**Cloning and Characterization of a Basic Cysteine-like protease (Cathepsin L1)
expressed in the gut of larval *Diaprepes abbreviatus* L. (Coleoptera:
Curculionidae)**

Sulley Ben-Mahmoud^a

John E. Ramos^b

Robert G. Shatters Jr.^b

Pierre Rougé^c

Charles A. Powell^a

Guy Smagghe^d

Dov Borovsky^{b*}

^aIndian River Research and Education Center, University of Florida, Fort Pierce, FL

^bUSDA-ARS, Ft. Pierce, FL 34945

^cUniversité de Toulouse, UPS, Institut de Recherche pour le Développement (IRD),
UMR 152 Pharma-Dev; Université Toulouse 3, Faculté des Sciences Pharmaceutiques,
F-31062 Toulouse cedex 09, France

^dUniversiteit Gent, Ghent 9000, Belgium

*corresponding author Tel.; +1-772-462-5945; Fax +1-772-462-5986

E. mail address: dovborovsky@gmail.com

Abstract

Diaprepes abbreviatus is an important pest that causes extensive damage to citrus in the USA. Analysis of an expressed sequence tag (EST) library from the digestive tract of larvae and adult *D. abbreviatus* identified cathepsins as major putative digestive enzymes. One class, sharing amino acid sequence identity with cathepsin L's, was the most abundant in the EST dataset representing 14.4% and 3.6% of the total sequences in feeding larvae and adults, respectively. The predominant cathepsin (Da-CTSL1) among this class was further studied. Three dimensional modeling of the protein sequence showed that the mature Da-CTSL1 protein folds into an expected Cathepsin L structure producing a substrate binding pocket with appropriate positioning of conserved amino acid residues. A full-length cDNA was obtained and the proCTSL1 encoding sequence was expressed in Rosetta™ *Escherichia coli* cells engineered to express tRNAs specific for eukaryotic codon usage. The Da-CTSL1 was expressed as a fusion protein with GST and His₆ tags and purified in the presence of 1% Triton X-100 by Ni-NTA affinity and size exclusion chromatography. Recombinant mature Da-CTSL1 (23 KDa) exhibits optimal activity at pH 8, rather than at acidic pH that was shown of all previously characterized cathepsins L. Substrate specificity supports the hypothesis that Da-CTSL1 is a unique basic cathepsin L and protease inhibitor studies also suggest unique activity, unlike other characterized acidic cathepsin Ls. This paper describes for the first time a prokaryotic expression system for the production of a functional eukaryotic cathepsin L 1 from larval gut of *D. abbreviatus*.

Keywords: *Diaprepes abbreviatus*, Basic Cathepsin L1, Bacterial expression, 3D modeling, enzyme characterization

Abbreviations:

3-D, Three dimensional; AFC, 7-Amino-4-trifluoromethylcoumarin; AMC, 7-amino-4-methylcoumarin; AeaTMOF, *Aedes aegypti* Trypsin Modulating Oostatic Factor; BME, β -mercaptoethanol; CA074, L-3-*trans*-(Propylcarbamyl)oxirane-2-carbonyl)-L-isoleucyl-L-proline methyl ester; CHAPS, 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate; Da-CTSL1, *Diaprepes abbreviatus* Cathepsin L1, *E. coli*, *Escherichia coli*; E-64, L-trans-Epoxy succinyl-leucylamido (4-guanidino) butane; EST, Expressed Sequence Tag; FR, Phenylalanine and Arginine; GPR, Glycine, Proline, and Arginine; GST, Glutathione S-transferase; IPTG, Isopropyl β -D-1-thiogalactopyranoside; MBP, Maltose Binding Protein, MEROPS, Peptidase Database; MES, 2-(*N*-morpholino)ethanesulfonic acid; mRNA, Messenger RNA, Ni-NTA, Nickel-Nitrilotriacetic acid, OGNG, Octyl Glucose Neopentyl Glycol; PAGE, Polyacrylamide Gel Electrophoresis, PCR, Polymerase Chain Reaction; pNA, p-nitroanilide; PP, Propeptide; RFU, Relative Fluorescent Unit; RNA, Ribonucleic Acid; SDS, Sodium dodecyl sulfate; Suc-AAPF, N-succinyl-alanine-alanine-proline-phenylalanine-p-nitroanilide; SUMO, Small Ubiquitin-like Modifier; TPCK, Tosyl phenylalanyl chloromethyl ketone; tRNA, Transfer RNA; VVR, Valine, Valine, and Arginine; Z, Benzyloxycarbonyl; Z-FY(tBu)DMK, (*N*-benzyloxycarbonyl-phenylalanyl-*t*-butyl-tryrosyl diazomethylketone

1. Introduction

The ability to break down proteins and complex carbohydrates, by digestive carbohydrases and proteases in the gut, into free amino acids and simple sugars, respectively is essential for growth and development of insects (Bown et al., 2004). In many coleopteran species such as the cowpea bruchid (*Callosobruchus maculatus* Fabricius) and the western corn rootworm (*Diabrotica virgifera virgifera* LeConte), cysteine proteases are the major digestive enzymes (Koiwa et al., 2000; Zhu-Salzman et al., 2003). In *D. abbreviatus*, the presence of both cathepsins (cysteine proteases) and serine proteases have been identified in digestive tract extracts, but only the trypsin-like enzyme was extensively characterized and was shown to be controlled by AeaTMOF (Yan et al., 1999).

Cathepsins belong to the papain family of proteins that include lysosomal proteins (Cristofolletti et al., 2005) that are involved in cellular protein catabolism (Kamboj et al., 1993; Stachowiak et al., 2004). Coleopterans and hemipterans utilize cathepsins to digest their food in the digestive tract lumen (Terra, 1990). Cathepsins are synthesized as pre-pro-enzymes and are secreted into the lumen by the gut epithelial cells. The pre-region serves as a signal peptide allowing the enzyme to enter the cellular secretory pathway targeting it for secretion. The pro-region is essential for the proper folding and for inhibiting proteolytic activity until it is cleaved in the gut's lumen by autocatalysis generating a mature active protein. *D. abbreviatus* midgut is partitioned into anterior and posterior regions, similar to the aphid, *Aphis gossypii* Glover (Deraison et al., 2004), and the weevil *Otiorhynchus sulcatus* Fabricius (Edwards et al., 2010). Although no direct pH measurements have been done to show

the pH of these regions, in the water weevil *Lissorhoptrus brevivirostris*, chymotrypsin like activity in the gut was the highest in the middle gut and posterior gut sections strongly indicating that in these sections the pH is basic, a requirement for chymotrypsin activity (Henandez et al. 2003). These observations suggest that acidic cathepsins would not be able to function in the posterior region of the gut.

All cathepsins are characterized by a cysteine residue at the active site (Turk et al., 2000) with a broad specificity for cleaving peptide bonds. Additionally, cathepsins are classified based on characteristics such as their activities with unique substrates and specific inhibitors, and pH optimum for activity (Kirschke, et al., 1977; Turk et al., 2000; Rawlings et al., 2012). *In vitro* enzymatic assays used to characterize cathepsins feature small peptide substrates consisting of a sequence of amino acid residues linked to a measurable chromogenic or fluorogenic leaving group (Tchoupé et al., 1991). The preferred amino acid sequences of substrates for various proteases including cathepsin L are determined via Positional Scanning Synthetic Combinatorial Libraries (Cotrin et al., 2004; Choe et al., 2006). Cathepsin L has been shown to have substrate tolerance for hydrophobic residues in the P2 and P3 positions (Otto and Schirmeister, 1997) with a preference for leucine or arginine at the P2 position (Clara et al., 2011). A protease database (MEROPS) groups various proteases according to their chemical properties including the sequence of amino acids in suitably cleaved substrates (Rawlings et al., 2012). Using information garnered from the MEROPS database, and from the characterization studies of the recombinant cathepsin L1 (Da-CTSL1) we show for the first time that Da CTSL1 is a unique basic enzyme that could function in the presumed basic posterior region of *D. abbreviatus* gut.

