

An investigation of novel traits in the order Lepidoptera through the use of proteomic methods

By

Desiree L. Harpel

Submitted to the graduate degree program in Ecology and Evolutionary Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Master of Arts.

Chairperson: James Walters

John Kelly

Jennifer Gleason

Date Presented: April 22, 2015

The Thesis Committee for Desiree L. Harpel
certifies that this is the approved version of the following thesis:

An investigation of novel traits in the order Lepidoptera
through the use of proteomic methods

Chairperson: James Walters

Date approved: July 28, 2015

Abstract

This thesis investigates two biological novelties found in Lepidoptera through the use of shotgun proteomics. The first chapter delves into an adaptation of pollen feeding found in *Heliconius* butterflies. We used shotgun proteomics to discover the saliva proteome, and furthermore, discover protein candidates that could be used in digestion and degradation of pollen. The second chapter explores the phenomenon of sperm heteromorphism. Although sperm heteromorphism occurs sporadically throughout many species, this heteromorphism is uniquely found in all Lepidoptera besides the most basal species. Using shotgun proteomics, we identified and functionally annotated the sperm proteome for *D. plexippus*, the monarch butterfly. Both of these projects furthered research about novel Lepidopteran traits through the use of a molecular approach.

The first thesis chapter reports an initial shotgun proteomic analysis of saliva from *Heliconius melpomene*. While most adult Lepidoptera use flower nectar as their primary food source, butterflies in the genus *Heliconius* have evolved the novel ability to acquire amino acids from consuming pollen. Little is known about the molecular mechanisms of this complex pollen feeding adaptation. Using liquid-chromatography tandem mass-spectrometry, we confidently identified 31 salivary proteins. Further bioinformatic annotation of these salivary proteins indicated the presence of four distinct functional classes: proteolysis (10 proteins), carbohydrate hydrolysis (5), immunity (6), and “housekeeping”(4). These results offer a first glimpse into the molecular foundation of *Heliconius* pollen feeding and provide a substantial advance towards comprehensively understanding this striking evolutionary novelty.

The second chapter explores the phenomenon of sperm heteromorphism in *D. plexippus*. Lepidoptera have sperm heteromorphy, in which males produce both a fertilizing (eupyrene) and non-fertilizing sperm (apyrene). Apyrene sperm lacks DNA and a nucleus and is produced in higher quantities than eupyrene sperm. The recent availability of several completely sequenced Lepidopteran genomes presents a novel opportunity to apply proteomic methods to characterize the protein content of these two heteromorphic sperm morphs. Here I report an analysis of the sperm proteome from *Danaus plexippus*, the monarch butterfly. I have analyzed a “commixed” sample of combined apyrene and eupyrene sperm as well as isolated fractions each sperm type. Notably, this study is the first to use shotgun proteomics to independently characterize the protein content of purified apyrene and eupyrene sperm. Comparison of the commixed sample to comparable proteomes of commixed *Manduca sexta* (Hornworm moth) sperm and monomorphic *Drosophila melanogaster* sperm revealed few functional differences but high levels of gene turnover. Functional annotations were assessed through the use of Blast2GO software. Next, we delved into the differences in proteins found between the apyrene and eupyrene sperm types, and found few differences functionally and molecularly between the two proteomes. These results reveal a first pass analysis of the sperm proteome found in *D. plexippus*, and offer insights into the next steps to be taken to further our knowledge on the phenomenon of sperm heteromorphism.

Acknowledgments

I would like to sincerely thank my advisor and mentor, Dr. James Walters, for the amount of time and effort he has devoted to training me in the ways of science. Not only has he trained me in the methodology of research, but has also taught me how to keep an open mind, to question everything, to think about the bigger picture, and to keep my head up when things do not go as planned. Without his advice and mentoring, I would have never completed these projects and become the researcher I am today.

I would also like to thank my committee members, Dr. Jennifer Gleason and Dr. John Kelly for their support and advice along the way. I am deeply grateful to work with such excellent scientists that have such differing viewpoints and areas of expertise. Their counsel and support has made me a more rounded scientist.

I would also like to thank my lab members and the graduate students of Ecology and Evolutionary Biology for their amazing support and helpful advice on my research. I have never been to a University where the sharing of ideas and research was so encouraged among the student body. I would like to thank in particular Kaila Colyott, Alex Erwin and Jeremy Forsythe as they went above and beyond to support me in every way they could, from idea and food sharing to encouragement when things became tough. They and the rest of the graduate students have made my experience at the University of Kansas one that I will never forget.

Table of Contents

<u>Introduction</u>	1
<u>Chapter 1</u>	2
Introduction	3
Methods	6
Results and Discussion	10
Conclusion	20
References	21
<u>Chapter 2</u>	26
Introduction	27
Methods	30
Results and Discussion	33
Conclusion and Further Directions	44
References	47

Introduction

Many evolutionary biologists are driven to understand the overwhelming biodiversity on Earth. This diversity can be seen through the wide array of organisms that populate the planet, down to the amazing range of cell types, such as sperm. This thesis is a direct reflection of that need to understand diversity through novel adaptations in Lepidoptera, the moths and butterflies.

The order Lepidoptera displays a vast amount of diversity among species through wing coloration, habitat preference, flight patterns and even food sources, and among other insect orders through wing orientation, life cycle differences and sperm cell types. This diversity has always made Lepidoptera an interesting candidate for organismal studies. Recently, with the increased genomic tools as well as many published Lepidoptera genomes, Lepidoptera are becoming model organisms to use for molecular work as well.

In the following chapters, novel traits of Lepidoptera are explored through the use of shotgun proteomics. Shotgun proteomics is a methodology to identify proteins in a complex biological sample all at once. The first chapter explores the novel ability of the genus *Heliconius* to digest pollen. This project takes an in depth look at the proteins present in *Heliconius melpomene* saliva to begin to understand which proteins may play a role in the protein digestion of pollen granules. Chapter two explores the phenomenon of heteromorphic sperm, which is present in almost all Lepidoptera. This research delves into understanding the two sperm types and the reason for which they may exist.

Chapter 1:

Pollen feeding proteomics: salivary proteins of the passion flower butterfly, *Heliconius melpomene*

Introduction

Most adult Lepidoptera use flower nectar as their primary food source. Nectar is typically rich in water and carbohydrates but quite limited as a source of amino acids (H. G. Baker, 1975; H. G. Baker and I. Baker, 1977; 1973). Consequently, most Lepidopteran species primarily acquire nutritional protein as larvae feeding on leafy plant material, storing nitrogen and essential amino acids for use during pupation and adulthood (Dunlap-Pianka et al., 1977). Intriguingly, a striking exception to this general pattern is found among butterflies in the genus *Heliconius*, the passion flower butterflies. In addition to nectar feeding, adult *Heliconius* butterflies feed on pollen, a trait with a single origin in this genus (Beltran et al., 2007; Brown, 1981; Gilbert, 1972). Pollen has high nitrogen and essential amino acid content, providing *Heliconius* butterflies with a substantial source of nutritional resources typically thought to constrain adult lepidopteran reproduction and longevity (Dunlap-Pianka et al., 1977; Gilbert, 1972; O'Brien et al., 2003). Accordingly, *Heliconius* butterflies are unusually long-lived, with adult life-spans known to last beyond six months (Gilbert, 1972). Females lay eggs at a moderate and continuous rate throughout adulthood without the reproductive or ovarian senescence characteristic of related butterflies. Carbon isotope analysis has demonstrated that essential amino acids from pollen are directly incorporated into eggs, and excluding pollen from adult *Heliconius* results in dramatic reductions of life-span and fecundity (Dunlap-Pianka et al., 1977; O'Brien et al., 2003). Thus pollen feeding clearly represents a remarkable evolutionary innovation that catalyzed dramatic changes in the physiology and life-history of *Heliconius* butterflies. However, many aspects of this adaptation remain enigmatic and in particular it remains unclear how amino acids are captured from the pollen.

Heliconius butterflies do not directly ingest pollen grains. Rather, pollen is collected and stored on the outside of the proboscis (Fig. 1), which has an array of unusually dense and long sensory bristles which presumably facilitate pollen collection and retention (Krenn and Penz, 1998). A suite of behavioral adaptations are also associated with pollen feeding, including sophisticated flower handling and a stereotypical coiling-uncoiling of the proboscis that agitates the collected pollen load (Krenn, 2008; Krenn et al., 2009; Penz and Krenn, 2000). During this pollen processing, saliva is exuded from the proboscis into the pollen and ingested some time later, presumably transporting free amino acids back into the butterfly's digestive tract.

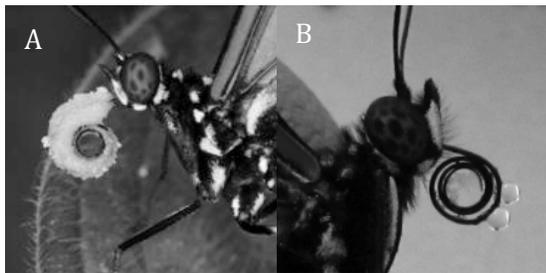


Figure 1: A) *Heliconius* butterfly with a large load of pollen on the proboscis. B) Saliva droplets exuded onto proboscis after stimulation with microscopic glass beads during saliva collection

There has been considerable uncertainty regarding the exact mechanism by which amino acids are released from the pollen grains. Early Hypotheses favored a “passive” process. In the initial description of *Heliconius* pollen feeding, Gilbert (1972) suggested that germination of pollen

when moistened on the proboscis was sufficient to release free amino acids (Gilbert, 1972). Later Erhardt & Baker (1990) proposed a diffusion process. However, a recent demonstration that proboscis coiling-uncoiling causes substantial mechanical disruption of pollen grains undermines these “passive” hypotheses, indicating instead that *Heliconius* butterflies actively degrade their pollen (Krenn et al., 2009). Additionally, colorimetric assays of proteolytic activity clearly show *Heliconius* saliva contains proteases that likely degrade pollen enzymatically to complement mechanical disruption (Eberhard et al., 2007). Thus the behavior of pollen processing in saliva acts as an extra-oral digestion (Krenn et al., 2009), but the proteins involved in this process remain unknown.

Here we report an initial investigation into the molecular components of pollen feeding. Using liquid chromatography mass spectrometry (LC-MS) “shotgun” proteomics, we analyzed the protein content of saliva from *Heliconius melpomene*. We confidently identified more than thirty proteins from *Heliconius* saliva, including several putatively secreted proteins with predicted proteolytic function. Also prevalent were proteins predicted to function in carbohydrate hydrolysis and immunity. These results lay the foundation for future investigations into the molecular origins and mechanisms of *Heliconius* pollen feeding.

