------------------------------|------------|------------------------------------------------------------------------------------------------------------------------------------------------|
| Fungi         | Ascomycota   | Pezizomycotina | NEUCR | Neurospora crassa            | 5141       | broadinstitute.org/annotation/genome/neurospora                                                                                               |
| Fungi         | Ascomycota   | Saccharomycotina | YEAST | Saccharomyces cerevisiae     | 4932       | fungi.ensembl.org/Saccharomyces_cerevisiae                                                                                                    |
| Fungi         | Basidiomycota| Ustilaginomycotina | USTMA | Ustilago maydis              | 5270       | broadinstitute.org/annotation/genome/ustilago_maydis                                                                                           |
| Metazoa       | Arthropoda   | -             | DROME | Drosophila melanogaster      | 7227       | metazoa.ensembl.org/Drosophila_melanogaster                                                                                                   |
| Metazoa       | Arthropoda   | -             | ANOGA | Anopheles gambiae            | 7165       | metazoa.ensembl.org/Anopheles_gambiae                                                                                                         |
| Metazoa       | Arthropoda   | -             | BOMMO | Bombyx mori                  | 7091       | silkworm.genomics.org.cn                                                                                                                      |
| Metazoa       | Chordata     | Cephalochordata | BRAFL | Branchiostoma floridae      | 7739       | genome.jgi-psf.org/Brafl1/Brafl1.home.html                                                                                                     |
| Metazoa       | Chordata     | Craniata      | DANRE | Danio rerio                  | 7955       | ensembl.org/Danio_rerio                                                                                                                        |
| Metazoa       | Chordata     | Craniata      | MOUSE | Mus musculus                 | 10090      | ensembl.org/Mus_musculus                                                                                                                       |
| Metazoa       | Chordata     | Craniata      | HUMAN | Homo sapiens                 | 9606       | ensembl.org/Homo_sapiens                                                                                                                       |
| Metazoa       | Chordata     | Tunicata      | CIOIN | Ciona intestinalis           | 7719       | ensembl.org/Ciona_intestinalis                                                                                                                  |
| Metazoa       | Cnidaria     | -             | NEMVE | Nematostella vectensis      | 45351      | genome.jgi-psf.org/Nemvel/Nemvel.download.ftp.html                                                                                              |
| Metazoa       | Echinodermata | -             | STRPU | Strongylocentrotus purpuratus | 7668       | ncbi.nlm.nih.gov/projects/genome/guide/sea_urchin                                                                                               |
| Metazoa       | Nematoda     | -             | CAEEL | Caenorhabditis elegans      | 6239       | ensembl.org/Caenorhabditis_elegans                                                                                                             |
| Metazoa       | Nematoda     | -             | CAEBR | Caenorhabditis briggsae     | 6238       | metazoa.ensembl.org/Caenorhabditis_briggsae                                                                                                   |
| Metazoa       | Platyhelminthes | -         | SCHMA | Schistosoma mansoni        | 6183       | schistodb.net                                                                                                                                   |
| Viridiplantae | Streptophyta | -             | ARATH | Arabidopsis thaliana        | 3702       | ebi.ac.uk/integr8                                                                                                                               |
## TABLE

Functional annotation of leishmanolysins

| PhylomeDB new | PhylomeDB old | UniProt accession | Original ID | Length | Product | Pfam accession | Function | PubMed ID | Organism | Notes |
|---------------|---------------|-------------------|------------|--------|---------|----------------|----------|-----------|----------|--------|