2. Materials and Methods

2.1. Bacterial Strains, Plasmids, and Primers

Escherichia coli strains OneShot TOP 10F' and DH5- α (Invitrogen, Grand Island, NY) were used for sub-cloning, purification, and sequencing of plasmid DNA. The plasmid used for checking DNA sequences was pCRTM2.1 (Invitrogen). Expression plasmids used was pET-41a (+) (EMD Millipore) which permits the expression of proteins as fusion products, with the incorporation of a polyhistidine tag or a glutathione-S transferase tag (pET-41a(+)). A plasmid purification kit (Qiagen, Valencia, CA) was used for all purifications according to the manufacturer's protocol. An IPTG (Isopropyl β -D-1-thiogalactopyranoside)-inducible derivative of *E. coli* strain BL21 (DE3), RosettaTM (DE3) pLysS Competent Cells, purchased from EMD Millipore, was used to express recombinant proteins. Bacterial strains were grown in Luria broth (LB) media at 37 °C except after inducing with IPTG for protein synthesis and expression at which time cells were grown at 25 °C for 4 h to 24 h. Primers GstproCatL1-f (forward) (5'-CTCAATCACTAGTACCGTCTTGGAGGAGACAGGTGTCAAAT-3'), and GstproCL1-r (reverse) (5'-GCAGTCACTCGAGTATAATTGGATATACGGTATCTTCAGCAACTCC-3') that were synthesized by IDT[®] (San Diego, CA) and were used for cloning Da-CTSL1.

2.2 Preparation of EST libraries and Digestome

Total RNA was isolated from excised *D. abbreviatus* digestive tracts of: a. five weeks old larvae raised on artificial diet (Beavers, 1982), b. newly emerged teneral adults, and, c. adults feeding on citrus flush for 3 weeks after emergence. Digestive tracts from 10 psyllids were pooled and RNA extracted using RNeasy Midi kit (Qiagen.,

CA, USA) following the manufacturer recommended protocol for animal tissue extractions. Excised digestive tracts were stored frozen at -80°C until used, and homogenized in liquid nitrogen prior to the RNA purification. Poly A+ RNA was isolated from the total RNA using Oligotex Direct mRNA Mini kit following manufacturer recommended protocol (Qiagen, CA, USA). Poly A + enriched extracts were used for PCR-based amplified cDNA synthesis using the Super SMART™ PCR cDNA Synthesis Kit following manufacturer protocol (Clontech, CA, USA). The optimized cDNA products obtained after using amplification parameters and 19 amplifications cycles recommended by the manufacturer were separated by agarose gel electrophoresis (1% gel). After electrophoresis, the gel segment corresponding to cDNA of 400 bp to 10 Kb was excised and the cDNAs purified using a QIAquick gel extraction kit (Qiagen, CA, USA) and cloned into pCR®4-TOPO using Topo TA cloning kit (Life Technologies, NY, USA) and transformed into *E. coli* One Shot® Mach1™-T1^R chemically competent cells (Life Technologies, NY, USA). The EST libraries are available at the USDA Ft. Pierce FL USA .

2.3 cDNA sequence analysis.

Sequence analysis was done using Sequencher DNA analysis software (Gene Codes, Ann Arbor, MI). Sequencing results were trimmed to remove vector and linker sequences and to remove low quality bases. Resulting sequences were contiged using the Assemble Automatically function of sequencher 4.0 with the following settings: Assembly Algorithm: Dirty Data; Optimize Gap placement with Use ReAligner and Prefer 3' Gap Placement and with Minimu Match Percentage set at 85% and Minimum

Overlap set at 50. Sequences were annotated using results from the NCBI batch BLAST service and by using BLASTN and BLASTX.

2.4 Analysis of *Da-CTS L1* transcript in the gut

The level of *Da CTS L1* transcript in different segments of the larval gut was followed by RT-qPCR. Total RNA was extracted from 10 gut segments of foreguts, anterior midguts, posterior midguts and hindguts (as above), . Each RT-qPCR analysis was done on extracted RNA (100 ng) using the QuantiTect SYBR Green RT-PCR kit (Qiagen). Each RT-qPCR reaction was run in duplicates including a control reaction without reverse transcriptase to check for DNA contamination. Primers: DCL-1forward 5'-AGTTGAAACACGGCAAGACC-3' and DCL-1 reverse 5'-GTGCTCGGTGGTGTGAAAT-3' were used to amplify a fragment (227 bp) Gene expression was normalized with the expression of a 18S rRNA of *D.abbreviatus* using primers pair: DRW-18S forward- 5'-GCTCATTAATCAGTTATGG-3' and DRW-18S reverse- 5'-ATCAGCGTAAAGTTATTCAG-3' amplifying a similar size fragment (196 bp). After RT-qPCR, a single melt peak was observed and the product was verified by sequencing. The gene expression of the the target *Da CTSL1* transcript was used to calculate fold differences from the *D. abbreviatus* 18S rRNA in different segments of the larval gut using the $\Delta\Delta C_T$ method (Livak and Schmittgen, 2001).

2.5 Cathepsin L 1 Constructs

The DNA sequences of the pro- and active regions of *DaCatL1* were amplified from a cDNA library and cloned into the expression vector pET-41a (+) DNA. The

primers GstproCL1-f (incorporating the cut site for *SpeI*) and GstproCL1-r (incorporating the cut site for *XhoI*) were used to amplify DaCatL1 using standard polymerase chain reaction (PCR) protocols. The following thermal cycling conditions were used: 95 °C for 10 min, followed by 30 cycles of 95 °C for 15 sec, 57 °C for 30 sec and 72 °C for 30 sec, and a final extension at 72 °C for 10 min. The size of the amplified DNA was verified by agarose gel electrophoresis, the band cut from the agarose gel and the Da-CTSL1 DNA fragment recovered from the agarose gel using Wizard[®] SV Gel and PCR Clean-Up System (Promega, Madison, WI) following manufacturer's instructions. The Da-CTSL1 fragment was then cut with *SpeI* and *XhoI* (New England Biolabs[®], Ipswich, MA), and cloned into the expression vector pET-41a. One Shot Top 10 (Invitrogen) competent *E. coli* cells were transformed with pET-41a (100 ng) carrying Da-CSTL1 and grown on LB media containing kanamycin (50 µg/ml). Single colonies were selected and grown overnight at 37 °C in LB media containing antibiotic in a shaking incubator at 225 rpm, plasmid (pET41-a) was then extracted using a plasmid extraction kit (Qiagen) and sequenced by the genomics core facility at the USDA-USHRL, and University of Florida-Interdisciplinary Center for Biotechnology Research (UF-ICBR), Gainesville. Several colonies contained the Da-CTSL1 insert were identified and stored in glycerol at -80 °C.

2.6 Expression of Recombinant Da-CTSL1

Rosetta[™] (DE3) pLysS Competent Cells (EMD Millipore) cells were transformed with pET-41a carrying Da-CTSL1 sequence. Colonies with the engineered vector were initially grown in 5 mL at 37 °C, 225 rpm in LB medium in the presence of 30 µg/ml chloramphenicol and 50 µg/ml kanamycin, to OD₆₀₀ of 0.6. Isopropyl β-D-1-thiogalactopyranoside (IPTG) (1 mM) was then added to induce synthesis of the

recombinant protein. Controls were similarly grown, except the IPTG induction step was omitted. The temperature during the induction stage was maintained at 25 °C for 4 to 6 h. Cells were harvested by centrifugation at 16,000 rpm for 5 min at 4 °C in an Eppendorf's refrigerated bench-top 5702 R centrifuge, re-suspended in 1 ml phosphate buffer saline (50 mM NaH₂PO₄, 300 mM NaCl, pH 7.0), broken in 2-ml tubes containing acid-washed glass beads (180 µm) in a FastPrep Instrument for 1 min and centrifuged at 14,000 rpm for 10 min at 4 °C. The supernatant was collected and aliquots (20 µL each) run on sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) using 4-12% NuPAGE® Novex® Bis-Tris Gel gradient (1.0 mm thick) (Life Technologies, Carlsbad, CA). Da-CTSL1 fused to GST ran as a protein band at 60 KDa (Fig. 6, A). To synthesize large amounts of recombinant proteins, transformed *E. coli* (4-6 L) were grown and harvested as before. Harvested *E. coli* cells from 4 liter culture were resuspended in 50 mL lysis buffer (50 mM Tris-HCl buffer pH 7.0 containing 300 mM NaCl and 1 mM βMercatoethanol) and lysed using a cell disruptor (Branson 450 Sonifier - SmithKline) using multiple cycles of 30 sec sonication and 90 sec rest for 45 min with a constant cycle output of 2 at 4⁰C. To improve solubility 1% Triton X-100 was added to all buffers after lysing the bacterial cells. SDS-PAGE was used to identify the recombinant protein followed by western blot analysis (Figs. 6, A,B).

