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

Hang Chen¹,², Yu Cheng Zhu³, R. Jeff Whitworth¹, John C. Reese¹, and Ming-Shun Chen¹,⁴†

¹Department of Entomology, Kansas State University, Manhattan, KS 66506, USA
²Research Institute of Resource Insect, Chinese Academy of Forestry, Kunming 650224, China
³USDA-ARS-JWDSRC, PO Box 346/141 Exp Stn Rd, Stoneville, MS 38776, USA
⁴Hard 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 downregulated 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
Hessian fly larvae used in this research were derived from a field collection from Scott County, Kansas in 2005 (Chen et al., 2009). The insects have been maintained on seedlings of Hessian fly-susceptible wheat variety ‘Karl 92’ in greenhouse since then. The majority of flies were biotype GP although biotypes virulent to known R genes were also found in low frequencies (Chen et al., 2009).

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

The overall conservation of different trypsins, chymotrypsins, and cysteine proteases in an organism and across different species facilitated the identification of new proteases within a genome. To identify potential new Hessian fly protease genes, the draft Hessian fly genome sequence (http://agripestbase.org/hessianfly/) was searched using Blastx with known Hessian fly trypsins and chymotrypsins (Zhu et al., 2005), two partial sequences of Hessian fly cysteine proteases obtained from a gut transcriptome, and Drosophila serine- and cysteine-protease as queries (Matsumoto et al., 1995; Ross et al., 2003). From the resulting list of similar sequences, every 15th sequence was retrieved for another round of blasting. All the amino acid sequences encoded by predicted genes with significant sequence similarity (E-value < 0.01) from each of queries were retrieved. Repeated blasting was carried out until no new hit with E-value < 0.01 could be found. Introns were identified by comparing genomic sequence with a corresponding cDNA sequence or a putative transcript predicted by MAKER2 (Holt and Yandell, 2011). If no cDNA sequence or transcript was available for a particular gene, the intron/exon boundary was determined manually by a Blastx alignment between the Hessian fly gene sequence and the query protein sequence. Sequences were discarded if they could not be annotated. The identity 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-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 94°C; 60 s at 55°C; 120 s at 72°C. 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™ 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 chymotrysins, 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 chymotrysins 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
presence of specificity determinant residue D\textsuperscript{312} (Figure 2A), while chymotrypsins usually have G or S at corresponding position (residue 298, Figure 2B) (Hedstrom et al., 1992; Wang et al., 1993). Trypsins cleave the protein chain on the carboxyl side of arginine or lysine, whereas chymotrypsin cleaves on the carboxyl side of aromatic amino acids (Terra and Ferreira, 1994).

An alignment of the 18 putative chymotrypsins is given in Figure 2B. The catalytic triad H\textsuperscript{119}, D\textsuperscript{185}, and S\textsuperscript{304} are conserved in all putative chymotrypsins. However, the specificity determinant residue G/S\textsuperscript{298} is not found in some members at the exact position in the alignment. Specifically, the specificity determinant residue is Y in MDP-15 and MDP-16, F residue in MDP-17, and H residue in MPD-84 and MDP-85. These five proteins are either chymotrypsin homologs without enzymatic activity, or the specificity determinant residue is located in nearby positions, which were not correctly positioned in the alignment. Like trypsins, the putative chymotrypsins are also highly diversified. The two most closely related proteins, MDP-15 and MDP-16, share 81.5% sequence identity (Table S3). The two most diversified proteins, MDP-7 and MDP-85, share only 18.3% identity. The majority of the proteins share 22-32% sequence identity.

