Serine and cysteine protease-like genes in the genome of a gall midge and their interactions with host plant genotypes

Hang Chen, Yu Cheng Zhu, R. Jeff Whitworth, John C. Reese, and Ming-Shun Chen

Published Version Information

Citation: Chen, H., Zhu, Y. C., Whitworth, R. J., Reese, J. C., & Chen, M.-S. (2013). Serine and cysteine protease-like genes in the genome of a gall midge and their interactions with host plant genotypes. Insect Biochemistry and Molecular Biology, 43(8), 701-711.

Digital Object Identifier (DOI): doi:10.1016/j.ibmb.2013.05.006

Publisher's Link: http://www.sciencedirect.com/science/article/pii/S0965174813001008

For: Insect Biochemistry and Molecular Biology

Serine and Cysteine Protease-like Genes in the Genome of a Gall Midge and Their Interactions with Host Plant Genotypes*

Hang Chen[1,2], Yu Cheng Zhu[3], R. Jeff Whitworth[1], John C. Reese[1], and Ming-Shun Chen[1,4†]

[1]Department of Entomology, Kansas State University, Manhattan, KS 66506, USA

[2]Research Institute of Resource Insect, Chinese Academy of Forestry, Kunming 650224, China

[3]USDA-ARS-JWDSRC, PO Box 346/141 Exp Stn Rd, Stoneville, MS 38776, USA

[4]Hard Winter Wheat Genetics Research Unit, USDA-ARS, 4008 Throckmorton Hall, Kansas State University, Manhattan, KS 66506

*Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the U.S. Department of Agriculture. USDA is an equal opportunity provider and employer.

Keywords: Hessian fly genome, *Mayetiola destructor*, serine protease, cysteine protease, gut proteinases.

†Communication should be addressed to Ming-shun Chen, Email: mchen@ksu.edu, ming-shun.chen@ars.usda.gov. Phone: 785-532-4719. Fax: 785-532-6232.

Abstract

Proteases play important roles in a wide range of physiological processes in organisms. For plant-feeding insects, digestive proteases are targets for engineering protease inhibitors for pest control. In this study, we identified 105 putative serine- and cysteine-protease genes from the genome of the gall midge *Mayetiola destructor* (commonly known as Hessian fly), a destructive pest of wheat. Among the genes, 31 encode putative trypsins, 18 encode putative chymotrypsins, seven encode putative cysteine proteases, and the remaining may encode either other proteases or protease homologues. Developmental stage- and tissue-specific expression profiles of the genes encoding putative trypsins, chymotrypsins, and cysteine proteases were determined by quantitative reverse-transcription PCR. Comparative analyses of stage- and tissue-specific expression patterns suggested that several genes are likely to encode digestive proteases in the *M. destructor* larval gut, including genes encoding putative trypsins MDP3, MDP5, MDP9, MDP24, MDP48, MDP51, MDP57, MDP61, MDP71, and MDP90; genes encoding putative chymotrypsins MDP1, MDP7, MDP8, MDP18, MDP19, and MDP20; and genes encoding putative cysteine proteases MDP95 and MDP104. The expression of some protease genes was affected by plant genotypes. Genes encoding trypsins MDP3, MDP9, and MPD23, chymotrypsins MDP20 and MDP21, and cysteine proteases MDP99 and MDP104 were upregulated in *M. destructor* larvae feeding in resistant plants, whereas genes encoding trypsins MDP12, MDP24, and MDP33, and chymotrypsins MDP8, MDP15, and MDP16 were downregualted in *M. destructor* larvae feeding in resistant plants. This study provides a foundation for further comparative studies on proteases in different insects, and further characterization of *M. destructor* digestive proteases and their interactions with host plants, as well as potential targets for transgenic wheat plants.

1. Introduction

Proteases are involved in various physiological and biochemical processes such as signal transduction, digestion, development, and defense reactions (Neurath and Walsh, 1976; Barrett et al., 1998; Neurath, 1999). In plant-feeding insects, proteases are involved in plant – insect interactions and participate in the arms race between hosts and herbivores (Bown et al., 1997; Pechan et al., 2002; Jongsma and Beekwilder, 2011). In herbivorous insects, proteases produced in salivary glands can be injected into host plants for pre-oral digestion and for removing toxic proteins produced by plants for defense (Miles, 1999; Eberhard et al., 2007). In the gut of most insects, the main role of proteases is to digest proteins ingested from host plants (Shukle et al., 1985; Lopes et al., 2006; Wright et al., 2006; Sato et al., 2008). During the long course of co-evolution, plants have developed defense mechanisms that suppress protease activities in the insect digestive system. Specifically, plants enhance the production of various protease inhibitors that reduce activities of insect digestive proteases in response to herbivory (Moura and Ryan, 2001; Habib and Fazili, 2007; Chen, 2008). Accordingly, these plant inhibitors could be targets for genetic engineering to increase plant resistance to herbivorous pests (Burgess et al., 1994; Murdock et al., 1988; Lawrence and Koundal, 2002). However, insects have also developed counter-defense mechanisms including up-regulation of constitutively expressed digestive proteases, induction of proteases that are normally not expressed, and/or by synthesizing inhibitor-insensitive proteases in response to plant defense (Jongsma et al., 1995; Bown et al., 1997; Mazumdar-Leighton and Broadway, 2001; Ahn et al., 2004; Brioschi et al., 2007). In order to make plant protease inhibitors an effective tool for pest management, a comprehensive understanding of protease composition in insect genomes and their regulatory

mechanisms for expression in response to plant defense is needed to design feasible strategies for engineering plant inhibitors.

One of the gall midges, *Mayetiola destructor*, commonly known as the Hessian fly, is one of the most destructive arthropod pests of wheat (Hatchett et al., 1987; Buntin, 1999; Pauly 2002). The digestive proteases in Hessian fly larvae are apparently targets for plant defense under natural conditions because a range of protease inhibitors are highly upregulated in resistant wheat seedlings during incompatible interactions (Liu et al., 2007; Wu et al., 2008). The upregulated protease inhibitors in resistant wheat plants may be part of the defense mechanism that results in the death of Hessian fly larvae (Stuart et al., 2012). Transcriptomic analyses have identified several genes encoding trypsins and chymotrypsins that are exclusively or highly expressed in the gut of Hessian fly larvae (Zhu et al., 2005). Protease activity has also been detected in gut extract from Hessian fly larvae (Shukle et al., 1985). These observations indicate that digestive proteases in Hessian fly larval gut could be targets for engineering effective plant inhibitors to increase wheat resistance to the pest. A better understanding of the composition of proteases in the Hessian fly genome, major digestive proteases in different larval instars, and the expression dynamics of the protease genes in response to plant defense should provide useful information for utilizing protease inhibitors to enhance host resistance for controlling Hessian fly damage. The objectives of this research were to: 1) Identify putative serine- and cysteine-protease genes in the Hessian fly genome; 2) Identify digestive proteases in Hessian fly larval gut; 3) To determine expression patterns of major digestive protease genes at different larval stages; and 4) Determine changes in expression of protease genes in response to plant defense.

2. Materials and methods

2.1. Hessian fly

92 Hessian fly larvae used in this research were derived from a field collection from Scott
93 County, Kansas in 2005 (Chen et al., 2009). The insects have been maintained on seedlings of
94 Hessian fly-susceptible wheat variety 'Karl 92' in greenhouse since then. The majority of flies
95 were biotype *GP* although biotypes virulent to known R genes were also found in low
96 frequencies (Chen et al., 2009).

97 **2.2. Identification of putative protease genes from the Hessian fly genome**

98 The overall conservation of different trypsins, chymtrypsins, and cysteine proteases in an
99 organism and across different species facilitated the identification of new proteases within a
100 genome. To identify potential new Hessian fly protease genes, the draft Hessian fly genome
101 sequence (http://agripestbase.org/hessianfly/) was searched using Blastx with known Hessian
102 fly trypsins and chymotrypsins (Zhu et al., 2005), two partial sequences of Hessian fly cysteine
103 proteases obtained from a gut transcriptome, and *Drosophila* serine- and cysteine-protease as
104 queries (Matsumoto et al., 1995; Ross et al., 2003). From the resulting list of similar sequences,
105 every 15th sequence was retrieved for another round of blasting. All the amino acid sequences
106 encoded by predicted genes with significant sequence similarity (*E*-value < 0.01) from each of
107 queries were retrieved. Repeated blasting was carried out until no new hit with *E*-value < 0.01
108 could be found. Introns were identified by comparing genomic sequence with a corresponding
109 cDNA sequence or a putative transcript predicted by MAKER2 (Holt and Yandell, 2011). If no
110 cDNA sequence or transcript was available for a particular gene, the intron/exon boundary was
111 determined manually by a Blastx alignment between the Hessian fly gene sequence and the
112 query protein sequence. Sequences were discarded if they could not be annotated. The identity
113 of the annotatable sequences was further confirmed using MotifScan (Yusim et al., 2004) and

ScanProsite tool (Gattiker et al., 2002) to reveal their characteristic sequence patterns and putative enzyme active sites.

