Comparative analyses of the salivary gland secretomes from related species of the gall midge family Cecidomyiidae

by

Zainab Abdalhussin Ali Al-Jbory

B.S., Baghdad University, Iraq, 1989
M.S., Baghdad University, Iraq, 1997

AN ABSTRACT OF A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Entomology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Abstract

The tools for arthropods with sucking-mouth parts to attack hosts are mainly in the saliva. For plant-sucking insects, these salivary secretions are primarily produced in the salivary glands. Secreted proteins (also referred to as salivary gland secretomes) are among the important components in the saliva of sucking insects. Gall midges (Cecidomyiidae), a large family of plant-sucking insects, apparently secrete proteins (some of them are effector proteins) into host tissues, inducing various forms of plant outgrowth (galls). Three major insect pest species in the genera *Mayetiola*, the stem gall midges, are known to produce saliva that can reprogram plant cells and manipulate the host plant growth, causing serious damage to the plants of small grains. The three pest species are the Hessian fly (*Mayetiola destructor*), the barley midge (*Mayetiola hordei*), and the oat midge (*Mayetiola avenae*). Another economically important species of this gall midge family is the wheat midge (*Sitodiplosis mosellana*). It is a major insect pest of spring wheat and feeds on wheat heads, causing damage to the developing wheat seeds.

A global analysis of the salivary gland secretome of first instar larvae of the Hessian fly, (a member of Mayetiola and) a model species for studying insect-plant interactions, has previously revealed a large number of genes encoding Secreted Salivary Gland Proteins, so called SSGPs. For comparison, we conducted analyses on transcripts encoding SSGPs from salivary glands of the first instar larvae of the wheat midge, barley midge, and oat midge.

In the first chapter, a transcriptomic analysis of wheat midge has been conducted. In this analysis, a total of 3,500 cDNA clones were sequenced, and 1,301 high quality sequences were obtained and approximately 25% of the cDNAs (with high quality sequences) encoded SSGPs. The SSGPs were grouped into 97 groups based on sequence homology. Among the SSGP-encoding transcripts, 206 encoded unique proteins with no sequence similarity to any known

protein and 29 encoded proteins similar to known proteins including proteases, serpines, thioesterases, ankryins, and feritins. The compositions of SSGP transcripts from the wheat midge were then compared with that of Hessian fly. The analyses have identified many common characteristics between the species. Despite these commonalities, no sequence similarity was found between SSGPs from wheat midge and those from Hessian fly, suggesting that SSGPs from these two insect species perform different functions to manipulate host plants.

The second chapter contains results of comparative transcriptomic analyses on the barley and oat midges. A total of 2570 cDNA clones were sequenced from the barley midge, and 743 were high quality cDNA sequences, and the analysis identified 458 cDNA clones encoding SSGPs, of these, 178 encoded unique proteins (also called unigenes). Transcripts encoding SSGPs were grouped into 51 groups based on sequence homology. A total of 3226 cDNA clones were sequenced from oat midge, and 718 cDNA sequences were high quality and used for further analysis. The analysis identified 450 cDNA clones encoding SSGPs. Among the SSGP-encoding transcripts, 194 are unigenes, which were placed into 50 groups.

The compositions of SSGP transcripts from the barley and oat midges were then compared with that of Hessian fly. The analysis identified five groups containing 102 (57.3%) unigenes from barley midges and seven groups containing 107 (55.1%) unigenes from oat midges which encode SSGPs that are conserved among the three species. The SSGPs conserved among the three midges are from family one (SSGP-1), family 4 (SSGP-4), family 11 (SSGP-11), and family 71 (SSGP-71). The SSGPs conserved among the three species indicate conserved functions such as a role in plant manipulation.

Some SSGP unigenes were found to be conserved between only two species. Specifically, there were eight gene groups which are conserved between two species. Within

these eight groups 19 (10.7%) unigenes from the barley midge and 25 (12.9%) unigenes from the oat midge were found to be conserved between only the barley and oat midges, whereas no homologues have been found in the Hessian fly. The remaining unigenes encode SSGPs that are unique to different midge species. The highly divergent SSGP groups that have been identified with no homology among the three midges indicate potential roles of these SSGPs in host specification.

Due to the important roles of effector proteins in insect-plant interactions for gall midge species and since no insect effector protein have been identified directly from infested plant tissues so far, I have chosen one of the SSGP family, SSGP-1, which are conserved among all three gall midge species, for further analysis in chapter 4. Members in family SSGP-1 are also the most abundantly expressed at the transcript level. Based on Hessian fly data, family 1 contains seven genes and are named SSGP-1A1, SSGP-1A2, SSGP-1B1, SSGP-1C1, SSGP-1C2, SSGP-1D1, and SSGP-1E1. To detect the presence of these proteins in the infested wheat tissues, and to identify probable targets from wheat that interact with the SSGPs in the feeding site, we have generated and purified recombinant proteins for five of the seven proteins, namely SSGP-1A2, SSGP-1B1, SSGP-1C1, SSGP-1D1, and SSGP-1E1 (since SSGP-1A1 and SSGP-1C2 are very similar to SSGP-1A2 and SSGP-1C1, respectively). Antibodies were produced for the recombinant proteins for western blot analyses and indirect immunostaining.

Immunostaining on dissected tissues including salivary glands, guts, and Malpighian tubules from 3-day old larvae, was conducted with antibodies against the five SSGPs, and detected a specific localization of all proteins in salivary glands except SSGP-1E1, which exhibited a weak signal in the foregut, in addition to localization in salivary glands. Western blot analyses demonstrated that these five proteins were expressed in larvae at all stages. The

continuous production of these proteins suggests that they play roles in initiation and maintenance in Hessian fly infestation. Consistent with their effector functions, these five proteins were detected for the first time in infested wheat tissues based on western blot analyses.

To identify possible target proteins from host plants that interact with SSGP-1 family proteins, *in vitro* pull down assays were performed. Putative interacting targets for SSGP-1A2, SSGP-1B1, and SSGP-1C1 have been identified by LC-MS/MS. These putative interaction target proteins included uncharacterized proteins, ribosomal proteins, a lipoxygenase, and a tubulin. Identification of these putative targets provided a base for further confirmation of their interaction with Hessian fly effectors in the future.

Comparative analyses of the salivary gland secretomes from related species of the gall midge family Cecidomyiidae

by

Zainab Abdalhussin Ali Al-jbory

B.S., Baghdad University, Iraq, 1989
M.S., Baghdad University, Iraq, 1997

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Entomology
College of Agriculture

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2018

Approved by:

Co-Major Professor
Ming-Shun Chen

Approved by:

Co-Major Professor
C. Michael Smith

Copyright

Abstract

The tools for arthropods with sucking-mouth parts to attack hosts are mainly in the saliva. For plant-sucking insects, these salivary secretions are primarily produced in the salivary glands. Secreted proteins (also referred to as salivary gland secretomes) are among the important components in the saliva of sucking insects. Gall midges (Cecidomyiidae), a large family of plant-sucking insects, apparently secrete proteins (some of them are effector proteins) into host tissues, inducing various forms of plant outgrowth (galls). Three major insect pest species in the genera *Mayetiola*, the stem gall midges, are known to produce saliva that can reprogram plant cells and manipulate the host plant growth, causing serious damage to the plants of small grains. The three pest species are the Hessian fly (*Mayetiola destructor*), the barley midge (*Mayetiola hordei*), and the oat midge (*Mayetiola avenae*). Another economically important species of this gall midge family is the wheat midge (*Sitodiplosis mosellana*). It is a major insect pest of spring wheat and feeds on wheat heads, causing damage to the developing wheat seeds.

A global analysis of the salivary gland secretome of first instar larvae of the Hessian fly, (a member of Mayetiola and) a model species for studying insect-plant interactions, has previously revealed a large number of genes encoding Secreted Salivary Gland Proteins, so called SSGPs. For comparison, we conducted analyses on transcripts encoding SSGPs from salivary glands of the first instar larvae of the wheat midge, barley midge, and oat midge.

In the first chapter, a transcriptomic analysis of wheat midge has been conducted. In this analysis, a total of 3,500 cDNA clones were sequenced, and 1,301 high quality sequences were obtained and approximately 25% of the cDNAs (with high quality sequences) encoded SSGPs. The SSGPs were grouped into 97 groups based on sequence homology. Among the SSGP-

encoding transcripts, 206 encoded unique proteins with no sequence similarity to any known protein and 29 encoded proteins similar to known proteins including proteases, serpines, thioesterases, ankryins, and feritins. The compositions of SSGP transcripts from the wheat midge were then compared with that of Hessian fly. The analyses have identified many common characteristics between the species. Despite these commonalities, no sequence similarity was found between SSGPs from wheat midge and those from Hessian fly, suggesting that SSGPs from these two insect species perform different functions to manipulate host plants.

The second chapter contains results of comparative transcriptomic analyses on the barley and oat midges. A total of 2570 cDNA clones were sequenced from the barley midge, and 743 were high quality cDNA sequences, and the analysis identified 458 cDNA clones encoding SSGPs, of these, 178 encoded unique proteins (also called unigenes). Transcripts encoding SSGPs were grouped into 51 groups based on sequence homology. A total of 3226 cDNA clones were sequenced from oat midge, and 718 cDNA sequences were high quality and used for further analysis. The analysis identified 450 cDNA clones encoding SSGPs. Among the SSGP-encoding transcripts, 194 are unigenes, which were placed into 50 groups.

The compositions of SSGP transcripts from the barley and oat midges were then compared with that of Hessian fly. The analysis identified five groups containing 102 (57.3%) unigenes from barley midges and seven groups containing 107 (55.1%) unigenes from oat midges which encode SSGPs that are conserved among the three species. The SSGPs conserved among the three midges are from family one (SSGP-1), family 4 (SSGP-4), family 11 (SSGP-11), and family 71 (SSGP-71). The SSGPs conserved among the three species indicate conserved functions such as a role in plant manipulation.

Some SSGP unigenes were found to be conserved between only two species. Specifically, there were eight gene groups which are conserved between two species. Within these eight groups 19 (10.7%) unigenes from the barley midge and 25 (12.9%) unigenes from the oat midge were found to be conserved between only the barley and oat midges, whereas no homologues have been found in the Hessian fly. The remaining unigenes encode SSGPs that are unique to different midge species. The highly divergent SSGP groups that have been identified with no homology among the three midges indicate potential roles of these SSGPs in host specification.

Due to the important roles of effector proteins in insect-plant interactions for gall midge species and since no insect effector protein have been identified directly from infested plant tissues so far, I have chosen one of the SSGP family, SSGP-1, which are conserved among all three gall midge species, for further analysis in chapter 4. Members in family SSGP-1 are also the most abundantly expressed at the transcript level. Based on Hessian fly data, family 1 contains seven genes and are named SSGP-1A1, SSGP-1A2, SSGP-1B1, SSGP-1C1, SSGP-1C2, SSGP-1D1, and SSGP-1E1. To detect the presence of these proteins in the infested wheat tissues, and to identify probable targets from wheat that interact with the SSGPs in the feeding site, we have generated and purified recombinant proteins for five of the seven proteins, namely SSGP-1A2, SSGP-1B1, SSGP-1C1, SSGP-1D1, and SSGP-1E1 (since SSGP-1A1 and SSGP-1C2 are very similar to SSGP-1A2 and SSGP-1C1, respectively). Antibodies were produced for the recombinant proteins for western blot analyses and indirect immunostaining.

Immunostaining on dissected tissues including salivary glands, guts, and Malpighian tubules from 3-day old larvae, was conducted with antibodies against the five SSGPs, and detected a specific localization of all proteins in salivary glands except SSGP-1E1, which

exhibited a weak signal in the foregut, in addition to localization in salivary glands. Western blot analyses demonstrated that these five proteins were expressed in larvae at all stages. The continuous production of these proteins suggests that they play roles in initiation and maintenance in Hessian fly infestation. Consistent with their effector functions, these five proteins were detected for the first time in infested wheat tissues based on western blot analyses.

To identify possible target proteins from host plants that interact with SSGP-1 family proteins, *in vitro* pull-down assays were performed. Putative interacting targets for SSGP-1A2, SSGP-1B1, and SSGP-1C1 have been identified by LC-MS/MS. These putative interaction target proteins included uncharacterized proteins, ribosomal proteins, a lipoxygenase, and a tubulin. Identification of these putative targets provided a base for further confirmation of their interaction with Hessian fly effectors in the future.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my sincere gratitude to my co-major advisors Drs. Ming-Shun Chen and Dr. Charles Michael Smith for their continuous support of my PhD study and related research, and for their valuable and constructive suggestions during the planning and development of this research work. Their willingness to give their time so generously has been appreciated.

Besides my advisors, I would like to thank the rest of my thesis committee, Drs. Jeff Whitworth and Gerald Reeck for their insightful comments and encouragement, Dr. Anita Dille for serving as the chairperson for my examination committee.

I would like to offer my special thanks to Dr. John Reese who provided me with great tools and support for my dissection work. Dr. Yoonseong Park for his assistance with the immunostaining experiments.

My special thanks to Dr. Marion Harris and Kirk Anderson at North Dakota state University whose assistance was important in facilitating the salivary glands dissection from wheat midge larvae. Dr. Mustapha El-Bouhssini at ICARDA who provided me with samples of salivary glands for barley midge and oat midge.

I would also like to extend my thanks to Shauna Dendy for her assistance during my research work in the greenhouse. I am grateful to Lisa Pacha, the USDA Program Support Assistant for her sincere assistance in my project.

My special thanks to Dr. John Ruberson, the head department of Entomology for his support for my PhD program during these five years. I would also like to extend thanks to the precious services provided by the main office staff and Kent Hampton.

Special thanks given to all faculties, graduate students in the Department of Entomology, and my colleagues in Dr. Chen's lab for offering help and support.

I am immensely grateful to the Ministry of Higher Education of Iraq for providing me with the scholarship for my PhD studies.

Finally, I would like to thank my family: my husband and my sons for their support and encouragement throughout my studies.

Dedication

This thesis is dedicated to my family. I have a special feeling of gratitude to my husband Maher for his unwavering support throughout this long journey; to my sons, Fawwaz, Ahmed, and Abdulrahman for their encouragement and support throughout the doctorate program. This work is also dedicated to the souls of my parents for their love and care, and whose good examples have taught me to work hard for the things that I aspire to achieve, and also to my sisters and brothers who supporting me all the time.

Chapter 1 - Background Information and Literature Review

The gall midge family Cecidomyiidae

Cecidomyiidae is one of the largest families of Diptera and is best known as gall midges. Included in this family are some of the most destructive pests of grains, fruits, and vegetables as well as important predators of aphids, scale insects and mites. Larvae of this family show a great range of feeding habits, including fungivory, herbivory, and predation on various arthropods (Gagne´ 2010). The family has an ancient origin, with a fossil from the Jurassic Period more than 150 million years ago (Yukawa and Rohfritsch 2005). There are ~4000 gall midge species identified and the majority of the members are gall-inducing species. Cecidomyiidae underwent explosive diversification in the Cretaceous period coincident with the appearance of flowering plants. They are known to diversify both through host plant shifts (Price 2005) and through ecological partitioning of a single plant (Joy and Crespi 2007).

As with most parasites, gall midge life cycles are highly synchronized with those of their hosts (Gagne´ 1989), the galls can be produced on different parts of the host plant such as the buds, stems, leaves, flowers, and fruit of dicotyledons, monocotyledons, gymnosperms, ferns, and mushrooms (Gagne´ 1994). Most species are adapted to only one type of plant tissue on a single plant species or a closely related group of species, and certain plants are hosts to multiple, but monophyletic gall midge species (Joy and Crespi 2007). Compared with other gall-forming insect taxa, the gall midges have colonized the widest variety of plants (~ 89 plant families in North America) (Price 2005). Cecidomyiidae belongs to the lower Diptera (Nematocera), which is believed to have arisen during the cretaceous era and is probably closer in evolution to mycetophilids and fungus gnats (Bertone et al. 2008). Gall midges parasitize host plants in many ways similar to those of biotrophic or hemibiotrophic plant pathogens, including a permanent

feeding site, ability to modulate gene expression, induction of either plant susceptibility or qualitative resistance, and the presence of effector proteins (Chen et al. 2008; Stuart et al. 2012; Zhao et al. 2015). In respect to reproduction, adult females of gall midges have high potential fecundity. Female flies emerge with a full cohort of mature eggs and typically have less than one day to deposit those eggs before they die (Harris et al. 2003). Although adults are short-lived, gall midges have another non feeding life stage, the last larval instar, which behaves like a fungal spore, and can survive extended periods of harsh environmental conditions (Barnes 1956).

Due to the ancestral origin of this family, this group provides valuable information on the evolution of primitive groups to the more derived gall-inducing subfamilies (Gagne´ 1989). Ultimately, understanding the evolutionary origin of gall midges may help to explain insect-plant interactions based on a gene-for-gene relationship, which is characteristic of insect- host plant interactions of some important species of the gall midge family (Stuart et al. 2015).

Genera Mayetiola

Mayetiola is a Palearctic genus containing 29 described species, which are adapted to a particular genus of plants in the Gramineae family (Skuhrava 1986). *Mayetiola* are species distinctive for the presence of a puparium and the larvae live in stems of grasses. They are either monophagous, restricted to one host, or oligophagous, restricted to a few closely related host species (Gagne´ 1989).

Cereal crops, including wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), and oat (*Avena sativa*) are of high economic importance by providing a daily source for human nourishment. Although crop yields have increased worldwide in recent decades due to advanced technology that has been widely used in breeding and cultivation, cereal crops are known to host many herbivorous insects (Johnson et al. 1978). Among these insect pests, several stem gall midge species belong to the genera *Mayetiola*, including the Hessian fly (*Mayetiola destructor*),

the barley midge (*Mayetiola hordei*), and the oat midge (*Mayetiola avenae*). These species are important insects annually causing severe damage to their host plants in specific regions, or worldwide (Gagne´ 2010). For many years, the taxonomic status of *Mayetiola* and how to confirm identification of its species was a controversial issue among scientists. A lot of identification work has been extensively investigated as there may be confusion with other described and undescribed species of *Mayetiola* due to the similarity in biology, damage symptoms, and overlap host ranges (Gagne´ et al. 1991).

Importance, distribution, and host range of *Mayetiola* species & wheat midge

Mayetiola species have long been known for their economic importance. Among the three species, Hessian has the widest distribution and widest host range. Barley midge and oat midge have a narrower host range and their occurrence is regional. However, like Hessian fly, barley and oat midges have the potential to spread beyond their current exisiting regions. Many governments keep tight quarantine restrictions to prevent introducing these species to their lands. For instance, USA has a strict quarantine against barley and oat midges. Australia also has strict procedure to prevent Hessian fly and barley midge from entering their land. The wheat midge is a significant pest of both bread and durum wheat, and is distributed throughout many wheat-growing regions in the Northern Hemisphere with a proven record as an invasive species (Berzonsky et al. 2003)

Hessian fly is one of the most destructive insect pests of wheat. In USA and many other countries, controlling the Hessian fly is a crucial factor in wheat production. Hessian fly causes average yield losses from 5 to 10% annually (Buntin 1999). In the USA, loss due to Hessian fly infestations has been estimated ranged from 6 to 30% in eight northwest Kansas counties (Wiseman and Webster 1999). In Georgia, estimated yield loss was worth $28 million from 1988 to 1989 (Buntin et al. 1992). In Indiana, significant losses occur almost yearly in susceptible

cultivar fields (Patterson et al.1990). Around the turn of the 20th century, this insect destroyed millions of bushels of wheat in Missouri (Boyd and Bailey 2000). In South Carolina, wheat losses were more than $4 million per year from 1984 to 1989 (Chapin 2009), and losses of 21 bushels per infested acre occurred in 1985 in Alabama (Flanders et al. 2008).

Hessian fly is believed to have originated in West Asia in the Fertile Crescent and spread to neighboring regions (Bouhssini et al. 2009). Currently, Hessian fly is a major threat to wheat production in North America, New Zealand, North Africa, southern Europe, and northern Kazakhstan (Pauly 2002). In North America, Hessian fly was accidently introduced from the southern Caucasus region of Russia by Hessian troops during the American Revolutionary War of 1776 (Walton 1920). Since the initial settlement of the insect in Long Island, New York, it has spread to most wheat-growing areas in the United States (Watson 2005). So far, many states reported Hessian fly infestations, including Kansas, Nebraska, Georgia, Oklahoma, Texas, Louisiana, South Carolina, Pennsylvania, North Carolina, Virginia, Alabama, and Missouri (Chen et al. 2009; Cambron et al. 2010). It was first detected in Kansas in 1871. During the first half of the present century, only about one out of every four years was designated as being free of serious fly injury (Brooks 1993)

The preferred host for the Hessian fly is wheat but it can also develop on rye, barley, triticale, oat, and some grasses, if wheat is not available, such as Aegilops, Lolium, Elytrigia, Bromus, Elymus, and some species of Agropyron. Elytrigiarepens (*Elymusrepens*) is an alternative host and it might have been the actual original host to Hessian fly in Europe (Barnes 1956). In Morocco, Hessian fly is mainly a pest of wheat, but occasionally infests barley, on which a morphologically distinct species, barley stem gall midge, is the main pest (Gagné et al. 1991).

Barley midge is a major pest of barley and causes serious damage to barley production. When the barley midge becomes established in a region or country, infestation rates and yield losses can become significant. In Morocco, barley fields may be infested by both barley midge and Hessian fly, although barley midge is the more important one. Yield loss of barley in Morocco has been estimated to be about 35% due to the damage caused by barley midge according to the annual report of ICARDA (ICARDA 2002). Losses caused by the barley midge are similar to losses caused by Hessian fly on wheat in Morocco (Lhaloui et al. 1992). Approximately 30 to 50% of the barley crops in Libya and Tunisia are infested by the barley midge (ICARDA 2002).

Limited information is available on the distribution of the barley stem gall due to confusion with Hessian fly. In USA, the insect hasn't been documented yet in barley growing areas, however, barley midge has been recorded in the Mediterranean regions of southern Europe (Italy, Spain, United Kingdom, and France) and northern Africa (Morocco, Algeria, Libya, and Tunisia) (Makni et al. 2000; Gagne et al. 1991; Lhaloui et al. 1988).

Barley is the preferred host for the barley midge (Gagné et al. 1991). This pest, however, has also been recorded on oat, wheat and rye (*Secale cereale*). Defining the host range of the barley midge has been difficult due to confusion with the Hessian fly.

Oat midge is a major pest of oats, causing injury in some regions of oat areas. An earlier report on the damage caused by oat midges was documented in autumn by Marchal (1893), on winter oats at Rambaud near Poitiers and also in oat fields near Vendome, France. Years later, another heavy infestation had been reported near Vienne in the same neighborhood (Barnes 1956). Recently, a preliminary survey in Morocco of field plants in 2014 revealed that up to 30%

of oat plants were infested and killed by this pest (personal contact with Dr. Mustapha El bouhssini, ICARDA).

Limited information is available about the dispersal of the oat midge worldwide. Like barley midge, oat midge hasn't been documented yet in USA. However, damage has been reported in other countries in Europe including Italy, Finland, Russia, Hungary, and Great Britain. In North Africa, the insect has been recorded on oats in Algeria (Barnes 1956).

Oats are the preferred host for the oat midge. The insect has also been recorded on the wild type (*Avena fatua*) (Bagnall and Harrison 1918). Marchal (1893) showed that although oat midges laid eggs both on wheat and oats, they exhibit great preference for oats and larvae failed to develop on wheat (Barnes 1956).

Wheat midge is one of the most destructive pests of wheat in the northern hemisphere. Attacks by the wheat midge can significantly reduce harvest yield and grain quality (Harris et al. 2017). Many countries in Europe have experienced serious outbreaks resulting in significant damage. In the United Kingdom, crop losses exceeded $25 million in 1931 and $50 million in 2004, in spite of insecticide application (Oakley et al. 2005). Annual losses in Canada during 2000-2010 were approximately $60 million per year. In the USA, the wheat midge outbreak in the mid-1990s caused crop losses on more than 725,000 acres of wheat that were valued in excess of $27 million to North Dakota (Knodel and Ganehiarachchi 2016).

It is believed that the wheat midge originated in Europe and first record in 1741 in England, then introduced to the North American continent in the 1800s and was established in Quebec by 1828 (Lamb et al. 1999). First outbreak occurred in the northeast and midwest of North America in the 1850s. Since then, it has been recorded in various locations throughout the Old World and New World (Harris et al. 2003). In North America, the wheat midge has been

reported in Indiana (Barnes 1956), Minnesota, Montana, and North Dakota. Many areas in Western Canada have been found with this pest, including Alberta, Saskatchewan, Manitoba and British Columbia (Berzonsky et al. 2003). In the old world continents, wheat midge also spread to China and Japan (Barnes 1956; Harris et al. 2017). Host range of the wheat midge includes 16 other species in the genus *Triticum*, including rye and triticale. Forage and wild grasses also are hosts (Berzonsky et al. 2003; Harris et al. 2017)

Symptoms of plants attacked by gall midge species

In general, the three species of *Mayetiola* cause similar symptoms on infested cereal plants. However, there are some differences at the feeding site on the stem that can be used to accurately distinguish the three species. The major difference is that each species has preferred host species. Hessian fly rarely attacks oat plants while oat midge has not been found in wheat plants. Both Hessian fly and barley midge can attack barley, but barley midge larvae do induce a kind of out-growth of tissues and the larvae stick to the expanded gall-tissue very tightly (Gagne´ et al. 1991). On the other hand, wheat midges have a different feeding site. Newly hatched larvae feed on the seeds during anthesis in the early stages of development, resulting in shriveled wheat kernels (Lamb et al. 2000).

Hessian fly: In the field, symptoms such as stunted wheat plants and dead tillers in the seedling stage, small or deformed heads, uneven plant height and maturity, or lodged stems at the mature plant stage indicate the possibility of Hessian fly damage (Brooks 1993). A single maggot feeding on a plant for three consecutive days can irreversibly stunt a young plant or tiller. Flies emerging in the fall can severely stunt or kill early planted wheat, resulting in severe stand loss. In warm winters, wheat can be infested by Hessian fly in Dec.-Jan in southern regions of the USA. Jointed wheat is then re-infested by the spring generation in March- April (Chapin 2009). Tillers infested during jointing may form small, poorly filled heads. When larvae attack

higher nodes, straw strength is weakened, resulting in lodging. In the USA damage from Hessian fly is greatest in winter wheat planted early, before the so-called fly free date, and in spring wheat planted late, in synchrony with a spring generation of the pest (Cook and Veseth 1991). Contributing factors to severe infestations include the use of susceptible varieties over large areas, early planting, unusually warm weather in November- December, reduced tillage into wheat stubble, and abundant volunteer wheat due to the lack of rotation (Walton 1920).

Barley midge: The barley midge can cause significant infestations of barley, and occasionally wheat, rye, and oats. Plants infested by the barley midge can be found throughout the growing season on both young and mature plants. Symptoms in young seedlings are mainly a yellowing of new growth and can occasionally result in death of the plant. In mature plants, the barley midge feeds at the base of the plant between the leaf sheath and the stem, producing the characteristic swellings (galls), leading to stem weakening, and loss of grain yield and quality (Parker et al. 2001).

Oat midge: Oat plants infested by the oat midge occasionally exhibit some characters such as swollen stems, bulbous, either basally or at a node, and usually white gregarious larvae or brown puparia clustered together under the leaf sheath. There are two types of damage depending on the plant age (Barnes 1956). When young plants are attacked, a small bulb is formed at the base of the plant and the leaves of the central shoot turn yellow and dry up before opening. The plant subsequently dies unless a new tiller grows out. In the case of plants at a more advanced growth stage, larvae may be found at the 1st or 2nd node, or less often, at the 3rd or 4^{th} nodes. Marchal and Ricchello, recorded two generations of oat midges on oats in France and Italy (Barnes 1956), respectively. First emergence and flight of adults occurred in late April in France and in early March in Italy, So then the progeny of this generation attack both spring

and winter oats. Second emergence takes place in October and November in France, and progeny of this generation attack autumn or winter planted oats (Barnes 1956).

Wheat midge: On the Northern Great Plains, the wheat midge primarily infests wheat, which is seeded in the spring and harvested in late summer (Knodel and Ganehiarachchi 2016). The wheat midge is a seed feeder and infests a wheat plant during heading through early flowering. The attacked seed is misshaped and shrunken. Degree of damage depends on number of larvae and where the larva feeds, with attack near the embryo causing the greatest damage (Barnes 1956). Since the amount of food taken by the larva is constant, smaller seeds are affected more than larger seeds. It is common to find 1 to 3 larvae inside each floret, but up to 30 larvae can mature on a single seed (Harris et al. 2017). Over 40% of the seed weigh less than 8 mg and these seeds are usually lost during harvest. The remaining bigger seeds are harvested but can have other problems including premature sprouting, reductions in germination, loss of early vigor of seedlings, and poor grain quality. When attacked by wheat midge larvae both bread and durum wheat show undesirable changes in protein levels and dough strength (Berzonsky et al. 2003).

Biology and feeding mechanism of gall midge species

Biology and life cycle of the gall midge species: Hessian fly, barley midge, oat midge, and wheat midge, are very similar across all stages. The differences are in length of time for each developmental stage. Like other Dipterans, gall midges have four life stages: eggs, larvae, pupae, and adults. The gall midge species are known to cause damage by feeding of maggots but adults do not feed. These insect species have a relatively long larval stage. The first instar attacks the plant by injecting salivary secretions into plant tissue to establish a permanent feeding site. The second instar feeds vigorously on plant tissue through the feeding site established by the first instar. The third instar becomes a non-feeding puparium stage. The first instar is the most critical

stage since it paves the way for the second instar larvae to feed. Larvae of the gall midges are sessile and suck up the liquid from the reprogrammed cells (the gall) (Stuart et al. 2012). The adult stage of the life cycle is short and it is specialized for reproduction. During the adult stage, the insect will not feed and will rely on internal energy stores and the uptake of water. In this time the fly must mate, disperse, locate hosts, and oviposit. The placement of eggs determines not only which plant the larva feeds upon but also where on the plant feeding will occur (Gagné, 1989).

Hessian fly: The Hessian fly has a short life cycle of about 28 days at 20˚C. There are various numbers of generations (2-5) per year in the field, depending on temperature and host plant availability (Walton 1920). Adults live only for 2 to 3 days, just long enough to mate and lay eggs. Females deposit oblong red eggs individually in parallel grooves on the adaxial leaf surface. Eggs may develop and hatch into first instar larvae in 3 to 4 days. Larva (maggot) is the harmful stage and usually lasts for 20 to 21 days at 20˚C and go through three instars. Newer leaves are preferred for egg laying, and a female fly may lay between 50 to 400 eggs (In 3 to 4 days at 20˚C a reddish-brown larva or maggot emerges from an egg). This first instar larvae crawl down the upper leaf surface all the way to the base and live between leaf sheaths by establishing a permanent feeding site within 2 to 3 days. The first instar larva is the most critical stage in the interaction between the Hessian fly and its host plant. Second instar larvae keep feeding for 4 to 5 days at 20˚C. After an inside molting, the larvae stop feeding and go through the third instar (non-feeding instar). The exoskeleton becomes dark brown called flaxseed. After 6 to 7 days, the puparium enters dormant state to turn to a pupa, which lasts for 7 to 10 days at 20˚C (Harris et al. 2003; Brooks 1993).

Barley midge: Although the biology and life cycle of the barley midge is similar to that of the Hessian fly, all stages and instars of the barley midge required longer developmental time than those of the Hessian fly. Generation time for non-diapausing barley midges is about 45 days at 18±1 ^{o}C, and under natural field conditions in Morocco, the insect had three generations (Lhaloui 1995). Unlike the Hessian fly, barley midge females lay between 115 to 265 eggs and they lay a high proportion of their pale red eggs on the abaxial leaf surface and stems. Eggs need seven days to complete development and hatching. The reddish neonates crawl down the leaves to reach the base of the stem and establish the feeding site, just like the Hessian fly. After 3 days of feeding the first instar larvae turns white with cylindrical bodies. The first instar of barley gall midge needs 9 to 10 days to complete. Soon after molting, second instar larvae increase in size and gall tissue gradually formed around the larvae and each larva became completely embedded in a pea size gall. Unlike Hessian flies, 2nd instar larvae of barley gall midge can survive and form galls on leaves, second instar larvae requires 12 days to develop into the third instar. However, as observed in the Hessian fly, the third instar of barley midge is a non-feeding stage and larvae of both species reversed their position inside the puparia, so the emerging adults would be oriented with their heads upwards. Time to complete the third instar is 10 days longer and pupal stage of barley midge is 12 days longer than that observed in Hessian fly, which is 8 and 6 days, respectively, at 18±1 ^{o}C (Lhaloui 1995).

Oat midge: The life cycle of the oat midge is very similar to that of both Hessian fly and barley midge. The oat midge has 2 to 3 generations in Europe owing to extra favorable conditions of food supply, warmth, and humidity. Adults are short-lived, surviving up to 4 days, and females lay 120 to 180 pale red eggs on the upper surface of the leaves. In the autumn generation egg stage lasts 15 to 20 days (Barnes 1956). Upon hatching, larvae migrate down the

leaf to the nearest axil and establish a feeding site at a position as far down as possible above a node. The larvae enter the second instar after molting and continue feeding and growth. In this stage it's very difficult to distinguish the larvae of oat midge from those of Hessian fly. The number of larvae found together and the development of the gall vary with their position on the stem. A characteristic bulb is formed and as many as 20 to 30 larvae may be found together when the attack is at ground level. When the attack is at the first or second node, fewer larvae occur together and the bulb is less developed, but still quite obvious. When the larva completes its growth, its skin swells and becomes brown to start the third instar or dormant stage, this stage is the longest, lasting 6 months in summer and a little less during the winter. The pupal stage lasts 10 to 30 days, and mating take place soon after emergence (Barnes 1956).

Wheat midge: The adult wheat midge is an orange-colored, 2 to 3 millimeters long. The wheat midge has only one generation per year in the northern states of USA. Like the Hessian fly and the other two gall midges, wheat midges have four life stages: egg, larva, pupa, and adult. Adults emerge from overwintering sites in the soil about the last week of June or first week of July when wheat plants are just beginning to head (Harris et al. 2003). Like Hessian fly, wheat midge adults do not feed and only live for 3 to 4 days. Mated females deposit eggs either individually or in groups of three to five on the spikelets of wheat heads. Eggs are elongate, whitish and very small. Newly hatched larvae feed on the surface of the primordial seed for 10 to 14 days to complete two larval instars. Like Hessian fly, the third instar is a non-feeding stage in which the larvae enter a period of dormancy within the floret. Length of dormancy period depends on moisture levels. After rain or heavy dew, larvae break dormancy and become active again (Harris et al. 2017). It wiggles out of its 'skin', drops to the soil, burrows to a depth of several centimeters, and forms a silken cocoon. This obligatory diapause of the third instar larva

occurs over the winter months. In the spring the larva emerges from the cocoon, crawls towards the soil surface, and then pupates just below the soil surface. Adults emergence coincides with the formation of wheat heads (Harris et al. 2017). The wheat midge exhibits an unusual capacity for long-term dormancy in the soil, lasting up to 12 years (Barnes (1956).

Virulence of gall midges

Fossil records show earlier existence and rapid evolution of some of them to synchronize the evolutionary change in flora kingdom which indicates high potency in adaptation and virulence of this family (Price 2005). Interaction between some gall midge species and their host plants have been studied previously to better understand the virulence of these species and plant resistance. Hessian fly is becoming a model species for studying insect-plant interactions, because of the Hessian fly's importance in agriculture, intriguing behavior, ease of maintenance in culture, and relatively well- characterized genetics (Harris et al. 2003). Studies have revealed that resistance genes (R-genes) in wheat cultivars could continuously be overridden by larvae change the composition of proteins in their saliva. This high adaptation of Hessian fly shortens the durability of R-genes, and costs the wheat industry money, and many years of efforts in deploying these genes (Gould 1998). Insect - plant interactions of barley midges and oat midges haven't been studied yet, and no resistance genes have been identified for either species. However, it's generally accepted that larvae of the stem gall midge species feed exclusively on cell content of their hosts and that their injected saliva plays a role in reprograming plant cells and manipulating the host plant growth (Hatchett et al. 1990).

On the other hand, little is known regarding detailed feeding mechanisms of wheat midges. However, deploying wheat cultivars with the resistance gene (Sm1) is known to significantly limit kernel damage due to wheat midge (Blake et al. 2014; Smith et al. 2014).

Feeding mechanism of Hessian fly

For many years, the feeding mechanism of Hessian fly larvae has been known. Although stunting and dark green leaves of infested plants have long been recognized as symptoms of feeding, obvious damage couldn't be observed since larvae live within a plant and have the ability to keep the plant alive (Haseman 1930). Hessian fly larvae are incapable of physically rupturing plant cells (Painter 1951). Refai (1955 and 1956) studied the mechanism of larvae feeding by visual and audio means, and found that larval secretions initially softened cell walls of seedlings and then feed by sucking the fluid. A functional model of the feeding mechanism of the first instar has been proposed based on the hypothesis that salivary fluids are injected into the plant via highly specialized mandibles and that these substances react with plant cell components and cause physiological alterations in the plant that allows the larva to suck up liquids through their functional mouth parts (Hatchett et al. 1990). Recent studies provide further support for the crucial role of the first instar larvae in manipulating growth and metabolism of the host plant (Chen et al. 2008).

Upon hatching, a larva moves down along a wheat leaf blade, enters the space between two leaf-sheaths or a leaf sheath and the stem, and continues to migrate all the way down to the base, where the larva establishes a permanent feeding site. The mandibles of first instar larvae have been minutely modified to resemble the end of a hypodermic needle, and they are grooved on the internal lateral surface. The groove extends from the tip of the mandible internally into a basal hole (Stuart et al. 2012). This modification in mandibles is essential and serves as piercing organ analogous to a short stylet in aphids, or haustorium in pathogens (Torto-Alalibo et al. 2009). During feeding, the larva first presses its head and mouth parts against the epidermal cells. The minute mandibles are then inserted into the plant tissue, and salivary secretion flows through an opening duct to fill the elongated grooves on the lateral surface of the mandibles,

which deliver salivary secretions into plant cells to effect extraoral digestion and plant changes. The mandibles then are retracted and after a short latent period the larva cups its mouth cavity tightly over the feeding site, brings the pharyngeal food tube in contact with the epidermal cells, and sucks liquid from the plant (Hatchett et al. 1990). Therefore, it's strongly believed that mandibles of the first instar larvae of Hessian fly have evolved functionally into highly specialized structures for salivary secretion and injection. The mandible blades of the second instar are quite different from that of the first instar. The mandible blades of the second instar are broadly rounded, curve outwardly at their apexes, and appear widely separated distally (Gagne´ and Hatchett 1989). This structural difference suggests that the mandibles of the second instar don't function in piercing and that the second instar may not inject salivary fluids into the plant further because physiological alterations in the plants are induced entirely by feeding of the first instar and the basal region of the salivary glands begins to degenerate in the second instar (Stuart and Hatchett 1987). The immobile second instar may simply suck the plant liquids that exude from the permanent feeding site of the first instar. Thus the feeding mechanism of the second instar may involve only ingestion (Hatchett et al. 1990).

Phenotypically, the feeding mechanisms of Hessian fly are similar to those of barley midge and oat stem midge. Although, these three species share commonality in feeding mechanisms, they have mechanisms that are species-specific. They all feed between leaf-sheaths close to the base of the stem, inhibit host growth, and eventually kill the plant (Hatchett et al. 1990). The strongest evidence that Cecidomyiid larvae ingest liquid food is the substantial reduction in their head and considerable modification of the mouth parts (Hatchett et al. 1990).

Structures of Hessian fly salivary glands

The morphology and cytology of the salivary glands were described in detail to correlate to the larval growth and development (Stuart and Hatchett 1987). Hessian fly salivary glands are

composed of two distinct regions, the duct of each gland extended distally from a common duct to a group of 18-24 relatively large cells which composed a distinct region of the gland, the basal region. Distal to the basal region is a long cylindrical region, the so called filament region, which consists of 80 to 100 cells arranged in two rows opposed to each other across a lumen (Stuart and Hatchett 1987). During the early feeding period of larval development, the basal region constitutes the greatest proportion of the salivary glands and has the greatest growth rates at the first 4 to 6 days of 1st instar. The basal region, however, decays in later instars, and it constitutes only a smaller proportion of the gland during the nonfeeding third instar period (Cartwright et al. 1959; Asavanich and Callun 1979). These morphological changes of the salivary glands support the conclusion that the basal region is the probable source of substances that are believed to cause stunting of wheat plants (Stuart and Hatchett 1987).

