Characterization of mid-gut proteases in the social aphid, *Pemphigus obesinymphae*

Jessica Mezzanotte
Junior
BSCI 286 Research Report
Fall 2008

Introduction

The process of convergent evolution, or the occurrence of similar adaptive traits in species without a common ancestry (Wood *et al.*, 2003), has major implications for evolutionary biology. Not only can the identification of convergence help resolve inaccuracies in phylogeny reconstruction due to homoplasy (Wood *et al.*, 2003), it also represents one of the most convincing cases for adaptation in evolutionary biology. In a field where a key focus is the question of how adaptation occurs (Jost *et al.*, 2008), discovering the molecular changes that underlie convergent evolution could be crucial in understanding the process of adaptive evolution.

Although the genetic basis of adaptation remains poorly understood, the specific traits that separate species from each other and allow them to survive in alternative environments are often obvious. One such adaptation is advanced sociality in insects. While ants, bees, and termites are some of the most well-known social insect species, an insect that the Abbot lab studies, *Pemphigus obesinymphae* (Hemiptera: Aphidomorpha:Pemphigidae), is a North American species of social aphid that forms and inhabits structures on plants called galls (Costa 2006). A female aphid, the fundatrix, induces the formation of the gall on the host plant and produces genetically identical offspring that go through a series of instars before they reach maturity and leave the gall (Costa, 2006). Additionally, *P. obesinymphae* was the first North American aphid species known to produce soldiers (Costa 2006), which are first-instar nymphs that perform the altruistic function of defending the gall when attacked by predators. When a potential predator approaches the gall, soldier aphids swarm out and attack the predator by stabbing it with their enlarged mouthparts, called stylets. The soldiers continue their attack until the predator either leaves or dies, and many of the soldier aphids also die defending their gall.

This process, although present in several other species of social aphids, is curious. How do aphids, often significantly smaller than the enemies they attack, successfully defend their galls?

In an unusual twist, a 2004 study of a Japanese species of social aphid, *Tuberaphis styraci* (Hemiptera: Aphidomorpha: Hormaphididae), revealed that a homolog of a cathepsin B-like cysteine protease has acquired a novel, venomous function: aphid soldiers specifically express cathepsin B, or *catB*, during attack (Kutsukake *et al.* 2004). Kutsukake *et al.* (2004) found that soldiers express *catB* in their midgut, which is then secreted into the gut cavity. It is then expelled through the stylet and injected into potential predators. They concluded that *catB* is a major component of *T. styraci* venom, and it is used—not as a digestive enzyme as previously believed—but as a venomous secretion designed for the altruistic defense of clonemates.

CatB may thus underlie the adaptation and evolution of sociality in aphids. Because *P. obesinymphae* and *T. styraci* are two very divergent species, the use of *catB* as a venom in *P. obesinymphae* would represent a case of convergent evolution between the two. Studies of *P. obesinymphae* last semester investigated this question, and three paralogs of *catB* were actually found to be produced by *P. obesinymphae* (**Figure1**). Although it is unknown at this time whether or not these paralogs of *catB* are components of *P. obesinymphae* venom, the existence of these paralogs still gives insight into aphid evolution, for a 2007 study by Rispe *et al.* on the pea aphid, *Acyrthosiphon pisum* (Hemiptera: Aphidomorpha: Aphididae)—a highly diverse and divergent species from *P. obesinymphae*—discovered 28 copies of a *catB*-like gene in its genome. Rispe *et al.* (2007) also hypothesized that a large amplification of cathepsin B genes occurred in an ancestor of the Aphididae, so the fact that at least 3 *catB* paralogs are present in *P.*

obesinymphae presents the possibility that this gene family expansion occurred before the Aphididae-Pemphigidae split.