2.3. RNA isolation and PCR analyses

Total RNA was extracted from either whole flies or specific fly tissues using RNeasy Micro Kit according to the procedure provided by the manufacturer (QIAGEN, Maryland, USA). Whole body samples were extracted from larvae of 1, 3, 6, and 12 days old, respectively, pupae and adults. Tissue-specific samples were extracted from dissected guts, fat bodies, salivary glands, Malpighian tubules, and the remaining carcass. These tissues were obtained from 3-day-old larvae (first instar).

Primers were designed using the Beacon Designer 7.0 software (Biosoft, Palo Alto, CA). Primer sequences were listed in Table S1. RNA samples were treated with RNase-free DNase-I (Promega, Madison, WI) to remove potential DNA contamination. The RNA was then reverse-transcribed into cDNA using $(oligo\text{-}dT)_{20}$ primers with the SuperScript® III First-Strand Synthesis System (Invitrogen, Carlsbad, CA, USA).

For semi-quantitative PCR, amplification was carried out for 25 to 40 cycles, depending on the intensity of PCR products, as follows: 60 s at 94ºC; 60 s at 55ºC; 120 s at 72ºC. DNA fragments from the PCR reactions were separated on 1.5 to 2% agarose gels depending on the size of the product and stained with (0.5 μg/mL) ethidium bromide. Actin was used as a control for normalization. DNA bands were photographed with a Bio Doc-It^{TM} System (UVP, Upland, CA) and band intensity was determined using Photoshop CS image analysis software (Adobe Systems Incorporated, San Jose, California).

For quantitative real-time PCR (qPCR), amplification was performed with iQ SYBR Green Supermix on a iCycler real time detection system (Bio-Rad, Hercules, CA). Each reaction was carried out with 2 μl of a 1/40 (v/v) dilution of the first cDNA strand, 0.5 μM of each primer

in a total volume of 25 µl. The cycling conditions were: 95°C for 5 min followed by 45 cycles of denaturation at 95°C for 20 s, annealing and extension at 62°C to 64.5°C, depending on the primer set, for 45 s. At the end of the cycles, PCR amplification specificity was verified by obtaining a dissociation curve, derived by cooling the denatured samples to 55°C and raising the temperature 0.5°C for 10 s for each cycle, for a total of 80 cycles until reaching 95°C. The PCR products were analyzed on 1.5% agarose gels, and subsequently purified and sequenced to confirm faithful amplification. Actin was selected as a reference for normalization of template concentration. Three independent biological replicates were carried out for each treatment.

Statistical significance for the log-transformed arbitrary expression values was analyzed by ANOVA using the PROCMIXED procedure of SAS (SAS institute Inc., SAS/STAT User's Guide, Version 9.13). Tukey's pairwise comparisons based on Student's range statistics were then conducted. Tukey's 95% simultaneous confidence intervals for pair-wise comparisons were used to separate data into groups with significant differences.

2.4. Sequence data processing and phylogenetic analysis

Molecular weight calculations and pI prediction of mature proteins were carried out with the 'Compute pI/Mw tool' (http://us.expasy.org/tools/pi_tool.html, Bjellqvist et al., 1993). Signal peptide cleavage sites were predicted using SignalP 4.1 Server (Petersen et al., 2011). Multiple alignments of gene data matrices and protein sequences were generated using similarity calculated with ClustalW (Larkin et al., 2007) and BioEdit (Ver. 5.09; Hall, 1999). Molecular and phylogenetic analyses were conducted using MEGA Version 5 software (Tamura et al., 2011). The phylogeny was inferred by using the Maximum Likelihood (ML) method for pairwise distance calculation. Phylogenetic tree was constructed using Neighbor Joining (NJ) and BIONJ algorithms (Tamura et al., 2004). Bootstrap analyses were conducted (1000

repeats) by the majority rule (70% or above) for grouping to ensure nodal reliability (Felsenstein, 1985).

3. Results

3.1. Serine and cysteine protease-like genes in the Hessian fly genome

Blast searches of the Hessian fly genome with known Hessian fly and *Drosophila* proteases (see Materials and Methods) identified 105 serine- and cysteine- protease genes. These genes were named *Mayetiola destructor* protease 1 (*MDP1*) to *MDP105*. Analysis of enzymatic active sites of the predicted proteins revealed that 94 of them are serine protease-like proteins and the remaining 11 are cysteine protease-like proteins. Of the 94 serine protease-like proteins, 31 were identified to be putative trypsins, 18 putative chymotrypsins, 29 trypsin/chymotrypsin homologs (Kwon et al., 2000), and 16 truncated proteins that could not be classified (Figure S1). Among the 11 cysteine protease-like proteins, seven of them possess all functionally critical residues (see below) and therefore were taken as putative cysteine proteases, whereas the other four are homologs that do not have all the functionally important residues conserved (Figure S1).

Phylogenetic analysis of Hessian fly serine proteases along with known serine protease sequences from *Drosophila melanogaster* revealed five major groups (Figure 1A). All putative Hessian fly chymotrypsins are clustered within groups 1 and 2. Thirteen putative Hessian fly trypsins are in group 1, while the remaining 18 trypsins are scattered into the other four groups. The putative Hessian fly cysteine-proteases were analyzed along with known cysteine proteases from other insects (Figure 1B). Hessian fly cysteine proteases are clustered with Drosophila cysteine proteases except MDP95, MDP96, and MDP97, which either form an independent group or clustered together with cysteine proteases from other insects.

Other characteristics of the putative trypsins, chymotrypsins, and cysteine proteases are given in Table 1, including predicted molecular weight (MW), isoelectric point (pI), number of amino acids, and first hit information of BLAST searches (GenBank accession no., scores, E-value, and name of the organism). Ten of the 31 putative trypsin precursors are full length (starting with methionine), and 9 of them were predicted to have 17-23 residue signal peptide, suggesting that most trypsins are not membrane-bound. Similarly, most putative chymotrypsin precursors have 19-21 residue signal peptide, and most cysteine-protease precursors have 16-22 residue signal peptide.

3.2. Putative trypsin genes

By using blast similarity search of GenBank and MotifScan and ScanProsite tools, sequence pattern and functional motifs for trypsin were revealed, including charge relay system or active site residues H, D. and S, and three pairs of cysteines for disulfide bonds to confirm the trypsin gene identity. An alignment of the 31 putative trypsins is shown in Figure 2A. The specificity determinant residue D^{312} for trypsins is conserved in all proteins. The catalytic triad H^{104}, D^{185}, and S^{318} residues are also conserved in all members. The overall sequence identities among the putative trypsins are very low. The two most closely related proteins, MDP-3 and MDP-5, exhibit 61.3% sequence identity (Table S2). The two most diversified proteins, MDP-58 and MDP-79, exhibit only 20.7% sequence identity. The majority of members share 25 – 35% identity.

3.3. Putative chymotrypsin genes

Chymotrypsins share major sequence patterns and functional motifs with trypsins. The major difference between a trypsin and a chymotrypsin is that trypsins are characterized by the

206 presence of specificity determinant residue D^{312} (Figure 2A), while chymotrypsins usually have
207 G or S at corresponding position (residue 298, Figure 2B) (Hedstrom et al., 1992; Wang et al.,
208 1993). Trypsins cleave the protein chain on the carboxyl side of arginine or lysine, whereas
209 chymotrypsin cleaves on the carboxyl side of aromatic amino acids (Terra and Ferreira, 1994).
210 An alignment of the 18 putative chymotrypsins is given in Figure 2B. The catalytic triad H^{119},
211 D^{185}, and S^{304} are conserved in all putative chymotrypins. However, the specificity determinant
212 residue G/S^{298} is not found in some members at the exact position in the alignment. Specifically,
213 the specificity determinant residue is Y in MDP-15 and MDP-16, F residue in MDP-17, and H
214 residue in MPD-84 and MDP-85. These five proteins are either chymotrypsin homologs without
215 enzymatic activity, or the specificity determinant residue is located in nearby positions, which
216 were not correctly positioned in the alignment. Like trypsins, the putative chymotrypsins are
217 also highly diversified. The two most closely related proteins, MDP-15 and MDP-16, share
218 81.5% sequence identity (Table S3). The two most diversified proteins, MDP-7 and MDP-85,
219 share only 18.3% identity. The majority of the proteins share 22-32% sequence identity.