However, when the relative growth rates of the basal and filament regions were compared, the filament region showed no increase in size during the first 3 days of larval feeding. The relative growth rate of the filament region exceeds that of the basal region only after day 12, the beginning of the nonfeeding period, and reached its greatest rate on day 15 of larval development (Stuart and Hatchett 1987). Based on these lines of evidence, it is likely that the filament region has primarily a nonfeeding function.

Secretions of salivary glands and saliva

Most of the previous studies have suggested that secretions of Hessian fly salivary glands are the effectors for reprogramming host-plant tissues in compatible interactions with susceptible wheat plants and are the sources for the avirulence gene products that elicit resistance in incompatible interactions with resistant wheat plants (Smith 2005). Thus, due to the importance of the Secreted Salivary Gland Proteins (SSGPs) in Hessian fly virulence and biotype differentiation, previous research efforts have revealed that there are hundreds of the SSGPs

produced in the salivary glands (Liu et al. 2004; Chen et al. 2004; 2006). Further expanded transcriptome analysis revealed that approximately 60% of transcripts in salivary glands of Hessian fly first instar larvae encode SSGPs (Chen et al. 2008). Later, genomic sequencing revealed more than 7% of predicted genes in the Hessian fly genome encode effector-like proteins (Zhao et al. 2015). These findings support the conclusion that proteins in larval saliva are mainly specialized for synthesizing SSGPs (Chen et al. 2008). Most of SSGPs identified are exclusively expressed in salivary glands. The functions of these salivary gland-specific proteins are unknown and their structures are unique. There are, however, several types of proteins that are expressed in the salivary glands and gut. Among these proteins are proteinases (Zhu et al. 2005) and proteinase inhibitors (Maddur et al. 2006). A secreted lipase-like protein encoded by the gene MdesL1 is also expressed in salivary glands as well as in other tissues (Shukle et al. 2009). At the evolutionary level, analyzing transcripts encoding SSGPs from different geographical populations of Hessian fly indicated greater diversity in the Israeli population than those from the USA, which suggests that the divergent SSGPs may represent ancestral types (Johnson et al. 2009). On the other hand, unusual conservation patterns have been discovered in several families of SSGPs of the Hessian fly, in which the coding regions are highly diversified and the non-coding regions are highly conserved. These findings suggested that this rapid diversification in the regions encoding mature SSGPs have evolved under high selection pressure from host plants for functional adaptation (Chen et al. 2010).

In order to develop (resistant wheat) cultivars with resistant genes, scientists have exerted efforts to better understand insect- plant interactions. In plant - insect interactions many plant pathogens and parasitic insects possess an effector-based mechanism to attack host plants and promote virulence via secreted effector proteins (Miles 1999; Harris et al. 2015), and therefore

many effectors have been identified and characterized from parasitic insects (Stuart 2015). For instance, a large number of the SSGPs produced in salivary glands of Hessian fly larvae have been identified to play effector roles once injected into plant tissues (Chen et al. 2008, 2010; Zhao et al. 2016). In addition, several avirulence effectors have been cloned from the Hessian fly and all of them were SSGPs (Aggarwal et al. 2014; Zhao et al. 2015, 2016). Many secreted proteins have also been identified in the saliva of several aphid species (Thorpe et al. 2016), and many of these aphid proteins act as effectors either to suppress or trigger plant defense responses (Elzinga et al. 2014). In other cases, studying the genetic responses to the susceptible and resistant reactions in rice gall midge larvae (*Orseolia oryzae)* indicates that effector proteins participate gall induction (Sinha et al. 2012). An effector triggering rice resistance to the Asian rice gall midge has been cloned from rice (Stuart 2015).

Like the Hessian fly, many other gall midges attack host plants by injecting saliva into host tissues as well. Therefore, analyzing transcripts in the salivary glands and identifying those proteins with a secretion signal peptide is an efficient way to identify putative effector proteins. The availability of these putative effector genes provides a foundation for further studying and characterizing the roles of these genes in insect plant interactions.

References

Asavanich, A.P. and Gallun, R.L.1979. Duration of feeding by larvae of the Hessian fly and growth of susceptible wheat seedlings. Annals of the Entomological Society of America, 72(2), 218-221.

Bagnall, R.S. and Harrison, J.W.1918. XIII. A Preliminary Catalogue of British Cecidomyidae (Diptera) with special reference to the Gall-midges of the North of England. Ecological Entomology, (65), 346-348.

Barnes, H.F.1956. Gall midges of economic importance. Vol. VII: Gall midges of cereal crops. London, UK: Crosby Lockwood.

Bertone, M.A., Courtney, G.W. and Wiegmann, B.M. 2008. Phylogenetics and temporal diversification of the earliest true flies (Insecta: Diptera) based on multiple nuclear genes. Systematic Entomology, 33(4), 668-687.

Berzonsky, W.A., Ding, H., Haley, S.D., Harris, M.O., Lamb, R.J., McKenzie, R.I.H., Ohm, H.W., Patterson, F.L., Peairs, F.B., Porter, D.R. and Ratcliffe, R.H. 2003. Breeding wheat for resistance to insects. Plant Breeding Reviews, (22), 221-296.

Blake, N.K., Stougaard, R.N., Bohannon, B., Weaver, D.K., Heo, H.Y., Lamb, P.F., Nash, D., Wichman, D.M., Kephart, K.D., Miller, J.H. and Reddy, G.V.P. 2014. Registration of 'Egan'wheat with resistance to orange wheat blossom midge. Journal of Plant Registrations, 8(3), 298-302.

Bouhssini, M., Chen, M.S., Lhaloui, S., Zharmukhamedova, G. and Rihawi, F. 2009. Virulence of Hessian fly (Diptera: Cecidomyiidae) in the Fertile Crescent. Journal of Applied Entomology, 133(5), 381-385.

Boyd, M.L., Bailey,W.C. 2000. Hessian fly Management on Wheat. MU guide, Missouri Univ. Extension. http://extension.missouri.edu/p/g7180.pdf.

Brooks, H.L. 1993. The Hessian fly understanding its importance and management in wheat production. Kans.State Univ. Coop.Ext. Serv. MF-1076.

Buntin, G.D., Ott, S.L. and Johnson, J.W.1992. Integration of plant resistance, insecticides, and planting date for management of the Hessian fly (Diptera: Cecidomyiidae) in winter wheat. Journal of economic entomology, 85(2), 530-538.

Buntin, G.D. 1999. Hessian fly (Diptera: Cecidomyiidae) injury and loss of winter wheat grain yield and quality. Journal of Economic Entomology, 92(5), 1190-1197.

Cambron, S.E., Buntin, G.D., Weisz, R., Holland, J.D., Flanders, K.L., Schemerhorn, B.J. and Shukle, R.H. 2010. Virulence in Hessian fly (Diptera: Cecidomyiidae) field

collections from the southeastern United States to 21 resistance genes in wheat. Journal of economic entomology, 103(6), 2229-2235.

Cartwright, W.B., Caldwell, R.M. and Compton, L.E., 1959. Responses of Resistant and Susceptible Wheats to Hessian Fly Attack 1. Agronomy Journal, 51(9), 529-531.

Chapin, J.W. 2009. Hessian fly: a pest of wheat, triticale, barley and rye. Clemson Ext. Bull. http://www. clemson.edu/extension/rowcrops/small_grains/pdfs/hessian_fly.pdf.

Chen, M.S., Liu, X., Yang, Z., Zhao, H., Shukle, R.H., Stuart, J.J. and Hulbert, S. 2010. Unusual conservation among genes encoding small secreted salivary gland proteins from a gall midge. BMC evolutionary biology, 10(1), p.296.

Chen, M.S., Echegaray, E., Whitworth, R.J., Wang, H., Sloderbeck, P.E., Knutson, A., Giles, K.L. and Royer, T.A. 2009. Virulence analysis of Hessian fly populations from Texas, Oklahoma, and Kansas. Journal of economic entomology, 102(2), 774-780.

Chen, M.S., Zhao, H.X., Zhu, Y.C., Scheffler, B., Liu, X., Liu, X., Hulbert, S. and Stuart, J.J. 2008. Analysis of transcripts and proteins expressed in the salivary glands of Hessian fly (Mayetiola destructor) larvae. Journal of insect physiology, 54(1), 1-16.

Chen, M.S., Fellers, J.P., Zhu, Y.C., Stuart, J.J., Hulbert, S., El-Bouhssini, M. and Liu, X. 2006. A super-family of genes coding for secreted salivary gland proteins from the Hessian fly, Mayetiola destructor. Journal of Insect Science, 6(1).

Chen, M.S., Fellers, J.P., Stuart, J.J., Reese, J.C. and Liu, X. 2004. A group of related cDNAs encoding secreted proteins from Hessian fly [Mayetiola destructor (Say)] salivary glands. Insect Molecular Biology, 13(1), 101-108.

Cook, R.J. and Veseth, R.J.1991. Wheat Health Management. Plant Health Series, USDA

Elzinga, D.A., De Vos, M. and Jander, G. 2014. Suppression of plant defenses by a Myzus persicae (green peach aphid) salivary effector protein. Molecular Plant-Microbe Interactions, 27(7), 747-756.

Flanders, K.L., Buntin, G.D., Mask, P.L. 2008. Biology and management of Hessian fly in wheat. Ala. Coop. Ext. Serv. Bull. ANR-1069. 4 pp.

Gagne´, R.J. 2010. Update for a Catalog of the Cecidomyiidae (Diptera) of the World. Washington, DC: Entomol. Soc. Wash. 544 pp.

Gagne´, R.J. 1994. The Gall Midges of the Neotropical Region. Ithaca, NY: Comstock Publ. Assoc. 352 pp.

Gagné, R.J., Hatchett, J.H., Lhaloui, S., Bouhssini. M. 1991. Hessian fly and barley stem gall midge, two different species of *Mayetiola* (Diptera: Cecidomyiidae) in Morocco. Annals of the Entomological Society of America, 84(4), 436-443.

Gagne´, R.J. 1989. The plant-feeding gall midges of North America. Ithaca (NY): Cornell University Press.

Gagné, R.J. and Hatchett, J.H. 1989. Instars of the Hessian fly (Diptera: Cecidomyiidae). Annals of the Entomological Society of America, 82(1), 73-79.

Gould, F. 1998. Sustainability of transgenic insecticidal cultivars: integrating pest genetics and ecology. Annual review of entomology, 43(1), 701-726.

Harris, M.O, Jacob, J., Brown, P. and Yan, G. 2017. Wheat pests: introduction, rodents and nematodes. 443-466. 10.19103/AS.2016.0004.24.

Harris, M.O., Friesen, T.L., Xu, S.S., Chen, M.S., Giron, D., Stuart, J.J. 2015. Pivoting from Arabidopsis to wheat to understand how agricultural plants integrate responses to biotic stress. J. Exp. Bot. 66 (2): 513-531.

Harris, M.O., Stuart, J.J., Mohan, M., Nair, S., Lamb, R.J. and Rohfritsch, O. 2003. Grasses and gall midges: plant defense and insect adaptation. Annual Review of Entomology, 48(1), 549-577.

Haseman, L. 1930. The Hessian fly larva and its method of taking food. Journal of Economic Entomology, 23(2), 316-321.

Hatchett, J.H., Kreitner, G.L. and Elzinga, R.J. 1990. Larval mouthparts and feeding mechanism of the Hessian fly (Diptera: Cecidomyiidae). Annals of the Entomological Society of America, 83(6), 1137-1147.

ICARDA. 2002. ICARDA Annual Report 2001. International Center for Agricultural Research in the Dry Areas, Aleppo, Syria. Iv,112 pp.

Johnson ,A., Chen,M.S., Morton, P., Cambron ,S. and Shukle, R.H .2009. Analyzing the diversity of secreted salivary gland transcripts in Hessian fly populations from Israel and the United States. NCB-ESA North Central Branch Entomological Society of America.

Johnson, V.A., Briggle, L.W., Axtel, J.D., Bauman, L.F., Leng, E.R. and Johnston, T.H. 1978. Grain crops. *In* M. Milner, N.S. Scrimshaw & D.I.C. Wang, eds. *Protein resources and technology*, p. 239-255. Westport, CT, USA, AVI Publishing.

Joy, J.B. and Crespi, B.J. 2007. Adaptive radiation of gall-inducing insects within a single host-plant species. Evolution, 61(4), 784-795.

Knodel, J.J. and Ganehiarachchi, M. 2016. Integrated pest management of the wheat midge in North Dakota. NDSU Extension Service, North Dakota State University.

Lamb, R.J., Tucker, J.R., Wise, I.L. and Smith, M.A.H. 2000. Trophic interaction between *Sitodiplosis mosellana* (Diptera: Cecidomyiidae) and spring wheat: implications for yield and seed quality. The Canadian Entomologist, 132(5), 607-625.

Lamb, R.J., Wise, I.L., Olfert, O.O., Gavloski, J. and Barker, P.S. 1999. Distribution and seasonal abundance of Sitodiplosis mosellana (Diptera: Cecidomyiidae) in spring wheat. The Canadian Entomologist, 131(3), 387-397.

Lhaloui, S.M. 1995. Biology, host preference, host suitability and plant resistance studies of the Barley midge and Hessian fly (Diptera: Cecidomyiidae) in Morocco. PhD Dissertation, Kansas State University, Manhttan KS, 567pp.

Lhaloui, S.M., Bouhssini, M.E., Amri, A., Miller, R., Kamel, A., Mekni, M. 1988. Premier Rapport sur la Surveillance de la Cecidomyie sur ble, orge, et avoine en Algerie et Tunisie. Annual Research Report, Aridoculture Research Center, Settat, Morocco.

Lhaloui, S.M., Buschman, L., Bouhssini, M., Amri, A., Hatchett, J.H., Keith, D., Starks, K. and El Houssaini, K. 1992. Infestations of Mayetiola s(Diptera: Cecidomyiidae) in bread wheat, durum wheat and barley: results of five annual surveys in the major cereal growing regions of Morocco. Al Awamia, 77, 21-53.

Liu, X., Fellers, J.P., Wilde, G.E., Stuart, J.J. and Chen, M.S. 2004. Characterization of two genes expressed in the salivary glands of the Hessian fly, Mayetiola destructor (Say). Insect biochemistry and molecular biology, 34(3), 229-237.

Maddur, A.A., Liu, X., Zhu, Y.C., Fellers, J.P., Oppert, B., Park, Y., Bai, J., Wilde, G.E. and Chen, M.S. 2006. Cloning and characterization of protease inhibitor-like cDNAs from the Hessian fly Mayetiola destructor (SAY). Insect molecular biology, 15(4), 485-496.

Makni, H., Marrakchi, M. and Pasteur, N. 2000. Biochemical characterization of sibling species in Tunisian Mayetiola (Diptera: Cecidomyiidae). Biochemical systematics and ecology, 28(2), 101-109.

Miles, P.W. 1999. Aphid saliva. Biological Reviews, 74(1), 41-85.

Oakley, J.N., Talbot, G., Dyer, C., Self, M.M., Freer, J.B.S., Angus, W.J., Barrett, J.M., Feuerhelm, G., Snape, J., Sayers, L. and Bruce, T.J.A. 2005. Integrated control of wheat blossom midge: variety choice, use of pheromone traps and treatment thresholds. HGCA Project Report, 363.

Painter, R.H. 1951. Insect resistance in crop plants. The Macmillan Company New York.

Parker, B.L., Bouhssini, M., Skinner, M. 2001. Field Guide: Insect Pests of Wheat and Barley in North Africa, West and Central Asia. International Centre for Agricultural Research in the dry Areas. Aleppo, Syria. 120 pp.

Patterson, F.L., Shaner, G.E., Ohm, H.W. and Foster, J.E. 1990. A historical perspective for the establishment of research goals for wheat improvement. Journal of Production Agriculture, 3(1), 30-38.

Pauly, P.J. 2002. Fighting the Hessian fly. Environmental History, (7), 385-507.

Price, P.W., 2005. Adaptive radiation of gall-inducing insects. Basic and Applied Ecology, 6(5), 413-421.

Refai, F. Y., E. T. Jones and B. S. Miller. 1955. Some biochemical factors involved in the resistance of the wheat plant to attack by the Hessian fly. Cereal Chemistry, (32), 437-451.

Refai, F.Y., Miller, B.S., Jones, E.T. and Wolfe, J.E. 1956. The feeding mechanism of Hessian fly larvae. Journal of Economic Entomology, 49(2), 182-184.

Sinha, D.K., Nagaraju, J., Tomar, A., Bentur, J.S. and Nair, S. 2012. Pyrosequencing-based transcriptome analysis of the Asian rice gall midge reveals differential response during compatible and incompatible interaction. International journal of molecular sciences, 13(10), 13079-13103.

Shukle, R.H., Mittapalli, O., Morton, P.K. and Chen, M.S. 2009. Characterization and expression analysis of a gene encoding a secreted lipase-like protein expressed in the salivary glands of the larval Hessian fly, Mayetiola destructor (Say). Journal of insect physiology, 55(2), 105-112.

Skuhravá, M.1986. Cecidomyiidae. In: Soós A, ed. *Catalogue of Palaearctic Diptera*. Volume 4 Sciaridae - Anisopodidae. Budapest, Hungary: AkadémiaiKiadó.

Smith, C.M. 2005. Plant Resistance to Arthropods - Molecular & Conventional Approaches. Springer, the Netherlands.

Stuart, J.J. and Hatchett, J.H. 1987. Morphogenesis and cytology of the salivary gland of the Hessian fly, Mayetiola destructor (Diptera: Cecidomyiidae). Annals of the Entomological Society of America, 80(4), 475-482.

Stuart, J.J., Chen, M.S., Shukle, R. and Harris, M.O. 2012. Gall midges (Hessian flies) as plant pathogens. Annual review of phytopathology, 50, 339-357.

Stuart, J. 2015. Insect effectors and gene-for-gene interactions with host plants. Current Opinion in Insect Science, 9, 56-61.

Thorpe, P., Cock, P.J. and Bos, J. 2016. Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC genomics, 17(1), p.172.

Torto-Alalibo, T., Collmer, C.W., Lindeberg, M., Bird, D., Collmer, A. and Tyler, B.M. 2009. Common and contrasting themes in host cell-targeted effectors from bacterial, fungal, oomycete and nematode plant symbionts described using the Gene Ontology. BMC microbiology, 9(1), p.S3.

Walton, W.R. 1920. The Hessian Fly and how to Prevent Losses from it (No. 1083). US Department of Agriculture.Farmers Bulletin 1083.

Watson, S. 2005. Hessian ßy problems have been increasing in recent years in the Central Plains. Wheat Farmer/Row Crop Farmer, (9), 4-5.

Wiseman, B.R. and Webster, J.A. 1999. Economic, environmental, and social benefits of resistance in field crops. Proceedings Thomas say publications in entomology **Series**, Entomological Society of America. Lanham, Maryland.

Yukawa, J., Rohfritsch, O. 2005. Biology and ecology of gall inducing Cecidomyiidae (Diptera). In: Raman A, Schaefer CW, Withers TM. (eds). Biology, ecology, and evolution of gall-inducing arthropods. Science Publishers Inc., Enfield (NH), p. 273–304.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R. and Batterton, M. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Current Biology, 25(5), 613-620.

Zhao, C., Shukle, R., Navarro-Escalante, L., Chen, M., Richards, S. and Stuart, J.J. 2016. Avirulence gene mapping in the Hessian fly (Mayetiola destructor) reveals a protein phosphatase 2C effector gene family. Journal of insect physiology, 84, 22-31.

Zhu,Y.C, Liu.X., Maddur.A.A., Oppert.B., Chen,M.S.2005. Cloning and characterization of chymotrypsin and trypsin-like cDNAs from the gut of the Hessian fly [*Mayetiola destructor* (Say)]. Insect biochemistry and molecular biology, 35(1), pp.23-32.

Chapter 2 - Transcriptomic analyses of the secreted proteins from the salivary glands of the wheat midge (*Sitodiplosis mosellana*) larvae

Abstract

Both the wheat midge (*Sitodiplosis mosellana*) and the Hessian fly (*Mayetiola destructor*) belong to a group of insects called gall midges (Diptera: Cecidomyiidae) and both are destructive pests of wheat. From Hessian fly larvae, a large number of genes have been identified to encode Secreted Salivary Gland proteins (SSGPs), which are presumably critical for the insect to feed on and manipulate host plants. For comparison, we conducted an analysis on transcripts encoding SSGPs from the first instar larvae of the wheat midge. A total of 3,500 cDNA clones were sequenced, from which 1,301 high quality sequences were obtained. Approximately 25% of the cDNAs with high quality sequences encoded SSGPs. The SSGPs were grouped into 97 groups based on sequence homology. Among the SSGP-encoding transcripts, 206 encoded unique proteins with no sequence similarity to any known protein and 29 encoded proteins similar to known proteins including proteases, serpines, thioesterases, ankryins, and feritins. Most (~80%) SSGP-encoding genes appear under strong selection for mutations that generate amino acid changes within the coding region. Identification and characterization of SSGPs in wheat midge larvae provide a foundation for future work to reveal molecular mechanisms behind wheat midge - wheat interactions and the role of these putative effector proteins in insect virulence. Availability of the SSGP transcripts will also facilitate comparative analyses of insect effectors from related species.

Introduction

The orange wheat blossom midge (wheat midge), *Sitodiplosis mosellana* (Géhin) (Diptera: Cecidomyiidae), is one of the most destructive pests of wheat in the northern hemisphere (Berzonsky et al. 2003; Doane and Olfert 2008). In the USA, wheat midge outbreaks have been recorded on spring wheat in the northern states of Minnesota, Montana, and North Dakota. In the mid-1990s, spring wheat losses were estimated at more than $27 million in North Dakota (Knodel and Ganehiarachchi 2016). Wheat midge larvae can feed on developing seeds of both bread and durum wheat (Ding et al. 2000; Harris et al. 2003). Wheat midges have four life stages: egg, larva, pupa, and adult. Females lay eggs on the surface of wheat heads. Newly hatched larvae feed on developing kernels for two to three weeks. The first two instars are the damaging stages. Mature third instars drop from wheat heads after rain or heavy dew in August to move into 2-4 inches deep in soil for overwintering. In unfavorable conditions, larvae can remain dormant and survive in cocoons for more than 10 years (Harris et al. 2003).

Very little is known regarding detailed feeding mechanisms of wheat midge larvae. Like other gall midges, wheat midge larvae are thought to inject saliva into wheat developing seeds, resulting in shriveled wheat kernels (Lamb et al. 2000). Many plant pathogens and parasitic insects possess an effector-based mechanism to attack host plants and promote virulence via secreted effector proteins (Shorthouse and Rohfritsch 1992; Miles 1999; Harris et al. 2015; Toruno et al. 2016). In plant - insect systems that have gene-for-gene interactions (namely for every resistance gene in host plants, there is a corresponding avirulence gene in the insect), many effectors from parasitic insects have been identified and characterized (Harris et al. 2015; Stuart 2015). For instance, in the Hessian fly, a large number of Secreted Salivary Gland Proteins (SSGPs) have been identified and many of them are likely to play effector roles once injected

into plant tissues. Approximately 60% of transcripts in salivary glands of Hessian fly first instar larvae encode SSGPs (Chen et al. 2008). Later, genomic sequencing revealed more than 7% of predicted genes in the Hessian fly genome encode effector-like proteins (Zhao et al. 2015). Several avirulence effectors have been cloned from the Hessian fly and all of them were SSGPs (Aggarwal et al. 2014; Zhao et al. 2015, 2016). Many secreted proteins have also been identified in the saliva of several aphid species (Thorpe et al. 2016), and many of these aphid proteins act as effectors either to suppress or trigger plant defense responses (Elzinga et al. 2014). Therefore, identification of SSGPs from insects provides an efficient way to identify putative effectors of insect species.

Whether the interaction between wheat midge and wheat follows a gene-for-gene model remains to be investigated. However, a highly effective resistance gene, named *Sm*1, to the wheat midge was discovered in winter wheat genotype in 1996 (Barker and McKenzie 1996), and wheat cultivars with *Sm*1 can significantly limit kernel damage and yield loss (Smith et al. 2014; Blake et al. 2014). The existence of a major resistance gene in wheat suggests that a gene-for-gene relationship is possible in the wheat midge - wheat interaction. Two groups of SSGPs in wheat midge larvae have been reported previously (Chen et al. 2010). However, large scale identification of SSGP-encoding genes in the wheat midge has not been conducted. The objective of this study is to conduct a more extensive analysis of SSGPs from dissected salivary glands of first instar wheat midge larvae via a transcriptomic approach.

Materials and Methods

Insects and salivary gland preparation

The insect population used in this research was derived from a colony consisting approximately 20,000 individuals collected from Divide County in North Dakota in 2013. The colony has been maintained in greenhouse in North Dakota State University since then. Seeds of Roblin hard red spring wheat, an early maturing Canadian variety that is susceptible to wheat midge, were planted in a greenhouse at North Dakota State University in Fargo, ND, to rear the wheat midge. Wheat plants were maintained at 20°C with 18:6h light/dark cycles. Meanwhile, dormant pupae at 4C have been placed at room temperature to breakdown the dormancy stage for adult's emergence. When wheat plants were at Zadok's growth stage 55-59 (the inflorescence is half or more emerged from the sheath), two or more gravid females were placed into a glass cylinder covering an individual wheat head. After 24 hours of exposure to the wheat midge females, the glass cylinder was removed and the head was covered with a glassine pollination bag to help protect the eggs from desiccating as they develop. Egg hatch and larval migration to their larval feeding sites on the surface of the developing seed occurs three days after oviposition. For RNA analysis, 3 to 4 day old wheat midge larvae were collected from wheat heads with the aid of an 20X dissecting microscope. Salivary glands were obtained by dissecting first instar larvae in saline buffer. Dissection was achieved by pulling away the anterior tip of a larva with a pair of forceps while holding the posterior end of the larva steady with another pair of forceps. The salivary glands of the larva move out of the cascade during this process along with other mouthpart tissues. Clean salivary glands were then obtained by removing unwanted mouthpart tissues. For RNA analysis and cDNA library construction, the dissected glands were

transferred into TRI reagent™ (Molecular Research, Inc., Cincinnati, OH) and frozen in liquid nitrogen as soon as they were obtained.

cDNA library construction and sequencing

Total RNA was isolated from 300 pairs of salivary glands using TRI reagent™ following the protocol provided by the manufacturer. RNA quality and integrity were assessed using a Bioanalyzer (Agilent Technologies, Santa Clara, CA). cDNA libraries were constructed using a 'SMART™' library construction kit from Clontech (Palo Alto, CA) as described by Chen et al. (2004). Briefly, cDNA inserts were ligated into the pPCRXL-TOPO plasmid contained in a TOPO TA cloning kit (Invitrogen, Carlsbad, CA) instead of a phage vector. Individual clones were picked up for plasmid DNA isolation, which were sequenced with the M13 forward and reverse primers following the Sanger DNA sequencing method via a commercial contract (GENEWIZ, South Plainfield, NJ).

Sequence analysis

Vector sequences were trimmed from raw reads after cDNA clones were sequenced. Sequences from sense- and antisense-directions were aligned to examine if a clone was sequenced fully from both directions. If no overlap was found between the sense and antisense reads, new primers were synthesized for further sequencing. Cluster analyses of cDNAs were conducted using BlastStation-Local 64 program. We have identified unigenes and groups in our analysis. Each unigene represents a unique structure of protein which may contains multiple redundant clones, and then unigenes were sorted in groups based on sequence similarity.

Open reading frames (ORF) were identified using the ORF finder. Sequence alignment and similarity analysis were performed using various BLAST programs (http://www.ncbi.nlm.

nih.gov/). Initial database search was conducted with BLASTN and BLASTX. Sequence alignments with E-values greater than 10^3 were considered to have no meaningful sequence similarity between the two sequences. Sequence alignments with E-values smaller than 10^{10} were considered that two sequences share significant similarity. Sequence alignments with E-values between 10^3 and 10^{10} were further examined individually to determine if two sequences share similarity based on the length and gaps of the alignments. Analysis for secretion signal peptides was carried out using the SignalP v4.1 (Center for Biological Sequence Analysis, Technical University of Denmark, http:// www.cbs.dtu.dk/services/SignalP/).

Calculation of synonymous and nonsynonymous mutation rates

The percentages of synonymous and nonsynonymous mutations were calculated based on sequence alignments of members within a group. For example, the percentages of nonsynonymous mutations were derived by dividing the number of nonsynonymous mutations by the number of total mutations among group members. If there are multiple members that share the same mutation at the same position, that mutation counted only once. However, if different members have two or more different mutations at the same position, then the mutation was counted as two or more.

Results

Composition of transcripts obtained from dissected salivary glands

A total of 3523 cDNA clones were sequenced. After removing clones with small inserts and bad quality sequences, 1,301 cDNA sequences were retained. Among these cDNAs, 330 (25.3%) encode SSGPs and the remaining 971 (74.6%) encode proteins without a typical secretion signal peptide. Among the SSGP-encoding cDNAs, 235 encoded unique proteins with no sequence similarity to any known sequences in Genbank, whereas 33 encoded proteins with

sequence similarity to known proteins such as carboxypeptidases, peptidases, lysosomal thioesterases, serpins, ankyrins, and ferritins (Appendix A- Table S1).

Among the 971 non-SSGP transcripts, 295 (30.4%) encode proteins with no meaningful (E-values greater than 10^3) sequence similarity to any proteins in Genbank, 321 (33.1%) encode proteins with sequence similarity to proteins with unknown function, and the remaining 353 (36.4%) encode proteins with sequence similarity to proteins with various functions. For the transcripts encoding known proteins, 153 (43.1%) are proteins with functions in protein synthesis and the remaining 205 (57.7%) with other house-keeping functions, including energy-metabolic enzymes, structural proteins, transporters, and others (Appendix A-Table S2).

SSGP classification

SSGP-encoding transcripts were sorted into 97 groups according to the sequence similarity among the cDNAs and derived proteins (Appendix A- Table S1). Among the 97 groups, 66 have either a single clone or multiple clones that encode the same protein. The remaining 31 groups have multiple clones that encode at least two different proteins. Proteins within a group share at least 30% amino acid identity and have a highly conserved secretion signal peptide. Proteins between different groups share no meaningful (E>10^3) sequence similarity and have a completely different secretion signal peptide (Genbank accession numbers for the ESTs from JZ971833 to JZ972161). Figure 1 shows amino acid sequence alignments of two representative groups. Both groups have a highly conserved signal peptide and a more diversified mature protein. The overall conservation among group members particularly in the signal peptide region suggest that the transcripts within a group may have been derived from genes that share the same evolutionary origin and, therefore, can be considered the same gene

family (Figure 2.1). Some sequence variation may have also resulted from different alleles of the same gene.

Sequence variations among group members

Group members among those with significant sequence variations were divided intomature protein (MP)-coding region, signal peptide (SP)-coding region, and non-coding regions, and percentages of nucleotides with sequence variation in each region were analyzed (Table 2.1). Among these three regions, sequence variation in the SP-coding region was the lowest except group 24, likely due to the functional constraint of the secretion role of signal peptides. Variation rates in MP-coding region and non-coding regions were much higher (Appendix A- Table S3, Figure S2). To examine if group members were under selection pressure for diversification, the percentages of non-synonymous and synonymous mutations in the MP-coding region were also analyzed. Over 70% of nucleotide substitutions were non-synonymous (Table 2.1).

Discussion

Many insects inject effectors into host tissues to manipulate plants including suppressing host defense, inhibiting plant growth, and reprogramming plant metabolism (Stuart 2015). Some insects also inject effectors into host tissues for pre-digesting food before ingestion and for various other functions (Miles 1999; Harris et al. 2015). The salivary glands of insects are the main tissue to produce effector proteins for host injection. Therefore, analyzing transcripts in the salivary glands and identifying those proteins with a secretion signal peptide is an efficient way to identify putative effector proteins (Chen et al. 2004, 2008). In this study, we analyzed the composition of transcripts in salivary glands of the first instars of the wheat midge through a traditional Sanger sequencing approach. There are two reasons to follow a traditional sequencing

approach in this study. First, previous studies have shown that effector genes from gall midges are conserved unconventionally (Chen et al. 2010), which would cause problems in correctly assembling short sequence reads from high throughput sequencing. Second, the wheat midge is an understudied species genomically and may be difficult to annotate small transcript fragments.

Our analysis resulted in the identification of 97 groups of transcripts encoding SSGPs. Among these groups, 64% (62 groups) are singletons, indicating that our analysis is very preliminary and further sequencing more clones is likely to identify much more unique SSGP transcripts. The most abundant group is group 1, which has 48 unique transcripts (99 including redundant sequences). SSGP proteins encoded by group 1 transcripts share no sequence similarity with any known sequences in Genbank, and therefore, the functions of this group of genes remain to be determined. The fact that members among this group have been under strong positive selection indicates that this group of genes are likely to play important roles in the wheat midge-wheat interaction. Other abundant transcript groups include group 2, group 3, group 4, group 12, group 13, group 24, group 29, group 40, group 45, and group 67 (Appendix A- Table S1)

There are commonalities and differences between the putative SSGPs from wheat midge larvae and those from Hessian fly larvae, a species that has been studied more extensively for SSGP-encoding genes (Chen et al. 2004, 2008, 2010; Zhao et al. 2015, 2016). A commonality is that most of the SSGPs are small peptides (50 - 150 amino acid residues), and those small SSGPs share no sequence similarity with any known proteins in available databases. In addition, SSGP-encoding genes from both the wheat midge and Hessian fly appear to be under strong diversifying selection pressure. Evidence for this is the fact that over 70% of point mutations among group members are nonsynonymous (Table 2.1). A similar phenomenon was also found

in Hessian fly SSGP-encoding genes, where over 80% of point mutations among group members were nonsynonymous (Chen et al. 2004). The fast-evolving nature of SSGP-encoding genes in both insect species is another indicator that these genes are involved in interactions with their host plants (Thompson 1998). There is no sequence similarity between SSGPs from wheat midge larvae and those from Hessian fly larvae, suggesting that SSGPs from these two insect species perform different biochemical functions and have different mechanisms to manipulate host plants. In addition, many SSGP-encoding genes from Hessian fly exhibit an unconventional conservation pattern, in which the 5'- and 3'-non-coding regions and introns are highly conserved, whereas the regions encoding mature proteins are highly diversified (Chen et al. 2010; Zhao et al. 2015). No such unconventional conservation pattern was found among group members of SSGP-encoding genes from the wheat midge.

In addition to small SSGPs, there are a few transcripts that encode secreted proteins with sequence similarity to known proteins, which include proteases, protease inhibitors, lysosomal thioesterases, ankyrins, and ferritins. Whether these proteins are injected into host plants or secreted into body fluid of the insect remains to be determined. Proteases and protease inhibitors have also been found in saliva from other insect species (Miles 1999; Chen et al. 2008; Liu et al. 2016). Proteases could act as digestive enzymes for pre-oral digestion of food before ingestion, whereas protease inhibitors could neutralize defense proteases from host plants (Pechan et al. 2002). Lysosomal thioesterases, ankyrins, and ferritins play house keeping functions inside insects. However, some proteins with house-keeping functions in insects can also be injected into host plants and play effector roles in insect - plant interactions (Miles 1999).

In summary, we have conducted a global analysis on genes expressed in the salivary glands of first instars of the wheat midge for the first time and identified numerous genes

encoding SSGPs. The availability of the putative effector genes provides a foundation for further research to characterize the roles of these genes in wheat midge and wheat interactions. For example, the cDNAs could be used to produce recombinant proteins for various biochemical assays, or for antibody production to analyze tissue distribution both within the insect bodies or host tissues if they are injected into plants during feeding. The availability of these genes is also useful for comparative analysis of salivary proteins from different insect species.

References

Aggarwal, R., Subramanyam, S., Zhao, C., Chen, M.S., Harris, M.O. and Stuart, J.J. 2014. Avirulence effector discovery in a plant galling and plant parasitic arthropod, the Hessian fly (Mayetiola destructor). PLoS One, 9(6), p.e100958.

Barker, P.S. and McKenzie, R.I.H. 1996. Possible sources of resistance to the wheat midge in wheat. Canadian Journal of Plant Science, 76(4), pp.689-695.

Berzonsky, W.A., Ding, H., Haley, S.D., Harris, M.O., Lamb, R.J., McKenzie, R.I.H., Ohm, H.W., Patterson, F.L., Peairs, F.B., Porter, D.R. and Ratcliffe, R.H. 2003. Breeding wheat for resistance to insects. Plant Breeding Reviews, (22), 221-296.

Blake, N.K., Stougaard, R.N., Bohannon, B., Weaver, D.K., Heo, H.Y., Lamb, P.F., Nash, D., Wichman, D.M., Kephart, K.D., Miller, J.H. and Reddy, G.V.P. 2014. Registration of 'Egan'wheat with resistance to orange wheat blossom midge. Journal of Plant Registrations, 8(3), 298-302.

Chen, M.S., Liu, X., Yang, Z., Zhao, H., Shukle, R.H., Stuart, J.J. and Hulbert, S. 2010. Unusual conservation among genes encoding small secreted salivary gland proteins from a gall midge. BMC evolutionary biology, 10(1), p.296.

Chen, M.S., Zhao, H.X., Zhu, Y.C., Scheffler, B., Liu, X., Liu, X., Hulbert, S. and Stuart, J.J. 2008. Analysis of transcripts and proteins expressed in the salivary glands of Hessian fly (Mayetiola destructor) larvae. Journal of insect physiology, 54(1), 1-16.

Chen, M.S., Fellers, J.P., Stuart, J.J., Reese, J.C. and Liu, X. 2004. A group of related cDNAs encoding secreted proteins from Hessian fly [Mayetiola destructor (Say)] salivary glands. Insect Molecular Biology, 13(1), 101-108.

Ding, H., Lamb, R.J. and Ames, N. 2000. Inducible production of phenolic acids in wheat and antibiotic resistance to Sitodiplosis mosellana. Journal of Chemical Ecology, 26(4), pp.969-985.

Doane, J.F. and Olfert, O. 2008. Seasonal development of wheat midge, Sitodiplosis mosellana (Géhin)(Diptera: Cecidomyiidae), in Saskatchewan, Canada. Crop protection, 27(6), pp.951-958.

Elzinga, D.A., De Vos, M. and Jander, G. 2014. Suppression of plant defenses by a Myzus persicae (green peach aphid) salivary effector protein. Molecular Plant-Microbe Interactions, 27(7), pp.747-756.

Harris, M.O., Friesen, T.L., Xu, S.S., Chen, M.S., Giron, D., Stuart, J.J. 2015. Pivoting from Arabidopsis to wheat to understand how agricultural plants integrate responses to biotic stress. J. Exp. Bot. 66 (2): 513-531.

Harris, M.O., Stuart, J.J., Mohan, M., Nair, S., Lamb, R.J. and Rohfritsch, O. 2003. Grasses and gall midges: plant defense and insect adaptation. Annual Review of Entomology, 48(1), 549-577.

Knodel, J.J. and Ganehiarachchi, M. 2016. Integrated pest management of the wheat midge in North Dakota. NDSU Extension Service, North Dakota State University.

Lamb, R.J., Tucker, J.R., Wise, I.L. and Smith, M.A.H. 2000. Trophic interaction between Sitodiplosis mosellana (Diptera: Cecidomyiidae) and spring wheat: implications for yield and seed quality. The Canadian Entomologist, 132(5), 607-625.

Liu, X., Zhou, H., Zhao, J., Hua, H., and He, Y. 2016. Identification of the secreted watery saliva proteins of the rice brown planthopper, Nilaparvata lugens (Stål) by transcriptome and Shotgun LC–MS/MS approach. Journal of insect physiology, (89), 60-69.

McKenzie, R.I.H., Lamb, R.J., Aung, T., Wise, I.L., Barker, P., Olfert, O.O. and McIntosh, R.I., 2002. Inheritance of resistance to wheat midge, Sitodiplosis mosellana, in spring wheat. Plant Breeding, 121(5), pp.383-388.

Miles, P.W. 1999. Aphid saliva. Biological Reviews, 74(1), 41-85.

Pechan, T., Cohen, A., Williams, W.P. and Luthe, D.S. 2002. Insect feeding mobilizes a unique plant defense protease that disrupts the peritrophic matrix of caterpillars. Proceedings of the National Academy of Sciences, 99(20), 13319-13323.

Smith, M.A.H., Wise, I.L., Fox, S.L., Vera, C.L., DePauw, R.M. and Lukow, O.M. 2014. Seed damage and sources of yield loss by Sitodiplosis mosellana (Diptera: Cecidomyiidae) in resistant wheat varietal blends relative to susceptible wheat cultivars in western Canada. The Canadian Entomologist, 146(3), 335-346.

Smith, M.A.H., Lamb, R.J., Wise, I.L. and Olfert, O.O.2004. An interspersed refuge for Sitodiplosis mosellana (Diptera: Cecidomyiidae) and a biocontrol agent Macroglenes penetrans (Hymenoptera: Pteromalidae) to manage crop resistance in wheat. Bulletin of entomological research, 94(2), 179-188.