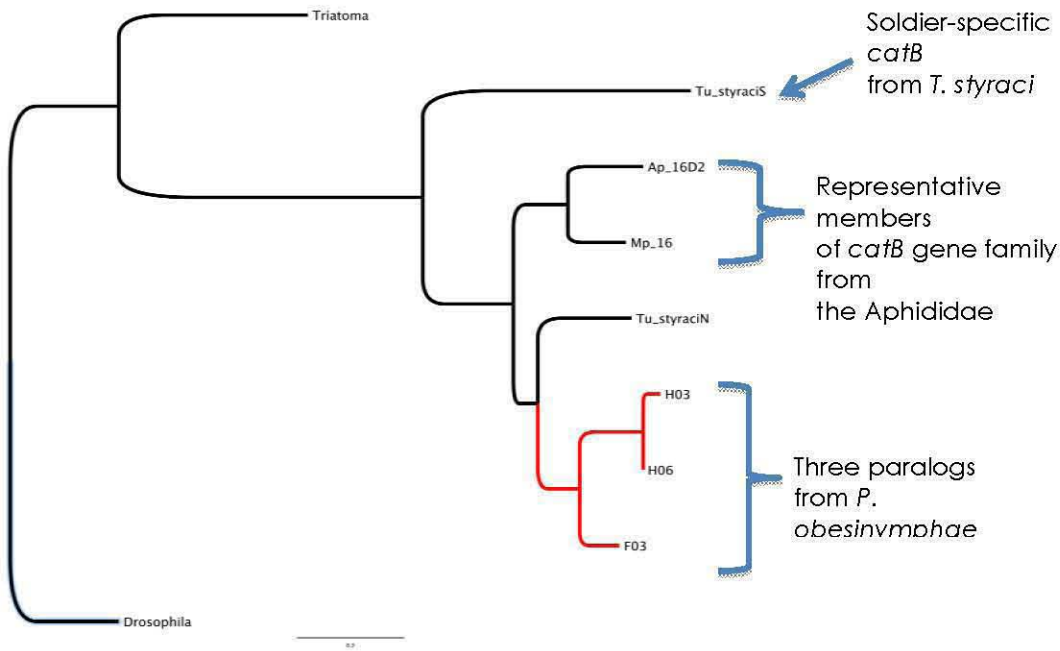


Figure 1: Phylogeny of *catB*. A phylogeny of different homologs of *catB*, including the 3 paralogs of *catB* obtained from *P. obesinymphae*, shown in red.

We are in the process of characterizing cathepsin B in *P. obesinymphae*. Meanwhile, experiments on the effects of *P. obesinymphae* on *Drosophila* larvae, which were used as mock predators of the gall, found that a majority of the attacked and deceased larvae exhibited extensive melanization around the sites where the aphid soldiers inserted their stylets (**Figure 2**). Typically, in *Drosophila*, melanization functions as an immune response to wounds or pathogens, and it is achieved through the conversion of prophenoloxidase to phenoloxidase, which activates a pathway that leads to melanin synthesis (Tang *et al.* 2006). The phenoloxidase activation is caused by two identified melanization proteins, MP1 and MP2, which are serine proteases (Tang *et al.* 2006). The extensive melanization and subsequent death of *Drosophila* larvae that were attacked by *P. obesinymphae* soldiers **(Figure 2)** suggested the possibility that

some type of serine protease was present in a *P. obesinymphae* secretion, rather than *catB*. If serine proteases are present, they could explain the extensive melanization patterns seen in the *Drosophila* larvae and could explain the larvae's eventual death. For example, *Drosophila* lines lacking serpin protease inhibitors required to regulate the phenoloxidase activation cascade suffer almost complete lethality when exposed to the over-expression of serine proteases (Tang *et al.* 2006). Possibly, aphids are secreting midgut-expressed serine proteases when they attack larval enemies, causing a rapid and systemic melanization response. The goal of this research was to complement ongoing studies of cysteine proteases by generating a cDNA library of serine proteases in *P. obesinymphae* in order to catalog and characterize serine paralogs. The experimental plan involved searching for serine paralogs that exhibit unusually high rates of amino acid substitutions. Venom genes typically exhibit high rates of evolution due to positive selection, rates which can be used as indicators of divergent serine paralogs in *P. obesinymphae* that are possibly co-opted for venomous functions.

Figure 2: The extensive melanization caused after a *P. obesinymphae* attack of a *Drosophila* larvae. The image on the left shows the body of a *Drosophila* larvae after being attacked by soldier aphids. Extensive melanization is present and appears as large, dark splotches on the larvae's body. A few aphids are even still present on the larvae. The image on the right is a close-up of the numerous melanin spots that occur in *Drosophila* after being attacked by *P. obesinymphae* soldiers.

Materials and Methods

Aphid Preparation. Fresh galls were cut open with a scalpel, and 11 galls were "induced" by placing a *Drosophila* larva, which acts as a mock-predator, inside of them. The aphids in these galls were allowed to react to and attack the *Drosophila* larvae for about 20 minutes; then, the *Drosophila* were removed and placed in 1.5mL Eppendorf tubes. The aphids were removed from their galls with the use of a small paintbrush, and they were divided into two groups: first instars and second through fifth instars. These groups were placed into 1.5mL Eppendorf tubes and, along with the attacked *Drosophila*, were stored at -80°C. An additional 12 galls were dissected in this manner, but they were left "uninduced," meaning that they were not presented with the mock-predator *Drosophila.*

RNA Isolation. Approximately 0.05g of aphids from three of the groups, the first instar induced group, the second through fifth instar induced group, and the first instar uninduced group, were placed into separate 1.5mL Eppendorf tubes. The three samples were labeled 1.In, 2.In, and 1.Un, respectively. A pestle was inserted into each tube, and the bottom of each tube was immediately submerged in liquid nitrogen for several seconds. After each tube was removed, the aphids were ground to powder using the pestle, and the process was repeated to create a fine powder. The ground aphid tissue was used in an RNA extraction and isolation reaction using Total RNA Isolation Reagent (TRIR) (Thermo Fisher Scientific, Inc.) according to the manufacturer's instructions. The RNA concentration was determined using a NanoDrop Spectrometer. The samples were purified using RNeasy Mini Kit (QIAGEN, Inc.) according to the manufacturer's instructions. The purified sample concentrations were determined using the NanoDrop Spectrometer.

cDNA Generation. The GeneRacer Kit SuperScript III RT Model (Invitrogen Life Technologies) was used to synthesize RACE-ready cDNA with known priming sites at the 5' and 3' ends from the purified RNA, according to the manufacturer's instructions. Seven µL of RNA from each of the three samples were used in this process, and the resulting cDNA from each sample was used in four RACE PCR reactions to amplify the 5' and 3' ends of the cDNA. The primers used in these reactions included the GeneRacer 5' Primer, which was paired in two separate reactions with two gene-specific primers (GSPs), SerR1 and SerR2, and the GeneRacer 3' Primer, which was paired in two separate reactions with the two GSPs, SerL1 and SerL2.