Protein quantification was done using the Quick Start™ Bradford method (Bio-Rad, Hercules, CA), except when proteins were solubilized in 1% Triton-X100 in which the RC DC™ Protein Assay (Bio-Rad) was used. Synergy™ HT Multi-Mode Microplate Reader (BioTek, Winooski, VT) was used to read absorbance.

Western blot analysis of recombinant proteins was done using the Xcell II blot module from Invitrogen and proteins were transferred after SDS PAGE onto a nitrocellulose membrane. Monoclonal antibodies raised against His₆- tagged Epitope (Thermo Scientific, Waltham, MA), were used to identify Da-CTSL1 and Western Lightning *Plus* ECL kit (PerkinElmer, Waltham, MA) was used to visualize proteins on the nitrocellulose membrane with the aid of a KODAK imager (Image Station 440CF).

2.7 Purification of Recombinant Proteins

Purification of histidine-tagged recombinant proteins from the crude extract was done using 3 mL of Ni-NTA Superflow resin (Qiagen) equilibrated with 5 mL of 50 mM Tris buffer, pH 7, 300 mM NaCl, 1 mM BME, 1% Triton, 10% Glycerol. To the affinity resin (3 mL) 50 ml of crude extract was added and gently mixed in the cold for an hour. Unbound proteins were removed by centrifugation at low speed (2,000 rpm, 5 min, 4 °C), and bound proteins were washed with buffer (5 times) by mixing in the cold as before with 5 volumes of buffer. Recombinant proteins were eluted from the Ni-NTA resin with 250 mM imidazole in the wash buffer, the Ni-NTA resin was centrifuged at low speed and supernatant removed and aliquots (20 µL; 144 µg protein) analyzed by SDS-PAGE and western blotting for Da CTSL1. To remove the pro-region of the Da-CTSL1 the recombinant protein was concentrated to 10 mL by centrifugation using centrifugal filters of M_r 10,000 cut off (Millipore, MA, USA) and the pH 7.0 buffer was exchanged by repeated increasing and decreasing the volume 4 to 5 fold with 50 mM Tris-HCl buffer, pH 8.0 containing 300 mM NaCl, 1mM βmercaptoethanol and 1% Triton-X100). After the final exchange cycle the activated recombinant protein (10 mL, 0.4 mg protein) was purified by size-exclusion chromatography (Hiprep™ 16/60 Sephacryl™

S-100 HR) with an ÄKTA™FPLC™ system (GE HealthCare Life Sciences, Piscataway Township, NJ). The column was maintained at constant pressure of 21 psi and elution speed of 0.8/min. Fractions with Da-proCTSL1 and Da-CTSL1 were collected and concentrated under N₂ pressure using a filtration cell (Amicon® Ultra-15 ml 10, 000 MW cut-off, EMD Millipore), and aliquots (200 µL) stored in the presence of 20% glycerol at - 80 °C.

2.8 Enzymatic Assays

To find out the optimal substrate for recombinant DA CTSL1 several small fluorogenic and chromogenic peptide substrates, Z-GPR-AMC (Z= benzyloxycarbonyl, GPR = glycine, proline, arginine, AMC = 7-amino-4-methylcoumarin), Z-FR-AFC (FR= phenylalanine, arginine, AFC= 7-Amino-4-trifluoromethylcoumarin), Z-VVR-AMC (VVR= valine, valine, arginine), Suc-AAPF-pNA (N-succinyl-alanine-alanine-proline-phenylalanine-p-nitroanilide), H-Arg-pNA.2HCl, and Z-FR-pNA.HCl (all obtained from Enzo® Life Sciences, Farmingdale, NY) were tested for Da-CTSL1 biological activity by using fluorometry (Wallac 1420, Victor3™- PerkinElmer). Z-FR-AFC was found to be the best substrate and was used in all subsequent assays using 96-well plates (Corning® - Corning, NY) containing a final assay volume of 150 µl per well. The procathepsin L1 was activated at pH 8.0 by buffer exchange and purified by size exclusion chromatography and 3 µl of activated enzyme (7 µg) were used per assay; no activation was observed at acidic pH. The fluorescent assays were conducted at excitation/emission wavelengths of 355 nm/535 nm, 25 °C, with a final substrate concentration of 12.5 µM, and a molar ratio of substrate to Da-CTSL1 of 4:1. The chromogenic assays were conducted at an absorbance of 405 nm with final substrate

concentration of 80 μ M and a molar ratio of substrate to Da-CTSL1 of 26 :1.

Fluorescence was monitored at 5 min intervals for 120 min and the reaction rate at molar ration of substrate to enzyme of 4:1 is linear up to 120 min (Fig. 8A). To find the optimum pH, the following buffers were used: 50mM sodium acetate buffer (pH 3-4), 50 mM MES buffer (pH 5-6), and 50 mM Tris-HCl (pH 7-9). For proteinase inhibitor assays, the enzyme-inhibitor mixtures (without substrate) were pre-incubated at room temperature for 10 min prior to the assay. All assays were repeated at least 3- times unless stated otherwise. The inhibition of Da-CTSL1 was assessed using the following protease inhibitors: E-64 (L-trans-Epoxy succinyl-leucylamido (4-guanidino) butane), Z-FY (tBu)DMK (*N*-benzyloxycarbonyl-phenylalanyl-*t*-butyl-tyrosyl diazomethylketone), CA074, cystatin and chymostatin at 50 μ M each and aprotinin at 50 nM or 50 μ M. The molar ratio of the inhibitors to Da CTS L1 used in the enzymatic reaction is 17:1 when 50 μ M of inhibitors were used and 1:17 when 50 nM was used for aprotinin.

2.9 pH determinations in the gut

Full length of 10 intact digestive tracts were dissected out from *D. abbreviatus* larvae (5-10 weeks old) , rinsed 3 times in deionized water and the foregut, anterior midgut, posterior midgut and hindgut regions were sectioned. . Identical sections were combined and placed in Eppendorf tubes containing 1 mL of deionized water. The tubes briefly vortexed, centrifuged and the pH of the supernatants measured using a pH meter with a microprobe. The experiment was repeated 3 times and the results are expressed as means of 3 determination with SEM of 5-10%.

2.10 Molecular modeling of Ca CTSL1

Homology modeling of the procathepsin L1 proteinase from the gut of *Diaprepes abbreviatus* (Da-proCTSL1), was performed with the YASARA Structure program (Krieger et al., 2002) running on a 2.53 GHz Intel core duo Macintosh computer. Different models of Da-proCTSL1 were built from the X-ray coordinates of the procathepsin L1 from *Fasciola hepatica* (PDB accession number 2O6X) (Stack et al., 2008) the larval midgut procathepsin L3 of *Tenebrio molitor* (PDB accession number 3QT4) (Beton et al., 2012), the Cys25Ala mutant of human procathepsin S (PDB accession number 2C0Y) (Kaulmann et al., 2006), the human procathepsin L (PDB accession number 1CS8) (Coulombe et al., 1996), and the activated cathepsin L of *Toxoplasma gondii* in complex with the propeptide (PDB accession number 3F75) (Larson et al., 2009), used as templates, respectively. Finally, a hybrid model of Da-proCathL1 was built up from the five previous models. PROCHECK (Laskowski, 1993) was used to assess the geometric quality of the three-dimensional models. In this respect, all of the residues of Da-proCTSL1 were correctly assigned in the allowed regions of the Ramachandran plot except for 9 Gly residues which occur in the non-allowed region of the plot. Using ANOLEA (Melo and Feytmans, 1998) to evaluate the model, only 12 residues of Da-proCathL1 over 306 exhibited energy over the threshold value. Both residues are located in the loop region connecting α helices and β sheets. Molecular cartoons were drawn using the UCSF Chimera package (Pettersen et al., 2004).

3. Results

3.1. Digestive Tract Proteases

cDNA clones were constructed from EST libraries of the digestive tracks of *D. abbreviatus* 5 week old larvae (1504 clones), teneral adults (1545 clones), and fed adults (1538 clones) (Table 1) to observe changes in gene expression by following mRNA abundance in digestive tracts isolated from different life stages. Based on abundance of specific sequences in the cDNA libraries, different classes of proteases are preferentially expressed at different life stages (Table 2). Unfed teneral adults synthesize negligible amount of proteases in their gut, whereas different classes of proteases are synthesized by the guts of 5 weeks old larvae and feeding adults (Table 2). The majority of the sequenced protease fragments from 5 week old larval and feeding adult guts (80% and 84%, respectively) represent cathepsin L homologues (1-24; Table 2). Using 5 week old larval gut we determined the amount of all cathepsin L homologues in of the cDNA clones showing that a single cathepsin L homolog (Da-CTSL1) represents 64% of all the expressed Da-CTSL homologues from the cDNA clones obtained from the gut (Fig. 1A) and, therefore, was selected for cloning, and characterization. Transcript abundance ratios in relationship to 18S rRNA using RT-qPCR and $\Delta\Delta C_T$ show that 1000 fold more DA-CTSL1 transcripts are found in the foregut and anterior midgut than in the posterior and hindgut segments (Fig. 1B). The pH in the foregut and anterior midgut is acidic, whereas the pH in the posterior midgut and the hindgut is 7.5 ± 0.7 and 6.8, respectively indicating the only region in the gut that is basic is the posterior midgut which is longer than the other segments (Fig. 1 B and C).