Methods

Butterfly care, saliva collection and preparation

Heliconius melpomene aglaope were purchased as pupae from commercial providers (Stratford Butterfly Farms, Stratford-Upon-Avon, Warwickshire, UK) and reared in a temperature and humidity controlled greenhouse at the University of Cambridge's Madingley Field Station, Madingley, UK. Butterflies were kept in cages 1.5 m tall, 1.5 m wide, by 1m deep and provisioned with artificial nectar consisting of 10% sucrose solution in water augmented with 5 g/L Critical Care Formula (Vetark Professional, Winchester UK). In order to minimize contamination of saliva samples with food or pollen proteins, the butterflies were not provided with plants or another pollen source. Additionally, for at least 36 hours before sampling, the Critical Care Formula supplement was removed from the artificial nectar.

Saliva samples were collected by applying a small amount of water-moistened glass beads ($\leq 106 \mu\text{M}$, Sigma-Aldrich, St Louis, MO, USA) to the proboscis with an insect pin and then washing the proboscis and beads into a 1.5 μL microcentrifuge tube using a pipettor. Typically the application of beads or even just the manipulation of the proboscis with a pin caused visible droplets of saliva to be exuded from the proboscis, usually from the outer edge proximal to the head (Fig 1). The same 150 μL of deionized water was used repeatedly to rinse saliva and beads from the proboscis of 8-10 butterflies per round of collection. Two rounds of collection were performed in one day, separated by 1.5 h, using the same 150 μL diH₂O. Sampling on two different days provided a pair of biological replicates for proteomic analysis.

Each of the two 150 μ L samples was vacuum-centrifuged at 60°C to reduce volume to 50 μ L. 20 μ L per sample was kept for polyacrylamide gel electrophoresis, and the remaining 30 μ L was submitted for direct shotgun proteomic analysis via LC-MS.

Protein gel electrophoresis

For polyacrylamide gel electrophoresis, 2.6 vol sample were mixed with 1 vol 4 \times NuPAGE LDS Sample Buffer (Invitrogen) and 0.4 vol 10 \times NuPAGE Reducing Agent (0.5 M dithiothreitol; Invitrogen). The samples were heated to 70°C for 10 min, loaded on 4–12% NuPAGE Bis-Tris 1.0mm precast gels (Invitrogen) and electrophoresed in NuPAGE MOPS running buffer at 4 mA/gel for about 100 min. Gels were then fixed and silver stained using standard methods, followed by imaging on a flat-bed scanner.

Mass spectrometry and analysis

Each biological replicate was split into two technical replicates, so a total of four LC-MS experiments were performed. Samples were digested and analyzed *in toto*, one experiment per replicate, without prior gel fractionation. Samples submitted for LC-MS analyses were dried down and resolubilised in 20 mL of 50 mM ammonium bicarbonate. Proteins were then reduced (5 mM DTT) and alkylated (15mM iodoacetamide) before being digested overnight with trypsin. The samples were then dried and resuspended in 20 mL 0.1% formic acid and pipetted into a sample vial and placed in the LC autosampler.

All LC-MS experiments were performed using a nanoAcquity UPLC (Waters Corp., Milford, MA) system and an LTQ Orbitrap Velos hybrid ion trap mass spectrometer (Thermo Scientific, Waltham, MA). Separation of peptides was performed by reverse-phase

chromatography using at a flow rate of 300 nL/min and a Waters reverse-phase nano column (BEH C18, 75 mm i.d. x 250 mm, 1.7 mm particle size). Peptides were loaded onto a pre-column (Waters UPLC Trap Symmetry C18, 180 mm i.d x 20mm, 5 mm particle size) from the nanoAcquity sample manager with 0.1% formic acid for 3 minutes at a flow rate of 10 mL/min. After this period, the column valve was switched to allow elution of peptides from the pre-column onto the analytical column. Solvent A was water + 0.1% formic acid and solvent B was acetonitrile + 0.1% formic acid. The linear gradient employed was 5-50% B in 60 minutes.

The LC eluant was sprayed into the mass spectrometer by means of a New Objective nanospray source. All m/z values of eluting ions were measured in an Orbitrap Velos mass analyzer, set at a resolution of 30000. Data dependent scans (Top 20) were employed to automatically isolate and generate fragment ions by collision-induced dissociation in the linear ion trap, resulting in the generation of MS/MS spectra. Ions with charge states of 2+ and above were selected for fragmentation. Post-run, the data was processed using Protein Discoverer (version 1.2, ThermoFisher) and converted to mascot generic format (.mgf) files for subsequent database searching.

Mass spectra analysis

MS/MS spectra were searched against the *H. melpomene* predicted protein set (downloaded from butterflygenome.org, last updated June 4, 2012) using the Mascot search engine (Perkins et al., 1999) . The search parameters were as follows: digestive enzyme- trypsin, maximum missed cleaves- 2, fixed modifications- carbamidomethyl, variable modifications- oxidation (M), peptide mass tolerance- 25 ppm, fragment mass tolerance- .8 Da, mass values- monoisotopic, instrument type- ESI-TRAP. The *cRAP* database (via The Global Proteome Machine, www.thegpm.org), last updated February 29, 2012, was also included to search for

contaminants in the samples. A false discovery rate (FDR) was calculated by simultaneously searching spectra against a decoy database created by reversing the sequences of the *H. melpomene* protein set. Proteins were identified using peptide and protein identifications validated through Scaffold 4.0 (Searle, 2010). Peptide threshold was established at 90% and protein threshold at 95%, using the Peptide Prophet algorithm and Protein Prophet algorithms respectively, with at least two unique peptide matches required in each sample. Protein and peptide FDR were 0% to ensure high confidence in identifications. Relative abundances of proteins were estimated as the mean of normalized spectral counts, as calculated by the Scaffold software.

Functional predictions

Proteins identified via LC-MS were functionally annotated bioinformatically using sequence homology. Proteins were searched against the NCBI non-redundant protein database using BLASTP (Altschul et al., 1990). Proteins were also submitted to InterproScan (Zdobnov and Apweiler, 2001). For each protein identified, putative function was manually assigned after reviewing and integrating bioinformatic search results.

Results and Discussion

SDS-PAGE

Protein electrophoresis revealed a relatively sparse collection of proteins present in the

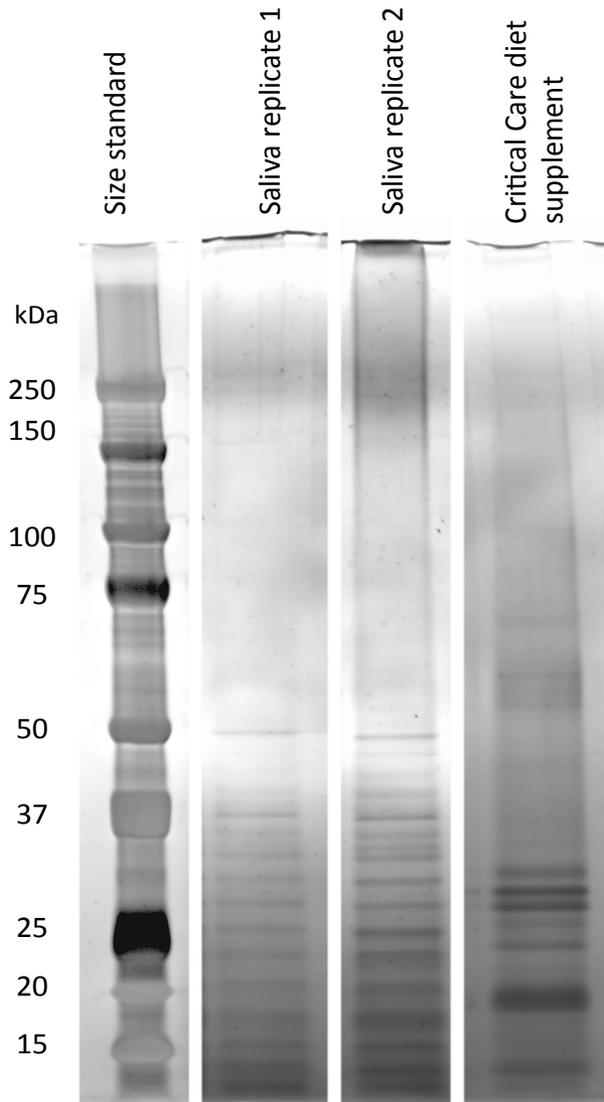


Figure 2: PAGE analysis of *H. melpomene* saliva and Critical Care Formula diet 605 supplement. Size standard is in kiloDaltons(kDa).

saliva (Fig. 2). Only about 20 distinct bands were visible in the saliva sample. Notably, none of the bands were concordant with bands observed in the dietary supplement, indicating that the saliva was not contaminated with Critical Care Formula diet supplement.

Shotgun Proteomics

After filtering the protein hits by significance using Scaffold and removing all contaminant protein hits, a total of 31 proteins were confidently identified from *H. melpomene* adult saliva. Results are summarized in Table 1. There was substantial consistency between biological replicates, with 24 proteins (77%) identified in both samples. Technical replication was also reasonably consistent, with 22 proteins

(70%) identified in all four replicates. We also identified and discarded a few obvious contaminant proteins in the filtered LC-MS results (e.g. human keratin, pig trypsin).

One clear prediction about salivary proteins is that they are secreted extracellularly and therefore should contain a signal peptide at the N-terminus (Scheele et al., 1978). As expected, signal peptides predicted by Signal-P (via InterproScan) were found in 20 of the salivary proteins (Petersen et al., 2011). This is probably an underestimate because four of 11 proteins without predicted signal peptides were represented by problematic gene models that lacked start codons. Missing start codons likely reflects errors in the underlying genome assembly on which gene models were built because our manual inspection could not identify obvious start codons. Otherwise, “complete” proteins without signal peptides tended to have “housekeeping” functions and are likely to be *Heliconius*-derived contaminants rather than true salivary proteins (see section below on “*housekeeping*” proteins).

The identified proteins could be divided into four groups based on function: proteolysis, carbohydrate hydrolysis, immunity, and “housekeeping”. Additionally, several proteins could not be functionally annotated and were lumped into a fifth group of proteins with unknown function.