| Phy000V136_SCHMA | Sch0003844 | C4QLC9 | Smp_090100 | 582 aa | SmPepM8; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V132_SCHMA | Sch0003840 | C4QLD0 | Smp_090110 | 567 aa | Metalloproteinase, putative; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V7EC_SCHMA | Sch00112282 | C4PZH6 | Smp_127030 | 729 aa | Protease family m8 leishmanolysin, putative; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V66N_SCHMA | Sch0010672 | C4Q3Z2 | Smp_135530 | 1,137 aa | Protease family m8 leishmanolysin, putative; Leishmanolysin-2 (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V3J5_SCHMA | Sch0007130 | C4QDA7 | Smp_153930 | 438 aa | Protease family m8 leishmanolysin, putative; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V1NS_SCHMA | Sch0004597 | C4QJ10 | Smp_167070 | 281 aa | Protease family m8 leishmanolysin, putative; Leishmanolysin-2 (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V1NQ_SCHMA | Sch0004595 | C4QJ12 | Smp_167090 | 699 aa | Protease family m8 leishmanolysin, putative; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V1NP_SCHMA | Sch0004594 | C4QJ13 | Smp_167100 | 363 aa | Expressed protein; Leishmanolysin-2 (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V1NN_SCHMA | Sch0004592 | C4QJ15 | Smp_167120 | 755 aa | Protease family m8 leishmanolysin, putative; Leishmanolysin-2 (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V135_SCHMA | Sch0003843 | - | Smp_171330 | 553 aa | Protease family m8 leishmanolysin, putative | PF01457 | - | - | Schistosoma mansoni | This sequence has genomic location overlap with Smp_090100. |
| Phy000V134_SCHMA | Sch0003842 | - | Smp_171340 | 543 aa | Protease family m8 leishmanolysin, putative | PF01457 | - | - | Schistosoma mansoni | This sequence has genomic location overlap with Smp_090110. |
| Phy000V0V8_SCHMA | Sch0003550 | C4QM23 | Smp_173070 | 572 aa | Protease family m8 leishmanolysin, putative; Invadolysin (M08 family) | PF01457 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy00003ZL_ANOGA | Aga0005169 | Q7QD48 | ENSANGP0000010959 | 621 aa | - | PF01457 | Predicted function | 12364791 | Anopheles gambiae | - |
| Phy0001NXC_ARATH | Ath0029631 | Q67ZD0 | Q67ZD0 | 841 aa | Major surface like glycoprotein | PF07974; PF01457 | Predicted function | 11130714 | Arabidopsis thaliana | - |
| Phy000V9EW_BOMMO | Bom0001650 | - | BGIBMGA001650-PA | 568 aa | - | - | - | Bombyx mori | - |
| PhyloMeDB_  | PhylomeDB_ | UniProt_ | Original_ | Length | Product | Pfam_ | Function | PubMed_ | Organism          | Notes                  |
|------------|------------|----------|------------|--------|---------|--------|----------|---------|------------------|------------------------|
| new        | old        | accession| ID         |        |         | accession|          |         |                  |                        |
| Phy000WY55_BRAFL | Bra0014763 | -        | prot14763  | 1,029 aa | -       | -      | -        | -       | Branchiostoma floridae | -                      |
| Phy000XJ65_BRAFL | Bra0042485 | C3YKD4  | prot42485  | 585 aa  | Putative uncharacterized protein | PF01457 | Predicted function | 18563158 | -                   |                        |
| Phy0002WFLCAEBR | Cbr0010832 | Q61YG1  | Q61YG1     | 663 aa  | Leishmanolysin-like peptidase | PF01457 | Predicted function | 14624247 | Caenorhabditis briggsae | -                      |
| Phy000362E_CAEEL | Cel0010125 | O62446  | Y43F4A.1a | 663 aa  | Leishmanolysin-like peptidase | PF01457 | Predicted function | 9851916  | Caenorhabditis elegans | -                      |
| Phy0005S4Q_DROME | Dme0007085 | Q9VH19  | CG3953-PA  | 683 aa  | Leishmanolysin-like peptidase; Invadolysin | PF01457 | Experimental evidence | 15557119, 10731132, 12537572, 12537569 | -                      |
| Phy0006MGV_DANRE | Dre0029534 | B0S5M4  | ENSDARP00000059717 | 618 aa | Novel protein similar to human and mouse leishmanolysin-like (metallopeptidase M8 family) | - | - | - | Danio rerio | -                      |
| Phy0008EM5_HUMAN | Hsa0022469 | Q96KR4  | ENSP00000328829 | 665 aa | Leishmanolysin-like peptidase | PF01457 | Experimental evidence | 1664997, 15557119 | Homo sapiens | -                      |
| Phy0009V90_MOUSE | Mms0018458 | Q8BMN4  | ENSMUSP00000023497 | 681 aa | Leishmanolysin-like peptidase | PF01457 | Experimental evidence | 16141072, 15489334 | Mus musculus | -                      |
| Phy000W4QT_NEMVE | Nem0003341 | A7ST90  | prot3341   | 563 aa  | Predicted protein | PF01457 | Predicted function | 17615350 | Nematostella vectensis | -                      |
| Phy000WFDM_NEMVE | Nem0017544 | A7RW33  | prot17544  | 624 aa  | Predicted protein | PF01457 | Predicted function | 17615350 | Nematostella vectensis | -                      |
| Phy000VKH7_STRPU | Str0001413 | - | XP_001186658.1 | 510 aa  | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VL96_STRPU | Str0002463 | - | XP_001189142.1 | 1,096 aa | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VLQO_STRPU | Str0003126 | - | XP_001203887.1 | 529 aa  | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VLQO_STRPU | Str0008229 | - | XP_785477.2 | 529 aa  | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VWGT_STRPU | Str0021529 | - | XP_001193341.1 | 265 aa  | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VXTF_STRPU | Str0025031 | - | XP_001192223.1 | 1,015 aa | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |
| Phy000VKH7_STRPU | Str0041478 | - | XP_001180406.1 | 510 aa  | -       | -      | -        | -       | Strongylocentrotus purpuratus | -                      |

*a*: new internal identifier in PhylomeDB; *b*: old internal identifier in PhylomeDB; *c*: UniProt accession number; *d*: original identifier in the database from which the proteome data was downloaded; *e*: amino acid (aa) sequence length; *f*: functional annotation in SchistoDB and UniProt; *g*: protein sequence domain(s) identified in the Pfam database; *h*: functional information in the literature available in UniProt; *i*: identifier in PubMed (PMID); *j*: scientific name of source organism; *k*: notes from this work.
