SEXUAL SELECTION AND THE EVOLUTION OF SEMINAL FLUID PROTEINS IN *HELICONIUS* BUTTERFLIES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

James Richmond Walters

August 2009

SEXUAL SELECTION AND THE EVOLUTION OF SEMINAL FLUID PROTEINS IN *HELICONIUS* BUTTERFLIES

James Richmond Walters, Ph. D.

Cornell University 2009

Reproductive proteins tend to diverge unusually rapidly between species. This pattern is frequently attributed to post-mating sexual selection. However, despite many well-characterized examples of rapidly evolving reproductive proteins, little data exist which directly address this widely invoked hypothesis. The *Heliconius* genus of butterflies offers a good opportunity to examine this hypothesis by contrasting patterns of reproductive protein evolution between clades with divergent mating systems: adult-mating and pupal-mating. With few exceptions, pupal-mating females mate only once, which severely limits the opportunity for post-mating sexual selection to act. In contrast, adult-mating females mate repeatedly throughout life, providing ample opportunity for post-mating sexual selection to drive the adaptive evolution of reproductive proteins. Thus theory predicts that reproductive protein evolution should be slower in the pupal-mating clade relative to the adult-mating clade.

Focusing initially on two species, *H. erato* (pupal-mating) and *H. melpomene* (adult-mating), I used a combination of expression, bioinformatic, and proteomic to identify 51 putative seminal fluid proteins. Evolutionary rate estimates based on pairwise alignments reveal these *Heliconius* seminal fluid proteins evolve more rapidly than a set of ~300 'non-reproductive' proteins derived from wing tissue. To further explore evolutionary patterns of reproductive proteins in *Heliconius* and to

compare mating systems, I sequenced 20 seminal fluid protein genes from 10 more *Heliconius* species and an outgroup. Applying codon-site models to these data indicated three proteins with dN/dS > 1, strongly implicating positive selection in the rapid evolution of at least a few *Heliconius* seminal fluid proteins. Comparison evolutionary rates between clades yielded the result that, contrary to predictions, the average evolutionary rate of seminal fluid proteins is greater among pupal-mating *Heliconius*. These results suggest that positive selection and relaxed constraint can generate conflicting signals when examining patterns of reproductive protein evolution across mating systems. As predicted, some loci may show elevated evolutionary rates in promiscuous taxa relative to monandrous taxa resulting from adaptations to post-mating sexual selection. However, when monandry is derived (as in *Heliconius*), the opposite pattern may result from relaxed constraint of loci formerly influenced by post-mating sexual selection.

BIOGRAPHICAL SKETCH

James (Jamie) Walters was born and raised in the Maryland suburbs of Washington, D.C. He was educated in the public schools of Montgomery Country, graduating from Bethesda – Chevy Chase High School in 1996. He matriculated the following fall at Bowdoin College in Brunswick, Maine and graduated in 2000, *Summa cum Laude*, with a B.A. and honors in Biology and a minor in French Language. His senior honors thesis, conducted under Dr. Michael Palopoli, focused on the molecular evolution of a sperm protein in nematode round worms and strongly influenced the trajectory of his scientific career.

During college Jamie developed a keen interest in outdoor recreation, education, and leadership which he maintains (and maintains him) to this day. He was deeply involved with the Bowdoin Outing Club, leading and participating in dozens of hiking, biking, and paddling trips. Most significantly, he became an avid whitewater kayaker, which was supported by spending summers teaching kayaking and outdoor skills at the Valley Mill Summer Camp in Darnestown, Maryland. This experience provided Jamie with the opportunity to become manager of the fledgling Valley Mill Kayak School for two summers and the intervening winter immediately upon graduating from Bowdoin. By the end of his second and final summer at VMKS (its fourth operating season), the school finally turned a profit and Jamie was ready to turn back to academia and research.

In the fall of 2001 Jamie took a part-time research technician position at the University of Maryland, College Park, in the laboratory of Jerry Wilkinson conducting evolutionary genetics research on Malaysian stalk-eyed flies. He applied to graduate school that same fall, arriving at Cornell a year later with Rick Harrison as his graduate advisor. Completing his doctorate at Cornell shaped his future academic and

professional life in many profound ways, but two facets of this experience stand out. First was his decision to conduct research in *Heliconius* butterflies, a taxon which has progressed substantially as a genomic model system and around which a particularly open and collaborative research community has developed. Second was his effort to acquire computational and bioinformatic skills during the course of his research. The combination of these two things set the stage for his being awarded a National Science Foundation post-doctoral training fellowship in bioinformatics. This award will take Jamie to the laboratory of Chris Jiggins at University of Cambridge, UK for two years of post-doctoral research and training. This will be followed by third and final year in the laboratory of Hunter Fraser at Stanford University, Palo Alto, California.

There is one other noteworthy event during Jamie's time at Cornell and in Ithaca which is important to include here. Almost exactly two years before completing his degree, he met Meg Jamieson. Jamie and Meg art to be married on July 25, 2009.

Dedicated to:

My mother,

who – by nature and nurture both – gave me interest and ability,

My father,

for unflagging support, enthusiasm, and approval,

and Meg,

for incomprehensibly profound love and devotion.

ACKNOWLEDGMENTS

I have received tremendous support from many, many people in pursuing my doctorate and graduate research. Foremost on this list is Rick Harrison, my advisor, who provided unerring guidance and assistance with matters great and small, scientific and bureaucratic, financial and pedagogical, and who, above all, was concerned that I keep happy and healthy while under his watch. A prominent supporting role in all of this was played by the other two members of my committee, Andy Clark and Kelly Zamudio. I will not soon forget their optimistic stance in the face of daunting tasks.

Steve Bogdanowicz and Jose Andres paved the way for me on many fronts in the lab and provided invaluable wisdom and guidance every time a pipette was involved in my work. All members of the Harrison lab with whom I overlapped contributed in important ways to this effort and I am grateful to the collaborative spirit that infuses our lab, department, and Cornell. Other Cornellians who generously gave shared time and knowledge to help me include (in no particular order): Tim Sackton, Sarah Stockwell, Marta del Campo, Bryan Danforth, Adam Seipel, Amy McCune, Tony Greenberg, Anthony Fiumera, Nathan Clark, Nadia Singh, Dara Torgerson, Laura Sirot, Alex Wong, and many others whom I'm surely and sadly forgetting here.

None of this work could have been done without the collaboration of the *Heliconius* research community. Owen McMillan, Larry Gilbert, and Chris Jiggins, were unhesitating and remarkably generous in their assistance. This was also true of all the members of their labs, especially Catalina Estrada, Alejando Merchan, Riccardo Papa, Brian Counterman, Mathieu Joron, Moises, Durrell Kapan, and Marcus Kronforst. I am also deeply indebted to Cheryl Tyndall and the Niagara Parks Butterfly Conservatory for making their resources and butterflies available to me time and again.

I owe a huge debt of gratitude to Jarek Pillardy and Qi Sun of the Cornell theory center as well as Brian Mlodzinski for helping me navigate and troubleshoot all sorts of computational challenges, not the least of which was my own ignorance.

Funding this work came from the National Science Foundation, in the form of a Graduate Research Fellowship, a Doctoral Dissertation Improvement Grant, and also Rick's grants focusing on seminal fluid proteins in *Gryllus*. Additional financial support came from Cornell administrated grants including the Andrew W. Mellon funds, Sigma Xi, Orenstein, and EEB departmental research allowances.

Outside of the immediate scientific sphere, there were many other wonderful friends and compatriots who made the good things in life better and the bad things in life bearable. My roommates at 314 #5 over the years certainly rank highly on this list: Laura Lyon, Beth Johnson, Michael Wunsch, Rayna Bell, and Dan Rabosky. The Cornell Outdoor Education community provided a haven of like-minded compatriots where I found many of my closest and most valued friends in Ithaca: Brendan Kelley, Iori Ueki, Brad Treat, Kara Spillman, Ben Blakely, Alana Jonat, and Lindsay Watkins. The yoga crowd was also a wonderful community of friends, with Kelly Wolfe and Drew Dolgert at the core.

Two folks, John Skrovan and Linda Warner, kept me in sound mind and body, respectively, while I navigated the pitfalls of graduate school.

Two other people merit special thanks: Dan Rabosky and Drew Dolgert. These two transcend the roles of friend and mentor, having blessed me with perspective and solutions in the face of both intellectual and emotional crises.

Of course I could not have accomplished any of this without the love and support of my parents. And speaking of love and family, there is no end to my gratitude for the patience and devotion of my fiancé, Meg Jamieson, who cheered me on selflessly in the home stretch.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

The work presented here was conceived and conducted in the midst of a revolution in biological research – the genomic revolution. The rapidly advancing technologies fueling this revolution increasingly allow unprecedented opportunities for investigating biological diversity. For evolutionary biologists, the sudden and tremendous increase in biological sequence data has opened completely novel ways to investigate similarities and differences between organisms. I like to think of such research as doing "genomic natural history", where inferences arise from comparisons of the size, content, complexity, and diversity of genomes between groups of organisms (for instance, between prokaryotes and eukaryotes) or between groups of genes. Such investigations of genomic diversity, typically lumped together under the rubric of "comparative genomics", have yielded many surprising observations, several of which remain poorly understood.

My thesis research specifically concerns one such observation: the unusually rapid and frequently adaptive molecular evolution of reproductive proteins. The data and analyses presented here address the leading hypothesis explaining this phenomenon, namely that post-mating sexual selection is the primary evolutionary process underlying this pattern. I have focused on testing this hypothesis in *Heliconius* butterflies, a neo-tropical genus consisting of about 40 species and exhibiting a striking dichotomy in mating systems. About half of the species exhibit an unusual pupal-mating behavior where females mate during or immediately after emerging from their chrysalis and typically mate only once. This provides very little opportunity for post-mating sexual selection to influence the evolution of reproductive proteins in these species. In contrast, the females in the remainder of *Heliconius* species mate as fully mature adults and typically mate more than once, providing

ample opportunity for sexual selection to act on reproductive proteins. Thus, my thesis addresses two related questions. First, is there evidence for relatively rapid and adaptive evolution among reproductive proteins in *Heliconius* butterflies as observed in many other taxa? Second, does the evolutionary rate of reproductive proteins differ between the two mating systems present in *Heliconius* butterflies in a way that is consistent with post-mating sexual selection elevating the evolutionary rates of these proteins?

Answering these questions has proceeded in three stages, which correspond to the three chapters of my dissertation. The first challenge in pursuing this research was to identify a representative sample of reproductive proteins in *Heliconius* butterflies. Doing so is related here in the first and second chapters. The first chapter describes the construction and sequencing of genetic libraries generated from male reproductive tissues in two *Heliconius* species, one from each of the two mating systems. These results are the foundation for the second chapter, in which genes sampled from these tissues are evaluated as to whether they encode reproductive proteins. The second chapter also establishes that putatively identified reproductive proteins are indeed rapidly evolving in *Heliconius* butterflies. It is in the third chapter where I test for differences in evolutionary rate among these reproductive proteins between the two mating systems.

Ultimately my results do indicate a difference in evolutionary rates between the two mating systems, but this difference cannot be interpreted as unequivocal support for the simple explanation that post-mating sexual selection elevates the evolutionary rate of reproductive proteins. Rather, my results show a decoupling of rapid evolution from adaptive evolution and suggest that post-mating sexual selection by itself does not provide a comprehensive explanation for the rapid and adaptive evolution of reproductive proteins.

CHAPTER 1

EST ANALYSIS OF MALE ACCESSORY GLANDS FROM *HELICONIUS* BUTTERFLIES WITH DIVERGENT MATING SYSTEMS

Abstract

Heliconius butterflies possess a remarkable diversity of phenotypes, physiologies, and behaviors that has long distinguished this genus as a focal taxon in ecological and evolutionary research. Recently *Heliconius* has also emerged as a model system for using genomic methods to investigate the causes and consequences of biological diversity. One notable aspect of *Heliconius* diversity is a dichotomy in mating systems which provides an unusual opportunity to investigate the relationship between sexual selection and the evolution of reproductive proteins. As a first step in pursuing this research, we report the generation and analysis of expressed sequence tags (ESTs) from the male accessory gland of *H. erato* and *H. melpomene*, species representative of the two mating systems present in the genus *Heliconius*. We successfully sequenced 933 ESTs clustering into 371 unigenes from *H. erato* and 1033 ESTs clustering into 340 unigenes from *H. melpomene*. Results from the two species were very similar. Approximately one-third of the unigenes showed no significant BLAST similarity (E-value $<10^{-5}$) to sequences in GenBank's non-redundant databases, indicating that a large proportion of novel genes are expressed in *Heliconius* male accessory glands. In both species only a third of accessory gland unigenes were also found among genes expressed in wing tissue. About 25% of unigenes from both species encoded secreted proteins. This includes three groups of highly abundant unigenes encoding repetitive proteins considered to be candidate seminal fluid proteins; proteins encoded by one of these groups were detected in *H. erato* spermatophores. This collection of ESTs will serve as the foundation for the

future identification and evolutionary analysis of male reproductive proteins in *Heliconius* butterflies. These data also represent a significant advance in the rapidly growing collection of genomic resources available in *Heliconius* butterflies. As such, they substantially enhance this taxon as a model system for investigating questions of ecological, phenotypic, and genomic diversity.

Introduction

One of the most promising and productive research approaches in contemporary biology involves deploying modern genomic methods to investigate the origin, maintenance, and function of biological diversity present in natural populations. Research efforts in this nascent field of evolutionary and ecological functional genomics (EEFG) generally can be split into two categories (Feder & Mitchell-Olds 2003, Mitchell-Olds *et al.* 2008). One approach studies natural populations of the few taxa (or their close relatives) that are already well-established laboratory model systems, making use of the extensive molecular genetic and genomic resources available for such organisms (e.g. *Drosophila* and *Arabidopsis*) (Shimizu & Purugganan 2005, Markow & O'Grady 2007). The alternative approach focuses on taxa which may be less tractable from a methodological perspective but which offer superb opportunities to investigate interesting and important ecological and evolutionary phenomena. In the case of such emerging model taxa, the development of genomic resources such as genetic libraries, linkage maps, and sequence databases are necessary and fundamental first steps in any EEFG research program (Feder & Mitchell-Olds 2003).

Heliconius butterflies stand out among such emerging model taxa for their extensive history in ecological and evolutionary research (Bates 1862, Gilbert 1972, Brown 1981, Mallet 1989, Deinert *et al.* 1994, Brower 1997, Jiggins *et al.* 2001,

Kronforst *et al.* 2007). The genus *Heliconius*, consisting of ~40 neotropical species, contains a remarkable diversity of phenotypes, behaviors, and physiologies, all of which have evolved relatively recently (Brower 1994, Brower 1997, Flanagan *et al.* 2004, Beltran *et al.* 2007). The most conspicuous and well-studied aspect of this diversity is the variation and convergence/mimicry of wing-color patterns present both within and between species (Benson 1972, Mallet 1989, Nijhout *et al.* 1990, Jiggins & McMillan 1997, Jiggins *et al.* 2001, Kapan 2001, Joron *et al.* 2006, Reed *et al.* 2008). Indeed it is the efforts to identify the genetic basis of this wing pattern diversity that have thus far driven the recent development of genomic resources for *Heliconius* butterflies (Joron *et al.* 2006, Jiggins *et al.* 2008, Beldade *et al.* 2008). The accumulation of such resources now provides a strong precedent for investigating additional aspects of *Heliconius* diversity.

Here we present the first genomic foray into facets of *Heliconius* diversity other than wing pattern. We focus on a striking dichotomy of mating system found within the genus and sample the transcriptome of male reproductive tissues from species representative of the two mating systems. In particular, we choose as our focal species the co-mimetic *H. erato* and *H. melpomene*. These two species have an average synonymous site divergence of 14.5%, do not interbreed, show extensive parallel radiations of wing patterns, and are the primary focus for wing pattern research in the genus (Papanicolaou *et al.* 2005, Joron *et al.* 2006, Jiggins *et al.* 2008). Consequently, these species possess the vast majority of genomic resources available for *Heliconius*. Both species have BAC libraries, linkage maps, and extensive collections of expressed sequence tags (ESTs) generated from wing tissue and curated in a lepidopteran-specific database (Jiggins *et al.* 2005, Kapan *et al.* 2006, Joron *et al.* 2006, Papanicolaou *et al.* 2008, Jiggins *et al.* 2008).

Heliconius erato and *H. melpomene* represent the two divergent mating systems found in the genus. About half of *Heliconius* species, including *H. erato*, exhibit an unusual pupal mating behavior: females are mated before or during eclosion and typically mate only once (i.e. females are monandrous). Otherwise, *Heliconius* butterflies, such as *H. melpomene*, mate as fully developed adults and regularly mate more than once (i.e. females are polyandrous) (Brown 1981, Gilbert 1991, Deinert *et al.* 1994, Brower 1997, Beltran *et al.* 2007). *Heliconius* species fall evenly into two major clades which correspond perfectly with mating system (Beltran *et al.* 2007).

This difference in mating system can engender very different intensities in sexual selection between the two clades. For instance, the pupal mating system drives extremely intense pre-mating sexual selection; males compete vigorously for mating position on the female chrysalis (Deinert *et al.* 1994). In contrast, the lack of remating in pupal mating females likely precludes most aspects of post-mating sexual selection such as sperm competition and cryptic female choice (reviewed in (Birkhead & Pizzari 2002). Therefore this phylogenetically concordant split in mating systems presents an unusual opportunity to explore hypotheses relating sexual selection and the molecular evolution of reproductive proteins.

Reproductive proteins include proteins mediating gametic interactions or those found in seminal fluid. These proteins tend to diverge rapidly between related species and often evolve via positive Darwinian selection (reviewed in (Swanson & Vacquier 2002a, Swanson & Vacquier 2002b, Clark *et al.* 2006). This is a pattern widely observed across the animal kingdom and also often in plants. It is commonly hypothesized that post-mating sexual selection is the primary evolutionary process underlying this pattern (Civetta & Singh 1998, Torgerson *et al.* 2002, Galindo *et al.* 2003, Andres *et al.* 2006, Panhuis *et al.* 2006, Haerty *et al.* 2007). However, there are

very few data currently available that directly address this hypothesis (but see (Dorus *et al.* 2004, Herlyn & Zischler 2007).