Proteolytic proteins

Ten identified proteins were found to play a role in proteolysis, encompassing a range of functions including protein degradation, cleaving small peptide bonds, and proteolytic inhibition. The seven proteases are primary candidates for playing a role in the digestion of pollen granules. These include serine proteases, cysteine proteases, astacins, and a carboxypeptidase. All three serine proteases appear to have trypsin-like or chymotrypsin-like properties based on BLAST-based homology and protein domain predictions. Intriguingly, both HMEL006217-PA and HMEL017107-PA show close similarity (i.e. strong BLAST hits) to the Cocoonase protein from *Bombyx mori* (silkworm). Cocoonase is a well-characterized trypsin-like protease secreted by the proboscis during eclosion to weaken the cocoon silk and facilitate emergence (Kafatos et al.,

1967; Yamamoto et al., 1999). The function of Cocoonase homologs in butterflies, which lack silken cocoons, remains unknown. In the case of *Heliconius* it is tempting to speculate that these proteases, which presumably have an evolutionary history of expression in the proboscis, were evolutionarily co-opted to function in pollen digestion.

Protein	Putative Function	Average Percent Coverage(A)	Unique Spectra(B)	Signal Peptide(C)	Biological Replication(D)	Technical Replication(E)	Mean Spectrum Count(F)
Proteolysis							
HMEL003607	trypsin-like protease	4.925	2	Absent	1	2	2.52
HMEL006217	trypsin-like protease	22.5	19	Uncertain	2	4	50.04
HMEL007847	carboxypeptidase	6.55	5	Absent	2	4	11.65
HMEL010316	astacin	9.9	6	Present	2	4	10.65
HMEL015078	astacin	9.8	7	Present	2	4	12.45
HMEL013718	trypsin inhibitor	18.75	3	Present	2	3	3.88
HMEL017107	trypsin-like protease	15	5	Absent	2	4	13.55
HMEL002374	cysteine protease inhibitor	22	3	Present	2	4	5.60
HMEL007577	cysteine protease inhibitor	32.25	4	Present	2	4	10.76
HMEL014517	cysteine protease	22.5	5	Present	2	4	18.96
Carbohydrate hydrolysis							
HMEL005611	β -fructofuranosidase	23.75	16	Present	2	4	31.22
HMEL005612	β -fructofuranosidase	10	9	Present	2	4	14.82
HMEL011728	glycerophosphoryl diester phosphodiesterase	0.625	2	Present	1	1	1.18
HMEL014479	hydrolase	6.5	9	Uncertain	2	4	16.20
HMEL014593	β -hexosaminidase	1.55	2	Absent	1	1	1.20
Immunity							
HMEL005769	REPAT gene	39.5	5	Present	2	4	9.24
HMEL013458	hemolin	1.2	2	Present	1	1	1.62
HMEL010053	GMC oxidoreductase	2.9	3	Uncertain	2	4	2.98
HMEL010482	alpha crystalline / HSP20	31.5	8	Present	2	4	19.79
HMEL002661	lysozyme	21	4	Present	2	4	6.23
HMEL016918	β 1,3 glucanase	2.325	2	Present	1	2	1.38
Housekeeping or other function							
HMEL002092	yellow-d	3.55	2	Present	1	2	1.60
HMEL010248	Splicing factor	1.53	2	Absent	2	2	1.33

HMEL013620	actin	5.2	3	Absent	2	4	3.80
HMEL015393	CAP domain	37	3	Uncertain	2	4	2.98
Unknown Function							
HMEL008913	Unknown	38.5	4	Present	2	4	7.20
HMEL008915	Unknown	24	4	Present	2	4	79.49
HMEL014907	Unknown	7.5	2	Present	1	2	2.53
HMEL015039	Unknown	17	3	Present	2	4	44.52
HMEL015041	Unknown	15	2	Present	2	4	6.05
HMEL010245	Unknown	7.025	3	Present	2	4	3.77

Table 1 Functional and Bioinformatic Results: A) Percent of protein covered by matched peptides. . B) Total count of unique spectra found for each protein identification. C) “Uncertain” indicates an incomplete gene model prevents informative predictions. D) Number of biological replicates (out of two) in which the protein was present E) Number of technical replicates (out of four) in which the protein was present F) Mean of normalized spectral counts found for each protein, indicating the protein’s relative abundance in the sample.

Carboxypeptidases hydrolyze peptide bonds at the carboxy-terminal end of a peptide or protein and are also known for their digestive roles (Bown and Gatehouse, 2004). Similarly, astacins often play an important role in extracellular protein digestion (Foradori et al., 2006). Thus this suite of secreted proteases together potentially provides a rich cocktail for breaking down pollen proteins and releasing free amino acids for consumption.

The cysteine and trypsin inhibitors inactivate cysteine and serine proteases, respectively, by binding to the protein’s active site and rendering it inactive (Eguchi, 1993). The two cysteine protease inhibitors identified here appear to be related to the well-characterized *Bombyx* Cysteine Protein Inhibitor (BCPI) (Yamamoto et al., 1999). BCPI-like proteins likely originated from the inhibitory propeptide region of a cysteine proteinase that is typically cleaved to release the proteolytic function of the mature peptide. These BCPI-like proteins function as “stand alone” inhibitors of cathepsin-L type cysteine proteases (Kurata et al., 2001). Another such protein was proteomically identified as a constituent of seminal fluid in *Heliconius erato*; the putative *H.*

melpomene ortholog of this seminal protein is clearly distinct from these two salivary cysteine protease inhibitors, sharing only ~70% amino acid identity with either (Wallow and Harrison, 2010). It thus appears that these propeptide-derived cysteine protease inhibitors are commonly deployed for extra-cellular regulation of proteolysis in *Heliconius* butterflies. Nonetheless, it is difficult to predict what role, if any, these cysteine and trypsin protease inhibitors play in pollen digestion. One plausible alternative function is in pathogen defense. Many insect protease inhibitors are known to target pathogen-derived proteases or are upregulated after pathogen exposure, presumably providing defense against infection (Kanost, 1999; Rai et al., 2010; Zhao et al., 2012). An immunity-related function of salivary protease inhibitors would be consistent with our observing several other immunity-related salivary proteins (see below).

A distinct lack of molecular characterization of other butterfly saliva proteomes leads to difficulty in making comparisons across pollen and non-pollen feeding Lepidoptera. However, a study performed by (Feng et al., 2013) gave insight into the honeybee saliva proteome. Honeybees are another insect that consumes both pollen and nectar, presenting interesting parallels to *Heliconius*. Honeybees have a mostly carbohydrate rich diet (nectar), which is reflected in the proteins found in their proteome. Both proteomes contain proteins relating to both proteolytic activity and carbohydrate hydrolysis, but *Heliconius* appears to have relatively more proteins related to proteolytic activity and fewer to carbohydrate hydrolysis.

Carbohydrate hydrolysis

Five proteins identified in *H. melpomene* saliva are predicted to be varieties of glycoside hydrolases that appear to play a role in carbohydrate hydrolysis (Withers, 2001). The two β -fructofuranosidases function in breaking down sucrose into fructose and glucose by cleaving the

O-C bond. Until recently, β -fructofuranosidases were thought to be absent from animals despite being found among bacteria, fungi, and plants. However, pairs of these proteins have been identified in several lepidopteran species, apparently having arisen via horizontal transfer from bacteria (Daimon et al., 2008). Previously, these β -fructofuranosidases have primarily been associated with larval gut, so their presence in adult saliva is consistent with a role in digestion but also marks a distinct expansion of their known functional milieu.

The remaining three glycoside hydrolases (glycerophosphodiester phosphodiesterase, β -hexosaminidase, and hydrolase) all appear to have relatively general functions in sugar metabolism. This is not unexpected given that *Heliconius* butterflies consume substantial quantities of sugar-rich plant nectar along with pollen.

Immune function

Another six *H. melpomene* salivary proteins likely play a role in immune response. Two of these, lysozyme and β -1,3 glucanase, are glycoside hydrolases that have secondarily evolved to function in immune response (Davis and Weiser, 2011). Lysozymes are common antimicrobial proteins that function to degrade bacterial cell walls; they are well known components of insect immune responses, including in Lepidoptera (Callewaert and Michiels, 2010; Jiang et al., 2010). Proteins that bind β -1,3glucan function as pathogen recognition proteins that tend to target gram-negative bacteria. Several such proteins have been identified in moths and butterflies (Fabrick et al., 2004). These proteins are usually isolated from hemolymph, but have also been found in the saliva and digestive tracts of other insects (Pauchet et al., 2009).

REPAT and hemolin are Lepidopteran specific immune proteins that have shown increased expression in response to pathogen infection in caterpillars of several species (Hernández-

Rodríguez et al., 2009; Terenius et al., 2009; Yamamoto et al., 1999). Also implicated in insect immune response are heat shock proteins, such as alpha crystalline, that are important in keeping essential proteins from unfolding (Pirkkala et al., 2001). Hsp20/alpha crystalline has been found in the salivary glands of other insects and is known to regulate proteins when the organism's temperature exceeds 25°C (Arrigo and Ahmadzadeh, 1981). Finally, we have tentatively assigned an immunity-related function to the one identified salivary glucose-methanol-choline (GMC) oxidoreductase gene. GMC oxidoreductases comprise a large and diverse protein family whose members play a variety of often poorly understood roles in developmental processes, glucose metabolism, and immune function (Iida et al., 2007). In Lepidoptera this protein family is particularly diverse and many members seem to play a role in immune response (Sun et al., 2012). Thus we have grouped this protein with other immunity-related proteins, but much additional research would be necessary to confidently characterize the true function of this particular GMC oxidoreductase.

Housekeeping and other functions

Proteins functioning in proteolysis, sugar metabolism, and immunity are reasonably expected to be found in saliva. We additionally identified in our samples several proteins that seemingly have little relevance to expected salivary functions, or are generally of ambiguous function. Foremost among these is actin, known for its role in muscle contraction and cytoskeletal structure generally, but not expected to function outside of cells (Dominguez and Holmes, 2011). Actin is a ubiquitous and highly abundant protein, so may easily have contaminated the saliva samples. Similarly, an identified serine-arginine-rich splicing factor protein typically functions in RNA splicing and gene expression and is probably best considered a contaminant (Long and Caceres, 2009).