**3.4. Putative cysteine-protease genes**

Like trypsins and chymotrypsins, cysteine proteases are also involved in the physiological protein breakdown, but they are optimally active in the slightly acidic condition (Turk et al., 2001). Cysteine proteases also have different active site residues compared to those in trypsins and chymotrypsins. By using MotifScan and ScanProsite, the identity of several cysteine proteases were confirmed with the presence of active residue cysteine and two additional active site residues, histidine and asparagines (Dufour, 1988). An alignment among the seven putative cysteine-proteases is given in Figure 2C. The alignment at the N-terminal region exhibits very different sequences with little similarity. The C-terminal region, in
comparison, is relatively conserved. The three important residues C^{391}, H^{570}, and N^{591} at the
active site are located at the C-terminal region and are conserved in all members. Sequence
identity between these proteins ranged from 22 to 64% with the exception of MDP-99 and MDP-100, which are identical (Table S4). MDP-99 and MDP-100 are encoded by two tandem genes (AEGA01013770).

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

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

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

Since the Hessian fly larva is the only feeding stage and 3-day-old larvae are most active, the tissues of 3-day larvae including the gut, salivary glands, fat bodies, Malphigian tubules, and the remaining carcass were dissected for RT-PCR analyses. The results of larval tissue analysis also revealed different patterns of transcript abundance with the different protease genes in different tissues (Figure 4). For the trypsin genes, MDP48 was exclusively expressed in the larval gut. MDP9, MDP3, MDP72, MDP51, MDP24, and MDP23 exhibited higher transcript levels in gut tissues than in other tissues. MDP90, MDP5, and MDP57 exhibited higher 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 phylogenetic 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
defense (Bown et al., 1997). Wheat plants with an effective Hessian fly-resistance gene produce elevated levels of different types of protease inhibitors in response to Hessian fly attack, whereas the expression of inhibitor genes is suppressed in susceptible plants (Liu et al., 2007; Wu et al., 2008). Therefore, genes encoding digestive proteases may exhibit differential expression patterns in Hessian fly larvae feeding in susceptible plants from those feeding in resistant plants, potentially due to the change in concentrations of protease inhibitors in host plants. The change in expression levels of these protease genes could be due to a reprogramming of digestive arsenal in the Hessian fly larval gut in response to elevated levels of protease inhibitors produced in resistant plants. The impact of host genotypes on the expression levels of these protease genes further indicates that these genes encode digestive proteases.

Interestingly, our data also indicate that the identified Hessian fly protease genes are highly diversified. The majority of protease genes share sequence identity of less than 35% (Tables S2, S3, S4). With the exception of two identical cysteine proteases, MDP-99 and MDP-100, the highest amino acid sequence identity shared by two putative trypsins is 61.3% (MDP-3 and MDP-5); and the highest amino acid sequence identity shared by two identified chymotrypsins is 81.5% (MDP-15 and MDP-16). This observation is far different from an earlier analysis of protease transcripts from the Hessian fly larval gut (Zhu et al., 2005), which identified many transcripts encoding trypsins and chymotrypsins that share over 90% amino acid identity. Large numbers of transcripts encoding very similar, but distinct proteins has also been found in other insects (Zhu et al., 2003; Coates et al., 2006). These observations suggest that similar, but distinct trypsins and chymotrypsins revealed from transcriptomic analyses are encoded by different alleles instead of similar genes. An insect population that keeps a large 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
cysteine-protease genes in the Hessian fly genome, and the determination of their expression profiles in different developmental stages and different fly tissues of the insect provide the foundation for further characterization of Hessian fly digestive proteases and their interactions 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


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
Figure 1B
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 chymotrysins. 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 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
### Table 1. Characteristics of putative chymotrypsins, trypsins, and cysteine proteases.