Ultimately we will use the dichotomous mating systems in *Heliconius* to test for a relationship between intensity of post-mating sexual selection and evolutionary rates of reproductive proteins. To do this it is first necessary to identify reproductive proteins in *Heliconius* butterflies. Here again we take our cues from previous EEFG research, though this time not from *Heliconius* but from *Drosophila* fruit flies and also from two genera of crickets. In these taxa researchers focused on proteins secreted by the accessory glands – part of the male reproductive tract – into seminal fluid. Early work in this field focused on indirect criteria such as the presence of a signal peptide and accessory gland biased expression to identify genes encoding accessory gland proteins (ACPs) which were assumed to be transferred to females in seminal fluid. Using modest numbers of ESTs generated from cDNA libraries enriched for male-biased transcripts, this work identified dozens of ACPs in these species (Swanson *et al.* 2001, Andres *et al.* 2006, Braswell *et al.* 2006). Subsequent studies at the protein level verified that many of these ACPs are transferred to females in seminal fluid (Herndon & Wolfner 1995, Bertram *et al.* 1996, Andres *et al.* 2008, Findlay *et al.* 2008, Sirot *et al.* 2008). These proteins have diverse and often dramatic effects on female reproductive physiology and behavior, including stimulating egg-laying, facilitating sperm storage, and inducing refractoriness to remating (Gillott 2003, Ram & Wolfner 2007). Moreover, in *Drosophila melanogaster*, genetic variation in seminal fluid proteins has been correlated with sperm competitive ability, indicating an important link between reproductive protein evolution and sexual selection (Clark *et al.* 1995, Fiumera *et al.* 2005).

In this paper we present parallel analyses of ESTs generated from the male accessory glands of the pupal mating *H. erato* and the adult mating *H. melpomene*.

These data constitute an important first step towards identifying a set of seminal fluid proteins in *Heliconius* butterflies and using these genes to examine the relationship between post-mating sexual selection and the molecular evolution of reproductive proteins. They also contribute significantly to the development of *Heliconius* butterflies into a sophisticated model system for genomic explorations of ecological and evolutionary phenomena.

Materials and Methods

RNA isolation and cDNA library construction

Male accessory glands were dissected from 11 adult male *Heliconius erato petiverana* (from stocks maintained at the University of Puerto Rico, Rio Piedras) and 10 adult male *Heliconius melpomene rosina* (from stocks maintained at the University of Texas, Austin). Tissue samples were placed immediately in TRIZOL reagent (Invitrogen, Carlsbad, CA) and homogenized. These and other subsequent total RNA extractions were done using TRIZOL and following the manufacturer's protocol.

Two directional cDNA libraries were constructed, one for each species, using the Creator SMART cDNA library kit (Clontech BD Bioscience, Mountain View, CA). Briefly, first-strand cDNA was reverse transcribed from 1.2 μg (*H. erato*) and .7 μg (*H. melpomene*) total RNA. Second-strand synthesis and amplification of cDNA pools for library construction were accomplished via Long Distance-PCR using the following cycling program: 1 min denaturation at 95°, 20 cycles of 30 sec at 95° then 6 min at 68°, and a final extension step of 6 min at 68°. Primers provided by the manufacturer were used for these reactions.

Seventy-five μl of the PCR-amplified cDNA were cleaned with Qiaquick PCR clean-up kits (Qiagen, Valencia, CA) and digested with SfiI. Digested cDNA was

electrophoresed on 1.2% TBE agarose gels and size-selected for transcripts >800 bp in length by gel extraction using a Qiaquick gel purification kit (Qiagen, Valencia, CA). The size-selected cDNA was ligated into the pDNR-LIB vector and used to transform electromax DH5α *E. coli* cells (Invitrogen, Carlsbad, CA) via electroporation with 2.5 kV/cm, 200 ohms, and 25 μF. Recombinant colonies were grown on chloramphenicol-selective LB agar medium. The *H. erato* and *H. melpomene* libraries contained $7x10^6$ and $1.3x10^5$ cfus respectively.

Library screening and EST sequencing

To enrich for transcripts expressed primarily in male tissue, both libraries were screened with cDNA generated from female abdominal tissue and only non-hybridizing clones were sequenced. Aliquots were plated at low density on chloramphenicol-selective LB agar medium and grown overnight at 37°. Colony lifts were made on Hybond XL membranes (Amersham Biosciences, Piscataway, NJ). Cells were lysed and DNA was fixed to the membrane by dry-cycle autoclaving at 250° (5 min sterilize, 5 min dry) followed by baking at 80° for two hours.

Probe for screening was generated from total RNA isolated from a single female abdomen. Four μg total RNA were used in a first-strand reverse transcription and subsequent second-strand synthesis/PCR amplification following the method of Chenchik *et al*. (Chenchik *et al.* 1998). Approximately 50 ng amplified cDNA was labeled with 32-P dCTP using the RADprime labeling kit (BioRad).

Before hybridization, membranes were soaked in 2x SSC solution and then incubated for 2 hours at 65° in hybridization buffer (0.5% BSA, 1mM EDTA, 7% SDS, 0.5M sodium phosphate). After 2 hours of pre-hybridization the radio-labeled female cDNA was added to the buffer and incubation continued overnight at 65°. Following hybridization, membranes were washed twice for 20 min with 1x

SSC/0.5% SDS at 65°, rinsed twice with 2x SSC at room temperature, dried, and imaged with x-ray film using a 5-day exposure.

In addition to screening with female cDNA, the *H. melpomene* library was simultaneously screened for four highly abundant transcripts and the stuffer fragment from the pDNR vector. The four highly abundant transcripts were identified by random sequencing of 356 clones before any hybridization screen. PCR primers were designed to amplify a portion of these 5 templates and 5 ng of amplicon from each, purified with a Qiaquick PCR clean-up kit, were combined and labeled with P^{32}-dCTP using the RADprime labeling kit (BioRad). This probe was added to the hybridization buffer at the same time as the probe generated from female cDNA.

Clones which failed to hybridize were manually picked into 50 µl 5mM Tris (pH 8.0) and lysed by heating at 99° for 5 min. One µl of this "boil prep" was used as template in a 10 µl PCR reaction using m13 primers (Clontech), platinum Taq polymerase (Invitrogen, Carlsbad, CA) and the following cycling program: initial denaturation of 95° (2 min), 35 cycles of 95° (50 sec) then 52° (1 min) then 72° (1 min), and a final extension of 72° (4 min). PCR amplified inserts were enzymatically cleaned with EXOSAP and single-pass sequenced from the 5' end using the ABI Prism BigDye Terminator Cycle Sequencing chemistry and a vector specific primer, SeqPrim3. Sequencing reactions were analyzed on an ABI 3730 automated sequencer.

EST analysis

EST data analysis was automated using the PartiGene suite of bioinformatic software (Parkinson *et al.* 2004). Raw sequences were trimmed of vector sequence, low-quality base calls, and poly-A tails (cutoff of 12 contiguous A's). Trimmed sequences >100 bp in length were clustered into putative unique gene objects (unigenes). Consensus sequences from each unigene were annotated via BLAST

searches to public databases (e.g. GenBank, SwissProt). Local BLAST databases were also used for all-vs-all BLAST searches to identify related sequences within and between libraries. Unigenes from wing tissues were downloaded from ButterflyBase (Papanicolaou *et al.* 2008). All BLAST searches were performed using the parallel-BLAST server hosted by the Cornell University computational biology service unit (cbsuapps.tc.cornell.edu). BLAST results were organized and analyzed using relational databases developed in Microsoft Access (Microsoft Corp., Redmond, WA). We screened the unigenes for nine *Heliconius* repetitive elements using the RepeatMasker software (Smit *et al.* 2004, Papa *et al.* 2008).

Putative open reading frames (ORFs) were identified and translated using the PartiGene suite's application prot4EST (Wasmuth & Blaxter 2004). Prot4EST utilizes several different methods for ORF prediction, including a hidden Markov model (HMM) approach implemented in ESTScan, which requires a large training set of complete coding sequences (Lottaz *et al.* 2003). At the time of analysis, a dataset of this type was not publicly available for any Lepidopteran species, so a 'simulated transcriptome' was generated for HMM training (J. Wasmuth, Personal Communication) (Cutter *et al.* 2006). First, codon usage statistics were estimated from pooled wing and accessory gland *Heliconius erato* unigenes for which coding sequences could be reliably identified via BLAST. Next, a 'simulated transcriptome' was generated by reverse-translating the *D. melanogaster* proteome using codon usage statistics estimated for *Heliconius erato*. The resulting data set was then submitted as a training set for ESTScan.

About one third of automatically predicted ORFs were manually inspected and, if necessary, edited using the Aligner (CodonCode Corp., Dedham, MA) or BioEdit (Hall 1999) software packages. These unigenes received this extra attention either because they were included in the set of orthologs for evolutionary analysis or because

they corresponded to the highly abundant tyrosine and asparagine rich proteins (see Results & Discussion for further information). For these unigenes, automated ORF predictions were replaced with manually edited versions for Gene Ontology annotations.

Where possible, Gene Ontology (GO) classifications were assigned to each protein translation based on BLASTX (E-value<10^{-5}) similarity to entries in a GO-annotated database (UNIPROT). GO annotations were summarized using 'GO-Slim' terms (Ashburner *et al.* 2000). This process was automated using the Annot8r application in the PartiGene package (Parkinson *et al.* 2004).

Secretory signal sequence peptides were predicted with the SignalP software (Nielsen *et al.* 1997, Bendtsen *et al.* 2004).

Patterns of tissue specific expression

We examined patterns of tissue-specific expression for a few unigenes of particular interest. Differences in expression were assayed via RT-PCR from three different tissues: male abdomen, male thorax, and female abdomen. PCR primers were designed within the predicted ORF of each unigene assayed. Primers were designed with the Primer3 software (Rozen & Skaletsky H.J. 2000). Total RNA was isolated from three adult male and female butterflies. A standard concentration of total RNA from each of these RNA extractions (*H. erato*, 1 μg; *H. melpomene*, 0.5 μg) was treated with DNase (Invitrogen, Carlsbad, CA) and reverse transcribed into single stranded cDNA using poly-T primers, SuperScript III Reverse Transcriptase (Invitrogen), and following the manufacturer's protocol. One μl of a 3-fold dilution of this cDNA was used as template in a 20 μl touch-down PCR with the following cycling parameters: initial denaturation of 95°C (2 min), 12 cycles of 95°C (30 sec) then 65-53°C (30 sec, decreasing one degree per cycle) then 72°C (2 min), 23 cycles

of 95°C (30 sec) then 53°C (30 sec) then 72°C (2 min), and a final extension of 72°C (4 min). For each set of primers an equal amount of PCR amplicon (between 4 and 9 µl) from each of the nine templates was electrophoresed on a 1.2% agarose gel, stained with ethidium bromide and visualized under UV light.

Spermatophore collections and proteomic analysis

H. erato individuals used in this experiment were taken from breeding stocks maintained at the Niagara Butterfly Conservatory, Niagara Falls, Ontario, Canada. Matings were performed in a 3m x 3m x 3m screen cage inside a green house. Females recently emerged from their chrysalis were placed in the cage with several males taken from larger rearing populations. The cage was checked for coupled pairs approximately every 30 min. Coupled butterflies were placed in individual plastic boxes until they separated, after which males were discarded and the spermatophore was immediately dissected out of the female's bursa copulatrix. Dissections were performed in ice-cold insect Ringer's solution. A total of 12 spermatophores were homogenized in a single microfuge tube containing 75 µl cold Phosphate Buffered Saline solution and centrifuged at 4°C for 15 minutes at 13,000 rpm. The resulting supernatant was stored at -80°C and sent to the Genome BC Proteomics Centre (University of Victoria, Canada) for two-dimensional liquid chromatography tandem mass-spectrometry (2d LC/MS) proteomic analysis. Initial separation of the spermatophore protein sample was performed with strong cation exchange (SCX) high performance liquid chromatography (HPLC). These SCX fractions were then analyzed on a Hybrid Quadrupole-TOF LC/MS/MS Mass Spectrometer (*QStar Pulsar I*, Applied Biosystems, Foster City, CA) with data acquired automatically using the Analyst QS 1.0 software (ABI MDS SCIEX, Concord, Canada). The resulting spectra were searched using the MASCOT 2.0 software (Matrix Science, Boston, MA) against

a protein database generated from *Heliconius* unigene sequences. The protein database, created using custom Perl scripts, consisted of all ORFs ≥ 10 amino acids long from all three forward reading frames from the combined *H. erato* and *H. melpomene* accessory gland and wing unigenes. It contained approximately 180,000 protein sequences derived from *Heliconius* unigenes as well as likely contaminants: pig trypsin and human keratin.

Results and Discussion

Library construction and EST assembly

The accessory glands from 11 adult male *H. erato* and 10 adult male *H. melpomene* were dissected from live butterflies and pooled within species to generate two tissue-specific directional cDNA libraries. The *H. erato* and *H. melpomene* libraries contained $7x10^6$ and $1.3x10^5$ colony-forming units, respectively. To enrich for transcripts expressed primarily in male tissue, both libraries were screened with cDNA generated from conspecific female abdominal tissue, and only non-hybridizing clones were sequenced. About 1150 clones were sequenced from each species to generate a collection of ESTs which were trimmed of low quality reads and poly-A tails, clustered, and assembled into contigs. We presume these assembled clusters, or unique gene objects (*unigenes*), represent distinct transcripts. Results were very similar in both species. *H. erato* yielded 371 unigenes and *H. melpomene* yielded 340. The two libraries were comparable in number of high quality ESTs, average read length, and the frequency spectrum of ESTs per unigene (Tables 1.1 and 1.2). In both libraries the vast majority of unigenes were represented by a single EST and ~90% of unigenes corresponded to 3 or fewer ESTs (Table 1.2).

Table 1.1. Summary of EST, BLAST, and SignalP analyses from *H. erato* and *H. melpomene* male accessory gland cDNA libraries.

	Heliconius erato		*Heliconius melpomene*	
EST results				
Number of clones sequenced	1152		1148	
High Quality ESTs [1]	936	(81%)	1033	(89%)
Number of unigenes	371		340	
Average sequence length (base pairs)	641		597	
BLAST results[2]				
Unigenes with significant BLAST hits to GenBank protein or nucleotides[3] (E-value < 10^{-5})	257	(69%)	235	(69%)
Unigenes with significant BLASTX hits to GenBank proteins (E-value < 10^{-5})	216	(58%)	218	(64%)
Unigenes with significant BLASTN hits to GenBank nucleotides (E-value < 10^{-5})	151	(40%)	150	(44%)
SignalP results[2]				
Unigenes with predicted signal peptides[4]	86	(24%)	92	(28%)

[1]ESTs > 100 bp after trimming raw sequences of vector sequence, low-scoring base calls, and poly-A tails. [2]Percentages given in reference to total number of unigenes. [3]Includes results from both BLASTN and BLASTX searches. [4]We required a positive result from both hidden markov models and neural network methods implemented in SignalP.

Table 1.2. Frequency distribution of ESTs per unigene from *H. erato* and *H. melpomene* male accessory gland cDNA libraries.

ESTs per unigene	*H. erato*	*H. melpomene*
1 (singletons)	319	256
2	13	31
3	5	12
4	9	10
5-10	14	17
11-20	7	9
21-50	3	3
>51	1	2
Total:	371	340

Unigene Annotation

We annotated the unigenes using BLASTX and BLASTN to search for similar sequences in GenBank's protein and nucleotide non-redundant databases; a significance cut-off of E-value < 10^{-5} was used for both searches. Results are summarized in Table 1.1.

Overall, 69% of unigenes in each species yielded significant BLAST hits to GenBank sequences (*H. erato*: 257 of 371 unigenes; *H. melpomene*: 235 of 340). This suggests nearly a third of the unigenes obtained from each species may correspond to novel and previously undescribed genes. In both species many unigenes with significant BLASTN hits to GenBank lacked significant BLASTX hits (*H. erato*: 41 unigenes, 11%; *H. melpomene*: 17 unigenes, 5%). These discrepancies could be explained in one of two ways: 1) these unigenes corresponded to ribosomal RNA or 2) these unigenes contained a *Heliconius* specific novel repetitive element and were similar only to a few other *Heliconius* sequences in GenBank containing such repeats (Papa *et al.* 2008) (see section below: *Novel Heliconius repetitive elements*; unigene sequences were not masked for GenBank BLASTs). Nineteen unigene pairs were reciprocal best-BLAST-hits between *H. erato* and *H. melpomene* and also showed no significant similarity to sequences in GenBank.

We used SignalP to identify protein-coding unigenes containing a predicted signal peptide sequence (Nielsen *et al.* 1997). ACPs are extracellularly secreted proteins and are therefore expected to have a signal peptide (Swanson *et al.* 2001, Braswell *et al.* 2006). Results were again similar between libraries, with 86 (24.4%) secreted proteins in *H. erato* and 92 (27.8%) in *H. melpomene* (Table 1.1).

Gene Ontology

Where possible, we assigned Gene Ontology (GO) annotations to protein-coding unigenes using the Annot8r application in the PartiGene software package (Ashburner *et al.* 2000, Parkinson *et al.* 2004). Annot8r assigns GO terms to unigenes based on BLASTX similarity (E-value < 10^{-5}) to proteins with known GO annotations; results are summarized via GO-slim terms corresponding to broad functional classes. GO annotations fall into three independent categories (Biological Process, Molecular Function, and Cellular Component) and a single sequence may be annotated in any or all categories. Moreover, a single sequence may be associated with multiple GO annotations within a single category, giving rise to more GO-annotations than sequences annotated (Table 1.3).

Table 1.3. Counts of GO-slim annotations from *H. erato* and *H. melpomene* accessory gland unigenes broken down by ontological category. Unigenes could be assigned more than one annotation both within and between categories.

	H. erato		*H. melpomene*	
	GO-slim Annotations	Unigenes Annotated[1]	GO-slim Annotations	Unigenes Annotated[1]
Molecular Function	225	169 (46%)	243	183 (54%)
Biological Process	157	126 (34%)	177	151 (44%)
Cellular Component	84	82 (23%)	122	121 (36%)
Total	466	187 (50%)	542	203 (60%)

[1] Percentages given in reference to total number of unigenes.