Somewhat more ambiguous is the presence of yellow-d, a member of the *yellow* protein family. The function of Yellow proteins is poorly understood, though clearly some members play a role in melanization (Drapeau, 2001; Ferguson et al., 2010). In *B. mori*, yellow-d appears to be ubiquitously expressed and also contains a predicted signal peptide (Xia et al., 2006). The annotation of the yellow-d gene model from the *H. melpomene* genome did not indicate the presence of a signal peptide. However, comparison with a sequence generated from ESTs (GenBank accession ADX87351) clearly indicates that the genome-based model is truncated and that *H. melpomene* yellow-d does contain a signal peptide. Thus, while the molecular function of this and other yellow proteins remains largely unknown, it seems reasonable to consider yellow-d as normally present in *H. melpomene* saliva.

The Cysteine-rich secretory proteins, antigen 5 and pathogenesis related (CAP) proteins are taxonomically diverse with an equally diverse set of functions, making it difficult to predict any particular function for this one protein found in *H. melpomene* saliva (Gibbs et al., 2008). CAP proteins are typically secreted extracellularly, but in the case of this one salivary CAP, the predicted gene model was incomplete at the N-terminus and therefore uninformative regarding the presence of a signal peptide.

Unknown function

Finally, six proteins found in the sample could not be functionally characterized at any level, other than all of them exhibiting a predicted signal peptide. One of these, HMEL010245-PA, showed extensive homology to similar proteins present in many other insect species, though none of these were functionally annotated. The remaining five proteins appear to be extremely taxonomically restricted. HMEL015039-PA and HMEL015041-PA are a pair of closely linked paralogs, separated by ~8Kbp, suggesting they arose via tandem duplication. Strikingly, a variety

of BLAST strategies have yielded no significant homology (e-val < 0.01) to any other protein or nucleotide sequences.

The remaining three uncharacterized proteins, HMEL008913-PA, HMEL008915-PA, and HMEL014907-PA, are another set of paralogs. The similarity and apparent tandem duplication of HMEL008913-PA and HMEL008915-PA suggest HMEL014907-PA is the most distantly related of the three paralogs. In this case, the only clearly homologous loci identified were a pair of paralogs from the monarch butterfly, KGM_02914 & KGM_02913, that also appear to be tandemly duplicated. Otherwise these proteins lacked both Blast and InterproScan hits, although each had a signal peptide. These groups of Nymphalid-specific, perhaps even *Heliconius*-specific, secreted proteins in the saliva are very intriguing in light of *Heliconius* pollen feeding.

Conclusions and future directions

The results presented here offer a first glimpse into the molecular foundation of *Heliconius* pollen feeding and provide a substantial advance towards comprehensively understanding this striking evolutionary novelty. The observation of several proteolytic enzymes supports the emerging view that *Heliconius* butterflies actively degrade pollen and consume released amino acids via extra-oral digestion (Krenn et al., 2009). These results also highlight the importance of salivary digestion of sugars for nectar-feeding insects as well as oral ingestion of pathogens as a common infection route that is actively defended via immune-related proteins in the saliva.

Our results open several different avenues for productive future research. One route for better understanding the molecular basis of *Heliconius* pollen feeding would be experimental characterization of the several proteins with ambiguous or unknown function via cloning and *in vitro* expression or targeted knock-outs (e.g. CRISPR; Sander and Joung, 2014). Complementing this, comparative proteomic analysis would inform the evolutionary history of this adaptation. Specifically, contrasting the salivary protein content of related taxa that do not pollen feed would highlight unique *Heliconius* salivary proteins that are most likely to reflect molecular adaptations to pollen feeding. If whole genome assemblies become available for related species that exclusively feed on pollen, then comparative genomics can reveal the relative importance of genetic novelty versus cooption and redeployment of existing genes in the evolution of pollen feeding. Broadly speaking, our results presented here demonstrate that proteomic and genomic analysis of *Heliconius* pollen feeding hold great potential for researching the molecular genetic basis of a complex physiological adaptation.

References

- Arrigo, A.P., & Ahmadzadeh, C., 1981. Immunofluorescence localization of a small heat shock protein (hsp-23) in salivary-gland cells of *Drosophila melanogaster*. *Molecular & General Genetics*, 184(1), 73-79.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Baker, H.G., 1975. Sugar Concentrations in Nectars from Hummingbird Flowers. *Biotropica* 7, 37–41.
- Baker, H.G., Baker, I., 1973. Amino-Acids in Nectar and Their Evolutionary Significance. *Nature* 241, 543–545.
- Baker, H.G., Baker, I., 1977. *Coevolution of animals and plants*. Austin: University of Texas Press.
- Beltran, M., Jiggins, C.D., Brower, A.V., Bermingham, E., Mallet, J., 2007. Do pollen feeding, pupal-mating and larval gregariousness have a single origin in *Heliconius* butterflies? Inferences from multilocus DNA sequence data. *Biological Journal of the Linnean Society* 92, 221–239.
- Bown, D.P., Gatehouse, J.A., 2004. Characterization of a digestive carboxypeptidase from the insect pest corn earworm (*Helicoverpa armigera*) with novel specificity towards C-terminal glutamate residues. *Eur J Biochem* 271, 2000–2011.
- Brown, K.S., Jr, 1981. The biology of *Heliconius* and related genera. *Annu. Rev. Entomol.* 26, 427–457.
- Callewaert, L., Michiels, C.W., 2010. Lysozymes in the animal kingdom. *Journal of Biosciences* 35, 127–160.
- Daimon, T., Taguchi, T., Meng, Y., Katsuma, S., Mita, K., Shimada, T., 2008. - Fructofuranosidase Genes of the Silkworm, *Bombyx mori*: Insights into enzymatic adaptation of *B. Mori* to toxic alkaloids in mulberry latex. *Journal of Biological Chemistry* 283, 15271–15279.
- Davis, K.M., Weiser, J.N., 2011. Modifications to the Peptidoglycan Backbone Help Bacteria To Establish Infection. *Infection and Immunity* 79, 562–570.
- Dominguez, R., Holmes, K.C., 2011. Actin Structure and Function. *Annu. Rev. Biophys.* 40, 169–186.

- Drapeau, M.D., 2001. The Family of Yellow-Related *Drosophila melanogaster* Proteins. *Biochemical and Biophysical Research Communications* 281, 611–613.
- Dunlap-Pianka, H., Boggs, C.L., Gilbert, L.E., 1977. Ovarian Dynamics in Heliconiine Butterflies: Programmed Senescence versus Eternal Youth. *Science* 197, 487–490.
- Eberhard, S.H., Hrassnigg, N., Crailsheim, K., Krenn, H.W., 2007. Evidence of protease in the saliva of the butterfly *Heliconius melpomene* (L.) (Nymphalidae, Lepidoptera). *Journal of Insect Physiology* 53, 126–131.
- Eguchi, M., 1993. Protein Protease Inhibitors in Insects and Comparison with Mammalian Inhibitors. *Comp. Biochem. Physiol., B* 105, 449–456.
- Fabrick, J.A., Baker, J.E., Kanost, M.R., 2004. Innate Immunity in a Pyralid Moth: Functional Evaluation of Domains from a -1,3-Glucan Recognition Protein. *Journal of Biological Chemistry* 279, 26605–26611.
- Feng, M., Fang, Y., Bin Han, Zhang, L., Lu, X., Li, J., 2013. Novel aspects of understanding molecular working mechanisms of salivary glands of worker honeybees (*Apis mellifera*) investigated by proteomics and phosphoproteomics. *Journal of Proteomics* 87, 1–15.
- Ferguson, L.C., Green, J., Surridge, A., Jiggins, C.D., 2010. Evolution of the Insect Yellow Gene Family. *Molecular Biology and Evolution* 28, 257–272.
- Foradori, M.J., Tillinghast, E.K., Smith, J.S., Townley, M.A., Mooney, R.E., 2006. Astacin family metallopeptidases and serine peptidase inhibitors in spider digestive fluid. *Comp. Biochem. Physiol. B, Biochem. Mol. Biol.* 143, 257–268.
- Gibbs, G.M., Gibbs, G.M., Gibbs, G.M., Gibbs, G.M., Roelants, K., Roelants, K., Roelants, K., Roelants, K., O'Bryan, M.K., O'Bryan, M.K., O'Bryan, M.K., O'Bryan, M.K., 2008. The CAP Superfamily: Cysteine-Rich Secretory Proteins, Antigen 5, and Pathogenesis-Related 1 Proteins—Roles in Reproduction, Cancer, and Immune Defense. *Endocrine Reviews* 29, 865–897.
- Gilbert, L.E., 1972. Pollen Feeding and Reproductive Biology of *Heliconius* Butterflies. *Proc. Natl. Acad. Sci. U.S.A.* 69, 1403.
- Hernández-Rodríguez, C.S., Ferré, J., Herrero, S., 2009. Genomic structure and promoter analysis of pathogen-induced repeat genes from *Spodoptera exigua*. *Insect Molecular Biology* 18, 77–85.
- Iida, K., Cox-Foster, D.L., Yang, X., Ko, W.-Y., Cavener, D.R., 2007. Expansion and evolution of insect GMC oxidoreductases. *BMC Evolutionary Biology* 7, 75.

