MCP Papers in Press. Published on February 21, 2008 as Manuscript M700560-MCP200

Proteomic and phylogenetic analysis of the cathepsin L protease family of the helminth pathogen, *Fasciola hepatica*: expansion of a repertoire of virulence-associated factors.

Mark W. Robinson[1,2], Jose F. Tort[1,3], Jonathan Lowther[1], Sheila M. Donnelly[1], Emily Wong[1], Weibo Xu[1], Colin M. Stack[1,4], Matthew Padula[5], Ben Herbert[5], John P. Dalton[1]

[1]Institute for the Biotechnology of Infectious Diseases (IBID), University of Technology Sydney (UTS), Level 6, Building 4, Corner of Thomas & Harris Street, Ultimo, Sydney, NSW 2007, Australia.

[2]School of Medical Sciences, Institute of Medical Sciences, University of Aberdeen, Aberdeen, Scotland, UK, AB25 2ZD

[3]Departamento de Genetica, Facultad de Medicina, UDELAR, Gral. Flores 2125, CP 11800, Montevideo, Uruguay

[4]presently at, University of Western Sydney (UWS)

[5]Proteomics Technology Centre of Expertise, University of Technology Sydney (UTS), Level 6, Building 4, Corner of Thomas & Harris Street, Ultimo, Sydney, NSW 2007, Australia.

Corresponding author: Mark Robinson. Institute for the Biotechnology of Infectious Diseases, University of Technology Sydney, Building 4, Harris Street, Ultimo, Sydney, New South Wales 2007, Australia. Tel: 61-2-95144127; Fax: 61-2-9514201; Email: mark.robinson-2@uts.edu.au

Running title: Proteomic analysis of *F. hepatica* virulence factors

Key words: *Fasciola*, gene duplication, proteomics, cathepsin L, evolution, secretion

Abbreviations: ES, excretory-secretory; NEJ, newly excysted juvenile

Copyright 2008 by The American Society for Biochemistry and Molecular Biology, Inc.

Summary

Cathepsin L proteases secreted by the helminth pathogen *Fasciola hepatica* have functions in parasite virulence including tissue invasion and suppression of host immune responses. Using proteomic methods alongside phylogenetic studies we have characterised the profile of cathepsin L proteases secreted by adult *F. hepatica* and, hence, identified those involved in host-pathogen interaction. Phylogenetic analyses showed that the *Fasciola* cathepsin L gene family expanded by a series of gene duplications followed by divergence which gave rise to three clades associated with mature adult worms (Clades 1, 2, and 5) and two clades specific to infective juvenile stages (Clades 3 and 4). Consistent with these observations our proteomics studies identified representatives from Clades 1, 2 and 5 but not from Clades 3 and 4 in adult *F. hepatica* secretory products. Clades 1 and 2 account for 67.39% and 27.63% of total secreted cathepsin Ls, respectively, suggesting that their expansion was positively driven, and that these proteases are most critical for parasite survival and adaptation. Sequence comparison studies revealed that the expansion of cathepsin Ls by gene duplication was followed by residue changes in the S2 pocket of the active site. Our biochemical studies show that these changes result in alterations in substrate binding and suggest that the divergence of the cathepsin L family produced a repertoire of enzymes with overlapping and complementary substrate specificities that could cleave host macromolecules more efficiently. Although the cathepsin Ls are produced as zymogens containing a prosegment and mature domain, all secreted enzymes identified by MS were processed to mature active enzymes. The prosegment region was highly conserved between the Clades except at the boundary of prosegment and mature enzyme. Despite the lack of conservation at this section, sites for exogenous cleavage by asparaginyl endopeptidases and a Leu-Ser↓His motif for autocatalytic cleavage by cathepsin Ls were preserved.

Introduction

The helminth pathogens *Fasciola hepatica* and *F. gigantica* are the causative agents of liver fluke disease (fasciolosis) in sheep and cattle. Whilst infections of *F. hepatica* occur predominantly in regions with temperate climates, the parasite has been reported on all continents (except Antarctica) as a result of introduction by European settlers. In contrast, *F. gigantica* infections are largely restricted to tropical regions (1). Fasciolosis is also an important food-borne zoonotic disease of humans with estimates of 2.4 – 17 million people infected worldwide, while a further 91.1 million people are currently living at risk of infection (2-4). Human disease is particularly prevalent in the Andean countries of South America, Egypt, Iran and Vietnam where farming practices allow infected animals to roam among plants used for consumption (3, 4). Following ingestion of contaminated vegetation infective parasite larvae migrate from the intestine into the liver where they cause significant tissue injury and induce immunological-related damage before they move into the bile ducts. The parasites can remain for up to 1-2 years in the bile ducts of cattle and as long as 20 years in sheep (1).

Studies in our laboratory have shown that the most predominant molecules secreted by *F. hepatica* parasites *in vitro* are cathepsin L cysteine proteases (5, 6) and a recent analysis of bile taken from animals harbouring adult parasites confirmed that these enzymes also represent the majority of protein produced *in situ* (7). The secretion of proteases facilitates migration of the parasite through host tissue and the degradation of host macromolecules to provide essential free amino acids for the parasite (6). Furthermore, while it has been known for several decades that Fasciolid parasites secrete a variety of molecules that suppress the immune responses of their host (8-10), cathepsin L proteases are considered the principle participants; the parasite enzymes cleave host antibodies specifically in the hinge region to prevent antibody-mediated cell damage (5) and alter the function of cells of the innate and

adaptive cellular immune systems to suppress the development of protective Th1-driven responses (11).

The *F. hepatica* cathepsin L proteases are represented by a large gene family that expanded within the genus *Fasciola* by a series of gene duplications that resulted in a monophyletic group consisting of several discreet clades (12). The functional diversity of the various members of the gene family, and their relationship to pathogen virulence and host adaptation are of particular interest (6, 12). Using molecular clock analysis, Irving *et al.* (12) estimated that the duplications and divergence of the family occurred over the last 135 million years and the timing of duplications correlates with the evolution of rodents, ruminants and higher mammals. However, most of the duplications took place approximately 25 million years ago (MYA), at about the time climatic conditions favoured the development of grasslands and the expansion of common hosts of *F. hepatica* and suggests that the divergence of the cathepsin L protease family was important in the evolution and adaptation of the parasite to a wider host range (12). At the molecular level, this divergence involved changes in residues within the enzymes' active site, in particular at positions that are known to occupy the S_2 subsite and are critical to determining the substrate specificities of the proteases (6, 12, 13). It was suggested that these changes gave rise to proteases with overlapping and complementary specificities that allowed the parasite to degrade a wider variety of macromolecules (6, 12).

In the present study, we have analysed and characterised the profile of cathepsin L proteases secreted by adult *F. hepatica* by 2-dimensional gel electrophoresis (2-DE) and MS to determine the relative importance of the various cathepsin L groups to parasite virulence and adaptation. We found that these parasites secreted cathepsin L proteases that were representative of all three adult Clades identified by phylogenetics; however, the proteases of Clade 1 (FhCL1) and Clade 2 (FhCL2) were by far the most predominant which is consistent

with their greater divergence and expansion in the family. Proteases of Clades 3 and 4 that were identified only from the infectious juvenile stages were not detected in the adult stage secretory products suggesting a specific role for these proteases in initiating host infection through the intestinal wall. A sub-Clade of Clade 1 (FhCL1C) for which genes are known in *F. gigantica* but not in *F. hepatica* were also not represented in the *F. hepatica* parasite's secretory products and supports the suggestion of Irving *et al.* (12) that this sub-clade expanded after the separation of the two species. Comparative biochemistry and sequence alignments show that the *F. hepatica* repertoire of virulence–associated cathepsin Ls was established by a process of gene duplication followed by refinement of the active site residues that influence substrate specificity. Clade-specific variations also took place at the boundary between pro-segment and mature enzyme but specific cleavage sites required for activation of the cathepsin L zymogens are preserved.

Experimental Procedures

Alignments and Phylogenetic analysis.

Phylogenetic trees were created using 32 selected *F. hepatica* and *F. gigantica* cathepsin L DNA sequences. *Carica papaya* papain (GenBank accession number M15203) was used to root the Fasciolid sequences. All of the nucleotide sequences used for tree construction encoded the full length cathepsin L precursor protein including the prosegment region, but excluding the signal peptide. The DNA sequences were initially aligned using CLUSTAL W (14) and the trees were created using the boot-strapped (1000 trials) neighbour-joining method of MEGA version 4.0 (15), using the Kimura 2 parameter model with uniform rates for all sites. The GenBank accession numbers of the cathepsin L sequences used for alignment and phylogenetic analyses are as follows: FhCL 1A_ie1 (U62288), FhCL 1A_pe (AF490984), FhCL 1A_au3 (L33771), FhCL 1A_pl (AY277628), FhCL 1A_pt1

(AY519971), FhCL 1A_tr (AY573569), FhCL 1B_nl1 (AY279092), FhCL 1B_ar (AY029229), FgCL 1C_id (AF510856), FhCL 1C_jpA (AB010923), FhCL 1C_jpB (AB009306), FgCL 1C_cn (EF36899), FgCL 1C_thA (AF112566), FgCL 1C_thB (AF239264), FhCL 2_jn3 (AB010924), FhCL 2_ie2 (U62289), FhCL 2_chC (Z22765), FgCL 2_thD (AF239266), FhCL 3_nl64 (AJ279093), FgCL 3_thG (AF419329), FhCL 3_nl22 (AJ279091), FhCL 3_uy5 (EU287914), FhCL 3_uy9 (EU287915), FhCL 3_pl1 (EU191984), FhCL 3_pl2 (EU195859), FhCL 4_uy7 (EU287916), FhCL 4_uy18 (EU287917), FgCL 4_thH (AY428949), FgCL 5_thC (AF239265), FhCL 5_au4 (L33772) and FhCL 5_au5 (AF271385). The naming scheme reflects the different clades identified and the origin of the sequences represented by the internet country domain.