| PhylomeDB_new | PhylomeDB_old  | UniProt_accession | Original_ID | Length | Product                                                                 | Pfam_accession | Function                  | PubMed_ID | Organism                  | Notes                                                                 |
|--------------|---------------|-------------------|-------------|--------|--------------------------------------------------------------------------|----------------|---------------------------|-----------|---------------------------|------------------------------------------------------------------------|
| Phy000UYL5_SCHMA | Sch0000377     | C1M2P1            | Smp_119130  | 263 aa | Elastase 1a, putative; cercarial elastase (S01 family)                   | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UYXF_SCHMA | Sch0000833     | C4QSP7            | Smp_115980  | 212 aa | Cercarial elastase HP1, putative; cercarial elastase (S01 family)       | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UZ5I_SCHMA | Sch0001129     | C4QS06            | Smp_187200  | 145 aa | Tryptase gamma precursor, putative; cercarial elastase (S01 family)     | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UAZ3_SCHMA | Sch0001310     | C1M1Z6            | Smp_112090  | 263 aa | Elastase 2a, putative; cercarial elastase (S01 family)                  | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UZB8_SCHMA | Sch0001364     | C4QRE6            | Smp_185190  | 141 aa | Elastase 1a, putative                                                   | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UZBD_SCHMA | Sch0001369     | C1M1X0            | Smp_185150  | 148 aa | Cercarial elastase (S01 family)                                         | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000UZPL_SCHMA | Sch0002010     | C4QQG8            | Smp_106910  | 186 aa | Elastase 1b, putative; cercarial elastase (S01 family)                  | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V3PR_SCHMA | Sch0007382     | C4QCN4            | Smp_056680.2| 138 aa | Cercarial elastase (S01 family); trypsin precursor, putative            | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V3PS_SCHMA | Sch0007383     | -                 | Smp_152560.2| 190 aa | Tryptase precursor, putative                                            | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V5A7_SCHMA | Sch0009474     | C4Q6V2            | Smp_141450  | 199 aa | Serine protease, putative; subfamily S1A unassigned peptidase (S01 family) | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V7LJ_SCHMA | Sch0012547     | -                 | Smp_006520  | 263 aa | Elastase 2b, putative                                                   | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V7LK_SCHMA | Sch0012548     | C4PY5S8           | Smp_006510  | 263 aa | Serine protease, putative; cercarial elastase (S01 family)              | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy000V7LL_SCHMA | Sch0012549     | -                 | Smp_194890  | 265 aa | Elastase, truncated protein, possible pseudogene                        | PF00089        | Predicted function        | 19606141 | Schistosoma mansoni       | -                                                                      |
| Phy00000UL_ANOGA | Aga0001101     | A7UTT9            | ENSANGP00000021783 | 305 aa | -                                                                       | PF00089        | Predicted function        | 12364791 | Anopheles gambiae         | -                                                                      |
| Phy00000V0_ANOGA | Aga0001116     | Q7Q6S1            | ENSANGP00000015859 | 275 aa | -                                                                       | PF00089        | Predicted function        | 12364791 | Anopheles gambiae         | -                                                                      |
| PhylomeDB new<sup>a</sup> | PhylomeDB old<sup>b</sup> | UniProt accession<sup>c</sup> | Original ID<sup>d</sup> | Length<sup>e</sup> | Product<sup>f</sup> | Pfam accession<sup>g</sup> | Function<sup>h</sup> | PubMed ID<sup>i</sup> | Organism<sup>j</sup> | Notes<sup>k</sup> |
|-------------------------|-------------------------|-----------------------------|------------------------|-----------------|-------------------|-----------------|----------------|----------------|----------------|----------------|
| Phy00006TW_ANOGA | Aga0008852 | Q7PZ86 | ENSANGP00000020262 | 440 aa | CLIP-domain serine protease subfamily A | PF00089 | Predicted function | 12364791 | Anopheles gambiae | - |