220 **3.4. Putative cysteine-protease genes**

221 Like trypsins and chymotrypsins, cysteine proteases are also involved in the
222 physiological protein breakdown, but they are optimally active in the slightly acidic condition
223 (Turk et al., 2001). Cystene proteases also have different active site residues compared to those
224 in trypsins and chymotrypsins. By using MotifScan and ScanProsite, the identity of several
225 cysteine proteases were confirmed with the presence of active residue cysteine and two
226 additional active site residues, histidine and asparagines (Dufour, 1988). An alignment among
227 the seven putative cysteine-proteases is given in Figure 2C. The alignment at the N-terminal
228 region exhibits very different sequences with little similarity. The C-terminal region, in

229 comparison, is relatively conserved. The three important residues C^{391}, H^{570}, and N^{591} at the
230 active site are located at the C-terminal region and are conserved in all members. Sequence
231 identity between these proteins ranged from 22 to 64% with the exception of MDP-99 and MDP-
232 100, which are identical (Table S4). MDP-99 and MDP-100 are encoded by two tandem genes
233 (AEGA01013770).

234 **3.5. Stage-specific expression of trypsin, chymotrypsin, and cysteine protease genes**

235 Quantitative reverse-transcription PCR (RT-PCR) analyses revealed different patterns of
236 transcript abundance with different protease genes in Hessian flies at different developmental
237 stages (Figure 3). There were only a few genes, such as MDP-95 and MDP-100 (Figure 3C),
238 whose transcripts were relatively equally distributed in larvae, pupae and adults. The majority of
239 the genes exhibited higher levels of transcripts at a certain stage(s) of fly development. Overall,
240 more genes exhibited higher transcript levels in 3- to 12-day old larvae and fewer genes
241 exhibited higher transcript levels in adults and 1-day old larvae.

242 **3.6. Tissue-specific expression of trypsin, chymotrypsin, and cysteine protease genes**

243 Since the Hessian fly larva is the only feeding stage and 3-day-old larvae are most active,
244 the tissues of 3-day larvae including the gut, salivary glands, fat bodies, Malphigian tubules, and
245 the remaining carcass were dissected for RT-PCR analyses. The results of larval tissue analysis
246 also revealed different patterns of transcript abundance with the different protease genes in
247 different tissues (Figure 4). For the trypsin genes, MDP48 was exclusively expressed in the
248 larval gut. MDP9, MDP3, MDP72, MDP51, MDP24, and MDP23 exhibited higher transcript
249 levels in gut tissues than in other tissues. MDP90, MDP5, and MDP57 exhibited higher
250 transcript levels in both the gut and salivary gland samples. For chymotrypsin genes, MDP1,

MDP8, and MDP2 exhibited higher transcript levels in the gut sample than in other tissues. MDP7, MDP20, and MDP21 exhibited higher transcript levels in both the gut and salivary gland samples than in other tissues. None of the cysteine protease genes exhibited predominant abundance in the gut or salivary gland samples.

3.7. Host plant genotypes affect the expression of several protease genes

To examine if plant defense affects the expression of protease genes, Hessian fly larvae were reared on the near-isogenic lines Newton and Molly. Newton is a Hessian fly susceptible wheat line with no Hessian fly resistance genes, while Molly is a back-cross offspring of Newton, but contains the resistance gene *H13* (Patterson et al., 1994). Hessian fly larvae of avirulent biotypes die without development in resistant Molly plants. The majority of protease genes showed no significant differences in transcript abundance in Hessian fly larvae reared on susceptible Newton and resistant Molly (data not shown). However, seven protease genes exhibited increased transcript abundance and six protease genes exhibited decreased transcript abundance in larvae feeding on resistant Molly plants (Figure 5).

The protease genes with increased transcript levels in larvae feeding in resistant Molly plants included genes encoding trypsins MDP3, MDP9, and MDP23; Chymotrypsins MDP20 and MDP21; and cysteine proteases MDP99 and MDP104 (Figure 5A). For the three trypsin genes, *MDP3* and *MDP9* were expressed at low levels in larvae feeding in susceptible Newton, but transcript levels increased approximately three-fold in both 1-day and 3-day old larvae feeding in resistant Molly. For *MDP23*, transcript was essentially undetected in larvae feeding in susceptible plants, but was abundantly expressed in both 1-day and 3-day old larvae feeding in resistant plants. For the chymotrypsin genes, both *MDP20* and *MDP21* were undetectable by RT-PCR in larvae feeding on susceptible Newton, but were transiently upregulated to high levels

in 1-day old larvae feeding in resistant Molly plants. For the cysteine protease genes, transcript for *MDP99* was not detectable in larvae feeding in susceptible Newton plants, but was upregulated to high level in larvae feeding in resistant plants. The transcript of *MDP104* was expressed at low level in 1- and 3-day old larvae feeding in susceptible plants, and was upregulated to higher levels in larvae feeding in resistant plants. The protease genes with decreased transcript abundance in larvae feeding in resistant plants included genes encoding trypsins *MDP12*, *MDP24*, and *MDP33*; and chymotrypsins *MDP8*, *MDP15*, and *MDP16* (Figure 5B). These protease genes were expressed abundantly at least in 3-day old larvae feeding in susceptible plants, but were essentially undetectable in larvae feeding in resistant plants.

4. Discussion

Our main interest was to identify proteases involved in the digestive system of Hessian fly larvae, the only feeding stage of the insect. Stage- and tissue-specific expression analyses revealed that trypsins MDP3, MDP5, MDP9, MDP24, MDP48, MDP51, MDP57, MDP61, MDP71, and MDP90, chymotrypsins MDP1, MDP7, MDP8, MDP18, MDP19, and MDP20, and cysteine proteases MDP95 and MDP104 are candidates as digestive enzymes in Hessian fly larvae. Interestingly, the majority (11 out 16) of the putative digestive trypsins and chymotrypsins belong to the phynogentic group 1, and the remaining belongs to group 5 (Figure 1A). Group 1 contains the largest number of Hessian fly trypsins and chymotrypsins. We speculate that the expansion of group 1 is related with Hessian fly adaptation to changes in host plants. Consistent with this speculation, all trypsin and chymotrypsin genes except MDP33 that were either up- or down-regulated on resistant host plants belong to this group (Figure 1A). Phytophagous insects alter the expression of different digestive enzymes in response to plant

297 defense (Bown et al., 1997). Wheat plants with an effective Hessian fly-resistance gene produce
298 elevated levels of different types of protease inhibitors in response to Hessian fly attack, whereas
299 the expression of inhibitor genes is suppressed in susceptible plants (Liu et al., 2007; Wu et al.,
300 2008). Therefore, genes encoding digestive proteases may exhibit differential expression
301 patterns in Hessian fly larvae feeing in susceptible plants from those feeding in resistant plants,
302 potentially due to the change in concentrations of protease inhibitors in host plants. The change
303 in expression levels of these protease genes could be due to a reprogramming of digestive arsenal
304 in the Hessian fly larval gut in response to elevated levels of protease inhibitors produced in
305 resistant plants. The impact of host genotypes on the expression levels of these protease genes
306 further indicates that these genes encode digestive proteases.

307 Interestingly, our data also indicate that the identified Hessian fly protease genes are
308 highly diversified. The majority of protease genes share sequence identity of less than 35%
309 (Tables S2, S3, S4). With the exception of two identical cysteine proteases, MDP-99 and MDP-
310 100, the highest amino acid sequence identity shared by two putative trypsins is 61.3% (MDP-3
311 and MDP-5); and the highest amino acid sequence identity shared by two identified
312 chymotrypsins is 81.5% (MDP-15 and MDP-16). This observation is far different from an
313 earlier analysis of protease transcripts from the Hessian fly larval gut (Zhu et al., 2005), which
314 identified many transcripts encoding trypsins and chymotrypsins that share over 90% amino acid
315 identity. Large numbers of transcripts encoding very similar, but distinct proteins has also been
316 found in other insects (Zhu et al., 2003; Coates et al., 2006). These observations suggest that
317 similar, but distinct trypsins and chymotrypsins revealed from transcriptomic analyses are
318 encoded by different alleles instead of similar genes. An insect population that keeps a large
319 number of different protease gene alleles must gain certain adaptive advantages. For

phytophagous insects, the adaptive advantage of maintaining allelic diversity in digestive protease genes is most likely towards adaptation to possible changes in protease inhibitors in host plants.