Overall we assigned GO annotations to 187 (50%) and 203 (60%) unigenes from *H. erato* and *H. melpomene*, respectively. With one exception, the distribution

of annotations across GO-slim summary terms is quite similar between the two species for all three GO categories (Figures 1.1-1.3). The one exception is the class "structural molecule activity" in the category Molecular Function (Figure 1.1). The proportion of *H. melpomene* annotations in this class is twice that in *H. erato*. This discrepancy clearly results from a greater number of ribosomal proteins represented in the *H. melpomene* ESTs, although it is unclear whether this reflects any biologically significant difference between the two species.

Novel Heliconius *repetitive elements*

Recently, Papa *et al*. identified nine novel, short (200-600 bp) *Heliconius* specific repetitive elements in BAC sequences from *H. melpomene* and *H. erato* (Papa *et al.* 2008). We used RepeatMasker to identify and mask these repetitive elements in both accessory gland and wing unigenes (Smit *et al.* 2004). Overall, each of the nine repeats were identified among the unigenes, but not all were present in each library (Table 1.4).

As reported by Papa *et al.*, repeat #7 was by far the most abundant and was significantly more common in *H. erato* (detected in 4.3% of unigenes) than *H. melpomene* (2.2%; one-tailed test of proportions, $p<.001$). To better characterize the nature of these *Heliconius* repeats we further examined repeat #7. Instances of this repeat typically fell outside ORF predictions from the PartiGene software, suggesting that when present in transcribed sequence it occurs in 3' or 5' untranslated regions of genes, not in the coding sequence. For one *H. erato* accessory gland unigene (Her00086), three of five ESTs lacked the repeat sequence; this indicates that for at least one locus the repeat motif is polymorphic (i.e. present/absent) among individuals pooled for library construction. Finally, BLAST searches in GenBank revealed repeat #7 was present in the introns of two additional *Heliconius* species (*H. doris*, *mannose*

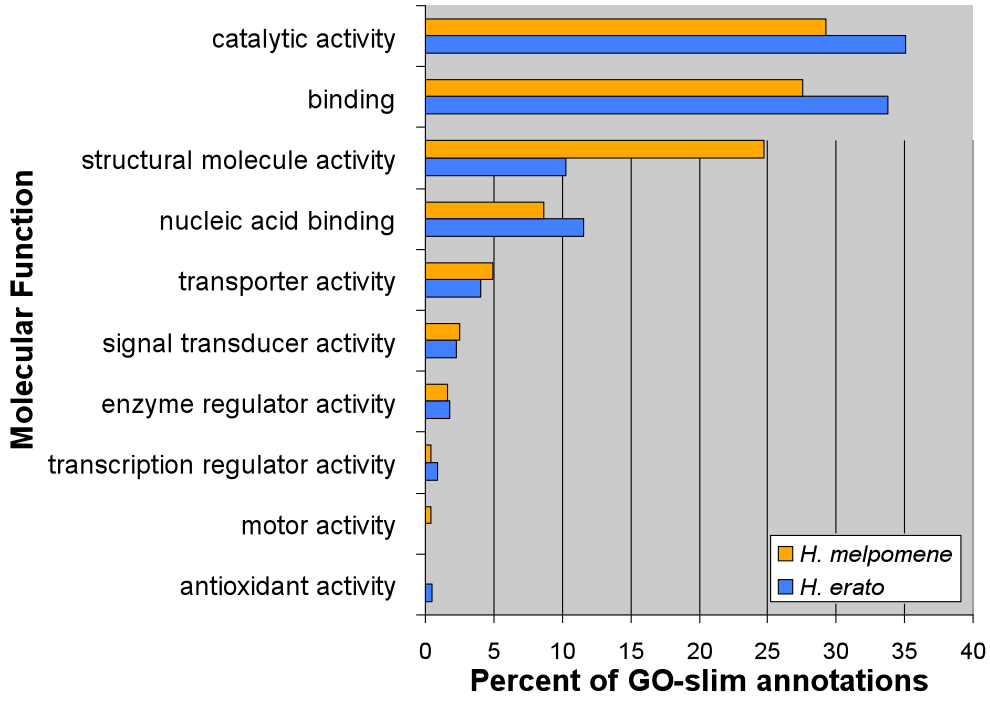

Figure 1.1. Molecular function GO-slim annotations from *H. melpomene* and *H. erato* accessory gland unigenes. Percentages are in reference to total molecular function GO-slim annotations. Not all unigenes could be annotated and some received multiple annotations.

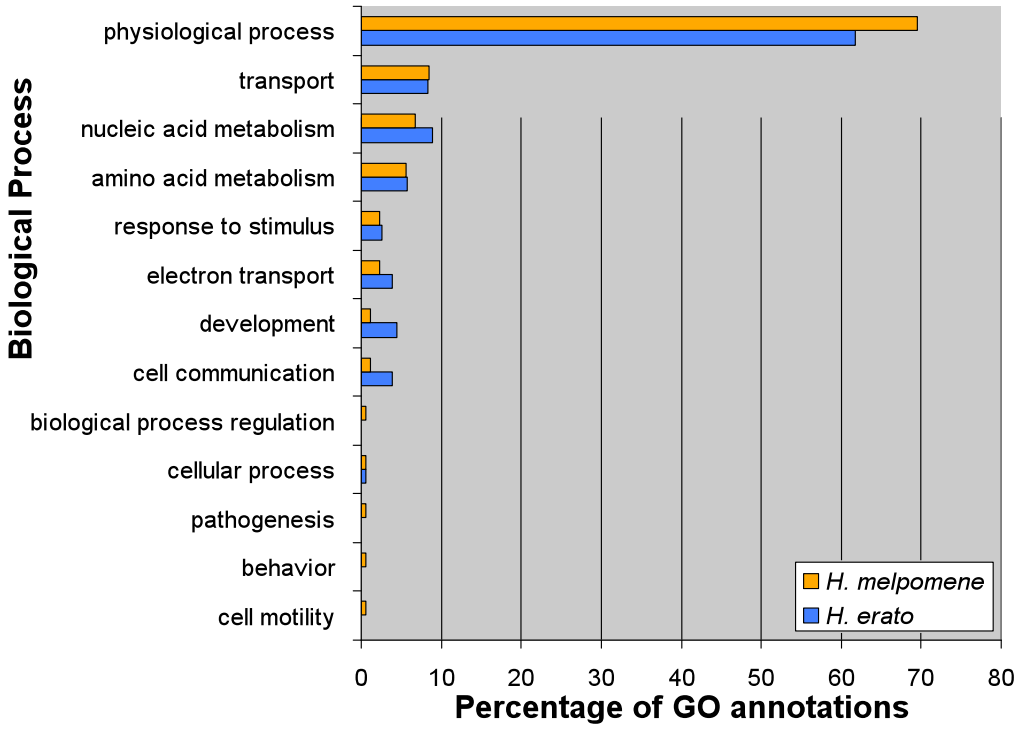

Figure 1.2. Biological process GO-slim annotations from *H. melpomene* and *H. erato* accessory gland unigenes. Percentages are in reference to total biological process GO-slim annotations. Not all unigenes could be annotated and some received multiple annotations.

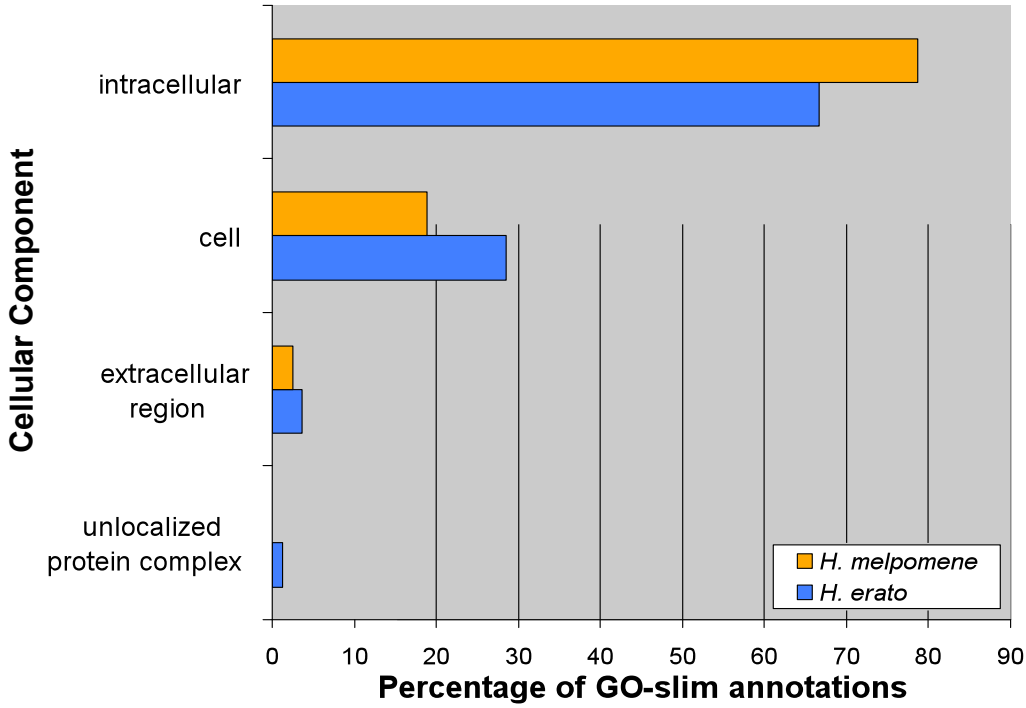

Figure 1.3. Cellular component GO-slim annotations from *H. melpomene* and *H. erato* accessory gland unigenes. Percentages are in reference to total cellular component GO-slim annotations. Not all unigenes could be annotated and some received multiple annotations.

Table 1.4. Counts of repetitive elements masked among unigenes from *H. erato* and *H. melpomene* accessory glands and developing wing tissue libraries. Repetitive element labeling numbers correspond to those used in Papa *et al*. 2008.

	Heliconius Repetitive Elements								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
H. erato									
Accessory Glands	0	2	1	3	0	0	26	3	0
Wing	5	10	19	13	3	5	275	40	1
total	5	12	20	16	3	5	301	43	1
H. melpomene									
Accessory Glands	0	0	1	1	0	1	8	1	0
Wing	0	4	0	2	1	1	41	6	1
Total	0	4	1	3	1	1	49	7	1

phosphate isomerase, [GenBank:AF413748]; *H. himera*, *dopa decarboxylase I* [GenBank:AY437779]). These results are consistent with the interpretation of Papa *et al.* that these repetitive elements likely arise from the replication and insertion of transposable elements that are common among *Heliconius* butterflies.

These repeats present a practical problem when using BLAST to identify homologous unigenes within and between *Heliconius* species. Such searches may generate significant alignment scores between unigenes either because the transcribed genes are truly homologous or because a repetitive element occurs in both transcripts. We therefore used the masked unigenes for all BLAST searches between *H. erato* and *H. melpomene* libraries. We assume that significant similarity scores produced using these masked unigenes indicate homologous transcripts and not spurious similarity due to sharing of repetitive sequence.

Comparisons of Accessory Gland and Wing Libraries

We used the criterion of high-scoring reciprocal best BLAST hits (RBBH; E-value < 10^{-10}) to explore overlaps in the transcripts sampled from accessory gland and wing libraries (Figure 1.4). For this analysis we assume that a highly significant RBBH between unigenes indicates that these transcripts originate from the same locus (within species) or orthologous loci (between species). However, we fully recognize that such questions of identity and orthology can only be conclusively determined in the context of complete genome sequences and that the modest number of ESTs generated for some of these libraries hardly represents an exhaustive profiling of the tissue's transcriptome. Nonetheless, contrasting the overlap in ESTs sampled between species or tissues is useful for identifying qualitative differences and similarities in the results. For instance, comparing wing and accessory gland RBBHs between species as

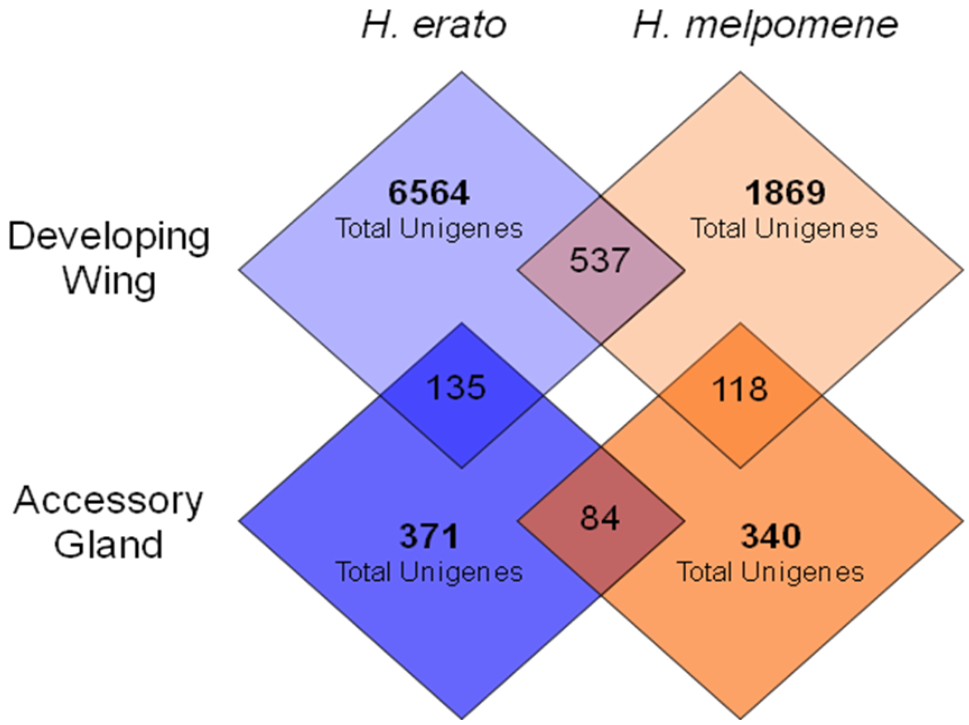

Figure 1.4. Counts of unigenes found in common between male accessory gland and developing wing cDNA libraries. Numbers in overlapping areas of diamonds are counts of reciprocal best BLAST hits (E-value < 10^{-10}) between unigenes from each library. Numbers in non-overlapping areas of diamonds are the total number of unigenes obtained from each library.

a percentage of total *H. melpomene* unigenes yields similar results: ESTs sampled from accessory glands (84/340 = 24.7%) and from wing (537/1869 = 28.7%) show similar proportions of unigenes shared between species. We used *H. melpomene* unigenes as the denominator in this comparison because *H. melpomene* has fewer ESTs sampled from both tissues, which we assume is a limiting factor in identifying RBBHs. These results suggest consistency between libraries, but must be interpreted with caution due to differences in library construction and EST sampling. We

anticipate that the numbers of genes identified in common between these species will increase dramatically as more ESTs become available.

A useful comparison can also be made between tissues within species. Genes common among accessory gland ESTs but absent from wing ESTs are promising candidates for encoding seminal fluid proteins. Encouragingly, the results from both species clearly indicated that a majority of transcripts sampled from the accessory glands were not present in wing tissue unigenes (Figure 1.4). In *H. erato* only 141 accessory gland unigenes had high-scoring BLAST hits (E-value < 10^{-10}) to wing unigenes, of which 135 were RBBHs. In *H. melpomene* there were 125 high-scoring BLAST hits to wing unigenes, with 117 RBBHs. Therefore in both species about 65% of accessory gland unigenes were not found among transcripts sampled from wing tissue.

Considered broadly, this lack of overlap indicates that the set of genes sampled from accessory glands is qualitatively different from the set of genes previously sampled from developing wing tissues. The wing libraries were not screened or subtracted and were well-sampled (*H. erato*, 17,573 ESTs; *H. melpomene,* 4,976 ESTs) (Papanicolaou *et al.* 2008), so this discrepancy in genes sampled from the two tissues likely reflects two phenomena, one biological and one methodological. Biologically, it might be that patterns of gene expression are quite different between these two tissues; the differences in sampled genes therefore may reflect substantial differences in transcript abundances. However, verifying this would require much deeper EST sampling and a methodologically consistent approach for profiling transcripts (e.g. microarrays). But apart from any underlying biological differences, our sampling method was also explicitly biased: we probed our libraries with female cDNA and sequenced non-hybridizing clones in order to enrich our ESTs for male-specific transcripts. Although we do not have unbiased samples for comparison, the

relatively low and stable proportion of unigenes shared between wing and accessory gland ESTs suggests that the enrichment for male-specific transcripts was at least moderately successful and that the resulting accessory gland ESTs will prove to be a useful resource for identifying seminal fluid proteins. Nonetheless, the enrichment process was clearly not perfect. For instance, many accessory gland unigenes showed highly significant BLAST hits to well-known 'housekeeping' genes which presumably exhibit little differential expression between sexes (e.g. c*ystathionine beta-synthase*, *elongation factor 1-α*, ribosomal proteins, and ribosomal RNAs).

Highly abundant transcripts are repetitive, secreted proteins

Both the *H. erato* and *H. melpomene* libraries contained a few unusually abundant transcripts (i.e. >20 ESTs per unigene, Maximum: *H. erato*, 168; *H. melpomene*, 133; Table 1.2). BLAST searches within libraries revealed that these abundant transcripts were highly similar to several other transcripts found at lower abundances. Overall there were three such groups each composed of about ten unigenes. Each group encoded highly repetitive proteins, two with a repeat structure rich in tyrosine and one rich in asparagine; all had a predicted signal peptide (Figure 1.5). These same three groups were identified in both species and sequences within groups were clearly similar between species. Similarities were also evident in the repeat structure present in the two tyrosine-rich groups. Unfortunately, extensive indel variation and the repetitive nature of these sequences precluded reliable alignments among any of these unigenes. Therefore robust inferences of homology between these sequences were not possible either within groups or between species.