Shorthouse, J.D. and Rohfritsch, D. 1992. Biology of insect-induced galls. New York : Oxford University Press.

Stuart, J. 2015. Insect effectors and gene-for-gene interactions with host plants. Current Opinion in Insect Science, 9, 56-61.

Thompson, J.N. 1998. Rapid evolution as an ecological process. Trends in ecology & evolution, 13(8),329-332.

Thorpe, P., Cock, P.J. and Bos, J. 2016. Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC genomics, 17(1), p.172.

Toruno, T.Y., Stergiopoulos, I., Coaker, G. 2016. Plant-pathogen effectors: cellular probes interfering with plant defenses in spatial and temporal manners. Annual review of phytopathology, 54, pp.419-441.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R. and Batterton, M. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Current Biology, 25(5), 613-620.

Zhao, C., Shukle, R., Navarro-Escalante, L., Chen, M., Richards, S. and Stuart, J.J. 2016. Avirulence gene mapping in the Hessian fly (Mayetiola destructor) reveals a protein phosphatase 2C effector gene family. Journal of insect physiology, 84, 22-31.

Figure 2.1 Amino acid alignments of two representative groups A and B. The boundary between predicted signal peptide and mature proteins is indicated by an arrow.

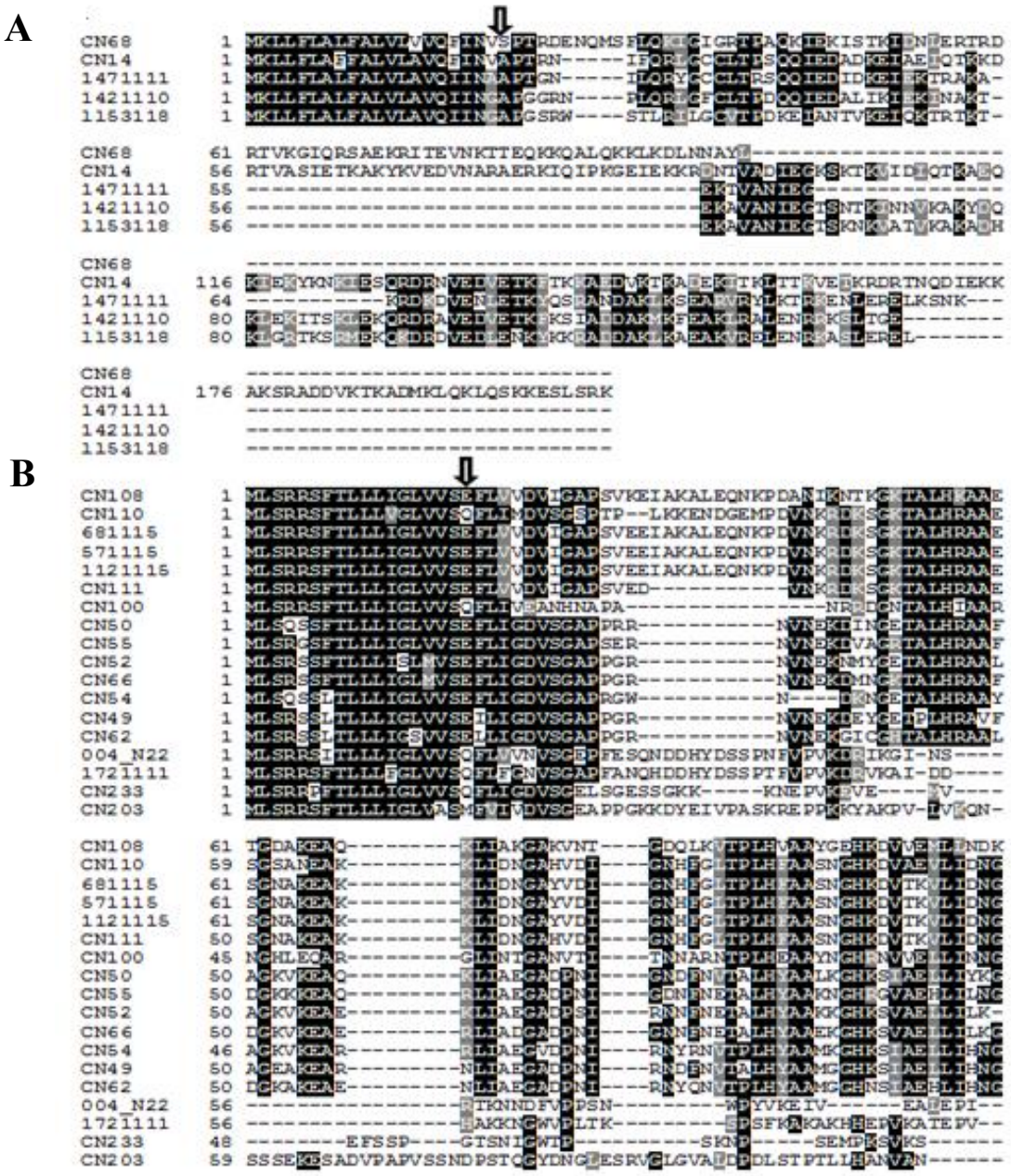


Table 2.1 Analysis of sequence variation among group or sub-group members from wheat midge. MP - Mature Protein, SP - Signal Peptide

Group or sub-group	% Nucleotides with mutations among group or sub-group members			% non-synonymous Mutations in MP coding
	MP-coding	SP-coding	Non-coding	
Group 1, sub-group 1	26.5	3.3	35.0	77.5
Group 1, sub-group 2	17.0	8.3	20.4	80.5
Group 24	7.2	4.2	0	82.6
Group 29	14.3	9.3	9.7	88.9
Group 40	11.9	1.6	11.7	71.9

Chapter 3 - Comparative transcriptomic analyses of secreted proteins from the salivary glands of three related stem gall midge species

Abstract

The stem gall midge genus *Mayetiola*, Cecidomyiidae, causes serious damage to small grain crops. Larvae of these species inject their saliva into host plants to reprogram plant metabolic pathways, suppress host defenses, and induce the formation of galls. In this study, an analysis on transcripts encoding **S**ecreted **S**alivary **G**land **P**roteins (SSGPs), the putative effector proteins, from the first instar larvae of the barley midge (*M. hordei*) and oat midge (*M. avenae*) was conducted. A total of 2,570 cDNA clones were sequenced from barley midge and 743 high quality cDNA sequences were retained. The analysis identified 458 cDNA clones encoding SSGPs. Among the SSGP-encoding transcripts, 178 encoded unique proteins (also called unigenes). Transcripts encoding SSGPs were divided into 51 groups based on sequence homology. A total of 3,226 cDNA clones were sequenced from the oat midge and 718 high quality cDNA sequences were retained. The analysis identified 450 cDNA clones encoding SSGPs, and among the SSGP-encoding transcripts, 194 are unigenes. Transcripts encoding SSGPs were sorted into 50 groups. Overall, more than 60% of the cDNAs from the salivary gland of these species encoded SSGPs. The composition of SSGP transcripts from the barley and oat midges were then compared with that of the Hessian fly (*Mayetiola destructor*), which has been studied previously. Comparative analyses identified five groups with 102 (57.3%) unigenes from barley midges and seven groups with 107 (55.1%) unigenes from the oat midge encode SSGPs that are conserved among the three midge species. Among the conserved groups are members belonging to some of the well characterized families from Hessian fly, including SSGP

family one (SSGP-1), the most abundant group in all three gall midges; SSGP family seventy-one (SSGP-71), the largest arthropod family with effectors mimicking plant pathogenic bacteria; SSGP family eleven (SSGP-11); and SSGP family four (SSGP-4). Conserved effectors among the three midges indicate conserved functions, which are most likely involved in plant manipulation.

There were also eight conserved groups between only barley midge and oat midge, with 19 (10.7%) unigenes from the barley midge and 25 (12.9%) unigenes from the oat midge were found with no corresponding homologues in the Hessian fly. The remaining unigenes encode SSGPs that are unique to different midge species. These divergent SSGP groups with no homology among the three species may play roles in host specification. Identification of conserved and divergent putative effectors among these three species may reveal the potential roles of these effectors in insect virulence and host specification.

Introduction

Cereal crops are attacked by many herbivorous insects (Johnson et al. 1978). Among these insect pests are stem gall midge species from the genus *Mayetiola,* including the Hessian fly (*Mayetiola destructor* Say), the barley stem gall midge (barley midge, *M. hordei* Keiffer), and the oat stem midge (oat midge, *M. avenae* Marchal). These species are major insects causing severe damage yearly to their host plants in specific regions or worldwide (Gagne´ 2010).

Gall midges have four life stages: eggs, larvae, pupae, and adults. The Hessian fly, barley midge, and oat midge cause damage to host plants with the first two instar larvae (also called maggots) whereas third instar larvae, pupae, and adults do not feed. First instar larvae attack plants by injecting salivary secretions into plant tissues to establish a permanent feeding site. Second instar larvae feed vigorously on plant tissues at the feeding site established by the first

instar. The third instar becomes a non-feeding puparium stage. The first instar is the most critical stage since it paves the way for the second instar larvae to feed. Larvae of the three gall midges are sessile and they feed by sucking up liquid from the nutritive tissue (the gall) (Stuart et al. 2012). The adult stage of the life cycle is short (surviving for 1-3 days), and is specialized only for reproduction.

Morphologically, the three gall midge species are very similar at all four developmental stages. However, each species has its own preferred host plants. The preferred host for Hessian fly is wheat, although it can also survive on certain cultivars of barley and oat if no wheat plants are available (Gagné et al. 1991). Previous reports suggested that the preferred host for barley midge is barley, but it has also been recorded on wheat, oat, and rye (Gagné et al. 1991). The preferred host of oat midge is oat, but it can also occur on the wild plant *Avena fatua*, and wheat (Barnes 1956). However, in the field, the barley midge survives and develops well only on barley (Lhaloui 1995). Similarly, the oat midge can feed and survive well on oat, but not on wheat or barley (Barnes 1956).

Although each gall midge having its own preferred host plants, the three species cause almost identical symptoms to infested plants. Infested susceptible plants are stunted due to irreversible inhibition of plant growth, the color of leaves become dark green due to increased chlorophyll, and the attacked plant will eventually die if no new tillers are produced. The molecular mechanisms required for the three midges to parasitize different host species, yet to cause essential identical symptoms to each host, remain to be delineated.

Interactions between the Hessian fly and wheat have been studied relatively extensively due to the importance of SSGPs in Hessian fly virulence and biotype differentiation. Previous research efforts have revealed that there are hundreds of the SSGPs produced in the salivary

glands (Liu et al. 2004; Chen et al. 2004; Chen et al. 2008). Expanded transcriptomic analyses have revealed a high proportion (about 60%) of transcripts in salivary glands of first instar larvae that encode SSGPs (Chen et al. 2008). Furthermore, genomic sequencing revealed more than 7% of predicted genes in the Hessian fly genome encode effector-like proteins (Zhao et al. 2015). The great diversity in Hessian fly SSGPs has been indicated in many studies (Chen et al. 2008, 2010; Johnson et al. 2009). A large number of Hessian fly SSGPs contain consensus characteristics of parasite effectors (Zhao et al. 2015). In addition, several avirulence effectors have been cloned from the Hessian fly and all of them are SSGPs (Aggarwal et al. 2014; Zhao et al. 2015, 2016).

Thus, and based on the fact that the Hessian fly, barley midge, and oat midge have the same feeding mechanism and result in the same symptom on host plants, yet each species attacks only certain host plants and can't survive on the others. The hypotheses for this study were: 1) some conserved effectors are injected into host tissues to manipulate host plants resulting in the same symptoms among the three midges, including inhibition of plant growth, suppression of host defense, and inducing the formation of nutritive tissues at the feeding site; and 2) some unique effectors in each of the three midges allow each species to infest different host plants to suppress different defense mechanisms from different host plants and to utilize different nutrients. The objectives of this study are to identify SSGP-encoding transcripts from the first instar larvae of barley midge, to identify SSGP-encoding transcripts from first instar larvae of oat midge, and to compare these SSGPs among the Hessian fly, barley midge, and oat midge for identification of common and unique putative effectors among these three species.

Materials and Methods

Insect rearing conditions & Salivary gland dissection

Salivary gland preparation was conducted in the Entomological Research Laboratory at the International Center for Agricultural Research in the Dry Area (ICARDA) in Rabat, Morocco. Three hundred pairs of salivary glands of barley midge and oat midge were collected in TRI reagent™ (Molecular Research, Inc., Cincinnati, OH) from first instar (3days old) larvae of each species, and the dissected glands were frozen in liquid nitrogen as soon as they were obtained. The samples were then shipped to the USDA Hessian fly research laboratory at Manhattan, Kansas, USA in 2016, for further processing and analyses.

RNA extraction, cDNA library construction, and DNA sequencing

Total RNA was isolated from 300 pairs of salivary glands using TRI reagent™ following the protocol provided by the manufacturer. RNA quality and integrity were assessed using a TapeStation Bioanalyzer (Agilent Technologies, Santa Clara, CA). Then the RNA samples were reverse-transcribed to cDNAs. The cDNA samples were amplified using a 'SMART™' library construction kit from Clontech (Palo Alto, CA) as described by Chen et al. (2004). Briefly, amplified cDNA inserts were ligated into the pPCRXL-TOPO plasmid contained in a TOPO TA cloning kit (Invitrogen, Carlsbad, CA). The ligated plasmids were then transformed into individual bacteria. Bacterial clones were picked up individually for plasmid DNA isolation, which were sequenced with the M13 forward and reverse primers following the Sanger DNA sequencing method via a commercial contract (GENEWIZ, South Plainfield, NJ).

Sequence analysis

Vector sequences were trimmed manually from raw reads after cDNA clones were sequenced. Sequences from sense- and antisense-directions were aligned to examine if a clone was sequenced fully from both directions using pairwise alignment, Blast tool in NCBI website. If no overlap was found between the sense and antisense reads, new primers were synthesized for further sequencing. In our analysis, we have identified unigenes and groups. Each unigene represents a unique structure of protein which may contains multiple redundant clones, and then unigenes were sorted in groups based on sequence similarity.

A local database was established for the cDNAs' sequences from barley and oat midges, separately. Cluster analyses of these cDNAs were conducted using BlastStation-Local 64 program (https://www.blaststation.com/intl/en/local64.php). The clustered groups were verified further by using a multiple alignment tool from the website https://www.ebi.ac.uk/Tools/msa/clustalo/, and by generating maximum likelihood phylogenetic tree for representative sequences of these groups using the MEGA 5.2.2 program. Sequence alignment and similarity analysis were performed using various BLAST programs (http://www.ncbi.nlm. nih.gov/). Initial database search was conducted with BLASTN and BLASTX. Sequence alignments with E-values greater than 10^3 were considered to have no meaningful sequence similarity between the two sequences. Sequence alignments with E-values smaller than 10^{10} were considered that two sequences share significant similarity. Sequence alignments with E-values between 10^3 and 10^{10} were further examined individually to determine if two sequences share similarity based on the length of regions with sequence similarity and gaps of the alignments. In order to identify the clustered groups and searching for the secreted signal peptide, open reading frames (ORF) were identified using the ORF finder

(https://www.ncbi.nlm.nih.gov/orffinder/). Analysis for secretion signal peptides was carried out using the SignalP v4.1 (Center for Biological Sequence Analysis, Technical University of Denmark, http:// www.cbs.dtu.dk/services/SignalP/).

Calculation of synonymous and nonsynonymous mutation rates

The percentages of synonymous and non-synonymous mutations were calculated based on sequence alignments of members within a group. For example, the percentages of nonsynonymous mutations were derived by dividing the number of nonsynonymous mutations by the number of total mutations among group members. If there are multiple members that shared the same mutation at the same position, that mutation counted only once. However, if different members have two or more different mutations at the same position, then the mutations at the same location were counted as two or more.

Results

Composition and classification of transcripts from barley midge

A total of 2570 cDNA clones were sequenced. After removing clones with small inserts and bad quality sequences, 743 cDNA sequences were retained. Among these cDNAs, 458 (61.6%) encode SSGPs and the remaining 285 (38.4%) encode proteins without a typical secretion signal peptide. The 458 SSGP-encoding cDNAs were sorted into 178 unigenes based on sequence similarity, and each unigene encoded a unique protein. Among the 178 unigenes, 102 (57.3%) had sequence similarity to SSGPs from Hessian fly, 19 (10.7%) had sequence similarity to SSGPs from oat midge, and 57 (32%) with no sequence similarity to any known sequences in Genbank (Appendix B- Table S1).

Among the 285 non-SSGP transcripts, 94 (33%) encoded proteins with no sequence similarity to any proteins in Genbank, 58 (20.3%) encoded proteins with sequence similarity to

proteins with unknown function, and the remaining 133 (46.7%) encoded proteins with sequence similarity to proteins with various functions. For the transcripts encoding known proteins, 58 (43.6%) were proteins with functions in protein synthesis and the remaining 75 (56.4%) had other house-keeping functions, including energy-metabolic enzymes, structural proteins, and transporters (Appendix B- Table S2).

SSGP-encoding transcripts were classified into 51 groups according to sequence similarity among the cDNAs and derived proteins (Appendix B- Table S1). Among the 51 groups, 34 had either a single clone or multiple clones that encode the same protein. The remaining 17 groups had multiple clones that encode at least two different proteins. Proteins within a group shared at least 30% amino acid identity and had a highly conserved secretion signal peptide. Proteins between different groups shared no meaningful (E>10^3) sequence similarity and had a completely different secretion signal peptide (Genbank accession numbers for the ESTs would be available upon publication). Figure 3.1 shows amino acid sequence alignments of two representative groups. Both groups have a highly conserved signal peptide and a more diversified mature protein. The overall conservation among group members particularly, in the signal peptide region, suggests that the transcripts within a group may have been derived from genes that share the same evolutionary origin and, therefore, can be considered the same gene family. Some sequence variation may have also resulted from different alleles of the same gene (Figure 3.1).

Composition and classification of transcripts from oat midge

A total of 3,226 cDNA clones were sequenced from oat midge. After removing clones with small inserts and bad quality sequences, 718 cDNA sequences were retained. Among these

cDNAs, 450 (62.7%) encode SSGPs and the remaining 268 (37.3%) encode proteins without a typical secretion signal peptide. The 450 SSGP-encoding cDNAs were sorted into 194 unigenes based on sequence similarity, and each unigene encodes a unique protein. Among the 194 groups, 107 (55.1%) had sequence similarity to SSGPs from Hessian fly, 25 (12.9%) had sequence similarity to SSGPs from barley midge, and 62 (32%) had no sequence similarity to any known sequences in Genbank (Appendix B- Table S4).

Among the 268 non-SSGP transcripts, 105 (39.2%) encoded proteins with no sequence similarity to any proteins in Genbank, 55 (20.5%) encoded proteins with sequence similarity to proteins with unknown function, and the remaining 108 (40.3%) encoded proteins with sequence similarity to proteins with various functions. For the transcripts encoding known proteins, 63 (58%) function in protein synthesis and the remaining 45 (42%) had other house-keeping functions, including energy-metabolic enzymes, structural proteins, and transporters (Appendix B- Table S5).

SSGP-encoding transcripts were classified into 50 groups according to sequence similarity among the cDNAs and derived proteins (Appendix B- Table S4). Among the 50 groups, 31 had either a single clone or multiple clones that encode the same protein. The remaining 19 groups had multiple clones that encode at least two different proteins. Proteins within a group shared at least 30% amino acid identity and had a highly conserved secretion signal peptide. Proteins between different groups shared no meaningful ($E>10^{3}$) sequence similarity and had a completely different secretion signal peptide (Genbank accession numbers for the ESTs would be available upon publication). Figure 3.2 shows amino acid sequence alignments of two representative groups, both with a highly conserved signal peptide and a more diversified mature protein. The overall conservation among group members, particularly in the

signal peptide region, suggests that transcripts within a group may have been derived from genes that share the same evolutionary origin and, therefore, can be considered the same gene family. Some sequence variation may have also resulted from different alleles of the same gene (Figure 3.2).

Sequence variations among group members of both species

Group members among those with significant sequence variations in both barley midge and oat midge were divided into mature protein (MP)-coding region, signal peptide (SP)-coding region, and non-coding regions, and percentages of nucleotides with sequence variation in each region were analyzed (Table 3.1 and 3.2). Among these three regions, sequence variation in the SP-coding region was the lowest in both species, likely due to the functional constraint of the secretion role of signal peptides. Variation rates in MP-coding regions were higher than non-coding regions in both species (Appendix B- Table S3 and S6,). To examine if group members were under selection pressure for diversification, the percentages of non-synonymous and synonymous mutations in the MP-coding region were also analyzed, and 80% of nucleotide substitutions were non-synonymous in barley midge, and 77% were non-synonymous in oat midges (Table 3.1 and 3.2).

SSGPs conserved among two or three species

To identify SSGPs conserved among the three species or between two of them, comparative analyses were conducted via local blasting and sequence alignments. Many SSGP groups from both barley and oat midges share sequence similarity with the previously identified SSGP families from Hessian fly (Table 3.3) (Chen et al. 2008, 2010). The conserved families identified among the three species include SSGP-1, SSGP-4, SSGP-11, SSGP-71.

Family SSGP-1, (family one) the most abundant group of SSGPs in Hessian fly, was previously identified with several subgroups, including SSGP-1A, SSGP-1B, SSGP-1C, SSGP-1D, and SSGP-1E (Chen et al. 2008). In this study, we identified many SSGPs from barley and oat midges that share sequence similarity to members belonging to the Hessian fly SSGP-1 family. There are 26 unigenes from barley midge that share high similarity ($E<10^{3}$) to Hessian fly SSGP-1C1, and 5 unigenes encode proteins with similarity ($E<10^{3}$) to SSGP-1D1. From oat midge, there are 18 unigenes encode proteins that share high similarity ($E<10^{5}$) to Hessian fly SSGP-1C1, and 5 unigenes encode proteins that share high similarity ($E<10^{4}$) to SSGP-1D1 (Table 3). Comparative alignment for members from both barley midges and oat midges with groups SSGP-1C1 (Figure 3.3.A), and SSGP-1D1 (Figure 3.3.B) showed a highly conserved structure in both signal peptide and mature protein regions among the three species.

We have also identified SSGPs which are conserved either between barley midge and Hessian fly or between oat midge and Hessian fly. There were 47 unique genes from barley midge encoding proteins that share sequence similarity to Hessian fly SSGP-1C2, and 67 unigenes from oat midge encoding proteins that share similarity with Hessian fly SSGP-1A2 (Table 3.3).

Family SSGP-4, (family four) from Hessian fly has 11 identified groups: SSGP-4A, SSGP-4B, SSGP-4C, SSGP-4D, SSGP-4E, SSGP-4F, SSGP-4G, SSGP-4H, SSGP-4I, SSGP-4J, SSGP-4K (Chen et al. 2008). In this study, we have identified SSGPs from both barley midge and oat midge homologous to family SSGP-4 members. We found 8 unigenes from barley midge, and 3 unigenes from oat midge encoding proteins with sequence similarity to SSGP-4 Hessian fly members (Table 3.3). However, all these unigenes are homologous to group SSGP-4A, and no homologues have been found in the other groups.

Family SSGP-11, Hessian fly SSGP family 11 includes three groups: SSGP-11A, SSGP-11B, and SSGP-11C (Chen et al., 2006; 2008). In this study, we identified 11 unigenes from barley midge and 8 unigenes from oat midge encoding proteins homologous to SSGP-11 (Table 3.3). All these unigenes are homologous to group SSGP-11B, except one unigene from oat midge, which is highly conserved with ($E<10^3$) similarity to group SSGP-11C. Comparative alignments of SSGP-11B members from the three midges showed that those proteins share high similarity in both the signal peptide and mature protein regions (Figure 3.4).

Family SSGP-71, SSGP-71 has 426 members identified from Hessian fly. Among the Hessian fly SSGPs, 14% of these proteins are partially orthologous to other arthropods, whereas 86% have no homology to any organisms (Zhao et al. 2015). Members belonging to this family have been found to encode larger proteins (over 400 amino acids) and, unlike other SSGPs, are relatively well dispersed throughout the genome, often in triplets (Zhao et al. 2015). Structural analyses showed that SSGP-71 mature proteins contain a cyclin-like F box domain near the N terminus and a series of leucine-rich repeats (LRRs). This suggests that proteins belonging to this family and are necessary for galling and are similar to bacterial plant pathogen effectors.

In this study, we identified 3 unigenes from barley midge and 2 unigenes from oat midge (Table 3.3) homologous to Hessian fly SSGP-71 proteins. These homologues from both barley and oat midges are truncated at the 3' end. However, comparison alignments revealed that they are homologous (Figure 3.5).

Our analysis identified eight groups of SSGPs conserved between only two species. These are groups 19, 20, 23, 24, 29, 32, 40 and 51 from the barley midge and groups 8, 15, 24, 39, 23, 21, 30 and 35 from oat midge (Table 3.4). In these eight groups 19 (10.7%) unigenes from the barley midge and 25 (12.9%) unigenes from the oat midge were found to be conserved

between the barley and oat midges, whereas no homologues were found in the Hessian fly. Based on NCBI database, the functions of these eight groups are not known, except two groups (groups 24 and 51) from barley midge and groups 30 and 35 from oat midge, have been identified as larval cuticle protein and peptidyl-prolyl cis-trans isomerase, respectively (Table 3.4).

SSGPs unique to each species

To understand how different gall midge species successfully infest different host plants within the Gramineae family, comparative analyses of transcripts encoding divergent SSGPs among the three gall midge species were conducted. Our comparative analysis showed that many of the SSGP groups from either the barley midge or oat midge that had no similarity to SSGPs from each other, nor to those from Hessian fly. There were 37 groups with 56 unigenes unique to the barley midge. Among these groups 4, 5, 6, 13, 15, 18, 33, 34 had at least two or more unigenes. The remaining groups had only one single gene in each group (Appendix B-Table S1).

There were 37 groups with 62 unigenes that were unique to oat midge. Among the 37 groups, ten groups (groups 2, 3, 4, 6, 9, 12, 14, 16, 17, and 19) had at least two unigenes. The remaining groups had only one single gene (Appendix B-Table S4). All these groups from both barley and oat midges have been identified with unknown functions in the NCBI database, with the exception of group 35 from barley midge, which has been identified encoding a kinase domain protein (Appendix B -Table S1).

Discussion

Like other plant-sucking insects, gall midges secrete effector proteins into their host, inducing various forms of plant outgrowth (galls) (Dieleman 1969; Hori 1992). In the genus *Mayetiola*, the three stem midges, Hessian fly, barley midge, and oat midge, share the same

feeding mechanism and causes the same symptoms, but each species infest a different host plant. This study conducted an initial analysis on the genetic mechanisms behind gall induction and host specificity for these three midges, using transcripts obtained from salivary glands of first instar larvae of barley and oat midges, and transcripts previously identified from Hessian fly first instar larvae (Chen et al. 2008, 2010).

Comparative blast analyses of the total transcripts from the three species found that many of the SSGP groups from both barley and oat midges share sequence similarity with previously identified SSGP families from Hessian fly. More than 50% of the unigenes from both barley and oat midges are conserved, and the majority of these are homologous to SSGP families 1, 4, 11, and 71.

Among these four families, SSGP-1 has the highest portion of conserved unigenes, with 78 of 102 unigenes from barley midge, and 90 of 107 unigenes from oat midge. As observed in Hessian fly, members of SSGP-1 were the most abundant in transcripts from barley and oat midges as well (Appendix B- Table S1 and S4). Many members of SSGP-1 have been identified with effectors characteristics, including: short peptides with 50-150 amino acid residues, highly hydrophobic secreted signal peptide on the N-termini, no homology to any known proteins, and they have short repeated peptides in the mature protein region (Chen et al. 2008). This study also identified such characteristics among the SSGPs from barley and oat midges. Comparative transcriptomic and proteomic analyses have also identified effectors shared among three related aphid species, revealing that some well characterized aphid effectors such as MpC002 and Me10 are highly conserved among species belonging to the same family or even across different families (Thorpe et al. 2016). The conservation of SSGP-1 family members among the three midge species suggests common roles in these gall midges, which might involve

in inducing nutritive cells, altering physiological pathways, and irreversibly stunting the growth of the host plant.

In addition to these common characteristics, many SSGPs among the three midges exhibited an unconventional conservation pattern. This phenomenon was previously identified in Hessian fly (Chen et al. 2010; Zhao et al. 2015), in which sequence alignments of many SSGPs members showed that the 5'- and 3'-non-coding regions and introns were highly conserved, whereas the regions encoding mature proteins were highly diversified. These significant findings suggest that many of the SSGPs were under high selection pressure for functional adaptation during co-evolution with the host plant. Our analysis showed the same unusual pattern among SSGP-encoding genes from both barley midge and oat midge (Figure 3.6 and 3.7). Sequence alignments of cDNA for the 5'- and 3' un-translational regions provided evidence that these regions were highly conserved among the three species (Figure 3.8).

SSGPs in SSGP-11 were previously characterized with less abundant transcripts and few members from Hessian fly. However, SSGPs that belong to this family have also been identified as having effector characteristics. Unlike SSGP-1, however, amino acids sequences belonging SSGP-11 lack short repeated peptides at the C-termini. This study identified homologues from barley midge and oat midge that share sequence similarity to groups SSGP-11B (Figure 3.4) and SSGP-11C of Hessian fly. Previous analysis of RNA expression profile from Hessian fly revealed that transcripts belonging to this family were exclusively expressed in the larvae, and Northern blot analysis indicated that these genes were predominantly expressed in salivary glands (Liu at al. 2004).

Genomic analysis of Hessian fly revealed that ~13% of the genes encode putative gall effectors, and that SSGP-71 is the largest known arthropod gene family (Zhao et al. 2015).

Members of SSGP-71 are large proteins with cyclin-like F box domains near the N terminus and a series of leucine-rich repeats (LRRs) (Zhao et al. 2015). F box domains are commonly associated with LRRs, and both domains mediate protein-protein interactions in a variety of species. Similarity in the structure of SSGP-71 to both ubiquitin E3 ligases in plants and E3-ligase-mimicking effectors in plant pathogenic bacteria suggest that proteins in this family are necessary for galling and resemble bacterial plant pathogen effectors. Moreover, it is believed that SSGP-71 proteins are a novel class of F-box-LRR mimics that enable the insect to hijack the plant proteasome in order to produce nutritive tissue, defeat basal plant immunity and stunt plant growth (Zhao et al. 2015). In this study, we identified homologues from barley midge and oat midge that share sequence similarity to Hessian fly SSGP-71 proteins (Figure 3.5.A and 3.5.B). The conservation of SSGP-71 members among the three gall midge species also suggests conserved functions in interactions between gall midges and host plants.

Evidence that SSGPs are under selection pressure were identified among the three midges. We have found 80% and 77% of point mutations among group members are non-synonymous in barley midges and oat midges respectively (Tables 3.1 and 3.2), and over 80% of point mutations among group members from Hessian fly were nonsynonymous (Chen et al. 2004). The fast-evolving nature of SSGP-encoding genes in all three species is another indicator that these genes are involved in co-evolution interaction with their host plants (Thompson 1998).

In addition to the SSGPs conserved among the three gall midges, 32% of unigenes from barley midge and oat midge share no homology to SSGPs from Hessian fly or any other species (Appendix B- Tables S1 and S4). The uniqueness of SSGPs in different species could be due to the fact that some of them have not been identified due to low coverage of transcripts in our studies on barley and oat midges. Indeed, less than 1,000 high quality cDNA clones were

obtained from each species. More comprehensive studies could identify more SSGPs common among the three insects. If some of the SSGPs are indeed unique to each species, they must perform unique functions in gall midge - plant interactions. Considering that each gall midge species parasitizes different host plants, these unique putative effectors might be responsible for host specificities.

Conclusion

We conducted a global analysis on genes expressed in the salivary glands of first instars of the barley midge and oat midge for the first time. Our transcriptomics analyses have identified many putative effectors that are produced in the saliva of these two species. When comparing these putative effectors from barley midge and oat midge with those from Hessian fly, we identified many conserved putative effectors among the three gall midges. Interestingly, many of the conserved effectors from barley midge and oat midge were homologous to SSGP groups that were characterized with effector characteristics from Hessian fly, that is, family SSGP-1, the most abundant group (Chen et al. 2008), and family SSGP-71, the largest arthropods family (Zhao et al. 2015). These conserved groups among the three species lead us to believe that these SSGPs might have conserved functions sharing similar mechanisms in infestation and causing the similar damages and symptoms in different host plants.

To better understand the role of these conserved SSGPs in specific plant-gall midge interactions, we have selected family SSGP-1, which appeared to be the most abundant group among the three species, for further characterization. Generating recombinant proteins for members from this family would be a useful tool to produce antibodies for detection of these proteins in infested tissues, and to identify interacting targets from host tissues (Chapter 4). We believe that getting insight about the functions of these conserved effectors could be useful to reveal feeding mechanisms.

We have also identified many divergent putative effectors among the three midges, and we believe some of these unique effectors might have roles in host specification. In order to determine whether these unique SSGPs are indeed species-specific, further investigation of putative effectors under a broader search will be required.

References

Aggarwal, R., Subramanyam, S., Zhao, C., Chen, M.S., Harris, M.O. and Stuart, J.J. 2014. Avirulence effector discovery in a plant galling and plant parasitic arthropod, the Hessian fly (Mayetiola destructor). PLoS One, 9(6), p.e100958.

Barnes, H.F. 1956. Gall midges of economic importance. Vol. VII: Gall midges of cereal crops. London, UK: Crosby Lockwood.

Chen, M.S., Liu, X., Yang, Z., Zhao, H., Shukle, R.H., Stuart, J.J. and Hulbert, S. 2010. Unusual conservation among genes encoding small secreted salivary gland proteins from a gall midge. BMC evolutionary biology, 10(1), p.296.

Chen, M.S., Zhao, H.X., Zhu, Y.C., Scheffler, B., Liu, X., Liu, X., Hulbert, S. and Stuart, J.J. 2008. Analysis of transcripts and proteins expressed in the salivary glands of Hessian fly (Mayetiola destructor) larvae. Journal of insect physiology, 54(1), 1-16.

Chen, M.S., Fellers, J.P., Zhu, Y.C., Stuart, J.J., Hulbert, S., El-Bouhssini, M. and Liu, X. 2006. A super-family of genes coding for secreted salivary gland proteins from the Hessian fly, Mayetiola destructor. Journal of Insect Science, 6(1).

Chen, M.S., Fellers, J.P., Stuart, J.J., Reese, J.C. and Liu, X. 2004. A group of related cDNAs encoding secreted proteins from Hessian fly [Mayetiola destructor (Say)] salivary glands. Insect Molecular Biology, 13(1), 101-108.

Dieleman, F,L.1969. Effects of Gall midge infestation on plant growth and growth regulating substances. Entomologia Experimentalis et Applicata, 12: 745–749. doi:10.1111/j.1570-7458.1969.tb02568.x

Gagne´, R.J. 2010. Update for a Catalog of the Cecidomyiidae (Diptera) of the World. Washington, DC: Entomol. Soc. Wash. 544 pp.

Gagné, R.J., Hatchett, J.H., Lhaloui, S., El Bouhssini. M. 1991. Hessian fly and barley stem gall midge, two different species of Mayetiola (Diptera: Cecidomyiidae) in Morocco. Annals of the Entomological Society of America, 84(4), 436-443.

Hori, K. 1992. Insect secretions and their effect on plant growth, with special reference to hemipterans. Biology of Insect-Induced Galls, Oxford University Press, pp.157-170.

Johnson ,A., Chen,M.S., Morton, P., Cambron ,S. and Shukle, R.H .2009. Analyzing the diversity of secreted salivary gland transcripts in Hessian fly populations from Israel and the United States. NCB-ESA North Central Branch Entomological Society of America.

Johnson, V.A., Briggle, L.W., Axtel, J.D., Bauman, L.F., Leng, E.R. and Johnston, T.H. 1978. Grain crops. In M. Milner, N.S. Scrimshaw & D.I.C. Wang, eds. Protein resources and technology, p. 239-255. Westport, CT, USA, AVI Publishing.

Lhaloui, S.M. 1995. Biology, host preference, host suitability and plant resistance studies of the Barley midge and Hessian fly (Diptera: Cecidomyiidae) in Morocco. PhD Dissertation, Kansas State University, Manhttan KS, 567pp.

Liu, X., Fellers, J.P., Wilde, G.E., Stuart, J.J. and Chen, M.S. 2004. Characterization of two genes expressed in the salivary glands of the Hessian fly, Mayetiola destructor (Say). Insect biochemistry and molecular biology, 34(3), 229-237.

Stuart, J.J., Chen, M.S., Shukle, R. and Harris, M.O. 2012. Gall midges (Hessian flies) as plant pathogens. Annual review of phytopathology, 50, 339-357.

Thompson, J.N. 1998. Rapid evolution as an ecological process. Trends in ecology & evolution, 13(8),329-332.

Thorpe, P., Cock, P.J. and Bos, J. 2016. Comparative transcriptomics and proteomics of three different aphid species identifies core and diverse effector sets. BMC genomics, 17(1), p.172.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R. and Batterton, M. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Current Biology, 25(5), 613-620.

Zhao, C., Shukle, R., Navarro-Escalante, L., Chen, M., Richards, S. and Stuart, J.J. 2016. Avirulence gene mapping in the Hessian fly (Mayetiola destructor) reveals a protein phosphatase 2C effector gene family. Journal of insect physiology, 84, 22-31.

Figure 3.1 Amino acid sequence alignments of two representative groups, A and B, from barley midge. The boundary between predicted signal peptide and mature proteins is indicated by an arrow.

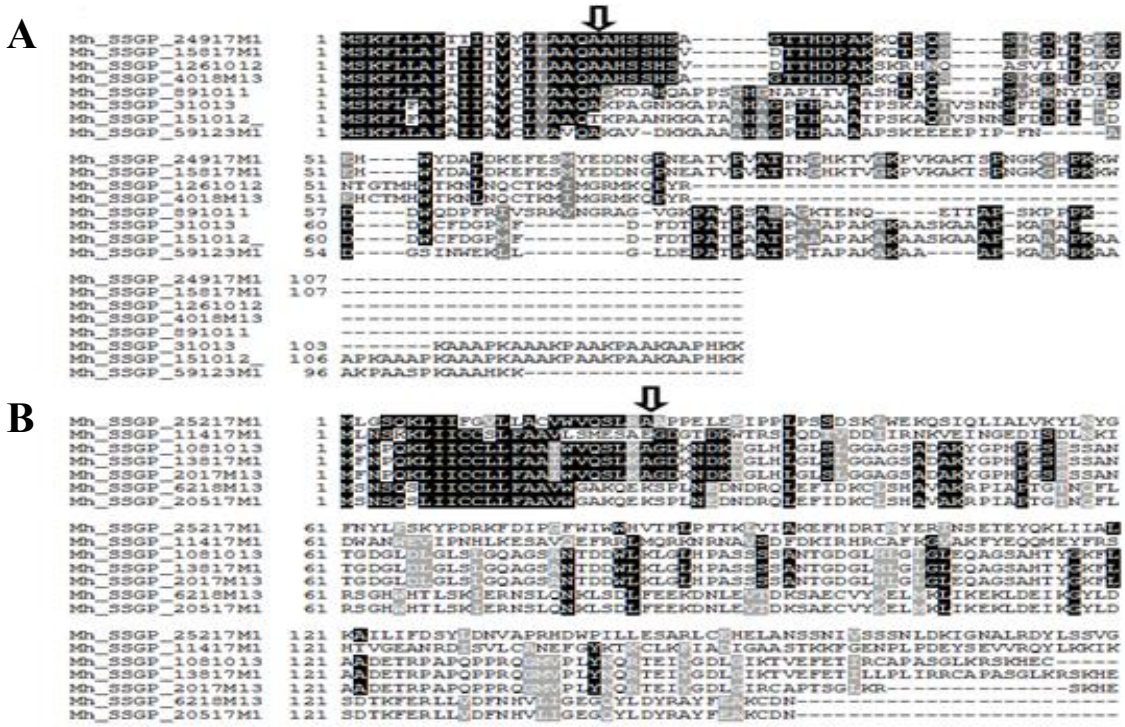


Figure 3.2 Amino acids sequence alignments of two representative groups, A and B, from oat midge. The boundary between predicted signal peptide and mature proteins is indicated by an arrow.