The prosegment regions of *Fasciola hepatica* and *Fasciola gigantica* cathepsin L proteinases were aligned using CLUSTAL W (14). The 26 amino acid sequences used (residues P1 to P91, fluke cathepsin L numbering) were truncated by removal of the predicted N-terminal signal peptide and the mature enzyme sequence. Amino acid consensus sequences were created by MULTALIN (16) and inserted manually into the alignment. The residues lining the S2 pocket of the active site of the various fluke cathepsin Ls were identified using a combination of primary sequence alignments (14) and analysis of the recently determined atomic structure of *F. hepatica* cathepsin L1 (17; PDB ID: 2O6X) and are shown in Table 2.

Preparation of F. hepatica excretory-secretory proteins.

Adult *Fasciola hepatica*, 16 weeks post-infection, were recovered from the liver tissue and bile ducts of Merino sheep (experimentally infected with 200 metacercariae) and washed in pre-warmed (37°C) 0.1 M PBS pH 7.3. Flukes were then transferred to pre-warmed (37°C) RPMI 1640 medium (Invitrogen) containing 2 mM L-glutamine, 30 mM HEPES, 0.1% (w/v) glucose and 2.5 µg/ml gentamycin and incubated for 8 h at 37°C. The culture medium

containing *F. hepatica* excretory and secretory (ES) proteins was pooled and concentrated using Centricon columns (Millipore) with a 3 kDa molecular weight cut-off to a final concentration of 1 mg/ml and stored in aliquots at -20°C.

2-dimensional electrophoresis and gel imaging

F. hepatica ES proteins (100 µg) were precipitated with 5 volumes of room temperature acetone and recovered by centrifugation at 5000 x g for 10 min. The protein pellets were resuspended in ProteoPrep (Sigma) extraction solution No.4 (7M urea, 2M thiourea, 1% $C7B_zO$, 40 mM Tris). The samples were reduced with 5 mM tributylphosphine and alkylated with 20 mM acrylamide monomer in a single 90 minute step. Excess acrylamide was subsequently quenched by the addition of 10 mM DTT. For separation in the first dimension, samples (150 µl) were actively loaded by re-hydrating 11 cm pH 4-7 ReadyStrip IPG strips (Bio-Rad) with 100 µl of 7 M urea, 2 M thiourea, 1 % $C7B_zO$ for 30 min. IEF was carried out using a 3 h convex ramp from 100 V to 3 kV with a further 5 h linear ramp to 10 kV where the voltage was held until 100 kVh was reached using an IsoelectrIQ^2 IEF device (Proteome Systems). After focussing, IPG strips were equilibrated in 7 M urea, 250 mM Tris-HCl (pH 8.8) and 1 % SDS (w/v) for 25 min. For separation in the 2nd dimension, the IPG strips were laid on 4-12 % Criterion XT gels (Bio-Rad) and run at 200 V for 45 min or until completion. After separation, proteins were visualised by staining with Flamingo fluorescent protein stain (Bio-Rad). Stained 2-D gels were imaged with a PharosFX laser imaging system (Bio-Rad) and normalised spot quantities were determined using PDQuest v8.01 software (Bio-Rad). Spot quantity represents the total intensity of a spot in an image and corresponds to the total amount of protein within a spot in a gel. In the present study, Gaussian analysis was used to determine the spot quantities using the formula: spot height x π x σ_x x σ_y where spot height is the peak of Gaussian representation of the spot, σ_x is the standard deviation of the Gaussian

distribution of the spot in the x axis and σ_y is the standard deviation of the Gaussian distribution of the spot in the y axis.

Mass spectrometry.

For excision of peptide spots, gels were overstained with colloidal Coomassie blue G250 (Sigma) overnight before destaining with 10 % methanol (v/v) and 7 % acetic acid (v/v). Selected protein spots were excised using an EXQuest robotic spot cutter with PDQuest software (Bio-Rad). The excised spots were in-gel digested with trypsin (Promega) and the peptides were analysed by nano liquid chromatography electrospray ionisation tandem mass spectrometry (nanoLC-ESI-MS/MS) using a Tempo nanoLC system (Applied Biosystems) with a C18 column (Vydac) coupled to a QSTAR Elite QqTOF mass spectrometer running in IDA mode (Applied Biosystems). Peak list files generated by the Protein Pilot v1.0 software (Applied Biosystems) using default parameters were exported to a local MASCOT v2.1.0 (Matrix Science) search engine for protein database searching.

Database searching.

MS/MS data was used to search 3239079 entries in the MSDB (20060809) database using MASCOT v2.1.0 (Matrix Science) with the enzyme specificity set to trypsin. Propionamide (acrylamide) modification of cysteines was used as a fixed parameter and oxidation of methionines was set as a variable protein modification. The mass tolerance was set at 1.0 Da for precursor ions and 0.3 Da for fragment ions. Only 1 missed cleavage was allowed. Matches achieving a molecular weight search (MOWSE) score >70 were considered to be significant (18, 19). However, other criteria were considered in assigning a positive identification including concordance between the calculated theoretical molecular mass and pI values of the protein and the observed position of the peptide by 2-DE. In order to account for

matches to multiple members of the *Fasciola* cathepsin family, peptides specific to individual enzymes or clades were looked for (see results section).

Enzyme assays and kinetics with fluorogenic peptide substrates

The activity of recombinant *F. hepatica* cathepsin L1 (FhCL 1A_ie1) and cathepsin L2 (FhCL 2_ie2) enzymes (20) was determined by a fluorometric assay using the synthetic substrates Z-Phe-Arg-NHMec, Z-Leu-Arg-NHMec and Z-Pro-Arg-NHMec. Initial rates of hydrolysis of the fluorogenic dipeptide substrates were measured by monitoring the release of the fluorogenic leaving group (NHMec) at an excitation wavelength of 370 nm and an emission wavelength of 460 nm using a Bio-Tek KC4 microfluorometer. The kinetic constants k_{cat} and K_M values were determined by nonlinear regression analysis. Initial rates were obtained at 37°C over a range of substrate concentrations spanning K_M (0.2 – 200 μM) and at fixed enzyme concentrations (0.5 – 5 nM). Assays were performed in 0.1 M PBS pH 7.3, and 100 mM sodium acetate buffer, pH 5.5, each containing 2.5 mM DTT and 2.5 mM EDTA.

Results

Phylogenetic analysis of Fasciola cathepsin L sequences

The evolutionary relationship between *F. hepatica* and *F. gigantica* cathepsin L gene sequences (46 in all) deposited in the public databases and at http://www.sanger.ac.uk/Projects/Helminths/ was investigated at the molecular level by constructing a bootstrapped neighbour-joining tree. Fourteen of the *Fasciola* cathepsin L sequences were not full-length cDNAs and/or differed by only a small number of nucleotides and thus likely represent different alleles rather than individual genes (12, 21). This is possible, and perhaps likely, given the occurrence of triploid Fasciolids (with an extra allele available) in both temperate and tropical regions (22-24). However, Grams *et al*. (21)

estimated from Southern blot analysis that at least 10 cathepsin L genes, formed by duplication events, exist in *F. gigantica* but that other, more divergent, sequences would not have been detected as a result of the stringent hybridisation conditions used. Until significant genomic sequence information is available for a member of the genus, the contribution of allelic variation to the current repertoire of *Fasciola* cathepsin Ls remains to be fully determined.

A phylogenetic analysis of 24 *F. hepatica* and 8 *F. gigantica* full-length sequences revealed that these separated into five well-supported clades that arose by a series of gene duplications (Figure 1). The two initial gene duplications separated the cathepsins isolated from the infective newly excysted juvenile (NEJ) parasites (Clade 3, FhCL3 and Clade 4, FhCL4) from three clades expressed in the adult worm stage (clades FhCL1, 2, and 5). Following this, another gene-duplication led to the separation of the adult clades FhCL1 and FhCL5 from clade FhCL2. The phylogenetic tree also showed that the *Fasciola* clade FhCL1 has undergone the greatest expansion and is represented by three distinct sub-clades: FhCL1A, FhCL1B and FhCL1C.