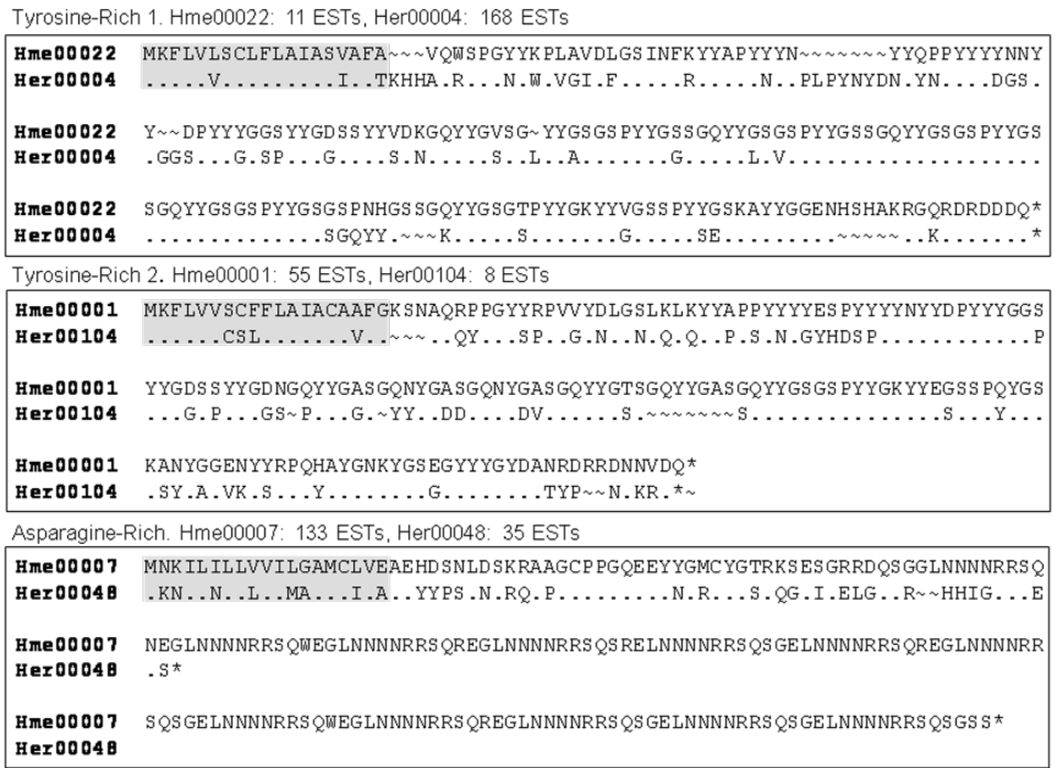

Figure 1.5. Amino acid alignments between *H. melpomene* and *H. erato* of candidate spermatophore proteins. These sequences are from the most abundant transcripts in each of three groups of highly abundant, repetitive proteins observed in the male accessory gland cDNA libraries. Dots (.) indicate identity between sequences; tildes (~) represent alignment gaps; filled boxes denote predicted signal sequences.

It is worth noting that the *H. erato* transcripts from the third (asparagine-rich) group exhibit some exceptions to the patterns uniting these groups of proteins. First, they completely lacked the repetitive asparagine-rich C-terminus motif that characterizes their *H. melpomene* counterparts (Figure 1.5). Nonetheless, the *H. erato* transcripts were clearly homologous to the non-repetitive N-termini of the *H. melpomene* sequences. Second, there were only two unigenes in this *H. erato* group while the other groups contained around ten unigenes. Nonetheless, one of these two unigenes, Her00048, was comprised of 35 ESTs and was the third most abundant transcript sampled from that library.

Sequences from these three groups of transcripts did exhibit some weak but significant similarity to sequences and protein domains in public databases (determined via BLAST and InterProScan). However, after inspecting these results we concluded that these similarities did not reflect true homology. Rather, these significant scores arose spuriously from matches to the repetitive motifs found in these sequences. No similar sequences were found among wing unigenes.

Accessory gland ESTs facilitate identifying reproductive proteins

In insects, most work identifying seminal fluid proteins has focused on two major criteria: enriched expression in accessory glands and the presence of a computationally predicted signal peptide (Wolfner *et al.* 1997, Swanson *et al.* 2001). Genes (and their encoded proteins) meeting these two criteria are commonly called ACPs (accessory gland proteins) and early work in *Drosophila* using western blots generally supported the assumption that these proteins are transferred to females in seminal fluid (Herndon & Wolfner 1995, Bertram *et al.* 1996, Wolfner 1997). More recent proteomic studies have broadly confirmed this assumption but have also revealed that many genes encoding seminal fluid proteins show significant expression outside of accessory glands (Andres *et al.* 2008, Findlay *et al.* 2008, Sirot *et al.* 2008). In light of this precedent, we focused on the highly abundant tyrosine- and asparagine-rich transcripts to demonstrate the utility of our ESTs for identifying *Heliconius* ACPs and seminal fluid proteins.

The high abundance of these transcripts in the accessory gland make them obvious candidates for being ACPs encoding seminal fluid proteins. The presence of a signal peptide in all groups meets one of the major criteria for identifying insect ACPs. None of these sequences were found among ESTs generated from developing wing tissue in either species; this absence, contrasted with their abundance among

accessory gland ESTs, provides support for the criterion of accessory-gland biased expression. We further evaluated this criterion using reverse transcription PCR (RT-PCR) to amplify these transcripts from male and female abdomen and also male thorax. Species-specific primers were designed to fall in regions of robust alignment between all members of each group so that tests of tissue-specific patterns of expression were inclusive of all transcripts and were therefore conservative. We used primers designed for *α-tubulin* as a positive control. RT-PCR results were similar for all three groups of transcripts in both species: there was robust amplification from male abdomen but weak or no amplification from male thorax and female abdomen (Figure 1.6). In contrast, *α-tubulin* amplified robustly from all tissues. These three observations, 1) the presence of a predicted signal peptide, 2) the discrepancy in EST abundance between wing and accessory gland tissue, and 3) the tissue-specific patterns of expression, are consistent with these transcripts being ACP genes and suggest they encode seminal fluid proteins.

We used shotgun peptide sequencing (2d-LC/MS) to search for these candidate seminal fluid proteins in *H. erato* spermatophores, the proteinaceous packet containing sperm and seminal fluid transferred from males to females during copulation. In *Heliconius*, spermatophores can be easily and cleanly dissected from freshly mated females; we crushed the spermatophores in buffer, pelleted the remnants via centrifugation, and reserved the supernatant for analysis. Tandem mass spectra generated from this supernatant were searched against protein translations of the combined *H. erato* and *H. melpomene* accessory gland and wing unigenes. This search yielded a significant match ($p < .005$) to the Tyrosine-Rich 1 group of proteins from *H. erato* (Figure 1.5), which includes the single most abundant transcript (168 ESTs) sampled from the accessory glands. This result confirms that at least one of the three groups of transcripts encodes a seminal fluid protein. More generally, it

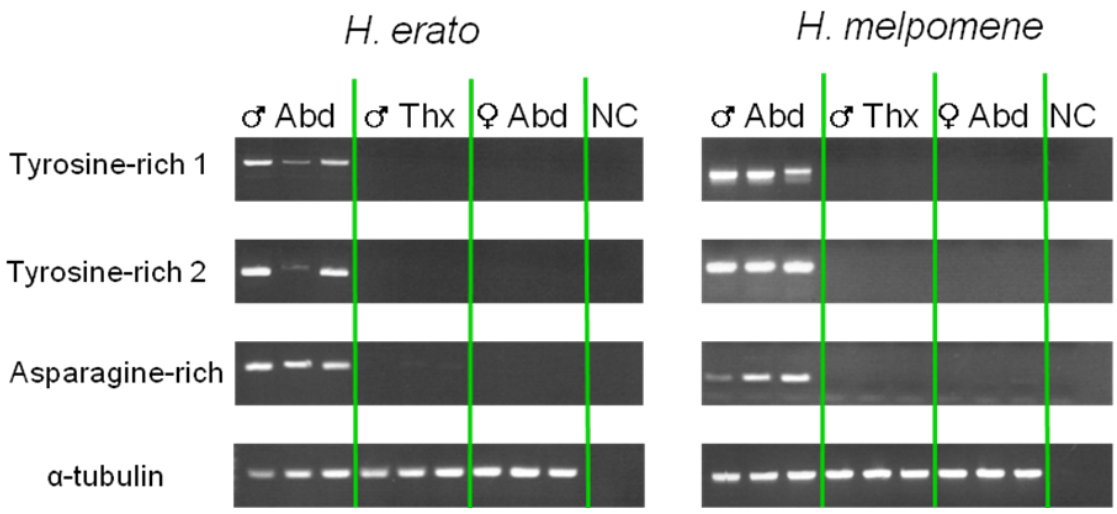

Figure 1.6. Tissue specific patterns of expression for candidate spermatophore proteins assayed via reverse transcription PCR. Patterns of expression for the three groups of highly abundant accessory gland transcripts and *α-tubulin* were assayed in three males (abdomen and thorax) and three females (abdomens only). PCR primers were designed to amplify universally from all transcripts in each of the three groups, two corresponding to tyrosine-rich proteins and one to asparagine-rich proteins. First-strand cDNA synthesized from equal concentrations of total RNA was used as template. For each primer set, equal amounts of PCR amplicon (ranging between 4 and 9 μl) were electrophoresed on 1.2% agarose gels. NC = negative control (no template added to PCR mix).

demonstrates the utility of these accessory gland ESTs as a resource for identifying seminal fluid proteins in *Heliconius* butterflies.

Unfortunately, due to the lack of similarity to other known sequences it is difficult to predict the molecular function of the Tyrosine-Rich 1 group of proteins or the two others which were not detected in the 2d-LC/MS experiment. However, we note the similarity between our results and other studies in crickets reporting abundant, hyper-variable, repetitive, secreted proteins with accessory gland biased expression and which were present in the spermatophore (Andres *et al.* 2006, Braswell *et al.* 2006, Andres *et al.* 2008). These authors speculate that the abundance and repetitive nature of these proteins suggest they are structural components of the spermatophore, which are generally known to be encoded by male insect accessory glands (Gillott 2003, Braswell *et al.* 2006). Although the tyrosine- and asparagine- rich *Heliconius* proteins do not appear to be homologous to the cricket proteins or another spermatophore protein reported in beetles (Paesen *et al.* 1992, Feng & Happ 1996), if these *Heliconius* proteins are structural components of the spermatophore it offers a possible explanation for the failure to detect the two additional groups in our proteomic assay. These proteins are unlikely to be water-soluble and the centrifugation step could have removed most of the spermatophore's structural components from the supernatant which was analyzed. Future work on the biochemical properties and structure of these three proteins will be informative in this matter, as will specifying where specifically in the spermatophore the Tyrosine-Rich 1 proteins are located. Alternatively, it may be that the two undetected proteins are not present in the spermatophore and are not seminal fluid proteins, in which case useful biological insights will likely arise from investigating this functional difference between these otherwise similar proteins. In either case, this combined approach using focused EST sequencing, *in silico* and *in vitro* expression assays, and proteomic

analyses has successfully identified novel and noteworthy *Heliconius* proteins for future research.

Conclusion

We report the successful sequencing of 936 ESTs, corresponding to 371 unigenes, and 1033 ESTs, corresponding to 340 unigenes, from the male accessory glands of *H. erato* and *H. melpomene*, respectively. Overall the results from the two species were very similar; our analyses did not reveal any obvious patterns that might reflect differences between the pupal and adult mating system. Approximately one-third of these unigenes showed no significant BLAST similarity to sequences in GenBank's non-redundant databases, indicating that a large proportion of novel genes are expressed in *Heliconius* male accessory glands. In both species only a third of accessory gland unigenes were also found among unigenes derived from wing tissue. About 25% of unigenes from both species encode secreted proteins. This includes three distinct groups of unigenes which consist of a few highly abundant transcripts and several additional similar but less-abundant transcripts all differentiated by extensive indel variation. Patterns of tissue-specific expression suggest that they are ACPs; proteomic analysis confirmed the presence of proteins from one of these groups in the spermatophore.

These EST sequences lay the foundation for future research investigating the patterns and process of molecular evolution among reproductive proteins in *Heliconius* butterflies. In particular, the striking dichotomy in mating systems offers a promising opportunity to explore the role of post-mating sexual selection in contributing to the rapid evolution of reproductive proteins. More generally, *Heliconius* butterflies are a remarkable system for investigating patterns of genetic diversity in the context of well-characterized ecological and phenotypic diversity. The

two species studied here are also the focal taxa for research examining the genetic basis of wing pattern diversity in *Heliconius*. Our results comprise the first major expansion of genomic-scale research into other aspects of *Heliconius* biology. They therefore mark a significant advance in the development of these species, and the *Heliconius* genus, as model systems for connecting various aspects of genomic, phenotypic, and ecological diversity.

REFERENCES

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Molecular Biology and Evolution **23**:1574-1584.

Andres, J. A., L. S. Maroja, and R. G. Harrison. 2008. Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets. Proceedings of the Royal Society B-Biological Sciences **275**:1975-1983.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat.Genet. **25**:25-29.

Bates, H. W. 1862. Contributions to an insect fauna of the Amazon Valley. Transactions of the Linnean Society **23**:495-566.

Beldade, P., W. O. McMillan, and A. Papanicolaou. 2008. Butterfly genomics eclosing. Heredity **100**:150-157.

Beltran, M., C. D. Jiggins, A. V. Z. Brower, E. Bermingham, and J. Mallet. 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.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology **340**:783-795.

Benson, W. W. 1972. Natural Selection for Mullerian Mimicry in Heliconius erato in Costa Rica. Science **176**:936-939.

Bertram, M. J., D. M. Neubaum, and M. F. Wolfner. 1996. Localization of the Drosophila male accessory gland protein Acp36DE in the mated female suggests a role in sperm storage. Insect Biochemistry and Molecular Biology **26**:971-980.

Birkhead, T. R., and T. Pizzari. 2002. Postcopulatory sexual selection. Nature Reviews Genetics **3**:262-273.

Braswell, W. E., J. A. Andres, L. S. Maroja, R. G. Harrison, D. J. Howard, and W. J. Swanson. 2006. Identification and comparative analysis of accessory gland proteins in Orthoptera. Genome **49**:1069-1080.

Brower, A. V. 1994. Rapid morphological radiation and convergence among races of the butterfly Heliconius erato inferred from patterns of mitochondrial DNA evolution. Proc.Natl.Acad.Sci.U.S.A **91**:6491-6495.

Brower, A. V. Z. 1997. The evolution of ecologically important characters in Heliconius butterflies (Lepidoptera: Nymphalidae): A cladistic review. Zoological Journal of the Linnean Society **119**:457-472.

Brown, K. S. 1981. The Biology of *Heliconius* and related genera. Annual Review of Entomology **26**:427-456.

Chenchik, A., Y. Y. Zhu, L. Diatchenko, R. Li, J. Hill, and P. D. Siebert. 1998. Generation and use of high quality cDNA from small amounts of total RNA by SMART PCR. *in* P. D. Siebert, and J. W. Larrick editors. Gene Cloning and Analysis by RT-PCR. BioTechniques Books.

Civetta, A., and R. S. Singh. 1998. Sex-related genes, directional sexual selection, and speciation. Molecular Biology and Evolution **15**:901-909.

Clark, A. G., M. Aguade, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in Sperm Displacement and Its Association with Accessory-Gland Protein Loci in *Drosophila melanogaster* . Genetics **139**:189-201.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

Cutter, A. D., J. D. Wasmuth, and M. L. Blaxter. 2006. The evolution of biased codon and amino acid usage in nematode genomes. Mol.Biol.Evol. **23**:2303-2315.

Deinert, E. I., J. T. Longino, and L. E. Gilbert. 1994. Mate Competition in Butterflies. Nature **370**:23-24.

Dorus, S., P. D. Evans, G. J. Wyckoff, S. S. Choi, and B. T. Lahn. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nature Genetics **36**:1326-1329.

Feder, M. E., and T. Mitchell-Olds. 2003. Evolutionary and ecological functional genomics. Nat.Rev.Genet. **4**:651-657.

Feng, X., and G. M. Happ. 1996. Isolation and sequencing of the gene encoding Sp23, a structural protein of spermatophore of the mealworm beetle, Tenebrio molitor. Gene **179**:257-262.

Findlay, G. D., X. H. Yi, M. J. MacCoss, and W. J. Swanson. 2008. Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. Plos Biology **6**:1417-1426.

Fiumera, A. C., B. L. Dumont, and A. G. Clark. 2005. Sperm competitive ability in Drosophila melanogaster associated with variation in male reproductive proteins. Genetics **169**:243-257.

Flanagan, N. S., A. Tobler, A. Davison, O. G. Pybus, D. D. Kapan, S. Planas, M. Linares, D. Heckel, and W. O. McMillan. 2004. Historical demography of Mullerian mimicry in the neotropical Heliconius butterflies. Proc.Natl.Acad.Sci.U.S.A **101**:9704-9709.

Galindo, B. E., V. D. Vacquier, and W. J. Swanson. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proceedings of the National Academy of Sciences of the United States of America **100**:4639-4643.

Gilbert, L. E. 1991. Biodiversity of a Central American Heliconius community: pattern, process, and problems. Pages 403-427 *in* P. W. Price, T. M. Lewinsohn, G. W. Fernandes, and W. W. Benson editors. Plant-Animal Interactions Evolutionary Ecology in Tropical and Temperate Regions. John Wiley and Sons.

Gilbert, L. E. 1972. Pollen Feeding and Reproductive Biology of *Heliconius* Butterflies. Proceedings of the National Academy of Sciences of the United States of America **69**:1403-&.

Gillott, C. 2003. Male accessory gland secretions: Modulators of female reproductive physiology and behavior. Annual Review of Entomology **48**:163-184.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. R. Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh. 2007. Evolution in the fast lane: Rapidly evolving sex-related genes in drosophila. Genetics **177**:1321-1335.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series **41**:95-98.

Herlyn, H., and H. Zischler. 2007. Sequence evolution of the sperm ligand zonadhesin correlates negatively with body weight dimorphism in primates. Evolution Int.J.Org.Evolution **61**:289-298.

Herndon, L. A., and M. F. Wolfner. 1995. A Drosophila Seminal Fluid Protein, Acp26Aa, Stimulates Egg-Laying in Females for 1 Day After Mating. Proceedings of the National Academy of Sciences of the United States of America **92**:10114-10118.