The amino acid sequence of Da-CTSL1 shows the presence of both a putative pre and pro regions (amino acids 1-15 and 15-109, respectively, Fig. 2A) and conserved active site residues, (Q, C, H and N, Fig. 2) as was shown in other cathepsins (Bown et al., 2004, Fonseca et al., 2012). ClustalW analysis and a phylogenetic tree show that Da-CTSL1 has closer sequence similarity to *Tenebrio molitor*, another coleopteran, than to blood sucking insects like mosquitoes, body louse or lepidopterans (*Spodoptera frugiperda* and *Helicoverpa armigera*) (Fig. 2C). A BLASTP search in the NCBI database showed that Da-CTSL1 has high sequence similarities with several known Coleopteran and human cathepsins (Table 3).

3.2. 3D Molecular Modeling of Da-CTSL1

The procathepsin L1 proteinase from *Diaprepes abbreviatus* Da-CTSL1 consists of a short N-terminal propeptide built up of three α helical, linked to a longer C-terminal catalytic domain which possesses the two catalytically active Cys120 and His255 amino acid residues located in a pocket at the center of the predicted mature protein catalytic groove (Fig. 3A). The 3D modeling supports the concept that both domains are organized in such a way that the propeptide (PP) becomes inserted into the groove and completely masks the catalytic pocket of the active domain, inhibiting the proteinase activity of Da-CTSL1 (Fig. 3B,C). The enzymatic inhibition results from steric hindrance at the catalytic groove where the propeptide region prevents the substrate from accessing the active residues C120 and H255. The model also predicts that the N-terminal propeptide (PP) does not have direct interaction with the active amino acid residues C120 and H255 at the catalytic pocket. Instead, the propeptide becomes H-bonded to a few amino acid residues located along the groove. The 3D model also

predicts that a network of fourteen hydrogen bonds involving D231, T233, L253, N254, E255 and E300 residues, allows the PP to anchor at the Da-CTSL1 catalytic groove . Residues R71, L76, S77, S78, S79 and K80 of PP participate in this H-bonding network. The proteolytic cleavage at pH 8.0 of the propeptide frees the catalytic groove and activates Da- CTSL1.

3.3. Synthesis and Purification of Da-CTSL1

Several attempts to express *D. abbreviatus* Da-CTSL1 and Da-CTSL2 in *Pichia pastoris* (pPICZ- α , Invitrogen), a Human *In Vitro* Translation Kit (Thermo Scientific), and in *E. coli* (pET200 Directional TOPO[®] Expression Kit, Life Technologies) failed. Similarly, Smith and Gottesman (1989) reported that expression of recombinant human Cathepsin L in *E. coli* yielded only 0.521 mg of cathepsin L from 6.73 gr of bacterial lysate proteins. The enzyme had to be recovered from the insoluble proteins of the bacterial lysate, solubilized and renatured. To overcome these difficulties and increase enzyme solubility, Da- proCTSL1 was cloned into pET-41a and expressed in *E. coli* as a fusion protein with a GST at the N-terminus and His₆ tags at the C-terminus, respectively (Fig. 4) after induction with IPTG. This strategy allowed us to remove the GST moiety during the activation at pH 8.0. The D-GST- proCTSL1 was purified by Ni-NTA chromatography, activated by buffer exchange at pH 8 and purified on a size exclusion chromatography column (Fig. 5). Aliquots from each purification step were collected and run on SDS-PAGE (Fig. 6A) showing the presence of the inactive Da-proCTSL1 with GST and His₆ tags (M_r 60KDa, Fig. 6A lanes b,c) and the activated Da-CTSL1 with the His₆ tag (M_r 23 KDa) after column exchange

1
2
3
4 393 chromatography at pH 8.0 and size exclusion chromatography (Fig. 6A lane d). After
5
6
7 394 activation, the activated Da-CTSL1 starts to digest the Da-GST-proCTSL1 that was not
8
9 395 activated yet, resulting in protein bands that are not found when the purified Da-GST-
10
11
12 396 proCTSL1 was run on SDS PAGE (Fig 6A, lanes c, d). Monoclonal antibodies against
13
14 397 the His₆-tag identified Da-GST-proCTSL1 as a major band and a minor band of
15
16 398 activated Da-CTSL1 after the Ni affinity chromatography (Fig.6B, left lane). These
17
18
19 399 bands are missing in Ni-NTA purified extract of *E. coli* cells that were transformed with
20
21 400 an empty plasmid (Fig. 6B, right lane). Mass spectrum analysis identified 34 unique
22
23 401 peptides (Fig. 6C, yellow highlight) indicating 100% probability that the expressed
24
25
26 402 purified protein is Da-CTSL1 fused to GST and His₆ (Fig. 6C). Mass spectra analyses
27
28
29 403 of the activated and the proenzyme showed that the proenzyme is cleaved between
30
31 404 L109 and A110 (Fig. 6C). The determination of the pre-pro region of the enzyme was
32
33 405 done by homology with other cathepsins (Fonseca et al., 2012).

34
35
36 406 Even though the GST tag kept Da-proCTSL1 soluble in the bacterial cell
37
38 407 preventing it from precipitating in the bacterial inclusion bodies, the recombinant
39
40
41 408 enzyme rapidly fell out of solution in low salt LB cultures, during induction and
42
43 409 expression at various temperatures (25 – 37 °C), and in the presence of different
44
45 410 concentrations of several surfactants (amphipol, CHAPS and OGNG) during bacterial
46
47
48 411 cell lysis. Da-GST-proCTSL1 stayed in solution only in the presence of Triton X-100
49
50 412 (1%). Removing Triton X-100 with HiPPR[®] Detergent Removal Spin Columns (Thermo
51
52
53 413 Scientific, MA, USA) caused a rapid loss of Da-GST-proCTSL1 (data not shown).

54
55 414 The Ni affinity chromatograph purified Da-GST-proCTDL1 was activated at pH
56
57
58 415 8.0 and purified by size exclusion chromatography. The non-activated Da-GST-

proCTSL1-His₆ was eluted between 40-48 mL (Fig.7 peak A) and exhibited M_r 60 kDa on SDS-PAGE (results not shown). The activated Da-CTSL1 was eluted between 52 to 64 mL (Fig. 7 peak B) with M_r 23 kDa on SDS-PAGE (Fig. 6A, lane d) and is biologically active at pH 8 when incubated with Z-FR-AFC (Fig. 8).

3.4 Cathepsin L1 Substrate Specificity and pH optimum

Da-CTSL1 hydrolyzes the general cathepsin L, B and H substrate, Z-FR-AFC at pH 8.0 9-folds better than at pH 5.5 and the enzymatic activity at pH 8.0 is linear for 120 min using a molar ratio of substrate to enzyme of 4:1 (Fig 8A and B) . The substrates Z-GPR-AMC and Z-VVR-AMC were not hydrolyzed at pH 5.5 or 8.0 by Da-CTSL1 (Fig. 10). Because Cathepsin B readily hydrolyzes Z-FR-AFC all incubation mixtures contained 50 µM cathepsin B inhibitor CA074. Because activation and higher activity of Da-CTSL1 was observed at pH 8.0 and not at the expected pH 5.5 (Fig. 6A) Da-CTSL1 was tested at different pHs. Incubation of Da-CTSL1 with Z-FR-AFC at a range of different pHs shows that the optimal activity of Da-CTSL1 is at pH 8 (Fig. 9). These results indicate that Da-CTSL1 is an alkaline cathepsin in contrast to the majority of cathepsins that show optimal activity at acidic pH range. Several other substrates that are cleaved by chymotrypsin and aminopeptidases that are active in the alkaline range like Cathepsin G substrate SucAAPF-pNA, cleaved by chymotrypsin-like proteases, and HR-pNA that is cleaved by Cathepsin H and other aminopeptidases are not cleaved by Da-CTSL1 (data not shown) indicating the Da-CTSL1 is not a general alkaline protease but a unique cathepsin L that is mainly active in the alkaline portion of the larval posterior midgut (Fig. 1B and C).