It is noteworthy that all clades contain sequences from both *F. hepatica* and *F. gigantica*. However, sub-clades FhCL1A and FhCL1B are composed exclusively of *F. hepatica* sequences whereas clade FhCL1C contains only one *F. hepatica* cathepsin L. This latter sequence was identified from a Japanese isolate that is most likely a *F. hepatica*/*F. gigantica* hybrid (12, 23), which indicates that clade FhCL1C may be exclusive to *F. gigantica.* This phylogenetic analysis, therefore, indicates that the early duplication events in the cathepsin L gene family occurred before the speciation of the *F. hepatica* and *F. gigantica* fasciolids and that expansion of sub-clades FhCL1A/FhCL1B and FhCL1C occurred after the segregation of these two species. Irving *et al.* (12) made a similar observation and suggested

that divergence of the FhCL1 clade reflected adaptation of the 'temperate' *F. hepatica* and the 'tropical' *F. gigantica* to different host species (see Discussion).

Identification of cathepsin L proteases secreted by adult F. hepatica.

ES proteins secreted by isolated adult *F. hepatica* during *in vitro* cultivation were precipitated from culture supernatants and analysed by 2-DE. Using the current sample preparation method and pH range, 30 major peptide spots and 50 less intense peptides of varying molecular mass and pI values were visible on 2-D gels of *F. hepatica* ES proteins (Figure 2A). To identify the secreted cathepsin L enzymes, these 80 peptide spots were excised from 2-D gels and a proteomic analysis was carried out using nanoLC-ESI-MS/MS. The resulting ion mass data was used to search databases. In line with previous proteomic studies of helminth secretory proteins, a MOWSE score >70 was considered to be significant (18, 19).

Although the predicted molecular mass of all *F. hepatica* cathepsin Ls (mature form) used to generate the phylogram is approximately 24 kDa, the predicted pI values differ considerably and range from 4.54 for FhCL 2_chC to 8.13 for FhCL 3_nl64 (based on conceptual translation of the cDNAs). Fifteen of the major peptide spots were definitively identified as *F. hepatica* cathepsin L proteases (Table 1; Figure 2B). These displayed a similar observed molecular mass of 24 kDa whilst the pI values ranged from 4.36 (spot 20) to 6.42 (spot 7). A further 8 peptides (spots 8, 9, 13, 14, 15, 16, 18 and 19) were also matched to fluke cathepsin Ls, albeit these were assigned on the basis of single-peptide matches and at present their identity could be regarded as tentative (data not shown). The position of spots 18 and 19 on the 2-D gels suggests that they had not resolved correctly while spots 8, 9, 13, 14, 15 and 16 are likely the result of proteolytic degradation since they migrate at a low molecular size. The 15 definitively identified cathepsin Ls represented sub-clade FhCL1A (7 spots), sub-clade FhCL1B (4 spots), clade FhCL2 (3 spots) and clade FhCL5 (1 spot) (Figure 2A).

Representatives of Clades FhCL3 and FhCL4 were not detected in the adult ES proteins. However, this substantiates our phylogenetic data since sequences derived from the genes belonging to these clades have been isolated only from infective juveniles and therefore may not be expected to be expressed and secreted by adult parasites (25). *F. hepatica* enzymes representing Clade FhCL1C were also not identified; this phylogenetic Clade was only detected in *F. gigantica* or *F. hepatica/F. gigantica* hydrids (see section above).

Fasciola cathepsin Ls are highly conserved at the amino acid level thus it is likely that the tryptic peptides will match to multiple members of the enzyme family. In order to account for this potential redundancy, matches to sequence-specific or clade-specific peptides were looked for. All peptide spots that were identified as clade 1 enzymes (with the exception of spot 5) matched with a clade 1-specific sequence VTGYYTVHSGSEVELK (ion *m*/*z* 590). The tryptic digest of spot 5 displayed an ion following MS/MS (*m*/*z* 590.37) that is likely to be the same peptide although it did not match as a result of the highly stringent criteria used for database searching in the current study. Another clade 1-specific peptide, NSWGLSWGER (ion *m*/*z* 596), was matched in almost all these spots (with the exception of spots 1 and 17). Spot 1 was identified as FhCL1A_pe and this was supported by the presence of an ion (*m*/*z* 621) corresponding to an amino acid sequence of NSWGSYWGER that is found only in this enzyme. Moreover, there was an almost exact match between the theoretical and observed molecular mass and pI values for this peptide spot. Spots 2 and 3 were identified as *Fasciola* cathepsin Ls FhCL1A_pt1 and FhCL1A_pt2 respectively. The mature enzymes show 96% sequence identity at the amino acid level but were distinguished by the presence of a FhCL1A_pt2-specific ion (*m*/*z* 835 with 2 oxidised methionines) in the mass spectra obtained from spot 3. In the absence of any FhCL1A_pt1-specific ions from spot 2, its assignment as this enzyme must remain tentative although the presence of the 590 and 596 ions support its identification as a clade 1 enzyme. Spots 4-7 were assigned to

FhCL1A_tr. The amino acid sequence ESGYVTGVK that was matched with the 470 ion of spots 4 and 5 is also present in the primary sequences of FhCL1A_pt2 and FhCL1B_nl1. However, no FhCL1A_pt2- or FhCL1B_nl1-specific ions were detected in the mass spectra obtained from these spots supporting their identification as FhCL1A_tr enzymes. The 470 ion was not observed for spots 6 and 7 but the presence of the 590 and 596 ions again support their identification as clade 1 enzymes. Four spots were identified as FhCL1B_nl1 (spots 10-12 and 17). An ion (*m*/*z* 724) matching an amino acid sequence of FGLETESSYPYR was present in the mass spectra from all four spots. Although this sequence is also present in clade 5 cathepsin Ls, the presence of clade 1-specific ions (*m*/*z* 590 and *m*/*z* 596) supports their assignment as FhCL1B_nl1 enzymes. A further 3 spots were identified as the clade 2 enzyme FhCL2_chC (spots 20-22). The mass spectra from these spots contained several ions that match clade 2-specific amino acid sequences including DYYYVTEVK (*m*/*z* 590), VTGYYTVHSGDEIELK (*m*/*z* 604), LTHAVLAVGYGSQDGTDYWIVK (*m*/*z* 798) as well as an ion (*m*/*z* 856) that matched with the FhCL2_chC-specific amino acid sequence ASASFSEQQLVDCTR. The theoretical and observed molecular mass and pI values were also in close agreement for these spots. Consequently, the assignment of spots 20-22 as FhCL2_chC is considered to be very robust. A single spot (spot 23) was assigned to the clade 5 enzyme, FhCL5_au5. The mass spectra from spot 23 contained 2 ions (*m*/*z* 724 and 863) that match peptides from clade 5 cathepsin Ls but also match FhCL1B_nl1 and clade 2 sequences respectively. The absence of any other clade 1- or clade 2-specific ions together with the excellent match between the theoretical and observed molecular mass / pI values strongly supports the identification of spot 23 as a clade 5 cathepsin L.

Relative expression levels of secreted F. hepatica *cathepsin Ls*

Densitometry was performed on peptide spots from several imaged 2-D gels of *F. hepatica* ES proteins to quantify the level of expression of each enzyme relative to each other. The raw data for each spot was converted to a percentage of the total amount of cathepsin L in the gels (Fig. 1; Table 1). The clade FhCL1 cathepsins accounted for 67.39% of total cathepsin protein detected, with sub-clades FhC1A and FhCL1B representing 35.30% and 32.09% of this figure, respectively. The clade FhCL2 accounted for 27.63% of the total secreted cathepsin protein levels whereas the single spot representing clade FhCL5 enzymes accounted for 4.98%.

Variation in the S2 subsite of the active site and its influence on enzyme kinetics

The substrate specificity of cathepsin L cysteine proteases is determined by the composition and arrangement of amino acids that create the biochemical characteristics of the S2 subsite of the active site (26). Using primary sequence alignments and analysis of the atomic structure of *F. hepatica* cathepsin L1, the key residues in this pocket that interact with the P2 amino acid of the substrate have been identified as those situated at positions 67, 68, 133, 157, 160 and 205 (17). A comparison of the amino acids that occupy these positions in the various phylogenetic clades of the *F. hepatica* cathepsin L family is presented in Table 2 and reveals a number of substitutions that could have a critical impact on their substrate preferences. In particular, most variation between the clades occurs in residues at positions 67 (Leu, Tyr, Trp or Phe) and 205 (Leu, Val or Phe), residues considered to be most important for substrate recognition. Variation was also observed at position 157 which lies at the opening of the S2 pocket and acts as a ‘gatekeeper’ residue that influences the type of residue that can access the pocket (17). Interestingly, the only difference found between the sub-clades FhCL1A and

FhCL1B of clade FhCL1 is seen at position 157; however, the amino acid residing here was always a hydrophobic Leu or a Val.