- Jiang, H., Vilcinskas, A., Kanost, M.R., 2010. Immunity in Lepidopteran Insects. *Invertebrate Immunity* 708, 181–204.
- Kafatos, F.C., Tartakoff, A.M., Law, J.H., 1967. Cocoonase I. Preliminary characterization of a proteolytic enzyme from silk moths. *Journal of Biological Chemistry* 242, 1477–1487.
- Kanost, M.R., 1999. Serine proteinase inhibitors in arthropod immunity. *Developmental & Comparative Immunology* 23, 291–301.
- Krenn, H.W., 2008. Feeding behaviours of neotropical butterflies (Lepidoptera, Papilionoidea). *Denisia, zugleich Kataloge der oberösterreichischen Landesmuseen Neue Serie* 88, 295–304.
- Krenn, H.W., Eberhard, M.J.B., Eberhard, S.H., Hinkl, A.-L., Huber, W., Gilbert, L.E., 2009. Mechanical damage to pollen aids nutrient acquisition in *Heliconius* butterflies (Nymphalidae). *Arthropod-Plant Interactions* 3, 203–208.
- Krenn, H.W., Penz, C.M., 1998. Mouthparts of *Heliconius* butterflies (Lepidoptera: Nymphalidae): a search for anatomical adaptations to pollen-feeding behavior. *International Journal of Insect Morphology and Embryology* 27, 301–309.
- Kurata, M., Yamamoto, Y., Watabe, S., Makino, Y., Ogawa, K., Takahashi, S.Y., 2001. *Bombyx* cysteine proteinase inhibitor (BCPI) homologous to propeptide regions of cysteine proteinases is a strong, selective inhibitor of cathepsin L-like cysteine proteinases. *Journal of biochemistry* 130, 857–863.
- Long, J.C., Caceres, J.F., 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem. J.* 417, 15.
- O'Brien, D.M., Boggs, C.L., Fogel, M.L., 2003. Pollen feeding in the butterfly *Heliconius charitonia*: isotopic evidence for essential amino acid transfer from pollen to eggs. *Proceedings of the Royal Society B: Biological Sciences* 270, 2631–2636.
- Pauchet, Y., Freitak, D., Heidel-Fischer, H.M., Heckel, D.G., Vogel, H., 2009. Immunity or Digestion: Glucanase Activity in a Glucan-Binding Protein Family from Lepidoptera. *Journal of Biological Chemistry* 284, 2214–2224.
- Penz, C.M., Krenn, H.W., 2000. Behavioral adaptations to pollen-feeding in *Heliconius* butterflies (Nymphalidae, Heliconiinae): an experiment using Lantana flowers. *Journal of Insect Behavior* 13, 865–880.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.
- Petersen, T.N., Brunak, S., Heijne, von, G., Nielsen, H., 2011. SignalP 4.0: discriminating signal

- peptides from transmembrane regions. *Nature Publishing Group* 8, 785–786.
- Pirkkala, L., Nykänen, P., Sistonen, L., 2001. Roles of the heat shock transcription factors in regulation of the heat shock response and beyond. *FASEB J.* 15, 1118–1131.
- Rai, S., Aggarwal, K.K., Mitra, B., Das, T.K., Babu, C.R., 2010. Purification, characterization and immunolocalization of a novel protease inhibitor from hemolymph of tasar silkworm, *Antheraea mylitta*.
- Sander, J.D., Joung, J.K., 2014. Crispr-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology* 32, 347–355.
- Scheele, G., Dobberstein, B., Blobel, G., 1978. Transfer of proteins across membranes. *Eur J Biochem* 82, 593–599.
- Searle, B.C., 2010. Scaffold: A bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 10, 1265–1269.
- Sun, W., Shen, Y.-H., Yang, W.-J., Cao, Y.-F., Xiang, Z.-H., Zhang, Z., 2012. Insect Biochemistry and Molecular Biology. *Insect Biochemistry and Molecular Biology* 42, 935–945.
- Terenius, O., Popham, H.J.R., Shelby, K.S., 2009. Bacterial, but not baculoviral infections stimulate Hemolin expression in noctuid moths. *Developmental & Comparative Immunology* 33, 1176–1185.
- Wallow, J.G., Harrison, R.G., 2010. Combined EST and Proteomic Analysis Identifies Rapidly Evolving Seminal Fluid Proteins in *Heliconius* Butterflies. *Molecular Biology and Evolution* 27, 2000–2013.
- Withers, S.G., 2001. Mechanisms of glycosyl transferases and hydrolases. *Carbohydrate Polymers* 44, 325–337.
- Xia, A.-H., Zhou, Q.-X., Yu, L.-L., Li, W.-G., Yi, Y.-Z., Zhang, Y.-Z., Zhang, Z.-F., 2006. Identification and analysis of YELLOW protein family genes in the silkworm, *Bombyx mori*. *BMC Genomics* 7, 195.
- Yamamoto, Y., Watabe, S., Kageyama, T., Takahashi, S.Y., 1999. Purification and characterization of *Bombyx* cysteine proteinase specific inhibitors from the hemolymph of *Bombyx mori*. *Arch. Insect Biochem. Physiol.* 42, 119–129.
- Zdobnov, E.M., Apweiler, R., 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848.
- Zhao, P., Dong, Z., Duan, J., Wang, G., Wang, L., Li, Y., Xiang, Z., Xia, Q., 2012. Genome-

Wide Identification and Immune Response Analysis of Serine Protease Inhibitor Genes in the Silkworm, *Bombyx mori*. PLoS ONE 7, e31168.

Chapter 2:

Sperm heteromorphy in *Danaus plexippus*: A molecular approach addresses an unsolved paradox

Introduction

A primary goal in evolutionary biology is to understand the functional and selective significance of biological diversity. This includes divergence in behavior, physiology, morphology, and numerous other characteristics that differentiate species. Reproductive traits often exhibit unusually high diversity and rapid evolution, a pattern typically attributed to sexual selection (Swanson and Vacquier, 2002a). The high diversity of different sperm morphologies is one example of a highly diversified reproductive trait. Sperm cell morphology ranges from the classic head-and-tail shape of mammalian sperm to the amoebic sperm of nematodes. This diversity can be counterintuitive, as sperm play a key role in fertilization and disrupting this process through the results of new mutations could result in lower fitness. However, mutations conferring advantages in fertilization should yield a high fitness reward and cause intense sexual selection in the form of sperm competition, the competitive process between sperm from different males to fertilize the eggs of one female (Parker, 1970; Swallow and Wilkinson, 2002).

Although sperm competition may explain much of the observed variation in sperm morphology between species, there are still phenomena that remain unexplained. One example of this is sperm heteromorphy, when the male produces a fertilizing sperm (eusperm) and a morphologically distinct non-fertilizing counterpart (parasperm). In most taxa parasperm have both a nucleus and nuclear DNA, but are recognizable by a difference in shape or size and the inability to fertilize. This phenomenon is found in several phyla, but in no case is it obvious that sperm dimorphism relates to sperm competition. Indeed, the function and origins of parasperm remains a distinct quandary for evolutionary biologists (Swallow and Wilkinson, 2002). Taxa with heteromorphic sperm include a few species of Diptera, Hemiptera, Hymenoptera and Coleoptera, but these species are scattered within their orders with no clear pattern (Swallow and

Wilkinson, 2002). In Lepidopteran (moths and butterflies) sperm heteromorphy is noteworthy for two reasons: 1) the non-fertilizing parasperm lack both nuclear DNA and a nucleus and 2) heteromorphic sperm is common to the entire order (Swallow and Wilkinson, 2002).

Many hypotheses exist to explain the role in sperm heteromorphism in Lepidoptera, but contradicting evidence leaves no definite answers (Cook and Gage, 1995; Swallow and Wilkinson, 2002). These hypotheses can be summarized into two distinct categories: sperm competition or facilitation. Most research has focused on sperm competition, specifically the idea of apyrene sperm being used as cheap filler that delays female remating. Apyrene sperm lack nuclear DNA and a nucleolus, and may be less energetically costly to produce. Males make more apyrene sperm than eupyrene (Swallow and Wilkinson, 2002). Therefore, male butterflies may sparingly use costly eupyrene sperm and instead fill the female's spermatheca with cheap filler sperm. This would in turn cause a delay in remating, as the female believes she has enough viable sperm to fertilize her eggs. In male Indian meal moths (*Plodia interpunctella*) higher amounts of apyrene sperm were transferred to virgin females than to mated females (Cook and Gage, 1995). This finding not only indicates that males can alter their ejaculates, but do so when the advantage of full paternity is higher. In cabbage white butterflies (*Pieris napi*) a positive correlation between a delay in remating time and the amount of apyrene sperm stored was found, suggesting that high quantities of apyrene sperm would lead to higher male fitness (Wedell and Cook, 1999). Both of these findings support evidence that apyrene sperm is being used as a filler to delay remating in female Lepidoptera to ensure higher paternity. Supporting the facilitation hypothesis, Sahara (2003) showed in silkmoths (*Bombyx mori*) that in the absence of apyrene sperm, the nucleated (eupyrene) sperm fertilizes less than .01% of eggs. This suggests that

apyrene sperm are needed for fertilization, though this experiment has never been repeated in any other species and the generality of the result remains unknown.

Despite many descriptive and organismal studies, no clear answer has been found to explain apyrene sperm function. In fact, the previously mentioned studies make understanding apyrene sperm function even more confusing as each study points to a different answer. This project aims to take a different, more molecular approach in hopes to shed more light on the mystery of apyrene sperm. Characterization of the two sperm types at the molecular level is lacking in Lepidoptera. Recent advances in functional genomic technologies and the increasing availability of fully sequenced Lepidopteran genomes now present an excellent opportunity to characterize the molecular composition of Lepidopteran sperm, with an ultimate goal of assessing the differences between the eupyrene and apyrene sperm proteomes. The research presented here represents a foray towards this ultimate goal by functionally characterizing the sperm proteome of the Monarch butterfly (*Danaus plexippus*). *D. plexippus* was chosen because it has a comprehensively assembled and annotated genome, and because samples were an easily accessible resource thanks to Monarch Watch at the University of Kansas.

We used shotgun proteomics to analyze the protein composition of apyrene/eupyrene mixed sperm directly isolated from males (henceforth referred to as the “commixed sample”). We compared this to previously published sperm proteomes from *Manduca sexta* (Tobacco hornworm moth) and also *Drosophila melanogaster* (Wasbrough et al., 2010; Whittington et al., 2015). Finally, we offer a preliminary description of the differences observed in shotgun proteomic analyses of purified apyrene and eupyrene sperm, to offer insight into the role of apyrene sperm through a molecular approach.

Methods

Butterfly care

Lab raised *D. plexippus* were donated to the Walters' Lab as pupae by the University of Kansas's Monarch Watch research station. Butterflies were reared as described in Karr & Walters (2015).

Sperm Extraction/Gel/LC-MSMS

Sperm was first dissected out of male butterflies and analyzed as a single commixed sample and also as isolated apyrene and eupyrene samples according to the methods reported in Karr & Walters (2015). Samples were run on a PAGE gel to separate proteins based on their electrophoretic mobility. Details of PAGE analysis are reported in Karr & Walters. (2015). PAGE gels were partitioned into four slices and each slice was subjected to LC-MS/MS analysis as described in Harpel et al. (2015).

Mass spectra analysis

Proteins in sperm samples were identified using the Mascot search engine (Perkins et al., 1999). MS/MS data were queried against the *D. plexippus* official gene set 2 (OGS2) predicted protein set (downloaded from <http://monarchbase.umassmed.edu>, last updated in 2012). Search parameters and downstream analysis of proteins via Scaffold were as described in Harpel et al. (2015).