The GSPs were designed specifically for the RACE reaction using an alignment of serine proteases based off of serine protease-like ESTs from the *Acyrthosiphon pisum* EST database (http://www.aphidests.org/). Several alignments were created of different serine proteases found in *A. pisum* compared to known serine proteases in organisms such as *Drosophila*, *Anopheles gambiae,* and *Homo sapiens*, and two alignments showed enough homology to design primers using Primaclade (http://www.umsl.edu/services/kellogg/primaclade.html) and Primer3 (http://frodo.wi.mit.edu/). The GSPs had characteristics recommended by the GeneRacer manufacturers.

The four PCR amplifications were carried out at volumes of 25µL, containing 1X Invitrogen 10X buffer, 0.8mM of each deoxynucleoside triphosphate and $MgCl_2$, 0.5µM of each primer, 0.05 U/µL Hot STARTAQ polymerase, sterile PCR-grade water, and approximately 5 to 10ng of the cDNA. Reaction conditions were one cycle at 95°C for 2 minutes 30 seconds; 35 cycles of 95°C for 20 seconds, 56°C for 20 seconds, and 72°C for 45 seconds; followed by one cycle at 72°C for 2 minutes 30 seconds. 8µL of the PCR products were visualized by electrophoresis and ethidium bromide staining under UV light on 1% agarose gels.

Since the first PCR reactions produced smears instead of distinct bands, two nested PCR reactions were performed using the GeneRacer 5' Nested Primer and SerR1 and SerR2 primers to amplify the 5' cDNA ends. These reactions were performed using the same reagents and reaction conditions as previously listed, except for the different primers and the replacement of the 56°C for 20 seconds step with a higher temperature of 60°C for 20 seconds step, and a 72°C for 20 minutes extension cycle was added to the end of the PCR program. 8μL of each PCR reaction product were visualized in the same manner as before.

cDNA Cloning. The results of the PCR reaction using the GeneRacer 5' Nested Primer and SerR1 and SerR2 for sample 1.In, and the results of the PCR reaction using the GeneRacer 5' Nested Primer and SerR1 for sample 2.In were cloned via a pCR2.1-TOPO vector (Invitrogen Life Technologies) and TOPO TA cloning kit using Top 10 chemically competent cells, according to the manufacturer's instructions.

Positive colonies were screened for the insert using PCR amplification reactions at 10μL, using the reagent concentrations previously given. The colonies were screened for the insert in both directions using the primer pair of the GeneRacer 5' Nested Primer and the vector primer *IVT7* and the primer pair of either SerR1 or SerR2 and *IVT7*. The reaction conditions were 1 cycle at 95°C for 3 minutes; 30 cycles of 94°C for 30 seconds, 52°C for 30 seconds, and 72°C for 30 seconds; and a final cycle at 72°C for 20 minutes. The results of each reaction were visualized on 1% agarose gels as described. The colonies that screened positive for the insert were amplified in 20μL PCR reactions using the reagent concentrations and reaction conditions listed above and the Invitrogen vector primers *IVT7* and *M13RIVG*. 8μL of the products of each reaction were visualized on 1% agarose gels using the methods described, and the products of the successful reactions were submitted for DNA sequencing at The University of Arizona

sequencing facility with the Invitrogen vector primer *IVT7*. Sequencher Version 4.5 was used to group the sequences into different groups, or contigs, from which one or two representative sequences were used in BLASTn searches. BLASTn searches were used to compare the sequences to known serine-like proteins.

Results

The overall goal of this experiment was to generate a cDNA library of serine protease-like genes in *P obesinymphae*. The major components of the experiment involved extracting RNA from aphids, creating a cDNA library from the extracted RNA, using PCR to amplify the serine protease genes in the cDNA, and cloning the resulting genes into bacteria for subsequent sequencing in order to obtain the cDNA sequences for the different genes.

The sample concentrations after the RNA extraction were found to be 1164.1ng RNA/μL, 1292.6ng RNA/μL, and 1239.8ng RNA/μL for samples 1.In, 1.Un, and 2.In, respectively. After RNA purification, the concentrations were determined to be 368ng/μL, 71.5ng/μL, and 355.4ng/μL for samples 1.In, 2.In, and 1.Un, respectively.

The GSPs for RACE PCR were created from two alignments of serine proteases, one of which included members of the chromotrypsin family of serine proteases and the other of which included members of the prolyl oligopeptidase family of serine proteases. The primers were designed to have a high GC content in order to obtain the highest possible annealing temperature, were between 23-28 nucleotides in length, and had fewer than 3 G or C residues in the last five bases on the 3' ends in order to reduce the replication of DNA at non-target sites. The four primers, along with their annealing temperatures and other relevant data, are listed in **Table 1.**

Primer Name	Sequence (5' to 3')	Annealing Temperature (°C)	Length (bases)
SerR1	GTCAGAGAACGGGCCAGGACCATAG	69.92	25
SerR2	AGTTGTCCAGCAACGAGCTCCAGTG	69.94	25
SerL1	TTCCACGTATGCGGAACGCTATCTG	69.83	25
SerL2	ATGGGATGTCAACCACGATGTGGAG	70.13	25

Table 1: Gene-Specific Primers and Properties. The four GSPs designed for this experiment. All of the primers had a high GC content to optimize the annealing temperature and were at least 25 base pairs in length.