To demonstrate the influence of the variations found in the S2 pocket of the active site, substrate-kinetic studies were performed using purified recombinant proteases derived from sequences representing members of the major secreted *F. hepatica* clades, FhCL1 (FhCL 1A_ie1) and FhCL2 (FhCL 2_ie2). Kinetic constants were obtained for three fluorogenic peptide substrates possessing different residues at their P_2 sites: Z-Phe-Arg-NHMec, Z-Leu-Arg-NHMec and Z-Pro-Arg-NHMec (Table 3). Overall, *F. hepatica* cathepsin L1 showed an affinity (k_{cat}/K_m) for the substrates in the following order: Z-Leu-Arg-NHMec > Z-Phe-Arg-NHMec >> Z-Pro-Arg-NHMec whereas for *F. hepatica* cathepsin L2 the order of affinity was Z-Leu-Arg-NHMec >> Z-Phe-Arg-NHMec ≈ Z-Pro-Arg-NHMec. Both enzymes showed higher affinities for substrates Z-Phe-Arg-NHMec and Z-Leu-Arg-NHMec at lower pH (on average a 2 fold increase in affinity at pH 5.5 compared with pH 7.3). For both these substrates the catalytic rates (k_{cat}/K_m) of cathepsin L1 was markedly greater than cathepsin L2 (approximately 25 times greater for Z-Phe-Arg-NHMec and on average 8 times greater for Z-Leu-Arg-NHMec at both pH values). However, while cathepsin L1 and L2 showed a lower affinity for the substrate containing a proline at the P_2 position, cathepsin L2 displayed a greater preference for Z-Pro-Arg-NHMec with an approximately 6 fold greater affinity for this substrate at pH 5.5 and 3 fold greater affinity at pH 7.3 than cathepsin L1.

Bioinformatic analysis of Fasciola cathepsin L prosegments

Cathepsin Ls are stored in specialised secretory vesicles within the parasites gut epithelial cells as inactive zymogens consisting of a prosegment and mature enzyme domain (27). The prosegment is removed by catalytic cleavage following secretion into the parasite intestine to

reveal an active mature protease (28). Our MS studies show that all cathepsin L proteases secreted by adult *F. hepatica* lack the prosegment and are thus fully-processed mature enzymes.

Phylogenetic analyses using only the prosegment domains of the *Fasciola* cathepsin Ls used in the present study produced a tree similar to that presented in Fig. 1 indicating parallel adaptation of the prosegment and mature domains (data not shown). An alignment of the prosegments (residues P1 to P91) shows that the N-terminal and intermediate regions (residues P1 to P70) of these enzymes are remarkably conserved across all cathepsin L clades (Figure 3). What is particularly noteworthy, however, is that the C-terminal portion of the *F. hepatica* cathepsin L prosegments (21 residues, P70 to P91) shows striking variability between the phylogenetic clades but is conserved within each clade. This is particularly evident in the final five residues which form the boundary between the prosegment and mature enzyme and gives each clade a signature sequence as shown in Figure 3. Despite the lack of conservation in this signature section, each still reserves an Asn (N) residue and, thereby, retains a highly specific cleavage site for the *trans*-processing enzyme asparaginyl endopeptidase which initiates or 'kick-starts' the enzyme activation process (28). Another highly conserved feature within the non-conserved C-terminal region of the prosegment is a Leu-Ser↓His motif that we have shown is essential for autocatalytic cleavage between cathepsin L proteases once activation is initiated by the asparaginyl endopeptidase (20). As can be seen in Figure 3 this cleavage site is conserved in all the enzyme clades of adult *F. hepatica*, with the notable change to Leu-Ser↓Arg in the adult FhCL2 Clade and to Leu-Ser↓Asp in the infective juvenile FhCL3 Clade.

Discussion

The longevity of a parasite species in nature is dependent on its ability to invade new hosts (29). *F. hepatica*, a helminth parasite of cattle and sheep, is of European origins but its geographical distribution has expanded over the last five centuries as a result of global colonisations by Europeans, and the associated continual export of livestock. The parasite's expansion, however, has been greatly facilitated by what seems to be a remarkable ability to adapt to new hosts; thus the parasite can develop, mature and produce viable off-spring even in very recently encountered species such as llama and alpaca in South America, camels in Africa and kangaroos in Australia (3). *F. hepatica* also infects a wide variety of wild animals including deer, rabbits, hare, boars, beavers and otters which, collectively, are major reservoir host populations that contribute significantly to the world-wide dissemination of the disease and to its local transmission patterns. *F. gigantica* diverged from *F. hepatica* 17 MYA (12) and penetrated more tropical regions in Asia and the Far East where it is the predominant parasitic disease of cattle and water buffalo (4). In the last 15 years human fasciolosis has emerged as an important zoonotic disease in the Andean countries of South America, Egypt, Thailand and Iran, all regions where animal disease is highly prevalent and farm management practices allow contamination of edible aquatic plants by the infective juvenile larvae (3, 30).

Cathepsin L proteases most likely played, and continue to play, a critical role in adapting these helminth parasites to new host species (6, 12, 13). *Fasciola* cathepsin Ls have several well-defined functions that are essential to parasite-host biology including degradation of host macromolecules and suppression of immune responses (reviewed in 6). In the present study we have shown that adult *F. hepatica* secretes a range of these proteases, with varying substrate specificities. Previous proteomic studies have shown that the only proteins secreted by adult worms residing in the bile ducts are cathepsin L proteases (7), and our *in vitro* experiments show that these are secreted in abundance (with other minor proteins that may be

artefacts of the culture method), estimated at 0.5 – 1.0 µg per parasite per hour (28). This high level of enzyme production is not surprising when it is considered that once inside the bile duct the parasite needs to digest a large quantity of host red blood cells (350 x 10^6/hr) to support the enormous production of progeny (30-50,000 eggs/day/worm; 28). The heavy reliance on cathepsin Ls is also reflected in a high gene transcription rate and, therefore, approximately 10% of cDNAs in a ~5000 EST library prepared from adult fluke mRNA were found to encode cathepsin L proteases (28; http://www.sanger.ac.uk/Projects/Helminths/).

Phylogenetic analysis of *Fasciola* cathepsin L gene sequences showed that this family of proteases arose by gene duplication and divergence into five main clades (6, 12 and this study). However, it was not clear whether proteases from all of these clades are secreted by the parasite and, hence, play a role in host-parasite interaction. Therefore, we carried out a proteomic characterisation of the secretory products of adult parasites, with particular focus on identifying proteases. We found peptide spots representative of the adult cathepsin L Clades FhCL1, FhCL2 and FhCL5 but no members of the FhCL3 and FhCL4 Clades were identified. This suggests that expression of *Fasciola* cathepsin Ls is developmentally regulated and is consistent with the specific expression of FhCL3 and FhCL4 genes by juvenile flukes. Alternatively, the FhCL3 and FhCL4 genes may encode proteases that are not secreted extra-corporeally by adult *F. hepatica* but, instead, perform internal functions such as in protein turnover, membrane biogenesis or egg production.

The three adult cathepsin L clades were not represented in equal proportions at the protein expression level in the secretions, and this somewhat corresponded to the diversity of genes found in the phylogenetic studies (Fig. 1 and 2; Table 1). Clades FhCL1 and FhCL2 contain eight and four members in the current gene family and account for 67.39% (11 spots) and 27.63% (3 spots) of the total secreted cathepsin Ls observed by 2-DE respectively. In contrast, the Clade FhCL5 contained three members but was represented by a single peptide

spot and contributed only 4.98% to the total amount of secreted proteases. The level of gene divergence and protein production presumably reflects the relative need for each enzyme to perform the functions they play during parasite infection.

An analysis of the critical S1 and S2 active site residues of the cathespin L family indicate that all fluke cathepsin Ls are functional (i.e. no degenerate active sites) which implies a positive evolutionary drive to create a repertoire of functionally active proteases. However, examination of the S2 subsite residues that are responsible for determining substrate specificity clearly demonstrates positive selection at the three most influential positions, i.e at residues 67, 157 and 205. Our biochemical data using recombinant versions of a Clade 1 (FhCL1A) and Clade 2 (FhCL2) proteases presented in Table 3 show how significant changes in these positions can be; for example, FhCL1A (Leu67, Val157 and Leu205) cleaves substrates with the hydrophobic residues (Phe and Leu) in the P2 position with catalytic rates (k_{cat}/K_m) that are 25- and 8-fold greater, respectively, than FhCL2 (Tyr67, Leu157 and Leu205). In contrast, FhCL1A exhibits a low preference for substrates with Pro in the P2 position. This suggests that the active site of the Clade 1 enzymes have opened up to accommodate larger residues and can cleave these with greater efficiency. FhCL2, on the other hand, readily accommodates a P2 Pro residue suggesting that the evolutionary trajectory followed by members of Clade 2 favoured active site changes that allowed a better accommodation of the short and bulky hydrophobic residue Pro. Recently, Stack *et al.* (17) correlated the ability to accommodate Pro in the S2 subsite in FhCL2 with the capacity to cleave native collagen, which contains a repeat motif of Gly-Pro-X motif. Since collagen is a predominant component of interstitial matrices, including the bile duct wall, this property would have provided the parasite with the tools for degrading and penetrating host tissues, thus enabling it to feed on blood from underlying vessels. Interestingly, FhCL5 which diverged from FhCL2 prior to its divergence from FhCL1 (see Fig. 1) exhibits an intermediate

S2 subsite (Leu67, Leu157 and Leu205) and does not readily accommodate substrates with a P2 proline residue (31). Thus, it can be envisaged that *Fasciola* evolved a series of enzymes with overlapping substrate specificities to create a more efficient tissue-degrading protease secretory system.