Functional predictions

Proteins were functionally annotated and, in some cases, also manually curated to assign function. Sperm protein sequences were queried against the NCBI non-redundant protein database using BLASTP (Altschul et al., 1990) and also submitted to InterproScan (Zdobnov and

Apweiler, 2001). We then used Blast2Go to assign gene ontologies to novel proteins via sequence homology (Conesa et al., 2005).

Analysis of orthology

Putative orthologs were identified using ProteinOrtho (Lechner et al., 2011). While a robust assessment of orthology requires phylogenetic analysis, this is currently impractical for a genome-wide analysis. Nonetheless, for the sake of brevity, we refer to these computationally predicted putative orthologs simply as “orthologs” throughout the remainder of this paper.

We concentrated orthology analyses on the two insects with comparable proteomic analysis of sperm, *M. sexta* and *D. melanogaster*, which allowed a dissection of genetic orthology in the context sperm proteomes (Wasbrough et al., 2010; Whittington et al., 2015). Like *D. plexippus*, *M. sexta* has apyrene and eupyrene sperm; the available sperm proteome is from a commixed sample. *D. melanogaster* has only a single monomorphic sperm type.

We quantified the extent of orthology in the *D. plexippus* sperm proteome relative to the genomic background via bootstrap analysis. Specifically, we compared the proportion of 706 commixed *D. plexippus* sperm proteins with orthologs to a genomic null distribution of proportions generated from 1000 random samples of 706 proteins from the entire *D. plexippus* proteome. This analysis was performed twice, once each for orthology relative to *M. sexta* and *D. melanogaster*.

In a separate analysis, we cross-referenced protein orthology with sperm proteomes in the three species. We calculated the proportion of orthologous proteins found between the commixed sperm sample and *M. sexta* or *D. melanogaster* proteomes. Sperm and non-sperm ortholog proportions were calculated using ProteinOrtho as a predictor for orthologous proteins

Analysis of separated apyrene and eupyrene sperm proteomes

Unique proteins for both the apyrene and eupyrene sperm proteomes were found and functionally annotated using the B2G software. Sperm and non-sperm ortholog proportions were calculated using ProteinOrtho as a predictor for orthologous proteins.

Results and Discussion

The discussion of results will focus initially on the commixed sperm sample. This is directly comparable to the published *M. sexta* commixed sperm proteome and the *D. melanogaster* sperm proteome, providing a robust framework for comparisons. We will subsequently consider the differences in proteome composition between the separated eupyrene and apyrene sperm protein samples.

Shotgun proteomics

After removing all contaminants from the sample and filtering protein hits by significance, a total of 706, 667 and 398 proteins were found in the commixed, eupyrene and apyrene *D. plexippus* sperm samples, respectively (Table 1). Each of these samples were run through the LC-MS/MS process separately which accounts for the differences in total proteins found, as the methodology is not perfect in capturing all proteins found in any one sample. This results some differences in proteins found with each run. A total of 831 unique sperm proteins were found across all samples.

	Count of unique proteins found in each sample	Count of shared proteins found in each sample	Count of total proteins found in each sample
Apyrene	50	348	398
Eupyrene	319	348	667
Commixed	114	592	706

Table 1 Protein Counts: Unique, shared and total protein counts for all three samples.

Functional analysis of the commixed sperm proteome

The functional composition of the commixed sperm sample was summarized using gene ontology (GO) terms. GO analysis offers a hierarchical characterization of functional annotation that is comparable across all forms of biological and medical research. We then performed Fishers Analysis of gene enrichment on the *D. plexippus* sperm proteome relative to the complete *D. plexippus* proteome (Figure 1)(Conesa et al., 2005). This test explores if two sets of genes are differ in their biological roles through over or under representation of gene categories. This analysis identified GO categories found in the sperm proteome that were over-represented in relation to the *D. plexippus* genome (two tailed $P \ll 0.05$). The majority of proteins had functions consisting of cell maintenance, cell energy, microtubules and mitochondrial processes. These results are consistent with protein functions reported as enriched in other sperm proteomes, as these GO terms are found as overrepresented in many sperm characterizations (Castillo et al., 2013; Gilany et al., 2011).

Direct comparison of GO categories present in the *D. plexippus* and *M. sexta* sperm proteomes also revealed high levels of functional conservation. As mentioned before, these results are consistent with other sperm proteome characterizations, as many sperm proteomes seemed to have high levels of functional overlap. Figure 2 shows the similarities of functional groups in the *D. plexippus* and *M. sexta* commixed sperm proteomes.

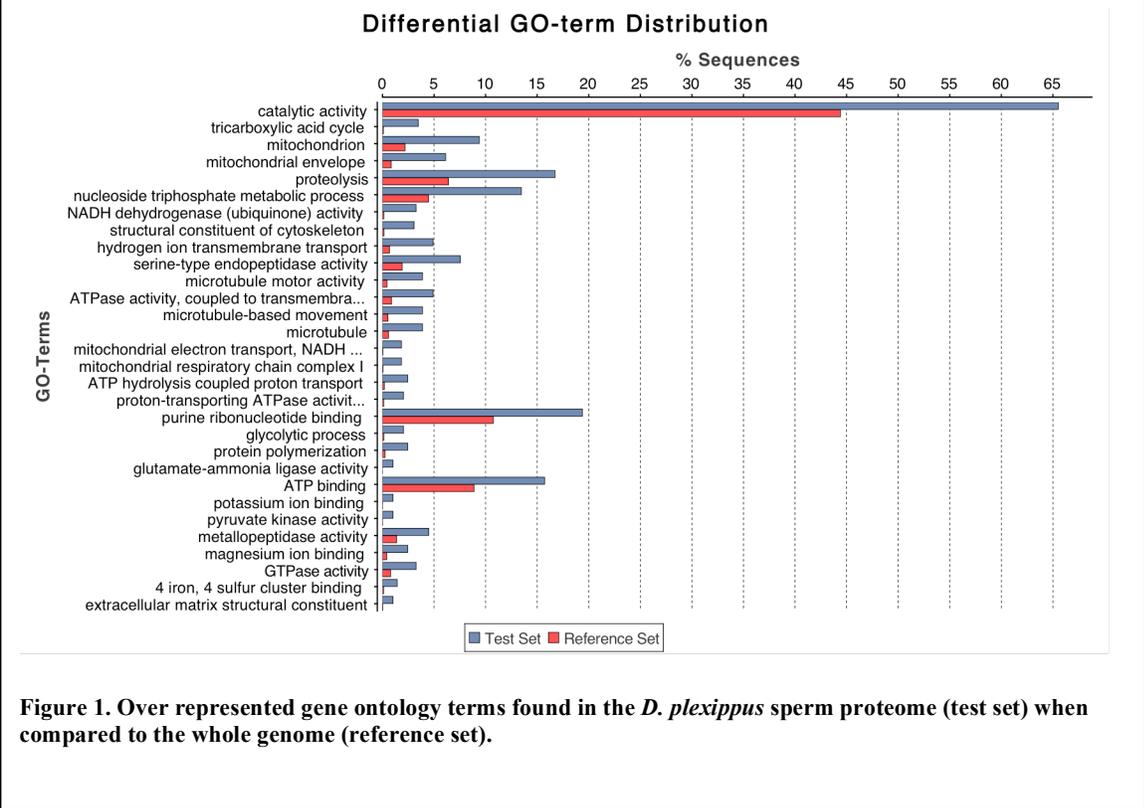
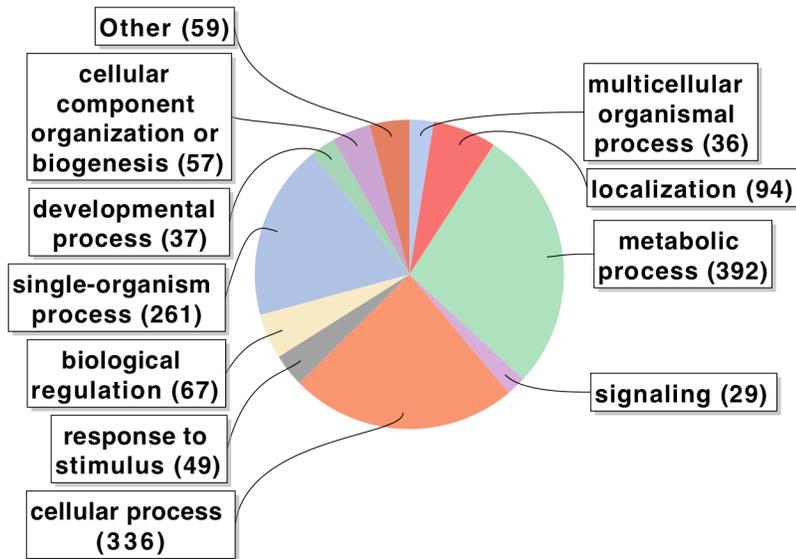


Figure 1. Over represented gene ontology terms found in the *D. plexippus* sperm proteome (test set) when compared to the whole genome (reference set).

D. Plexippus commixed gene ontology analysis



M. sexta commixed gene ontology analysis

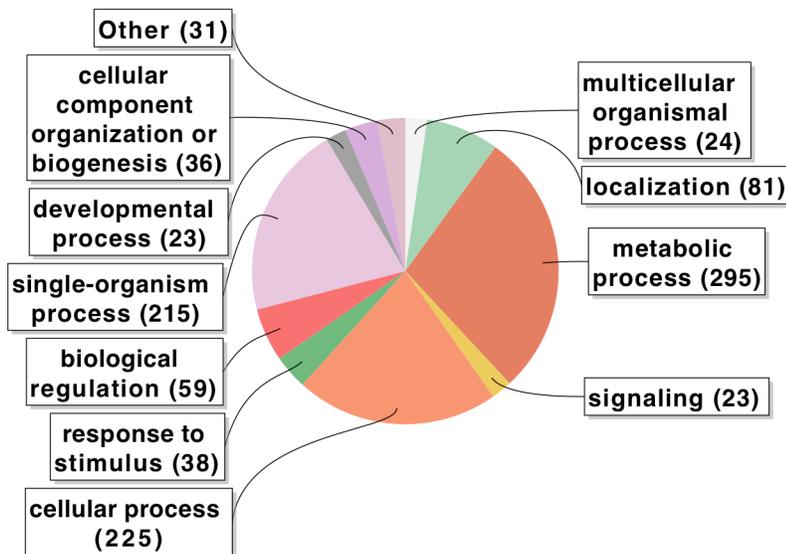


Figure 2 Gene ontologies from *D. plexippus* and *M. sexta*. This shows the relative amounts of proteins in certain functional categories. The numbers shown in the pie charts do not represent an actual count of proteins found in the sperm proteomes, but rather the number of GO-term assignments; a single protein may be assigned multiple GO-terms.