Jiggins, C. D., S. Baxter, W. O. McMillan, N. Chamberlain, and R. ffrench-Constant. 2008. Prospects for locating genes of interest in lepidopteran genomes: A case study of butterfly colour patterns. *in* M. R. Goldsmith, and F. Marec editors. Lepidopteran Molecular Biology and Genetics. CRC.

Jiggins, C. D., J. Mavarez, M. Beltran, W. O. McMillan, J. S. Johnston, and E. Bermingham. 2005. A Genetic Linkage Map of the Mimetic Butterfly Heliconius melpomene. Genetics **171**:557-570.

Jiggins, C. D., and W. O. McMillan. 1997. The genetic basis of an adaptive radiation: warning colour in two Heliconius species. Proceedings of the Royal Society of London Series B-Biological Sciences **264**:1167-1175.

Jiggins, C. D., R. E. Naisbit, R. L. Coe, and J. Mallet. 2001. Reproductive isolation caused by colour pattern mimicry. Nature **411**:302-305.

Joron, M., C. D. Jiggins, A. Papanicolaou, and W. O. McMillan. 2006. Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity **97**:157-167.

Kapan, D. D. 2001. Three-butterfly system provides a field test of mullerian mimicry. Nature **409**:338-340.

Kapan, D. D., N. S. Flanagan, A. Tobler, R. Papa, R. D. Reed, J. A. Gonzalez, M. R. Restrepo, L. Martinez, K. Maldonado, C. Ritschoff, D. G. Heckel, and W. O. McMillan. 2006. Localization of Mullerian mimicry genes on a dense linkage map of Heliconius erato. Genetics **173**:735-757.

Kronforst, M. R., L. G. Young, and L. E. Gilbert. 2007. Reinforcement of mate preference among hybridizing Heliconius butterflies. Journal of Evolutionary Biology **20**:278-285.

Lottaz, C., C. Iseli, C. V. Jongeneel, and P. Bucher. 2003. Modeling sequencing errors by combining Hidden Markov models. Bioinformatics **19**:ii103-ii112.

Mallet, J. 1989. The genetics of warning color in Peruvian hybrid zones of *Heliconius erato* and *Heliconius melpomene*. Proceedings of the Royal Society of London Series B-Biological Sciences **236**:163-&.

Markow, T. A., and P. M. O'Grady. 2007. Drosophila biology in the genomic age. Genetics **177**:1269-1276.

Mitchell-Olds, T., M. Feder, and G. Wray. 2008. Evolutionary and ecological functional genomics. Heredity **100**:101-102.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. vonHeijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering **10**:1-6.

Nijhout, H. F., G. A. Wray, and L. E. Gilbert. 1990. An Analysis of the Phenotypic Effects of Certain Color Pattern Genes in Heliconius (Lepidoptera, Nymphalidae). Biological Journal of the Linnean Society **40**:357-372.

Paesen, G. C., M. B. Schwartz, M. Peferoen, F. Weyda, and G. M. Happ. 1992. Amino acid sequence of Sp23, a structural protein of the spermatophore of the mealworm beetle, Tenebrio molitor. J.Biol.Chem. **267**:18852-18857.

Panhuis, T. M., N. L. Clark, and W. J. Swanson. 2006. Rapid evolution of reproductive proteins in abalone and Drosophila. Philos.Trans.R.Soc.Lond B Biol.Sci. **361**:261-268.

Papa, R., C. M. Morrison, J. R. Walters, B. A. Counterman, R. Chen, G. Halder, L. Ferguson, N. Chamberlain, R. Ffrench-Constant, D. D. Kapan, C. D. Jiggins, R. D. Reed, and W. O. McMillan. 2008. Highly conserved gene order and numerous novel repetitive elements in genomic regions linked to wing pattern variation in Heliconius butterflies. BMC Genomics **9**:345.

Papanicolaou, A., M. Joron, W. O. McMillan, M. L. Blaxter, and C. D. Jiggins. 2005. Genomic tools and cDNA derived markers for butterflies. Molecular Ecology **14**:2883-2897.

Papanicolaou, A., S. Gebauer-Jung, M. L. Blaxter, W. Owen McMillan, and C. D. Jiggins. 2008. ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Research **36**:D582-D587.

Parkinson, J., A. Anthony, J. Wasmuth, R. Schmid, A. Hedley, and M. Blaxter. 2004. PartiGene - constructing partial genomes. Bioinformatics **20**:1398-1404.

Ram, K. R., and M. F. Wolfner. 2007. Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology **47**:427-445.

Reed, R. D., W. O. McMillan, and L. M. Nagy. 2008. Gene expression underlying adaptive variation in Heliconius wing patterns: non-modular regulation of overlapping cinnabar and vermilion prepatterns. Proc.Biol.Sci. **275**:37-45.

Rozen, S., and Skaletsky H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. Pages 365-386 *in* S. M. S. Krawetz editor.

Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Shimizu, K. K., and M. D. Purugganan. 2005. Evolutionary and ecological genomics of Arabidopsis. Plant Physiol **138**:578-584.

Sirot, L. K., R. L. Poulson, M. C. McKenna, H. Girnary, M. F. Wolfner, and L. C. Harrington. 2008. Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, Aedes aegypti: Potential tools for control of female feeding and reproduction. Insect Biochemistry and Molecular Biology **38**:176-189.

Smit A.F.A., Hubley R. & Green P. RepeatMasker Open-3.0. 2004. Ref Type: Computer Program

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proceedings of the National Academy of Sciences of the United States of America **98**:7375-7379.

Swanson, W. J., and V. D. Vacquier. 2002a. Reproductive protein evolution. Annual Review of Ecology and Systematics **33**:161-179.

Swanson, W. J., and V. D. Vacquier. 2002b. The rapid evolution of reproductive proteins. Nature Reviews Genetics **3**:137-144.

Torgerson, D. G., R. J. Kulathinal, and R. S. Singh. 2002. Mammalian sperm proteins are rapidly evolving: Evidence of positive selection in functionally diverse genes. Molecular Biology and Evolution **19**:1973-1980.

Wasmuth, J. D., and M. L. Blaxter. 2004. Prot4EST: Translating Expressed Sequence Tags from neglected genomes. Bmc Bioinformatics **5**.

Wolfner, M. F. 1997. Tokens of love: Functions and regulation of Drosophila male accessory gland products. Insect Biochemistry and Molecular Biology **27**:179-192.

Wolfner, M. F., H. A. Harada, M. J. Bertram, T. J. Stelick, K. W. Kraus, J. M. Kalb, Y. O. Lung, D. M. Neubaum, M. Park, and U. Tram. 1997. New genes for male accessory gland proteins in Drosophila melanogaster. Insect Biochemistry and Molecular Biology **27**:825-834.

CHAPTER 2

COMBINED EST AND PROTEOMIC ANALYSIS IDENTIFIES RAPIDLY EVOLVING SEMINAL FLUID PROTEINS IN *HELICONIUS* BUTTERFLIES

Introduction

The introduction of Kimura's (1968) controversial theory of neutral molecular evolution coincided with the earliest uses of molecular genetics in evolutionary biology (Lewontin & Hubby 1966, Kimura 1968, Li 1997). This revolution in both theory and data fundamentally altered previous assumptions about how and when selection causes evolutionary change at the molecular level. Since that time, vast amounts of data have accumulated that support the major tenets of the neutral theory: rates of molecular evolution are relatively constant, adaptive molecular evolution is rare, and most variation both within and between species results from the random fixation (genetic drift) of neutral mutations (Fay & Wu 2003, Nei 2005). These data have also yielded notable exceptions to this general trend; some groups of genes evolve unusually rapidly, often due to positive selection. (Holmes 2004, Bustamante *et al.* 2005, Nielsen *et al.* 2005, Roth *et al.* 2005). However, for a few of these groups, the cataloging of evolutionary patterns has far outpaced our ability to confidently identify the underlying causes (*e.g.* natural vs. sexual selection; positive selection vs. relaxed constraint). The molecular evolution of reproductive proteins provides a clear example. These proteins diverge rapidly and often adaptively, but data directly addressing a cause for this pattern are scarce (Swanson & Vacquier 2002a, Swanson & Vacquier 2002b, Clark *et al.* 2006, Turner & Hoekstra 2008).

A reproductive protein is any protein acting after copulation to mediate gamete usage, gamete storage, signal transduction, or fertilization (Swanson & Vacquier 2002b). Because mutations affecting these processes should have profound fitness

consequences, the rapid evolution of reproductive proteins may at first seem difficult to explain. Given the large potential for deleterious consequences of mutations in reproductive proteins, why do we not observe a pattern of extreme constraint? One possible explanation is that sexual selection in promiscuous mating systems will rapidly fix advantageous mutations. For this reason, studies documenting rapid, adaptive reproductive protein evolution routinely invoke sexual selection in explaining this observation (Civetta & Singh 1998, Swanson *et al.* 2001, Swanson & Vacquier 2002b, Galindo *et al.* 2003, Dorus *et al.* 2004, Haerty *et al.* 2007). However, despite the popularity of this hypothesis, empirical support is limited and several reasonable alternative hypotheses exist (Swanson & Vacquier 2002a). This discrepancy arises, in part, because most well-characterized examples of reproductive protein evolution come from only a few model species of primates, rodents, and flies in the genus *Drosophila* (Turner & Hoekstra 2008). This relatively narrow taxonomic sampling confines the use of comparative studies to address the role of sexual selection in reproductive protein evolution.

The seminal fluid proteins (SFPs) *Drosophila melanogaster* and its close relatives are a classic example of rapid, adaptive reproductive protein evolution. Electrophoretic studies on the earliest identified SFPs revealed them to be on average more divergent than other proteins examined (Thomas & Singh 1992, Civetta & Singh 1995). Contemporary analyses using DNA sequence data have confirmed this pattern across much larger samples of SFP loci (Swanson *et al.* 2001, Mueller *et al.* 2005, Haerty *et al.* 2007). Furthermore, these studies revealed that rates of nonsynonymous substitution per site (resulting in an amino acid change; d_N) often exceeded those of synonymous substitution per site (silent change; d_S). A d_N/d_S ratio (symbolized as ω) greater than one is evidence of adaptive evolution; it is the expected pattern if most mutations are fixed because they increase fitness (Yang & Bielawski 2000).

Unfortunately, the mating systems of *D. melanogaster* and its close relatives are homogeneous and offer little opportunity to compare SFP evolution under different regimes of sexual selection (Markow 2002). Intriguingly, descriptions of SFPs in the more promiscuous cactophilic *Drosophila* do suggest elevated rates of evolution relative to the *melanogaster* group, though this has not been formally tested (Wagstaff & Begun 2005, Wagstaff & Begun 2007, Almeida & DeSalle 2008, Almeida & DeSalle 2009).

Efforts to address this issue in mammals have been less limited by opportunities for informative comparisons. There are now several studies in mammals suggesting or formally inferring positive correlations between the evolutionary rate of reproductive proteins and intensity of sexual selection (Kingan *et al.* 2003, Jensen-Seaman & Li 2003, Dorus *et al.* 2004, Clark & Swanson 2005, Herlyn & Zischler 2007, Ramm *et al.* 2008). However, these studies have been exclusively confined to primates and murine rodents and typically focus on only one or a few loci, limiting the generality of the conclusions.

These examples highlight the need to expand sampling of reproductive proteins into additional taxa offering informative contrasts between mating systems. They also emphasize the importance of including multiple reproductive proteins in such assays in order to accurately assess the strength of correlations between evoultionary rate and sexual selection. Meeting these research objectives requires efficiently identifying reproductive proteins *de novo* in the absence of a fully sequenced genome. Recent work in several genetic and agricultural model systems (*e.g.* fruit fly, honey bee, cow, human, *etc.*) now indicates two promising and complementary approaches for identifying reproductive proteins more broadly. The first approach involves generating expressed sequence tags (ESTs) from reproductive tissues and using patterns of tissue-specific expression combined with bioinformatic

annotations as criteria to identify putative reproductive proteins. For instance, loci where expression is biased towards male reproductive tissues and where the encoded protein possess a predicted signal peptide (indicating extracellular secretion) are extremely likely to be SFPs. (Swanson *et al.* 2001, Swanson *et al.* 2004, Wagstaff & Begun 2005, Andres *et al.* 2006, Braswell *et al.* 2006, Davies & Chapman 2006, Walters & Harrison 2008, Almeida & DeSalle 2009). This approach is inherently indirect; inference of a protein's involvement in reproductive functions arises via an "argument from consistency" of several criteria rather than any direct observation of function. The second approach to identifying reproductive proteins offers a more direct inference. Proteomic analyses combining liquid chromatography with mass spectrometry can identify thousands of proteins present in complex biological samples (Steen & Mann 2004, Karr 2008). This method depends on having a protein sequence database to which peptide mass spectra can be matched, so it is most effective in organisms with complete and well-annotated genome sequences. Nonetheless, the method can be used effectively in the absence of a complete genomic sequence by comparing spectra to protein predictions from ESTs (Aagaard *et al.* 2006, Clark *et al.* 2007, Andres *et al.* 2008, Brautigam *et al.* 2008). Proteomic analyses have proven particularly effective in rapidly identifying protein components of the ejaculate from many different taxa, both in sperm and extracellular SFPs (Swanson *et al.* 2001, Fung *et al.* 2004, Collins *et al.* 2006, Kelly *et al.* 2006, Dorus *et al.* 2006, Karr 2007, Andres *et al.* 2008, Findlay *et al.* 2008, Sirot *et al.* 2008). Direct proteomic detection of SFPs nicely complements the indirect approach because: 1) it allows a confirmation of the indirect inferences and 2) in non-model organisms transcriptome sequence (*i.e.* EST analysis) will be the starting point for both methods.

In this paper we present the combined results of both an indirect approach and a direct proteomic analysis (two-dimensional liquid chromatography tandem mass

spectrometry; 2dLC/MS) from two species of butterflies, both in the genus *Heliconius*. Lepidoptera (moths and butterflies) have an extremely rich history as study organisms in ecological and evolutionary research (Boggs *et al.* 2003). In particular, comparative analyses of mating systems are common among Lepidoptera because: 1) they exhibit a wide diversity of mating systems between species, 2) courtship behavior and mating are often conspicuous and easily observed, and 3) in most species the male-derived spermatophore persists indefinitely in the female reproductive tract after mating, providing a reliable life-long record of female mating history (Scott 1972, Drummond 1984, Deinert *et al.* 1994, Bissoondath & Wiklund 1995, Bissoondath & Wiklund 1996, Bergstrom & Wiklund 2002). For these reasons, the evolution and function of ejaculates have been carefully investigated in many Lepidopteran species, especially butterflies. The mass, protein content, and production rate of spermatophores transferred by males all correlate positively with female remating rate (Svard & Wiklund 1989, Bissoondath & Wiklund 1995). Male butterflies can adjust components of their ejaculate in response to sperm competition and risk (Wedell & Cook 1999, Andersson *et al.* 2004, Solensky & Oberhauser 2009). Also, many male derived compounds transferred during copulation clearly benefit females directly (Boggs 1979, Boggs *et al.* 1981, Karlsson 1995), suggesting that sexual conflict is not a major force in shaping the evolution of Lepidopteran ejaculates (Andres *et al.* 2006).

Despite this broad characterization of overall ejaculate evolution and function, little is known about individual proteins comprising Lepidopteran ejaculates. Most previous attempts to identify individual components transferred from males to females in spermatophores have focused on non-protein chemical compounds such as sodium, which functions as a dietary supplement to females (Smedley & Eisner 1996), and cyanogenic compounds, which play a role in chemical defense from predation (Cardoso & Gilbert 2007, Cardoso *et al.* 2009). However, there is also evidence in

Lepidoptera for a direct connection between SFPs and female reproductive physiology. Injecting into females individual purified protein fractions extracted from male accessory glands in the moth *Helicoverpa armigera* stimulates oogenesis and oviposition (Jin & Gong 2001).

Heliconius butterflies have played an integral role in several of these studies analyzing the content and function of spermatophores, which demonstrates their utility as a tractable organismal system for researching the evolutionary dynamics of ejaculates in the Lepidoptera. Moreover, the diversity of mating system comparisons possible across the Lepidoptera is mirrored within the genus *Heliconius*. There are approximately 40 *Heliconius* species, all neotropical, which fall evenly into two clades (Beltran *et al.* 2007). All species in one of these clades exhibit an unusual pupal-mating behavior; these females mate before or during eclosion (Gilbert 1976, Gilbert 1991). *Heliconius* females in the other clade mate as fully developed adults (Brown 1981, Gilbert 1991, Brower 1997). This distinct difference in mating types corresponds to several conspicuous post-mating traits. Behaviorally it is generally accepted that pupal-mating females do not remate (*i.e.* are monandrous) while adult-mating females do regularly remate (*i.e.* females are polyandrous) (Gilbert 1976, Deinert *et al.* 1994, Mallet *et al.* 1998, Deinert 2003, Cardoso *et al.* 2009). Morphologically, pupal-mating females lack *signa* Signa are scleritized rasp-like protrusions on the interior of the bursa copulatrix, which are believed to function in breaking down the spermatophore (Gilbert 2003, Galicia *et al.* 2008). The lack of signa probably relates closely to the observation that, unlike nearly all other butterflies, the spermatophores in pupal-mating *Heliconius* completely degrade in a relatively short period of time (Boggs 1979, Boggs 1981, Deinert 2003).

Beyond the studies discussed above, *Heliconius* butterflies have been a model system for research in evolutionary and ecological genetics, adaptation, and speciation

for over a century (Brown 1981, Brower 1997, Beltran *et al.* 2007). Recently the genomic resources have expanded dramatically for two species, *H. erato* (pupal-mating) and *H. melpomene* (adult-mating). These resources now include extensive collections of ESTs from wing tissue, bacterial artificial chromosome genomic libraries, and linkage maps (Jiggins *et al.* 2005, Kapan *et al.* 2006, Joron *et al.* 2006, Papanicolaou *et al.* 2008, Jiggins *et al.* 2008). They are primarily intended to aid in dissecting the genetic basis of the striking diversity and mimicry/convergence among wing coloration patterns that has long been the major focus of research in this genus (Joron *et al.* 2006). However, this rapidly increasing array of genomic resources in *Heliconius* lays the foundation for molecular-genetic investigations of other diversity present in the genus. In particular, the organismal and comparative framework represented generally by Lepidoptera and specifically by *Heliconius* provides a rich context for investigating the causes and consequences of reproductive protein evolution.