Our data suggested that two cysteine protease genes, *MDP95* and *MDP104*, may also play a role in digestion under certain conditions. The expression of these two genes was affected by host genotypes. Specifically, they were upregulated in larvae feeding in resistant plants. In addition, transcripts of these two genes were identified in cDNA libraries from gut tissue (Zhang et al., 2010). However, these two genes were also expressed in other tissues and in non-feeding stages of the insect (Figures 3 and 4), indicating other functions. So far, cysteine proteinases have been found in the midgut lumen (acting as digestive enzymes) only in hemipterans and coleopterans. Therefore, the digestive function of cysteine proteases in the gut of gall midges remains to be determined experimentally.

Ever since the discovery of protease inhibitor induction following insect attack by Ryan (1973), protease inhibitors have become targets of bioengineering for arthropod pest control (Lawrence and Koundal, 2002; Schlüter et al., 2010). Yet so far, no commercial cultivar with an engineered protease inhibitor has achieved effective, long-lasting defense to insect pests in the field. The difficulty in doing so comes from the multiplicity of diverse protease genes and their dynamic changes in response to elevated inhibitors in host plants. A better understanding of genome-wide composition of digestive proteases and the regulation of the expression of the encoding genes may help to optimize the strategy for engineering protease inhibitors. The availability of whole genome sequences of an increasing number of insect species provides us an opportunity to study the insect gut digestive enzymes and their interactions with host inhibitors globally, comparatively, and comprehensively. The identification of serine-protease and

343 cysteine-protease genes in the Hessian fly genome, and the determination of their expression
344 profiles in different developmental stages and different fly tissues of the insect provide the
345 foundation for further characterization of Hessian fly digestive proteases and their interactions
346 with host plants.

Acknowledgment

This contribution No. for this paper is 13-052-J from the Kansas Agricultural Experiment Station. Hessian fly voucher specimens (No. 150) are located in the KSU Museum of Entomological and Prairie Arthropod Research, Kansas State University, Manhattan, Kansas. The authors want to thank Drs. Kun Yan Zhu and Michael Smith for reviewing an earlier version of the manuscript. The research was partially supported by a grant from the U.S. Department of Agriculture (USDA NIFA 2010-03741) and a grant from the National Natural Sciences Foundation of China (grant no.30800105).

References

Ahn, J.E., Salzman, R.A., Braunagel, S.C., Koiwa, H., Zhu-Salzman, K., 2004. Functional roles of specific bruchid protease isoforms in adaptation to a soybean protease inhibitor. Insect Mol. Biol. 13, 649-657.

Barrett, A.J., Rawlings, N.D., Woessner, J.F., 1998. Handbook of Proteolytic Enzymes. Academic Press. New York.

Bjellqvist, B.,Hughes, G.J., Pasquali, Ch., Paquet, N., Ravier, F., Sanchez, J.-Ch., Frutiger, S. & Hochstrasser, D.F.(1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. Electrophoresis 14, 1023-1031.

Bown, D.P., Wilkinson, H.S. & Gatehouse, J.A., 1997. Differentially regulated inhibitor-sensitive and insensitive protease genes from the phytophagous insect pest, *Helicoverpa armigera*, are members of complex multigene families. Insect Biochem. Mol. Biol. 27, 625-638.

Brioschi, D., Nadalini, L.D., Bengtson, M.H., Sogayar, M.C., Moura, D.S., Silva-Filho, M.C. 2007. General up regulation of *Spodoptera frugiperda* trypsins and chymotrypsins allow its adaptation to soybean proteinase inhibitor. Insect Biochem. Mol. Biol. 37, 1283-1290.

Buntin, G.D., 1999. Hessian fly (Diptera: Cecidomyiidae) injury and loss of winter wheat grain yield and quality. J. Econ. Entomol. 92, 1190-1197.

Burgess, E.P.J., Main, C.A., Stevens, P.S., Christeller, J.T., Gatehouse, A.M.R., Laing, W.A., 1994. Effects of protease inhibitor concentration and combinations on the survival, growth and gut enzyme activities of the blackfield cricket, *Teleogryllus commodus*. J. Insect Physiol. 40, 803-811.

379 Chen, M.S., 2008. Inducible direct plant defense against insect herbivores. Insect Sci.15, 101-
380 114.
381 Chen, M.S., Echegaray, E., Whitworth, R.J., Wang, H., Sloderbeck, P.E., Knutson, A., Giles,
382 K.L. 2009. Virulence analysis of Hessian fly (*Mayetiola destructor*) populations from Texas,
383 Oklahoma, and Kansas. J. Econ. Entomol. 102, 774-780.
384 Coates, B.S., Hellmich, R.L., Lewis, L.C., 2006. Sequence variation in trypsin- and
385 chymotrypsin-like cDNAs from the midgut of *Ostrinia nubilalis*: methods for allelic
386 differentiation of candidate *Bacillus thuringiensis* resistance genes. Insect Mol. Biol. 15, 13-
387 24.
388 Dufour, E., 1988. Sequence homologies, hydrophobic profiles and secondary structures of
389 cathepsins B, H and L: comparison with papain and actinidin. Biochimie 70, 1335-1342.
390 Eberhard, S.H., Hrassnigg, N., Crailsheim, K., Krenn, H.W., 2007. Evidence of protease in the
391 saliva of the butterfly *Heliconius melpomene* (L.) (Nymphalidae, Lepidoptera). J. Insect
392 Physiol. 53, 126-131.
393 Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap.
394 Evolution 39, 783-791.
395 Gattiker A., Gasteiger E., Bairoch A., 2002. ScanProsite: a reference implementation of a
396 PROSITE scanning tool. Appl. Bioinformatics 1, 107-108.
397 (http://prosite.expasy.org/scanprosite/)
398 Habib, H., Fazili K.M., 2007. Plant protease inhibitors: a defense strategy in plants. Biotech.
399 Mol. Biol. Rev. 2, 68-85.
400 Hall, T.A., 1999. BioEdit: a user-friently biological sequence alignment editor and analysis
401 program for Windows 95/98/NT. Nucl. Acid. Symp. Series 41, 95-98.

Hatchett, J.H., Starks, K.J., Webster, J.A., 1987. Insect and mite pests of wheat. In: Wheat and Wheat improvement. Agronomy Monograph No. 13, 625–675.

Hedstrom, L., Szilagyi, L., Rutter, W.J., 1992. Converting trypsin to chymotrypsin: the role of surface loops. Science 255, 1249–1253.

Holt C., Yandell M., 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinform. 12, 491.

Jongsma, M.A., Beekwilder, J., 2011. Co-evolution of insect proteases and plant protease inhibitors. Curr. Protein Pept. Sci. 12, 437-447.

Jongsma M.A., Bakker P.L., Peters J., Bosch D., and Stiekema W.J.,1995. Adaptation of Spodoptera exigua larvae to plant proteinase inhibitors by induction of gut proteinase activity insensitive to inhibition, Proc. Natl. Acad. Sci. USA 92, 8041-8045.

Kwon, T.H., Kim, M.S., Choi, H.W., Joo, C.H., Cho, M.Y. and Lee,B.L., 2000. A masquerade-like serine proteinase homologue is necessary for phenoloxidase activity in the coleopteran insect, *Holotrichia diomphalia* larvae. Eur. J. Biochem. 267, 6188–6196.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G. 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21), 2947-2948.

Lawrence, P.K., Koundal, K.P., 2002. Plant protease inhibitors in control of phytophagous insects. Electronic J. Biotech. 5, 93-103.

Liu, X.L., Bai, J., Huang, L., Zhu, L., Liu, X., Weng, N., Reese, J.C., Harris, M., Stuart, J.J., Chen, M.S., 2007. Gene expression of different wheat genotypes during attack by virulent and avirulent Hessian fly (*Mayetiola destructor*) larvae. J. Chem. Ecol. 33, 2171-2194.

Lopes, A.R., Juliano, M.A., Marana, S.R., Juliano, L., Terra, W.R., 2006. Substrate specificity of insect trypsins and the role of their subsites in catalysis. Insect Biochem. Mol. Biol. 36, 130-140.

Matsumoto, I., Watanabe, H., Abe, K., Arai, S., Emori, Y., 1995. A putative digestive cysteine proteinase from *Drosophila melanogaster* is predominantly expressed in the embryonic and larval midgut. Eur. J. Biochem. 227, 582−587.

Mazumdar-Leighton, S. & Broadway, R.M., 2001. Identification of six chymotrypsin cDNAs from larval midguts of *Helicoverpa zea* and *Agrotis ipsilon* feeding on the soybean (Kunitz) trypsin inhibitor. Insect Biochem. Mol. Biol. 31, 633-644.

Miles, P.W., 1999. Aphid saliva. Biol. Rev. 74, 41-85.

Moura, D.S., Ryan, C.A., 2001. Wound-inducible proteinase inhibitors in pepper. Differential regulation upon wounding, systemin and methyl jasmonate. Plant Physiol. 126 (1), 289-298.