Comparison of sperm proteomes

We investigated patterns of orthology for the *D. plexippus* commixed sperm sample in relation to both commixed *M. sexta* and monomorphic *D. melanogaster* sperm samples.

Proportions of putative orthologous monarch sperm proteins are reported in Table 2.

Commixed Sample	<i>Proportion of orthologs shared with D. melanogaster</i>	<i>Proportion of orthologs shared with M. sexta</i>
Sperm Orthologs	.14	.46
Non-sperm orthologs	.01	.01
No orthology	.85	.53
Total Commixed proteins	706	706

Table 2: Orthology proportions for the commixed sample. Sperm orthologs are the orthologous proteins found in the sperm of all three species. Non-sperm orthologs represents the orthologs found in *D. plexippus* sperm, but not in the compared species' sperm proteomes.

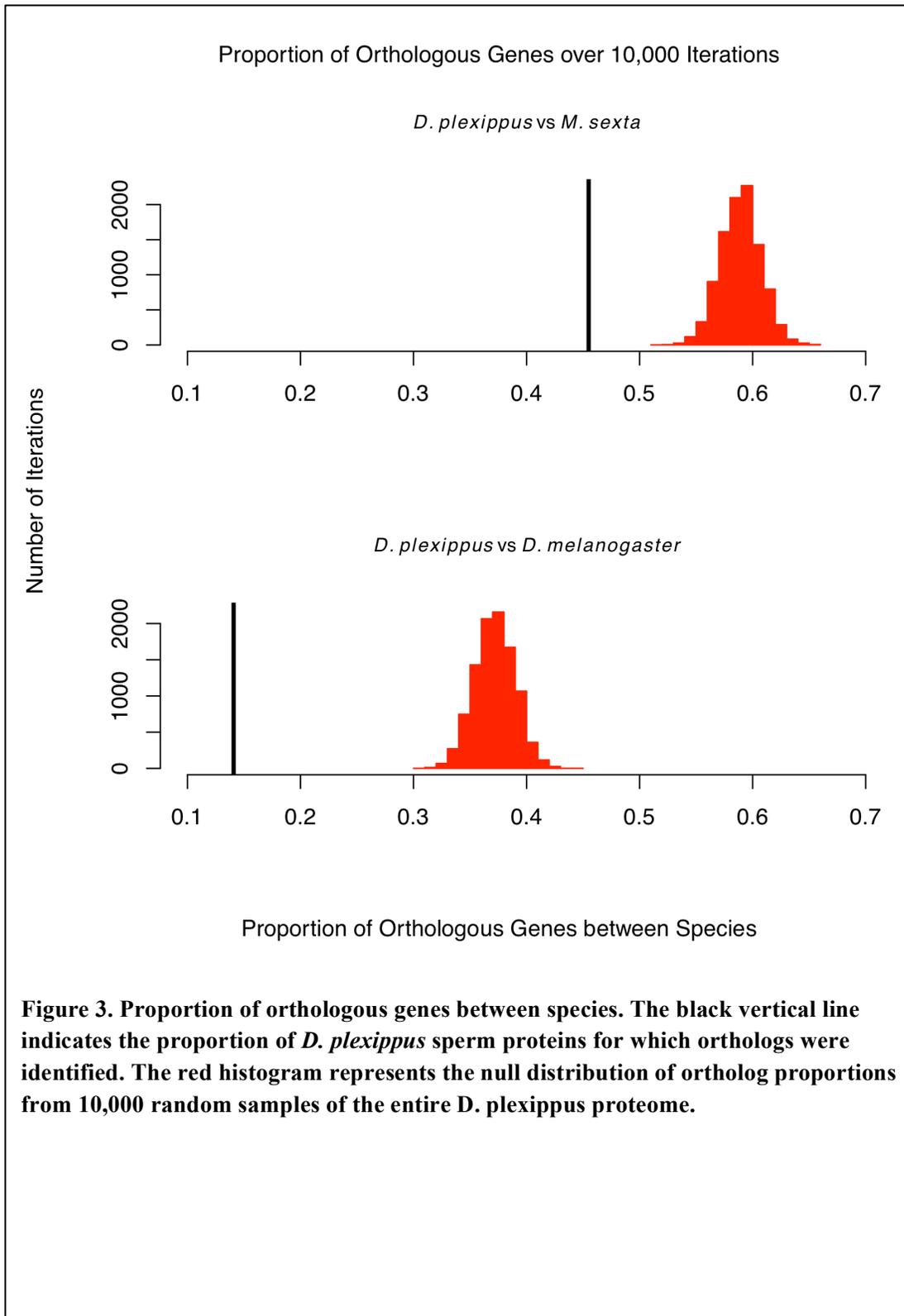
As expected, *D. plexippus* shares a higher proportion of orthologous sperm proteins, i.e. sequence similarity, with its closer relative, *M. sexta*. Strikingly, there seem to be relatively few non-sperm orthologs found in either species. These are proteins found in the sperm of *D. plexippus*, but not detected in the sperm proteomes of the other species. This pattern may arise either from a orthologous protein that was ancestrally a sperm protein that has since lost its function in sperm, or alternatively, a protein that has gained function in sperm. Less than half of the commixed sperm proteome has an orthologous match in *M. sexta*, and even fewer orthologs are found in *D. melanogaster*.

This pattern reveals substantial and ongoing turnover among insect sperm proteomes, as the more distant the relative from *D. plexippus*, a reduction in overall orthology is seen. It further suggests that current sperm proteins do not persist or exist in the genome when they are not functioning in sperm, at least with reference to *D. plexippus* sperm. Replicating this pattern in reciprocal analyses using *M. sexta* and *D. melanogaster* as the reference will greatly strengthen the confidence of this observation.

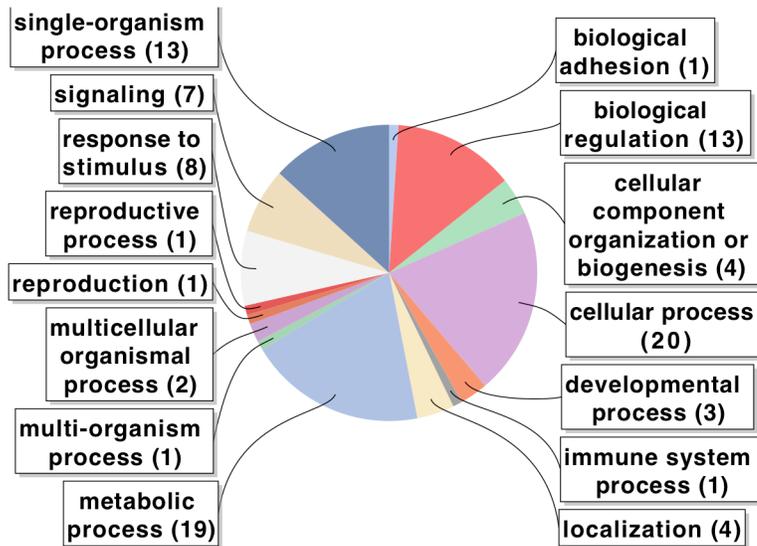
The *D. plexippus* sperm proteome showed a significantly lower ortholog proportion relative to the overall genomic ortholog proportion in the cases of both *M. sexta* and *D. melanogaster* (randomization test $P \ll 0.001$) (Figure 3). This indicates a lower proportion of orthology in the sperm proteome relative to the complete *D. plexippus* proteome.

The relative paucity of orthologs and high gene turnover in the sperm proteome is consistent with relatively rapid evolution that has been widely observed among reproductive proteins (Clark et al., 2005; Swanson and Vacquier, 2002b). Yet this is somewhat contradictory to the high levels of functional conservation seen in the GO analysis. One possible explanation is that many of common functional categories identified in the sperm proteome are highly conserved cell functions, such as cellular metabolism and locomotion. As such, these proteins contain deeply conserved functional motifs and sequences, implying paralogy. So while the exact proteins performing these functions may be regularly replaced in the sperm proteome, the overall functional classes remain highly conserved. In essence, fundamental protein domains are conserved despite relatively rapid evolution of the proteins performing these functions. There is precedent for this pattern of low orthology but high conservation of functional classes among other male reproductive proteins. In particular, the pattern of conserved structural and functional similarity in the face of rapidly changing protein sequences has been reported for many non-

sperm seminal fluid proteins (i.e., male accessory gland proteins) (Mueller et al., 2004; Walters and Harrison, 2010). Our results obtained for sperm proteins appear to extend the evidence for this phenomenon.



Unique Apyrene



Unique Eupyrene

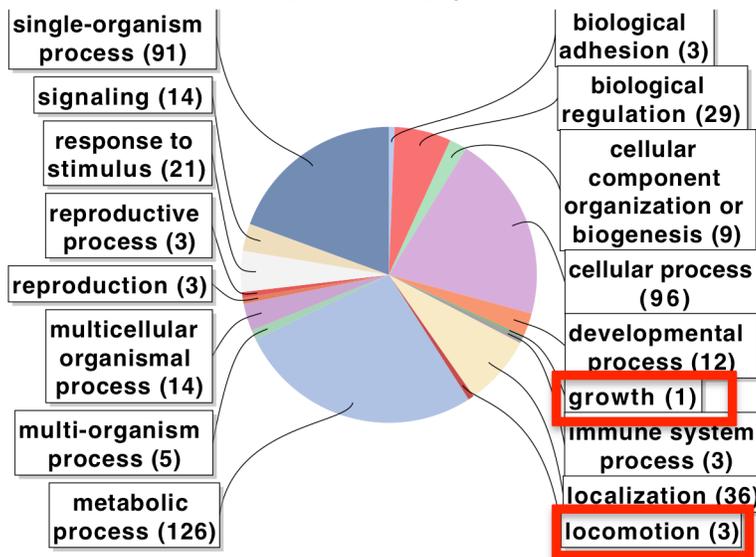


Figure 5 Gene ontology of unique apyrene and eupyrene sperm types. Red boxes indicate GO functional categories unique to the eupyrene sperm proteome.