The gel for the first RACE PCR reaction produced indistinct smears, so a second PCR using a higher annealing temperature and the GeneRacer Nested 5' and 3' primers was performed. This gel showed bands around 500-600 base pairs in size for both the SerR1 and SerR2 primers, which were paired with the GeneRacer 5' Nested primer, for the 1.In and 2.In samples (**Figure 3**), although the band produced by SerR2 for 2.In was considered to be too much of a smear and was therefore not useful. None of the reactions using the SerL1 or SerL2 primers were considered successful.

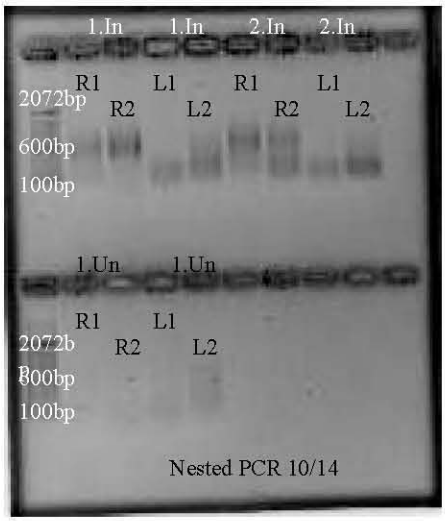


Figure 3: Nested PCR Amplification Results. The results of the nested PCR using the GeneRacer Nested 5' and 3' primers paired with their appropriate GSPs. Lane 1 in each row contains a 100bp ladder, and lanes 2-5 in row 1 contain the reaction for 1.In cDNA. Lanes 6-9 in row 1 contain the reaction for 2.In cDNA. Lanes 2-5 in row 2 contain the reaction for 1.Un cDNA, and this reaction was unsuccessful. Primers SerR1, SerR2, SerL1, and SerL2 are abbreviated as R1, R2, L1, and L2, respectively, in the figure, and R1 and R2 were both paired with the GeneRacer 5' Nested primer, while L1 and L2 were paired with the GeneRacer 3' Nested primer.

A total of 12 cloning reactions were performed, and 344 positive colonies were screened for the insert. The colonies that screened positive for the insert and that were used in high-volume PCR were screened a second time to ensure the success of the reaction (**Figure 4**). In the sample gel, lanes 38-40 are blank. Lanes 4, 5, 7, 8, 13, 14, 17-20, 22-24, 26-30, and 32-37 contained successful reactions, and their PCR products were submitted for sequencing. A total of 99 colonies were submitted for DNA sequencing, and 93 sequences were obtained.

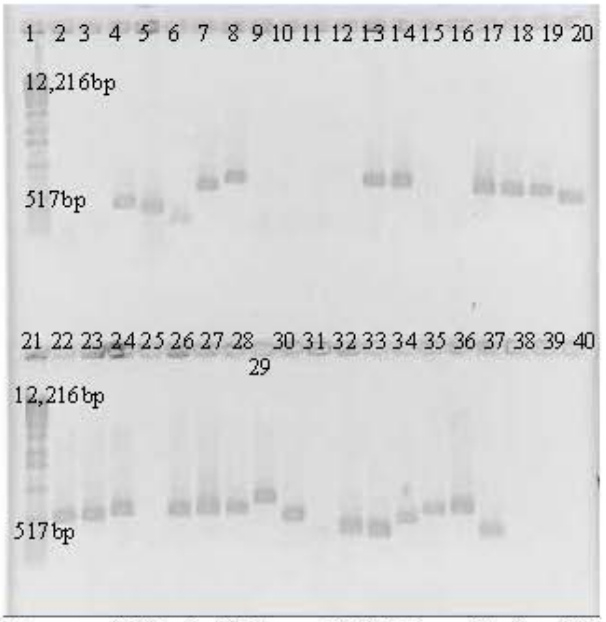


Figure 4: High-Volume PCR Results for 35 colonies. The results of a high-volume PCR for 35 of the colonies shown to have the insert. All lanes are numbered. Lanes 1 and 21 contain a 1kb ladder, and lanes 2-20 and 22-37 contain the PCR products of 35 colonies.

Once the sequences were sorted into contigs, a BLASTn search was performed on one or two sequences from each contig to see if they showed homology with any known serine proteases. **Table 2** shows the results of this search for representatives from each contig. Unfortunately, none of the sequences showed homology with any known serine proteases.