3.5 Effect of inhibitors on Da-CTSL1 activity

The activity of Da-CTSL1 was assayed at pH 8 in the absence or presence of known protease inhibitors by pre-incubating the enzyme and the inhibitors together for 10 min before adding the substrate and then following Da-CTSL1 activity in the presence of Z-FR-AFC for 120 min (Fig. 10). Aprotinin (serine protease inhibitor) at 50 nM weakly inhibited Da-CTSL1 at a molar ratio of enzyme to inhibitor of 17:1, however it inhibited Da-CTSL1 activity almost completely at a molar ratio of enzyme to inhibitor of 17:1 (data not shown). In the presence of the general cathepsin inhibitor (excluding cathepsin C and H) E-64 (50 μ M), the activity was inhibited by 26.6%, whereas a specific cathepsin L inhibitor Z-FY(tBu)DMK (50 μ M) caused higher inhibition (40.6%). The specific cathepsin B inhibitor (CA074) at 50 μ M inhibited the enzyme only by 25.3%. The general protease inhibitors cystatin and chymostatin (50 μ M each), which also inhibit cathepsin L, were more effective, inhibiting Da-CTSL1 activity, by 47.5% and 85.6%, respectively (Fig. 10). Since all the inhibitory reactions were linear throughout the incubation period of 120 min (results not shown), the poor inhibition by many of these inhibitors may indicate that at pH 8 these substrate may not be able to bind to the active site of Da-CTSL1. Similarly, plant cysteine protease inhibitors may not bind at a basic pH and this may help this polyphagous insect that feeds on many plants. However, this has not been shown yet.

Discussion

Diaprepes abbreviatus larvae are voracious feeders, and the nutrients consumed as part of their diet are broken by gut proteases and utilized for growth and development. Since proteases play an important role in larval and adult food digestion, it is important to characterize these proteases in order to be able to effectively control

1
2
3
4 462 them in future crop protection strategies. Numerous proteases are transcribed during
5
6 463 larval feeding stages and the majority of these proteases belong to the serine
7
8
9 464 (chymotrypsins, trypsins and serine carboxypeptidases) and cysteine (Cathepsins B
10
11 465 and L) protease families (Table 2) that are common in Coleoptera (Thie and
12
13
14 466 Houseman, 1990; Hernandez et al., 2003; Edwards et al., 2010). Of the cysteine
15
16 467 proteases, cathepsin L1 (Da-CTSL1) is encoded by the most abundant transcript that
17
18
19 468 comprises 64% of the total cathepsin transcripts considering the fact that 24 different
20
21 469 members of this gene superfamily were identified, and, therefore, Da-CTSL1 is
22
23
24 470 probably important in food digestion in the gut of 5 week old larvae (Fig. 1A). Assaying
25
26 471 segments of the larval gut by RT-qPCR show that Da-CTSL1 transcripts are 1000 fold
27
28
29 472 more abundant in the foregut and the anterior midgut than in the posterior mid gut
30
31 473 where the pH is basic (Fig. 1B). These findings strongly indicate that even though the
32
33 474 pro-enzyme is synthesized by the foregut and anterior midgut (Fig. 1C) it is not
34
35
36 475 activated until it reaches the posterior midgut which is basic (Fig. 1B). To better
37
38 476 understand its significance, we cloned, expressed and characterized Da-CTSL1.

39
40
41 477 Several expression systems were used to express cathepsin L *in vitro*. Bown et
42
43 478 al. (2004) and Fonseca et al. (2012) have expressed Cathepsin L from the gut of the
44
45 479 western corn rootworm (*Diabrotica virgifera*) and the sugarcane weevil, *Sphenophorus*
46
47
48 480 *levis* using the methanotropic yeast *Pichia pastoris*. The procathepsin L was secreted
49
50
51 481 by the cells into the medium during the fermentation of the cells. Bown et al. (2004)
52
53 482 reported that the secreted proenzyme was cleaved by the yeast proteases and was
54
55 483 found as the active cathepsin L, whereas Fonseca et al. (2012) reported that the
56
57
58 484 secreted *S. levis* proenzyme was not activated during the secretion into the medium.
59
60
61
62
63
64
65

Ishidoh and Kominami (1994) expressed procathepsin L in v-*Ha*-ras transformed NIH3T3 cells and Yamamoto et al. (1999) expressed *Bombyx mori* preproCathepsin L in *E. coli*. To study the activation process at the molecular level and to avoid precipitation of the recombinant protein in the bacterial inclusion bodies Yamamoto et al. (1999) used lengthy renaturation steps recommended by Smith and Gottesman (1989). We expressed the recombinant Da-proCTSL1 in Rosetta competent *E. coli* cells as a fusion product with a GST tag (Fig. 4). Even though the recombinant Da-GST-proCTSL1 did not precipitate in the bacterial inclusion bodies, the recombinant Da-CTSL1 was unstable in solution and tended to precipitate upon storage at 4 °C. In the future other polypeptide moieties could perhaps be fused to the enzyme to keep it soluble and retain its activity, such as the maltose binding protein (MBP) and the small ubiquitin-like modifier (SUMO) (Kapust and Waugh, 1999; Butt et al., 2005). It is also possible that because Da-CTSL1 is active at a basic pH unlike all other cathepsins that are active at acidic pH and lose activity at basic pH, Da-CTSL1 may not be stable and even precipitate at low or neutral pHs. To overcome the insolubility of Da-CTSL1, Triton X-100 (1%) was added to the enzyme keeping Da-CTSL1 in solution without interfering with its protease activity (Figs. 8-10). Although numerous alternative recombinant protein synthesis technologies are available (Hunt, 2005), expressing heterologous proteins in *E. coli* is simple and straight-forward. Bacteria provide the cheapest and fastest recombinant protein synthesis technology that can be easily scaled up.

The characterization of Da-CTSL1 shows that it is preferentially cleaves Z-FR-AFC (Fig. 8), which has a hydrophobic phenylalanine at P2 and Da-proCTSL1 is autocatalyzed at pH 8.0 releasing an activated Da-CTSL1 that exhibits optimal

activity at pH 8 indicating that the enzyme when secreted by the foregut and anterior midgut is only activated when it reaches the posterior midgut which has a basic pH (Figs. 9 and 1B, respectively), unlike other cathepsin Ls that have optimal activity in the acidic pH range (pH 5-6.5). Reports of basic cysteine proteases are rare (Otto and Schirmeister, 1997) and unusual for the cathepsins listed in the MEROPS database (Rawlings et al., 2012). Furthermore, no insect basic cathepsin L have been previously identified. Though Z-FR-AFC is an established cathepsin L substrate (Tchoupé et al., 1991), cathepsins B and H have affinities also for it. The general cathepsin inhibitor E-64 inhibits the activities of almost all cysteine proteases, including cathepsins B, H, and L (Barrett et al., 1982; Turk et al., 2012) but not cathepsin C (Rozman-Pungercar et al., 2003). Meanwhile, the inhibitor CA074 and its analogues specifically inhibit cathepsin B (Bogyo et al., 2000), whereas Z-Phe-Tyr(tBu)-diazomethylketone and analogues specifically inhibits cathepsin L (Shaw et al., 1993). Some of these inhibitors, among others, were used to characterize Da-CTSL1. To rule out cathepsin B, the specific cathepsin inhibitor, CA074, was assayed against Da-CTSL1. In the presence of CA074 (50 μ M), Da-CTSL1 still retained 75% of its activity (Fig. 10). The general cathepsin inhibitor E-64 and the specific cathepsin L inhibitor Z-FY-(tBu)DMK only showed limited inhibition of Da-CTSL1 (Fig. 10). The general cathepsin inhibitor chymostatin was effective at inhibiting Da-CTSL1 (Fig. 10), perhaps because it also inhibits serine proteases that are active at a basic pH. Additionally, aprotinin, a sole serine protease inhibitor, did not inhibit Da-CTSL1 at the same concentration range at which it inhibits serine proteases. Da-CTSL1 did not hydrolyze Z-GPR-AMC, Z-VVR-AMC, Suc-AAPF-pNA and H-Arg-pNA.2HCl probably because these substrates do not fit in the active