The FhCL1A and the FhCL1B sub-clades of FhCL1 consist exclusively of *F. hepatica* sequences. In contrast, the FhCL1C sub-clade is represented by four cDNA sequences from *F. gigantica* and only one from a *F. hepatica/F. gigantica* hybrid identified from a Japanese isolate (12, 23). This clear segregation suggests that the expansion of sub-clades FhCL1A/FhCL1B and FhCL1C occurred after the divergence of *F. hepatica* and *F. gigantica* species. Irving *et al.* (12) suggested that the separate expansion of these sub-clades in the two *Fasciola* species reflects the adaptation of these parasites to different host breeds and species: *F. hepatica* typically infects cattle and sheep in temperate areas whereas *F. gigantica* has penetrated tropical regions in Asia and the Far East where it is the predominant parasitic disease of cattle and water buffalo (4).

The significance of the group of cathepsin L proteases in Fasciolid biology is further emphasised by the complete absence of other endoproteases in the secretory proteome of adult flukes. Previous studies have detected significant levels of transcript for cathepsin B cysteine proteases in *F. hepatica* and *F. gigantica* infective juveniles and immature liver stages, but the expression of these in adult parasites was low (32-34). Additionally, immunoblotting experiments using antisera raised against a recombinant *F. hepatica* cathepsin B detected this protease in proteins secreted by immature parasites but not by adult parasites (35). These data imply that the endoproteolytic needs of the mature parasites residing in the bile duct, which primarily involves feeding on host blood since the parasite exists in an immunologically-safe location, are met solely by the cathepsin Ls (6). By contrast, the infective parasite larvae must not only penetrate the intestinal wall of the host,

but also make their way across the large tissue mass of the liver and at the same time defend itself from immune attack. Thus, a greater range of enzyme types may be required to complete these multiple tasks. A recent report using RNA interference methodology demonstrated that knocking down of either cathepsin B or cathepsin L transcripts in infective juveniles reduced their ability to penetrate the intestinal wall of the host and implied a role for both enzyme types in host invasion (36). Interference of cathepsin L expression, however, had the greater impact on invasion, possibly because the infective juvenile-specific cathepsin Ls are specifically adapted to this function. Cysteine proteases have also been implicated in the invasion process by the skin-penetrating parasites *Schistosoma mansoni* (37) and *Trichobilharzia reagenti* (38) pointing to a common mechanism of host invasion amongst these trematodes.

Gene duplication is considered a prime means by which all organisms generate molecules with new functions; however, the mechanism by which this occurs has become the topic of much debate, largely due to the increased availability of genome sequence information (39-42). In the much-cited early theory of gene duplication put forward by Ohno (43) the new function appears subsequent to the duplication event and the most common fate of one of the pair of the duplicated genes is loss of function. However, more recent theories suggest that ancestral genes possess the multiple functions prior to duplication which are then preserved and refined in the duplicated genes by subsequent positive selection to evolve paralogs with distinct functions (44, 45). It is likely that the expansion of the *Fasciola* cathepsin L gene family arose by a mechanism of duplication and refinement given the overlap in substrate specificity of the family members. In this scenario, the gene expressed by the infective juveniles, FhCL3, is the ancestral gene which possessed the combined properties of the FhCL4, FhCL2, FhCL5 and FhCL1 lineages it gave rise to. During the refinement process however, and in particular in the adult stage, it is possible that the FhCL5 clade,

which is represented by just a few genes and is poorly expressed, became reduced in importance and that its role was gradually overtaken by the expanding and highly expressed FhCL1 and FhCL2 enzymes. The partnership between the existing FhCL1 and FhCL2 enzymes would be expected to be maintained since each possesses quite distinct specificities that provide the adult parasite with the advantage of being able to cleave a wider variety of target protein/tissue substrates and with greater efficiency.

F. hepatica cathepsins are synthesised within gastrodermal epithelial cells and stored in secretory vesicles as inactive zymogens (46). The zymogens contain an N-terminal extension or prosegment that regulates cathepsin activity by binding to the substrate cleft (47) and acts as a molecular chaperone to ensure correct folding of the enzyme (48). Activation by removal of the prosegment must take place within the parasite gut before secretion into the medium as our studies show that all the cathepsin L proteases produced by adult *F. hepatica* lack the prosegment. We proposed that the cleavage events that lead to the removal of the prosegment takes place in two steps (a) a bimolecular process whereby a small number of cathepsin L molecules are *trans*-activated by another enzyme, asparaginyl endopeptidase which is also localised within the intestinal epithelial cells of *Fasciola* (49), through cleavages at the C-terminal side of Asn residues lying close to the junction of the prosegment and mature domain, followed by (b) the rapid cleavage of prosegments from other cathepsin Ls by the activated cathepsin L molecules through cleavage at a Leu-Ser↓His motif (20). Despite the high variability in the C-terminal region of the prosegment that forms a boundary with the mature domain, Asn residues and the Leu-Ser↓His motif are preserved in this section in all clades and, therefore, supports our idea that these cleavage sites are critical for zymogen processing and activation.

Our proteomic and phylogenetic analysis has shown that adult *F. hepatica* evolved a large repertoire of proteases of one mechanistic class, cysteine proteases, to perform various

critical functions that allow the parasite to survive within the host. It is likely that the expansion of the cathepsin L family was central to the success of this parasite in terms of its ability to infect and adapt to new hosts. Because of their pivotal role in Fasciolid biology cathepsin L proteases are, therefore, considered primary candidates for the development as first generation vaccines against fasciolosis (4, 6).

Acknowledgments

John P. Dalton is a recipient of the NSW Government BioFirst Award in Biotechnology. Mark W. Robinson is supported by a Wain International Travel Fellowship from the British Biotechnology and Biological Sciences Research Council (BBSRC) and research support grants from the ARC/NHMRC Research Network for Parasitology and Merial Animal Health Ltd.

References

1. Andrews, S.J. (1999) In *Fasciolosis* (Dalton, J.P., ed) pp 1-29, CABI, Oxford, United Kingdom.
2. Keiser, J. and Utzinger, J: (2005) Food-borne trematodiasis: an emerging public health problem. *Emerg. Inf. Dis.* **11**, 1507-1514.
3. Mas-Coma, S., Bargues, M.D. and Valero, M.A. (2005) Fascioliasis and other plant-bourne trematode zoonoses. *Int. J. Parasitol.* **35**, 1255-1278.
4. MacManus, D.P. and Dalton, J.D. (2006) Vaccines against the zoonotic trematodes *Schistosoma japonicum*, *Fasciola hepatica* and *Fasciola gigantica*. *Parasitology* **133**, S43-61.

5. Smith, A.M., Dowd, A.J., Heffernan, M., Robertson, C.D. and Dalton, J.P. (1993) *Fasciola hepatica*: a secreted cathepsin L-like proteinase cleaves host immunoglobulin. *Int. J. Parasitol.* **23**, 977-983.
6. Dalton, J.P., Neill, S.O., Stack, C., Collins, P., Walshe, A., Sekiya, M., Doyle, S., Mulcahy, G., Hoyle, D., Khaznadji, E., Moire, N., Brennan, G., Mousley, A., Kreshchenko, N., Maule, A.G. and Donnelly, S.M. (2003) *Fasciola hepatica* cathepsin L-like proteases: biology, function, and potential in the development of first generation liver fluke vaccines. *Int. J. Parasitol*. **33**, 1173-1181.
7. Morphew, R.M., Wright, H.A., LaCourse, E.J., Woods, D.J. and Brophy, P.M. (2007) Comparative proteomics of excretory-secretory proteins released by the liver fluke *Fasciola hepatica* in sheep host bile and during *in vitro* culture ex host. *Mol. Cell. Proteomics.* **6**, 963-972.
8. Mulcahy, G., O'Connor, F., Clery, D., Hogan, S.F., Dowd, A.J., Andrews, S.J. and Dalton, J.P. (1999) Immune responses of cattle to experimental anti-*Fasciola hepatica* vaccines. *Res. Vet. Sci.* **67**, 27-33.
9. Brady, M.T., O'Neill, S.M., Dalton, J.P. and Mills, K.H. (1999) *Fasciola hepatica* suppresses a protective Th1 response against *Bordetella pertussis*. *Infect. Immun*. **67**, 5372-5378.
10. Donnelly, S., O'Neill, S.M., Sekiya, M., Mulcahy, G. and Dalton, J.P. (2005) Thioredoxin peroxidase secreted by *Fasciola hepatica* induces the alternative activation of macrophages. *Infect. Immun*. **73**, 166-173.
11. O'Neill, S.M., Mills, K.H. and Dalton, J.P. (2001) *Fasciola hepatica* cathepsin L cysteine proteinase suppresses *Bordetella pertussis*-specific interferon-gamma production *in vivo*. *Parasite Immunol*. **23**, 541-547.