<table>
<thead>
<tr>
<th>Gene</th>
<th>MW</th>
<th>pI</th>
<th>Length (aa)</th>
<th>First hit (Accession)</th>
<th>Scores</th>
<th>E-value</th>
<th>Organism name</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Trypsins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDP3</td>
<td>27657</td>
<td>8.6</td>
<td>251</td>
<td>ABM26904</td>
<td>264</td>
<td>6E-85</td>
<td>Lutzomyia longipalpis</td>
</tr>
<tr>
<td>MDP5</td>
<td>27070</td>
<td>5.91</td>
<td>246</td>
<td>XP_001652944</td>
<td>249</td>
<td>2E-79</td>
<td>Aedes aegypti</td>
</tr>
<tr>
<td>MDP9</td>
<td>27872</td>
<td>8.92</td>
<td>251</td>
<td>ABM26904</td>
<td>246</td>
<td>4E-78</td>
<td>L. longipalpis</td>
</tr>
<tr>
<td>MDP9</td>
<td>26803</td>
<td>6.98</td>
<td>245</td>
<td>CAA80515</td>
<td>277</td>
<td>7E-90</td>
<td>A. gambiae</td>
</tr>
<tr>
<td>MDP4</td>
<td>29676</td>
<td>6.38</td>
<td>273</td>
<td>CAA79327</td>
<td>282</td>
<td>2E-91</td>
<td>A. gambiae</td>
</tr>
<tr>
<td>MDP10</td>
<td>26751</td>
<td>8.12</td>
<td>244</td>
<td>AAB66878</td>
<td>207</td>
<td>7E-63</td>
<td>An. stephensi</td>
</tr>
<tr>
<td>MDP12</td>
<td>31284</td>
<td>6.75</td>
<td>283</td>
<td>ACT21118</td>
<td>183</td>
<td>9E-53</td>
<td>Drosophila mojavensis</td>
</tr>
<tr>
<td>MDP13</td>
<td>27666</td>
<td>8.78</td>
<td>253</td>
<td>ACT21122</td>
<td>195</td>
<td>5E-58</td>
<td>D. mojavensis</td>
</tr>
<tr>
<td>MDP25</td>
<td>Truncate</td>
<td>?</td>
<td>?</td>
<td>NP_001166078</td>
<td>370</td>
<td>3E-124</td>
<td>Nasonia vitripennis</td>
</tr>
<tr>
<td>MDP26</td>
<td>25870</td>
<td>8.05</td>
<td>233</td>
<td>XP_001658407</td>
<td>305</td>
<td>9E-101</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP27</td>
<td>36858</td>
<td>8.4</td>
<td>331</td>
<td>XP_001658409</td>
<td>441</td>
<td>2E-153</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP28</td>
<td>35521</td>
<td>8.51</td>
<td>321</td>
<td>AEW46850</td>
<td>390</td>
<td>3E-132</td>
<td>Chilo suppressalis</td>
</tr>
<tr>
<td>MDP29</td>
<td>Truncate</td>
<td>?</td>
<td>?</td>
<td>NP_001650120</td>
<td>461</td>
<td>3E-155</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP34</td>
<td>10525</td>
<td>9.03</td>
<td>959</td>
<td>EFN77168</td>
<td>522</td>
<td>2E-168</td>
<td>Harpegnathos saltator</td>
</tr>
<tr>
<td>MDP35</td>
<td>Truncate</td>
<td>?</td>
<td>?</td>
<td>XP_001870897</td>
<td>607</td>
<td>0E+00</td>
<td>Culex quinquefasciatus</td>
</tr>
<tr>
<td>MDP59</td>
<td>38446</td>
<td>6.71</td>
<td>348</td>
<td>ADD18853</td>
<td>343</td>
<td>1E-112</td>
<td>Glossina morsitans</td>
</tr>
<tr>
<td>MDP61</td>
<td>43822</td>
<td>8.42</td>
<td>393</td>
<td>NP_649734</td>
<td>296</td>
<td>2E-93</td>
<td>D. melanogaster</td>
</tr>
<tr>
<td>MDP71</td>
<td>54246</td>
<td>8.29</td>
<td>491</td>
<td>XP_001654732</td>
<td>418</td>