Murdock, L.L., Shade, R.E., Pomeroy, M.A., 1988. Effects of E-64, acysteine proteinase-inhibitor, on cowpea weevil growth, development, and fecundity. Environ. Entomol. 17, 467-469.

Neurath, H., 1999. Proteolytic enzymes, past and future. Proc. Natl. Acad. Sci. USA. 96, 10962-10963.

Neurath, H., Walsh, K.A., 1976. Role of proteolytic enzymes in biological regulation (a review). Proc Natl Acad Sci USA.73, 3825-3832.

Patterson. F.L., Mass. F.B. III, Foster. J.E., Ratcliffe. R.H., Cambron. S., Safranski. G., Taylor. P.L., Ohm. H.W., 1994. Registration of eight Hessian fly resistant common winter wheat germplasm lines (Carol, Erin, Flynn, Iris, Joy, Karen, Lola, and Molly). Crop Sci. 34, 315-316.

Pauly, P.J., 2002. Fighting the Hessian fly. Environ. Hist. 7, 385-507.

448 Pechan, T., Cohen, A., Williams, W. P., Luthe, D. S., 2002. Insect feeding mobilizes a unique
449 plant defense protease that disrupts the peritrophic matrix of caterpillars. Proc. Natl. Acad.
450 Sci. USA 99, 13319-13323.
451 Petersen, T.N., Brunak, S., von Heijne, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal
452 peptides from transmembrane regions. Nature Methods, 8, 785-786.
453 http://www.cbs.dtu.dk/services/SignalP/
454 Ross, J., Jiang, H., Kanost, M.R., Wang, Y., 2003. Serine proteases and their homologs in the
455 *Drosophila melanogaster* genome: an initial analysis of sequence conservation and
456 phylogenetic relationships. Gene 304, 117-131.
457 Ryan, C.A., 1973. Proteolytic enzymes and their inhibitors in plants. Annu. Rev. Plant Physiol.
458 24, 173-96
459 Sato, P.M., Lopes, A.R., Juliano, L., Juliano, M.A., Terra, W.R., 2008. Subsite substrate
460 specificity of midgut insect chymotrypsins. Insect Biochem. Mol. Biol. 38, 628-633.
461 Schlüter, U., Benchabane, M., Munger, A., Kiggundu, A., Vorster, J., Goulet, M.-C., Cloutier,
462 C., Michaud, D., 2010. Recombinant protease inhibitors for herbivore pest control: a
463 multitrophic perspective. J. Exp. Bot. 61, 4169-4183.
464 Shukle, R.H., Murdock, L.L., Gallun, R.L., 1985. Identification and partial characterization of a
465 major gut proteinase from larvae of the Hessian fly, *Mayetiola destructor* (Say) (Deptera:
466 Cecidomyiidae).Insect Biochem. 15, 93-101.
467 Stuart J.J., Chen M.S., Shukle R., Harris M.O., 2012. Gall Midges (Hessian Flies) as Plant
468 Pathogens. Annu. Rev. Phytopathol. 50, 17.1–17.19.
469 Tamura K., Nei M., Kumar S., 2004. Prospects for inferring very large phylogenies by using the
470 neighbor-joining method. Proc. Natl. Acad. Sci. 101, 11030–11035.

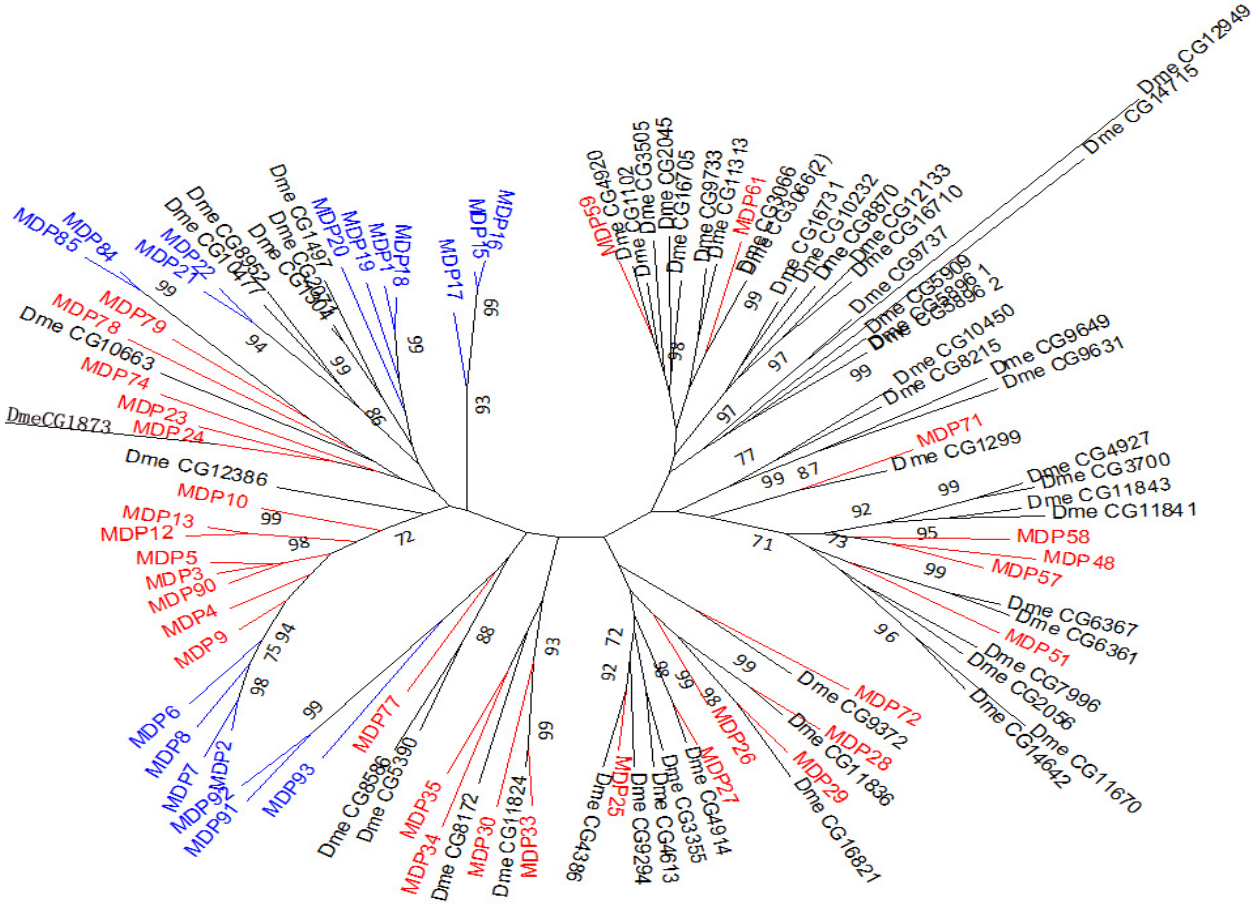
471 Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5:
472 Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary
473 Distance, and Maximum Parsimony Methods. Mol. Biol. Evol. 28, 2731-2739.
474 Terra, W.R., Ferreira, C., 1994. Insect digestive enzymes: properties, compartmentalization and
475 function. Comp. Biochem. Physiol. 109B,1–62.
476 Turk, V., Turk, B., Turk, D., 2001. Lysosomal cysteine proteases: facts and opportunities.
477 EMBO J. 20, 4629e4633.
478 Wang, S., Magoulas, C., Hickey, D.A., 1993. Isolation and characterization of a full-length
479 trypsin-encoding cDNA clone from the lepidopteran insect Choristoneura fumiferana. Gene
480 136, 375–376.
481 Wright, M.K., Brandt, S.L., Coudron, T.A., Wagner, R.M., Habibi, J., Backus, E.A., Huesing,
482 J.E., 2006. Characterization of digestive proteolytic activity in *Lygus hesperus* Knight
483 (Hemiptera : Miridae). J. Insect Physiol. 52, 717-728.
484 Wu J.X., Liu X.M., Zhang S.Z., Y.C. Zhu, Whitworth R.J., Chen M.S., 2008. Differential
485 responses of wheat inhibitor-like genes to Hessian fly, *Mayetiola destructor*, attacks during
486 compatible and incompatible interactions. J. Chem. Ecol. 34, 1005-1012.
487 Yusim K.S.J., Honeyborne I., Calef C., Goulder P.J., Korber B.T., 2004. Enhanced motif scan: A
488 tool to scan for HLA anchor residues in proteins. Los Alamos: Theoretical Biology and
489 Biophysics Group, Los Alamos National Laboratory. pp. 25–36. Publication number LA-
490 UR 04-8162. pp. (http://myhits.isb-sib.ch/cgi-bin/motif_scan)
491 Zhang, S.Z., Shukle, R., Mittapalli, O., Zhu, Y.C., Reese, J.C., Wang, H.Y., Hua, B.Z., Chen
492 ,M.S., 2010. The Gut Transcriptome of a Gall Midge, Mayetiola destructor. J. Insect
493 Physiol. 56, 1198-1206.