In this paper we present several substantial and important steps towards investigating the relationship between sexual selection and reproductive protein evolution in butterflies. First we report the identification of numerous putative SFPs in *Heliconius* and compare the identity and predicted function of these proteins to SFPs in other insects. Second, we demonstrate that this group of proteins is rapidly evolving relative to a large sample of non-reproductive proteins. Finally, we show at least one of these proteins is evolving adaptively, implicating a role for positive selection in explaining this rapid evolution.

Methods

IDENTIFICATION OF *HELICONIUS* SFPS

We employed two approaches for identifying SFPs in *Heliconius* butterflies. The first approach follows the set of 'indirect' criteria typically employed in identifying insect SFPs: 1) a pattern of gene expression consistent with expression primarily in the male accessory gland and 2) the presence of a computationally predicted signal peptide in the encoded protein, indicating the protein is secreted extracellularly (Swanson *et al.* 2001, Swanson *et al.* 2004, Wagstaff & Begun 2005, Andres *et al.* 2006, Braswell *et al.* 2006, Davies & Chapman 2006, Walters & Harrison 2008, Almeida & DeSalle 2009). The second approach 'directly' identifies SFPs via proteomic analysis of seminal fluid. In both cases our starting point was a collection of ~1100 ESTs (~350 unigenes) sequenced from each of two male accessory gland cDNA libraries constructed from *H. melpomene* and *H. erato* (Walters & Harrison 2008).

Indirect Criteria I: Candidate SFPs

Two distinct bioinformatic approaches were used to identify candidate SFP loci among unigenes present in the two *Heliconius* accessory gland libraries. First, all accessory gland unigene ORFs were 1) BLASTed against ~18,000 ESTs derived from *H. erato* imaginal wing tissue (Papanicolaou *et al.* 2005, Papanicolaou *et al.* 2008, Walters & Harrison 2008) and 2) assayed for the presence of a predicted signal peptide using the Signal P software (Nielsen *et al.* 1997, Bendtsen *et al.* 2004). Accessory gland unigenes lacking a significant BLASTn hit ($E< 10^{-10}$) to any wing EST but containing a predicted signal peptide were considered candidate SFPs; these

are likely to be secreted proteins that are expressed primarily or uniquely in the male accessory gland. The second bioinformatic approach was based on the protein functional class of unigene ORFs determined using InterProScan (Zdobnov & Apweiler 2001). Unigenes with predicted functions similar to known seminal fluid proteins were also considered candidate SFPs (Mueller *et al.* 2004), regardless of the criteria applied in the first approach described above.

Indirect Criteria II: Qualitative Gene Expression Assays

For each bioinformatically identified candidate SFP gene we qualitatively characterized expression patterns using RT-PCR. The goal of these experiments was to identify candidate SFP loci showing expression patterns consistent with being SFPs. Expression was assayed in three body segments: female abdomen, male abdomen, and thorax. The expectation was that true SFP genes should amplify strongly from male abdomen and only weakly or not at all from the other two segments. [We note, however, that proteomic analyses have recently identified SFP genes which do not follow this expected pattern of expression (Bebas *et al.* 2008, Findlay *et al.* 2008, Sirot *et al.* 2008) Locus-specific PCR primers were designed within predicted ORF sequence using Primer3 (Rozen & Skaletsky H.J. 2000). All PCR primers used in the research presented here are available from the authors upon request. Total RNA was extracted separately from the abdomen or thorax of each of three adult male or female *H. melpomene* (from stocks maintained at the University of Texas, Austin) and *H. erato* (from stocks maintained at the University of Puerto Rico, Rio Piedras). Extractions were performed using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and subsequently purified using RNeasy columns (Qiagen, Valencia, CA, USA). A standard concentration of total RNA from each of these RNA extractions (*H. erato*, 1 μg; *H. melpomene*, 0.5 μg) was treated with DNase (Invitrogen, Carlsbad, CA, USA)

and reverse transcribed into single stranded cDNA using poly-T primers, SuperScript III Reverse Transciptase (Invitrogen, Carlsbad, CA, USA), and following the manufacturer's protocol. This cDNA was diluted (*H. erato*, 10-fold; *H. melpomene*, 5-fold) and one µl was used as template in a 10 µl touch-down PCR with the following cycling parameters: initial denaturation of 95°C (2 min), 12 cycles of 95°C (30 sec) then 65-54°C (30 sec, decreasing one degree per cycle) then 72°C (2 min), 23 cycles of 95°C (30 sec) then 53°C (30 sec) then 72°C (2 min), and a final extension of 72°C (4 min). Three µl PCR amplicon were electrophoresed on a 1.2% agarose gel, stained with ethidium bromide and visualized under UV light. Patterns of expression were qualitatively scored as being 'consistent' or 'not consistent' with the pattern expected for SFPs. The gene α-tubulin was used as a positive control.

Direct Criteria: Proteomic Analyses

Complete details of spermatophore sample collection and proteomic analyses are as reported in (Walters & Harrison 2008). Briefly, spermatophores were collected from freshly mated *H. melpomene* (7 females) or *H. erato* (12 females), crushed in saline solution, and centrifuged to pellet sperm and the solid remnants of the spermatophore. The supernatant was sent to the Genome BC Proteomics Centre (University of Victoria, Canada) for 2d-LC/MS proteomic analysis. We searched the resulting spectra against a protein database generated from the combined *H. erato* and *H. melpomene* accessory gland unigene sequences using the MASCOT 2.0 software (Matrix Science, Boston, MA). The protein database, created using custom Perl scripts, consisted of all ORFs $\geq$ 40 amino acids from all three forward reading frames for each unigene as well as likely contaminants: pig trypsin and human keratin obtained from the IPI database (Kersey *et al.* 2004). Another similar database was generated and searched which included unigenes combined from both accessory gland

and wing ESTs. Protein 'hits' to the databases were determined using MASCOT's aggressive (MudPIT) scoring algorithm with a significance threshold of $p \leq 0.01$.

Identification of Homologous Loci and ORF sequencing

Homologous sequences between *H. melpomene* and *H. erato* were identified via manual inspection of reciprocal BLAST analyses of accessory gland unigenes between the two species. High scoring pairs of loci which were reciprocal best BLAST hits (RBBHs) were considered orthologous. For SFP genes initially isolated from only one species, cross-species PCR was attempted to amplify the orthologous sequence. Amplicons were enzymatically cleaned with EXOSAP, sequenced directly in both directions with ABI Prism BigDye Terminator cycle sequencing chemistry, and analyzed on an ABI 3730 automated sequencer. Base-calling and assembly of chromatograms were performed using the phred-phrap algorithm as implemented in the CodonCode Aligner software (CodonCode Corp, Dedham, MA). We assigned a unique identifier, such as 'HACP001' or 'HACP054', to each protein to use as a label and index. Complete ORFs were determined for most loci either by complete sequencing of cDNA library clones or by 5' and 3' RACE (Matz *et al.* 1999, Matz *et al.* 2003). RACE was also used to obtain complete ORF sequences from several additional *Heliconius* species for HACP004 and HACP0018.

SFP annotations

We used several approaches to bioinformatically annotate and characterize the identified *Heliconius* SFPs. First we queried (BLASTx) SFP sequences against GenBank's non-redundant protein database. Second, we searched for predicted protein domains among SFPs using InterProScan. Third, we submitted each SFP to the PHYRE protein fold recognition server (Bennett-Lovsey *et al.* 2008). Finally, we

queried (BLASTp) SFP sequences against the complete set of predicted *D. melanogaster* proteins and recorded the Gene Ontology annotations from FlyBase for the best hit ($E<10^{-3}$) (Tweedie *et al.* 2009). Based on results from these four approaches we assigned a single "summary" functional annotation for each locus.

We also searched specifically for similarity between *Heliconius* SFPs and other known insect SFPs by querying (BLASTp) *Heliconius* SFPs against the complete proteome of *D. melanogaster, A. mellifera*, and *A. aegypti*. Results from these BLAST searches were cross referenced with published sets of SFPs from these insects (Collins *et al.* 2006, Ram & Wolfner 2007, Findlay *et al.* 2008, Sirot *et al.* 2008).

EVOLUTIONARY ANALYSES

Pairwise Comparisons

We estimated evolutionary rates of 30 SFPs for which pairwise alignments > 50 amino acids in length were available in our data. For comparison to the SFPs, 363 alignments of 'control' loci (also > 50 amino acids) were obtained from unigenes derived from *H. melpomene* and H. *erato* wing ESTs (downloaded from ButterflyBase, January 2009). Pairs of RBBH (BLASTp; $E<10^{-10}$) unigenes were aligned as proteins with ClustalW (Thompson *et al.* 1994) and back-translated to the original DNA sequences for analysis using codon models. Maximum likelihood estimates of synonymous and nonsynonymous evolutionary rates for SFP and control loci were obtained using codon sites implemented in the codeml application in the PAML software package (Yang 1997). Third position GC content (GC3) and codon bias (effective number of codons; ENc) were estimated for each locus from both species using the CodonW software (http:// codonw.sourceforge.net).

We tested for differences in evolutionary rates between SFPs and 'control' loci using a permutation T-test and also an ANCOVA. For the permutation T-test, a null distribution of sample mean differences was generated by randomly shuffling the complete data set between two samples equal in size to initial SFP and control groups. For the ANCOVA, the full model fit nonsynonymous rate as the response predicted by synonymous rate, GC3, and ENc as covariates and gene class as a factor (dN ~ dS + GC3 + ENc + Class). The significance of predictor variables was tested using an ANOVA to compare the full model (including Class) and a reduced models (omitting class). All statistical analyses were conducted using the R statistical software package (R Development Core Team 2005).

Multiple Species Comparisons

Two SFPs, HACP004 and HACP018, showed pairwise $\omega = 0.5$, suggesting that expanding analyses to include data from additional species would likely reveal evidence for adaptive evolution at these loci (see Results and Discussion sections for further details). For these sequences we obtained complete ORF sequence (via RACE and sequencing as above) from several *Heliconius* species and at least one outgroup (details of species and sample data are listed in Table 2.1.) DNA alignments of the coding regions were generated using ClustalW to align the protein translations and the back-translated to the original DNA sequence. Using these alignments we tested for adaptively evolving codon sites ($\omega > 1$) using the maximum likelihood models implemented in PAML v. 4.2 (Yang 1997). We compared models M1a to M2a, M7 to M8, and M8a to M8 with significance of model fit determined via likelihood ratio tests. We performed these analyses assuming three different phylogenetic topologies for each locus. First, we used a topology based on a previously published multi-locus molecular phylogeny of the group (Beltran *et al.* 2007). We also performed these test

with topologies inferred via maximum likelihood and neighbor joining methodologies. These phylogenies were reconstructed using the DNAml and DNAdist applications in the PHYLIP software package as implemented in the BioEdit software package (Hall 1999, Felsenstein 2005).

Table 2.1. Species sample and collection information for multi-species analyses.

Taxon Code	Individual Identifier	Genus	species	Source
al	76	*Eueides*	*aliphera*	Collected March 2007 near Gamboa, Panama
ib	158	*Eueides*	*isabella*	Obtained from the Niagara Butterfly Conservatory, August, 2007
hl	106	*Heliconius*	*hecale*	Collected March 2007 near Canazas, Panama
is	102	*Heliconius*	*ismenius*	Collected March 2007 near Canazas, Panama
ao	427	*Neruda*	*aoede*	Collected Dec 2008 near Tarapoto, Peru
bu	426	*Heliconius*	*burneyi*	Collected Dec 2008 near Tarapoto, Peru
do	108	*Heliconius*	*doris*	Collected March 2007 near Gamboa, Panama
xa	347	*Heliconius*	*xanthocles*	Collected Dec 2008 near Tarapoto, Peru
ht	192	*Heliconius*	*hortense*	Obtained from Houston Butterfly Gardens, October 2006
hw	28	*Heliconius*	*hewitsoni*	Obtained from culture matained by L. Gilbert, Univ. of Texas Austin, April 2006
sr	101	*Heliconius*	*sara*	Collected March 2007 near Canazas, Panama
dm	357	*Heliconius*	*demeter*	Collected Dec 2008 near Tarapoto, Peru

Results

Identification of Heliconius *seminal fluid proteins*

We have identified 51 genes putatively encoding seminal fluid proteins in *Heliconius* butterflies (Table 2.2). This number reflects the combined results of both 'indirect' and 'direct' approaches to infer which accessory gland unigenes corresponded to SFPs, which we discuss independently below. Another three proteins putatively encoding spermatophore structural proteins were previously described elsewhere (Walters & Harrison 2008) and we do not discuss them further here.

Table 2.2. Characteristics of *Heliconius* SFPs identified using 'direct' (proteomic) and 'indirect' (expression & bioinformatic) criteria.

HACP	Functional Summary	ORF status (*melpomene*)	ORF status (*erato*)	Pairwise ω	Paralogy	ACP expression (*erato*)	ACP expression (*melpomene*)	2d-LC/MS (*erato*)	2d-LC/MS (*melpomene*)
1	chymotrypsin	Complete	Complete	Y	N	Consistent	Consistent	Y	N
2	chymotrypsin	Complete	Complete	Y	N	Consistent	Consistent	Y	N
3	chymotrypsin	Complete	Complete	Y	N	Consistent	Consistent	Y	N
4	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
5	NA	Complete	Complete	N	Y	Consistent	Consistent	Y	N
6	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	Y
7	chymotrypsin	Fragment	Fragment	Y	N	Consistent	Consistent	N	N
8	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
10	chymotrypsin	Fragment	Fragment	Y	N	Consistent	Consistent	Y	Y
11	NA	Complete	Complete	Y	N	Not Consistent	Consistent	Y	N
12	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
13	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
14	NA	Complete	Complete	N	Y	Consistent	Not Consistent	N	N
15	NA	Complete	Complete	N	Y	Consistent	Not Consistent	N	N
16	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
18	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	Y
20	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	N
21	NA	Complete	Complete	Y	N	Not Tested	Consistent	N	N
23	NA	absent	Complete	N	N	Consistent	Not Tested	N	N

Table 2.2 (Continued)

HACP	Functional Summary	ORF status (*melpomene*)	ORF status (*erato*)	Pairwise ω	Paralogy	ACP expression (*erato*)	ACP expression (*melpomene*)	2d-LC/MS (*erato*)	2d-LC/MS (*melpomene*)
24	NA	absent	Complete	N	N	Consistent	Not Tested	N	N
25	NA	absent	Complete	N	N	Consistent	Not Tested	N	N
26	chymotrypsin	Complete	Complete	Y	N	Consistent	Consistent	Y	N
27	chymotrypsin	Fragment	Complete	Y	N	Consistent	Not Tested	Y	Y
28	NA	Fragment	Complete	Y	N	Consistent	Consistent	N	N
29	NA	Fragment	Fragment	Y	N	Consistent	Consistent	N	N
30	NA	Complete	Complete	Y	N	Consistent	Consistent	Y	Y
31	oxidoreductase	absent	Fragment	N	N	Consistent	Not Tested	N	N
33	NA	Complete	Fragment	N	N	Not Tested	Consistent	N	N
34	NA	Complete	absent	N	N	Not Tested	Consistent	N	N
35	proteinase inhibitor	Complete	Fragment	Y	N	Not Tested	Consistent	N	N
36	NA	Complete	absent	N	N	Not Tested	Consistent	N	N
37	chymotrypsin	Complete	Fragment	Y	N	Consistent	Consistent	N	Y
38	chymotrypsin	Complete	Complete	Y	N	Consistent	Consistent	Y	Y
39	NA	Complete	Fragment	Y	N	Not Tested	Consistent	N	N
40	NA	Complete	absent	N	N	Not Tested	Consistent	N	N
41	NA	Fragment	Fragment	N	Y	Consistent	Consistent	N	N
43	Lipid transport	Complete	Complete	Y	N	Consistent	Consistent	N	N
44	NA	Fragment	Complete	Y	N	Consistent	Not Tested	N	N
46	NA	absent	Fragment	N	N	Consistent	Not Tested	N	N

Table 2.2 (Continued)

HACP	Functional Summary	ORF status (*melpomene*)	ORF status (*erato*)	Pairwise ω	Paralogy	ACP expression (*erato*)	ACP expression (*melpomene*)	2d-LC/MS (*erato*)	2d-LC/MS (*melpomene*)
47	Glycoside hydrolase	absent	Fragment	N	N	Consistent	Not Tested	N	N
48	NA	Complete	absent	N	N	Not Tested	Consistent	N	N
49	chymotrypsin	Complete	Fragment	Y	N	Not Tested	Consistent	N	N
50	NA	Complete	Fragment	N	N	Not Tested	Consistent	N	N
51	NA	Complete	absent	N	N	Not Tested	Consistent	N	N
53	NA	Fragment	Complete	Y	N	Consistent	Not Tested	N	N
54	Hormone binding	Complete	Fragment	Y	N	Consistent	Consistent	N	N
57	proteinase inhibitor	Complete	Fragment	N	Y	Not Consistent	Not Consistent	Y	N
58	Aldo/keto reductase	Fragment	Fragment	Y	N	Not Tested	Not Tested	Y	Y
59	proteinase inhibitor	Fragment	Fragment	Y	N	Not Consistent	Not Consistent	Y	Y
60	zinc ion binding	Fragment	Fragment	Y	N	Not Tested	Not Tested	N	Y
61	CRISP	Fragment	Undetermined	N	N	Not Tested	Not Tested	N	Y

Indirect inference of SFPs: Bioinformatic analysis and differential tissue expression

Focusing on the expected patterns of tissue-limited gene expression and extracellular secretion, we initially used a series of bioinformatic criteria to obtain a set of candidate SFP genes from the complete set of accessory gland library EST unigenes. To evaluate differential patterns of expression, we performed a BLAST-based *in silico* subtraction between unigenes from the accessory gland libraries and genes expressed in the imaginal wing disc of *H. erato*. In both accessory gland libraries 36% of accessory gland unigene ORFs (H. *erato*: 140; *H. melpomene*: 127) had significant BLAST hits ($E< 10^{-10}$) to wing ESTs. We excluded these unigenes from being candidate SFPs on the assumption that genes expressed in developing wing tissue are unlikely to also encode proteins transferred to females in seminal fluid. Extracellular secretion was inferred by the presence of a computationally predicted signal peptide (Nielsen *et al.* 1997, Bendtsen *et al.* 2004). In both species, about 25% of ORFs contained a predicted 5' signal peptide (H. *erato*: 86; *H. melpomene*: 92) (Walters & Harrison 2008).