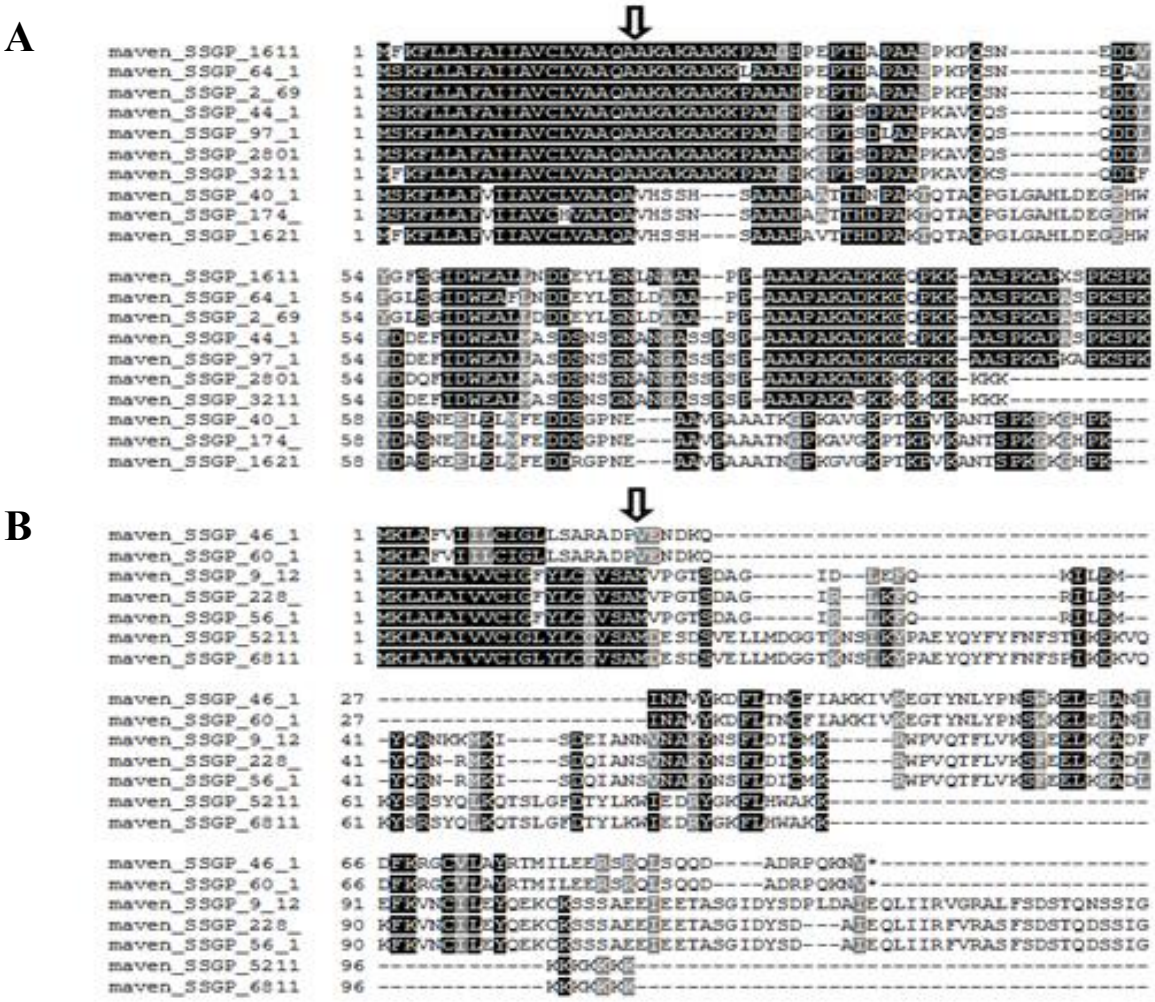


Figure 3.3 Amino acid alignments of two SSGP groups conserved among the three species. A: An alignment of proteins similar to Hessian fly SSGP-1C1 from both barley and oat midges. B: An alignment of proteins similar to Hessian fly SSGP-1D1 for members from both barley and oat midges.

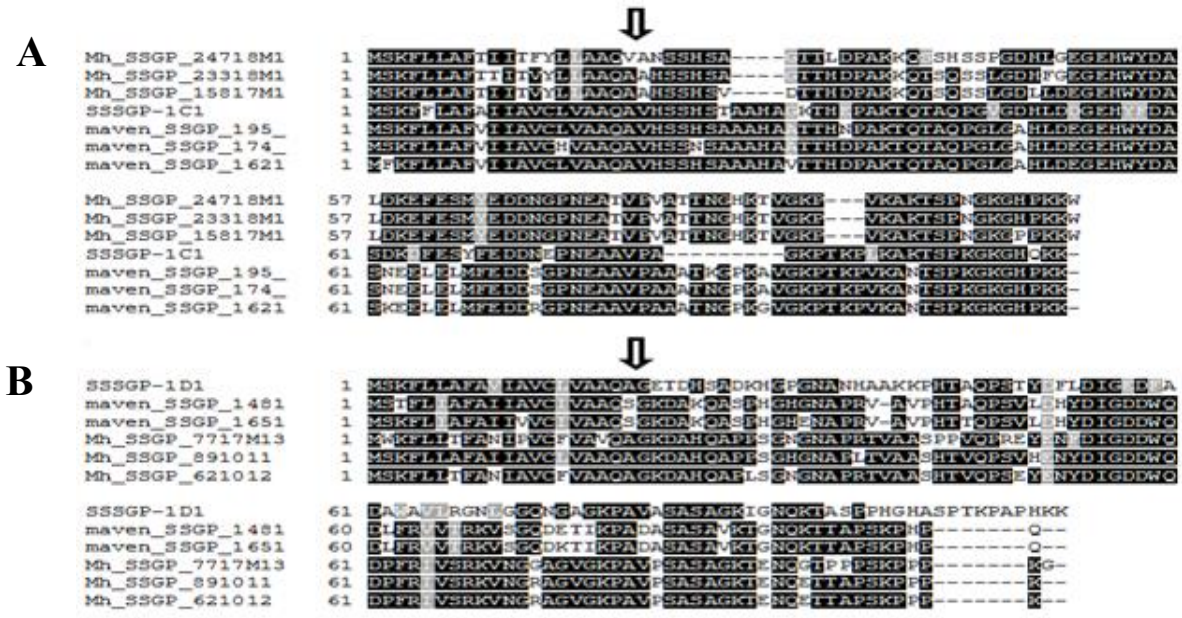


Figure 3.4 An alignment for proteins similar to Hessian fly SSGP-11B from barley and oat midges.

Figure 3.5 Sequence alignments of putative effectors proteins similar to Hessian fly SSGP-71 from barley and oat midges. A: SSGPs belonging to group 11 from the barley midge. B: SSGPs belonging to group 44 from the oat midge.

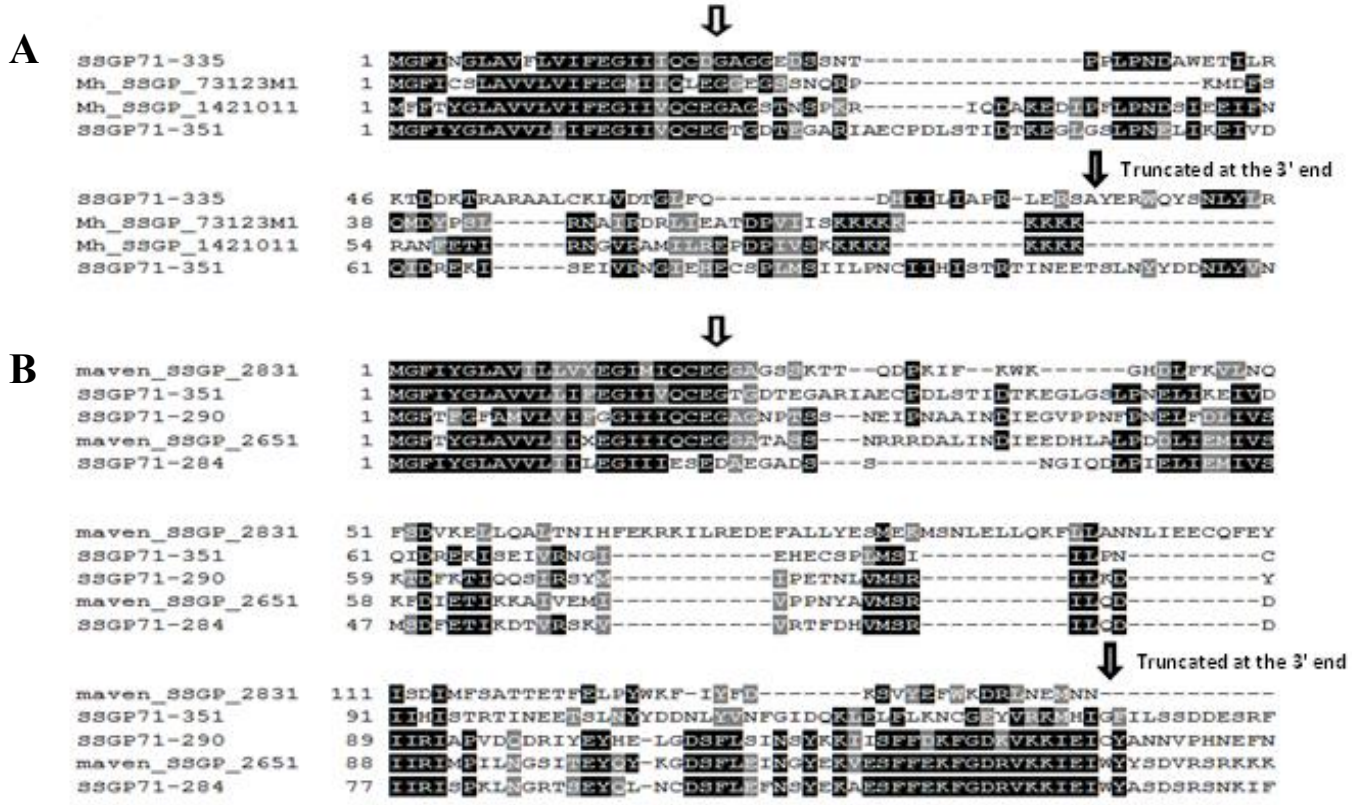


Figure 3.6 Sequence alignment of the cDNAs for members belonging to group 1 from the barley midge. The 5’ & 3’ UTRs, start, stop, and the mature protein have all marked in this figure.

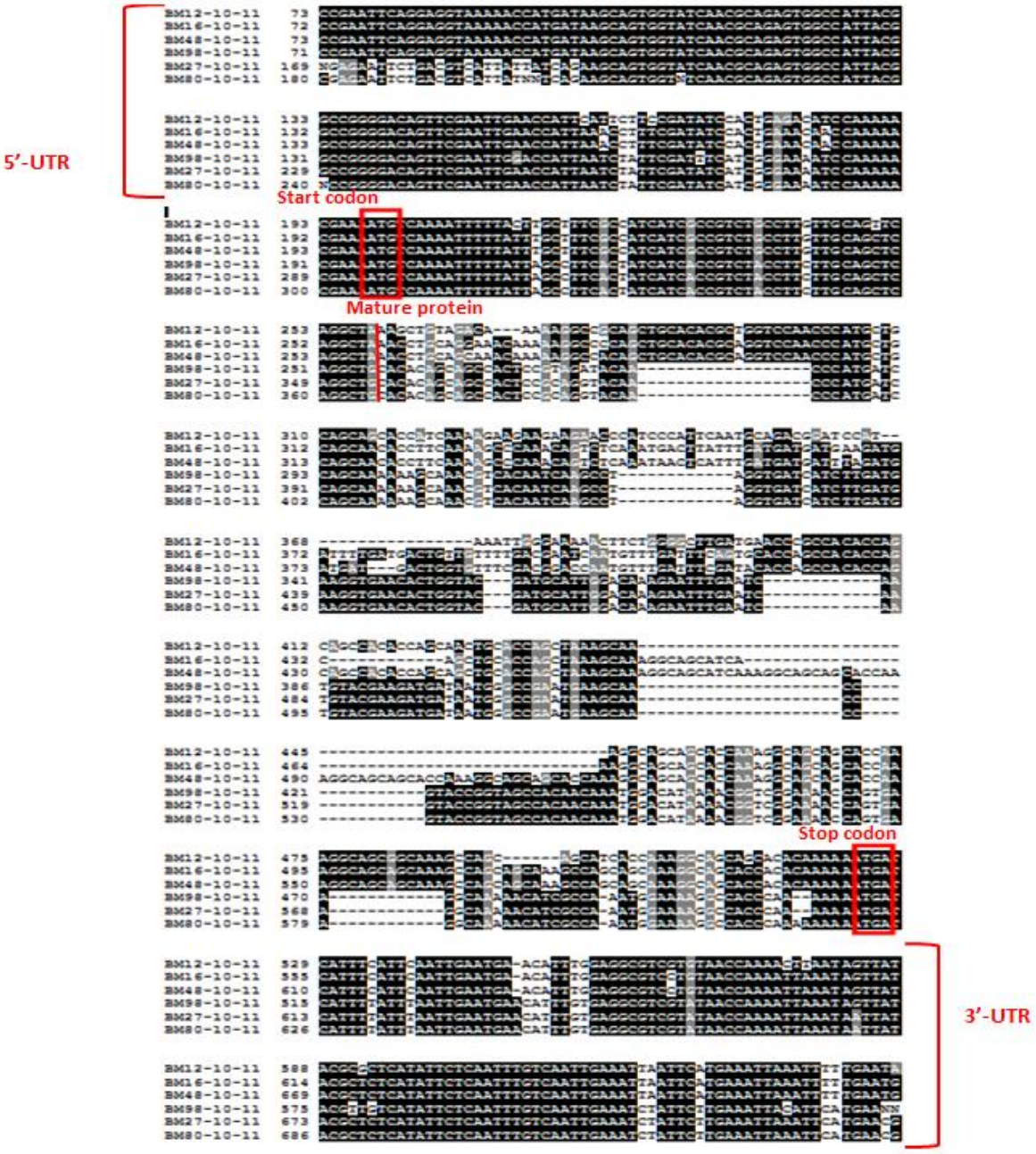


Figure 3.7 Sequence alignment of cDNAs for members belonging to group 1from the oat midge. The 5’ & 3’ UTR, start, stop, and the mature protein have all marked in this figure.

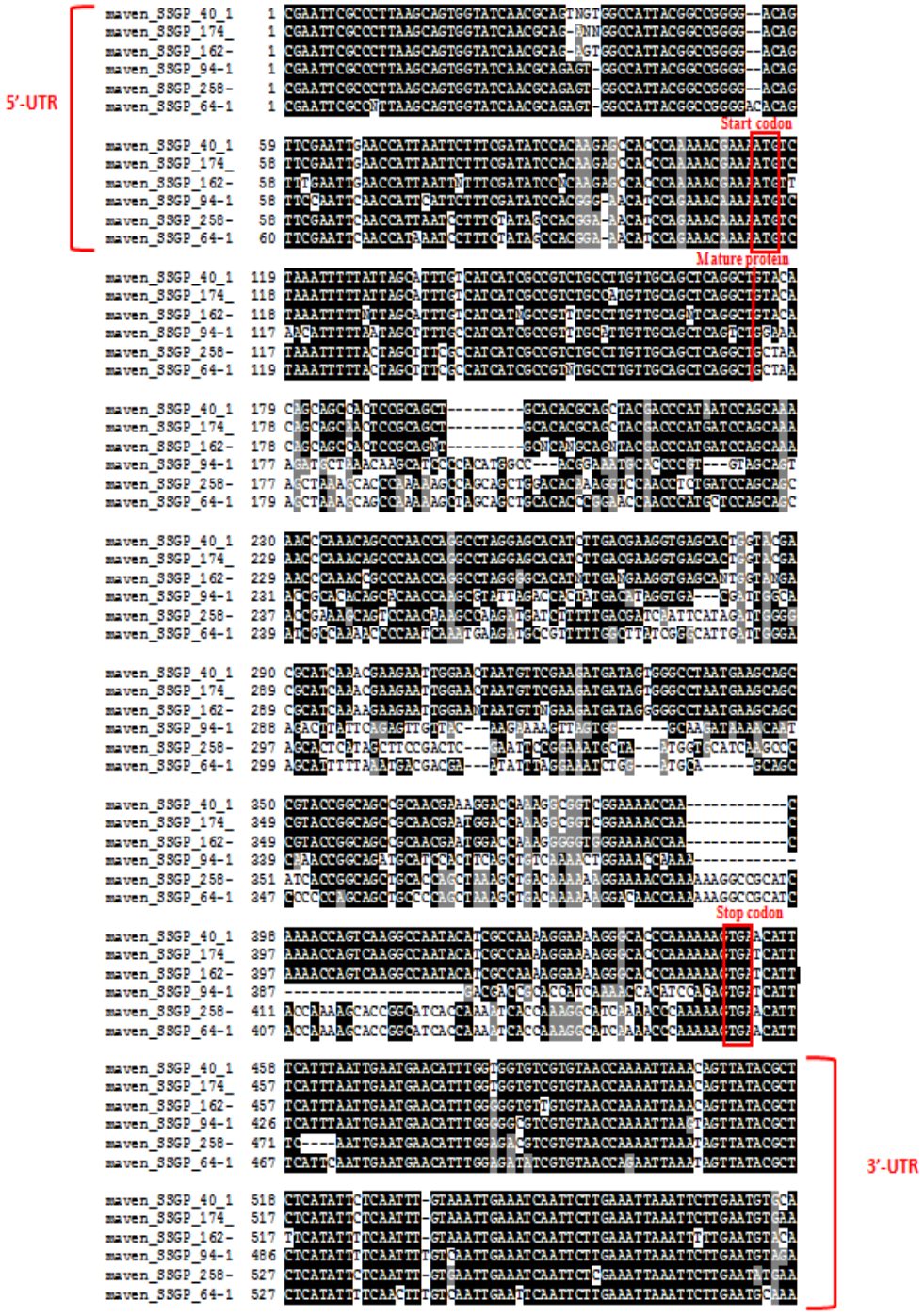


Figure 3.8 Sequence alignment of the 5’ & 3’ UTRs of cDNAs from the three gall midges, Hessian fly, barley midge, and gall midge. A: 5’ UTR . B: 3’ UTR.

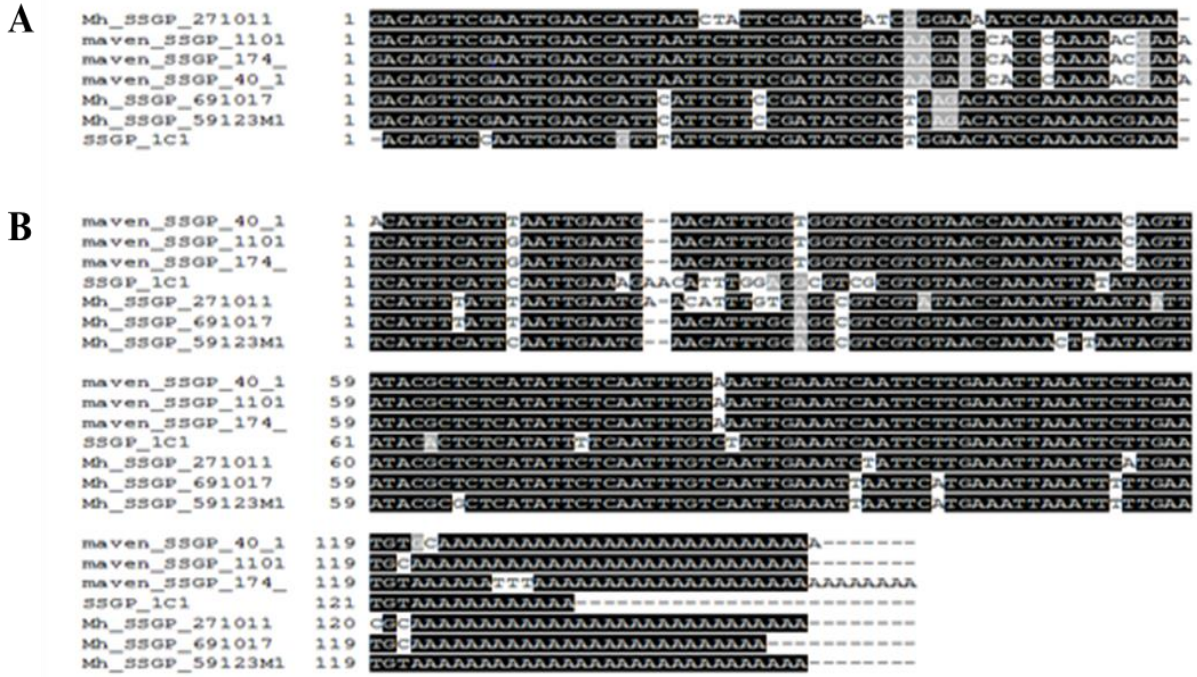


Table 3.1 Sequence variations among group members belonging to barley midge. MP - Mature Protein, SP - Signal Peptide.

Group #	% Nucleotides with mutations among group or sub-group members			% non-synonymous Mutations in MP coding
	MP-coding	SP-coding	Non-coding	
Group 1	25.4	5	5.9	72.2
Group 4	31.7	6.6	8.05	80.4
Group 31	24.26	1.7	3.3	86.2
Group 32	20.1	0	5.2	81.6
Group 33	11.9	3.7	8.1	80.35

Table 3.2 Sequence variations among group members belonging to the oat midge. MP - Mature Protein, SP - Signal Peptide.

Group #	% Nucleotides with mutations among group or sub-group members			% non-synonymous Mutations in MP coding
	MP-coding	SP-coding	Non-coding	
Group 1	17.2	6.6	11	83.7
Group 3	10.13	2	8.18	73.9
Group 8	7.8	0	3.8	66.6
Group 15	33.6	0	9.1	85.2

Table 3.3 Comparison of the conserved groups of SSGPs among all three species of gall midges, and between each barley midge or oat midge to Hessian fly.

Hessian fly BLAST hits	Oat midge			Barley midge		
	Gene	Number of uniqenes	Number of total sequences	Gene	Number of uniqenes	Number of total sequences
Conserved groups among the three species						
SSSGP-1C1 [Mayetiola destructor] 5e-55	Group 1	18	31	Group1	26	101
SSSGP-1D1 [Mayetiola destructor] 3e-24	Group 1	5	11	Group 1	5	5
SSGP-11 family protein [Mayetiola destructor	Group 5	8	14	Group 31	11	24
SSGP-71 [Mayetiola destructor]	Group 44	2	2	Group 11	3	4
SSGP-4A	Group 22	3	3	Group 28	8	10
secreted protein F [Mayetiola destructor]	Group 29	2	2	Group 39	2	3
Total		38			55	
Conserved groups between the Barley midge and Hessian fly						
SSSGP-1C2 [Mayetiola destructor] 6e-10				Group 1	47	154
Conserved groups between the Oat midge and Hessian fly						
SSSGP-1A2 [Mayetiola destructor] 4e-21	Group 1	67	241			
salivary secreted protein [Mayetiola destructor] 5e-43	Group 7	1	2			
putative secreted protein, partial [Mayetiola destructor] 1e-29	Group 48	1	1			
Total		69				

Table 3.4 Comparison of the conserved groups of SSGPs between only barley midge and oat midge.

First BLAST hit if any	Oat midge			Barley midge		
	Gene	Number of uniqenes	Number of total sequences	Gene	Number of uniqenes	Number of total sequences
Conserved groups between the Barley midge and the Oat midge						
Unknown protein	Group 8	6	19	Group 19	3	14
Unknown protein	Group15	7	15	Group 20	5	5
Unknown protein	Group24	6	7	Group 23	1	1
larval cuticle protein 12-like	Group35	1	1	Group 24	3	5
Unknown protein	Group39	1	1	Group 29	1	2
Unknown protein	Group23	1	2	Group 32	4	22
Unknown protein	Group21	2	7	Group 40	1	12
peptidyl-prolyl cis-trans isomerase FKBP2 precursor	Group30	1	1	Group 51	1	2
Total		25			19	

Chapter 4 - Tissue localization and interacting wheat targets of SSGP family 1 members from Hessian fly

Abstract

Like pathogens, Hessian fly and other gall midges deliver effectors into plant tissues to induce gall formation. Small secreted proteins have been found to play crucial roles in interactions between biotrophic or hemibiotrophic pathogens and plants. Hessian fly also has the same trophic styles mimicking pathogens, and many of the Secreted Salivary Gland Proteins (SSGPs) from Hessian fly larvae are small peptides with effector characteristics. Family SSGP-1, is one of these small secreted peptides. Members belonging to this family are among the genes that are most abundantly expressed based on most abundant transcripts. This family is also conserved among the Hessian fly, barley midge, and oat midge. High levels of expression and conservation of this family among different insect species indicate their important roles in Hessian fly-wheat interactions. At present, no direct evidence suggests that these proteins are injected into host tissues for effector functions. In this study, we presented for the first time several lines of evidence to indicate that SSGP-1 groups 1A, 1B, 1C, 1D, and 1E, are injected into host plants by Hessian fly larvae and are involved in Hessian fly - wheat interactions.

Immunostaining using antibody detected specific localization of five SSGP-1 proteins in large cells of the basal region of the salivary glands obtained from first instar Hessian fly larvae. Western blot analysis demonstrated that these five proteins were expressed in all larval stages. The continuous production of these proteins suggests that these proteins are necessary to initiate and maintain Hessian fly infestation. Consistent with their effector functions, these five proteins were detected in infested wheat tissues based on western blot analyses. To identify possible target proteins in host plants that interact with SSGP-1 family proteins, *in vitro* pull-down assays

were performed. Putative interacting targets for SSGP-1A, SSGP-1B, and SSGP-1C were identified by Liquid Chromatography Mass Spectrometry (LC-MS/MS). These putative interaction target proteins included a tubulin, a lipoxygenase, ribosomal proteins, and uncharacterized proteins. Identification of these putative targets provide a base for further confirmation of their interaction with Hessian fly effectors in future experiments.

Introduction

Larvae of the gall midge family are known to deliver effectors into plant tissues to induce gall formation, suppress plant defense, and inhibit host growth (Hatchett et al. 1990; Gagne´ et al. 1991; Harris et al. 2006). Evidence from Hessian fly, the mostly studied gall midge; suggest that larvae manipulate host plants by injecting effector proteins into host tissues during feeding stages (Chen et al. 2004, 2008, 2010; Zhao et al. 2015). Thus, due to the importance of the SSGPs in Hessian fly virulence and biotype differentiation, transcripts expressed in salivary glands of first instar larvae have been reported in many references (Liu et al. 2004; Chen et al. 2004, 2006). The SSGPs encoding cDNAs from Hessian fly were grouped into 71 families and groups according to sequence similarities. Many of these families were identified with large numbers of family members, such as family SSGP-1, SSGP-4, SSGP-9, and SSGP-71(Chen et al. 2008). Family SSGP-1 contains members that have the most abundant transcripts among SSGPs so far identified. Five groups of SSGPs were found belonging to this family. They are SSGP-1A, SSGP-1B, SSGP-1C, SSGP-1D, SSGP-1E (Figure 4.1) (Chen et al. 2008). Alignment of members belongs to these groups share high similarity in the signal peptide (Figure 4.1). Regarding the variability in the mature proteins, SSGP-1A share sequence similarity to SSGP-1B more than other groups, and group SSGP-1E are more diverse and has the largest structure.

SSGPs from these groups were identified with effector characteristics. For instance; they are short peptides 100-180 amino acid residues, have highly conserved hydrophobic secreted

signal peptide on the N-termini, are exclusively expressed in the crucial larval stage of infestation, and have no homology to any known proteins (Chen et al. 2008; 2010). Group SSGP-1A has 13 members with 114 amino acid residues. Group SSGP-1B has 31 members with 116 amino acid residues. Group SSGP-1C has 7 members with 103 amino acid residues. Group SSGP-1D has one member with 112 amino acid residues. And group SSGP-1E has one member with 175 amino acid residues (Chen et al. 2010).

Salivary glands of the barley midge and oat midge have SSGPs that are homologous to those from Hessian fly SSGP-1. Like in Hessian fly, these homologous SSGPs are also the most abundant in terms of transcripts. The conservation of SSGP-1 family members among the three gall midge species suggests their important roles in feeding and virulence.

However, the roles of SSGP-1 proteins are not known. In this study, we have selected five members from SSGP-1, including 1A2, 1B1, 1C1, 1D1, and1E1 from Hessian fly for further characterization. Our hypothesis is that members in SSGP-1 are injected in host plant, and they are key components in feeding mechanisms. Thus, we generated and purified recombinant proteins for these five members, and produced antibodies using purified recombinant proteins. Our objectives were 1) to determine localization of the five proteins in different tissues of the first instar larvae; 2) to detect the presence of these proteins in infested wheat tissues using western blots; and 3), to identify probable targets from wheat tissue at the feeding site using affinity resin column, so called pull-down assays, coupled with LC-MS/MS.

Materials and methods

Recombinant protein production and purification

Recombinant proteins for genes encoding 1A2, 1B1, 1C1, 1D1, and 1E1 were produced using the pET system as described below and following the manual provided by the manufacturer Novagen, Inc.

Gene construction: cDNA-encoding genes for the five proteins were inserted in pET-28a vector for cloning, vector map and the insert gene described in (Appendix C- Figure S8). For vector preparation, 1 µl of the vector transferred to 100 µl of the host bacteria BL21 on ice. The bacteria tube then heated for 30 s in a 42°C water bath, and replace on ice again for 2 min. The transformed bacteria were grown in LB broth overnight with shaking at 37°C. The plasmid were then isolated and purified from the LB broth using SpinPrep plasmid kit (Novagen, Inc). To harvest the inserted plasmid from bacteria, 1.5 ml overnight culture transferred to 1.5 ml tube, after centrifuging and discarding supernatant, the pelleted cells resuspended in 100 µl bacterial resuspension Buffer. Two hundred µl of Lysis Buffer added to the tubes and mixed. After 5 min. incubation at room temperature, 400 µl neutralization buffer was added and centrifuged for 10 min, then 650 µl of the supernatant was transferred to filter unit and centrifuged for 1 min. After discarding the filtrate, 650 µl reconstituted wash buffer was added, filters then transferred to receiver tubes, and 50 µl pre-warmed elute buffer added and incubated for 3 min at 50°C. To collect eluted plasmid DNA, final centrifugation for 1 min was carried out. Five µg of the purified vector was incubated with the restriction enzymes at 37ºC for two hours. To check the extent of digestion, 5 µl of samples was loaded on an agarose gel. To decrease non-recombinants background, alkaline phosphatase was added to the digestion reaction at the end, after diluting the enzyme in Tris-HCl, pH 9.0 buffer and incubated at 37ºC for one hour. To purify the digested vector, 20 µl of samples were loaded on an agarose gel, and DNA plasmid were recovered by cutting the band from the gel using a clean razor blade, and then gel bands were purified by SpinPrep Gel DNA kit (Novagen, Inc.). The gel slice was dissolved in 300 µl of GelMelt Solution per 100 mg of gel slice by incubation at 50°C water bath for 10 min. Seven hundred µl of the dissolved gel solution was transferred to SpinPrep Filter, and centrifuged. Multiple loads

of the melted gel slice applied until entire volume of melted gel solution has been passed through SpinPrep Filter. The flow-through from the receiver tube discarded, 650 µl of wash buffer added and centrifuged. The flow-through then discarded, and SpinPrep Filter transferred to Eluate Receiver Tube. 50 µl of pre-warmed (50°C) SpinPrep Elute Buffer added onto the SpinPrep Filter, and incubated for 3 min. Eluted DNA plasmid were collected by centrifuging.

Target genes were amplified by PCR with the specific primers for each gene, 1A2 Forward GCTGTAACTAAACATCCAGC, 1A2 Reverse TCACTTCTTCTTTGAGGCT; 1B1 Forward GCTAAACCTAAAAAAGGCA; 1B1 Reverse TGGATAAGAACGGGGGAGA; 1C1 Forward GTACACAGCAGCCATTCCA, 1C1Reverse CTAAAGGAAAGACCTTATCC; 1D1 Forward GGAGAAACAGATCATTCA, 1D1 Reverse TCACTTTTTATGTGGGGC; 1E1 Forward GTACAGGAACCACAAGCAT, 1E1 Reverse CAATGGTGGAGATCGTTCT. Then PCR products were digested with the restriction enzymes at 37ºC for two hours. To verify ligation between the vector and insert, 1 µl Ligation reaction diluted 1:10 in TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8.0) mixed with PCR master mix following the manual of PCR kit (Novagen, Inc). The PCR products were visualized on an agarose gel. To decrease contamination with the fragment of interest, the digested reaction loaded on an agarose gel and then purified following the manual of SpinPrep Gel DNA kit (Novagen, Inc).

To make the correct construct, 50 ng/µl (0.03 pmol) of the purified vector were ligated with 0.2 pmol of purified gene insert. After assembling all components from the DNA Ligation Kit, Novagen, Inc. in a volume of 10 µl, the ligation reaction then incubated at 16°C for two hours. To transform the ligation reaction in bacteria, 1µl of these reactions transform in bacterial strains BL21, and incubated on ice for 5 min. The samples then heated for 30 s in a 42°C water bath, and then replace on ice for 2 min. 100 µl SOC medium were added to the transformed

tubes, and 50 µl of each transformation were spread on LB agar plates. Plates then incubated at 37ºC overnight. For insert screening, transformed colonies picked up in 50 µl of sterile water. Bacteria cells then lysed by boiling tubes in water at 99°C for 5 min, and centrifuged to remove cell debris. 10 µl of the supernatant then amplified by PCR, and visualized on gel agarose for target inserts analyzing.

Gene sequencing: to verify reading frame by sequencing, clones with inserts were picked up, and plasmid form these clones were purified. The plasmid was isolated following the manual of SpinPrep™ Plasmid Kits, Novagen, Inc. as described in previous section. The plasmid was further purified with phenol preparation, by adding TE-buffered phenol to 100 µl and extracting twice with: CIAA (1:1; CIAA is 24 parts chloroform, 1 part isoamyl alcohol) and once with CIAA. Then the aqueous phase transferred to new tube for precipitation by adding 0.1 volume 3 M Na acetate and 2 volumes ethanol. After incubating and centrifuging DNA pellet rinsed twice with 70% ethanol, the pellet then dried and resuspend the DNA in 30 µl TE. Samples of purified plasmid for each gene have been sequenced via a commercial contract with Genewize, Inc.

Gene expression: for transformation into an expression host, successful reconstituted plasmids were transferred into expression host strain.1 µl of a 50-fold dilution (1 ng) of plasmid in TE buffer transferred in bacterial strains BL21(DE3), and incubated on ice for 5 min. The samples then heated for 30 s in a 42°C water bath, and then replace on ice for 2 min. One hundred µl SOC medium was added to the transformed tubes. Fifty µl of each transformation was spread on LB agar plates. Plates were then incubated at 37ºC overnight. A single colony were inoculated into 2 ml LB medium and incubated with shaking at 37°C. The culture then stored at 4°C overnight. Bacterial cells with plasmid were collected by centrifugation, and pellet were resuspended in 2 ml fresh medium and then used to inoculate 50 ml medium in Erlenmeyer

flask. For induction, a starter culture was prepared for each recombinant gene, 3 ml LB media were inoculated with a single colony from the plate, and incubated with shaking at 37°C at 250 rpm for four hours. Then 100 ml of LB media was added to 1 ml inoculated medium and incubated at 37°C with shaking at 250 rpm for two hours. One mM of IPTG (Isopropyl-Beta-d-Thiogalactopyranoside) was added to 100 ml cultures, and then were incubated at 37°C with shaking at 250 rpm for two hours. The expression of target genes then analyzed on 12% pre-casted SDS polyacrylamide gel from Life Technologies (Fresno, CA).

Protein purification: For protein harvesting, the induced culture was lysed using PopCulture Reagent (Novagen, Inc.). One tenth culture volume of PopCulture Reagent was added to 50 ml bacteria culture. Fourty U of rLysozyme was added after diluting with rLysozyme Dilution Buffer, and 25 U of Benzonase nuclease per 1 ml. Bacteria lysed then mixed well and incubated for 10 min at room temperature. To increase the lysis efficiency, the bacteria was lysed by sonicating with a microtip at the following settings: power level between 2–3, at 20–30% duty for 8–10 bursts. After quantifying the extract using Bradford Protein Assay from (Bio-Rad, Hercules, CA), and analyzing on SDS-PAGE, the recombinant proteins were then purified using Ni-NTA His-bind resin (Novagen, Inc). After assembling the column as described in the manual, the slurry of Ni-NTA resin was poured into the column. The columns were equilibrated at room temperature. Equal amounts of protein extracts with equilibration Buffer were mixed and added to the columns. The columns were then incubated for 30 minutes at room temperature. After centrifuging the flow-through was collected, as well after each wash steps for the resin. The resin was washed with the wash buffer. Then tagged proteins were eluted by centrifuging with imidazole buffer. Eluted proteins then quantified by Bradford Protein Assay from (Bio-Rad, Hercules, CA), and analyzed on SDS-PAGE. To remove imidazole from protein

samples, the mixtures were dialyzed using dialysis Cassettes (Thermo Scientific Slide-A-Lyzer). Proteins samples were lyophilized using Flexi-Dry MP Lyophilizer.

For final purification of the proteins, Histags from the fused proteins were cleaved using Thrombin (Thrombin Restriction Grade, Novagen.Inc.). One U/ ml thrombin enzyme was added to 1mg protein sample, after adding 5 µl of 10X Thrombin Cleavage Buffer (250 mM Tris-HCl, 1.5 M NaCl, 50 mM CaCl2). The reaction mixture was incubated at 20 ^{o}C for 4 hours. To determine the extent of cleavage, samples were analyzed by SDS-PAGE. After the cleavage reaction, thrombin was removed using streptavidin. Sixteen µl streptavidin agarose was added to 1U of enzyme, and incubated at room temperature for 30 min. The reaction mixture was then centrifuged in Spin Filter. Samples of the cleaved protein then quantified by Bradford Protein Assay, and analyzed on SDS-PAGE.

Production of antibodies

Then, antibody for the five samples of the purified recombinant proteins were generated via a commercial contract with GenScript (Piscataway, NJ).

Tissues dissection and whole-mount immunostaining

Insect and plant material: Hessian fly biotype GP, derived from a Kansas population (Chen et al. 2009), were used in this study. The insect population was maintained on the susceptible wheat Newton. All insects were maintained on wheat seedlings in growth chambers at 20^{o}C and 12:12 L:D h. (day/night) photoperiod until tested.

Tissues dissections: Salivary glands, guts, and Malpighian tubules were dissected from 3 days old larvae in phosphate-buffered saline. The dissected tissues were mounted onto concave glass slides. Staining was conducted following the protocol described by Šimo et al. (2009). Dissected tissues fixed overnight in 4% paraformaldehyde in PBS, and followed by three time

washes in PBS containing 1% Triton X-100 (PBST). Tissues were dehydrated in methanol/PBS series 20%, 40%, 60%, 80%, 100% each for 30 min, then incubated in 1x PBS,0.2% Triton X-100, 20% DMSO Dimethyl sulfoxide, 0.3M glycine for 10 min. Tissues were blocked in 1xPBS, 0.2%TritonX-100, 10% DMSO, 6% M normal goat serum (Sigma, St. Louis, Mo., USA) for 60 min. The tissues were then incubated in 1:4000 dilution primary antibody for 2 days. After several washes with 1xPBS/0.2%Tween-20 with 10ug/ml heparin, the tissues were incubated overnight with 1:1000 dilution secondary antibodies (Molecular Probes, Carlsbad, Calif., USA). After several washes with 1xPBS, 0.2%Tween-20 with 10ug/ml heparin. Stained tissues were mounted in glycerol. Negative controls included preabsorption of each antibody with its respective antigen have been carried for each antibody staining. Images for the stained tissues have been captured using a confocal microscope (Zeiss LSM-700).

Samples collected for Western blot analyses

Insect and plant materials, Two Western blot analyses were conducted, one using protein extracts from whole bodies of Hessian fly adults, pupae, and larvae at different time points, and a second using protein extracts from plant tissues at the larval feeding site. For both analyses, seeds of 'Newton' and 'Molly' wheat were planted in 10 flats containing PRO-MIX ‘BX’ potting mix (Hummert Inc., Earth City, MO). The seeds were geminated and wheat seedlings were maintained in a growth chamber programmed at 20^{o}C with a photoperiod of 12:12 (L:D) h. When wheat seedlings reached the 1.5 leaf stage (stage 11 on Zadoks scales), flats of 'Newton' and 'Molly' were infested with Hessian fly females by confining flies in a cheesecloth tent. After 4 to 5 days, eggs hatched into neonates that migrated into wheat plants. Two of the ten flats were retained without infestation as controls for both 'Newton' and 'Molly' and were maintained in the same conditions.