494 Zhu, Y.C., Liu, X., Maddur, M., Oppert, B., Chen, M.S., 2005. Cloning and characterization of
495 chymotrypsin- and trypsinlike cDNAs from the gut of the Hessian fly [*Mayetiola destructor*
496 (Say)]. Insect Biochem. Mol. Biol. 35, 23-32.
497 Zhu Y.C., Zeng F.R., Oppert B., 2003. Molecular cloning of trypsin-like cDNAs and
498 comparison of proteinase activities in the salivary glands and gut of the tarnished plant bug
499 *Lygus lineolaris* (Heteroptera: Miridae). Insect Biochem. Mol. Biol. 33, 889-899.
500

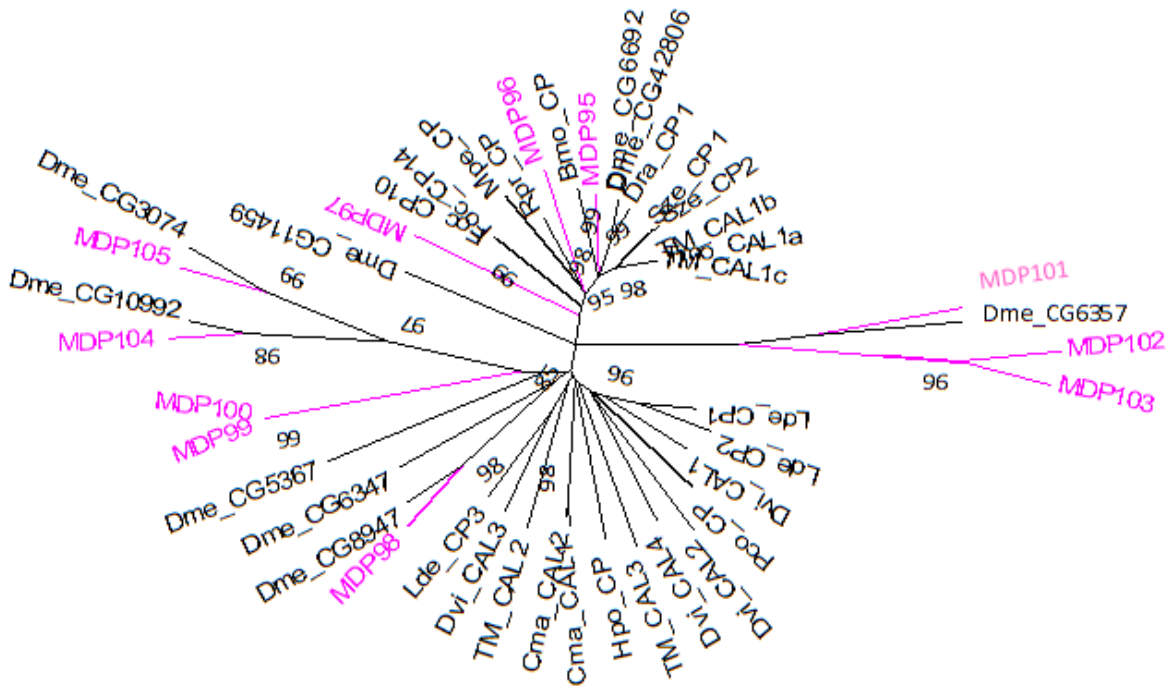
Figure Legends:

Figure 1. Phylogenetic analyses. A. A phylogenetic tree of identified Hessian fly trypsins and chymotrypsins along with serine proteinases from *Drosophila melanogaster*. Names in red color are trypsins, names in blue are chymotrypsins, and names in black are serine proteinases from *Drosophila*. **B**. A phylogenetic tree of identified putative Hessian fly cysteine-protease proteases along with known cysteine proteases from other insects. Hessian fly cysteine proteases are in pink color and the cysteine proteases from other insects are in black color. Dme, Tmo, Dro, Sze, Bmo, Mpe, Rpr, Foc, Hpo, Dvi, Lde, Pco, and Cma, represent cysteine proteases from *Drosophila melanogaster, Tenebrio molitor, Delia radicum, Sitophilus zeamais, Bombyx mori, Myzus persicae, Rhodnius prolixus, Frankliniella occidentalis, Hypera postica, Diabrotica virgifera, Leptinotarsa decemlineata, Phaedon cochleariae,* and *Callosobruchus maculates*, respectively.

Figure 1A



517 Figure 1B



518

Figure 2. Amino acid alignments of different types of proteases. A. Alignment of 31 identified putative trypsins. Functionally important residues H, D, and S (active sites) are boxed. Cysteine residues corresponding to the sites of the predicted disulfide bridges are marked with arrows at the bottom. The trypsin specificity determinant residue is indicated with (♦) on the top of the alignment. The activation site (consensus K/R-IVGG at position 41) is conserved in most of the putative trypsins. Hyphens represent alignment gaps. Trypsins MDP3, MDP4, MDP5, MDP9, MDP10, MDP12, MDP23, MDP24, MDP79, and MDP90 are full length and each has a predicted signal peptide. **B.** An alignment of the 18 identified putative chymotrypsins. Functionally important residues H, D, and S (active sites) are boxed. Cysteine residues corresponding to the sites of the predicted disulfide bridges are marked with arrows at the bottom. Chymotrypsin specificity determinant residue is indicated with (♦) on the top of the alignment. The activation site (consensus K/R-IVGG at position 65) is partially conserved in most of the putative trypsins. Hyphens represent alignment gaps. MDP1, MDP2, MDP6, MDP7, MDP8, MDP15, MDP16, MDP17, MDP18, MDP19, MDP21, MDP22, MDP84, and MDP85 are full length and each has a typical secretion signal peptide. **C.** An alignment of the seven putative cysteine-proteases. Functionally important residues C, H, and N (active sites) are boxed. Hyphens represent alignment gaps. All putative cysteine proteases are full length and each has a typical secretion signal peptide.