| Phy00008H9_ANOGA | Aga0010989 | Q7PWE3 | ENSANGP00000025294 | 248 aa | - | PF00089 | Predicted function | 12364791 | Anopheles gambiae | - |
| Phy00008HB_ANOGA | Aga0010991 | - | ENSANGP00000006368 | 365 aa | - | - | - | - | Anopheles gambiae | - |
| Phy000WN7C_BRAFL | Bra0000495 | C3ZSB4 | prot495 | 262 aa | Putative uncharacterized protein | PF00089 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000WUO5_BRAFL | Bra0010245 | - | prot10245 | 218 aa | - | PF00089 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000XJPI_BRAFL | Bra0043201 | C3YEH6 | prot43201 | 236 aa | Putative uncharacterized protein | PF00089 | Predicted function | - | Branchiostoma floridiae | - |
| Phy0005WZD_CROM | Dme0013359 | Q9FRS6 | CG40160-PA | 421 aa | - | PF00089 | Experimental evidence | 10731132, 12537568, 12537574, 12537573, 12537572, 16110336, 17569856, 17569867 | Drosophila melanogaster | - |
| Phy0007XNC_HUMAN | Hsa0000480 | P08861 | ENSP00000338369 | 270 aa | Chymotrypsin-like elastase family member 3B; elastase IIIB; elastase-3B | PF00089 | Experimental evidence | 2826474, 14702039, 16710414, 15489334, 3477287, 2675835, 3178837, 2753124, 2737288 | Homo sapiens | - |
| Phy00085MY_HUMAN | Hsa0010834 | - | ENSP00000234798 | 321 aa | - | PF00089 | Experimental evidence | - | Homo sapiens | - |
| Phy00085MZ_HUMAN | Hsa0010835 | - | ENSP00000344083 | 274 aa | - | PF00089 | Experimental evidence | - | Homo sapiens | - |
| Phy00085N1_HUMAN | Hsa0010837 | - | ENSP00000343577 | 275aa | - | PF00089 | Experimental evidence | - | Homo sapiens | - |
| Phy00085N2_HUMAN | Hsa0010838 | Q9BZJ3 | ENSP00000211076 | 242 aa | Tryptase delta; mast cell mMCP-7-like; tryptase-3 | PF00089 | Experimental evidence | 9920877, 11174199, 11157797, 15616553, 18854315, 12391231 | Homo sapiens | - |
| Phy0009LSN_MOUSE | Mms0006205 | Q14C59 | ENSMUSP00000042406 | 416 aa | Membrane protease serine 11B; airway trypsin-like protease 5 | PF00089 | Experimental evidence | 2826474, 14702039, 16710414, 15489334, 3477287, 2675835, 3178837, 2753124, 2737288 | Mus musculus | - |
| Phy0009QF6_MOUSE | Mms0012200 | Q91X79 | ENSMUSP0000023775 | 266 aa | Elastase 1, pancreatic | PF00089 | Experimental evidence | 15489334, 12040188 | Mus musculus | - |
| Phy0009R3K_MOUSE | Mms0013078 | - | ENSMUSP0000015576 | 243 aa | - | PF00089 | Experimental evidence | - | Mus musculus | - |
| Phy0009UUF_MOUSE | Mms0017933 | P11032 | ENSMUSP0000023897 | 260 aa | Granzyme A | PF00089 | Experimental evidence | 2976140, 1639378, 1460043, 15489334, 3292396, 2422755, 3555842, 3536355, 3260181 | Mus musculus | - |
| Phy000A4AP_MOUSE | Mms0030183 | Q9CQ52 | ENSMUSP00000024200 | 269 aa | Elastase 3, pancreatic | PF00089 | Experimental evidence | 10349636, 16140172, 15489334, 15328353 | Mus musculus | - |
| Phy000W6H_NEMVE | Nem0006289 | A7RP61 | prot6289 | 252 aa | Predicted protein | PF00089 | Predicted function | 17615350 | Nematostella vectensis | - |
| Phy000WDER_NEMVE | Nem0014980 | A7SGX2 | prot14980 | 299 aa | Predicted protein | PF00089, PF00629 | Predicted function | 17615350 | Nematostella vectensis | - |

---

*a*: new internal identifier in PhylomeDB; *b*: old internal identifier in PhylomeDB; *c*: UniProt accession number; *d*: original identifier in the database from which the proteome data was downloaded; *e*: amino acid (aa) sequence length; *f*: functional annotation in SchistoDB and UniProt; *g*: protein sequence domain(s) identified in the Pfam database; *h*: functional information in the literature available in UniProt; *i*: identifier in PubMed (PMID); *j*: scientific name of source organism; *k*: notes from this work.
### TABLE

Aspartic peptidases analyzed in this paper

| PhylomeDB_ new | PhylomeDB_ old | UniProt accession | Original ID | Length | Product | Pfam accession | Function | PubMed_ID | Organism | Notes |
|----------------|----------------|-------------------|-------------|--------|---------|----------------|----------|-----------|----------|-------|