Protein extraction from whole insects at different developmental stages: Adults, pupae, first instar, second instar, and third instar larvae of Hessian fly were collected from the susceptible flats 'Newton', and only first instar, second instar, and third instar larvae were collected from the resistant flats 'Molly'. All the collected samples were then frozen immediately in liquid nitrogen. The frozen insects were ground to powder with a high speed electric motor. Equal amounts of grounded tissues (200 mg) from all samples were weighted except in one case the sample of 1 da-old on resistant plant due to the lack of sample collection. Frozen tissues were solubilized into cold TCA-2ME-acetone solution. Protein precipitates were collected by centrifugation. After washing with cold TCA-2ME-acetone solution, the protein precipitates were air-dried and then dissolved into the so called R2D2 buffer (7 M urea, 2 M thiourea, 2% 3-[(3-cholamidopropyl) dimethyl-ammonio]-1-propane-sulfonate, 2% N-decyl-N,N-dimethyl-3-ammonio-1-propane-sulfonate, 20 mM dithiothreitol, 8 mM Tris(2-carboxyethyl) phosphine) and stored in -20°C for later western blot analyses.

Protein extraction from host tissues at the feeding site: Wheat seedlings containing 3da-old larvae were collected from infested 'Newton' and 'Molly' flats and washed clean. Roots were cut off with scissors and coleoptiles were removed from the seedlings with a pair of forceps. The first leaf-sheath was separated from the second leaf-sheath to expose larvae at the feeding site, and tissues containing larvae were soaked in 10 ml TE-SDS buffer (50 mM Tris and 2 mM EDTA, pH 8.0, with 0.1% SDS). During this process, insect proteins in wheat cells penetrated by Hessian fly larval mandibles were likely dissolved into the buffer along with wheat proteins. Hessian fly larvae in the TE buffer sank to the bottom of the solution. The upper part of the solution without Hessian fly larvae was transferred to a new tube. Control treatments from both 'Newton' and 'Molly' were also collected in the same method. The solution with proteins

was frozen in liquid nitrogen until use. After several collections in the same way, solutions containing proteins were combined, dialyzed against DI water, lyophilized, dissolved into sample buffer and stored in -20°C for later western blot analyses.

Electrophoresis and western blots

All protein extracts were quantified using a Bradford Protein Assay (Bio-Rad, Hercules, CA). Equal amounts of proteins were dissolved in sample buffer and separated on a 12% pre-casted SDS polyacrylamide gel from Life Technologies (Fresno, CA). Proteins on the gel were transferred onto nitrocellulose membrane using an electric device (Thermo Fisher, Fremont, CA). The membrane was blocked with 5% milk at room temperature for 1h., and incubated for 2h. with the antibody (1:10,000 dilution with PBST buffer) against one of the SSGP proteins. The membrane was then washed with 1% PBST for 1h. with a buffer change every 10 minutes, and incubated with HRP-conjugated secondary antibody (1:1000 dilution, Amersham, GE Healthcare Life Sciences, Pittsburgh, PA) for 1h. The membrane was washed again for 1h. with a buffer change every 10 minutes. Chemiluminescence was developed with a WesternSure$^{(R)}$ PREMIUM Chemiluminescent Substrate and visualized with a C-DiGit Blot Scanner (Li-Cor, Lincoln, NE).

Protein- protein interaction

Plant materials and treatments. After removing larvae from the infested seedlings using a fine brush, the wheat sheaths (10 mm) were collected from 3 days old seedlings of susceptible Newton and resistant Molly with/withour infestation. Wheat tissues were frozen immediately in liquid nitrogen. The frozen samples were ground to fine powder with the mortar and pestle and were stored at -80 ^{o}C until used. Total crude protein was extracted from each sample of the four treatments in a buffer containing 137 mM NaCl, 1.45 mM NaH2PO4.H2O, 20.5 mM Na2HPO4 pH 7.5. Cell debris was removed by centrifugation. Total protein

concentration was estimated by the Bradford assay (BioRad). Samples were stored at -20^{o}C until used. **S**amples for the purified recombinant proteins for the five proteins were prepared separately following the same protocol in previous paragraph of the materials and methods, and stored at -20^{o}C until used.

Pull-down assays: Pulldown assays were carried out for each one of the five proteins separately using Ni-NTA His-bind resin (Novagen, Inc). Briefly, 100 mg of a bait protein was immobilized on Ni-NTA column, washed with washed buffer and incubated with 0.5 g of total crude plant protein extract for 1 h at 4^{o}C. Each reaction was then washed before elution and dissolved in sample buffer and stored in -20°C for later SDS page analyses.

LC-MS/MS identification. Eluted samples from each reaction for the five proteins were separated on SDS page and stained with silver staining. SDS/PAGEs were examined for significant bands in each reaction. For LC-MS/MS identification significant bands from different reactions of the five proteins were cut from the gel and sent for protein identification through a commercial contract with ProtTech Inc (Phoenixville, PA). Mass spectrometric data were used to search against the most recent UniPortKB database http://www.uniprot.org with a ProtTech's ProtQuest software suite. Identified peptides for the target bands were further analyzed for confirming the identity and checking for similarity using alignment tools.

Results

Localization of SSGP-1 family members in different tissues of first instar larvae

Antibodies generated against five members of the SSGP-1 family were used for indirect immunostaining to reveal the presence of corresponding proteins in salivary glands, gut, and Malpighian tubules dissected from first instar larvae, as described in the methods section. As

shown in Figure 4.2, (A.1 to A.5), different levels of proteins were detected in the salivary glands, but no signals were detected in the gut or the Malpighian tubules for all proteins (Figure 4.2, B.1 to B.4) except 1E1(Figure 4.2, B.5.1), which exhibited weak signal in the foregut (Figure 4.2, B.5.2).

Overall, our immunostaining results revealed the following for all five proteins: 1) signals were detected only in the base region, 2) within the base region there was no signal in the proximal cells, and proteins localized in the cells of the base region showed different patterns. Localization of both 1A2 (Figure 4.2, A.1) and 1B1 (Figure 4.2, A.2) was most likely peripheral at the edge of the cells. In contrast, localization of 1E1 (Figure 4.2, A.5) was apparently central to between the cells. Both 1C1 (Figure 4.2, A.3) and 1D1 (Figure 4.2, A.4) most likely have an overall localization.

Western blot analyses of SSGP-1 family members in different life stages of Hessian fly

Protein bands were detected with expected molecular size for each one of the five members. The observed molecular size for the purified proteins on SDS page is approximately 20 kD for all five proteins except the larger protein 1E1, which is approximately 30 kD. The intensity of the bands was variable among the five proteins in the larval stages.

In the case of larvae fed on susceptible plant, SSGPs with higher sequence similarity, for instance proteins 1A2 (Figure 4.3,A) and 1B1(Figure 4.3,B), showed similar patterns of protein banding. In both cases, intensity increased as larvae advanced from age 1 to 10 days, with a much stronger band detected in 6 d-old larvae (2^{nd} instar). For 1C1 (Figure 4.3,C), bands were detected only at the late stages, namely 6 and 10 d-old. For 1D1 (Figure 4.3,D), intensity of the band started much stronger at the first three days and decreased as larvae advanced to 10d-old. For 1E1 (Figure 4.3,E), the largest protein among all, band were detected earlier in two day old

larvae. No bands could be detected for all proteins except protein 1D1 in pupae and adults (Figure 4.3,D). Protein 1D1 was with a strong band specifically in the adult stage. In the case of larvae fed on resistant plants, only two of the five proteins were detected, 1A2 (Figure 4.3,A) with faint band at 2d-old, and 1D1 (Figure 4.3,D), with a strong band in 2 and 3d-old larvae.

When RT-PCR used to measure transcript abundance of members of SSGP-1 in larvae fed susceptible wheat, the pattern of the expression was highly consistent with their corresponding proteins from our study (Chen et al., 2010).

Presence SSGP-1 proteins in host tissues at the feeding site

Protein bands were detected with expected molecular size for each of the five proteins in samples from both susceptible and resistant plants. The observed molecular size for the purified proteins on SDS PAGE was approximately 20 kD for all five proteins except the larger protein SSGP-1E1, which was approximately 30kD. The blotted membrane detected bands with good intensity for all five proteins in the tissues from resistant plants (Figure 4.4).

For 1A2 (Figure 4.4, A) and 1B1 (Figure 4.4,B), the two proteins that share more similarity among the others members, bands around 20 kD were detected from both resistant and susceptible plants. For 1C1 (Figure 4.4,C), two bands in size 20 kD were detected on the membrane, one with good intensity detected in the resistant plant, and the other was a weak band detected in the susceptible plants. Similar to 1C1, membranes of blotted 1D detected two bands in size 20 kD, a strong band detected from the resistant plant, and very weak band detected from the susceptible plant (Figure 4.4,D). For the larger protein 1E1 a more intense band has been detected in an approximate size of 30 kD from both the susceptible and resistant plants (Figure 4.4,E).

Identification of potential wheat targets that interact with Hessian fly SSGPs

1A2. Reactions eluted from incubation between susceptible wheat extract and 1A2 detected two clear, but weak bands, one band with approximately molecular size 30 kD, which was marked as SI-C-1A2 in the figure. The second band with approximately molecular size 28 kD and were marked as SI-D-1A2 (Figure 4.5, S1and S2). These two bands were specific to infested susceptible plants since they were not present in the corresponding non-infested susceptible tissues. For reactions eluted after incubation between resistant wheat extract and SSGP-1A2 a strong band was detected with approximate molecular size 15 kD, which was marked as RN-1A2. This band was specific to non-infested resistant plants since it was not present in the corresponding infested resistant plants (Figure 4.5, R1 and R2).

1B1. Similar experiments were performed with the 1B1 (Figure 4.6, S1 and S2). Three weak bands were detected in the infested susceptible plant tissues, and no corresponding bands were detected in non-infested susceptible tissues. The bands were marked as SI-D-1B1, SI-E-1B1, and SI-F-1B1, which have approximately molecular sizes of 120 kD, 100 kD, and 85 kD respectively. One weak band was detected in infested resistant plant tissues, and a different band was detected in non-infested resistant plant tissues. These bands were marked as RI-1B1, RN-1B1 (Figure 4.6, R1 and R2) with approximate molecular sizes of 18kD and 15kD, respectively.

1C1. Experiments with 1C1 reactions detected one weak band (Figure 4.7, S1 and S2) in the infested susceptible plant tissues, and no band was detected in the corresponding non-infested tissues. This band was marked as SI-A-1C1 with approximate molecular size of 22kD. No bands were found in the experiments with infested and non-infested resistant plant tissues.

1D1and1E1. Similar experiments were carried for 1D1 and 1E1, but no visual bands were detected in susceptible or resistant plant extracts.

Protein identification of target proteins by LC-MS/MS

Nine protein bands were analyzed by LC-MS/MS to identify specific proteins. Among these nine protein bands, five were reliably identified (Table 4.1). For the 1A2-identified protein bands (SI-C-1A2 and SI-D-1A2) (Figure 4.5, S1 and S2), five different proteins were identified from band SI-C-1A2, including two proteins identified as uncharacterized with molecular weights of 26.5kD, 29.8kD, and three identified as 40S ribosomal protein S8, ribosomal protein S4, ribosomal protein S3a with molecular weights of 25.2kD, 29.8kD, 29.9kD respectively (Table 4.1). One protein was identified from the band SI-D-1A2 and it was the 60S ribosomal protein L13 with molecular weight of 24.1kD (Table 4.1). Band RN-1A2 (Figure 4.5, R1 and R2) from resistant plants was identified as uncharacterized with molecular weight of 17.5kD (Table 4.1).

For 1B1 experiments, the only band SI-F-1B1, from the infested susceptible reaction among the five selected proteins, was reliably identified (Figure 4.6, S1 and S2). Three different proteins identified from the band SI-F-1B1, which were an uncharacterized protein with molecular weight of 93.8kD, a lipoxygenase with molecular weight of 89.4kD, and a tubulin alpha chain with molecular weight of 49.7kD (Table 4.1). For band SI-A-1C1 identified from the infested susceptible reaction of 1C1 (Figure 4.7, S1 and S2), two different proteins identified, both uncharacterized proteins with molecular weights of 24.4kD and 27.2kD (Table 4.1).

Discussion

Plant-parasitic insects including Hessian fly and other gall midges use similar mechanisms to attack host plants as observed in plant pathogens. First the larval mandibles have been modified for injection of effectors that is analogous to secretion systems such as type III

secretion pipes in plant bacterial pathogens (Harris et al. 2010). Second larvae inject effector proteins into host tissues during feeding to manipulate and control plant growth and metabolism (Chen et al. 2008, 2010; Zhao et al. 2015; Zhu et al. 2008). Small, secreted proteins have been found to play crucial roles in interactions between biotrophic or hemi-biotrophic pathogens and plants (Lyu et al. 2016). Indeed, SSGPs in SSGP-1 are small peptides with 50-180 amino acid residues and many effector characteristics (Chen et al. 2008, 2010). The conservation of SSGP-1 family members among Hessian fly, barley midge, and oat midge (data haven't published yet), appears to play an important role of these proteins in feeding and virulence. Evidence of effector characteristics in members of SSGP-1 has been provided by several previous studies (Liu et al. 2004; Chen et al. 2004, 2008, 2010).

However, so far none of these peptides have been detected from feeding sites of infested tissues, or insects. Our report is the first to detect five SSGP-1 proteins from infested wheat tissues, dissected tissues of first instar Hessian fly larvae, and other life stages. These proteins have been convincingly identified by immunostaining, western blots, pull down assays, and LC-MS/MS.

To determine the localization of SSGP-1 family members in different tissues of first instar larvae, polyclonal antibodies were used for immunostaining. As expected, all SSGP-1 family proteins were detected in the large cells of the basal region (Figure 4.2, A.1 to A.5), which is linked directly with mandibles. The fact that all cells in the whole base region express these SSGP proteins indicates that there is a high demand for their production due to their importance during the earlier stages of Hessian fly larval infestation. The localization of all SSGP proteins in the basal region, but not in the filament region suggests that they are injected into plant tissues and perform effector functions such as causing wheat plant stunt (Stuart and

Hatchett 1987). Even though all five proteins are localized in the base region of salivary glands in first instar larvae, their timing of expression and localization patterns are slightly different. These differences might indicate structural and functional differences among the five proteins.

Four of the five members of SSGP-1 were expressed in first, second, and third instars except SSGP-1E1 (Figure 4.3, A to E). These findings indicate that these SSGPs may perform other functions in addition for manipulating host plants and support in many aspects the idea that members of SSGP-1 share functional differences by acting as initiators and maintainers (Chen et al. 2010). This evidence supports the hypothesis that extension in protein production among larval stages is necessary to initiate and maintain infestation. Possible roles in maintaining infestation would be to fend off secondary infestation from environmental microbes that may kill wheat seedlings. Wheat plants attacked by Hessian fly larvae are physically weakened and are very vulnerable to secondary attack from surrounding micro-organisms, which may kill wheat plants. Hessian fly larvae require live plants to produce photo-assimilates as nutrition. Alternatively, abundant SSGPs produced in later instar Hessian fly larvae could aid larvae to obtain nutrients from deeper tissues by forming a passage pipe similar to stylet sheaths formed during aphid feeding (Will and Vilcinskas 2015). This possibility is logical since Hessian fly mandibles are very short and can hardly penetrate a single layer of cells. By forming a food passage pipe, Hessian fly larvae can obtain nutrients from surrounding tissues. Consistent with this possibility, the genes in the SSGP-1 family are the mostly abundantly expressed. Nearly 66% of total transcripts identified in Hessian fly salivary glands belong to SSGP-1 family.

The western blot analysis confirmed that these secreted proteins are injected in the plant tissues after 3d of feeding. In general, western blots efficiently detected all five SSGP-1 proteins in infested tissues from both susceptible and resistant wheat plants (Figure 4.4).

Taken together, these lines of evidence suggest effector roles of members of SSGP-1 family, which may include the induction of nutritive cells, suppression of plant defense, and inhibition of plant growth. Similar effector roles for small secreted peptides injected by plant pathogens in suppressing plant immunity system, and facilitating pathogen colonization have been supported by Lyu et al. (2016) and Qi et al. (2016).

Our western blots detected a strong presence of SSGP-1 family proteins at the Hessian fly attack site in resistant plants in comparison with that in susceptible plants (Figure 4.4, A to E). The exact reason for this observation remains to be determined. One of the important defense mechanisms in resistant plants after being attacked by pathogens is to increase the production of lignin or accumulate granular material in the space between the plasma membrane and the outer cell wall (Harris et al. 2010). The strengthened cell walls might prevent or decrease the penetration of effector proteins secreted by parasites to penetrate into deeper tissues, resulting in enhanced accumulation of parasite effector proteins on the plant surface, which can be enriched in western samples. Alternatively, Hessian fly larvae feeding on resistant tissues were under high stress, which might send a signal to the salivary glands to produce more proteins to be injected into resistant plants.

Several lines of evidence from our experiments indicate that SSGP-1 family members are injected in plant tissues by Hessian fly larvae. The *in vitro* pull-down assays performed identified several potential target proteins in 1A2, 1B1, and 1C1. The potential interacting targets (uncharacterized proteins, ribosomal proteins, a lipoxygenase, and a tubulin) await further confirmation in studies of their involvement in different functional pathways.

Conclusion

SSGP-1 family members are the most abundantly expressed proteins in the salivary glands of Hessian fly larvae. This family of proteins is conserved among the three species of gall

midges including Hessian fly, barley midge, and oat midge. Members in this family have been investigated in many studies, and much evidence has been provided about their important characteristics (Liu et al. 2004; Chen et al. 2004, 2008, 2010). In this study, we report new evidence for the first time about the role of five SSGP-1 family proteins from infested tissues that point to the direction that these proteins function as effectors and play crucial roles in Hessian fly infestation and virulence. Putative interacting targets for family members 1A2, 1B1, and 1C1 were identified by LC-MS/MS, and are available for further examination in the future.

References

Chen, M.S., Liu, X., Yang, Z., Zhao, H., Shukle, R.H., Stuart, J.J. and Hulbert, S. 2010. Unusual conservation among genes encoding small secreted salivary gland proteins from a gall midge. BMC evolutionary biology, 10(1), p.296.

Chen, M.S., Echegaray, E., Whitworth, R.J., Wang, H., Sloderbeck, P.E., Knutson, A., Giles, K.L. and Royer, T.A. 2009. Virulence analysis of Hessian fly populations from Texas, Oklahoma, and Kansas. Journal of economic entomology, 102(2), 774-780.

Chen, M.S., Zhao, H.X., Zhu, Y.C., Scheffler, B., Liu, X., Liu, X., Hulbert, S. and Stuart, J.J. 2008. Analysis of transcripts and proteins expressed in the salivary glands of Hessian fly (Mayetiola destructor) larvae. Journal of insect physiology, 54(1), 1-16.

Chen, M.S., Fellers, J.P., Zhu, Y.C., Stuart, J.J., Hulbert, S., El-Bouhssini, M. and Liu, X. 2006. A super-family of genes coding for secreted salivary gland proteins from the Hessian fly, Mayetiola destructor. Journal of Insect Science, 6(1).

Chen, M.S., Fellers, J.P., Stuart, J.J., Reese, J.C. and Liu, X. 2004. A group of related cDNAs encoding secreted proteins from Hessian fly [Mayetiola destructor (Say)] salivary glands. Insect Molecular Biology, 13(1), 101-108.

Gagné, R.J., Hatchett, J.H., Lhaloui, S., Bouhssini. M. 1991. Hessian fly and barley stem gall midge, two different species of Mayetiola (Diptera: Cecidomyiidae) in Morocco. Annals of the Entomological Society of America, 84(4), 436-443.

Hatchett, J.H., Kreitner, G.L. and Elzinga, R.J. 1990. Larval mouthparts and feeding mechanism of the Hessian fly (Diptera: Cecidomyiidae). Annals of the Entomological Society of America, 83(6), 1137-1147.

Harris, M.O., Freeman, T.P., Moore, J.A., Anderson, K.G., Payne, S.A., Anderson, K.M. and Rohfritsch, O. 2010. H-gene-mediated resistance to Hessian fly exhibits features of penetration resistance to fungi. Phytopathology, 100(3), 279-289.

Harris, M.O., Freeman, T.P., Rohfritsch, O., Anderson, K.G., Payne, S.A. and Moore, J.A. 2006. Virulent Hessian fly (Diptera: Cecidomyiidae) larvae induce a nutritive tissue during compatible interactions with wheat. Annals of the Entomological Society of America, 99(2), 305-316.

Liu, X., Fellers, J.P., Wilde, G.E., Stuart, J.J. and Chen, M.S. 2004. Characterization of two genes expressed in the salivary glands of the Hessian fly, Mayetiola destructor (Say). Insect biochemistry and molecular biology, 34(3), 229-237.

Lyu, X., Shen, C., Fu, Y., Xie, J., Jiang, D., Li, G. and Cheng, J. 2016. A small secreted virulence-related protein is essential for the necrotrophic interactions of Sclerotinia sclerotiorum with its host plants. PLoS pathogens, 12(2), p.e1005435.

Qi, M., Link, T.I., Müller, M., Hirschburger, D., Pudake, R.N., Pedley, K.F., Braun, E., Voegele, R.T., Baum, T.J. and Whitham, S.A. 2016. A small cysteine-rich protein from the Asian soybean rust fungus, Phakopsora pachyrhizi, suppresses plant immunity. PLoS pathogens, 12(9), p.e1005827.

Stuart, J.J. and Hatchett, J.H. 1987. Morphogenesis and cytology of the salivary gland of the Hessian fly, Mayetiola destructor (Diptera: Cecidomyiidae). Annals of the Entomological Society of America, 80(4), 475-482.

Šimo, L., Slovák, M., Park, Y. and Žitňan, D. 2009. Identification of a complex peptidergic neuroendocrine network in the hard tick, Rhipicephalus appendiculatus. Cell and tissue research, 335(3), 639-655.

Will, T. and Vilcinskas, A. 2015. The structural sheath protein of aphids is required for phloem feeding. Insect biochemistry and molecular biology, 57, 34-40.

Zhao, C., Escalante, L.N., Chen, H., Benatti, T.R., Qu, J., Chellapilla, S., Waterhouse, R.M., Wheeler, D., Andersson, M.N., Bao, R. and Batterton, M. 2015. A massive expansion of effector genes underlies gall-formation in the wheat pest Mayetiola destructor. Current Biology, 25(5), 613-620.

Zhu, L., Liu, X., Liu, X., Jeannotte, R., Reese, J.C., Harris, M., Stuart, J.J. and Chen, M.S., 2008. Hessian fly (Mayetiola destructor) attack causes a dramatic shift in carbon and nitrogen metabolism in wheat. Molecular Plant-Microbe Interactions, 21(1), 70-78.

Figure 4.1 Amino acid sequence alignment of members belong to family SSGP-1, the boundary between the secreted signal peptide and the mature protein indicated by the arrow

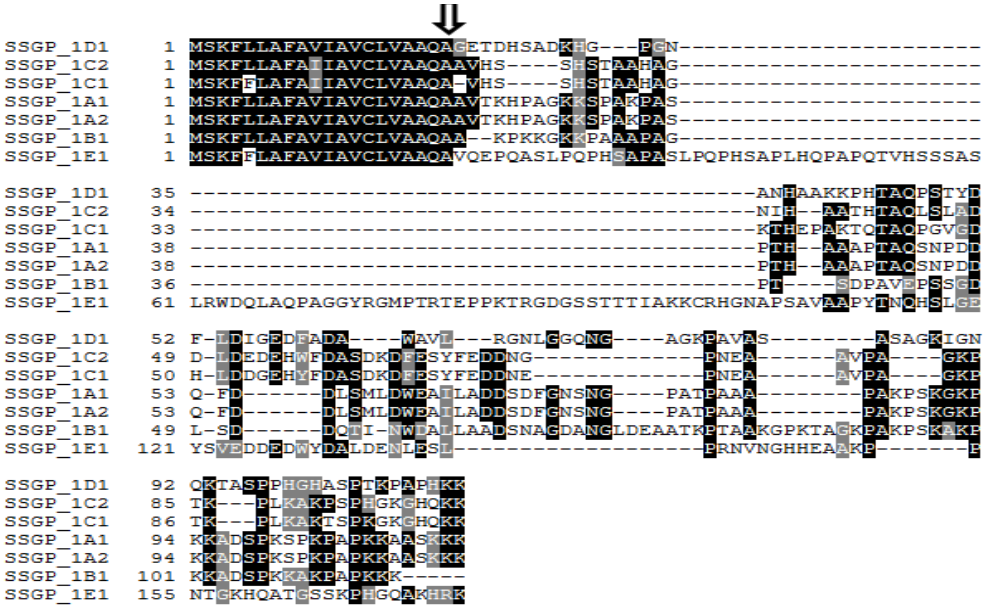
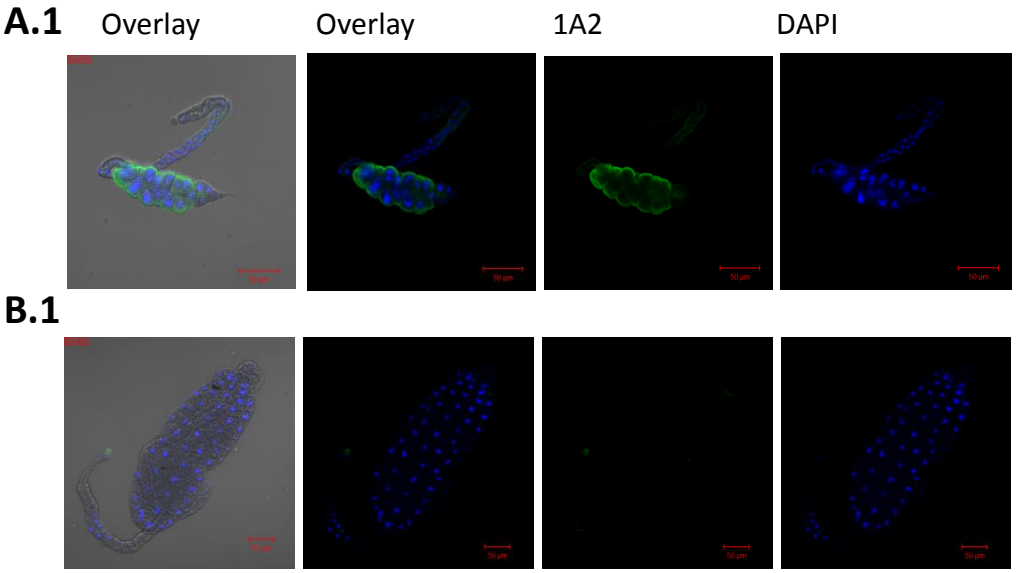
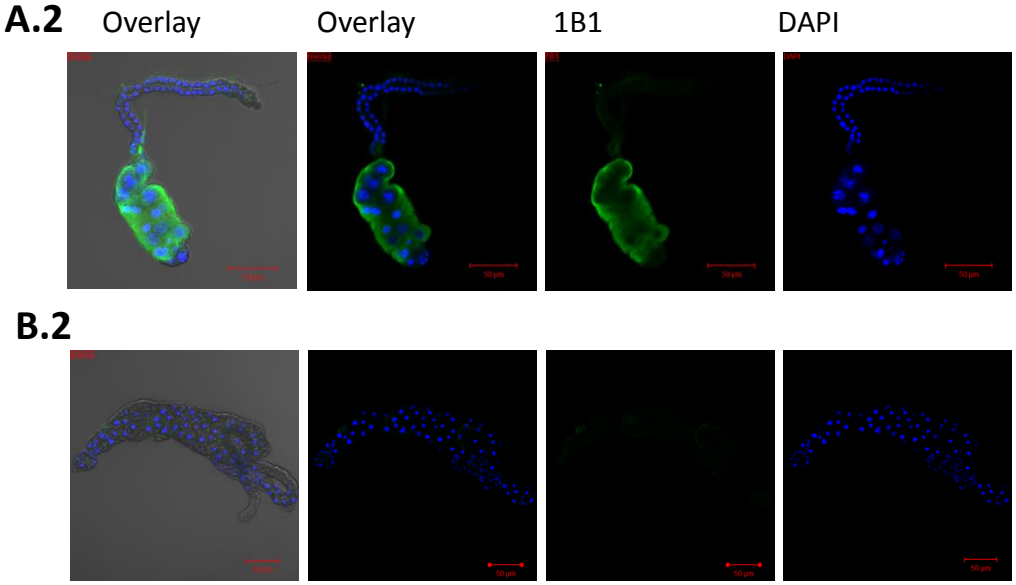
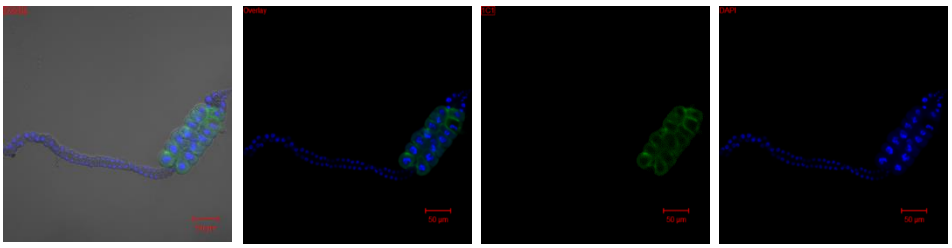


Figure 4.2 Antibody staining of the five members of SSGP-1 from dissected tissues of first instar larvae. A.1 to A.5 Antibody staining for salivary glands of the five members. B.1 to B.5 Antibody staining for the gut and Malpighian tubules of the five members. The green color is the antibody staining 1A2,1B1, 1C1, 1D1, 1E1; the blue color is the nucleus staining DAPI. Overlay combines both staining together. B.5.2 Enlarged view for the foregut images from B.5.1.

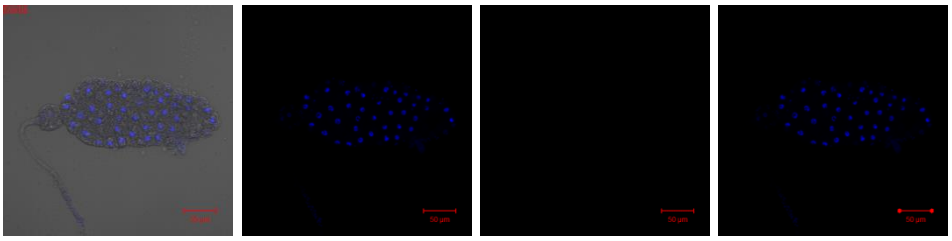




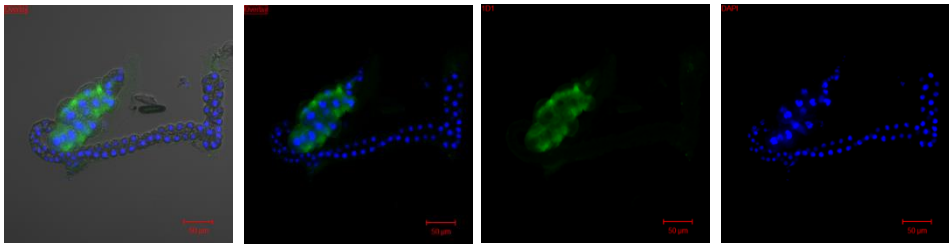
A.3 Overlay Overlay 1C1 DAPI



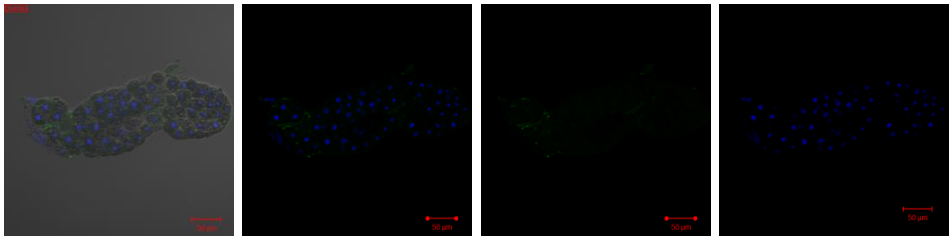
B.3



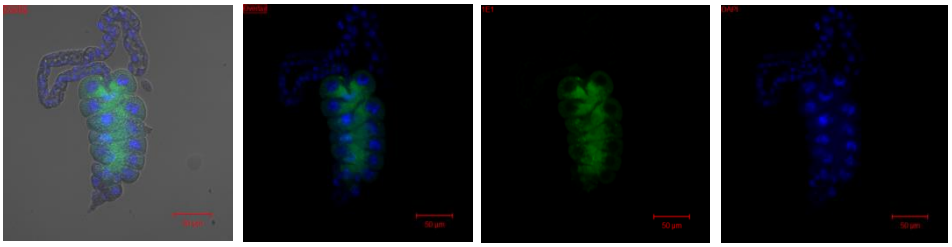
A.4 Overlay Overlay 1D1 DAPI



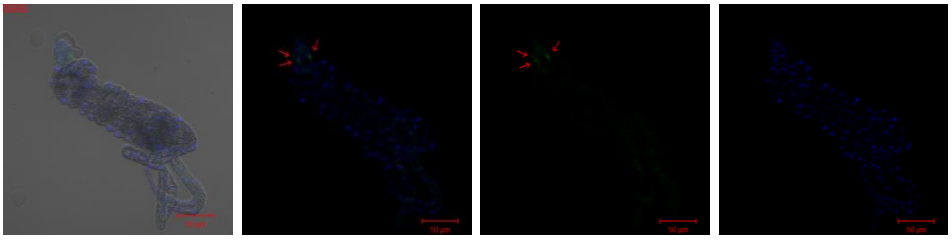
B.4



A.5 Overlay Overlay 1E1 DAPI



B.5.1



B.5.2

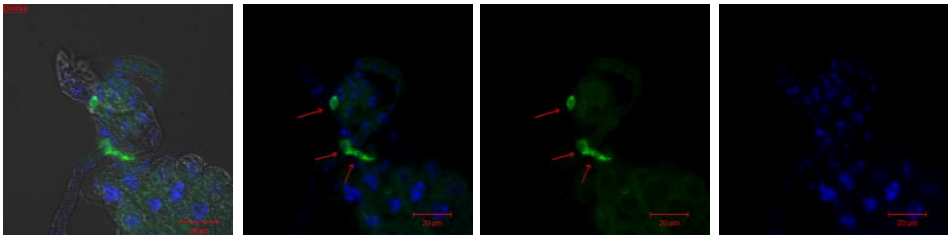
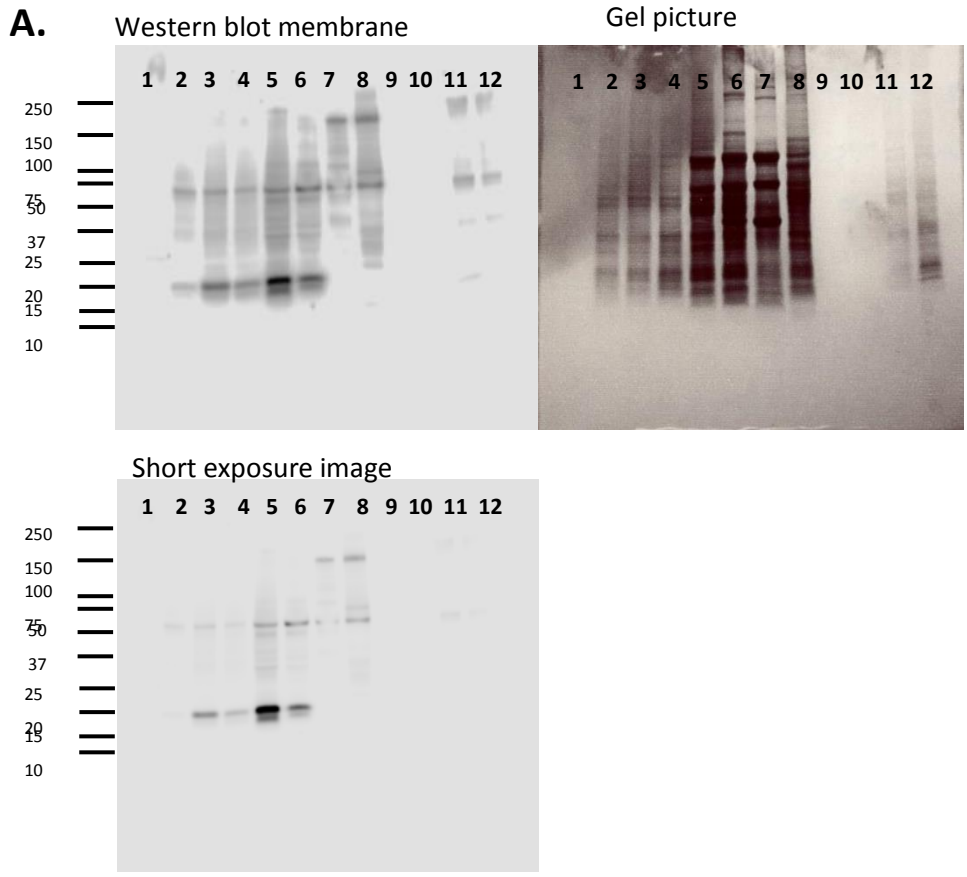
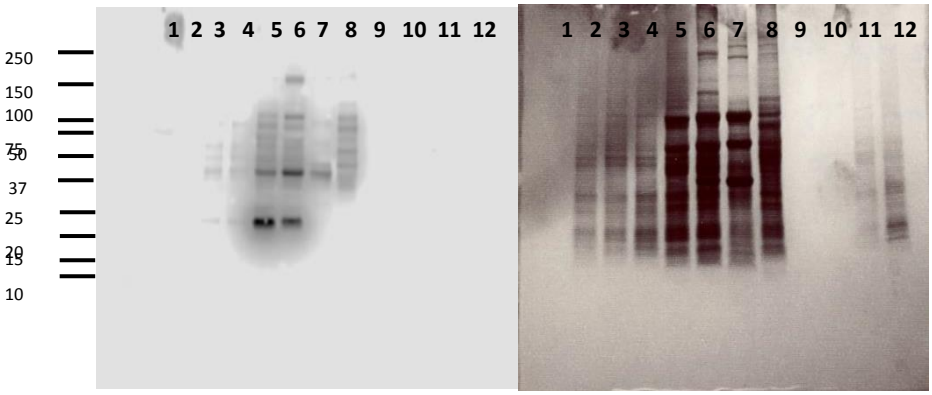


Figure 4.3 A to E, Detection of five proteins in protein extracts derived through Western blots from Hessian fly larvae at different ages on susceptible plant, and at three ages on resistant plant, plus adult and pupae stages. Number of lanes in each photo corresponds to: 1) Molecular marker. 2) Larvae in age 1day, 1st instar (susceptible plant). 3) Larvae in age 2 days, 1st instar (susceptible plant).4) Larvae in age 3 days, 1st instar (susceptible plant). 5) Larvae in age 6 days, 2nd instar (susceptible plant).6) Larvae in age 10 days, 3rd instar (susceptible plant). 7) Pupae stage. 8) Adult stage. 9) Blank lane. 10) Larvae in age 1 day, 1st instar (resistant plant). 11) Larvae in age 2 days, 1st instar (resistant plant). 12) Larvae in age 3 days, 1st instar (resistant plant). The short exposure image is for the same Western blot membrane but with less scanned exposure.