Figure 2A

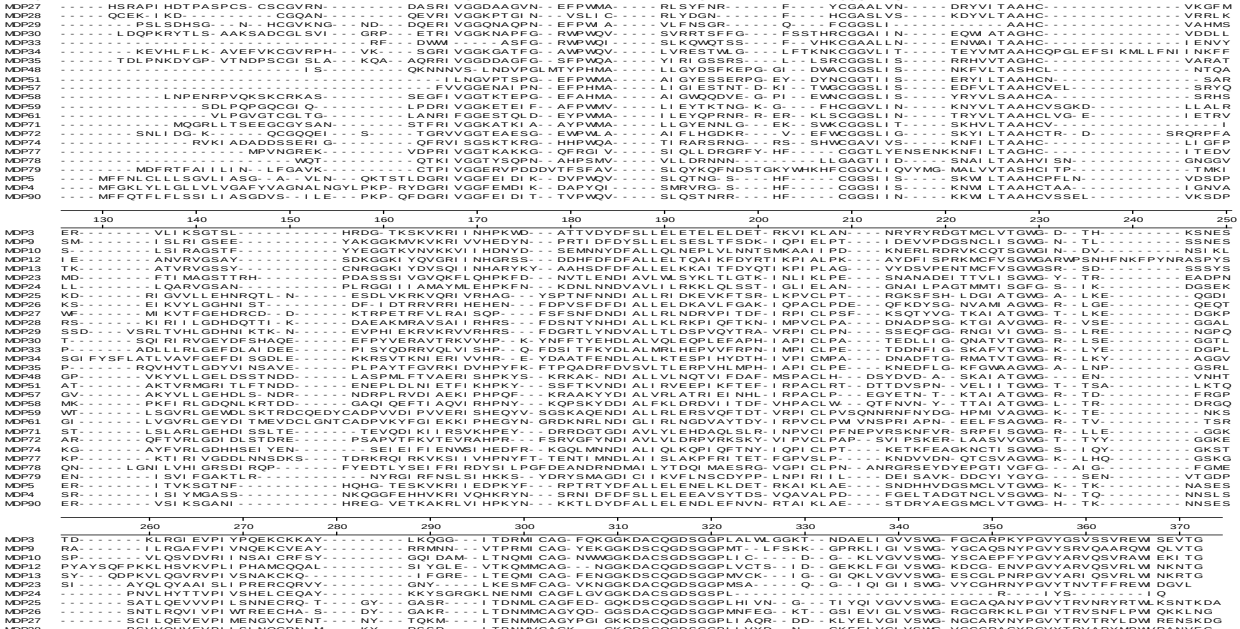
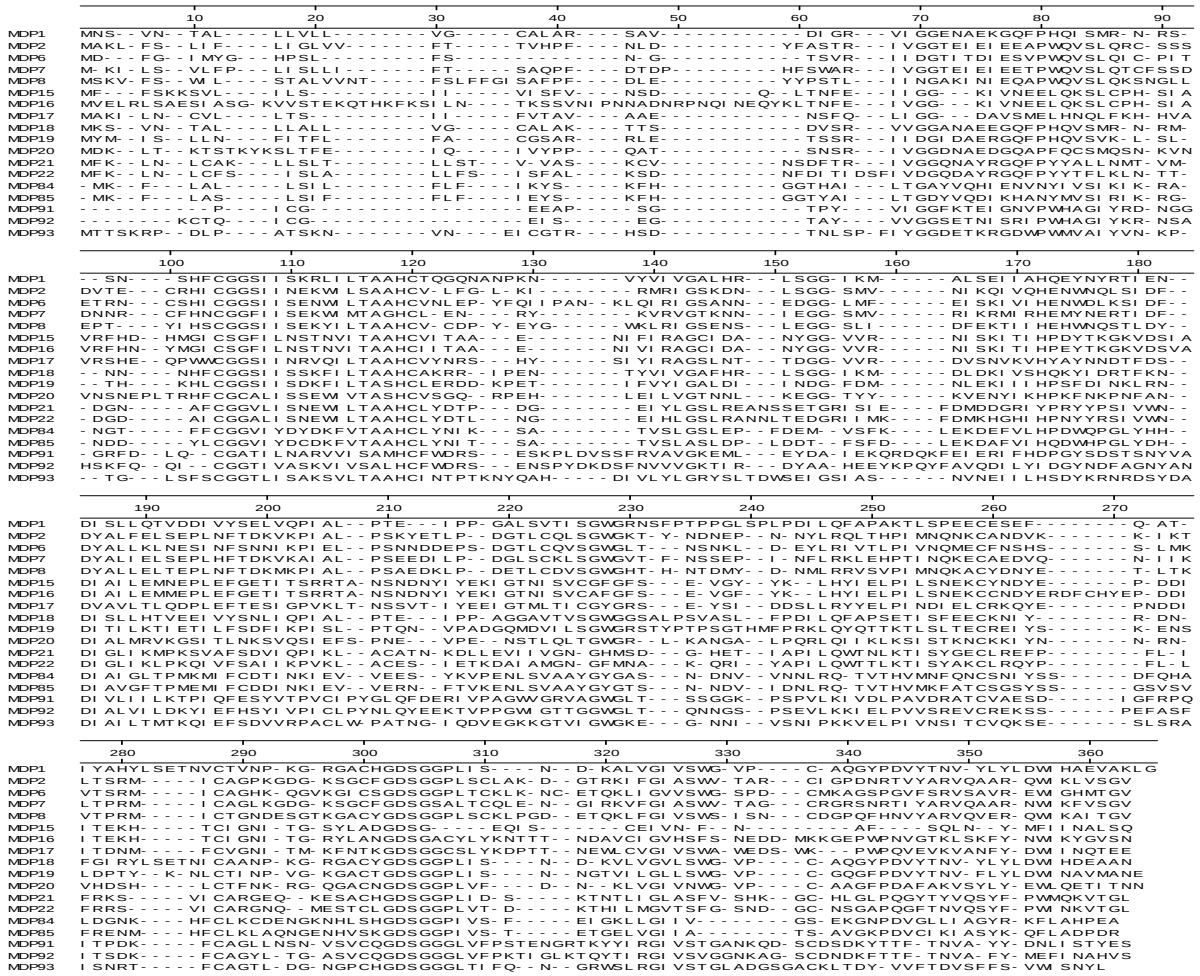


Figure 2B



539

540

541

Figure 3. RT-PCR analyses of various protease transcripts in insects at different developmental stages. **A**. Transcript abundance of putative trypsin genes. **B**. Transcript abundance of putative chymotrypsin genes. **C**. Transcript abundance of putative cysteine protease genes. 1L, 3L, 6L, 12L, P, and A on the top of images represent one-day, 3-day, 6-day, and 12-day old larvae, pupae, and adults, respectively. Gene names are given on the left of the images. Transcript abundance is shown in grey scale, with darker rectangles indicating higher transcript abundance.

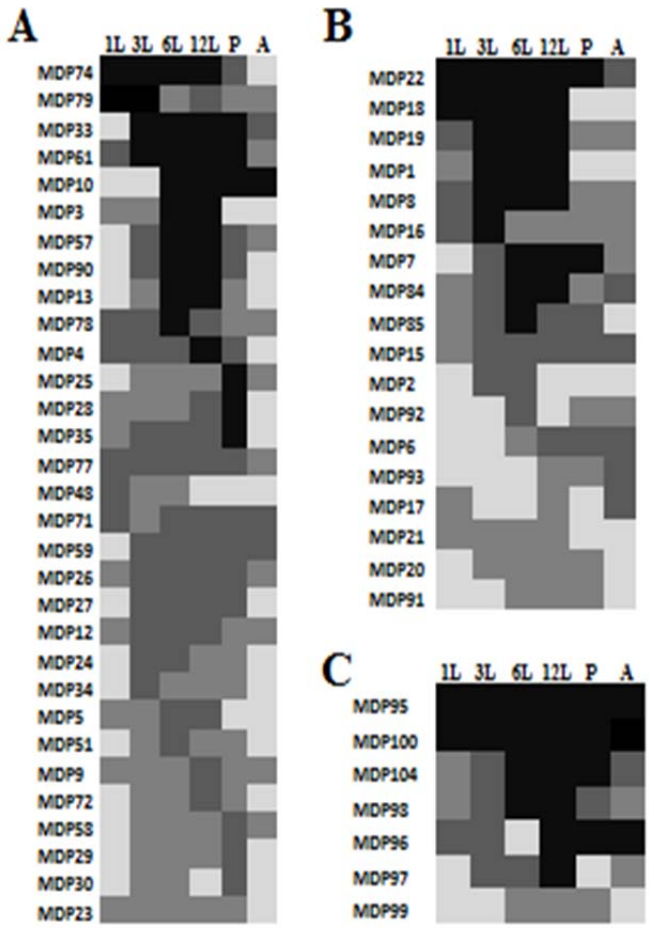


Figure 4. RT-PCR analyses of various protease transcripts in different tissues of three-day old larvae. **A**. Transcript abundance of putative trypsin genes. **B**. Transcript abundance of putative chymotrypsin genes. **C**. Transcript abundance of putative cysteine protease genes. MG, SG, FB, MT, and CC on the top of images represent mid-gut, salivary gland, fat body, Malphagian tubule, and the remaining carcass, respectively. Gene names are given on the left of the images. Transcript abundance is shown in grey scale, with darker rectangles indicating higher transcript abundance.