12. Irving, J.A., Spithill, T.W., Pike, R.N., Whisstock, J.C. and Smooker, P.M. (2003) The evolution of enzyme specificity in *Fasciola* spp. *J. Mol. Evol*. **57**, 1-15.
13. Tort, J., Brindley, P.J., Knox, D., Wolfe, K.H. and Dalton, J.P. (1999). Helminth proteinases and their associated genes. *Adv. Parasitol.* **43**, 161-266.
14. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673-4680.
15. Kumar, S., Tamura, K. and Nei, M. (1994) MEGA: Molecular Evolutionary Genetics Analysis software for microcomputers. *Comput. Appl. Biosci*. **10**, 189-191.
16. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*. **16**, 10881-10890.
17. Stack, C.M., Caffrey, C.R., Donnelly, S.M., Seshaadri, A., Lowther, J., Tort, J.F., Collins, P.R., Robinson, M.W., Xu, W., McKerrow, J.H., Craik, C.S., Geiger, S.R., Marion. R., Brinen, L.S. and Dalton JP. (In Press) Structural and functional relationships in the virulence-associated cathepsin L proteases of the parasitic liver fluke, *Fasciola hepatica*. *J. Biol. Chem.* Dec 26 [Epub ahead of print].
18. Robinson, M.W. and Connolly, B. (2005) Proteomic analysis of the excretory-secretory proteins of the *Trichinella spiralis* L1 larva, a nematode parasite of skeletal muscle. *Proteomics* **5**, 4525-4532.
19. Robinson, M.W, Greig, R., Beattie, K., Lamont, D. and Connolly, B. (2007) Comparative analysis of the excretory-secretory proteome of the muscle larva of *Trichinella pseudospiralis* and *Trichinella spiralis. Int. J. Parasitol.* **37**, 139-148.
20. Stack, C.M., Donnelly, S., Lowther, J., Xu, W., Collins, P.R., Brinen, L.S. and Dalton, J.P. (2007) The major secreted cathepsin L1 protease of the liver fluke, *Fasciola*

hepatica: a Leu-12 to Pro-12 replacement in the nonconserved C-terminal region of the prosegment prevents complete enzyme autoactivation and allows definition of the molecular events in prosegment removal. *J. Biol. Chem.* **282**, 16532-16543.

21. Grams, R., Vichasri-Grams, S., Sobhon, P., Upatham, E.S. and Viyanant, V. (2001) Molecular cloning and characterization of cathepsin L encoding genes from *Fasciola gigantica. Parasitol. Int*. **50**, 105-114.
22. Fletcher, H.L., Hoey, E.M., Orr, N., Trudgett, A., Fairweather, I. and Robinson, M.W. (2004) The occurrence and significance of triploidy in the liver fluke, *Fasciola hepatica. Parasitology* **128**, 1-4.
23. Itagaki, T. and Tsutsumi, K. (1998) Triploid form of *Fasciola* in Japan: genetic relationships between *Fasciola hepatica* and *Fasciola gigantica* determined by ITS-2 sequence of nuclear rDNA. *Int. J. Parasitol*. **28**, 777-781.
24. Terasaki, K., Noda, Y., Shibahara, T. and Itagaki, T. (2000) Morphological comparisons and hypotheses on the origin of polyploids in parthenogenetic *Fasciola* sp. *J. Parasitol.* **86**, 724-729.
25. Harmsen, M.M., Cornelissen, J.B., Buijs, H.E., Boersma, W.J., Jeurissen, S.H. and van Milligen, F.J. (2004) Identification of a novel *Fasciola hepatica* cathepsin L protease containing protective epitopes within the propeptide. *Int. J. Parasitol.* **34**, 675-682.
26. Turk, D., Guncar, G., Podobnik, M. and Turk, B. (1998) Revised definition of substrate binding sites of papain-like cysteine proteases. *Biol. Chem.* **379**, 137-147.
27. Roche, L., Dowd, A.J., Tort, J., McGonigle, S., MacSweeney, A., Curley, G.P., Ryan, T. and Dalton, J.P. (1997) Functional expression of *Fasciola hepatica* cathepsin L1 in *Saccharomyces cerevisiae*. *Eur. J. Biochem.* **245**, 373-380.

28. Dalton, J.P., Caffrey, C.R., Sajid, M., Stack, C., Donnelly, S., Loukas, A., Don, T., McKerrow, J., Halton, D.W. and Brindley, P.J. (2006) Proteases in Trematode Biology. In *Parasitic Flatworms: Molecular biology, Biochemistry, Immunology and Physiology*, A.G. Maule and N.J. Marks, eds. Pp 348-368. Oxford, United Kingdom: CAB International.
29. Littlewood, D.T.J. (2006) The Evolution of Parasitism in Flatworms. In *Parasitic Flatworms: Molecular biology, Biochemistry, Immunology and Physiology*, A.G. Maule and N.J. Marks, eds. pp1-36. Oxford, United Kingdom: CAB International.
30. Parkinson, M., O'Neill, S.M. and Dalton, J.P. (2007) Endemic human fasciolosis in the Bolivian Altiplano. *Epidemiol. Infect.* **135**, 669-674.
31. Smooker, P.M., Whisstock, J.C., Irving, J.A., Siyaguna, S., Spithill, T.W. and Pike, R.N. (2000) A single amino acid substitution affects substrate specificity in cysteine proteinases from *Fasciola hepatica*. *Protein. Sci*. **9**, 2567-2572.
32. Heussler, V.T. and Dobbelaere, D.A. (1994) Cloning of a protease gene family of *Fasciola hepatica* by the polymerase chain reaction. *Mol. Biochem. Parasitol*. **64**, 11-23.
33. Wilson, L.R., Good, R.T., Panaccio, M., Wijffels, G.L., Sandeman, R.M. and Spithill, T.W. (1998) *Fasciola hepatica*: characterization and cloning of the major cathepsin B protease secreted by newly excysted juvenile liver fluke. *Exp. Parasitol.* **88**, 85-94.
34. Meemon, K., Grams, R., Vichasri-Grams, S., Hofmann, A., Korge, G., Viyanant, V., Upatham, E.S., Habe, S. and Sobhon, P. (2004) Molecular cloning and analysis of stage and tissue-specific expression of cathepsin B encoding genes from *Fasciola gigantica*. *Mol. Biochem. Parasitol.* **136**, 1-10.
35. Law, R.H., Smooker, P.M., Irving, J.A., Piedrafita, D., Ponting, R., Kennedy, N.J., Whisstock, J.C., Pike, R.N. and Spithill, T.W. (2003) Cloning and expression of the

major secreted cathepsin B-like protein from juvenile *Fasciola hepatica* and analysis of immunogenicity following liver fluke infection. *Infect. Immun.* **71**, 6921-6932.

36. McGonigle, L., Mousley, A., Marks, N.J., Brennan, G.P., Dalton, J.P., Spithill, T.W., Day, T.A. and Maule, A.G. (2008) The silencing of cysteine proteases in *Fasciola hepatica* newly excysted juveniles using RNA interference reduces gut penetration. *Int. J. Parasitol*. **38**, 149-155.
37. Dalton, J. P., Clough, K. A., Jones, M. K., and Brindley, J. P. (1997). The cysteine proteinases of *Schistosoma mansoni* cercariae. *Parasitology.* **114**, 105-112.
38. Kašný, M., Dalton, J.P., Mikeš, L. and Horák, P. (2007). Comparison of cysteine and serine peptidase activities in *Trichobilharzia regenti* and *Schistosoma mansoni* cercariae. *Parasitology.* **22**, 1-11.
39. Lynch, M. (2002) Gene duplication and evolution. *Science* **297**, 945-947.
40. Lynch, M. and Katju, V. (2004) The altered evolutionary trajectories of gene duplicates. *Trends Genet.* **20**, 544-549.
41. Davis, J.C. and Petrov, D.A. (2004) Preferential duplication of conserved proteins in eukaryotic genomes. *PLOS. Biol.* **2**, 318-326.
42. Shakhnovich, B.E. and Koonin, E.V. (2006) Origins and impact of constraints in evolution of gene families. *Genome Res.* **16**, 1529-1536.
43. Ohno, S. (1970) *Evolution by gene duplication*. Springer-Verlag, Berlin, New York.
44. Piatigorsky, J. and Wistow, G. (1991) The recruitment of crystallins: new functions precede gene duplication. *Science* **24**, 1078-1079.
45. Hughes, A.L. (1994) The evolution of functionally novel proteins after gene duplication *Proc. R. Soc. Lond. B. Biol. Sci.* **256**, 119-124.
46. Collins, P.R., Stack, C.M., O'Neill, S.M., Doyle, S., Ryan, T., Brennan, G.P., Mousley, A., Stewart, M., Maule, A.G., Dalton, J.P. and Donnelly, S. (2004)

Cathepsin L1, the major protease involved in liver fluke (*Fasciola hepatica*) virulence: propetide cleavage sites and autoactivation of the zymogen secreted from gastrodermal cells. *J. Biol. Chem.* **279**, 17038-17046.