| Phy000V5ZW_SCHMA | Sch0010420 | C4Q4K4 | Smp_136840 | 94 aa | Cathepsin d, putative; subfamily A1A unassigned peptidase | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V5ZX_SCHMA | Sch0010421 | C4Q4K2 | Smp_136830.2 | 345 aa | Cathepsin d, putative; subfamily A1A unassigned peptidase (A01 family) EMBL CAZ30383.1 | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V5ZY_SCHMA | Sch0010422 | C4Q4K3 | Smp_136830.1 | 255 aa | Cathepsin d, putative; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V5ZZ_SCHMA | Sch0010423 | C4Q4K1 | Smp_136820 | 192 aa | Aspartic proteinase-related; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V60H_SCHMA | Sch0010441 | C4Q4I3 | Smp_136730 | 401 aa | Cathepsin D2-like protein; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V60I_SCHMA | Sch0010442 | C4Q4I2 | Smp_136720 | 242 aa | Aspartic proteinase-related; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V6MX_SCHMA | Sch0011276 | C4Q2B3 | Smp_132480 | 393 aa | Aspartyl proteases, putative; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V6MY_SCHMA | Sch0011277 | C4Q2B2 | Smp_132470 | 286 aa | Aspartyl proteases, putative; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V6MZ_SCHMA | Sch0011278 | C4Q2B1 | Smp_018800 | 278 aa | Aspartyl proteases, putative; subfamily A1A unassigned peptidase (A01 family) | PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V73F_SCHMA | Sch0011883 | C4Q0K4 | Smp_013040.1 | 428 aa | Aspartyl proteases, putative; cathepsin D (A01 family) | PF07966, PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy000V73G_SCHMA | Sch0011884 | C4Q0K5 | Smp_013040.2 | 430 aa | Aspartyl proteases, putative; cathepsin D (A01 family) | PF07966, PF00026 | Predicted function | 19606141 | Schistosoma mansoni | - |
| Phy00004AG_ANOGA | Aga0005560 | Q7PTQ9 | ENSANGP00000013568 | 389 aa | - | - | Predicted function | 12364791 | Anopheles gambiae | Alternative splicing product. |
| Phy000143C_ARATH | Ath0003927 | O04593 | O04593 | 433 aa | - | PF00026, PF05184, PF03489 | Predicted function | - | Arabidopsis thaliana | - |
| Phy00015EA_ARATH | Ath0005617 | O65453 | O65453 | 336 aa | Aspartic proteinase like protein | PF00026 | Predicted function | 10617198 | Arabidopsis thaliana | - |
| PhylomeDB_new | PhylomeDB_old | UniProt_accession | Original_ID | Length | Product | Pfam_accession | Function | PubMed_ID | Organism | Notes |
|---------------|--------------|------------------|-------------|--------|---------|----------------|----------|-----------|----------|-------|
| Phy00019QK_ARATH | Ath0011243 | Q8VYL3 | Q8VYL3 | 513 aa | - | PF00026, PF05184, PF03489 | Predicted function | 11130712 | Arabidopsis thaliana | - |
| Phy0001G2O_ARATH | Ath0019455 | Q9LQA9 | Q9LQA9 | 375 aa | - | PF00026 | - | - | Arabidopsis thaliana | - |
| Phy0001MDZ_ARATH | Ath0027638 | Q9XEC4 | Q9XEC4 | 508 aa | Putative aspartic proteinase | PF00026, PF05184, PF03489 | Predicted function | 10617198 | Arabidopsis thaliana | - |
| Phy0001PJF_ARATH | Ath0031722 | O65390 | O65390 | 506 aa | Putative aspartic proteinase | PF00026, PF05184, PF03489 | Experimental evidence | 11130712; 12093376 | Arabidopsis thaliana | - |
| Phy000VGW3_BOMMO | Bom0011344 | BGIBMGA011344-PA | - | 326 aa | - | - | - | - | Bombus mori | - |
| Phy000WP82_BRAFL | Bra0003157 | prot3157 | - | 423 aa | - | - | - | - | Branchiostoma floridiae | - |
| Phy000WRJZ_BRAFL | Bra0006192 | C3YUT2 | prot6192 | 388 aa | Putative uncharacterized protein | PF00026 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000WUPE_BRAFL | Bra0010298 | C3ZMY0 | prot10298 | 493 aa | Putative uncharacterized protein | PF00026 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000X67S_BRAFL | Bra0025329 | C3XQC3 | prot25329 | 392 aa | Putative uncharacterized protein | PF00026 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000X6BF_BRAFL | Bra0025469 | C3YB78 | prot25469 | 423 aa | Putative uncharacterized protein | PF00026 | Predicted function | 18563158 | Branchiostoma floridiae | - |
| Phy000XA8D_BRAFL | Bra0030645 | prot30645 | 488 aa | - | - | - | - | - | Branchiostoma floridiae | - |
| Phy000XFYF_BRAFL | Bra0038243 | prot38243 | 439 aa | - | - | - | - | - | Branchiostoma floridiae | - |
| Phy000WN7K_BRAFL | Bra0048871 | prot48871 | 398 aa | - | - | - | - | - | Branchiostoma floridiae | - |
| Phy000XO0W_BRAFL | Bra0049039 | prot49039 | 243 aa | - | - | - | - | - | Branchiostoma floridiae | - |
| Phy000XOPE_BRAFL | Bra0050000 | prot50000 | 387 aa | - | - | - | - | - | Branchiostoma floridiae | - |
| Phy002PUX_CAEBR | Cbr0002312 | Q60TT2 | 393 aa | - | - | - | - | - | Caenorhabditis briggsae | - |
| Phy002QAI_CAEBR | Cbr0002873 | A8XV46 | Q60W85 | 704 aa | CBR-ASP-2 protein; Cbr-asp-2 | PF00026 | Predicted function | 14624247 | Caenorhabditis briggsae | - |
| Phy002TGn_CAEBR | Cbr0006968 | A8XDB8 | Q61GA5 | 389 aa | CBR-ASP-6 protein; Cbr-asp-6 | PF00026 | Predicted function | 14624247 | Caenorhabditis briggsae | - |
| Phy0002TG9_CAEBR | Cbr0006969 | Q61GA7 | - | - | - | - | - | - | Caenorhabditis briggsae | - |
| Phy0002TM9_CAEBR | Cbr0007184 | A8XC79 | Q61H49 | 397 aa | CBR-ASP-3 protein | PF00026 | Predicted function | 14624247 | Caenorhabditis briggsae | - |
| PhylomeDB_{new} | PhylomeDB_{old} | UniProt accession | Original ID | Length | Product | Pfam accession | Function | PubMed_ID | Organism | Notes |