Sequence ID #	BLASTn Results	Sequence Type
B11_B11JESS33_T7_846084.ab1	NA	dirty seq
E08_E08ABB31_T7_846063.ab1	DQ005172	cytochrome c
C05_C05ABB4_T7_846037.ab1	NA	dirty seq
C08_C08ABB7_T7_846061.ab1	NA	dirty seq
A08_A08JESS16_T7_846059.ab1	NA	dirty seq
A09_A09JESS17_T7_846067.ab1	NA	dirty seq
A03_A03JESS5_T7_846019.ab1	NA	vector DNA only
F01_F01ABB36_T7_846008.ab1	XM_001942831	*A. pisum* predicted ribosomal protein
D03_D03ABB14_T7_846022.ab1	EF429249	cytochrome c
C03_C03ABB2_T7_846021.ab1	NA	dirty seq
D07_D07ABB18_T7_846054.ab1	NA	dirty seq
C11_C11ABB10_T7_846085.ab1	NA	dirty seq
F02_F02ABB37_T7_846016.ab1	NA	dirty seq
H12_H12ABB71_T7_846098.ab1	NA	dirty seq
G12_G12ABB59_T7_846097.ab1	NA	dirty seq
F03_F03ABB38_T7_846024.ab1	NA	dirty seq
F03_F03ABB38_T7_846024.ab1	NA	dirty seq
F09_F09ABB44_T7_846072.ab1	NA	dirty seq
G06_G06ABB53_T7_846049.ab1	NA	dirty seq

Table 2: Representative sequences from each contig and BLASTn results. The first column contains sequences from each contig that was created using Sequencher Version 4.5. A representative sequence was used from each contig, and a BLASTn search was performed. Two contigs contained sequences showing homology with cytochrome c, and one contig contained sequences showing homology with a predicted, commonly expressed ribosomal protein in *A. pisum*. The majority of the sequences in each contig, however, did not return BLASTn hits, as is indicated by "NA." These sequences, labeled "dirty seq" in the table, were found to contain unusable DNA.

Discussion

The goal of this study was to create a cDNA library of serine proteases in *P. obesinymphae* in order to characterize and catalog serine paralogs and to gain insight into the presence and potential uses of these genes in aphids, as was done with the previous discovery of 3 paralogs of cathepsin B in *P. obesinymphae*. Although no sequences have currently been found that show homology to known serine proteases, the presence of a large number of "dirty" sequences led to a further investigation that revealed that a large number of the sequences had

inserted in a "backwards" orientation, meaning that a long strand of thymine nucleotides (converted from the poly-A tail that was present on the initial extracted RNA) were present towards the beginning of a majority of the sequences. Since the presence of these poly-Ts would have interfered with sequencing due to destabilization of Taq polymerase, the sequences need to be screened in the opposite direction using the Invitrogen vector primer *M13RIVG*. This work is currently underway (due 12/09/2008), and the results should yield a larger number of useful sequences that will need to be analyzed by a BLASTn search.

There are numerous varieties of serine proteases; in the *MEROPS* Peptidase Database (http://merops.sanger.ac.uk/), serine proteases are divided into 13 clans, which are groups of homologous families of different types of serine proteases (Rawlings *et al.* 2008). The two sets of primers used in this study represented only two families of serine proteases, chromotrypsins and prolyl oligopeptidases, therefore it would be useful in future studies to design a larger number of primers for use in RACE PCR that would represent at the very least a minimum of one family from each of the 13 clans of serine proteases. There is also the possibility, in the case of sequences undergoing rapid amino acid evolution (which would indicate a change in selective pressures), that highly diverged sequences would be difficult to uncover unless specific primers using only highly conserved regions of the proteins were used. In this case, it would be useful to first extract different types of serine proteases from *P. obesinymphae* and to then create an alignment of them that could be used to generate highly specific primers capable of detecting conserved regions of divergent copies of serine proteases in *P. obesinymphae*.

While the initial discovery of *catB* in *P. obesinymphae* has numerous implications for the evolutionary origins of aphids, inspection of the amino acid alignment of the 3 paralogs and homologous sequences from GenBank does not indicate an unusual rate of amino acid evolution,

suggesting that it is unlikely that these paralogs have experienced directional selection in *P.* obesinymphae (**Figure 5**). Although studies on the presence of this protein in *P. obesinymphae* have yet to be completed, the presence of other types of proteases—especially if they show high rates of evolution—in *P. obesinymphae* could indicate that there is more than one mechanism of reaching the common product of venom evolution shared by *T. styraci*. If *P. obesinymphae* is found to express and use a venom through the use of proteases, this will represent an example of convergent evolution with a distantly related species, *T. styraci*, and will have numerous implications about the genetic basis of the evolution of venom use in aphids.