<td>4E-138</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP72</td>
<td>73460</td>
<td>7.49</td>
<td>657</td>
<td>XP_001662898</td>
<td>586</td>
<td>0E+00</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP74</td>
<td>Truncate</td>
<td>?</td>
<td>?</td>
<td>NP_729453</td>
<td>652</td>
<td>0E+00</td>
<td>D. melanogaster</td>
</tr>
<tr>
<td>MDP77</td>
<td>32416</td>
<td>8.72</td>
<td>296</td>
<td>EFN86055</td>
<td>167</td>
<td>7E-46</td>
<td>Ha. saltator</td>
</tr>
<tr>
<td>MDP78</td>
<td>42673</td>
<td>5.64</td>
<td>384</td>
<td>AAT09848</td>
<td>138</td>
<td>7E-34</td>
<td>Anthonomus grandis</td>
</tr>
<tr>
<td>MDP79</td>
<td>28179</td>
<td>6.92</td>
<td>257</td>
<td>ACH56915</td>
<td>110</td>
<td>7E-26</td>
<td>Simulium vittatum</td>
</tr>
<tr>
<td>MDP23</td>
<td>26008</td>
<td>7.14</td>
<td>235</td>
<td>CAA80516</td>
<td>169</td>
<td>7E-48</td>
<td>An. gambiae</td>
</tr>
<tr>
<td>MDP24</td>
<td>21824</td>
<td>9.3</td>
<td>203</td>
<td>AAU84270</td>
<td>137</td>
<td>1E-36</td>
<td>Culicoides sonorensis</td>
</tr>
<tr>
<td>MDP30</td>
<td>71237</td>
<td>8.73</td>
<td>647</td>
<td>XP_001651579</td>
<td>451</td>
<td>1E-152</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP33</td>
<td>62722</td>
<td>4.99</td>
<td>562</td>
<td>XP_001857202</td>
<td>390</td>
<td>1E-129</td>
<td>C. quinquefasciatus</td>
</tr>
<tr>
<td>MDP51</td>
<td>64951</td>
<td>5.69</td>
<td>592</td>
<td>E1H76340</td>
<td>221</td>
<td>3E-63</td>
<td>Danaus plexippus</td>
</tr>
<tr>
<td>MDP58</td>
<td>47573</td>
<td>6.59</td>
<td>423</td>
<td>AAD21841</td>
<td>284</td>
<td>1E-88</td>
<td>Ctenocephalides felis</td>
</tr>
<tr>
<td>MDP48</td>
<td>44640</td>
<td>5.72</td>
<td>400</td>
<td>XP_001655815</td>
<td>232</td>
<td>8E-69</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td>MDP57</td>
<td>40874</td>
<td>6.23</td>
<td>366</td>
<td>XP_001655816</td>
<td>291</td>
<td>5E-92</td>
<td>Ae. aegypti</td>
</tr>
<tr>
<td><strong>Chymotrypsins</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDP1</td>
<td>27569</td>
<td>5.58</td>
<td>256</td>
<td>ADR80135</td>
<td>256</td>
<td>2E-81</td>
<td>S. mosellana</td>
</tr>
<tr>
<td>MDP18</td>
<td>27138</td>
<td>6.17</td>
<td>251</td>
<td>E1H72680</td>
<td>196</td>
<td>2E-58</td>
<td>Danaus plexippus</td>
</tr>
<tr>
<td>MDP19</td>
<td>27280</td>
<td>5.67</td>
<td>246</td>
<td>ADR80135</td>
<td>215</td>
<td>1E-65</td>
<td>S. mosellana</td>
</tr>
<tr>
<td>MDP20</td>
<td>29230</td>
<td>9.04</td>
<td>264</td>
<td>AAD17493</td>
<td>256</td>
<td>1E-81</td>
<td>Anopheles darlingi</td>
</tr>
<tr>
<td>MDP6</td>
<td>29239</td>
<td>5.47</td>
<td>267</td>
<td>AAA97479</td>
<td>204</td>
<td>4E-61</td>
<td>A. stephensi</td>
</tr>
<tr>
<td>MDP8</td>
<td>31322</td>
<td>4.98</td>
<td>280</td>
<td>ADR80134</td>
<td>188</td>
<td>2E-94</td>
<td>S. mosellana</td>
</tr>
</tbody>
</table>