47. Coulombe, R., Grochulski, P., Sivaraman, J., Menard, R., Mort, J.S. and Cygler, M. (1996) Structure of human procathepsin L reveals the molecular basis of inhibition by the prosegment. *EMBO. J*. **15**, 5492-5503.
48. Cappetta, M., Roth, I., Diaz, A., Tort, J. and Roche, L. (2002) Role of the prosegment of *Fasciola hepatica* cathepsin L1 in folding of the catalytic domain. *Biol. Chem*. **383**, 1215-1221.
49. Adisakwattana, P., Viyanant. V., Chaicumpa, W., Vichasri-Grams, S., Hofmann, A., Korge, G., Sobhon, P. and Grams, R. (2007) Comparative molecular analysis of two asparaginyl endopeptidases and encoding genes from *Fasciola gigantica*. *Mol. Biochem. Parasitol*. 156, 102-116.
50. Vernet, T., Berti, P.J., de Montigny, C., Musil, R., Tessier, D.C., Menard, R., Magny, M.C., Storer, A.C. and Thomas, D.Y. (1995) Processing of the papain precursor. The ionization state of a conserved amino acid motif within the Pro region participates in the regulation of intramolecular processing. *J. Biol. Chem.* **270**, 10838-10846.
51. Kuk, S., Kaplan, M., Ozdarendeli, A., Tonbak, S., Felek, S. and Kalkan, A. (2005) *Fasciola hepatica* cathepsin L1 from a Turkish isolate is related to Asian isolates. *Acta. Parasitol.* **50**, 244-248.
52. Cornelissen, J.B., Gaasenbeek, C.P., Borgsteede, F.H., Holland, W.G., Harmsen, M.M. and Boersma, W.J. (2001) Early immunodiagnosis of fasciolosis in ruminants using recombinant *Fasciola hepatica* cathepsin L-like protease. *Int. J. Parasitol*. **31**, 728-737.

Figure 1: Bootstrapped (1000 trials) neighbour-joining phylogenetic tree showing the evolutionary relationship of *Fasciola hepatica* and *Fasciola gigantica* cathepsin L cDNA sequences. Numbers represent bootstrap values (given as percentages) for a particular node and values greater than 70% are shown. Branch lengths are proportional to distances. All nucleotide sequences used for tree construction encoded the cathepsin L prosegment. The tree is rooted to *Carica papaya* papain (GenBank accession number M15203). Brackets and figures represent the relative expression levels of each sub-clade (shown as a percentage of total cathepsin L levels visualised by 2-DE). The diverse origin of each sequence is indicated in the name by the internet country code of the location reported in the Genbank database.

Figure 2: (A) Typical 2-D profile of adult *Fasciola hepatica* ES proteins separated in the first dimension in the pH range 4-7 and then in the second dimension on a Criterion XT 4-12 % gradient gel (Bio-Rad). Gels were stained with Flamingo fluorescent protein stain and imaged with a PharosFX laser imaging system (Bio-Rad). Boxed regions show the positions of clade 1A, 1B, 2 and 5 *Fasciola* cathepsin Ls identified in the present study. (B) Detail from another 2-D gel of the boxed regions shown in panel A. Peptide spots selected for analysis by mass spectrometry are numbered and correspond to the identifications in Table 1.

Figure 3: ClustalW alignment of *Fasciola hepatica* and *Fasciola gigantica* cathepsin L prosegment amino acid sequences. Residues used for the alignment extend from the site of removal of the N-terminal signal peptide to the mature enzyme cleavage site. An overall consensus sequence for all *Fasciola* cathepsin Ls (FhCL consensus) is shown at the top of the alignment as well as consensus sequences for individual clades (CL1A consensus, CL1B consensus and so on). Gaps in the alignment are represented by a dash (-) and amino acids that are conserved in all sequences are indicated by a dot (.). Prosegment residues are in

capital letters whereas the first amino acids of the mature enzymes are in *lowercase letters*. The GXNXFXD motif (50) and Leu-Ser-His motif (20) required for processing of cathepsins are shown in ***bold letters***. The conserved Asn residues at the junction of the prosegment and mature domain are indicated in grey shading.

Table 1

Identification of adult *Fasciola hepatica* ES proteins by nanoLC-ESI-MS/MS.

Spot numbers refer to those shown in Figure 2B. The revised fluke cathepsin L nomenclature presented in the current study is given along with the existing GenBank name and accession number for each sequence. Relative expression levels of the cathepsin L protein spots (shown as a percentage of total cathepsin L levels in the gel) are shown. [a] Theoretical molecular mass and pI values were calculated using the primary amino acid sequence of the mature enzyme. [b] This peptide is found in FhCL1B_n11 and FhCL5 sequences. [c] This peptide is found in clade 2 and clade 5 sequences. [1] Rinaldi *et al*., (2006) Unpublished; [2,3] Castro *et al*., (2004) Unpublished; [4] (51); [5] (52); [6] (32); [7] (31).

Spot	*Fasciola* protein	Genbank name	GenBank Accession No.	Theoretical MW/pI[a]	Observed MW/pI	MOWSE score	Matched ion/peptide	Ions score	Relative Expression (%)
1	FhCL1A_pe	Fh1_6 [1]	AF490984	24.1/5.27	24.4/5.25	117	590.28 / VTGYYTVHSGSEVELK 621.24 / NSWGSYWGER	49 69	5.03
2	FhCL1A_pt1	CatIP1 [2]	AY519971	24.0/4.98	24.4/5.46	167	506.23 / ESGYVTEVK 590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	44 54 70	4.43
3	FhCL1A_pt2	CatIP2 [3]	AY519972	24.0/5.45	24.4/5.69	159	590.29 / VTGYYTVHSGSEVELK 835.04 / NLVGSEGPAAIAVDVESDFMMYR 596.25 / NSWGLSWGER	38 78 (2 oxid M) 42	10.53
4	FhCL1A_tr	CatL1Firat [4]	AY573569	23.9/5.66	23.1/5.20	161	470.22 / ESGYVTGVK 590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	49 55 57	3.17
5	FhCL1A_tr	CatL1Firat [4]	AY573569	23.9/5.66	23.1/5.42	120	470.23 / ESGYVTGVK 596.26 / NSWGLSWGER	56 64	7.48
6	FhCL1A_tr	CatL1Firat [4]	AY573569	23.9/5.66	25.0/6.03	101	590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	41 57	3.91
7	FhCL1A_tr	CatL1Firat [4]	AY573569	23.9/5.66	25.1/6.42	82	590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	36 46	0.75
10	FhCL1B_n11	CatL1 [5]	AJ279092	24.1/5.14	23.8/4.91	241	724.82 / FGLETESSYPYR[b] 511.25 / YNEQLGVAK 590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	95 68 40 37	10.02
11	FhCL1B_n11	CatL1 [5]	AJ279092	24.1/5.14	22.7/5.06	175	724.81 / FGLETESSYPYR[b] 590.28 / VTGYYTVHSGSEVELK 596.26 / NSWGLSWGER	81 38 53	5.92
12	FhCL1B_n11	CatL1 [5]	AJ279092	24.1/5.14	23.6/5.08	283	724.81 / FGLETESSYPYR[b] 511.25 / YNEQLGVAK 590.28 / VTGYYTVHSGSEVELK 596.27 / NSWGLSWGER	90 74 60 59	8.92
17	FhCL1B_n11	CatL1 [5]	AJ279092	24.1/5.14	12.1/4.92	119	724.81 / FGLETESSYPYR[b] 590.24 / VTGYYTVHSGSEVELK	85 34	7.23
20	FhCL2_chC	FhprC [6]	Z22765	24.5/4.54	23.9/4.36	197	590.26 / DYYYVTEVK 856.87 / ASASFSEQQLVDCTR 604.28 / VTGYYTVHSGDEIELK	47 116 33	12.92
21	FhCL2_chC	FhprC [6]	Z22765	24.5/4.54	24.2/4.73	335	590.26 / DYYYVTEVK 856.86 / ASASFSEQQLVDCTR 604.28 / VTGYYTVHSGDEIELK 798.39 / LTHAVLAVGYGSQDGTDYWIVK 863.85 / NSWGTWWGEDGYIR[c]	51 103 49 83 70	7.66
22	FhCL2_chC	FhprC [6]	Z22765	24.5/4.54	22.8/4.67	355	590.26 / DYYYVTEVK 856.86 / ASASFSEQQLVDCTR 604.28 / VTGYYTVHSGDEIELK 863.85 / NSWGTWWGEDGYIR[c] 882.40 / GNMCGIASLASVPMVAR[c] 890.40 / GNMCGIASLASVPMVAR[c]	40 109 43 56 100 106 (2 oxid M)	7.05
23	FhCL5_au5	CatL5 [7]	AF271385	24.4/4.75	22.8/4.73	151	724.81 / FGLETESSYPYR[b] 863.85 / NSWGTWWGEDGYIR[c]	75 77	4.98

Table 2

Residues forming the S2 active site of human and *F. hepatica* cathepsin L proteases.

Comparison of the residues from the S2 active site that contribute to differential substrate-binding in *Fasciola hepatica* cathepsin Ls (clades 1-5) and human cathepsin L. Residues were identified using primary sequence alignments and analysis of the atomic structure of *F. hepatica* cathepsin L1 (17; PDB ID: 2O6X).