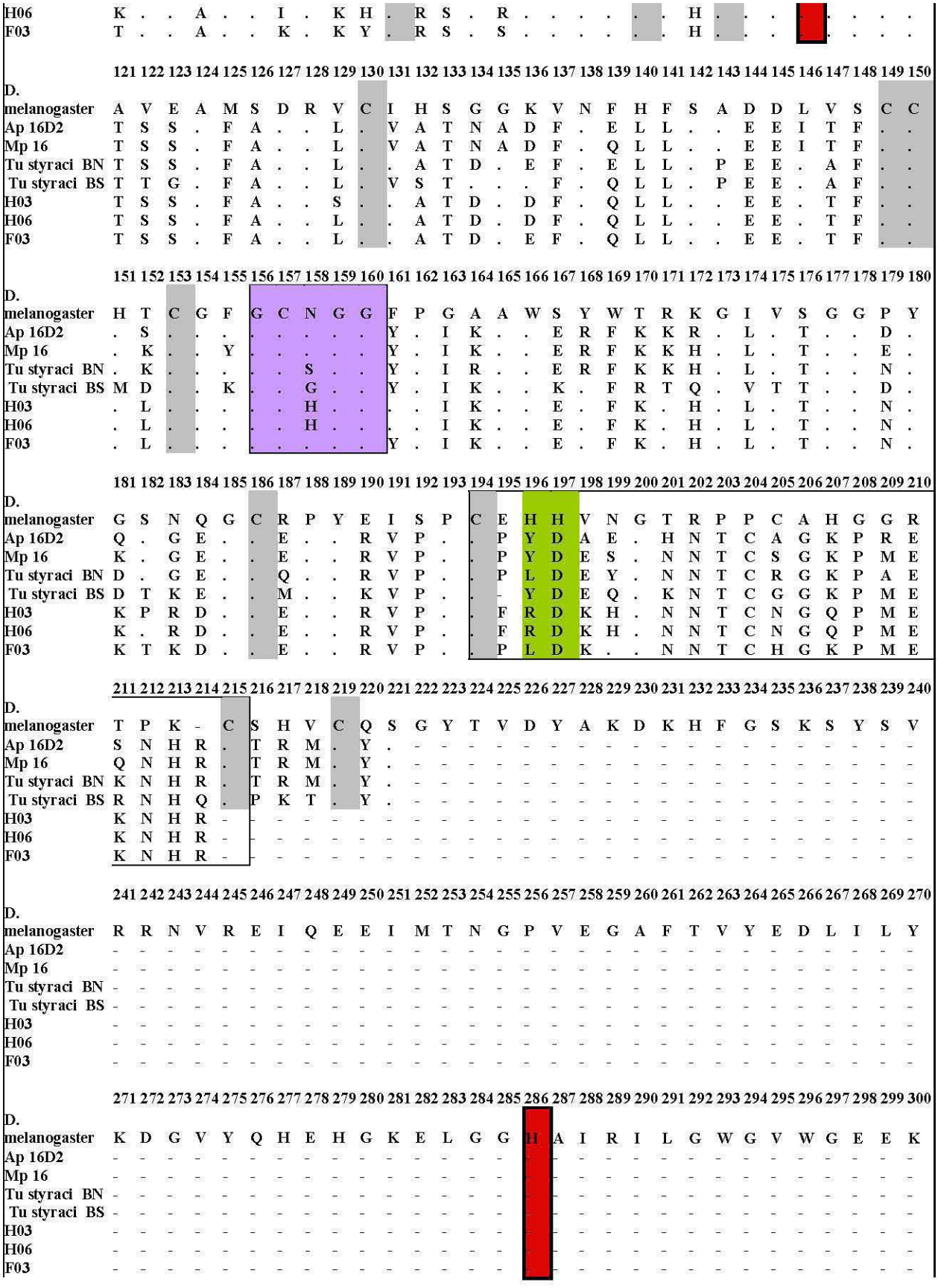
Figure 5: Annotated amino acid alignment of catB, including paralogs in *P. obesinymphae*

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
D. melanogaster	M	N	L	L	L	L	V	A	T	A	A	S	V	A	A	L	T	S	G	E	P	S	L	L	S	D	E	F	I	E
Ap 16D2	.	A	R	V	.	M	L	L	S	V	I	F	.	S	F	Y	L	T	E	Q	A	Y	F	.	Q	K	D	.	.	D
Mp 16	.	A	R	V	.	I	L	L	S	V	I	L	F	S	V	Y	M	T	E	Q	A	Y	F	.	E	K	D	Y	.	N
Tu styraci BN	.	I	R	.	V	V	L	L	S	V	V	L	F	S	V	Y	R	T	E	Q	A	Y	F	.	E	K	D	Y	.	N
Tu styraci BS	A	K	F	V	T	I	.	C	A	I	F	V	S	V	Y	.	A	E	P	T	L	Q	F	.	.	.	.	R	.	K
H03	V	D	R	F	.	I	L	L	L	.	L	L	F	S	V	Y	K	T	E	Q	A	Y	F	.	E	E	N	.	.	K
H06	.	D	R	F	.	I	L	L	L	.	L	L	F	S	V	Y	K	T	E	Q	A	Y	F	.	E	E	N	.	.	K
F03	.	A	R	F	F	I	L	L	S	V	I	L	F	S	V	Y	Q	T	E	Q	A	Y	F	.	E	K	S	.	.	D

	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
D. melanogaster	V	V	R	S	K	A	K	T	W	T	V	G	R	N	F	D	-	A	S	V	T	E	G	H	I	R	R	L	M	G
Ap 16D2	N	I	N	E	R	.	T	.	.	K	A	.	V	.	.	.	P	D	T	P	K	.	H	F	L	K	M	.	G	S
Mp 16	K	I	N	E	.	.	S	.	.	.	A	.	F	.	.	.	P	S	T	P	K	.	D	I	L	.	L	.	G	S
Tu styraci BN	Q	I	N	A	N	.	.	.	.	K	A	.	V	.	.	.	P	K	L	S	I	D	S	F	V	K	L	.	G	S
Tu styraci BS	Y	I	N	E	V	.	.	.	.	K	A	E	.	Y	.	-	P	.	N	-	.	S	E	E	Y	F	I	G	L	L
H03	Q	I	N	N	V	.	T	.	R	K	A	.	V	.	.	.	K	N	L	S	L	.	N	F	V	K	L	.	G	S
H06	Q	I	N	N	V	.	T	.	.	K	A	.	V	.	.	.	K	N	L	S	L	.	N	F	V	K	L	.	G	S
F03	Q	I	N	N	E	.	T	.	.	K	A	.	V	.	.	.	P	N	L	S	F	.	N	F	V	T	L	.	G	S