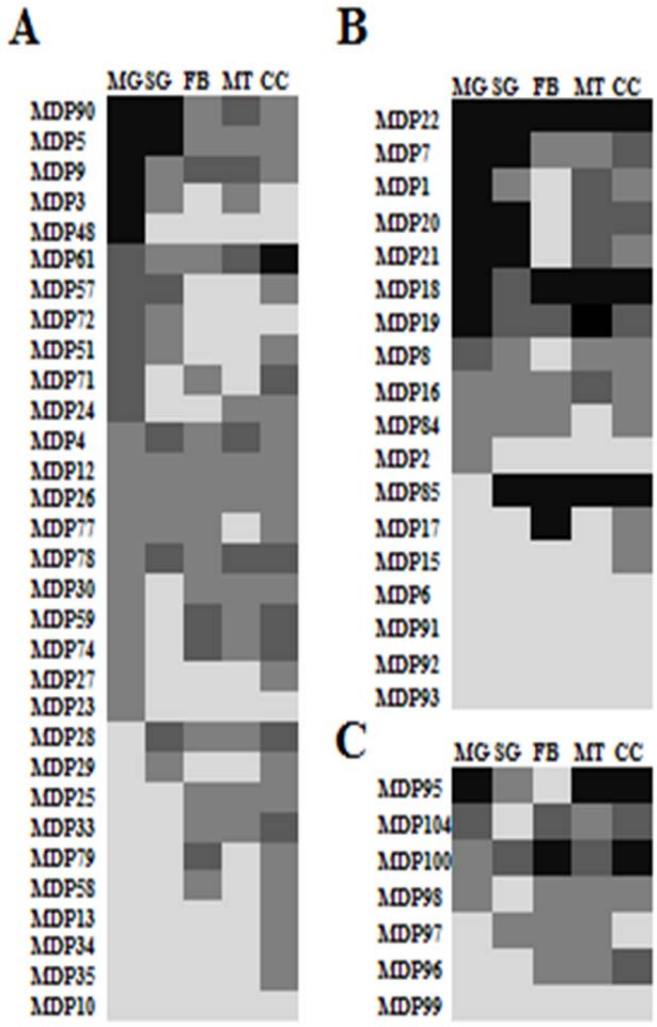
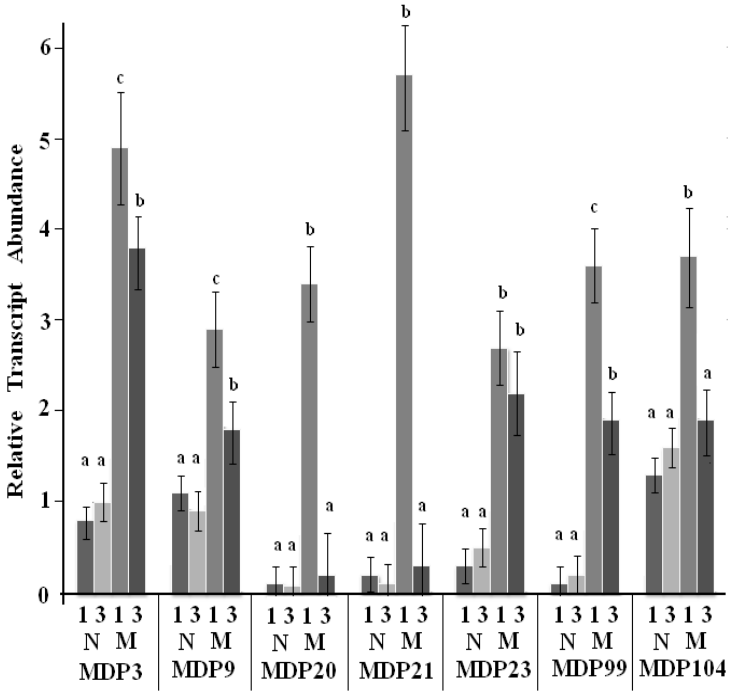
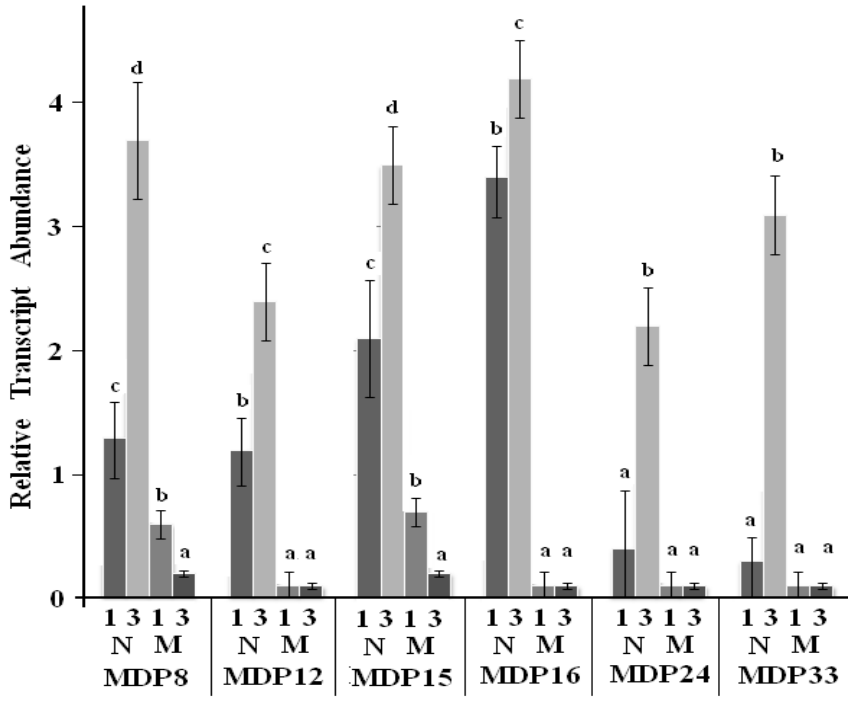


Figure 5. qPCR analyses of changes in transcript abundance of selected protease genes in Hessian fly larvae feeding in plants of different genotypes. A. Increased transcript abundance of protease genes in larvae feeding in fly resistant Molly (M) plants that contains the resistance gene *H13* in comparison with that in larvae feeding in susceptible isogenic Newton (N) plants (Patterson et al., 1994). B. Decreased transcript abundance of protease genes in larvae feeding in resistant Molly plants. The numbers 1 and 3 under the graphs represent samples from 1-day and 3-day larvae, respectively.

A



B



568 **Table 1. Characteristics of putative chymotrypsins, trypsins, and cysteine proteases.**

Gene	MW	pI	Length (aa)	First hit (Accession)	Scores	E-value	Organism name
Trypsins							
MDP3	27657	8.6	251	ABM26904	264	6E-85	*Lutzomyia longipalpis*
MDP5	27070	5.91	246	XP_001652944	249	2E-79	*Aedes aegypti*
MDP90	27872	8.92	251	ABM26904	246	4E-78	*L. longipalpis*
MDP9	26803	6.98	245	CAA80515	277	7E-90	*A. gambiae*
MDP4	29676	6.38	273	CAA79327	282	2E-91	*A. gambiae*
MDP10	26751	8.12	244	AAB66878	207	7E-63	*An. stephensi*
MDP12	31284	8.75	283	ACT21118	183	9E-53	*Drosophila mojavensis*
MDP13	27666	8.78	253	ACT21122	195	5E-58	*D. mojavensis*
MDP25	Truncate	?	?	NP_001166078	370	3E-124	*Nasonia vitripennis*
MDP26	25870	8.05	233	XP_001658407	305	9E-101	*Ae. aegypti*
MDP27	36858	8.4	331	XP_001658409	441	2E-153	*Ae. aegypti*
MDP28	35521	8.51	321	AEW46850	390	3E-132	*Chilo suppressalis*
MDP29	Truncate	?	?	XP_001650120	461	3E-155	*Ae. aegypti*
MDP34	105255	9.03	959	EFN77168	522	2E-168	*Harpegnathos saltator*
MDP35	Truncate	?	?	XP_001870897	607	0E+00	*Culex quinquefasciatus*
MDP59	38446	6.71	348	ADD18853	343	1E-112	*Glossina morsitans*
MDP61	43822	8.42	393	NP_649734	296	2E-93	*D. melanogaster*
MDP71	54246	8.29	491	XP_001654732	418	4E-138	*Ae. aegypti*
MDP72	73460	7.49	657	XP_001662898	586	0E+00	*Ae. aegypti*
MDP74	Truncate	?	?	NP_729453	652	0E+00	*D. melanogaster*
MDP77	32416	8.72	296	EFN86055	167	7E-46	*Ha. saltator*
MDP78	42673	5.64	384	AAT09848	138	7E-34	*Anthonomus grandis*
MDP79	28179	6.92	257	ACH56915	110	7E-26	*Simulium vittatum*
MDP23	26008	7.14	235	CAA80516	169	7E-48	*An. gambiae*
MDP24	21824	9.3	203	AAV84270	137	1E-36	*Culicoides sonorensis*
MDP30	71237	8.73	647	XP_001651579	451	1E-152	*Ae. aegypti*
MDP33	62722	4.99	562	XP_001857202	390	1E-129	*C. quinquefasciatus*
MDP51	64951	5.69	592	EHJ76340	221	3E-63	*Danaus plexippus*
MDP58	47573	6.59	423	AAD21841	284	1E-88	*Ctenocephalides felis*
MDP48	44640	5.72	400	XP_001655815	232	8E-69	*Ae. aegypti*
MDP57	40874	6.23	366	XP_001655816	291	5E-92	*Ae. aegypti*
Chymotrypsins							
MDP1	27569	5.58	256	ADR80135	256	2E-81	*Sitodiplosis mosellana*
MDP18	27138	6.17	251	EHJ72680	196	2E-58	*Danaus plexippus*
MDP19	27280	5.67	246	ADR80135	215	1E-65	*S. mosellana*
MDP20	29230	9.04	264	AAD17493	256	1E-81	*Anopheles darlingi*
MDP6	29239	5.47	267	AAA97479	204	4E-61	*A. stephensi*
MDP8	31322	4.98	280	ADR80134	188	2E-94	*S. mosellana*

569