The Separation of apyrene and eupyrene sperm types

The successful separation of apyrene and eupyrene sperm allows for further analyses to tease apart their differences to understand apyrene sperm function (Karr & Walters, 2015). There is considerable overlap in the proteins identified between the apyrene and eupyrene sperm proteomes (Table 1). In particular, the apyrene sperm protein set appears to mostly be a subset of eupyrene sperm, with only 50 unique proteins found in the apyrene sperm proteome. We then compared the unique protein sets of both the apyrene and the eupyrene sperm proteomes using GO functional annotations. Very few functional differences were found between the two proteomes (Figure 5). The only differences found between the two sperm proteomes in regards to functional annotations were a small number of annotations in the eupyrene sperm for locomotion and growth that were not detected in the apyrene proteome.

Since apyrene sperm have relatively few unique proteins and no unique functional annotations, no radically distinct function of apyrene sperm relative to eupyrene sperm is apparent, despite the obvious inability of apyrene sperm to transfer nuclear DNA to the egg. However, the lack of differences found in the apyrene proteome does not undermine support for either the cheap filler or facilitation hypotheses. Indeed, the hypothesis that apyrene sperm may act as cheap filler is particularly appealing given that apyrene sperm has the same functional annotations as eupyrene sperm, but has 40% fewer proteins in the proteome and no nuclear DNA. Since few differences were found between the apyrene and eupyrene sperm in functional annotations, apyrene sperm are less likely to have a novel function and are far cheaper to make.

Patterns of orthology within the unique apyrene and eupyrene proteins follow the same trend as the commixed sperm sample. Orthologous sperm proteins are in higher abundance than the orthologous non-sperm proteins (Table 3). Curiously, more protein orthologs were found in

comparisons to *D. melanogaster* than to *M. sexta*, but this could be a product of chance with such low overall protein numbers.

	Unique Apyrene		Unique Eupyrene		Shared	
	<i>M. sexta</i>	<i>D. mel</i>	<i>M. sexta</i>	<i>D. mel</i>	<i>M. sexta</i>	<i>D. mel</i>
Proportion of sperm ortholog	0.12	0.18	0.29	0.07	0.56	0.20
Proportion of non-sperm ortholog	0.0	0.02	0.01	0.01	0.02	0.01
Proportion with no orthology	0.88	0.80	0.70	0.92	0.42	0.79
Total protein count	50		398		348	

Table 3: Proportion of orthologous proteins in the separated sperm types.

Conclusion and Further Directions

This project provides pioneering insights into the sperm proteome of *D. plexippus* and the differences between apyrene and eupyrene sperm types. Functional categories across the *D. plexippus* and *M. sexta* sperm proteomes are similar to each other and also to sperm proteomes from more divergent species. Despite this apparently conserved function, patterns of orthology indicate more rapid turnover of orthologous sperm proteins than proteins found in the overall proteome. This suggests there is broad maintenance of gene function despite rapid turnover of gene orthology in the monarch sperm proteome.

Another possible explanation for a pattern of high turnover in orthologous proteins is relaxed evolutionary constraint on the apyrene sperm type. There have been mixed reports on the presence or absence of apyrene sperm in the family Micropterigidae, the basal group of Lepidoptera (Regier et al., 2013). One genus that has been closely studied is *Micropterix hubneri*. *Micropterix* completely lack apyrene sperm, and their eupyrene sperm resemble apyrene sperm in their wispy, unbundled nature (Sonnenschein and Häuser, 1990). Although only speculation, one could imagine an evolutionary transition between these basic unbundled eupyrene sperm types to a bundled eupyrene sperm type, and the apyrene sperm functioning as nothing more than an evolutionary leftover that has not dissipated yet. To test this hypothesis, a sperm collection would need to be collected for different members along the Micropterigidae family tree to test for the presence or absence of the apyrene sperm type. Furthermore, I would suspect the more basal eupyrene sperm to be a subset of the derived eupyrene sperm. One could test this using proteomics to uncover protein IDs and calculate the proportion of identical proteins found between the two eupyrene sperm sets.

The apyrene sperm proteome was found to be mostly a subset of eupyrene sperm. Comparisons of unique proteins found in apyrene versus eupyrene sperm did not clearly indicate a function of apyrene sperm distinct from eupyrene sperm.

Support for the cheap filler hypothesis could come from either the hypothetical situation of transitioning apyrene sperm and the finding that apyrene sperm are a smaller, less extensive subset of the eupyrene sperm type. If the hypothesis of transitioning apyrene sperm was found true, these transitioning apyrene sperm may be remnants of the old eupyrene sperm type and are being utilized as cheap filler in most Lepidopteran species. Alternatively, since apyrene sperm has far fewer proteins in the proteome than the eupyrene sperm and no nuclear DNA, it may be cheaper to produce than the more costly eupyrene sperm.

On the other hand, these results do not lend evidence for or against the facilitation hypothesis. This hypothesis is currently being further explored, as our lab is learning techniques to artificially inseminate *D. plexippus* females in the hopes of repeating the artificial insemination experiment on *Bombyx mori*. If fertilization success is drastically lowered in *D. plexippus* with the removal of apyrene sperm, more support will be lent to the facilitation hypothesis.

Another promising line of research that emerges is through findings of candidate proteins. Now that a full list of proteins has been found for apyrene sperm proteome, a closer look could reveal unique proteins that play a role in either the cheap filler, facilitation, or shed light on some new function. Genetic knockdowns of these proteins in behavior and mating experiments should illuminate which proteins found in apyrene sperm are important, and elucidate the overall role apyrene sperm have in reproductive processes. In a first glance, no

proteins in the apyrene sperm set stand out. However, protein abundance has not been yet been measured, and could play an important role in determining overall function. For example, if apyrene sperm had a higher abundance of locomotion proteins it could lend evidence to a highly mobile sperm type that could help in facilitation.

While much work lies ahead to unravel the paradox of sperm heteromorphy in Lepidoptera, the work presented here lays a broad and strong foundation for novel approaches to study this phenomenon with modern molecular genetic approaches.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Castillo, J., Amaral, A., Oliva, R., 2013. Sperm nuclear proteome and its epigenetic potential. *Andrology* 2, 326–338.
- Clark, N.L., Aagaard, J.E., Swanson, W.J., 2005. Evolution of reproductive proteins from animals and plants. *Reproduction* 131, 11–22.
- Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., Robles, M., 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674–3676. d
- Cook, P.A., Gage, M.J., 1995. Effects of risks of sperm competition on the numbers of eupyrene and apyrene sperm ejaculated by the moth *Plodia interpunctella* (Lepidoptera: Pyralidae). *Behavioral Ecology and Sociobiology* 36, 261–268.
- Gilany, K., Lakpour, N., Vafakhah, M., Sadeghi, M.R., 2011. The Profile of Human Sperm Proteome; A Mini-review. *Journal of Reproduction & Infertility* 12, 193–199.
- Harpel, D., Cullen, D.A., Ott, S.R., Jiggins, C.D., Walters, J.R., 2015. Pollen feeding proteomics: Salivary proteins of the passion flower butterfly, *Heliconius melpomene*. *Biological Insights from the Manduca sexta genome* 63, 7–13.
- Karr, T.L., Walters, J.R., 2015. Panning for sperm gold: Isolation and purification of apyrene and eupyrene sperm from lepidopterans. *Biological Insights from the Manduca sexta genome*.
- Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J., 2011. Proteinortho. *BMC Bioinformatics* 12, 124.
- Mueller, J.L., Ripoll, D.R., Aquadro, C.F., Wolfner, M.F., 2004. Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid. *Proceedings of the National Academy of Sciences* 101, 13542–13547.
- Parker, G., 1970. Sperm competition and its evolutionary consequences in the insects. *Biological Reviews* 45, 525–568.
- Perkins, D.N., Pappin, D.J., Creasy, D.M., Cottrell, J.S., 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567.
- Regier, J.C., Mitter, C., Zwick, A., Bazinet, A.L., Cummings, M.P., Kawahara, A.Y., Sohn, J.-C., Zwickl, D.J., Cho, S., Davis, D.R., Baixeras, J., Brown, J., Parr, C., Weller, S., Lees, D.C., Mitter, K.T., 2013. A Large-Scale, Higher-Level, Molecular Phylogenetic Study of the Insect Order Lepidoptera (Moths and Butterflies). *PLoS ONE* 8, e58568.
- Sahara, K., Takemura, Y., 2003. Application of artificial insemination technique to eupyrene and/or apyrene sperm in *Bombyx mori*. *J. Exp. Zool.* 297A, 196–200.
- Sonnenschein, M., Häuser, C.L., 1990. Presence of only eupyrene spermatozoa in adult males of the genus *Micropterix hübner* and its phylogenetic significance (Lepidoptera : Zeugloptera, Micropterigidae). *International Journal of Insect Morphology and Embryology* 19, 269–276.
- Swallow, J.G., Wilkinson, G.S., 2002. The long and short of sperm polymorphisms in insects. *Biol. Rev.* 77, 153–182.
- Swanson, W.J., Vacquier, V.D., 2002a. Reproductive protein evolution. *Annu. Rev. Ecol. Syst.* 33, 161–179.
- Swanson, W.J., Vacquier, V.D., 2002b. The rapid evolution of reproductive proteins. *Nature*

Reviews Genetics.

- Wallow, J.G., Harrison, R.G., 2010. Combined EST and Proteomic Analysis Identifies Rapidly Evolving Seminal Fluid Proteins in *Heliconius* Butterflies. *Molecular Biology and Evolution* 27, 2000–2013.
- Wasbrough, E.R., Dorus, S., Hester, S., Howard-Murkin, J., Lilley, K., Wilkin, E., Polpitiya, A., Petritis, K., Karr, T.L., 2010. *Drosophila melanogaster*. *Journal of Proteomics* 73, 2171–2185.
- Wedell, N., Cook, P.A., 1999. Butterflies tailor their ejaculate in response to sperm competition risk and intensity. *Proceedings of the Royal Society B: Biological Sciences* 266, 1033–1039.
- Whittington, E., Zhao, Q., Borziak, K., Walters, J.R., Dorus, S., 2015. Characterisation of the *Manduca sexta* sperm proteome: Genetic novelty underlying sperm composition in Lepidoptera. *Biological Insights from the Manduca sexta genome* 62, 183–193.
- Wiegmann, B.M., Mitter, C., Regier, J.C., Friedlander, T.P., Wagner, D.M., Nielsen, E.S., 2000. Nuclear Genes Resolve Mesozoic-Aged Divergences in the Insect Order Lepidoptera. *Molecular Phylogenetics and Evolution* 15, 242–259.
- Zdobnov, E.M., Apweiler, R., 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17, 847–848.