pocket of Da-CTSL1 (Fig. 3, B) and thus cannot be hydrolyzed either at acidic or basic
 pH as was shown for Z-GPR-AMC and Z-VVR-AMC (Fig. 8). The 3D modeling of Da-
 proCTSL1 predicts that the propeptide becomes inserted into the groove preventing the
 enzyme's substrate entry into the active groove and reaching C120 and H255 at the
 catalytic site (Fig. 3). The propeptide binding at the active groove is fortified by a
 network of fourteen hydrogen bonds (Fig. 3, B, C). Our 3D model of Da-proCTSL1 is
 supported by a report that in *Bombyx mori* incubation of a recombinant proregion of
 cathepsin L binds activated cathepsin L protecting it from denaturation in urea
 (Yamamoto et al., 1999). Mass spectrometry analysis (not shown) after SDS-PAGE
 analysis of purified Da-GST-proCTSL1 and Da-CTSL1 identified the cleavage site of the
 Da-proCTSL1 from Da-CTSL1 at L109 (Fig. 2) as was predicted from homology studies
 with other cathepsins (Fonseca et al., 2012). Our report shows that *D. abbreviatus*
 synthesizes in its gut a unique and basic Da-CTSL1 and the enzyme is not affected by
 chymotrypsin-specific inhibitor TPCK (Tosyl phenylalanyl chloromethyl ketone) (data not
 shown). Our results also suggest that 64% of all the gut total cathepsins is Da-CTSL1
 that is probably synthesized, activated and functions in the basic region of the insect's
 gut. Unlike acidic cathepsins which denature at basic pH Da-CTSL1 is stable and
 maximally active at pH 8 which allows it to synergize with *D. abbreviatus* trypsin that is
 active at a basic pH and is also synthesized by the gut of *D. abbreviatus* (Yan et al.,
 1999). The interesting implication in finding an unusual cathepsin L1-like protease that
 comprises 64% of all the cathepsin Ls found in *D. abbreviatus* allows this insect to
 become a highly polyphagous and invasive insect and precludes inhibitors that are
 commonly synthesized by plants to guard against damage from feeding insects from

inhibiting Da-CTSL1. This provides a mechanism for certain Coleoptera to support polyphagy and invasiveness.

Acknowledgements

This work has been supported by TSTAR USDA and Florida Citrus Advanced Technology Program grants to D.B., C.P and R.S. One of us (D.B.) is an established scientist fellow at the Oak Ridge Institute for Science and Education (ORISE). This research was supported in part by an appointment to the Agricultural Research Service (ARS) Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy (DOE) and the U.S. Department of Agriculture (USDA). ORISE is managed by ORAU under DOE contract number DE-AC05-06OR23100. All opinions expressed in this paper are the author's and do not necessarily reflect the policies and views of USDA, ARS, DOE, or ORAU/ORISE. We thank Dr. Lapointe at the USDA Ft. Pierce FL, for providing us with a photograph of the digestive tract of larval *D. abbreviatus*.

REFERENCES

- Barrett, A.J., Kembhavi, A.A., Brown, M.A., Kirschke, H., Knight, C.G., Tamait, M., Hanadat, K., 1982. Analogues as inhibitors of cysteine proteinases including cathepsins B , H and L. *Biochem. J.* 201, 189–198.
- Beavers, J. B., 1982. Biology of *Diaprepes abbreviatus* (Coleoptera: Curculinoidae) reared on an artificial diet. *Florida Entomologist* 65(2), 263-269.
- Bogyo, M., Verhelst, S., Bellingard-Dubouchaud, V., Toba, S., Greenbaum, D., 2000. Selective targeting of lysosomal cysteine proteases with radiolabeled electrophilic substrate analogs. *Chem. Biol.* 7, 27–38.

- Bown, D.P., Wilkinson, H.S., Jongsma, M. A., Gatehouse, J. A., 2004. Characterisation of cysteine proteinases responsible for digestive proteolysis in guts of larval western corn rootworm (*Diabrotica virgifera*) by expression in the yeast *Pichia pastoris*. *Insect Biochem. Mol. Biol.* 34, 305–320.
- Beton, D., Guzzo, C.R., Ribeiro, A.F., Farah, C.S., Terra, W.R. 2012. The 3D structure and function of digestive cathepsin L-like proteinases of *Tenebrio molitor* larval midgut. *Insect Biochem. Mol. Biol.* 42, 655-664.
- Butt, T.R., Edavettal, S.C., Hall, J.P., Mattern, M.R., 2005. SUMO fusion technology for difficult-to-express proteins. *Protein Expr. Purif.* 43, 1–9.
- Choe, Y., Leonetti, F., Greenbaum, D.C., Lecaille, F., Bogyo, M., Brömme, D., Ellman, J. a, Craik, C.S., 2006. Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *J. Biol. Chem.* 281, 12824–32.
- Clara, R.O., Soares, T.S., Torquato, R.J.S., Lima, C. a, Watanabe, R.O.M., Barros, N.M.T., Carmona, A.K., Masuda, A., Vaz, I.S., Tanaka, A.S., 2011. Boophilus microplus cathepsin L-like (BmCL1) cysteine protease: specificity study using a peptide phage display library. *Vet. Parasitol.* 181, 291–300.
- Coulombe, R., Li, Y., Takebe, S., Ménard, R., Mason, P., Mort, J.S., Cygler, M. 1996. Crystallization and preliminary X-ray diffraction studies of human procathepsin L. *Proteins* 25, 398-400.
- Cotrin, S.S., Puzer, L., de Souza Judice, W.A., Juliano, L., Carmona, A.K., Juliano, M.A., 2004. Positional-scanning combinatorial libraries of fluorescence resonance energy transfer peptides to define substrate specificity of carboxydipeptidases: assays with human cathepsin B. *Anal. Biochem.* 335, 244–52.
- Cristofolletti, P.T., Ribeiro, A.F., Terra, W.R., 2005. The cathepsin L-like proteinases from the midgut of *Tenebrio molitor* larvae: sequence, properties, immunocytochemical localization and function. *Insect Biochem. Mol. Biol.* 35, 883–901.
- Deraison, C., Darboux, I., Duportets, L., Gorojankina, T., Rahbé, Y., Jouanin, L., 2004. Cloning and characterization of a gut-specific cathepsin L from the aphid *Aphis gossypii*. *Insect Mol. Biol.* 13, 165–77.

- Edwards, M.G., Gatehouse, J. a, Gatehouse, a M.R., 2010. Molecular and biochemical characterisation of a dual proteolytic system in vine weevil larvae (*Otiorhynchus sulcatus* Coleoptera: Curculionidae). *Insect Biochem. Mol. Biol.* 40, 785–91.
- Fonseca, F.P.P., Soares-Costa, A., Ribeiro, A. F., Rosa, J. C., Terra, W. R., Henrique-Silva, F., 2012. Recombinant expression, localization and *in vitro* inhibition of midgut cysteine peptidase (*Sl-CathL*) from sugarcane weevil, *Sphenophorus levis*. *Insect Biochem. Mol. Biol.* 42, 58-69.
- Hernández, C. A, Pujol, M., Alfonso-Rubí, J., Armas, R., Coll, Y., Pérez, M., González, A., Ruiz, M., Castañera, P., Ortego, F., 2003. Proteolytic gut activities in the rice water weevil, *Lissorhoptrus brevisrostris* Suffrian (Coleoptera: Curculionidae). *Arch. Insect Biochem. Physiol.* 53, 19–29.
- Hunt, I., 2005. From gene to protein: a review of new and enabling technologies for multi-parallel protein expression. *Protein Expr. Purif.* 40, 1–22.
- Ishidoh, K., Kominami, K., 1994. Multi-step processing of procathepsin L in vitro. *FEBS Letters.* 352, 281-284.
- Kamboj, R.C., Pal, S., Raghav, N., Singh, H., 1993. A selective colorimetric assay for cathepsin L using Z-Phe-Arg-4-methoxy-beta-naphthylamide. *Biochimie* 75, 873–8.
- Kapust, R.B., Waugh, D.S., 1999. Escherichia coli maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci.* 8, 1668–74.
- Kaulmann, G., Palm, G.J., Schilling, K., Hilgenfeld, R., Wiederanders, B., 2006. The crystal structure of a Cys25->Ala mutant of human procathepsin S elucidates enzyme-prosequence interactions. *Protein Sci.* 15, 2619-2629.
- Kirschke, H., Langner, J., Wiederanders, B., Ansorge, S., Bohley, P., 1977. Cathepsin L: a new proteinase from rat liver lysosomes. *Eur. J. Biochem.* 74, 293-301.
- Koiwa, H., Shade, R.E., Zhu-Salzman, K., D'Urzo, M.P., Murdock, L.L., Bressan, R. a, Hasegawa, P.M., 2000. A plant defensive cystatin (soyacystatin) targets cathepsin L-like digestive cysteine proteinases (DvCALs) in the larval midgut of western corn rootworm (*Diabrotica virgifera virgifera*). *FEBS Lett.* 471, 67–70.
- Krieger, E., Koraimann, G., Vriend, G., 2002. Increasing the precision of comparative models with YASARA NOVA - a self-parameterizing force field. *Proteins* 47, 393-402.