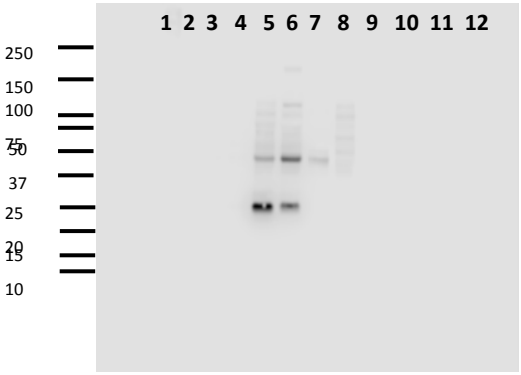


B. Western blot

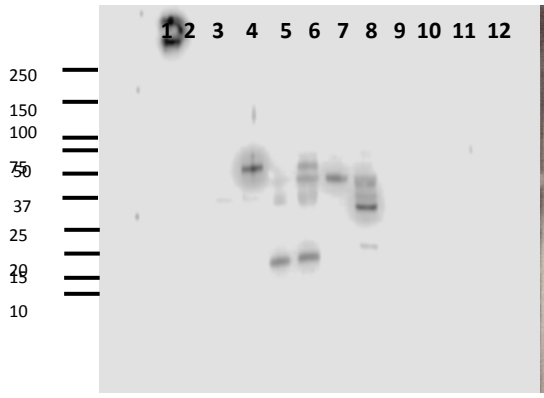


Gel picture

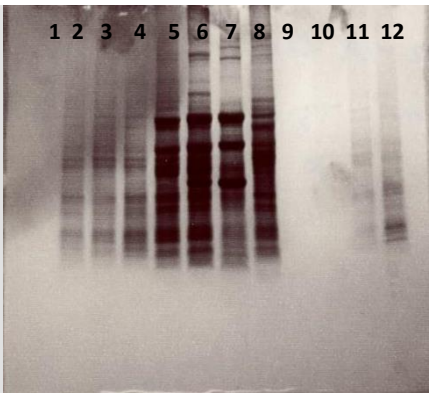
Short exposure image



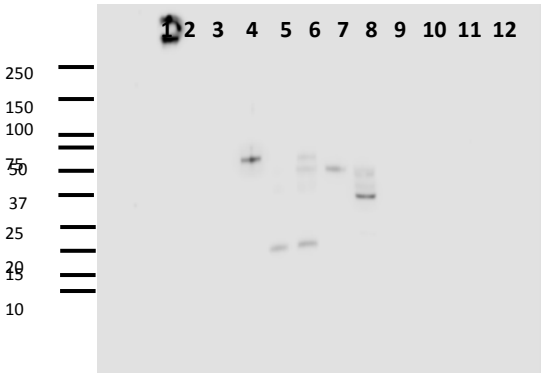
C. Western blot



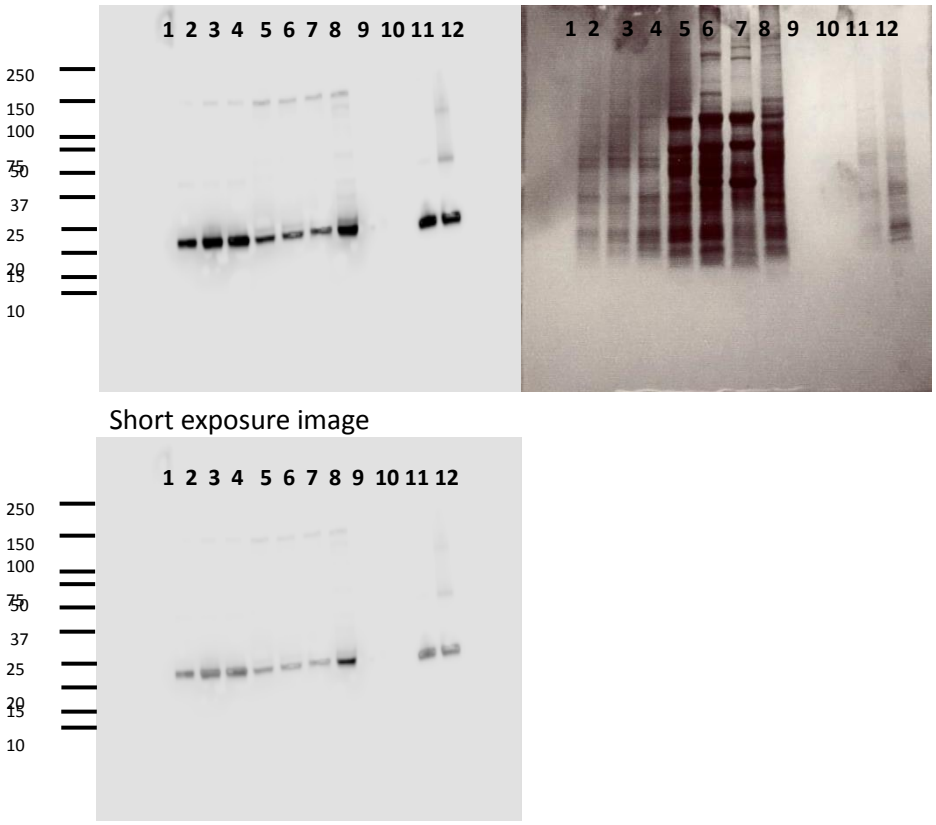
Gel picture



Short exposure image



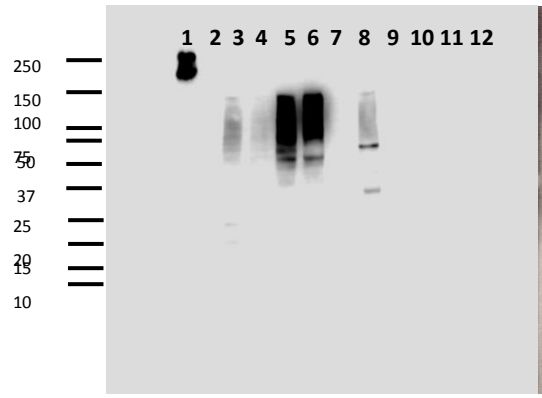
D. Western blot



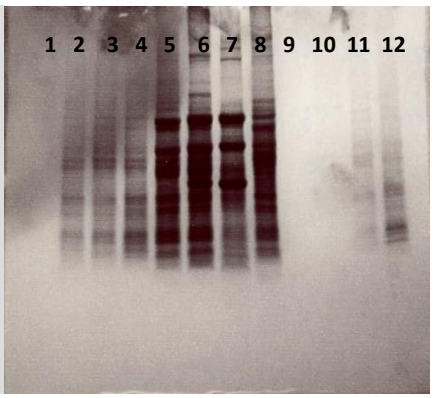
Gel picture

Short exposure image

E. Western blot



Gel picture



Short exposure image

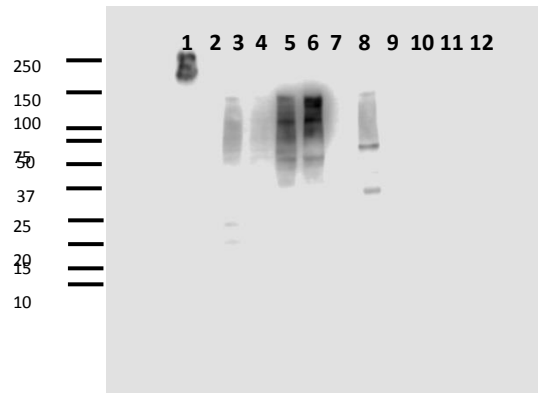
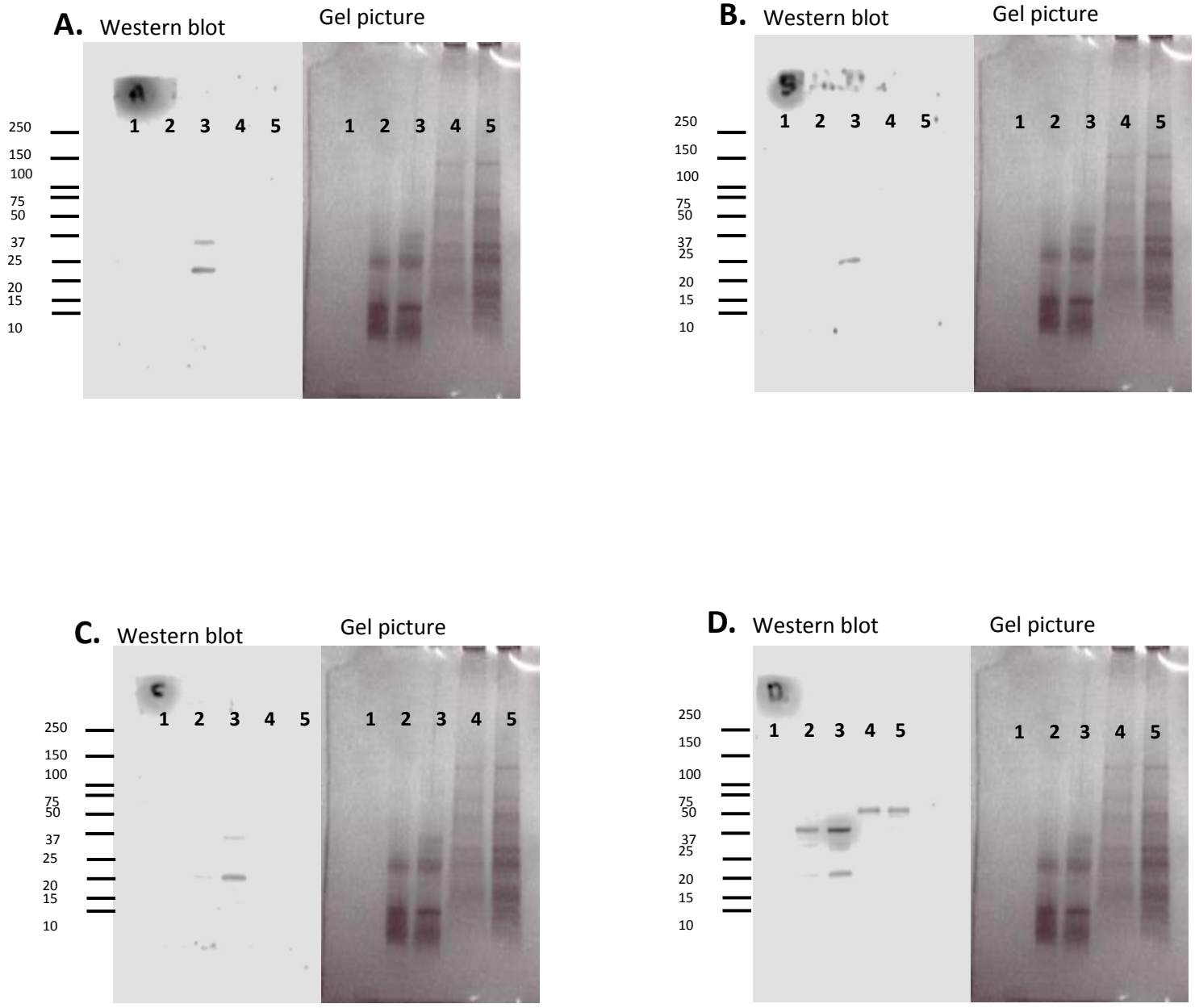


Figure 4.4 A to E, Detection of five proteins in protein extracts derived through Western blots from infested susceptible and resistant wheat tissues at the feeding site after 3days of successive feeding. Number of lanes in each photo corresponds to: 1) Molecular marker. 2) Infested susceptible tissues. 3) Infested resistant tissues. 4) Control; non-infested susceptible tissues. 5) Control, non-infested resistant tissues.



E. Western blot Gel picture

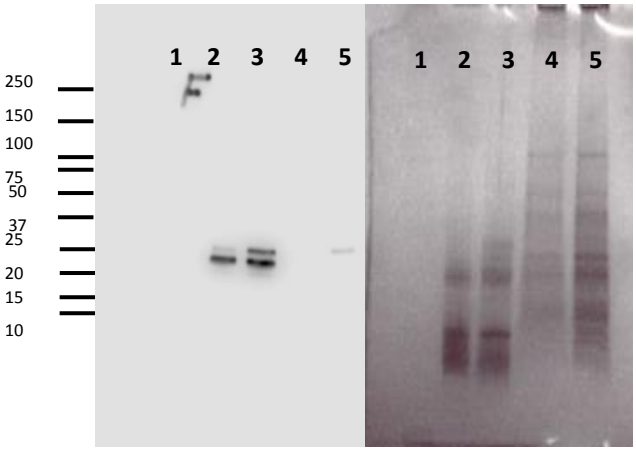


Figure 4.5 Detection of probable targets via Pull down assay. SDS page stained with silver staining was used to detect unique bands of interactions between purified recombinant proteins **(Bait 1A2)** from Hessian fly with protein extracts from different wheat treatments. S1 (unmarked) and S2 (marked) represent susceptible interactions of infested and non-infested tissues. 1) Molecular marker. 2) Eluted bait without interaction. 3) Ignore due to different preparation method 4) Interacted bait with extract from infested susceptible tissues. 5) Interacted bait with extract from non-infested susceptible tissues. 6) Correspond to lane 3, ignore. 7) Non-interacted eluted of infested susceptible tissues only (control). 8) Non- interacted eluted of non-infested susceptible wheat (control). R1 (unmarked) and R2 (marked) represent resistant interactions of infested and non-infested tissues. Number of lanes in each photo corresponds to 1) Molecular marker. 2) Eluted bait without interaction. 3) Interacted bait with extract from infested resistant tissues. 4) Non- interacted eluted of infested resistant tissues only (control). 5) Interacted bait with extract from non-infested resistant tissues. 6) Non- interacted eluted of non-infested resistant wheat (control)

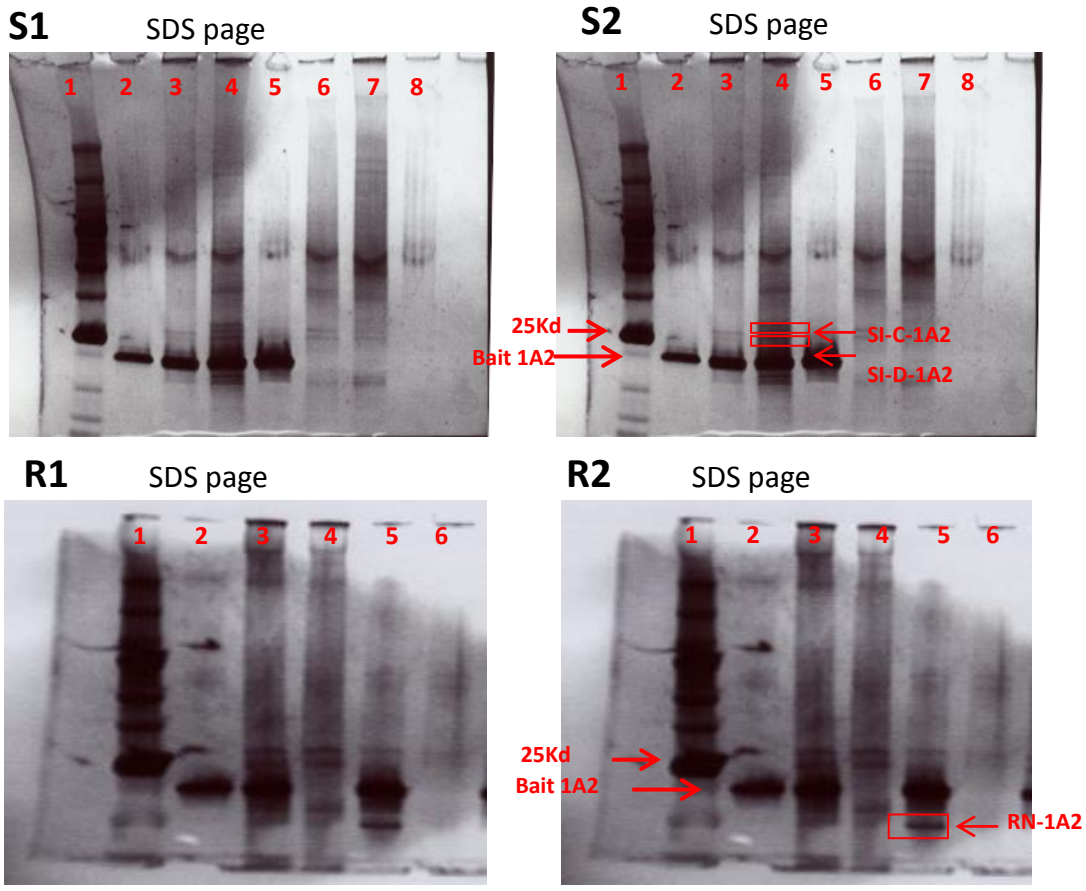


Figure 4.6 Detection of probable targets to 1B1 via Pull down assay. SDS page stained with silver staining was used to detect unique bands of interactions between purified recombinant proteins **(Bait 1B1)** from Hessian fly with protein extracts from different wheat treatments. S1 (unmarked) and S2 (marked) represent susceptible interactions of infested and non-infested tissues. 1) Molecular marker. 2) Eluted bait without interaction. 3) Interacted bait with extract from infested susceptible tissues. 4) Non- interacted eluted of infested susceptible tissues only (control). 5) Interacted bait with extract from non-infested susceptible tissues. 6. Non- interacted eluted of non-infested susceptible wheat (control). R1 (unmarked) and R2 (marked) represent resistant interactions of infested and non-infested tissues. 1) Molecular marker. 2) Eluted bait without interaction. 3) Interacted bait with extract from infested resistant tissues. 4) Non-interacted eluted of infested resistant tissues only (control). 5) Interacted bait with extract from non-infested resistant tissues. 6. Non- interacted eluted of non-infested resistant wheat (control)

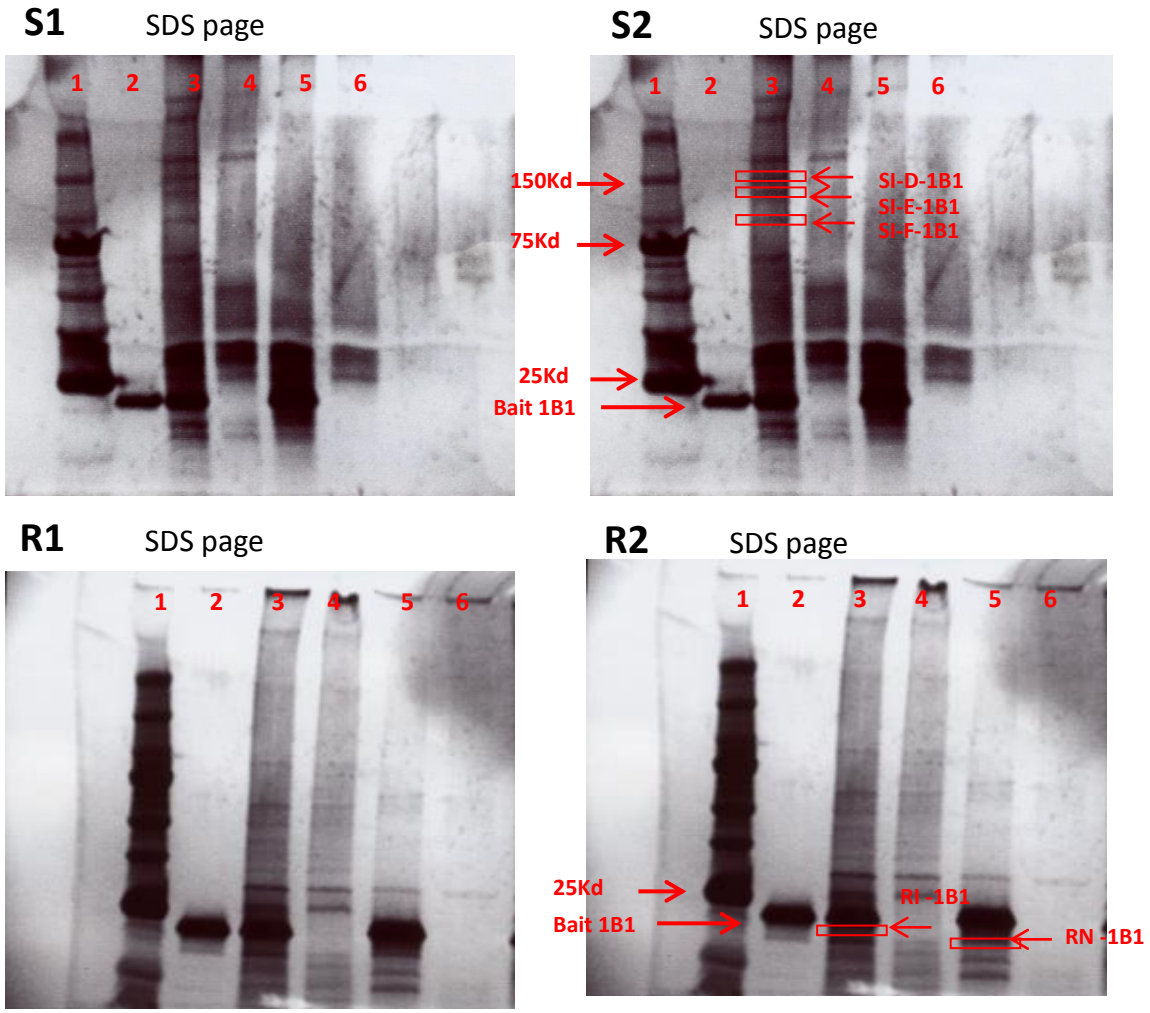


Figure 4.7 Detection of probable targets to 1C1 via Pull down assay. SDS page stained with silver staining was used to detect unique bands of interactions between purified recombinant proteins **(Bait 1C1)** from Hessian fly with protein extracts from different wheat treatments. S1 (unmarked) and S2 (marked) represent susceptible interactions of infested and non-infested tissues. 1) Molecular marker. 2) Eluted bait without interaction. 3) Interacted bait with extract from infested susceptible tissues. 4) Non- interacted eluted of infested susceptible tissues only (control). 5) Interacted bait with extract from non-infested susceptible tissues. 6. Non- interacted eluted of non-infested susceptible wheat (control).

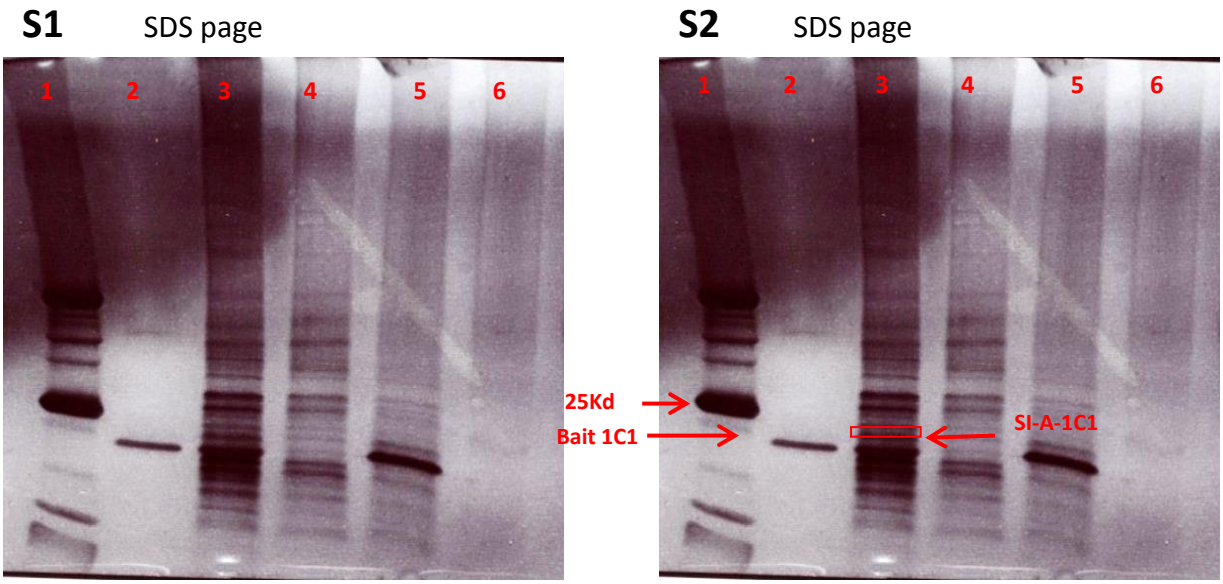


Table 4.1 Identification of putative target proteins that interacted with one of the five SSGP-1 members by LC-MS/MS. Protein bands represent the following reactions; SI-C-1A2 and SI-D-1A2, are bands of susceptible infested tissues interacted with 1A2. RN-1A2, is band of resistant non-infested tissues interacted with 1A2. SI-F-1B1, is band of susceptible infested tissues interacted with 1B1. SI-A-1C1, is band of susceptible infested tissues interacted with 1C1. MW, is molecular weight.

Protein bands	First Hit Accession	First Hit Protein Name [species]	Database MW	Observed MW
SI-C-1A2	A0A1D6CDE5	Uncharacterized protein, *Triticum aestivum*	26.5kd	30kd
	A0A1D5YTD0	Uncharacterized protein, *Triticum aestivum*	29.8kd	
	A0A1D5UKZ4	40S ribosomal protein S8 *Triticum aestivum*	25.2kd	
	A0A1D5SLI4	40S ribosomal protein S4 *Triticum aestivum*	29.8kd	
	A0A1D6RVF1	40S ribosomal protein S3a*Triticum aestivum*	29.9kd	
SI-D-1A2	A0A1D6DIT0	60S ribosomal protein L13 *Triticum aestivum*	24.1kd	28kd
RN-1A2	A0A1D5YSK2	Uncharacterized protein, *Triticum aestivum*	17.5kd	15kd
SI-F-1B1	A0A1D6C799	Uncharacterized protein, *Triticum aestivum*	93.8kd	85kd
	A0A1D5YDJ7	Lipoxygenase *Triticum aestivum*	89.4kd	
	Q9ZRB7|TBA	Tubulin alpha chain *Triticum aestivum*	49.7kd	
SI-A-1C1	W5GVR2	Uncharacterized protein, *Triticum aestivum*	24.4kd	25kd
	A0A1D5YHB0	Uncharacterized protein, *Triticum aestivum*	27.2kd	

Appendix A - Supplementary data, Chapter 2

Table S1. Distribution of SSGP transcripts among different groups belong to wheat midge (transcripts in each group and their related proteins are listed in order in Figure S1)

Gene	Number of unique sequences	Number of total sequences	First BLAST hit if any
Group 1	48	99	
Group 2	5	7	
Group 3	10	12	ankyrin repeat-containing protein [Persephonella marina]
Group 4	18	20	ankyrin repeat protein [Trichomonas vaginalis G3]
Group 5	1	1	
Group 6	1	1	
Group 7	1	1	
Group 8	1	1	
Group 9	1	1	
Group 10	1	1	
Group 11	2	3	
Group 12	5	13	
Group 13	8	10	
Group 14	3	3	
Group 15	1	1	
Group 16	1	1	
Group 17	1	1	
Group 18	1	1	
Group 19	1	1	
Group 20	1	1	
Group 21	1	1	putative ubiquinol-cytochrome c reductase complex
Group 22	2	2	
Group 23	2	2	
Group 24	5	8	
Group 25	1	1	
Group 26	1	1	
Group 27	2	2	
Group 28	1	2	
Group 29	4	14	
Group30	1	1	
Group31	1	1	
Group32	1	1	
Group33	1	1	
Group34	1	1	
Group35	1	1	

Group36	1	1	
Group37	1	1	
Group38	2	2	
Group39	1	5	
Group40	5	9	
Group41	2	2	
Group42	1	1	
Group43	3	3	
Group44	1	3	larval cuticle protein 8-like [Drosophila eugracilis] have signal peptide
Group45	6	6	
Group46	4	4	
Group47	1	1	
Group48	1	1	
Group49	1	1	larval cuticle protein 8-like [Bactrocera cucurbitae]
Group50	1	1	
Group51	2	2	
Group52	2	2	
Group53	1	1	
Group54	1	1	
Group55	1	1	
Group56	1	1	
Group57	2	2	
Group58	1	1	
Group59	1	1	
Group60	1	1	
Group61	2	2	
Group62	2	2	
Group63	2	2	
Group64	2	2	
Group65	2	2	
Group66	1	1	
Group67	9	12	
Group68	1	1	
Group69	1	1	
Group70	1	1	
Group71	3	3	
Group72	1	1	
Group73	1	1	
Group74	1	1	
Group75	1	1	
Group76	1	1	
Group77	3	3	
Group78	1	1	
Group79	2	2	

Group80	1	1	
Group81	1	1	
Group82	1	1	
Group83	1	1	
Group84	1	1	Heat shock 70 kDa protein cognate 3, partial [Ceratitis capitata]
Group85	1	1	ankyrin-3-like, partial [Rhagoletis zephyria]
Group86	1	1	uncharacterized protein Dwil_GK12749 [Drosophila willistoni]
Group87	1	1	Serpin B8, partial [Bactrocera latifrons]
Group88	1	1	salivary/fat body serine carboxypeptidase [Sitodiplosis mosellana]
Group89	1	1	carboxypeptidase B [Mayetiola destructor]
Group90	1	1	flexible cuticle protein 12-like [Aethina tumida]
Group91	1	1	AAEL004120-PA [Aedes aegypti]
Group92	1	1	putative lysosomal thioesterase ppt2 [Culex tarsalis]
Group93	1	1	
Group94	1	1	Ferritin, lower subunit [Anthurium amnicola]
Group95	1	1	carboxypeptidase A [Mayetiola destructor]
Group96	1	1	Sarcophaga pro-cathepsin B [Sarcophaga peregrina]
Group97	1	1	AAEL001498-PA [Aedes aegypti]
Total	235	330	

Table S2. Classification of proteins encoded by non-SSGP transcripts, wheat midge.

Group #	No.	First hit	E-value
A. Protein synthesis and folding			
1. Ribosomal protein			
20	2	40s ribosomal protein s6, partial [Nyssomyia neivai]	5e-57
22	2	ribosomal protein L11 [Bombyx mori]	3e-105
35	3	60s ribosomal protein l27a [Tabanus bromius]	9e-74
39	17	ribosomal protein L23 [Drosophila melanogaster]	2e-77
40	4	40S ribosomal protein S20 [Culex quinquefasciatus]	3e-71
41	2	40S ribosomal protein S23, partial [Ceratitis capitata]	6e-99
43	2	40s ribosomal protein s8, partial [Tabanus bromius]	3e-108
45	3	40s ribosomal protein s13 [Tabanus bromius]	9e-101
59	2	40S ribosomal protein S19a [Lucilia cuprina]	2e-70
63	6	60s ribosomal protein l39, partial [Haematobia irritans]	2e-22
93	2	ribosomal protein s26e [Nyssomyia neivai]	3e-66
117	2	60S ribosomal protein L37-A [Rhagoletis zephyria]	2e-26
121	3	60s ribosomal protein l37a [Haematobia irritans]	1e-33
CN6	1	ribosomal protein S15 [Diaphorina citri]	1e-62
CN19	1	ribosomal protein l3 isoform a [Tabanus bromius]	4e-111
CN26	1	60S ribosomal protein L19 [Drosophila busckii]	2e-34
CN114	1	ribosomal protein S27A [Drosophila melanogaster]	2e-78
CN121	1	ribosomal protein l23a, partial [Tabanus bromius]	5e-50
CN122	1	60s ribosomal protein l26 [Nyssomyia neivai]	2e-81
CN138	1	40s ribosomal protein s4 [Nyssomyia neivai]	5e-160
CN140	1	40S ribosomal protein S16 [Bactrocera cucurbitae]	6e-94
CN169	1	40s ribosomal protein s12 [Nyssomyia neivai]	6e-57
CN199	1	ribosomal protein s21e [Tabanus bromius]	1e-43
CN200	1	40S ribosomal protein S14, partial [Bactrocera latifrons]	2e-64
CN232	1	40S ribosomal protein S24 [Rhagoletis zephyria]	2e-60
CN236	1	ribosomal protein L38 [Psathyromyia shannoni]	2e-35
CN240	1	ribosomal protein L17 [Drosophila melanogaster]	1e-109
CN252	1	ribosomal protein l31 [Haematobia irritans]	6e-67
CN254	1	60s ribosomal protein l6 [Nyssomyia neivai]	3e-74
MIDGESAL_001_I23.F1	1	40S ribosomal protein S18-like [Rhagoletis zephyria]	8e-12
MIDGESAL_002_C02.F1	1	60S ribosomal protein L24 [Drosophila suzukii]	9e-46
MIDGESAL_002_C20.F1	1	ribosomal protein L32, isoform C [Drosophila melanogaster]	4e-84
MIDGESAL_002_N19.F1	1	ribosomal protein l18 isoform b [Tabanus bromius]	1e-99
1501114	1	60S ribosomal protein L21 [Musca domestica]	1e-94
7612M13F21	1	50S ribosomal protein L14 (chloroplast) [Xerophyllum tenax]	8e-15
CN23	1	60S ribosomal protein L35 [Drosophila serrata]	8e-65
CN93	1	40S ribosomal protein S3a [Rhagoletis zephyria]	2e-18
CN131	1	60S ribosomal protein L9 [Lucilia cuprina]	7e-120

CN204	1	40S ribosomal protein S28 [Tribolium castaneum]	2e-27
CN262	1	60s ribosomal protein l36-like isoform x1 [Tabanus bromius]	2e-44
MIDGESAL_002_G12.F1	1	ribosomal protein L4, partial [Drosophila albomicans]	3e-59
MIDGESAL_003_C23.F1	1	60S acidic ribosomal protein P0 isoform X2 [Rhagoletis zephyria]	4e-136
MIDGESAL_003_F04.F1	1	60S acidic ribosomal protein P1 [Plutella xylostella]	2e-35
MIDGESAL_003_E03.F1	1	60s ribosomal protein l35a [Tabanus bromius]	3e-58
MIDGESAL_003_H15.F1	1	40S ribosomal protein S3a-like [Plutella xylostella]	7e-54
MIDGESAL_003_K11.F1	1	40S ribosomal protein S10b, partial [Ceratitis capitata]	3e-57
MIDGESAL_003_N24.F1	1	probable 28S ribosomal protein S16, mitochondrial [Drosophila busckii]	2e-45
MIDGESAL_004_E05.F1	1	60S ribosomal protein L10a isoform X2 [Linepithema humile]	4e-66
MIDGESAL_004_J12.F1	1	28S ribosomal protein S11, mitochondrial [Aethina tumida]	3e-56
MIDGESAL_004_K05.F1	1	60s ribosomal protein l22, partial [Nyssomyia neivai]	7e-40
MIDGESAL_006_C01.F1	1	40s ribosomal protein s27 [Tabanus bromius]	2e-52
CN96	1	40s ribosomal protein sa, partial [Corethrella appendiculata]	8e-136
MIDGESAL_006_K07.F1	1	ribosomal protein L36A, isoform A [Drosophila melanogaster]	7e-54
MIDGESAL_007_I11.F1	1	ribosomal protein L10 [Riptortus pedestris]	4e-103
MIDGESAL_007_P01.F1	1	mitochondrial ribosomal protein L50 [Anopheles darlingi]	3e-48
MIDGESAL_008_C04.F1	1	S17e ribosomal protein [Dascillus cervinus]	1e-66
MIDGESAL_008_O18.F1	1	60s ribosomal protein l27 [Nyssomyia neivai]	1e-61
MIDGESAL_008_I06.F1	1	60S ribosomal protein L8-like [Diuraphis noxia]	4e-31
MIDGESAL_005_K20.F1	1	39S ribosomal protein L13, mitochondrial-like [Aedes albopictus]	6e-41
CN249	1	60S acidic ribosomal protein P2 [Drosophila busckii]	3e-28
18312M13F21	1	50s ribosomal protein l4 [Haemophilus parainfluenzae]	2e-27
105	2	ribosome-associated membrane protein 4 [Antheraea yamamai]	9e-33
2. Transcription & translation factors			
47	2	translationally-controlled tumor protein homolog [Ceratitis capitata]	2e-38
14	2	elongation factor 1 gamma [Aedes albopictus]	3e-77
7	3	regulator of rdna transcription protein 15, [Amblyomma aureolatum]	2e-34
84	2	reverse transcriptase [Zingiber officinale]	3e-15
CN198	1	Translation initiation factor 5a [Tabanus bromius]	1e-105
27612M13F21	1	reverse transcriptase [Zingiber officinale]	7e-5
CN250	1	elongation factor-1 alpha, partial [Corcyra cephalonica]	1e-14
MIDGESAL_001_M08.F1	1	elongation factor 1-beta [Lucilia cuprina]	6e-100
MIDGESAL_003_D17.F1	1	transcription elongation factor B polypeptide 2 [Musca domestica]	5e-54
MIDGESAL_003_L20.F1	1	elongation factor-1 alpha, partial [Lethocerus deyrollei]	1e-83
MIDGESAL_003_M18.F1	1	PR domain zinc finger protein 4 isoform X1 [Ictalurus punctatus]	5e-9
MIDGESAL_005_B12.F1	1	Transcription elongation factor SPT4 [Lucilia cuprina]	2e-31
MIDGESAL_005_O09.F1	1	eukaryotic translation initiation factor 3 subunit 9, partial [Aedes albopictus]	7e-33
MIDGESAL_006_G03.F1	1	cellular repressor of transcription, partial [Aedes albopictus]	5e-54
MIDGESAL_006_H06.F1	1	transcriptional coactivator yap1 isoform x1, partial [Nyssomyia neivai]	3e-29
MIDGESAL_008_D04.F1	1	translationally-controlled tumor protein homolog [Drosophila busckii]	1e-44
MIDGESAL_008_H24.F1	1	signal transducer and activator of transcription 5B isoform X6 [Pseudomyrmex gracilis]	2e-69

MIDGESAL_007_D15.F1	1	translation initiation factor 5B-like isoform X1 [Habropoda laboriosa]	5e-12
MIDGESAL_004_L17.F1	1	eukaryotic translation initiation factor [Anopheles aquasalis]	1e-17
CN221	1	zinc finger matrin-type protein CG9776-like [Aedes albopictus]	3e-11
11114	1	c2h2-type zn-finger protein [Tabanus bromius]	3e-53
37	7	reverse transcriptase [Zingiber officinale]	5e-25
3. Chaperones, protein folding and secretion			
116	2	ankyrin-3-like, partial [Rhagoletis zephyria]	2e-32
CN81	1	Heat shock protein mitochondrial [Nyssomyia neivai]	2e-46
120	2	Heat shock protein 70, partial [Calliphora vicina]	4e-18
CN43	1	ankyrin-3-like, partial [Rhagoletis zephyria]	5e-37
CN47	1	ankyrin-3-like, partial [Rhagoletis zephyria]	4e-8
MIDGESAL_004_D22.F1	1	molecular chaperone dnaj superfamily [Corethrella appendiculata]	7e-48
MIDGESAL_004_J03.F1	1	Heat shock protein 27 [Lucilia cuprina]	2e-13
MIDGESAL_006_N23.F1	1	chaperonin complex component tcp-1 epsilon subunit cct5 [Psorophora albipes]	4e-114
MIDGESAL_004_I21.F1	1	ankyrin repeat and protein kinase domain-containing protein 1 [Cricetulus griseus]	4e-15
MIDGESAL_002_B20.F1	1	prefoldin subunit 1 [Stomoxys calcitrans]	1e-41
4. RNA synthesis and processing			
MIDGESAL_001_L02.F1	1	peptidyl-tRNA hydrolase 2 mitochondrial [Culex tarsalis]	4e-32
MIDGESAL_002_M14.F1	1	asparagine--tRNA ligase, mitochondrial [Diachasma alloeum]	1e-44
MIDGESAL_003_D20.F1	1	U2 small nuclear ribonucleoprotein A [Anopheles darlingi]	6e-99
MIDGESAL_004_E06.F1	1	RNA polymerase ii transcription subunit 21 [Tabanus bromius]	8e-66
MIDGESAL_006_O06.F1	1	small nuclear ribonucleoprotein Sm D2 [Exaiptasia pallida]	5e-58
MIDGESAL_005_A22.F1	1	U6 snRNA-associated Sm-like protein LSm3 [Aedes albopictus]	2e-41
MIDGESAL_001_H19.F1	1	RNA-binding protein nova1/pasilla [Nyssomyia neivai]	2e-92
1641117	1	anti-RNA polymerase sigma factor SigE [Escherichia coli]	9e-87
B. House keeping **1. Energy metabolism**			
106	2	ATP synthase lipid-binding protein, mitochondrial [Musca domestica]	2e-65
137118	1	cytochrome b-c1 complex subunit 8 [Dendroctonus ponderosae]	2e-25
331114	1	formate dehydrogenase- seleno polypeptide subunit [Shigella sonnei	2e-172
CN78	1	cytochrome oxidase subunit 1, partial (mitochondrion) [Cecidomyiidae sp.]	1e-81
CN 85	1	cytochrome oxidase subunit I, partial (mitochondrion) [Asteromyia modesta]	1e-71
CN119	1	ATP synthase subunit b, mitochondrial [Lucilia cuprina]	5e-98
CN137	1	ATP synthase F0 subunit 6 (mitochondrion) [Mayetiola destructor]	3e-31
CN156	1	cytochrome oxidase subunit 3, partial (mitochondrion) [Sitodiplosis mosellana]	7e-77
CN224	1	dihydropteridine reductase dhpr/qdpr [Anopheles aquasalis]	1e-57
CN239	1	succinate dehydrogenase B, partial [Mayetiola destructor]	3e-63
1901229M13F21	1	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 9 [Musca domestica]	2e25
CN170	1	ATP synthase subunit e, mitochondrial [Stomoxys calcitrans]	6e-17
MIDGESAL_001_B13.F1	1	NADH dehydrogenase subunit 3 (mitochondrion) [Heleodromia immaculata]	9e-5
MIDGESAL_001_H05.F1	1	ATPase inhibitor A, mitochondrial [Drosophila biarmipes]	2e-21
MIDGESAL_001_H14.F1	1	cytochrome c oxidase subunit 6A, mitochondrial-like [Wasmannia auropunctata]	8e-30
MIDGESAL_001_J06.F1	1	ATP synthase subunit g, mitochondrial [Aethina tumida]	2e-39