|----------------|-----------------|------------------|-------------|--------|---------|----------------|----------|-----------|----------|--------|
| Phy0002UJ3_CAEEL | Cbr0008366      | A8X733           | Q61MA1      | 428 aa | Putative uncharacterized protein | PF00026 | Predicted function | 14624247 | Caenorhabditis briggsae | - |
| Phy0002VN7_CAEEL | Cbr0009810      | -                | Q61U20      | 393 aa | -       | -              | -         | -         | Caenorhabditis briggsae | - |
| Phy0002VQI_CAEEL | Cbr0009929      | -                | Q61UL2      | 446 aa | -       | -              | -         | -         | Caenorhabditis briggsae | - |
| Phy0002XGM_CAEEL | Cbr0012165      | -                | Q624R1      | 386 aa | -       | -              | -         | -         | Caenorhabditis briggsae | - |
| Phy000373I_CAEEL | Cel0011461      | -                | C11D2.2     | 394 aa | -       | -              | -         | -         | Caenorhabditis elegans | - |
| Phy00039HI_CAEEL | Cel0014557      | O16338           | F59D6.3     | 474 aa | Putative uncharacterized protein | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy00039HJ_CAEEL | Cel0014558      | O16339           | F59D6.2     | 380 aa | Putative uncharacterized protein | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003AKP_CAEEL | Cel0015968      | O01530           | F21F8.7.1   | 389 aa | Aspartic protease 6; asp-6 | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003AKQ_CAEEL | Cel0015969      | O01531           | F21F8.4.1   | 395 aa | Putative uncharacterized protein | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003AKR_CAEEL | Cel0015970      | O01532           | F21F8.3.2   | 393 aa | Aspartyl protease protein 5; asp-5 | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003ANL_CAEEL | Cel0016072      | Q22972           | F28A12.4.1  | 389 aa | Putative uncharacterized protein F28A12.4 | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003ARP_CAEEL | Cel0016220      | Q86NE1           | T18H9.2a    | 635 aa | Aspartyl protease protein 2; asp-2 | PF00026, PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003BH6_CAEEL | Cel0017137      | Q18020           | C15C8.3     | 428 aa | -       | -              | PF00026 | Experimental evidence | 17761667, 9851916 | Caenorhabditis elegans | - |
| Phy0003CNH_CAEEL | Cel0018660      | -                | ZK384.3     | 638 aa | -       | -              | -         | -         | Caenorhabditis elegans | - |
| Phy0003CP4_CAEEL | Cel0018719      | Q8MYN5           | Y39B6A.24.1 | 391 aa | -       | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003CP5_CAEEL | Cel0018720      | Q8MYN6           | Y39B6A.23   | 395 aa | -       | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003CP6_CAEEL | Cel0018721      | Q8MYN7           | Y39B6A.22   | 390 aa | -       | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003CP7_CAEEL | Cel0018722      | Q9TVS4           | Y39B6A.20.4 | 396 aa | Aspartic protease 1; asp-1 | PF07966, PF00026 | Predicted function | 10854422, 9851916 | Caenorhabditis elegans | - |
| Phy0003DVO_CAEEL | Cel0020251      | Q94271           | K10C2.3     | 410 aa | Putative uncharacterized protein | PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| Phy0003DXT_CAEEL | Cel0020328      | P55956           | H22K11.1    | 398 aa | Aspartic protease 3 | PF00026 | Predicted function | 9851916, 9150941, 17761667 | Caenorhabditis elegans | - |
| Phy0003EV3_CAEEL | Cel0021526      | Q21966           | R12H7.2     | 444 aa | Protein R12H7.2, confirmed by transcript evidence EMBL CAA90633.1; asp-4 | PF07966, PF00026 | Predicted function | 9851916 | Caenorhabditis elegans | - |
| PhylomeDB_new | PhylomeDB_old | UniProt_accession | Original_ID | Length | Product | Pfam_accession | Function | PubMed_ID | Organism | Notes |
|---------------|---------------|-------------------|------------|--------|----------|----------------|----------|-----------|----------|-------|
| Phy0004BS1_CIOIN | Cin0006471 | - | ENSCINP0000010874 | 430 aa | - | - | - | - | Ciona intestinalis | - |
| Phy0004DLX_CIOIN | Cin0008843 | - | ENSCINP0000014585 | 400 aa | - | - | - | - | Ciona intestinalis | - |
| Phy0005MTU_DROME | Dme0000213 Q9VQ11 | CG31928-PA | 418 aa | - | PF00026 | Predicted function | 10731132, 12537572 | Drosophila melanogaster | - |
| Phy0005MTV_DROME | Dme0000214 Q9VQ12 | CG33128-PA | 405 aa | - | PF00026 | Predicted function | 10731132, 12537572 | Drosophila melanogaster | - |