- Larson, E.T., Parussini, F., Huynh, M.H., Giebel, J.D., Kelley, A.M., Zhang, L., Bogyo, M., Merritt, E.A., Carruthers, V.B., 2009. *Toxoplasma gondii* cathepsin L is the primary target of the invasion-inhibitory compound morpholinurea-leucyl-homophenyl-vinyl sulfone phenyl. J. Biol. Chem. 284, 26839-26850.
- Laskowski R.A., MacArthur M.W., Moss D.S., Thornton J.M., 1993. PROCHECK: a program to check the stereochemistry of protein structures, J. Appl. Cryst. 26, 283-291.
- Livak, K.J., Schmittgen, T. D., 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{-ddC_T} method. Methods 25, 402-408.
- Melo, F., Feytmans, E., 1998. Assessing protein structures with a non-local atomic interaction energy, J. Mol. Biol. 277, 1141-1152.
- Otto, H.-H., Schirmeister, T., 1997. Cysteine Proteases and Their Inhibitors. Chem. Rev. 97, 133–172.
- Pettersen E.F., Goddard T.D., Huang C.C., Couch G.S., Greenblatt D.M., Meng E.C., Ferrin T.E., 2004. UCSF Chimera - a visualization system for exploratory research and analysis, J. Comput. Chem. 25, 1605-1612.
- Rawlings, N.D., Barrett, A.J., Bateman, A., 2012. MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 40, D343–50.
- Rozman-Pungercar, J., Kopitar-Jerala, N., Bogyo, M., Turk, D., Vasiljeva, O., Stefe, I., Vandenabeele, P., Brömme, D., Puizdar, V., Fonović, M., Trstenjak-Prebanda, M., Dolenc, I., Turk, V., Turk, B., 2003. Inhibition of papain-like cysteine proteases and legumain by caspase-specific inhibitors: when reaction mechanism is more important than specificity. Cell Death Differ. 10, 881–8.
- Shaw, E., Mohanty, S., Colic, a, Stoka, V., Turk, V., 1993. The affinity-labelling of cathepsin S with peptidyl diazomethyl ketones. Comparison with the inhibition of cathepsin L and calpain. FEBS Lett. 334, 340–2.
- Smith, S. M., Gottesman, M. M., 1989. Activity and detection analysis of recombinant human cathepsin L expressed in *Escherichia coli*. J. Biol. Chem. 264, 20487-20495.

- Stachowiak, K., Tokmina, M., Karpińska, A., Sosnowska, R., Wiczak, W., 2004. Fluorogenic peptide substrates for carboxydipeptidase activity of cathepsin B. *Acta Biochim. Pol.* 51, 81–92.
- 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., Dalton, J.P., 2008. Structural and functional relationships in the virulence-associated cathepsin L proteases of the parasitic liver fluke *Fasciola hepatica*. *J. Biol. Chem.* 283, 9896–9908.
- Tchoupe, J.R., Moreau, T., Gauthier, F., Bieth, J.G., 1991. Photometric or fluorometric assay of cathepsin B, L and H and papain using substrates with an aminotrifluoromethylcoumarin leaving group. *Biochim. Biophys. Acta* 1076, 149–51.
- Terra, W., 1990. Evolution Of Digestive Systems Of Insects. *Annu. Rev. Entomol.* 35, 181–200.
- Thie, N.M.R., Houseman, J.G., 1990. Identification of cathepsin B, D and H in the larval midgut of colorado potato beetle, *Leptinotarsa decemlineata* say (Coleoptera: Chrysomelidae). *Insect Biochem.* 20, 313–318.
- Turk, B., Turk, D., Turk, V., 2000. Lysosomal cysteine proteases: more than scavengers. *Biochim. Biophys. Acta* 1477, 98–111.
- Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., Turk, D., 2012. Cysteine cathepsins: from structure, function and regulation to new frontiers. *Biochim. Biophys. Acta* 1824, 68–88.
- Yamamoto, Y., Watabe, S., Kageyama, T., Takahashi, S. Y., 1999. Proregion of *Bombyx mori* cysteine proteinase functions as an intramolecular chaperone to promote proper folding of the mature enzyme. *Arch. Insect Biochem. Physiol.* 42, 167–178.
- Yan, X.H., De Bondt, H.L., Powell, C.C., Bullock, R.C., Borovsky, D., 1999. Sequencing and characterization of the citrus weevil, *Diaprepes abbreviatus*, trypsin cDNA. Effect of *Aedes* trypsin modulating oostatic factor on trypsin biosynthesis. *Eur. J. Biochem.* 262, 627–36.
- Zhu-Salzman, K., Koiwa, H., Salzman, R. a, Shade, R.E., Ahn, J.-E., 2003. Cowpea bruchid *Callosobruchus maculatus* uses a three-component strategy to overcome a plant defensive cysteine protease inhibitor. *Insect Mol. Biol.* 12, 135–45.

Figure 1

Figure 2

A

B

1 MKVFIAACLLVAVSATVLEETGVKFQAFKLKHGKTYKNQVEET 43
44 ARFNIFKDNLRAIEQHNVLVEEQGLVSYKKGINRFTDMTQEEFR 86
87 AFLTLSSSKKPHFNTTEHVLTGLAVPDSIDWRTKGQVTGVKD 128
129 QGNCGSCWAFSVTGSTEAAYYRKAGKLVSLSEQQLVDCSTD 169
170 INAGCNGGYLDETFTYVKS KGLEAESTYPYKGT DGSC KYSAS 211
212 KVVTKVSGHKSLKSEDENALLDAVGNVGPVSVAIDATYLSSE 254
255 SGIYEDDWCSPELNHGVLVVG YGTSNGKKYWIVKNSWGGSF 296
297 GESGYFRLLRGKNECGVAEDTVYPII 322

C

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8

Figure 9

Figure 10

Figure legends

Fig. 1. A. Pie Chart of the relative proportion of *D. abbreviatus* cathepsin L clones in the cDNA library of cathepsin L digestome proteases of 5-week old larvae. Cathepsin L groups at 99% identity were used in the construction of the pie-chart. Da-CTSL groups greater than 2% are labeled. Da-CTSL1 is the major cathepsin L (64%) of the library containing 24 cathepsin L groups that were identified. **B.** Transcript abundance of Da-CTSL1 showing fold difference from 18S rRNA in the foregut (FG), anterior midgut (AMG), posterior midgut (PMG) and hindgut (HG) calculated by the $\Delta\Delta C_T$ method. In each experiment RNA (100 ng) from gut segments were assayed by RT-qPCR. **C.** A light microscopy of digestive tract of larval *D. abbreviatus* showing the foregut (FG), anterior midgut (AMG), posterior midgut (PMG) and hindgut (HG) regions.

Fig. 2. *D. abbreviatus* pre-pro-CTSL1. A. schematic presentation of Cathepsin L1 with the pre region (amino acids 1-15) pro-region (amino acids 15-109) and the active enzyme (amino acids 109-322). **B.** The amino acids sequence of Da-pre-pro-CTSL1 (1-322). The four conserved amino acids at the active pocket (Q, C, H and N are red). The pre region is cleaved at A15 and the pro region is cleaved at L109 during releasing an active Da-CTSL1. **C.** A phylogenetic tree constructed by ClustalW (DNASTar v. 12) using the following sequences: *D. abbreviatus* (accession number ABG73217.1), *T. molitor* (accession number ABC88769.1), *Triatoma brasiliensis* (accession number ACF48469), *Aphis gossypii* (accession number CAD33266) , *Pediculus humanus corporis* (accession number FEB177709), *Spodoptera frugiperda* (accession number ADN19567), *Helicoverpa armigera* (accession number NP_203611),

Drosophila melanogaster (accession number AAF58311) , *Aedes aegypti* (accession number ABE72970), *Anopheles darling* (accession number ETN64675), and *Culex quinquefasciatus* (accession number EDS42770).