MIDGESAL_007_P19.F1	1	mitochondrial cytochrome c oxidase subunit 6b [Portunus trituberculatus]	6e-34
MIDGESAL_002_C10.F1	1	ubiquinone oxidoreductase [Tabanus bromius]	1e-33
CN192	1	NADH:ubiquinone oxidoreductase ndufb5/sgdh subunit [Corethrella appendiculata]	2e-52
MIDGESAL_002_J12.F1	1	ATP synthase, partial [Tabanus bromius]	2e-55
MIDGESAL_002_J15.F1	1	flavin reductase (NADPH) [Drosophila takahashii]	1e-75
MIDGESAL_002_O04.F1	1	vacuolar atpase m9, partial [Anopheles aquasalis]	5e-25
MIDGESAL_003_A15.F1	1	NADH dehydrogenase subunit 5, partial (mitochondrion) [Sitodiplosis mosellana]	3e-51
MIDGESAL_003_I05.F1	1	cytochrome b, partial (mitochondrion) [Sitodiplosis mosellana]	5e-8
MIDGESAL_003_K12.F1	1	cation-independent mannose-6-phosphate receptor [Culex tarsalis]	6e-29
MIDGESAL_006_F04.F1	1	NADH dehydrogenase [ubiquinone] iron-sulfur protein 4, mitochondrial [Aedes albopictus]	6e-69
MIDGESAL_006_F09.F1	1	ATP synthase subunit f, mitochondrial [Plutella xylostella]	4e-56
MIDGESAL_006_J21.F1	1	mitochondrial atp synthase coupling factor 6 [Tabanus bromius]	8e-42
MIDGESAL_007_H21.F1	1	ubiquinone biosynthesis protein COQ9, mitochondrial [Anopheles darlingi]	7e-14
MIDGESAL_008_N04.F1	1	f0f1-type atp synthase beta subunit, partial [Tabanus bromius]	5e-19
MIDGESAL_004_A08.F1	1	saccharopine dehydrogenase domain-containing protein [Aedes aegypti]	3e-80
MIDGESAL_003_P06.F1	1	peroxiredoxin-2 [Musca domestica]	5e-36
CN251	1	dolichyl-diphosphooligosaccharide--protein glycosyltransferase subunit 4 [Drosophila kikkawai]	7e-5
MIDGESAL_001_B24.F1	1	CDP-diacylglycerol--inositol 3-phosphatidyltransferase-like [Aedes albopictus]	4e-117
CN136	1	ectonucleotide pyrophosphatase/phosphodiesterase family member 5-like [Aplysia californica]	6e-10
MIDGESAL_004_G18.F1	1	lactosylceramide 4-alpha-galactosyltransferase [Drosophila biarmipes]	3e-18
MIDGESAL_004_J21.F1	1	glyceraldehyde 3-phosphate dehydrogenase, partial [Mayetiola destructor]	3e-36
MIDGESAL_002_M06.F	1	13-hydroxy-3-methylglutaryl-coenzyme A reductase [Aedes albopictus]	2e-64
MIDGESAL_003_M20.F1	1	glutaredoxin [Nyssomyia neivai]	1e-30
MIDGESAL_006_I19.F1	1	aldehyde dehydrogenase, isoform A [Drosophila melanogaster]	1e-138
2. Structural proteins			
38	6	histone 1, partial [Anopheles aquasalis]	4e-51
60	6	histone H4 [Aptenodytes forsteri]	3e-50
MIDGESAL_001_F20.F1	1	histone H1.2-like [Dinoponera quadriceps]	2e-28
MIDGESAL_004_M10.F1	1	histone H1.3 [Tribolium castaneum]	2e-28
MIDGESAL_002_J13.F1	1	glycine-rich cell wall structural protein 1.0-like [Stomoxys calcitrans]	8e-10
CN86	1	Myosin heavy chain, muscle [Bactrocera cucurbitae	0.0
CN139	1	myosin regulatory light chain ef-hand protein [Nyssomyia neivai]	7e-74
CN153	1	alpha tubulin [Nyssomyia neivai]	0.0
CN258	1	myosin heavy chain, muscle, partial [Drosophila navojoa]	4e-34
125	2	tropomyosin [Nyssomyia neivai]	1e-102
112	2	actin, partial [Laufeia concava]	1e-145
CN144	1	actin-depolymerizing factor 1 [Bombyx mori]	1e-98
CN225	1	myosin light chain alkali isoform X2 [Plutella xylostella]	2e-60
CN227	1	nuclear protein 1-like [Xenopus laevis]	3e-6
CN167	1	alpha-helical protein, partial [Tabanus bromius]	1e-25

CN152	1	Muscle-specific protein 300 kDa, isoform D [Drosophila melanogaster]	2e-26
MIDGESAL_007_I18.F1	1	muscle LIM protein at 60A, isoform E [Drosophila melanogaster]	2e-37
3. Protein processing and degradation			
CN149	1	ubiquitin-related modifier protein [Anopheles aquasalis]	7e-36
MIDGESAL_003_F16.F1	1	proteasome subunit beta type-4 [Trichogramma pretiosum]	4e-30
CN259	1	cathepsin l [Corethrella appendiculata] no SP	3e-44
MIDGESAL_001_I10.F1	1	proteasome subunit alpha type-7-1 [Musca domestica]	2e-64
MIDGESAL_002_A20.F1	1	ubiquitin-like protein 7 [Corethrella appendiculata]	2e-16
CN91	1	thrombin inhibitor infestin precursor, partial [Triatoma infestans]	3e-4
MIDGESAL_003_H01.F1	1	m13 family peptidase, partial [Tabanus bromius]	1e-53
MIDGESAL_005_I04.F1	1	e3 ubiquitin ligase cullin 1 component [Tabanus bromius]	2e-167
MIDGESAL_006_J17.F1	1	e3 ubiquitin-protein ligase rbbp6, partial [Culex tarsalis]	0.025
MIDGESAL_006_J15.F1	1	e3 ubiquitin-protein ligase lin [Anopheles aquasalis]	7e-38
MIDGESAL_007_C10.F1	1	Serine protease easter [Bactrocera latifrons] no SP	1e-7
MIDGESAL_008_J11.F1	1	trypsin inhibitor-like [Atta colombica] no SP	4e-11
4. Transportation			
MIDGESAL_006_D16.F1	1	translocon-associated complex trap gamma subunit [Nyssomyia neivai]	5e-93
CN172	1	monocarboxylate transporter [Nyssomyia neivai]	7e-7
MIDGESAL_002_H15.F1	1	transmembrane protein 85 [Anopheles darlingi]	2e-52
11712M13F21	1	ER retained protein [Culex quinquefasciatus]	1e-59
CN216	1	plasma membrane protein [Tabanus bromius]	7e-7
MIDGESAL_002_D18.F1	1	titin, partial [Aedes albopictus]	5e-56
MIDGESAL_002_H15.F1	1	transmembrane protein 85 [Anopheles darlingi]	2e-52
18	2	transmembrane protein [Anopheles aquasalis]	2e-12
MIDGESAL_002_H16.F1	1	snapin [Nyssomyia neivai]	2e-53
MIDGESAL_002_M03.F	1	rab subfamily protein of small gtpase [Nyssomyia neivai]	6e-127
75	2	Ras superfamily small GTPase Rab11 [Nilaparvata lugens]	3e-24
MIDGESAL_001_E01.F1	1	rab GTPase [Haematobia irritans]	1e-100
MIDGESAL_002_O01.F1	1	trafficking protein particle complex subunit 1 isoform X1 [Drosophila takahashii]	1e-27
MIDGESAL_003_B19.F1	1	preprotein translocase gamma subunit [Nyssomyia neivai]	2e-26
MIDGESAL_003_F09.F1	1	transport and golgi family organization [Nyssomyia neivai]	9e-41
MIDGESAL_003_I18.F1	1	acetyltransferase gnat family [Tabanus bromius]	3e-99
MIDGESAL_003_J24.F1	1	protein transporter [Culex tarsalis]	1e-27
MIDGESAL_004_F05.F1	1	copii vesicle protein [Nyssomyia neivai]	1e-113
MIDGESAL_006_N15.F1	1	transmembrane transport [Corethrella appendiculata]	1e-58
MIDGESAL_007_C02.F1	1	Golgi apparatus membrane protein-like protein CG5021 isoform X1 [Ceratitis capitata]	2e-27
MIDGESAL_008_B10.F1	1	protein transporter of the tram translocating chain-associating membrane [Corethrella appendiculata]	3e-20
MIDGESAL_008_E10.F1	1	vacuolar protein sorting 13D [Anopheles darlingi]	1e-76
CN134	1	translocon-associated protein subunit alpha [Drosophila arizonae]	1e-19
30	12	translocase of inner membrane 9b, isoform C [Drosophila melanogaster]	1e-19
Other proteins			
1. **Detoxification**			

21 confirm	2	glutathione s-transferase [Nyssomyia neivai] no signal peptide	1e-86
MIDGESAL_002_D04.F1	1	sigma GST [Mayetiola destructor] no signal peptide	8e-23
MIDGESAL_006_H01.F1	1	microsomal glutathione s-transferase 1-like protein [Haematobia irritans]	7e-38
MIDGESAL_001_C24.F1	1	cytochrome P450 6d5 [Aedes albopictus]	1e-85
2. Regulation			
CN73	1	creatine kinase, partial [Corethrella appendiculata]	6e-137
CN154	1	nucleoside diphosphate kinase [Orseolia oryzae]	1e-105
MIDGESAL_003_O16.F1	1	s-phase kinase-associated protein 1a [Tabanus bromius]	5e-80
MIDGESAL_004_F10.F1	1	cyclin-dependent kinase 9 [Galendromus occidentalis]	2e-12
3. Fatty acid proteins			
MIDGESAL_002_C15.F1	1	sn1-specific diacylglycerol lipase alpha [Aedes albopictus]	1e-14
Smosell_SSGP_105118	1	lipoprotein [Escherichia coli]	1e-72
MIDGESAL_006_C16.F1	1	acyl-coa reductase [Nyssomyia neivai]	5e-30
MIDGESAL_008_M07.F1	1	acyl-coa dehydrogenase [Culex quinquefasciatus]	2e-54
MIDGESAL_008_B23.F1	1	Peptidyl-prolyl cis-trans isomerase F, mitochondrial [Daphnia magna]	6e-85
MIDGESAL_003_N01.F1	1	phospholipase/carboxylesterase [Culex tarsalis]	3e-110
MIDGESAL_003_N01.F1	1	phospholipase/carboxylesterase [Culex tarsalis]	3e-110
MIDGESAL_003_D11.F1	1	malonyl-CoA-acyl carrier protein transacylase, mitochondrial [Agrilus planipennis]	3e-53
4. Amino acids synthesis & cycling proteins			
MIDGESAL_005_F08.F1	1	glycine hydroxymethyl transferase [Mayetiola destructor]	4e-38
MIDGESAL_007_H17.F1	1	arginase [Anopheles darlingi] no signal peptide	3e-07
MIDGESAL_007_L14.F1	1	bifunctional purine biosynthesis protein PURH [Stomoxys calcitrans]	3e-172
MIDGESAL_007_C14.F1	1	alpha-14-n-acetylglucosaminyltransferase [Nyssomyia neivai]	4e-28
MIDGESAL_006_P18.F1	1	ceramide glucosyltransferase [Corethrella appendiculata]	2e-43
MIDGESAL_006_M10.F1	1	thioredoxin-like protein [Tabanus bromius]	2e-69
MIDGESAL_004_I16.F1	1	N-acetyltransferase san [Stomoxys calcitrans]	6e-75
MIDGESAL_004_G24.F1	1	Protein catecholamines up, partial [Fopius arisanus]	8e-20
MIDGESAL_003_M21.F1	1	aspartate aminotransferase/glutamic oxaloacetic transaminase aat2/got1 [Psorophora albipes]	5e-51
MIDGESAL_002_I15.F1	1	aminoacylase acy1 [Nyssomyia neivai]	3e-99
5. Nero proteins & neurotransmission			
CN243	1	n-acetyl neuraminate lyase, partial [Rhodnius neglectus]	6e-50
MIDGESAL_003_F15.F1	1	potassium channel AKT1-like [Setaria italica]	2e-5
MIDGESAL_008_M12.F1	1	sn-12-diacylglycerol ethanolamine- and cholinephosphotransferase [Corethrella appendiculata]	1e-89
MIDGESAL_007_J20.F1	1	mesencephalic astrocyte-derived neurotrophic factor-like protein [Lasius niger]	5e-146
MIDGESAL_003_O08.F1	1	N-acetylneuraminate lyase B-like [Aedes albopictus]	1e-43
2071229M13F21	1	neural/ectodermal development factor IMP-L2 [Anopheles darlingi] no signal peptide	3e-18
6. Others			
78	2	Zgc:165536 protein [Danio rerio]	2e-15
124	2	single-pass membrane and coiled-coil domain-containing protein [Anoplophora glabripennis]	5e-13
MIDGESAL_001_I16.F1	1	juvenile hormone epoxide hydrolase 1 [Drosophila miranda]	8e-45
MIDGESAL_002_N12.F1	1	larval cuticle protein 8-like [Aethina tumida]	4e-16

MIDGESAL_003_B13.F1	1	chromatin assembly factor-i [Aedes albopictus]	2e-26
MIDGESAL_003_C15.F1	1	myofilin [Nyssomyia neivai]	6e-15
MIDGESAL_005_P06.F1	1	fucosyltransferase [Nyssomyia neivai]	6e-67
MIDGESAL_006_C05.F1	1	death-associated inhibitor of apoptosis 1 isoform X2 [Drosophila miranda]	7e-73
MIDGESAL_005_P18.F1	1	mini spindles, isoform B [Drosophila melanogaster]	8e-69
MIDGESAL_007_I23.F1	1	Short spindle protein 4 [Bactrocera dorsalis]	6e-129
MIDGESAL_006_J09.F1	1	adipokinetic hormone-like protein [Corethrella appendiculata]	7e-11
MIDGESAL_007_B18.F1	1	mucin-1-like [Ceratina calcarata] no signal peptide	3e-46
MIDGESAL_007_B10.F1	1	myophilin [Drosophila takahashii]	2e-109
MIDGESAL_007_F20.F1	1	gametocyte-specific factor 1 homolog [Aedes albopictus]	1e-16
MIDGESAL_007_O11.F1	1	protein targeting to golgi, partial [Corethrella appendiculata]	6e-25
MIDGESAL_008_H18.F1	1	troponin I isoform X11 [Musca domestica]	3e-60
MIDGESAL_008_J24.F1	1	pupal cuticle protein Edg-78E [Ceratitis capitata]	9e-24
MIDGESAL_008_L24.F1	1	Nucleoplasmin-like protein [Lucilia cuprina]	2e-29
MIDGESAL_007_P15.F1	1	23 kDa integral membrane protein [Drosophila biarmipes]	6e-08
MIDGESAL_007_D18.F1	1	endoplasmic reticulum resident protein 44 isoform X3 [Ceratitis capitata]	7e-85
MIDGESAL_007_B07.F1	1	ER lumen protein-retaining receptor [Nasonia vitripennis]	1e-95
MIDGESAL_006_E11.F1	1	superoxide dismutase [Bombus impatiens]	3e-56
MIDGESAL_005_N18.F1	1	Twinfilin [Bactrocera dorsalis]	1e-38
MIDGESAL_005_N08.F1	1	ring finger protein [Culex tarsalis]	3e-58
MIDGESAL_005_H16.F1	1	nucleoplasmin [Corethrella appendiculata]	1e-31
MIDGESAL_005_G16.F1	1	keratin, type II cytoskeletal 5 [Musca domestica]	1e-17
MIDGESAL_004_K12.F1	1	tetratricopeptide repeat protein 14 homolog [Cimex lectularius]	5e-9
MIDGESAL_004_L19.F1	1	caspase-1 [Culex quinquefasciatus]	4e-11
MIDGESAL_004_J11.F1	1	glycogenin [Anopheles aquasalis]	2e-22
MIDGESAL_004_F13.F1	1	dnaJ homolog subfamily C member 7 isoform X1 [Ceratitis capitata]	2e-11
MIDGESAL_003_L04.F1	1	armadillo/beta-catenin/plakoglobin, partial [Tabanus bromius]	1e-64
MIDGESAL_002_I11.F1	1	u6 snrna-associated sm-like protein [Nyssomyia neivai]	4e-54
MIDGESAL_002_C08.F1	1	cyanate hydratase [Spongiibacter tropicus]	6e-52
MIDGESAL_001_L19.F1	1	CDGSH iron-sulfur domain-containing protein 3, mitochondrial isoform X2 [Drosophila kikkawai]	1e-46
MIDGESAL_001_G01.F1	1	cubilin homolog [Aedes albopictus]	3e-45
CN238	1	proteinral odorant-binding protein 99a [Nyssomyia neivai]	2e-39
CN206	1	high mobility group protein D [Drosophila kikkawai]	3e-32
CN210	1	Rieske iron-sulfur protein 1 [Graphocephala atropunctata]	9e-68
CN206	1	Calponin [Nyssomyia neivai]	5e-60
CN46	1	Regucalcin [Bactrocera latifrons]	4e-17
CN80	1	PDZ and LIM domain protein 3 [Bactrocera cucurbitae]	9e-55
1291230M13F21	1	Ribulose bisphosphate carboxylase large chain [Morus notabilis]	4e-31
5413M13F21	1	phagocytosis engulfment [Psorophora albipes]	2e-94
2311115	1	crustapain-like [Dendroctonus ponderosae]	3e-16
1591115	1	cell death protein/oligosaccharyltransferase epsilon subunit [Corethrella appendiculata]	6e-56

Table S3. Single nucleotide and other small mutations among members in SSGP gene families.
Table S3-1: Sequence variations among group 1 members in family 1.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variations in the mature protein coding region			
22	**C**TC-**T**TC	L-F	1401111
29	A**G**A-A**A**A	R-K	211115, 1401111
38	T**G**C-T**T**C	C-F	1421110, 1601115
43,45	**C**T**G**-**A**T**C**	L-I	1401111
73,74	**GA**C-**CT**C	D-L	1421110
77,78	A**AG**-A**TA** A**AG**-A**CA**	K-I K-T	1421110, 1601115 171111
79	**A**AA- **G**AA	K-E	1401111,21115
94	**A**AC- **C**AC	N-H	1401111,21115
97	**T**CA-**G**CA	S-A	1601115, 1421110
111	AA**G**-AA**A**	_	1601115, 1421110
113	G**C**T-G**T**T	A-V	1401111,21115
115	**G**TT-**A**TT	V-I	171111
122	A**A**T-A**G**T	N-S	1601115
130	**G**GC-**A**GC	G-S	1601115
141	AA**T**-AA**A**	N-K	1401111, 211115
148	**G**TT-**A**TT	V-I	1601115, 1421110
152	A**C**T-A**A**T	T-N	1601115, 1421110
154	**A**AT-**G**AT	N-D	171111
163	**G**CC-**A**CC	A-T	171111
169,170	**GC**T-**TA**T	A-Y	1601115, 1421110
177	CA**G**-CA**T**	Q-H	211115, 1401111
183	TT**A**-TT**G**	_	1401111
188	A**A**A-A**G**A	K-R	211115, 171111
191	A**T**A-A**C**A	I-T	211115, 171111
201	AA**G**-AA**A**	_	1601115, 1421110
202	**A**TA-**T**TA	I-L	1601115, 1421110
222	AG**A**-AG**G**	_	1601115, 1421110
223,224,225	**AAG**-**GCT**	K-A	1601115, 1421110
231	GA**A**-GA**T**	E-D	211115, 1401111
235,237	**GTA** -**A**T**T**	V-I	211115, 1401111
241	**A**CT-**C**CT	T-P	1401111
248	T**A**T-T**T**T	Y-F	1601115, 1421110
250	**A**AA-**C**AA	K-Q	211115, 1401111
254	A**G**T-A**A**T	S-N	171111
257,258	A**GG**-A**AA** A**GG**-A**TA**	R-K R-I	171111 1421110, 1601115
269	G**T**A-G**C**A	V-A	1421110, 1601115
274	**T**TG-**A**TG	L-M	1421110, 1601115
279	AA**G**-AA**A**	_	1421110, 1601115
280,281	**TT**T-**GC**T	F-A	171111, 211115
290	A**A**A-A**G**A	K-R	171111

292	**G**TA-**C**TA	V-L	1421110, 1601115
298,299,300	**GCA-AAC** **GCA-AGC**	A-N A-S	211115, 1401111 171111
308	A**A**C-A**C**C	N-T	1401111, 171111
311	A**A**G-A**G**G	K-R	1421110, 1601115
314	A**A**A-A**G**A	K-R	1421110, 1601115
316,318	**A**A**G**-**G**A**A**	K-E	171111
320	A**G**T-A**A**T	S-N	171111
322	**C**TG-**T**TG	_	1421110, 171111
325,326	**TC**A-**GG**A	S-G	171111
328	**A**GA-**G**GA	R-G	1421110, 1601115
331,333	**A**A**A**-**G**A**G**	K-E	1601115
Sequence variations in the signal peptide coding region			
59	G**G**T-G**C**T	G-A	1401111,1601115
Sequence variations in the non-coding regions			
3	C-T		171111, 1421110
14	T-A		1421110, 1601115
22	Indel		1421110
23	A-G		1421110
24	Indel		1601115
25	T-G		1421110, 1601115
27	G-T		1421110
34	C-A		1421110
35	A-C		1421110
37	G-A		1421110, 1401111
38	C-G		1421110, 1401111
39	A-C		1421110, 1401111
40	T-A		1421110, 1401111
43	C-T C-A		1421110, 1401111,1601115
45	Indel		211115, 1601115
455	C-T		1421110
466	G-A		1421110, 1601115
479	Indel		171111, 1421110
482	G-A		1421110, 1601115
483	A-G A-T		1421110 1401111
489	T-C		1401111
492	G-C		1401111
496	T-G		1401111
507	T-C C-T		171111 1421110
512	C-T T-C		171111 1421110, 1401111
513	C-T T-C		171111 1421110, 1401111
534	A-C		1401111
539	Indel		1421110
548	T-G		1401111
551	G-A A-G		171111 1421110

585	C-A	1421110
586	A-T	1421110
587	A-T	171111
590	G-C	1401111
592	C-G	1401111
599	T-A	1421110
600	T-A	171111
603	A-T T-A	171111 1421110
604	T-A	171111
606	G-T	1401111
607	G-T C-G	1421110 1401111
611	T-C	171111
613	A-T T-A	171111 1421110
614	T-A C-T	171111 1421110
615	C-T A-C	171111 1421110
616	A-C T-A	171111 1421110, 1401111
617	T-A G-A	171111 1421110
618	G-A A-T A-G	171111 1421110 1401111
619	A-T T-G	171111, 1421110, 1401111
620	T-G	171111
623	A-T	171111
625	C-A	171111
628	T-A G-C G-A	171111 1421110 1401111
629	G-C	171111
631	G-A	1421110, 1401111

Table S3-2: Sequence variations among group 2 members in family 1.

Base position	Codon change	AA change	Clone name
Sequence variations in the mature protein coding region			
4	**C**CA-**G**CA	P-A	191118
13,14	**CA**G-**AC**G	Q-T	191118
16	**T**GG-**G**GG	W-G	10512M13F21,21115
19	**A**GT-**G**GT	S-G	21115
20	A**G**T-A**A**T	S-N	191118
29	C**G**A-C**A**A	R-Q	191118
33	AT**A**-AT**G**	I-M	191118
44	G**T**G-G**C**G	V-A	191118
54	GA**T**-GA**A**	D-E	191118
60	**G**AA-CA**A**	E-Q	191118
66	GC**G**-GC**A**	-	191118
76	**A**AG-**G**AG	K-E	561117
95	C**GC**-C**AG**	R-Q	191118
96	**A**CA-**G**CA	T-A	191118
110	A**A**G-A**G**G	K-R	21115
113	G**T**T-G**G**T	V-G	21115
117	GT**T**-GT**G**	-	21115
123	AA**T**-AA**G**	N-K	191118
126	AT**A**-AT**G**	I-M	21115
141	AA**A**-AA**T**	K-N	191118
144	AC**C**-AC**T**	-	191118
151	**G**CT-**C**CT	A-P	21115
155	A**A**T-A**C**T	N-T	191118
162	AA**T**-AA**G**	N-K	191118
165	GC**C**-GC**A**	-	191118

167	A**T**A-A**A**A	I-K	191118
168	AT**A**-AT**T**	-	21115
169,171	**T**A**T**-**A**A**A**	Y-K	191118
172	**G**AA-AA**A**	E-K	191118
177	AA**G**-AA**A**	-	191118
181,182	**TT**A-**AA**A	L-K	191118
184	**G**AA-**A**AA	E-K	191118
188	A**G**A-A**A**A	R-K	191118
190	**A**CA-**C**CA	T-P	21115
191	A**C**A-A**A**A	T-K	191118
197	A**G**C-A**C**C	S-T	21115
197	AG**C**-AG**T**	-	10512M13F21
198	AG**C**-AG**T**	-	10512M13F21
Sequence variation in the signal peptide coding region			
13	**T**TC-**A**TC	F-I	10512M13F21,21115,191118
52,53,54	**ATC**-**TCA**	I-S	191118
56,57	A**AT**-A**TA**	N-I	191118
Sequence variation in the non-coding regions			
17	C-T	10512M13F21,21115	
20	C-T	561117,10512M13F21	
25	T-G	21115	
28	T-G	21115	
31	A-T	561117,10512M13F21, 21115	
53	A-C	10512M13F21,21115	

Table S3-3: Sequence variations among family 24 members.

Base position	Codon change	AA change	Clone name
Sequence variation in the mature protein coding region			
128	A**G**C-A**C**C	S-T	1781229M13F21
132	TT**T**-TT**C**	-	1781229M13F21
133	**G**AT-**C**AT	D-H	1781229M13F21
170	C**G**C-C**C**C	R-P	1781229M13F21
40	**T**CT-**C**CT	S-P	1241230M13F21, CN187
131	T**T**T-T**C**T	F-S	1241230M13F21, CN187
234,235	**GA**A-**AC**A	E-T	1241230M13F21
236	G**A**A-G**C**A	E-A	CN187
244,246	**A**A**A**-**G**A**G**	K-E	CN187
253,254	**AA**A-**GC**A	K-A	CN187
256,257	**AA**A-**TC**A	K-S	CN187
259,260	**AA**A-**TC**A	K-S	CN187
131,132	T**TC**-T**CT**	S-F	1241230M13F21, CN187
235	**A**CA-**G**CA	T-A	CN187
240	AA**A**-AA**G**	-	CN187
252	AA**A**-AA**G**	-	CN187
Sequence variation in the signal peptide coding region			
43	**T**TG-**C**TG	-	1781229M13F21,1241230M13F21
54	TT**G**-TT**C**	L-F	1781229M13F21
Sequence variation in the non-coding regions			
None			

Table S3-4: Sequence variations among family 29 members.

Base position	Codon change	AA change	Clone name
Sequence variation in the mature protein coding region			
5	T**T**A-T**C**A	L-S	1081114
7,8,9	**AAA**-**TTT**	K-F	CN8
10,11	**CA**T-**TT**T	H-F	1081114,CN8
13,15	**G**A**C**-**T**A**T**	D-Y	CN8
16,18	**T**C**C**-**G**C**A**	S-A	CN8
27	AA**A**-AA**C**	K-N	CN8
37	**A**GA-**G**GA	R-G	CN8
44	**CC**A-**CA**A	P-Q	2151229M13F21,1081114,341115, CN8
49,50	**AC**T-**TT**T	T-F	CN8
58,59	**AA**A-**CC**A	K-P	CN8
86	C**C**A-C**T**A	P-L	CN8
95	G**C**C-G**T**C	A-V	CN8
112	**T**AC-**A**AC	Y-N	1081114
119	T**T**A-T**A**A	L- Stop codon	1801117
Sequence variation in the signal peptide coding region			
11	A**A**T-A**C**T	N-T	1081114
22	**C**TC-**T**TC	L-F	2151229M13F21
34	**A**TT-**G**TT	I-V	1081114
43	**A**TA-**T**TA	I-L	CN8
68	G**T**T-G**C**T	V-A	CN8
71	G**A**A-G**G**A	E-G	CN8
73	**G**CT-**A**CT	A-T	CN8
Sequence variation in the non-coding regions			
1,2,3	Indel	1801117,341115	
2	G-T	1081114	
4	C-G	2151229M13F21	

46	C-T	CN8	
47	C-A C-G	1081114 CN8	
48	C-G	1081114	
69	T-C	CN8	
86	T-C	1081114,341115,CN8	
107	T-A	1081114,CN8	
129	T-C	1081114,341115,CN8	
352	T-G	CN8	
430	T-C	341115	
431	T-C	341115	
432	G-A T-C	341115 1081114	

Table S3-5: Sequence variations among family 40 members.

Base position	Codon change	AA change	Clone name
Sequence variation in the mature protein coding region			
77	A**G**G-A**A**G	R-K	2701230M13F21
135	AG**C**-AG**G**	S-R	2701230M13F21
148	**G**AT-**T**AT	D-Y	2551115
173	T**C**T-T**T**T	S-F	2701230M13F21
215	Indel		171114,2551115
221,222,223	**TCG**-**GTC**	S-V	2701230M13F21
224,225,226	**GCC**-**GGC**	A-G	2701230M13F21
227,228,229	**TCT**-**CTC**	S-L	2701230M13F21
232	TT**C**-TT**T**	-	2701230M13F21
233,234,235	**GCA**-**CGC**	A-R	2701230M13F21
236,237,238	**TCA-ATC**	S-I	2701230M13F21
240	AA**A**-AA**G**	-	2551115,171114
243	AA**T**-AA**A**	N-K	2701230M13F21
244,245,246	**GCA**-**TGC**	A-C	2701230M13F21
247,249	**C**C**A**-**A**C**C**	P-T	2701230M13F21
250,251,252	**TCA**-**AGC**	-	2551115,2701230M13F21
249,250,251	**AGC**-**GCA**	S-A	171114
254,255	**CA**A-**AC**A	Q-T	2701230M13F21
257,259	**G**G**A**-**A**G**G**	G-R	2701230M13F21
262	AA**T**-AA**A**	N-K	2701230M13F21
263,265	**G**G**A**-**T**G**G**	G-W	2701230M13F21
266,267	**CA**A-**AC**A	Q-T	2701230M13F21
269	CG**G**-CG**A**	-	2551115
269,270,271	**CGA**-**ACG**	R-T	2701230M13F21

272,274	T**TG**-A**TT**	L-I	2701230M13F21
277	GG**T**-GG**G**	-	2701230M13F21
278,279	**GC**C-**TG**C	A-C	2701230M13F21
281,282,283	**TGT**-**CTG**	C-L	2701230M13F21
284,285	**AC**C-**TA**C	T-Y	2701230M13F21
287,288	**TC**C-**CT**C	S-L	2701230M13F21
290,292	**A**A**G**-**C**A**A**	K-Q	2701230M13F21
293,294	**CA**A-**GC**A	Q-A	2701230M13F21
297,298	A**CT**-A**AC**	T-N	2701230M13F21
299	**A**AA-**T**AA	K- Stop codon	2701230M13F21
Sequence variation in the signal peptide coding region			
58	**A**GC-**T**GC	S-C	2551115
Sequence variation in the non-coding regions			
1,2	Indel		2701230M13F21,171114
3	G-C		2551115
16,22	A-G		2551115
476	A-C		171114
480	A-T		171114
481	A-C		171114
482	A-T		171114
483	A-C		171114
486	A-G		171114
488	A-T		171114
489	A-T		171114
490	A-G		171114
491	A-C		171114
496	A-T		2551115

503	A-T	2551115
505	A-G	2551115

Appendix B - Supplementary data, Chapter 3

Table S1. Distribution of SSGP transcripts among different groups of barley midge (transcripts in each group and their related proteins are listed in order in Figure S1)

Gene	Number of unique sequences	Number of total sequences	First BLAST hit if any	E-value
Group 1 Subgroup E Subgroup F	78	260	SSSGP-1C2 [Mayetiola destructor] SSSGP-1D1 [Mayetiola destructor] SSSGP-1C1 [Mayetiola destructor]	6e-10 6e-26 3e-36
Group 2	1	1		
Group 3	1	1		
Group 4	2	3		
Group 5	4	5		
Group 6	5	6		
Group7	1	1		
Group 8	1	1		
Group 9	1	2		
Group 10	1	1		
Group 11	3	4	SSGP-71 [Mayetiola destructor]	2e-15
Group 12	1	2		
Group 13	2	18		
Group 14	1	1		
Group 15	4	4		
Group 16	1	2		
Group 17	1	1		
Group 18	3	8		
Group 19	3	14		
Group 20	5	5		
Group 21	1	1		
Group 22	1	1		
Group 23	1	1		
Group 24	3	5	larval cuticle protein 12-like [Anoplophora glabripennis]	6e-26
Group 25	1	1		
Group 26	1	1		
Group27	1	1		
Group28	8	10	secreted salivary gland protein SSGP-4A [Mayetiola destructor]	2e-04
Group29	1	2		
Group30	1	1		

Group31	11	24	SSGP-11B family protein [Mayetiola destructor]	7e-33
Group32	4	22		
Group33	3	5		
Group34	4	6		
Group35	1	2	kinase domain protein [Ichthyophthirius multifiliis]	6e-17
Group36	1	3		
Group37	1	2		
Group38	1	2		
Group39	2	3	secreted protein F [Mayetiola destructor]	4e-124
Group40	1	11		
Group41	1	1		
Group42	1	2		
Group43	1	2		
Group44	1	1		
Group45	1	1		
Group46	1	1		
Group47	1	1		
Group48	1	1		
Group49	1	1		
Group50	1	1		
Group 51	1	2		
Total	178	458		

Table S2. Classification of proteins encoded by non-SSGP transcripts belong to barley midge.

#	No.	First hit	E-value
C. Protein synthesis and folding			
5. Ribosomal protein			
20	4	40S ribosomal protein S23-like [Drosophila obscura]	2e-93
33	2	60S ribosomal protein L37-A [Zeugodacus cucurbitae]	5e-38
38	4	40S ribosomal protein S26 [Athalia rosae]	1e-66
47	5	ribosomal protein L38 [Psathyromyia shannoni]	5e-37
Mh_SSGP_1281012	1	60S ribosomal protein L30 [Aedes aegypti]	1e-69
60	2	40S ribosomal protein S21, putative [Pediculus humanus corporis]	4e-44
67	3	ribosomal protein S25 [Chrysomela tremula]	3e-45
Mh_SSGP_911014	1	ribosomal protein L37A, isoform A [Drosophila melanogaster]	4e-45
Mh_SSGP_371017	1	ribosomal protein L10Ab, isoform A [Drosophila melanogaster]	2e-65
74	2	ribosomal protein S19a, isoform A [Drosophila melanogaster]	8e-66
75	2	60S ribosomal protein L39-like [Bubalus bubalis]	2e-18
81	2	60S ribosomal protein L11 [Lucilia cuprina]	1e-121
88	3	60S ribosomal protein L35a [Drosophila takahashii]	7e-58
99	3	60S ribosomal protein L27a [Zeugodacus cucurbitae]	9e-72
108	4	ribosomal protein L3, isoform G [Drosophila melanogaster]	2e-38
Mh_SSGP_27118M13F21	1	40S ribosomal protein S8 [Athalia rosae]	9e-98
Mh_SSGP_1061012	1	60S ribosomal protein L8 [Leptinotarsa decemlineata]	5e-41
Mh_SSGP_1351013	1	40S ribosomal protein S20 [Culex quinquefasciatus]	1e-67
Mh_SSGP_1211017	1	60S ribosomal protein L14 [Drosophila suzukii]	4e-77
Mh_SSGP_1071014	1	60S ribosomal protein L27 [Bactrocera latifrons]	2e-57
Mh_SSGP_1151018	1	ribosomal protein S27, isoform A [Drosophila melanogaster]	1e-51
2. Transcription & translation factors			45
56	2	eukaryotic translation initiation factor 2 alpha subunit [Culex quinquefasciatus]	2e-24
Mh_SSGP_12718M13F21	1	eukaryotic translation initiation factor 3 [Culex quinquefasciatus]	1e-64
Mh_SSGP_161014	1	Nucleotidyl transferase [Sphingomonas sp. UNC305MFCol5.2]	1.0
3. Chaperones, protein folding and secretion			
Mh_SSGP_1391011	1	heat shock cognate 70 isoform C, partial [Aedes albopictus]	2e-83
105	2	peptidyl-prolyl cis-trans isomerase FKBP2 precursor [Mus musculus]	2e-48
90	2	peptidyl-prolyl cis-trans isomerase f, ppif [Riptortus pedestris]	5e-06
Mh_SSGP_918M13F21	1	heat shock cognate 70 protein, partial [Antheraea yamamai]	1e-17
Mh_SSGP_1191013	1	heat shock protein 83, partial [Stratiomys japonica]	2e-54
Mh_SSGP_21517M13F21	1	stress-associated endoplasmic reticulum protein 2 [Musca domestica]	2e-34

4. RNA synthesis and processing			
Mh_SSGP_21717M13F21	1	RNA polymerase subunit sigma-70 [Nocardiopsis sp. JB363]	9.0
D. House keeping			13
1. Energy metabolism			
46	5	NADH dehydrogenase [ubiquinone] 1 beta subcomplex subunit 5, mitochondrial isoform X1 [Dinoponera quadriceps]	9e-37
Mh_SSGP_51014	1	NADH -cytochrome b5 reductase-like [Aedes aegypti]	8e-32
58	2	AGAP013189-PA [Anopheles gambiae str. PEST]	6e-50
45	2	V-type proton ATPase subunit E [Bactrocera dorsalis]	5e-95
Mh_SSGP_1041012	1	cytochrome c oxidase subunit NDUFA4 [Lucilia cuprina]	1e-27
89	2	cytochrome c oxidase subunit 8B, mitochondrial-like [Ochotona princeps]	0.027
98	3	cytochrome oxidase subunit I, partial (mitochondrion) [Asteromyia laeviana]	9e-70
100	5	cytochrome c oxidase subunit I (mitochondrion) [Mayetiola destructor]	3e-99
102	2	NADH-ubiquinone oxidoreductase ashi subunit [Anopheles darlingi]	3e-36
109	2	mitochondrial NADH:ubiquinone oxidoreductase ESSS subunit [Anopheles darlingi]	2e-26
Mh_SSGP_16518M13F21	1	putative ATP synthase subunit f, mitochondrial [Plutella xylostella]	9e-65
Mh_SSGP_26317M13F21	1	GTP-binding nuclear protein Ran [Rhagoletis zephyria]	6e-62
Mh_SSGP_781018	1	acyltransferase AGPAT6 [Heliothis subflexa]	4e-37
Mh_SSGP_5517M13F21	1	enolase-like, partial [Diuraphis noxia]	8e-51
Mh_SSGP_20418M13F21	1	mannose-1-phosphate guanyltransferase alpha [Lucilia cuprina]	3e-96
Mh_SSGP_18017M13F21	1	4-(cytidine 5'-diphospho)-2-C-methyl-D-erythritol kinase [Butyrivibrio hungatei]	8.4
2. Structural proteins			
Mh_SSGP_731011	1	histone H1 [Rhynchosciara americana]	5e-21
84	2	histone deacetylase 8-like [Biomphalaria glabrata]	6e-13
Mh_SSGP_12518M13F21	1	adenosine-5'-phosphosulfate reductase alpha subunit, partial [uncultured marine bacterium]	0.003
3. Protein processing and degradation			
Mh_SSGP_481012	1	ubiquitin-conjugating enzyme E2 G1 isoform X2 [Stomoxys calcitrans	5e-09
44	2	serine/arginine repetitive matrix protein 2-like [Eurytemora affinis]	2e-7
Mh_SSGP_3317M13F21	1	serine protease inhibitor dipetalogastin-like [Nicrophorus vespilloides]	5e-09
Mh_SSGP_10418M13F21	1	Metallo carboxypeptidase inhibitor [Dorcoceras hygrometricum]	2e-08
77	3	putative signal peptidase 12kda subunit [Aedes albopictus] no signal peptide	2e-38
3	2	signal peptidase complex catalytic subunit SEC11C isoform 1 [Homo sapiens]	6e-96
4. Transportation			
Mh_SSGP_561013	1	zinc transporter [Culex quinquefasciatus]	3e-49
Mh_SSGP_16117M13F21	1	transport and Golgi organization protein 11 isoform X2 [Zootermopsis nevadensis]	3e-22

Mh_SSGP_261017	1	B-cell receptor-associated protein 31-like [Aegilops tauschii subsp. tauschii]	9e-43
Mh_SSGP_111014	1	transposase, partial [Salmonella enterica subsp. enterica serovar Kentucky]	7e-105
Mh_SSGP_36123M13F21	1	endoplasmic reticulum mannosyl-oligosaccharide 1,2-alpha-mannosidase [Drosophila bipectinata]	2e-04
E. Other proteins **1. Regulation**			
82	5	CHK1 checkpoint-like protein, partial [Helicoverpa armigera]	7e-19
94	2	serine/threonine-protein kinase/endoribonuclease IRE1-like protein [Sarcoptes scabiei]	7e-06
101	2	galactokinase [Danio rerio]	5e-14
112	2	Iron-sulfur cluster assembly enzyme [Drosophila melanogaster]	8e-84
Mh_SSGP_1451017	1	phosphatase type 2c [Culex quinquefasciatus]	3e-09
2. Others			
Mh_SSGP_6018M13F21	1	Zgc:158463 protein [Danio rerio]	4e-14
Mh_SSGP_15518M13F21	1	catenin alpha [Drosophila biarmipes]	1e-15
87	2	death-associated protein 1 [Drosophila willistoni]	2e-24
51	2	outer envelope pore protein 24, chloroplastic-like [Aegilops tauschii subsp. tauschii]	2e-28
Mh_SSGP_21217M13F21	1	nucleoside diphosphate kinase [Orseolia oryzae]	8e-80
Mh_SSGP_1311018	1	annexin D7-like [Aegilops tauschii subsp. tauschii]	1e-78
Mh_SSGP_24117M13F21	1	FI09342p, partial [Drosophila melanogaster]	4e-83
Mh_SSGP_1918M13F21	1	natterin-4 isoform X3 [Aedes aegypti]	1e-61
Mh_SSGP_771011	1	CLUMA_CG015009, isoform A [Clunio marinus]	1e-05
Mh_SSGP_24317M13F21	1	CLUMA_CG009854, isoform A [Clunio marinus]	0.019
Mh_SSGP_20317M13F21	1	Mde8i18_4 [Mayetiola destructor]	6e-04

Table S3. Single nucleotide and other small mutations among members in SSGP gene families of barley midge.