| Phy0005MTW_DROME | Dme0000215 Q9VQ13 | CG31926-PA | 410 aa | - | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005MTX_DROME | Dme0000216 Q9VQ14 | CG31661-PA | 393 aa | - | PF00026 | Predicted function | 10731132, 12537572 | Drosophila melanogaster | - |
| Phy0005NNB_DROME | Dme0001274 Q9VLK3 | CG13095-PA | 372 aa | - | Beta-site APP-cleaving enzyme; bace | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005NXM_DROME | Dme0001645 Q9VKP7 | CG6508-PA | 423 aa | - | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005NXN_DROME | Dme0001646 Q9VKP6 | CG17134-PA | 391 aa | - | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005SYG_DROME | Dme0008155 Q9VEK5 | CG17283-PA | 465 aa | - | PF00026 | Predicted protein | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005SYH_DROME | Dme0008156 Q9VEK4 | CG5860-PA | 370 aa | - | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005SYI_DROME | Dme0008157 Q9VEK3 | CG5863-PA | 395 aa | - | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005UQR_DROME | Dme0010470 Q9W5G3 | CG13374-PA | 407 aa | Pepsinogen-like; pcl | PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005XCX_DROME | Dme0013860 Q7K485 | CG1548-PA | 392 aa | Cathepsin D | PF07966, PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy0005Y92_DROME | Dme0015017 A1Z9Q9 | CG10104-PA | 404 aa | - | PF07966, PF00026 | Predicted function | 12537572, 10731132 | Drosophila melanogaster | - |
| Phy00060Q9_DANRE | Dre0001360 | - | ENSDARP00000055495 | 395 aa | - | - | - | - | Danio rerio | - |
| Phy00060RL_DANRE | Dre0001408 | - | ENSDARP00000012342 | 412 aa | - | - | - | - | Danio rerio | - |
| Phy00060RT_DANRE | Dre0001416 | - | ENSDARP00000013587 | 200 aa | - | - | - | - | Danio rerio | - |
| Phy00064F8_DANRE | Dre0006147 Q6XQJ0 | ENSDARP00000061335 | 395 aa | Renin | PF07966, PF00026 | Predicted function | 14645735 | Danio rerio | - |
| Phy00066AZ_DANRE | Dre0008586 Q8AWD9 | ENSDARP00000061373 | 398 aa | Ctsd protein | PF07966, PF00026 | Predicted function | - | Danio rerio | - |
| Phy00066BY_DANRE | Dre0008621 | - | ENSDARP00000061099 | 395 aa | - | - | - | - | Danio rerio | - |
| Phy00066F3T_DANRE | Dre0019992 | - | ENSDARP00000003409 | 503 aa | - | - | - | - | Danio rerio | - |
| Phy0007YI8_HUMAN | Hsa0001592 | ENSP00000304306 | 307 aa | - | - | - | - | Homo sapiens | - |
| Phy0007YIQ_HUMAN | Hsa0001610 | ENSP00000329601 | 391 aa | - | - | - | - | Homo sapiens | - |
| Phy0007ZEQ_HUMAN | Hsa0002762 | ENSP00000272190 | 406 aa | - | - | - | - | Homo sapiens | - |
| PhylomeDB_new | PhylomeDB_old | UniProt_accession1 | Original_ID | Length1 | Product1 | Pfam_accession1 | Function1 | PubMed_ID | Organism1 | Notes1 |
|----------------|-------------|-------------------|-------------|---------|----------|----------------|----------|----------|-----------|-------|
| Phy0007ZFY_HUMAN | Hsa0002806 | ENSP00000354337 | 401 aa | - | - | - | - | 3927292, 3588310, 2069717, 15489334, 8262386, 7935485, 12643545, 12754519, 16335952, 16670177, 17081065, 16263699, 19159218, 8467789, 8393577, 10716266, 16685649 | Homo sapiens | - |
| Phy0008WQ_HUMAN | Hsa0004706 | ENSP00000236671 | 412 aa | Cathepsin D | PF07966, PF00026 | Experimental evidence | - | Homo sapiens | - |
| Phy00081HX_HUMAN | Hsa0005469 | ENSP00000322192 | 388 aa | - | - | - | - | 15489334 | Homo sapiens | - |
| Phy00081HY_HUMAN | Hsa0005470 | B7ZW62 | ENSP00000235720 | 402 aa | Napsin-A | PF00026 | Experimental evidence | 9877162, 10580105, 10591213, 15489334 | Homo sapiens | - |
| Phy0008CLE_HUMAN | Hsa0019850 | Q9Y5Z0 | ENSP00000332979 | 518 aa | Beta-secretase 2; aspartyl protease 1; memapsin-1 | PF00026 | Experimental evidence | 10591213, 10838186, 10965118, 10683441, 10749877, 11083922, 11042159, 12040188, 12712197, 14574404, 15489334, 15857888, 11423558, 16181112, 16305800, 11316808, 16965550 | Homo sapiens | - |
| Phy0008H2Y_HUMAN | Hsa0025666 | P20142 | ENSP00000211310 | 388 aa | Gastricsin | PF07966, PF00026 | Experimental evidence | 3335549, 2909526, 2567697, 14574404, 15489334, 2515193, 6816595, 7714902, 9406551 | Homo sapiens | - |
| Phy0009JXB_MOUSE | Mms0003781 | O09043 | ENSMUSP0000002274 | 419 aa | Napsin-A | PF00026 | Experimental evidence | 9013890, 11082205, 15489334, 16949457 | Mus musculus | - |
| Phy0009L41_MOUSE | Mms0005319 | - | ENSMUSP0000008035 | - | - | - | - | Mus musculus | - |
| Phy0009NOR_MOUSE | Mms0008657 | Q3UKT5 | ENSMUSP0000073072 | 397 aa | Cathepsin E | PF07966, PF00026 | Experimental evidence | 10349636, 16141073, 11042159, 11076861, 12040188 | Mus musculus | - |
| Phy0009NQ4_MOUSE | Mms0008706 | P06281 | ENSMUSP000000788 | 402 aa | Renin-1 | PF07966, PF00026 | Experimental evidence | 6370686, 2685761, 2691339, 16141072, 15489334, 6089205, 6392850, 9030738, 6327270 | Mus musculus | - |