Combining these two criteria, the lack of a BLAST hit to wing ESTs and the presence of a signal peptide, we generated 95 candidate SFP genes (*H. erato*: 49; *H. melpomene*: 46). To this group we added an additional 17 *H. erato* unigenes and 11 *H. melpomene* candidates. These additional candidate SFPs were excluded by one or both of the above criteria, but they were predicted (via InterProScan) to have a function similar to that of other known seminal fluid proteins (Mueller *et al.* 2004). This produced a total of 123 candidate SFP unigenes.

Validation of candidate SFPs via tissue specific RT-PCR

Candidate SFP unigenes were assayed for tissue specific expression patterns via RT-PCR from standard concentrations of total RNA isolated from male abdomen,

male thorax, and female abdomen. Of the 123 candidates assayed, 65 (*H. erato*: 33; *H. melpomene*: 32) showed expression patterns consistent with being SFP genes; they amplified strongly from male abdomen but weakly or not at all from male thorax and female abdomen (Figure 2.1). Furthermore, many of these 65 proteins appeared to overlap between species and to have been independently isolated in parallel. Careful manual inspection of BLAST comparisons within and between species indicated that 46 unique loci were represented among the 65 candidates with SFP-consistent expression. We henceforth refer to these 46 confirmed candidates as *indirect* SFPs. Patterns of expression were consistent between species. There were 29 indirect SFP loci where expression had been assayed in both species and could be compared for consistency between species. Only three showed patterns of expression consistent with being an SFP in one species but not the other (Table 2.2).

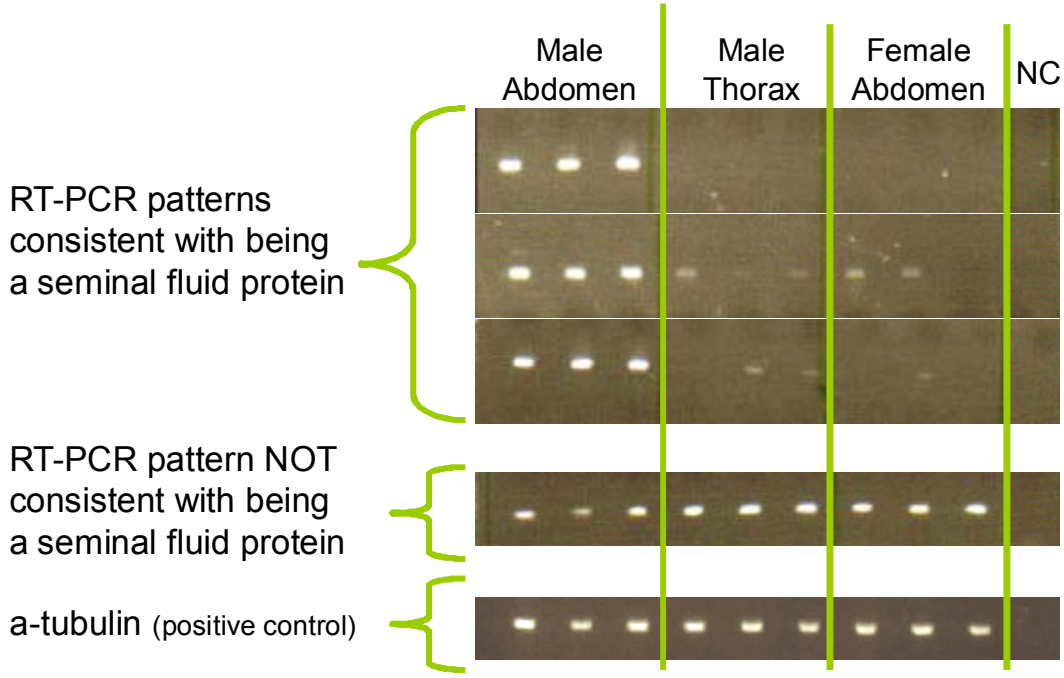

Figure 2.1. Examples of results from qualitative RT-PCR expression assays to determine whether candidate accessory gland unigenes showed patterns of expression as expected for seminal fluid proteins. For each unigene, three individuals were tested for three body segments. NC = negative control.

Direct inference of SFPs: Proteomic analysis.

Tandem mass spectrometry (2d-LC/MS) of spermatophores from *H. erato* and *H. melpomene* provided a means for directly identifying proteins passed to females during copulation. Here we limit our discussion to the results from searching the database consisting only of unigenes from accessory gland ESTs. Results from querying the database including wing ESTs were very similar, producing the same hits to accessory gland unigenes and also some additional matches to 'house keeping' genes (as described below).

In *H. erato* 25 different accessory gland unigene ORFs were significantly matched (protein score $p<0.01$) to peptides present in the spermatophore. Twenty-one of these proteins contained a predicted signal peptide and we therefore consider them putative SFPs. The remaining four proteins identified had, respectively, highly significant BLAST hits to *DOPA-decarboxylase, fructose-bisphosphate aldolase, cystathionine beta-synthetase,* and *cytochrome oxidase II*. The first two of these ORFs appeared to be complete at the 5' end and yielded negative results for predicted signal peptides. We presume these four proteins are not SFPs and reflect 'house keeping' genes which were present in sperm or other tissues which were inadvertently included in the spermatophore sample. In *H. melpomene* 10 unigenes were significantly matched to spermatophore peptides. Nine of these had a predicted signal peptide and we consider these SFPs. The remaining one matched the same *cystathionine beta-synthetase*-like unigene as in *H. erato*. In all cases the peptide matches corresponded to the ORF initially predicted by the PartiGene EST pipeline (Walters & Harrison 2008).

Comparing results between species revealed that seven of the proteomically identified SFPs were recovered from both species, yielding a total of 23 'directly' identified SFPs. Of these, 18 were also identified as SFPs via the indirect criteria

while five were not previously identified via indirect criteria (*H. erato*: 2; *H. melpomene*: 2; 1 shared).

Homology between H. erato and H. melpomene

We used two approaches to identify or obtain homologous SFP gene sequences between *H. erato* and *H. melpomene*. First, reciprocal BLASTing between EST libraries followed by intensive manual inspection of results revealed that many apparently orthologous pairs of SFP genes were already present among accessory gland unigenes from the two species. When a gene was present in the ESTs of only one species, we attempted to amplify a homologous fragment from the missing species using cross-species RT-PCR. These approaches yielded homologous sequences for 40 of the 51 putative SFPs identified overall. Five loci identified in *H. melpomene* and six from *H. erato* could not be successfully cross-amplified in the reciprocal species despite several attempts using several combinations of standard and degenerate primers.

Of the 40 SFPs with homologous sequence isolated from both species, 35 appear to be orthologous with no indication of paralogy in either species. However, there are five proteins which show evidence of paralogy or alternative splicing in at least one of the species where we have been unable to resolve the patterns of variation observed.

Homology to other species and functional annotations.

The majority of *Heliconius* SFPs could not be readily annotated and appear to be novel proteins; 31 showed no significant BLASTx similarity ($E<10^{-5}$) to proteins in GenBank (Table 2.2). We determined functional categories for the remaining 20 *Heliconius* SFPs using a combination of InterProScan, PHYRE, and BLAST similarity

to annotated *D. melanogaster* proteins. Exactly half of these were chymotrypsins, which was the most common functional class, followed by proteinase inhibitors, of which there were three. The frequency of observed functional groups are listed in Table 2.3.

Table 2.3 Counts of functional classes identified among *Heliconius* SFPs. Classes in **bold** include proteins detected in proteomic analyses.

Functional Class	Total
chymotrypsin	10
proteinase inhibitor	3
Aldo/keto reductase	1
CRISP	1
Glycoside hydrolase	1
Hormone binding	1
Lipid transport	1
oxidoreductase	1
zinc ion binding	1
Unknown	31
Grand Total	51

SFPs have been extensively surveyed in three other genome-enabled insects: Fruit fly (*D. melanogaster*), Honey bee (*A. mellifera*), and the yellow fever mosquito (*A. aegypti*) (Collins *et al.* 2006, Ram & Wolfner 2007, Findlay *et al.* 2008, Sirot *et al.* 2008). We BLASTed the *Heliconius* SFPs against the complete set of protein predictions for each of these insects and cross-referenced the top 30 hits with the SFPs identified from each of these insects (Table 2.4).

Table 2.4. Seminal fluid proteins (SFPs) detected in three other insect species which show significant BLASTp hits (Eval < 10^{-4}) to *Heliconius* SFPs.

	Genome Queried											
	D. melanogaster (fruit fly)				*A. aegypti* (Yellow fever mosquito)				*A. mellifera* (Honey Bee)			
***Heliconius* SFP**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**
H. erato												
HACP038_er	FBpp 0077991	8	51.6	14.6								
HACP049_er	FBpp 0086420	30	63.2	28.5	AAEL 014005	18	82	10				
HACP058_er									GB18109	3[c]	188	29
HACP059_er	FBpp 0085496	1	89.4	0	AAEL 008364	9	57	54				
HACP059_er	FBpp 0080979	14	79.3	10.1								
HACP059_er	FBpp 0079243	17	76.6	12.8								
HACP059_er	FBpp 0079094	19	73.9	15.5								
H. melpomene												
HACP003_me					AAE L014005	30						
HACP035_me	FBpp 0083821	4	52.4	11.1								
HACP038_me	FBpp 0077991	20	48.5	20.4								
HACP054_me[b]	FBpp 0082691	1		0								
HACP058_me									GB18109	3[c]	142	25

Table 2.4 (Continued)

	Genome Queried											
	D. melanogaster (fruit fly)				*A. aegypti* (Yellow fever mosquito)				*A. mellifera* (Honey Bee)			
***Heliconius* SFP**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**	**Gene ID**	**Hit Rank**	**bit score**	**Δscore[a]**
HACP059_me	FBpp 0085496	14	177	14	AAEL 008364	8	124	79				
HACP059_me	FBpp 0080979	15	167	24								
HACP059_me	FBpp 0079243	16	165	26	AAEL 002720	22	85.1	117.9				
HACP059_me	FBpp 0079094	19	147	44								
HACP061_me[b]	FBpp 0072229	1	51.2	0								

[a] Difference in alignmient bit score between the SFP BLAST hit listed in the table and top ranking hit returned from BLAST query
[b] Only BLAST hit returned from query to genome using a cutoff of Eval<1e-4.
[c] Only 8 BLAST hits were returned from query to genome using a cutoff of Eval<1e-4.

Only eight *Heliconius* SFPs showed significant similarity ($E<10^{-4}$) to SFPs from these three species and the majority of these were to similar only to fruit fly SFPs.

Evolutionary Analyses

We compared patterns of molecular evolution between *Heliconius* SFPs and a set of 363 'control' loci consisting of unigenes derived from ESTs sequenced from developing wing tissue of *H. erato* and *H. melpomene* (Papanicolaou *et al.* 2008). SFPs showed a clear pattern of accelerated evolution relative to the control loci (Table 2.5). Permutation T-tests yielded a highly significant difference ($p<0.001$) for dN, dS, and ω (the dN/dS ratio) between SFP and control loci (Figures 2.2 and 2.3). SFPs in both species also show significantly lower GC3 values and less codon bias compared to control loci (Figure 2.4 and Table 2.5). Note that the ENc statistic is interpreted such that higher values mean *less* codon bias.

Table 2.5. Mean parameter values (and standard deviations) for seminal fluid proteins (SFPs) and control proteins use in pairwise evolutionary comparisons between *H. erato* and *H. melpomene*.

	No. of Loci	dN	dS	ω	GC3		ENc	
					H. erato	*H. melpo-mene*	*H. erato*	*H. melpo-mene*
SFPs	30	0.0656 (0.039)	0.3010 (0.07)	.22 (.13)	0.35 (0.072)	0.36 (0.07)	53.08 (5.8)	53.66 (3.4)
Control	363	0.0175 (0.029)	0.2072 (0.012)	.08 (.11)	0.47 (0.14)	.47 (0.14)	50.8 (6.6)	50.89 (6.8)

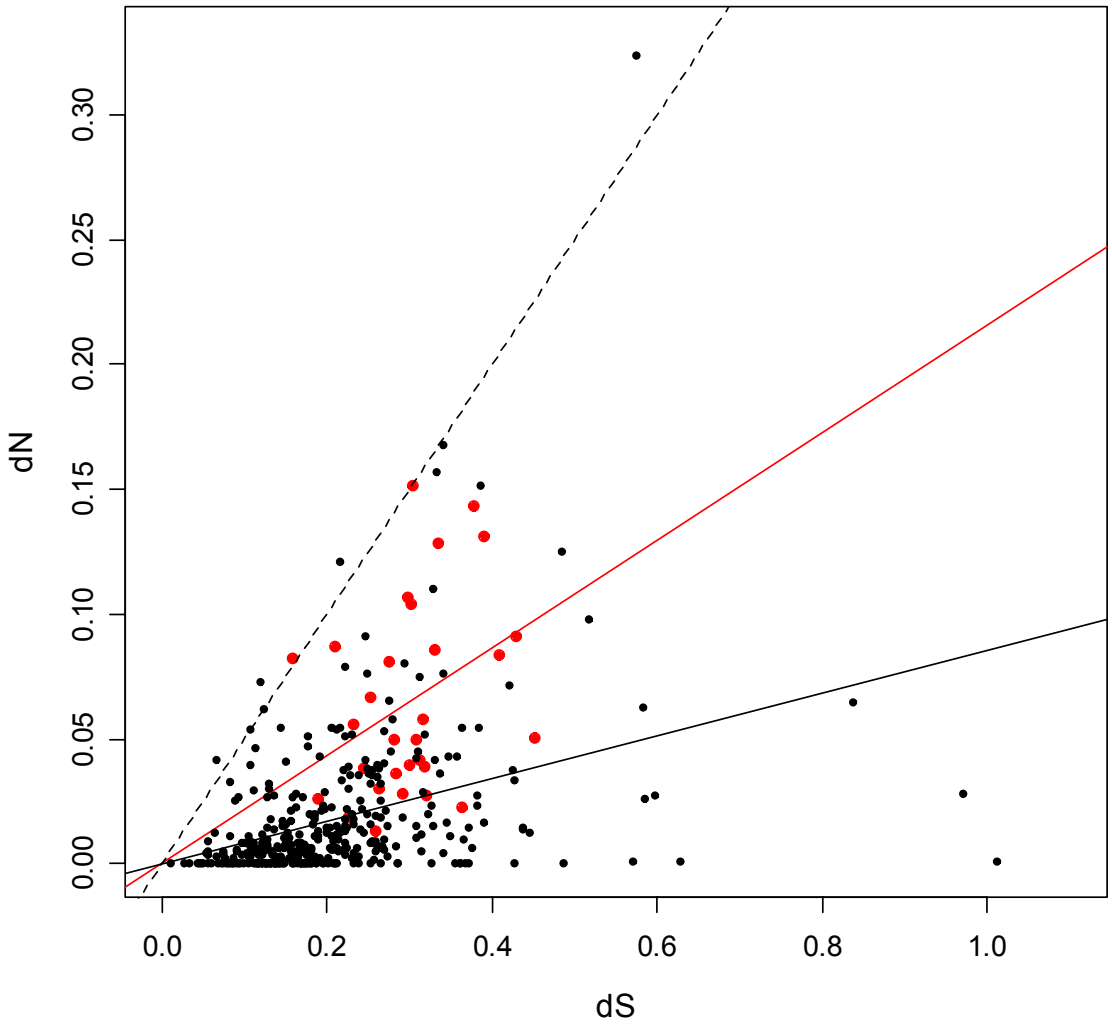

Figure 2.2. Nonsynonymous (dN) versus synonymous (dS) evolutionary rates for 363 'control' proteins (black) and 30 seminal fluid proteins (SFPs; red) estimated from pairwise alignments between *Heliconius erato* and *H. melpomene.* Solid lines represent least-squares regression lines (forced through the origin) for control proteins (black) and SFPs (red). The dashed line represents dN/dS = 0.5.