	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90
D. melanogaster	V	H	P	D	A	H	K	F	A	L	P	D	K	R	E	V	L	G	D	L	Y	V	N	S	V	D	E	L	P	E
Ap 16D2	K	G	V	Q	I	P	N	K	H	N	I	H	M	Y	K	T	H	D	A	A	.	D	.	L	F	G	R	I	.	R
Mp 16	K	G	V	Q	T	P	S	K	I	N	H	K	M	Y	K	S	E	D	K	E	.	D	.	L	F	G	R	I	.	K
Tu styraci BN	K	G	V	Q	.	A	.	Q	.	S	.	.	M	F	K	T	H	D	E	A	.	N	S	W	S	N	R	I	.	S
Tu styraci BS	G	S	R	G	Y	K	N	Y	T	N	E	V	E	I	K	K	Y	D	P	.	.	.	E	N	N	-	-	S	.	K
H03	K	G	V	E	S	A	.	A	.	S	.	.	.	F	K	T	F	D	E	V	.	S	Y	L	-	G	R	I	.	K
H06	K	G	V	E	S	A	.	A	.	S	.	.	.	F	K	T	F	D	E	V	.	S	Y	L	-	G	R	I	.	K
F03	R	G	V	Q	S	A	.	E	.	S	A	.	.	F	K	T	S	D	E	A	.	S	S	L	-	G	S	I	.	I

	91	92	93	94	95	96	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
D. melanogaster	E	F	D	S	R	K	Q	W	P	N	C	P	T	I	G	E	I	R	D	Q	G	S	C	G	S	C	W	A	F	G
Ap 16D2	H	.	.	A	.	R	K	.	R	R	.	H	.	.	.	A	V	.	.	.	.	N	.	.	.	.	.	.	M	A
Mp 16	K	.	.	A	.	.	K	.	R	H	.	T	.	.	.	A	V	.	.	.	.	N	.	.	.	.	.	.	I	A
Tu styraci BN	S	.	.	A	.	.	K	.	R	K	.	S	.	.	.	.	V	.	.	.	.	K	.	.	.	.	.	.	.	.
Tu styraci BS	Q	.	.	.	.	E	N	.	K	S	.	K	Q	.	.	H	.	.	.	.	.	N	.	.	.	.	.	S	.	S
H03	K	.	.	A	.	.	I	.	K	H	.	R	S	.	R	.	.	.	.	.	.	H	.	.	.	R	.	.	.	.


H06 K . . A . . I . K H . R S . R . . . . . H . . . . . .
F03 T . . A . . K . K Y . R S . S . . . . . H . . . . . .
121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150
D. melanogaster A V E A M S D R V C I H S G G K V N F H F S A D D L V S C C
Ap 16D2 T S S . F A . . L . V A T N A D F . E L L . . E E I T F . .
Mp 16 T S S . F A . . L . V A T N A D F . Q L L . . E E I T F . .
Tu styraci BN T S S . F A . . L . . A T D . E F . E L L . P E E . A F . .
Tu styraci BS T T G . F A . . L . V S T . . . F . Q L L . P E E . A F . .
H03 T S S . F A . . S . . A T D . D F . Q L L . . E E . T F . .
H06 T S S . F A . . L . . A T D . D F . Q L L . . E E . T F . .
F03 T S S . F A . . L . . A T D . E F . Q L L . . E E . T F . .
151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180
D. melanogaster H T C G F G C N G G F P G A A W S Y W T R K G I V S G G P Y
Ap 16D2 . S . . . . Y . I K . . E R F K K R . L . T . . D .
Mp 16 . K . . Y . . . . Y . I K . . E R F K K H . L . T . . E .
Tu styraci BN . K . . . . . S . . Y . I R . . E R F K K H . L . T . . N .
Tu styraci BS M D . . K . . G . . Y . I K . . K . F R T Q . V T T . . D .
H03 . L . . . . . H . . . . I K . . E . F K . H . L . T . . N .
H06 . L . . . . . H . . . . I K . . E . F K . H . L . T . . N .
F03 . L . . . . . Y . I K . . E . F K . H . L . T . . N .
181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210
D. melanogaster G S N Q G C R P Y E I S P C E H H V N G T R P P C A H G G R
Ap 16D2 Q . G E . . E . . R V P . . P Y D A E . H N T C A G K P R E
Mp 16 K . G E . . E . . R V P . . P Y D E S . N N T C S G K P M E
Tu styraci BN D . G E . . Q . . R V P . . P L D E Y . N N T C R G K P A E
Tu styraci BS D T K E . . M . . K V P . . - Y D E Q . K N T C G G K P M E
H03 K P R D . . E . . R V P . . F R D K H . N N T C N G Q P M E
H06 K . R D . . E . . R V P . . F R D K H . N N T C N G Q P M E
F03 K T K D . . E . . R V P . . P L D K . . N N T C H G K P M E
211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240
D. melanogaster T P K - C S H V C Q S G Y T V D Y A K D K H F G S K S Y S V
Ap 16D2 S N H R . T R M . Y . -
Mp 16 Q N H R . T R M . Y . -
Tu styraci BN K N H R . T R M . Y . -
Tu styraci BS R N H Q . P K T . Y . -
H03 K N H R -
H06 K N H R -
F03 K N H R -
241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270
D. melanogaster R R N V R E I Q E E I M T N G P V E G A F T V Y E D L I L Y
Ap 16D2 -
Mp 16 -
Tu styraci BN -
Tu styraci BS -
H03 -
H06 -
F03 -
271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300
D. melanogaster K D G V Y Q H E H G K E L G G H A I R I L G W G V W G E E K
Ap 16D2 -
Mp 16 -
Tu styraci BN -
Tu styraci BS -
H03 -
H06 -
F03 -