| Phy0009VNF_MOUSE | Mms0018977 | Q8JL18 | ENSMUSP0000043918 | 514 aa | - | - | - | Mus musculus | - |
| Phy0009XXL_MOUSE | Mms0021935 | Mms0022711 | Q9D106 | ENSMUSP0000077032 | 381 aa | Pepsinogen 5, group I | PF07966, PF00026 | Experimental evidence | 10349636, 16141073, 11042159, 11076861, 15489334, 10683441, 10749877, 11083922, 11042159, 12040188 | Mus musculus | - |
| Phy0009YJ5_MOUSE | Mms0024598 | Q8JL18 | ENSMUSP0000025647 | 387 aa | - | - | - | Mus musculus | - |
| Phy0009ZZK_MOUSE | Mms0024598 | Ncr0000966 | Q7SD30 | (NCU00994.2) | 434 aa | Endothiapepsin | PF00026 | Predicted function | 12712197 | Mus musculus | Neurospora crassa |
| Phy000AW60_NEUCR | Ncr0000966 | Q7SD30 | (NCU00994.2) | 396 aa | Vacular protease A, pep-4 | PF00026 | Experimental evidence | 8702999, 12655011, 12712197 | Neurospora crassa | - |
| PhylomeDB new\(^a\) | PhylomeDB old\(^b\) | UniProt accession\(^c\) | Original ID\(^d\) | Length\(^e\) | Product\(^f\) | Pfam accession\(^g\) | Function\(^h\) | PubMed ID | Organism\(^i\) | Notes\(^k\) |
|-------------------------|--------------------------|--------------------------|------------------|--------------|----------------|----------------|------------|-----------|-------------|-----------|
| Phy000B20G_NEUCR        | Ncr0008518               | Q7SCF6                   | (NCU08739.2)     | 439 aa       | Endothiapepsin| PF00026       | Predicted function | 12712197 | Neurospora crassa | -         |
| Phy000B3H7_NEUCR        | Ncr0010417               | A7UXG4                   | (NCU10907.2)     | 529 aa       | Predicted protein | PF00026 | Predicted function | 12712197 | Neurospora crassa | -         |
| Phy000W68V_NEMVE         | Nem0005466               | prot5466                 | 370 aa           | Predicted protein | PF07966, PF00026 | Predicted function | 17615350 | Nematostella vectensis | -         |
| Phy000CWVQ_YEAST        | See0001780               | P12630                   | YIL015W          | 587 aa       | Barrierpepsin | PF00026       | Predicted function | 3124102, 9169870, 14562106 | Saccharomyces cerevisiae | -        |
| Phy000CYQD_YEAST        | See0004179               | P32329                   | YLR120C          | 569 aa       | Aspartic proteinase 3 | PF00026 | Experimental evidence | 2183521, 9000053, 9169871, 8389368, 7779785, 7657670, 9417119, 11737827, 14617149, 14562106, 16087741, 18591427, 18573178, 9485427 | Saccharomyces cerevisiae | -        |
| Phy000CYQF_YEAST        | See0004181               | Q12303                   | YLR121C          | 508 aa       | Aspartic proteinase yapsin-3 | PF00026 | Experimental evidence | 9090053, 9169871, 10191273, 11016834, 16087741, 3537721, 3023936, 8948103, 9169875, 1618910, 18840499, 1959673, 9135120 | Saccharomyces cerevisiae | -        |
| Phy000D0B9_YEAST        | See0006227               | P07267                   | YPL154C          | 405 aa       | Saccharopepsin | PF00026       | Experimental evidence | 17080091 | Ustilago maydis | -         |
| Phy000F3MX_USTMA        | Uma0000064               | Q4PJ9                    | UM00064.1        | 397 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |
| Phy000F55V_USTMA        | Uma0002043               | Q4PCX0                   | UM02043.1        | 481 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |
| Phy000F59M_USTMA        | Uma0002178               | Q4PCI5                   | UM02178.1        | 452 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |
| Phy000F5L9_USTMA        | Uma0002597               | Q4PBB6                   | UM02597.1        | 603 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |
| Phy000F7DY_USTMA        | Uma0004926               | Q4P4N7                   | UM04926.1        | 418 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |
| Phy000F7LP_USTMA        | Uma0005205               | Q4P3V8                   | UM05205.1        | 768 aa       | Putative uncharacterized protein | PF00026 | Predicted function | 17080091 | Ustilago maydis | -         |

\(^a\): new internal identifier in PhylomeDB; \(^b\): old internal identifier in PhylomeDB; \(^c\): UniProt accession number; \(^d\): original identifier in the database from which the proteome data was downloaded; \(^e\): amino acid (aa) sequence length; \(^f\): functional annotation in SchistoDB and UniProt; \(^g\): protein sequence domain(s) identified in the Pfam database; \(^h\): functional information in the literature available in UniProt; \(^i\): identifier in PubMed (PMID); \(^j\): scientific name of source organism; \(^k\): notes from this work.