Table S3-1 : Sequence variations among group 1 members.

Base position	**Base change**	**Amino acid change**	**Clone name with base change**
Sequence variation in the mature protein coding region, group 1			
4	**C**CT-**G**CT	P-A	59123M13F21,691017
8	G**C**A-G**T**A	A-V	59123M13F21,691017
11	G**G**A-G**C**A	G-A	59123M13F21,691017,151012
13,14,15	INDEL		59123M13F21,691017
25 25	**G**CA-**C**CA **G**CA-**A**CA	A-P A-T	17123M13F21,31013 151012
39	GC**A**-GC**T**	-	59123M13F21
61	**A**CA-**G**CA	T-A	59123M13F21,691017
66	CC**A**-CC**T**	-	17123M13F21,31013,151012
74,75	G**CC**-G**AA**	A-E	59123M13F21,691017
76	**C**AA-**G**AA	Q-E	59123M13F21,691017
79,80	**GA**A-**AC**A	E-T	17123M13F21,31013,151012
83,84	G**AA**-G**TC**	E-V	17123M13F21,31013,151012
85,87	**T**CA-**C**CA	S-P	59123M13F21,691017
89,90	A**TC**-A**AT**	I-N	17123M13F21,31013,151012
91	**G**AC-**A**AC	D-N	31013, 151012
96	TT**C**-TT**A**	F-L	17123M13F21
97,98	**TT**T-**AA**T	F-N	59123M13F21,691017
101,102	G**AT**-G**CA**	D-A	59123M13F21,691017
105	GA**T**-GA**C**	-	59123M13F21,691017
107,108	G**AT**-G**GA**	D-G	59123M13F21,691017
110,111	T**CC**-T**TA**	S-L	31013, 151012
112,113,114	**GAT**-**ATA**	D-I	59123M13F21,691017
115	**G**AT-**A**AT	D-N	59123M13F21,691017

118,119	**GA**T-**TT**T Indel	D-F	17123M13F21 59123M13F21,691017
121,122,123	Indel		59123M13F21,691017,31013, 151012
124,125,126	Indel		59123M13F21,691017
129 127,128,129	TG**G**-TG**T** Indel	W-C	17123M13F21 59123M13F21,691017
135 133,134,135	TT**C**-TT**T** Indel	-	17123M13F21 59123M13F21,691017
140	G**A**A-G**G**A	E-G	31013, 151012
142,143 142	**AA**A-**CC**A **C**CA-**T**CA	K-P P-S	31013, 151012 17123M13F21
145,147	**A**T**G**-**C**T**T**	M-L	59123M13F21,691017
148,150	**T**T**T**-**C**T**G**	F-L	59123M13F21,691017
152,153	G**AT**-G**GG**	D-G	59123M13F21,691017
154,156	**T**T**C**-**C**T**T**	F-L	59123M13F21,691017
157,158	**GA**T-**AG**T	D-S	17123M13F21
160,161 160	**GA**A-**AC**A G**A**A-G**C**A	E-T E-A	31013, 151012 17123M13F21
165	CC**A**-CC**C**	-	59123M13F21,691017
180	**A**CA-**G**CA	T-A	17123M13F21
189	GC**A**-GC**T**	-	17123M13F21
190 191,192	**G**CT-**A**CT A**CT**-A**AA**	A-T T-K	59123M13F21 17123M13F21
196,197,198	**CCA**-**AAG**	P-K	17123M13F21
205	**G**CA-**T**CA	A-S	17123M13F21
217	**T**CA-**G**CA	S-A	59123M13F21,691017,17123M13F21
220,221,222	**AAG**-**CCA**	K-P	59123M13F21,691017,17123M13F21
316-321	Indel		59123M13F21,691017
322	**C**CA-**G**CA	P-A	59123M13F21,691017
325	**G**CA-**T**CA	A-S	59123M13F21,691017
328	**G**CA-**C**CA	A-P	59123M13F21,691017
340	**C**CA-**G**CA	P-A	59123M13F21,691017

Sequence variation in the signal peptide coding region			
16	**T**TT-**C**TT	F-L	59123M13F21,691017
53	G**C**T-G**T**T	A-V	59123M13F21,691017
58	**G**CT-**A**CT	A-T	151012
Sequence variation in the non-coding regions			
Base position	**Base change**		**Clone name with base change**
22	A-C		59123M13F21,691017
24	A-T		59123M13F21,691017
25	C-T		59123M13F21,691017
29	T-C		59123M13F21,691017
43	A-G		59123M13F21,691017
47	A-T		59123M13F21,691017
427	C-T		691017
430	C-T		691017
457	G-C		151012, 17123M13F21
469	T-C		59123M13F21
471	A-T		59123M13F21
485	T-G		59123M13F21
594	T-C		691017,31013

Table S3-2 : Sequence variations among group 4 members.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 4			
3	GC**A**-GC**C**	-	24718M13F21
4,5	**C**AC-**A**AC	H-N	24718M13F21
15	CA**C**-CA**T**	-	24718M13F21
20	G**C**A-G**T**A	A-V	15817M13F21,981011,1261012
23	G**G**T-G**A**T	G-D	15817M13F21,981011,1261012
32	C**A**T-C**T**T	H-L	24718M13F21
45	Indel		1261012
47,48	A**AG**-A**GC**	K-S	1261012
49	**C**AA-**A**AA	Q-K	1261012
52	**A**CG-**T**CG	T-S	24718M13F21
58,60	**C**A**A**-**A**A**T**	Q-N	1261012
60	CA**A**-CA**T**	Q-H	24718M13F21
61,62	**TC**A-**CA**A	S-Q	1261012
64,65	**AG**C-**GC**C	S-A	1261012
68	C**T**A-C**C**A	L-P	24718M13F21
67,68,69	**CTA**-**TAG**	L- Stop codon	1261012
71,72	G**GT**-G**TG**	G-V	1261012
76,77,78	**CAT**-**ATC**	H-I	1261012
77	C**A**T-C**T**T	H-L	15817M13F21
79,81	**C**T**T**-**T**T**G**	-	1261012
83	G**A**T-G**G**T	D-G	24718M13F21,24917M13F21
85,87	**G**A**A**-**A**A**G**	E-K	1261012
89,90	G**GT**-G**TG**	G-V	1261012
94,95,96	**CAC**-**ACT**	H-T	1261012
96	CA**C**-CA**T**	-	15817M13F21

98	Indel		4018M13F21
97	**G**GT-**T**GT	G-C	4018M13F21
100,101,102	**TAC**-**ACG**	Y-T	1261012,4018M13F21
102	TA**C**-TA**G**	Y-Stop codon	15817M13F21
103,104,105	**GAT**-**ATG**	D-M	1261012,4018M13F21
110	T**T**G-T**G**G	L-W	1261012,4018M13F21
112,113,114	**GAC**-**ACA**	D-T	1261012,4018M13F21
117	AA**A**-AA**G**	-	1261012,4018M13F21
118,120	**G**A**A**-**A**A**T**	E-N	1261012,4018M13F21
124,126	**G**A**A**-**A**A**T**	E-N	1261012,4018M13F21
127,128	**TC**A-**CA**A	S-Q	1261012,4018M13F21
133,134,135	**TAC**-**ACG**	Y-T	1261012,4018M13F21
136,138	**G**A**A**-**A**A**G**	E-K	1261012,4018M13F21
139,140,141	**GAT**-**ATG**	D-M	1261012,4018M13F21
146,147	A**AT**-A**TG**	N-M	1261012,4018M13F21
150	GG**G**-GG**C**	-	1261012,4018M13F21
152,153	C**CG**-C**GA**	P-R	1261012,4018M13F21
160,161	**GC**A-**CA**A	A-Q	1261012,4018M13F21
163,165	**A**C**C**-**C**C**G**	T-P	1261012,4018M13F21
168	GT**A**-GT**T**	-	24718M13F21
170	C**C**G-C**G**G	P-R	1261012,4018M13F21
201	GT**C**-GT**G**	-	15817M13F21
245	C**A**C-C**C**C	H-P	15817M13F21
249	CC**A**-CC**C**	-	24718M13F21
Sequence variation in the signal peptide coding region			
29	A**T**C-A**C**C	I-T	24917M13F21
37,38	**G**T**C**-**T**T**T**	V-F	24718M13F21
59	G**C**T-G**T**T	A-V	24718M13F21
Sequence variation in the non-coding regions			

Base position	Base change	Clone name with base change
13	T-C	24718M13F21
15	A-G	981011
25	C-G	24718M13F21
34	A-T	15817M13F21,981011
43	A-C	24718M13F21
380	A-C	24718M13F21
385	A-T	15817M13F21
408	A-G	15817M13F21
414	C-G	24718M13F21
434	A-G	15817M13F21,981011
442	C-T	15817M13F21,981011
444	C-G	15817M13F21,981011
452	C-T	24718M13F21
472	C-T	24718M13F21,15817M13F21
477	C-T	24718M13F21
487	A-C	15817M13F21,981011
492	A-G	24718M13F21

Table S3-3 : Sequence variations among group 31 members.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 31			
155,156	Indel		24517M13F21,2817M13F21
157	Indel		1211012
159	AA**A**-AA**T**	K-N	1211012
160,161,162	**TCA**-**AAT**	S-N	521018
163,164,165	**ATG**-**TGA**	M-W	1211012
166,167	**C**T**T**-**TT**G	-	521018
170,171	C**CA**-C**AG**	P-Q	1211012
172,174	**G**A**A**-**C**A**G**	E-Q	521018
175,167	**TT**T-**AA**T	F-T	521018
178	GT**C**-GT**G**	-	
179,180	T**GG**-T**TT**	W-F	521018
181	**G**GA-**A**GA	G-R	1211012
182,183	G**GA**-G**AG**	G-E	2817M13F21
185,186	A**AG**-A**GT**	K-S	1211012
186	AG**T**-AG**A**	S-R	521018
189	TT**T**-TT**G**	F-L	1211012
187,188	**TT**T-**AG**T	F-S	521018
190	**G**AA-**A**AA	E-K	1211012
197,198	A**GT**-A**AA**	S-K	521018
201	TT**T**-TT**G**	F-L	1211012
202	AA**A**-AA**G**	-	
207	AA**T**-AA**A**	N-K	521018
209	A**A**A-A**T**A	K-I	521018
213	AA**C**-AA**A**	N-K	521018
215	A**A**G-A**G**G	K-R	1211012
215,216	A**AG**-A**CA**	K-T	521018
217,218,219	**GCA**-**AGG**	A-R	521018

220,221,222 222	**CAG**-**AGA** CA**G**-CA**C**	Q-R Q-H	1211012 521018
224,225	A**AC**-A**CA**	N-T	1211012
226	AA**A**-AA**G**	-	521018
227,228	A**TT**-A**CA**	I-T	521018
229,231	**G**A**A**-**A**A**C**	E-N	1211012
232,233,234	**CAC**-**ACT**	H-T	1211012
Sequence variation in the signal peptide coding region			
54	GA**G**-GA**A**	-	521018
Sequence variation in the non-coding regions			
Base position	**Base change**		**Clone name with base change**
373	T-C		2817M13F21,521018
416 416	A-C A-T		24517M13F21,521018 2817M13F21
418	A-C		2817M13F21
419	A-T		2817M13F21

Table S3- : Sequence variations among group 32 members

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 32			
13	**C**TA-**T**TA	-	1021012, 1311014
26	G**T**T-G**C**T	V-A	1021012, 1311014
53	G**G**A-G**A**A	G-E	26117M13F21
82	**A**CT-**G**CT	T-A	1021012, 1311014
137	A**A**A-A**C**A	K-T	26117M13F21, 491013
150	Indel		26117M13F21, 491013, 1021012
153	AA**T**-AA**A**	N-K	1311014
154,155,156	**CTA**-**TCT**	L-S	1311014
158 158,159	A**T**T-A**C**T A**TT**-A**AC**	I-T I-N	1021012 1311014
163,164	**GA**A-**AG**A	E-R	1311014
166,167	**GA**A-**AG**A	E-R	1311014
169,171	**T**C**A**-**A**C**T**	S-I	1311014
172,173	**GA**A-**AG**A	E-R	1311014
175,177 177	**T**T**C**-**A**T**T** TT**C**-TT**A**	F-I F-L	1311014 1021012
178,179,180	**GAC**-**AGA**	D-R	1311014
184,186	**G**G**A**-**T**G**G**	G-W	1311014
187,189	**G**G**C**-**A**G**G**	G-R	1311014
190,192	**T**T**A**-**C**T**T**	-	1311014
193,195	**G**G**C**-**A**G**G**	G-R	1311014
198	TT**C**-TT**T**	-	1311014
199,200,201	**GTC**-**CGT**	V-R	1311014
203	**C**TA-**T**TA	-	1311014
206	T**G**G-T**T**G	W-L	1311014
208,210	**A**A**T**-**G**A**A**	N-E	1311014

216	TT**C**-TT**T**	-	1311014
221,222	G**CA**-G**GC**	A-G	1311014
225	AA**T**-AA**A**	N-K	1311014
226-240	Indel		1311014
Sequence variation in the signal peptide coding region			
None			
Sequence variation in the non-coding regions			
Base position	**Base change**		**Clone name with base change**
1,2	Indel		26117M13F21, 491013
3	Indel A-G		491013 26117M13F21
4	G-T		1021012, 1311014
349	T-C		1021012, 1311014
402	T-A		491013
408	T-A		1021012, 1311014
451	C-T		491013,1021012
452	G-C G-A		26117M13F21 1311014
453	C-A		1021012, 1311014

Table S3- : Sequence variations among group 33 members

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 33			
10,11	**AA**T-**CC**T	N-P	741012
14,15	A**CC**-A**TG**	T-M	741012
23	G**G**T-G**C**T	G-A	741012
28	**C**GT-**T**GT	R-C	13218M13F21,1091014
31	AA**T**-AA**C**	-	13218M13F21,1091014
34	**C**AT-**T**AT	H-Y	13218M13F21,1091014
40	**T**AC-**A**AC	Y-N	13218M13F21,1091014
91	**G**AA-**A**AA	E-K	13218M13F21,1091014
94	**A**AG-**G**AG	K-E	13218M13F21,1091014
148	TT**C**-TT**T**	-	13218M13F21
160	**G**AA-**A**AA	E-K	13218M13F21,1091014
169,171	**G**A**G**-**A**A**C**	E-N	741012
186	GT**C**-GT**G**	-	741012
193	**G**AT-**A**AT	D-N	741012
199	**T**TC-**C**TC	F-L	13218M13F21
215	A**A**C-A**T**C	N-I	741012
223	**A**AA-**C**AA	K-Q	13218M13F21,1091014
226	**A**AT-**T**AT	N-Y	741012
252	AA**T**-AA**C**	-	741012
253,254,255	Indel		13218M13F21,1091014
256,257,258	**GTT**-**AAA**	V-K	741012
259,260	**TG**G-**AA**G	W-K	741012
281	G**C**A-G**T**A	A-V	741012
290	A**A**C-A**T**C	N-I	13218M13F21,1091014
295	**A**AT-**G**AT	N-D	13218M13F21,1091014

316	**A**AA-**G**AA	K-E	741012
329,330	A**CA**-A**AT**	T-N	13218M13F21,1091014
334	GC**T**-GC**C**	-	741012
341	A**G**T-A**T**T	S-I	741012
344	C**T**A-C**C**A	L-P	741012
356	A**G**C-A**A**C	S-N	741012
362	G**A**T-G**C**T	D-A	741012
364	**A**AA-**G**AA	K-E	13218M13F21,1091014
373,374	**GT**G-**AA**G	V-K	13218M13F21,1091014
377	C**A**A-C**C**A	Q-P	741012
380	C**C**A-C**T**A	P-L	741012
388	**A**AA-**G**AA	K-E	13218M13F21,1091014
393	TT**C**-TT**A**	F-L	741012
399,401	**T**T**A**-**C**T**T**	-	741012
412	**A**TC-**G**TC	I-V	741012
421	**C**AG-**A**AG	Q-K	13218M13F21,1091014
443	G**A**A-G**G**A	E-G	13218M13F21,1091014
453	GC**T**-GC**C**	-	741012
454	**C**TA-**A**TA	L-I	741012
460	**A**AA-**G**AA	K-E	741012
Sequence variation in the signal peptide coding region			
14	T**T**A-T**C**A	L-S	741012
16	**A**TT-**G**TT	I-V	741012
Sequence variation in the non-coding regions			
Base position	**Base change**		**Clone name with base change**
2	A-G		1091014
3	C-A		1091014
4	A-C		1091014
5	T-A		1091014

6	A-T	1091014
593	C-T	13218M13F21, 741012
596	A-C	13218M13F21, 741012
598	T-A	1091014, 741012

Table S4. Distribution of SSGP transcripts among different groups belong to oat midge (transcripts in each group and their related proteins are listed in order in Figure S3)

Gene	Number of unique sequences	Number of total sequences	First BLAST hit if any	E-value
Group 1 Subgroup C Subgroup D	90	283	SSSGP-1A2 [Mayetiola destructor] SSSGP-1C1 [Mayetiola destructor] SSSGP-1D1 [Mayetiola destructor]	4e-21 5e-55 3e-24
Group 2	4	4		
Group 3	11	28		
Group 4	3	4		
Group 5	8	13	SSGP-11B family protein [Mayetiola destructor]	5e-34
Group 6	2	2		
Group 7	1	2	salivary secreted protein [Mayetiola destructor]	5e-43
Group 8	6	19		
Group 9	2	2		
Group 10	1	2		
Group 11	1	2		
Group 12	2	2		
Group13	1	2		
Group 14	2	2		
Group 15	7	15		
Group 16	2	4		
Group 17	7	9		
Group 18	1	4		
Group 19	2	2		
Group20	1	2		
Group 21	2	7		
Group22	3	3	secreted salivary gland protein SSGP-4A5 [Mayetiola destructor]	1e-08
Group 23	1	2		
Group 24	6	7		
Group 25	1	1		
Group26	1	1		
Group27	1	1		
Group28	1	1		
Group29	2	2	secreted protein F [Mayetiola destructor]	1e-65

Group30	1	1	peptidyl-prolyl cis-trans isomerase FKBP2 precursor [Danio rerio]	6e-62
Group 31	1	1		
Group 32	1	1		
Group33	1	1		
Group34	1	1		
Group35	1	1	larval cuticle protein 8-like [Zeugodacus cucurbitae]	2e-28
Group36	1	1		
Group 37	1	1		
Group38	1	1		
Group39	1	1		
Group40	1	1		
Group41	1	1		
Group42	1	1		
Group43	1	1		
Group44	2	2	Family 71 [Mayetiola destructor]	1e-19
Group45	1	1		
Group46	1	1		
Group47	1	1	putative secreted protein, partial [Mayetiola destructor]	1e-29
Group48	1	1		
Group49	1	1		
Group50	1	1		
Total	194	450		

Table S5. Classification of proteins encoded by non-SSGP transcripts belong to oat midge

#	No.	First hit	E-value
F. Protein synthesis and folding			
6. Ribosomal protein			
12	2	60S acidic ribosomal protein P0 [Lucilia cuprina]	1e-34
20	3	ribosomal protein S27, isoform A [Drosophila melanogaster]	3e-42
31	3	ribosomal protein S13, isoform A [Drosophila melanogaster]	1e-50
32	5	40S ribosomal protein S11 [Danio rerio]	4e-33
54	2	ribosomal protein L41 [Drosophila melanogaster]	6e-20
56	3	ribosomal protein L37A, isoform A [Drosophila melanogaster]	1e-58
72	3	ribosomal protein S24 [Drosophila melanogaster]	3e-78
maven_SSGP_88_12_15_M13F	1	ribosomal protein L27A, isoform A [Drosophila melanogaster]	7e-56
maven_SSGP_221_12_15_M13F	1	ribosomal protein S3A, isoform B [Drosophila melanogaster]	2e-59
maven_SSGP_233_12_15_M13F	1	ribosomal protein L11, isoform A [Drosophila melanogaster]	1e-123
maven_SSGP_86_12_16_M13F	1	ribosomal protein S26, isoform B [Drosophila melanogaster]	2e-65
maven_SSGP_108_12_16_M13F	1	ribosomal protein S4, isoform A [Drosophila melanogaster]	8e-154
maven_SSGP_198_12_16_M13F	1	ribosomal protein S23, isoform A [Drosophila melanogaster]	2e-99
maven_SSGP_123_12_18_M13F	1	ribosomal protein S8, isoform C [Drosophila melanogaster]	1e-28
maven_SSGP_77_11_28	1	ribosomal protein L17 isoform A [Lysiphlebus	7e-22
maven_SSGP_223_11_28	1	ribosomal protein L38, isoform A [Drosophila melanogaster]	1e-38
maven_SSGP_256_11_28	1	ribosomal protein L37a, isoform A [Drosophila melanogaster]	2e-48
maven_SSGP_168_11_29	1	ribosomal protein L24, isoform A [Drosophila melanogaster]	2e-83
maven_SSGP_60113M13F21	1	ribosomal protein L9, isoform A [Drosophila melanogaster]	6e-118
maven_SSGP_92113M13F21	1	ribosomal protein S21, isoform A [Drosophila melanogaster]	1e-40
maven_SSGP_126114M13F21	1	ribosomal protein S15, isoform B [Drosophila melanogaster]	4e-83
60	2	ribosomal protein S19a, isoform A [Drosophila melanogaster]	5e-60
2. Transcription & translation factors			
5	4	putative reverse transcriptase [Zingiber officinale]	4e-19
57	2	transcriptional repressor NrdR [Streptobacillus notomytis]	0.024
84	3	putative reverse transcriptase, partial [Cicer arietinum]	1e-21
3. Chaperones, protein folding and secretion			
44	2	heat shock protein 26, isoform A [Drosophila melanogaster]	3e-39
79	2	chaperone protein dnaJ 1, mitochondrial-like isoform X1 [Glycine max]	1e-08
85	2	heat shock 70-kDa protein cognate 3, isoform A [Drosophila melanogaster]	8e-32
34	4	signal peptidase complex subunit 1 [Microplitis demolitor]	2e-34
23	2	COP9 signalosome complex subunit 9-like [Aedes albopictus]	3e-27
82	2	type II secretion system protein GspG [Desulfovibrio bastinii]	4e-05
maven_SSGP_20_12_15_M13F	1	Secretion-associated Ras-related 1, isoform A [Drosophila melanogaster]	4e-95

4. RNA synthesis and processing			
83	2	Aspartyl-tRNA synthetase, mitochondrial, isoform A [Drosophila melanogaster]	1e-23
G. House keeping **Energy metabolism**			
18	2	ATP synthase subunit a-like [Protobothrops mucrosquamatus]	8e-04
76	2	cytochrome c oxidase assembly protein COX11, mitochondrial [Mus musculus]	2e-12
maven_SSGP_24113M13F21	1	GTP-binding protein [Drosophila melanogaster]	4e-40
maven_SSGP_123113M13F21	1	cytochrome c oxidase subunit 6A1, mitochondrial [Ooceraea biroi]	4e-42
maven_SSGP_167113M13F21	1	glyceraldehyde 3 phosphate dehydrogenase 2, isoform A [Drosophila melanogaster]	9e-56
86	2	glycosyl transferase family 1 [Butyrivibrio sp. XBB1001]	0.024
maven_SSGP_16119M13F21	1	adenylate kinase 3, isoform A [Drosophila melanogaster]	3e-72
maven_SSGP_171_12_18_M13F	1	AGAP007700-PA-like protein [Anopheles sinensis]	3e-21
maven_SSGP_247_12_18_M13F	1	G protein alpha s subunit, isoform A [Drosophila melanogaster]	3e-142
maven_SSGP_16_12_16_M13F	1	HIG1 domain family member 2A, mitochondrial [Danio rerio]	5e-13
2. Structural proteins			
62	4	sensor histidine kinase [Terrabacteria group]	3e-04
maven_SSGP_2_8_10_31	1	carbamoyl transferase [Escherichia coli str. K-12 substr. MG1655]	4e-162
maven_SSGP_180113M13F21	1	intronic protein 259, isoform A [Drosophila melanogaster]	7e-125
3. Protein processing and degradation			
maven_SSGP_120_12_18_M13F	1	small integral membrane protein 14 [Homo sapiens]	5e-13
13	4	N-acetylneuraminate lyase isoform X1 [Tribolium castaneum]	3e-19
maven_SSGP_77_12_18_M13F	1	Nedd8, isoform A [Drosophila melanogaster]	1e-48
4. Transportation			
maven_SSGP_217_11_29	1	putative transporter [Escherichia coli str. K-12 substr. MG1655]	3e-143
71	3	resistance-nodulation-cell division (RND) efflux transporter [Pseudomonas aeruginosa PAO1]	3e-102
maven_SSGP_2_8_12_13	1	Ef1gamma, isoform A [Drosophila melanogaster]	5e-98
H. Other proteins **Detoxification**			
21	1	chloramphenicol acetyltransferase [Promoter probe vector pEvoGlowRed]	5e-29
45	4	fluoroquinolone resistance protein, partial [Escherichia coli]	0.002
maven_SSGP_99113M13F21	1	superoxide dismutase 1, isoform A [Drosophila melanogaster]	2e-68
Others			

38	3	spider venom protein NPTX_B154 [Nephila pilipes]	1e-04
60	3	venom allergen 3-like, partial [Wasmannia auropunctata]	0.002
maven_SSGP_140_12_15_M13F	1	death-associated protein 1 [Drosophila busckii]	3e-32
maven_SSGP_178114M13F21	1	lethal (1) 10Bb [Drosophila melanogaster]	3e-96
maven_SSGP_116123M13F21	1	twisted bristles roughened eye [Drosophila melanogaster]	2e-42

Table S6. Single nucleotide and other small mutations among members in SSGP gene families of oat midge. Table S6-1 : Sequence variations among group 1 members.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 1			
1,2,3	**GCT**-**AAA**	A-K	1401128,2731218M13F,177113M13F,73119M13F
5,6	A**AA**-A**GC**	K-S	1401128,2731218M13F,177113M13F,73119M13F
7,9	**G**C**T**-**A**C**A**	A-T	1401128,2731218M13F,177113M13F,73119M13F
16	**G**CC-**C**CC	A-P	1041215M13F,2581215M13F
35	**G**GA-**G**CA	G-A	177113M13F,73119M13F,280113M13F
40	**A**AA-**G**AA	K-E	1401128
53	T**C**T-T**T**T	S-F	73119M13F,2731218M13F
59	C**C**A-C**T**A	P-L	9713M13F21
82	**C**AA-**A**AA	Q-K	32114M13F
97	**C**TT-**T**TT	L-F	32114M13F, 1041215M13F,73119M13F
106	**G**AT-**C**AT	D-H	73119M13F
109	**C**AA-**G**AA	Q-E	1401128,2731218M13F,9713M13F21,32114M13F,441215M13F
125	CC**G**-CC**A**	-	1041215M13F,2581215M13F
133 135	**A**TG-**C**TG AT**G**-AT**A**	M-L M-I	2731218M13F,9713M13F21 2581215M13F,1041215M13F
186	CC**G**-CC**A**	-	280113M13F,1401128,73119M13F,177113M13F
195	GC**A**-GC**C**	-	1041215M13F,441215M13F
209 210	GA**C**-GA**A** G**A**C-G**G**C	D-E D-G	2731218M13F,73119M13F 32114M13F
217,218	**AA**A-**GG**A	K-G	1401128,177113M13F,2581215M13F,9713M13F21, 441215M13F
220	**A**AA-**C**AA	K-Q	177113M13F,441215M13F
223,224	**AA**A-**CC**A	K-P	1401128,177113M13F,2581215M13F,9713M13F21, 441215M13F
231	AA**A**-AA**G**	-	9713M13F21, 441215M13F,1401128,177113M13F
232,233,234	**AAA**-**GCC**	K-A	1401128,177113M13F,2581215M13F,9713M13F21, 441215M13F
235,236	**AA**A-**GC**A	K-A	1401128,177113M13F,2581215M13F,9713M13F21, 441215M13F
238,239	**TC**A-**AA**A	S-K	2731218M13F,73119M13F,32114M13F,1041215M13F, 280113M13F
249	GC**A**-GC**C**	-	441215M13F
252	CC**G**-CC**A**	-	1401128,9713M13F21
253,254,255	**GCA**-**AAG**	A-K	1401128,9713M13F21
256 258	**G**CA-**T**CA TC**A**-TC**C**	A-S -	177113M13F,2581215M13F 441215M13F
Sequence variation in the signal peptide coding region			
5	T**C**T-TTT	S-F	2731218M13F,73119M13F,32114M13F,1041215M13F

16	**T**TA-**C**TA	-	73119M13F,177113M13F,2581215M13F,1041215M13F
21	GC**A**-GC**T**	-	73119M13F,177113M13F,2581215M13F,1041215M13F
39	GT**C**-GT**T**	-	2731218M13F,73119M13F
Sequence variation in the non-coding regions			
Base position	**Base change**		**Clone name with base change**
8	C-G		1041215M13F
17	C-A		177113M13F
21	T-A		177113M13F,73119M13F
27	G-T		177113M13F,73119M13F,2581215M13F,1041215M13F
31	G-T		2581215M13F,1041215M13F
35	T-G		2581215M13F,1041215M13F
36	C-T		73119M13F
37	T-C		2581215M13F,1041215M13F,2731218M13F,1401128
40	T-G		2581215M13F,1041215M13F
51	A-G		2581215M13F,1041215M13F
56	G-A		2581215M13F,1041215M13F
442	G-A		1401128
446	T-C		1401128,2581215M13F
447	A-G		1401128,2581215M13F
449	C-G		441215M13F
459	G-A		1401128,2581215M13F
467	T-C		1401128
471	T-C		1401128
495	T-G		1401128,2581215M13F
496	G-T		1401128,2581215M13F
497	T-G		1401128,2581215M13F
506	T-A		1401128,2581215M13F
515	T-C		2581215M13F
534	A-G		1401128,441215M13F

Table S6-2 : Sequence variations among group 3 members.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 3			
4,5	**GT**T-**AG**T	V-S	32113M13F
7	**G**TT-**T**TT	V-F	32113M13F
11	G**C**A-G**T**A	A-V	32113M13F
13	**C**TA-**T**TA	-	691215M13F
17	G**T**A-G**A**A	V-E	32113M13F
24	GA**A**-GA**T**	E-D	32113M13F
32	G**C**A-G**T**A	A-V	691215M13F,32113M13F,1861215M13F
54	AA**A**-AA**G**	-	1861215M13F
76	**G**AG-**A**AG	E-K	32113M13F
103	**C**AA-**A**AA	Q-K	691215M13F,1861215M13F
124	**G**AA-**A**AA	E-K	32113M13F
141	GA**A**-GA**T**	E-D	691215M13F,32113M13F,1861215M13F
142-150	INDEL		2411218M13F,681215M13F,18113M13F,1701215M13F, 58113M13F
166	**C**TA-**T**TA	-	1701215M13F, 58113M13F
182	G**C**A-G**T**A	A-V	32113M13F
213	CT**C**-CT**T**	-	18113M13F,1701215M13F, 58113M13F,1861215M13F
215	T**A**C-T**G**C	Y-C	32113M13F
219	GA**A**-GA**G**	-	2411218M13F
227	T**A**C-T**T**C	Y-F	691215M13F,32113M13F,1861215M13F
233	A**G**T-A**A**T	S-N	691215M13F,32113M13F,1861215M13F
236	A**G**A-A**A**A	R-K	32113M13F
237 239	**C**AT-**G**AT CA**T**-CA**G**	H-D H-Q	691215M13F,1861215M13F 32113M13F
250	**G**AT-**A**AT	D-N	691215M13F,32113M13F,1861215M13F
259	**G**AA-**C**AA	E-Q	691215M13F,32113M13F,1861215M13F
271	**G**AG-**A**AG	E-K	1701215M13F
291	**C**TA-**T**TA	-	691215M13F,32113M13F,1861215M13F
296	A**T**T-A**A**T	I-N	691215M13F,1861215M13F
301	**C**AT-**A**AT	H-N	691215M13F,1861215M13F
304	**A**AA-**G**AA	K-E	691215M13F,1861215M13F
313 315	**A**AA-**G**AA AA**C**-AA**A**	K-E N-K	32113M13F 691215M13F,1861215M13F
333	CA**A**-CA**G**	-	32113M13F
347	C**A**G-C**G**G	Q-R	2411218M13F
354	AA**G**-AA**C**	K-N	691215M13F,32113M13F,1861215M13F

359	T**T**G-T**G**G	L-W	1701215M13F
362	G**C**G-G**A**G	A-E	691215M13F,1861215M13F
368	AA**A**-AA**G**	-	1861215M13F
385	**G**AA-**A**AA	G-K	691215M13F,1861215M13F
397	**T**CA-**C**CA	S-P	691215M13F,1861215M13F
408	AG**T**-AG**G**	S-R	691215M13F,1861215M13F
430,431	**AG**A-**GA**A	R-E	32113M13F
439	**C**TA-**T**TA	G-	32113M13F
Sequence variation in the signal peptide coding region			
31	**G**CT-**A**CT	A-T	691215M13F
Sequence variation in the non-coding regions			
Base position	Base change		Clone name with base change
539	C-T		1861215M13F,18113M13F
562	A-C		32113M13F
564	A-T		32113M13F
565	T-A		32113M13F
569	T-A		32113M13F
571	C-T		32113M13F
574	A-T		691215M13F,1861215M13F
575	A-G		1861215M13F
576	T-C		691215M13F

Table S6-3 : Sequence variations among group 8 members.

Base position	**Base change**	**Amino acid change**	**Clone name with base change**
Sequence variation in the mature protein coding region, group 8			
16	**A**CT-**C**CT	T-P	281240, 28912
33	GA**T**-GA**G**	D-E	248113M13F
52	**C**GA-**G**GA	R-G	781218M13F
52,54	**C**G**A**-**A**G**G**	-	248113M13F
66	GG**C**-GG**G**	-	248113M13F
69	TC**A**-TC**G**	-	248113M13F
75	AC**C**-AC**G**	-	248113M13F
80	T**C**T-T**T**T	S-F	281240
100	**A**GC-**G**GC	S-G	248113M13F
148	**C**GT-**T**GT	R-C	248113M13F
168	CC**A**-CC**C**	-	281240, 28912
169	**C**GT-**G**GT	R-G	781218M13F
172	**C**GT-**G**GT	R-G	248113M13F
176	T**A**C-T**T**C	Y-F	248113M13F
Sequence variation in the signal peptide coding region			
None			
Sequence variation in the non-coding regions			
Base position	Base change		Clone name with base change
11	C-T		248113M13F
22	A-T		248113M13F
332	C-A		248113M13F
352	C-T		248113M13F
376	A-T		248113M13F
397	C-T		281240
401	C-A		248113M13F,28912

Table S6-4 : Sequence variations among group 15 members.

Base position	Base change	Amino acid change	Clone name with base change
Sequence variation in the mature protein coding region, group 15			
2	G**T**A-G**G**A	V-G	254113M13F
4,6	**C**A**C**-**A**A**A**	H-K	254113M13F
7,8,9	**AGC**-**GAT**	S-D	254113M13F
10,11,12	**AGC**-**GCT**	S-A	254113M13F
15	AA**A**-AA**C**	K-N	254113M13F
16,17,18	**TCC**-**CAA**	S-Q	254113M13F
22,24	**G**C**T**-**T**C**C**	A-S	254113M13F
25	**G**CA-**C**CA	A-P	254113M13F
30	CA**C**-CA**T**	-	254113M13F
32,33	G**CA**-G**GC**	A-G	254113M13F
35	G**C**T-G**T**T	A-V	254113M13F
40,41,42	INDEL		231218M13F,281239,51129
43,45	**A**C**C**-**G**C**A**	T-A	281239
47,48	C**AT**-C**CC**	H-P	281239
49	**G**AT-**A**AT	D-N	281239
52,53	**CC**A-**GT**A	P-V	281239
66	CA**A**-CA**C**	Q-H	281239
69	AC**A**-AC**C**	-	281239
72	GC**C**-GC**A**	-	281239
79	**G**GC-**A**GC	G-S	231218M13F,281239,51129
82	**C**TA-**G**TA	L-V	231218M13F,51129
87	GG**A**-GG**G**	-	231218M13F,281239,
89,90	G**CA**-G**AC**	A-D	231218M13F,281239,51129
93	CA**T**-CA**C**	-	51129
108	GA**G**-GA**C**	E-D	231218M13F,281239,51129
109,111	**C**A**C**-**G**A**T**	H-D	231218M13F
115,117	**T**A**C**-**C**A**A**	Y-Q	231218M13F,281239,51129
121,122	**GC**A-**TT**A	A-L	51129
125,126	T**CA**-T**TC**	S-F	231218M13F,281239,51129
130-138	INDEL		281239
131,132	G**AA**-G**TT**	E-V	231218M13F,281239,51129
134,135	G**AA**-G**TT**	E-V	231218M13F,281239,51129
139,140	**GA**A-**AG**A	E-R	254113M13F
142,143	**CT**A-**AA**A	L-K	254113M13F
145,147	**A**T**G**-**G**T**T**	M-V	254113M13F
161,162	A**GT**-A**AA**	S-K	254113M13F

162	AG**T**-AG**G**	S-R	28123
171	AA**T**-AA**A**	N-K	94114M13F
172,173,174	**GAA**-**CCG**	E-P	94114M13F
179,180	G**CC**-G**AT**	A-D	94114M13F
182	G**T**A-G**C**A	V-A	94114M13F
184,186	**C**C**G**-**T**C**C**	P-S	94114M13F
195	GC**A**-GC**T**	-	94114M13F
196,197,198	**ACG**-**GTC**	T-V	94114M13F
201	CA**C**-CA**T**	-	231218M13F,281239
205,206	**CC**A-**GG**A	P-G	254113M13F
210	AA**G**-AA**C**	K-N	254113M13F
211,212,213	**GCG**-**CAA**	A-Q	254113M13F,281239,51129
216	GT**C**-GT**G**	-	254113M13F
217,218,219	**GGA**-**ACG**	G-T	254113M13F
221,222	A**AA**-A**CC**	K-T	254113M13F
223	GG**A**-GG**G**	-	254113M13F, 281239,51129
226	**A**CA-**C**CA	T-P	254113M13F
240,242	**G**C**C**-**C**C**A**	A-P	254113M13F
243,245	AAT-CAG	N-Q	254113M13F
Sequence variation in the signal peptide coding region			
Non			
Sequence variation in the non-coding regions			
Base position	Base change		Clone name with base change
8	C-T		51129
9	G-C		254113M13F
14	G-C		254113M13F
22	A-C		254113M13F
40	A-G		254113M13F
41	A-G		254113M13F
44	G-A		254113M13F
46	C-A		254113M13F
48	C-T		231218M13F,281239
52	A-G		281239
57	G-A		281239
403	T-A		51129
432	T-G		231218M13F,281239
435	T-C		281239
468	C-T		51129
476	T-C		231218M13F,281239
484	INDEL		51129

487	A-C	51129
516	C-T	231218M13F,281239,51129
545	G-A	231218M13F,281239

Appendix C - Supplementary data, Chapter 4

Figure S8. pET-28a cloning vector. A, vector map showed the multiple cloning sites and location of restriction enzymes. B, sequence of the insert gene and the specific primers, arrows on the vectors sequence indicates to the selected restriction enzymes.