		Residues					
		67	**68**	**133**	**157**	**160**	**205**
Human cathepsin L		Leu	Met	Ala	Met	Gly	Ala
Adult:	FhCL 1A	Leu	Met	Ala	Val	Ala	Leu
	FhCL 1B	Leu	Met	Ala	Leu	Ala	Leu
	FhCL 1C	Leu	Met	Ala	Val/Leu	Ala	Leu
	FhCL 2	Tyr	Met	Ala	Leu	Ala	Leu
	FhCL 5	Leu	Met	Ala	Leu	Gly	Leu
Juvenile:	FhCL 3	Trp	Met	Ala	Val	Ala	Val
	FhCL 4	Phe	Met	Ala	Leu	Ala	Phe

Table 3

Enzyme kinetics of recombinant *F. hepatica* cathepsin L proteases.

Kinetic parameters for hydrolysis of peptidyl-MCA substrates by the recombinant *Fasciola hepatica* cathepsins L1 (FhCL 1A_ie1) and L2 (FhCL 2_ie2) at pH 7.3 and pH 5.5.

	pH 7.3			**pH 5.5**		
	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($M^{-1}s^{-1}$)	k_{cat} (s^{-1})	K_M (μM)	k_{cat}/K_M ($M^{-1}s^{-1}$)
Z-FR-NMec						
cathepsin L1	2.04 ± 0.43	3.03 ± 1.08	**673,267**	24.69 ± 1.30	24.18 ± 3.92	**1,021,092**
cathepsin L2	0.40 ± 0.03	16.02 ± 2.90	**24,969**	1.70 ± 0.07	39.94 ± 5.12	**42,564**
Z-LR-NMec						
cathepsin L1	11.02 ± 0.23	3.55 ± 0.32	**3,104,225**	36.52 ± 0.63	4.35 ± 0.21	**8,395,402**
cathepsin L2	3.16 ± 0.22	4.85 ± 0.99	**651,546**	1.62 ± 0.08	1.39 ± 0.20	**1,165,467**
Z-PR-NMec						
cathepsin L1	1.74 ± 0.22	181.86 ± 13.44	**9,568**	1.03 ± 0.04	191.21 ± 16.90	**5,387**
cathepsin L2	1.90 ± 0.17	75.21 ± 12.56	**25,263**	2.64 ± 0.13	84.03 ± 11.28	**31,417**

Figure 1

Phylogenetic relationships of the *Fasciola* cathepsin L gene family

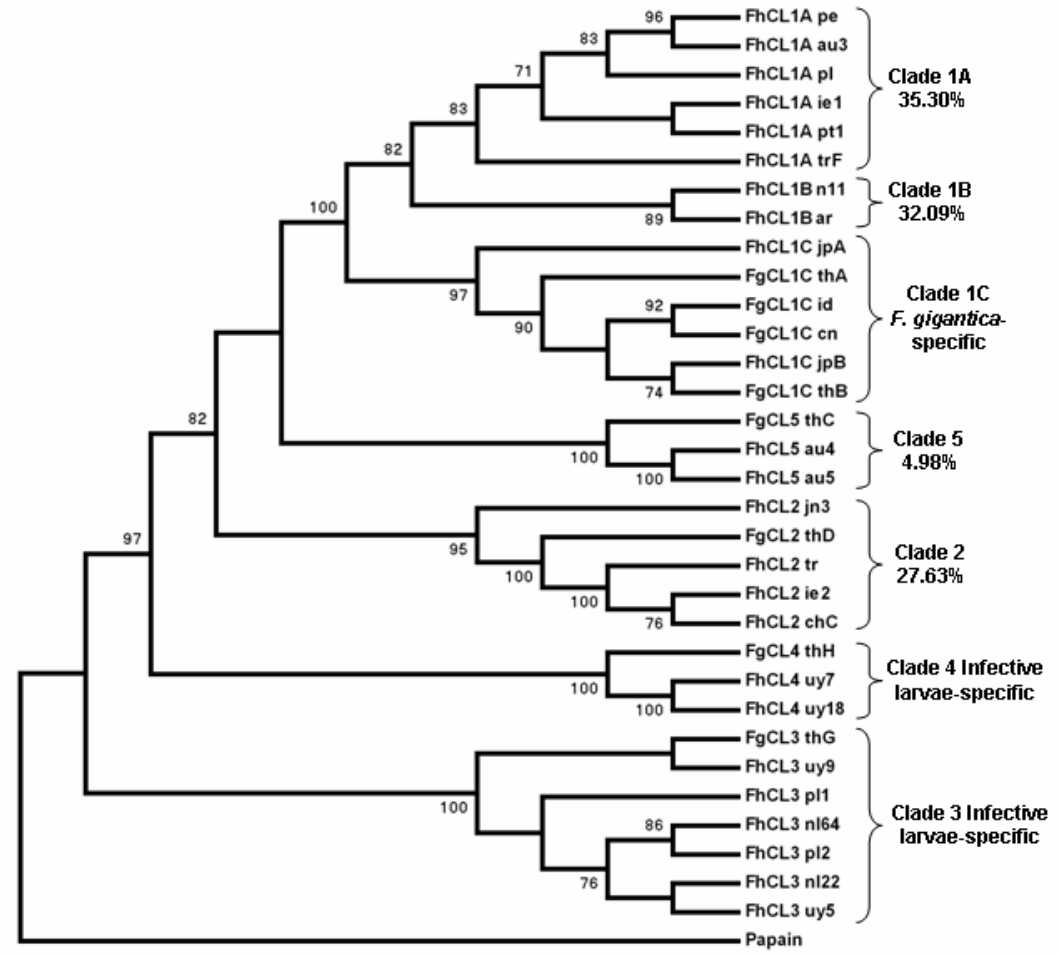

Figure 2

Representative 2-DE of adult *F. hepatica* ES proteins

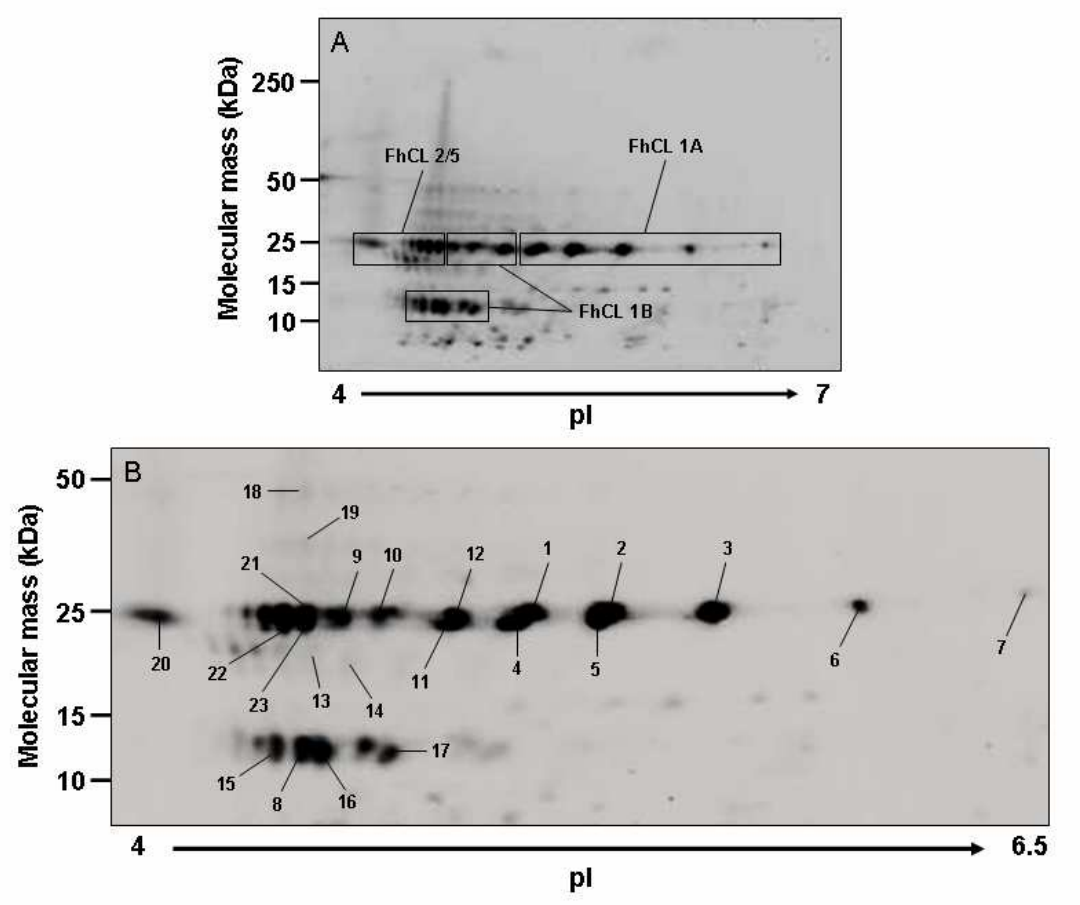

Figure 3

Primary sequence alignment of the prosegments from *Fasciola* cathepsin L proteases.