Fig. 3. Ribbon diagram showing the overall three-dimensional conformation of *D. abbreviatus* pro-CTSL1. **A.** The propeptide (PP) and the catalytic domain are colored violet red and yellow, respectively with the catalytically active C20 and H255 residues. **B.** Molecular surfaces showing the blocking of the catalytic groove (black dashed lines) of the activated Da-CTSL1 by its propeptide (PP) and the catalytically active C120 and H255 residues. **C.** Enlarged view of the blocking of the catalytic residues C120 and H255 by the C-terminal tail of the propeptide (PP) found in the activity groove of the catalytic domain of Da-CTSL1.

Fig. 4. Cloning strategy of Da-CTSL1. **A.** Expression vector pET-41a with His₆ and Gutathione S-Transferase (GST) tags on the 5' and 3' of the cloning region was digested with *Xho*I and *Spe*I. **B.** Da- proCTSL1 fragment was amplified by PCR and digested with *Xho*I and *Spe*I. **C.** Da-proCTSL1 fragment was then ligated into pET-41a vector between His₆ and GST. **D.** *E. coli* cells were transformed with pET-41a Da-proCTSL1, induced with IPTG and expressed a 60 kDa recombinant GST-pro-CTSL1-His₆ protein containing pro- and mature segments (10.9 kDa and 22.1 kDa, respectively) and His₆ and GST tags (1 kDa and 26 kDa, respectively). His₆=yellow, GST= red, mature CTSL1= dark green, Pro region= brown green.

Fig. 5. Purification and activation of recombinant Da-GST-proCTSL1-His₆. Proteins were extracted from transformed Rosetta competent bacterial cells in the presence of 1% Triton X-100. The 60 kDa fusion protein was purified by Ni-NTA affinity

chromatography. Autocatalysis and purification of mature Da-CTSL1 from the fusion protein was done at pH 8 using a size-exclusion chromatography

Fig. 6. A. SDS-PAGE analysis of recombinant Da-CTSL1 during its purification. Lanes: St= molecular weight markers, a= control, *E. coli* cells transformed with plasmid that did not carry the Da-proCTSL1 insert (150 μ g protein), b= crude bacterial extract (145 μ g protein) of recombinant *E. coli* containing Da-pro-CTSL1 (M_r 60KDa) (arrow), c= Da-proCTSL1 (144 μ g protein; M_r 60KDa) after affinity chromatography (arrow), d= Da-CTSL1 (50 μ g protein; M_r 23 kDa) after activation at pH 8.0 and purification by exclusion chromatography (arrow). **B.** Western blot analysis (144 μ g protein). Right lane (Da-CTSL1) after affinity chromatography of Da-GST-proCTSL1His₆ (M_r 60 kDa) upper band and activated Da-CTSL1 (M_r 23 kDa) lower band. Left lane is Western blot analysis of protein extract from *E.coli* cells (144 μ g protein) transformed without the Da-GST-proCTSL1His₆ insert (control) after affinity chromatography. **C.** Mass spectra analysis of Da-GST-proCTSL1-His₆. The GST sequence is found at the beginning of the mass spectrum and the His₆ sequence is found at the end of the mass spectrum. Yellow color shows positively identified peptides (34 unique peptides with 100% probability for a correct sequence).

Fig. 7. Size-exclusion chromatography of Da-CTSL1 after activation at pH 8.0. Protein peak A is Da-GST-proCTSL1-His₆ (M_r 60 kDa) eluted between 42 and 50 mL. Peak B is the activated Da-CTSL1 (M_r 23 kDa) eluted between 52 to 64 mL (double headed arrow). Protein was followed at 280 nm (blue line). Constant column pressure was maintained throughout the run (green line).

Fig. 8. A. Enzymatic activity of Da CTSL1 using the Fluorogenic substrate (Z-R-AFC) at different times of incubation (0-240 min) at pH 8.0. Activity is linear for 120 min at a molar ratio of substrate to enzyme of 4:1 . **B.** Fluorogenic substrate preference of activated Da-CTSL1 at acidic and basic pHs. Activities are expressed in relative fluorescent units (RFU) as means of 3-5 determinations $\pm$ SEM.

Fig. 9. pH activity profile of activated and purified recombinant Da-CTSL1 in the presence of Z-FR-AFC. Activities at different pHs are expressed in relative fluorescent units (RFU) as means of 3-5 determinations $\pm$ SEM after 30 min incubation of Da CTSL1 with Z-FR-AFC.

Fig. 10. Testing activated purified Da-CTSL1 in the presence of commercially available protease inhibitors. Enzymatic activities were followed using the fluorogenic substrate Z-FR-AFC in the presence of different inhibitors at pH 8.0 using molar ratios of substrate to enzyme of 4:1, and molar ratio of inhibitor to enzyme of 17:1. Activities are expressed in relative fluorescent units (RFU) as means of 3-5 determinations $\pm$ SEM and were linear throughout the incubation period.

Table 1. Complementary DNA clones constructed from expressed sequence tags from digestive tracts isolated from *D. abbreviatus* life stages.

Source of Digestive Tract Tissue	Number of cDNA Clones Sequenced
5 Week Old Larvae	1504
Unfed Teneral Adults	1545
Feeding Adults	1538

Total RNA was extracted from whole gut extracts, and reverse transcribed into cDNA and ligated into plasmids. The recombinant DNA was transformed into bacteria to construct a cDNA library.

Table 2. Breakdown of digestome proteases present in cDNA library constructed from life stages of *D. abbreviatus*.

Digestome Proteases	Number of cDNA clones		
	5 week old larvae	Unfed Teneral Adults	Feeding Adults
Aspartic Protease-1	3	0	0
Aspartic Protease-2	0	0	1
Cathepsin B-1	7	0	2
Cathepsin B-2	3	0	1
Cathepsin B-3	0	0	1
Cathepsin L-1-24	216 (84%)	1	56 (80%)
Chymotrypsin-1	2	0	0
Chymotrypsin-2	1	0	0
Chymotrypsin-3	1	0	0
ChymotrypsinB1	0	0	1
Serine Carboxypeptidase	0	0	2
Trypsin	17	0	0
Trypsin	1	0	3
Trypsin	1	0	2
Trypsin	2	0	0
Trypsin	2	0	0
Trypsin	1	0	0
Trypsin	0	0	1
xaa-pro dipeptidase-1	0	1	0

A number of digestive protease sequences found in the cDNA clones generated from the guts of 5-week old larvae, unfed teneral and feeding adults are compared with the number of cathepsin L sequences in the library. The proportion of cathepsin L forms in total cDNA clones is shown as a percentage in parentheses.

Table 3. BLASTP of *D. abbreviatus* cathepsin L1 (Da-CTSL1)

Description	Max Score	Query Cover	E-value	Identity	Accession
Cathepsin L 1 precursor [<i>Diaprepes abbreviatus</i>]	664	100	0.0	100	ABG73217.1
Cysteine proteinase [<i>Hypera postica</i>]	475	100	2e-165	68	AAD41105.1
Cysteine proteinase [<i>Anthonomus grandis</i>]	422	100	2e-144	62	AAR02406.1
Cathepsin L-like proteinase [<i>Diabrotica virgifera virgifera</i>]	311	100	5e-101	49	CAE47501.1
Cathepsin L-like proteinase [<i>Triatoma infestans</i>]	301	100	5e-97	48	AAR12010.1
Putative cathepsin L-like proteinase [<i>Tenebrio molitor</i>]	296	100	3e-95	45	ABC88769.1
Cysteine proteinase 1 [<i>Chrysomela tremulae</i>]	294	100	7e-97	48	ACP18830.1
Digestive cysteine protease intestain [<i>Leptinotarsa decemlineata</i>]	290	100	1e-92	46	ABM55480.1
Putative gut cathepsin L-like cysteine protease [<i>Callosobruchus maculatus</i>]	278	95	8e-91	47	AAQ11969.1
Cysteine proteinase [<i>Sitophilus zeamais</i>]	261	95	3e-84	43	BAA24444.1
Cathepsin S [<i>Homo sapiens</i>]	256	100	2e-81	46	AAC37592.1
Cathepsin L1 [<i>Homo sapiens</i>]	254	99	9e-81	44	AAI42984.1
Cathepsin K [<i>Homo sapiens</i>]	243	99	2e-76	42	CAG46520.1
Cathepsin X [<i>Homo sapiens</i>]	239	99	8e-75	41	AAA95998.1
Cathepsin H [<i>Homo sapiens</i>]	230	98	3e-71	40	AAL23961.1

The amino acid query sequence of Da-CTSL1 was compared to the protein sequence dataset of coleopteran and human proteins from the NCBI dataset. Max score reflects the best aligned sequence.