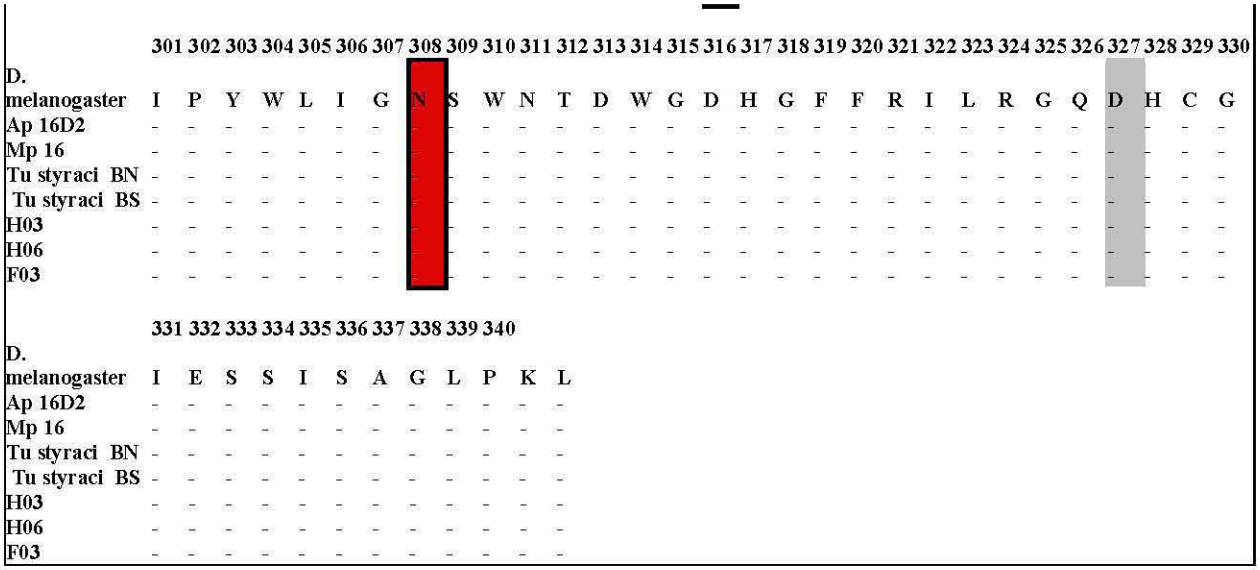


Figure 5: This is an annotated alignment of *catB* that compares the 3 paralogs found in *P. obesinymphae* to *catB* in *D. melanogaster, A. pisum* (Ap 16D2), the soldier-expressed and non-soldier forms in *T. styraci* (Tu styraci BS and BN, respectively), and a form found in the peach-potato aphid, *Myzus persicae* (Mp 16). Areas highlighted in red represent the highly-conserved active sites. The light purple boxed area represents the GCNGG motif, a highly conserved motif in cysteine proteases, and areas highlighted in gray are highly conserved, usually cysteine, residues. The area highlighted in green represents two histidine residues, which are highly conserved in *Drosophila* but apparently are not in aphids, and the occluding loop is boxed. Although the sequences for the 3 *P. obesinymphae* paralogs are incomplete, inspection of the 3 paralogs shows that they do not appear to have a high rate of amino acid substitution that would indicate directional selection.

References

Costa, JT (2006). The Other Social Insect Societies. Harvard University Press, Cambridge.

Jost, M., Hillis, D., Lu, Y., Kyle, J., Fozzard, H., and Zakon, H. (2008). Toxin-resistant sodium channels: parallel adaptive evolution across a complete gene family. *Molecular Biology and Evolution*. **25**(6), 1016-1024.

Kutsukake M, Shibao H, Nikoh N, Morioka M, Tamura T, Hoshino T et al. (2004). Venomous protease of aphid soldier for colony defense. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 11338-11343.

Rawlings, N.D., Morton, F.R., Kok, C.Y., Kong, J. & Barrett, A.J. (2008) *MEROPS*: the peptidase database. *Nucleic Acids Res* **36**, D320-D325.

Rispe C., Kutsukake M., Doublet V., Hudaverdian S., Legeai F., Simon J., Tagu D., and Fukatsu T. (2007). Large gene family expansion and variable selective pressures for cathepsin B in aphids. *Molecular Biology and Evolution* Advance Access, 1-35.

Tang, H., Kambris, Z., Lemaitre, B., and Hashimoto, C. (2006). Two proteases defining a melanization cascade in the immune system of *Drosophila*. *The Journal of Biological Chemistry* **281**, 28097-28104.

Wood, T., Burke, J., and Rieseberg, L. (2005). Parallel genotypic adaptation: when evolution repeats itself. *Genetica* **123**, 157-170.