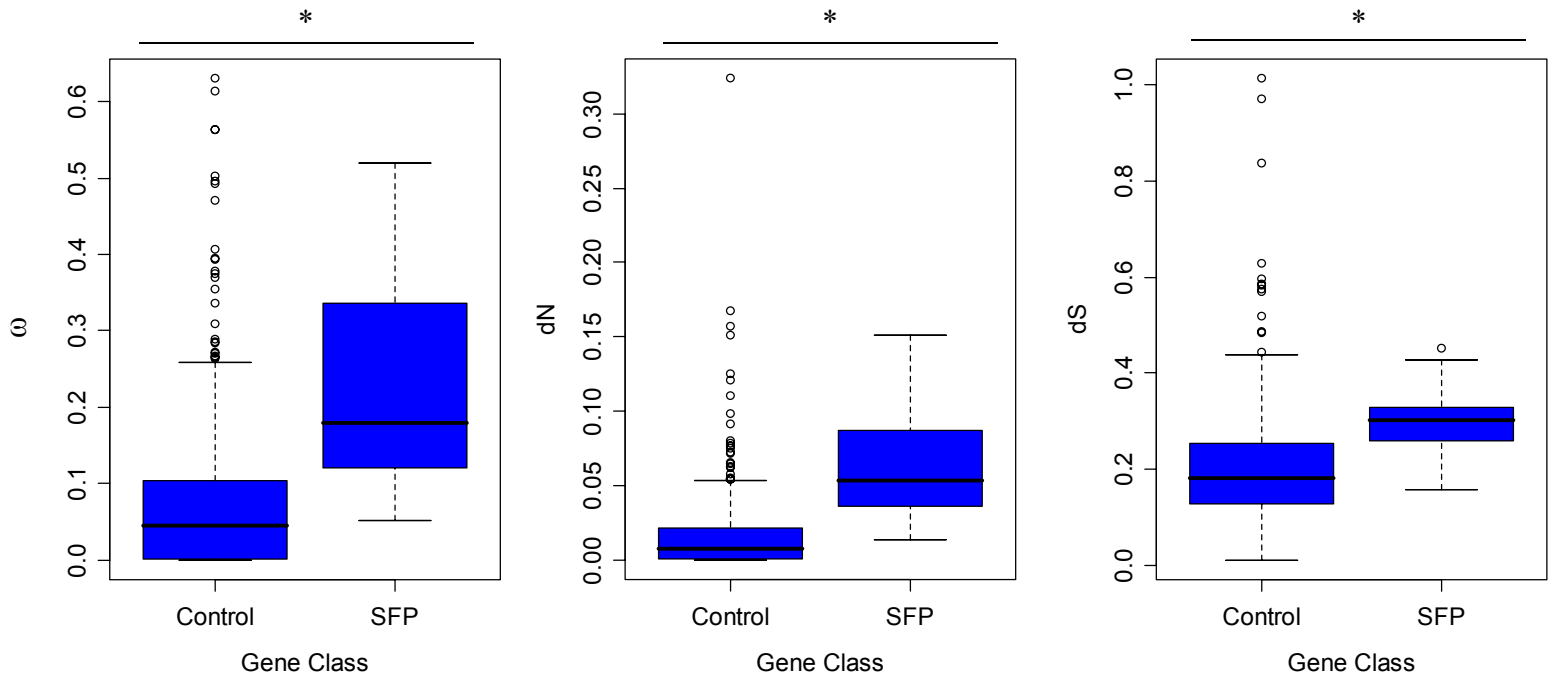

Figure 2.3. Box-Plot comparisons of evolutionary rates (dN, dS, and the dN/dS ratio [ω]) for control and SFPs estimated from pairwise alignments between *Heliconius erato* and *H. melpomene.* Significance: * = p <0.001

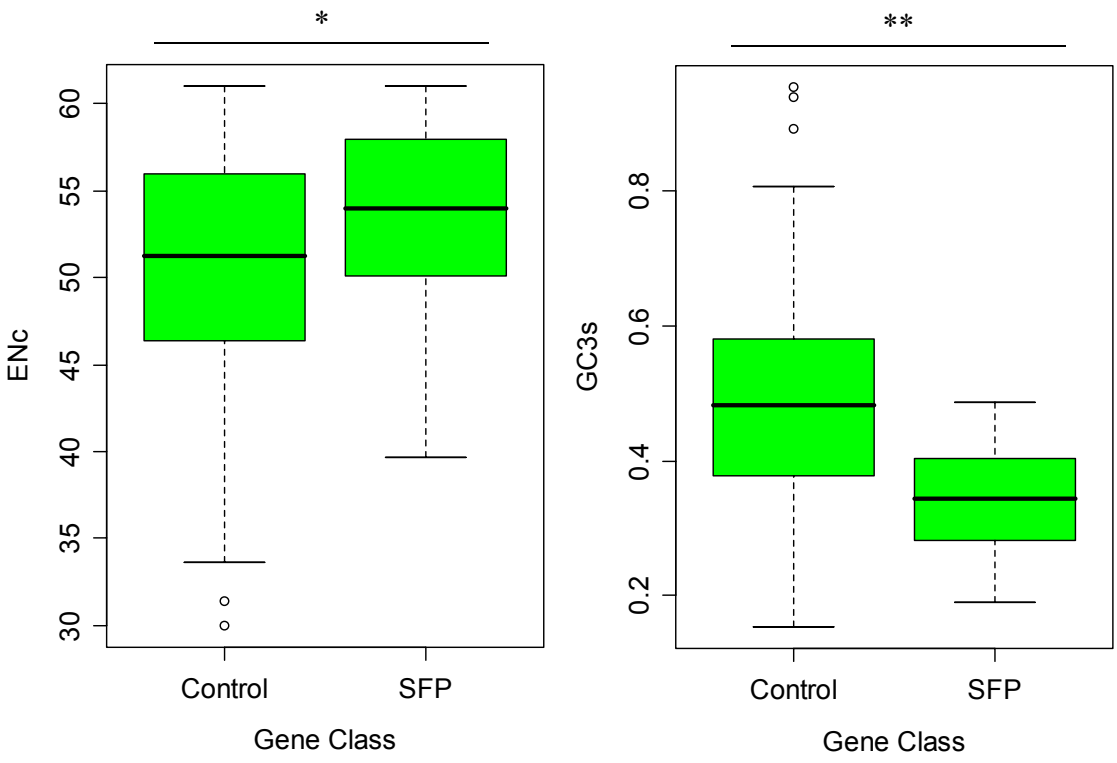

Figure 2.4. Comparison of codon bias (Effective number of codons; ENc) and third-position G/C content (GC3s) for control and SFP loci in *H. erato*. Comparable results were obtained with data from *H. melpomene*. Significance: * = $p < 0.05$; ** = $p < 0.01$.

In addition to the permutation tests, we used a regression framework to test for differences in rates of protein evolution between gene classes while also taking into account differences in GC content and codon bias. In order to meet model assumptions we used a natural log transformation of dN as the response variable. Also, 98 control loci with no observed nonsynonymous differences were discarded. This means the ensuing test for an elevated rate of nonsynonymous evolution was conservative since all control loci showing no protein divergence were excluded from the test. An ANOVA model comparison showed a significantly better fit ($p<0.001$) for a model which included the factor "gene class" as the final term in the model. The effect of being an SFP significantly increased the average rate of evolution above that of the control loci. Comparable results are obtained using estimates of GC3 and ENc from either *H. erato* or *H. melpomene*. There is also a similarly significant effect of gene class when the ANCOVA is parameterized with Ln(ω) as the response in place of Ln(dN); dS is dropped as a predictor in this case.

None of the SFPs assayed showed a clear signature of adaptive evolution (*i.e.* $\omega > 1$) based on the pairwise estimates of evolutionary rate. However, such pairwise comparisons are a highly conservative test for positive selection because the estimate of evolutionary rate is averaged across the entire proteins, potentially obscuring a signal of adaptive evolution that has occurred at specific codon sites (Anisimova & Kosiol 2009). A more powerful and sensitive approach is to use models which allow for selective pressures to vary across codon sites in the gene. Such site models require a multiple-species alignment. Loci showing $\omega \geq 0.5$ from pairwise estimates often show evidence of adaptive evolution when a site model is applied (Swanson *et al.* 2004, Clark & Swanson 2005). Two of 51 total SFP loci in our data meet this criterion: HACP004 and HACP018. For this reason, we sequenced these loci in

several other *Heliconius* and *Eueides* (the sister genus) species and tested for adaptive evolution using site models implemented in PAML (codeml).

We conducted these analyses assuming three different genealogies for each locus: a published species-level phylogeny as well as ML and NJ trees inferred directly from the data. The published and inferred phylogenies differed only slightly for both loci and the results of the codon site models were both qualitatively and quantitatively similar (Figure 2.5 and Table 2.6). For the sake of simplicity we focus here only on the results associated with the published species tree. HACP004 showed a strong signal of adaptive evolution for all three model comparisons ($p < 0.001$). The M8 model indicates 8.5% of codons sites are adaptively evolving with average $\omega = 4.3$. In contrast, HACP018 showed no evidence for adaptive evolution in its recent history. In this case, none of the model comparisons rejected the null hypothesis and estimates of ω were well below one.

Discussion

Identification of SFPs

By combining direct proteomic assays with indirect expression and bioinformatic criteria, we have identified 46 putative SFPs in *H. erato* and 45 in *H. melpomene.* These two groups of SFPs overlapped substantially, with 40 loci shared between species to give a total of 51 distinct SFPs. The two approaches used to identify these loci were also consistent. In both species, ~80% of SFPs identified via proteomics also met the indirect criteria for being SFPs. The proteomic method detected many fewer loci than the indirect approach, particularly in *H. melpomene* where only nine SFPs were detected. However, of these nine, seven overlap with direct SFPs from *H. erato* and six were also indirect SFPs. This deviation in

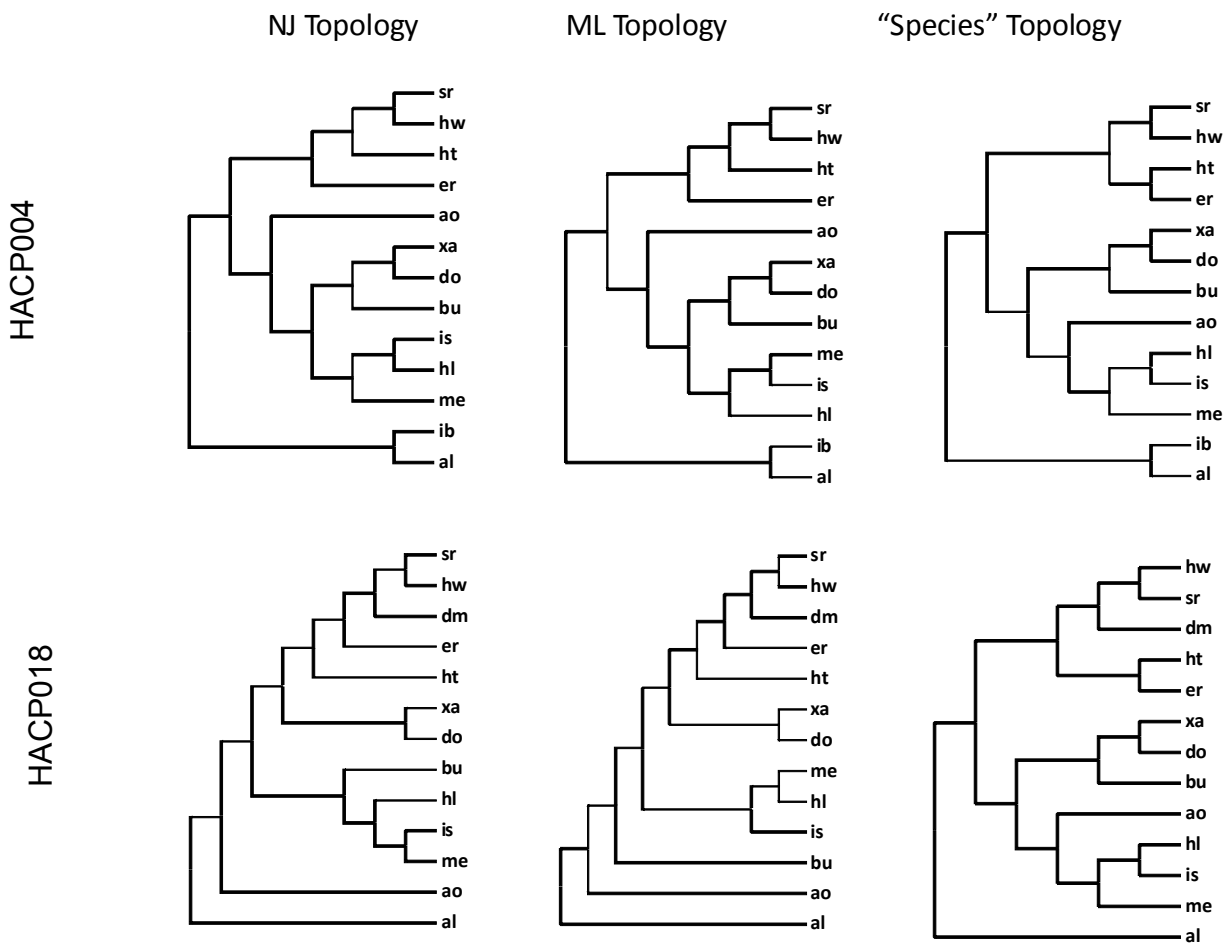

Figure 2.5. Genealogical relationships used in multi-species evolutionary analyses of HACP004 and HACP018. Tree topologies were inferred using with Neighbor Joining (NJ) or Maximum likelihood methods from the aligned sequences. The "species" topology follows Beltran *et al.* 2007. Species codes and collection information are found in Table 2.1

Table 2.6. Results from PAML codon-site tests for adaptive evolution for HACP004 and HACP018 applied assuming three different evolutionary histories among sampled taxa.

Model	lnL	-2ΔlnL	P value	# of params	Kappa	Tree Length	p1	ω1	p2	ω 2	p3	ω3	β p0	β p	β q	β ps	β ωs
Spp Tree																	
HACP004		taxa=13	codons=130														
0	1917.7			25	2.54	2.28	1.00	0.54	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	1864.1			26	2.56	2.44	0.63	0.07	0.37	1.00	0.00	NA	NA	NA	NA	NA	NA
2	1847.3	33.5	<0.001	28	2.92	2.66	0.59	0.07	0.33	1.00	0.08	4.56	NA	NA	NA	NA	NA
7	1866.7			26	2.55	2.44	NA	NA	NA	NA	NA	NA	1.00	0.13	0.17	0.00	NA
8	1847.8	37.6	<0.001	28	2.91	2.65	NA	NA	NA	NA	NA	NA	0.92	0.18	0.29	0.08	4.30
8a	1864.1	32.5	<0.001	27	2.55	2.44											
HACP018		taxa=13	codons=71														
0	739.6			25	2.58	1.53	1.00	0.30	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	736.6			26	2.55	1.54	0.85	0.18	0.15	1.00	0.00	NA	NA	NA	NA	NA	NA
2	736.0	1.0	0.59	28	2.58	1.56	0.96	0.24	0.03	1.00	0.02	3.35	NA	NA	NA	NA	NA
7	736.3			26	2.59	1.54	NA	NA	NA	NA	NA	NA	1.00	0.44	1.03	0.00	NA
8	735.4	1.8	0.41	28	2.59	1.56	NA	NA	NA	NA	NA	NA	0.98	0.89	2.47	0.02	3.58
8a	736.2	1.6	0.44	27	2.57	1.54											
ML Tree																	
HACP004		taxa=13	codons=130														
0	1905.6			25	2.59	2.25	1.00	0.53	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	1855.5			26	2.59	2.39	0.62	0.07	0.38	1.00	0.00	NA	NA	NA	NA	NA	NA
2	1841.1	28.8	<0.001	28	2.92	2.58	0.59	0.07	0.34	1.00	0.07	4.51	NA	NA	NA	NA	NA
7	1857.7			26	2.59	2.39	NA	NA	NA	NA	NA	NA	1.00	0.13	0.17	0.00	NA
8	1841.4	32.4	<0.001	28	2.91	2.58	NA	NA	NA	NA	NA	NA	0.92	0.18	0.28	0.08	4.27
8a	1855.5	28.1	<0.001	27	2.59	2.39											

Table 2.6 (Continued)

Model	lnL	-2ΔlnL	P value	# of params	Kappa	Tree Length	p1	ω1	p2	ω 2	p3	ω3	β p0	β p	β q	β ps	β ωs
HACP018		taxa=13	codons=71														
0	726.6			25	2.99	1.41	1.00	0.32	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	725.9			26	3.09	1.43	0.74	0.15	0.26	1.00	0.00	NA	NA	NA	NA	NA	NA
2	725.9	0	1.00	28	3.09	1.43	0.74	0.15	0.20	1.00	0.06	1.00	NA	NA	NA	NA	NA
7	725.1			26	3.05	1.42	NA	NA	NA	NA	NA	NA	1.00	0.50	1.01	0.00	NA
8	725.1	0	1.00	28	3.05	1.42	NA	NA	NA	NA	NA	NA	1.00	0.50	1.01	0.00	1.00
8a	725.1	0	1.00	27	3.05	1.42											
NJ Tree																	
HACP004		taxa=13	codons=130														
0	1905.9			25	2.63	2.25	1.00	0.54	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	1854.8			26	2.62	2.40	0.62	0.07	0.38	1.00	0.00	NA	NA	NA	NA	NA	NA
2	1840.0	29.63	<0.001	28	2.96	2.59	0.59	0.07	0.34	1.00	0.07	4.54	NA	NA	NA	NA	NA
7	1857.1			26	2.62	2.39	NA	NA	NA	NA	NA	NA	1.00	0.12	0.16	0.00	NA
8	1840.5	33.27	<0.001	28	2.95	2.59	NA	NA	NA	NA	NA	NA	0.92	0.17	0.26	0.08	4.29
8a	1854.8	28.66	<0.001	27	2.62	2.40											
HACP018		taxa=13	codons=71														
0	731.4			25	2.97	1.43	1.00	0.34	0.00	NA	0.00	NA	NA	NA	NA	NA	NA
1	729.0			26	3.07	1.45	0.67	0.10	0.33	1.00	0.00	NA	NA	NA	NA	NA	NA
2	729.0	0.00	1.00	28	3.07	1.45	0.67	0.10	0.23	1.00	0.10	1.00	NA	NA	NA	NA	NA
7	728.4			26	3.03	1.45	NA	NA	NA	NA	NA	NA	1.00	0.31	0.55	0.00	NA
8	728.4	0.00	1.00	28	3.03	1.45	NA	NA	NA	NA	NA	NA	1.00	0.31	0.55	0.00	3.09
8a	728.4	0.00	1.00	27	3.03	1.45											

proteomics results between *H. erato* and *H. melpomene* may reflect differences in the quality of the proteomic samples; the sample of spermatophores was smaller in *H. melpomene* and some of those mated females had been frozen prior to dissection.

We interpret the overlap in results between species and between methods as indicating that we have accurately identified a representative set of SFPs in *Heliconius* butterflies. These proteins therefore provide a good foundation for subsequently characterizing the function and diversity of SFPs in this genus. There is a growing interest in identifying the protein constituents of seminal fluid in a wide range of taxa, many of which lack a relevant reference genome (Wagstaff & Begun 2005, Andres *et al.* 2006, Braswell *et al.* 2006, Davies & Chapman 2006, Andres *et al.* 2008, Almeida & DeSalle 2009, Reinhardt *et al.* 2009). These results confirm this combination of EST analysis, bioinformatics, expression assays, and proteomics as an effective way to identify SFPs in organisms when a complete genome sequence is not available.

SFP function and homology

Two striking patterns are apparent among the functional annotations of *Heliconius* SFPs. First, as might be expected for rapidly evolving proteins, a majority of SFPs could not be functionally annotated in any way, including having no significant BLAST hits ($E<10^{-5}$) to GenBank. Although we employed several different methods to infer the function of *Heliconius* SFPs, such inferences typically rely on primary sequence similarity to other sequences (or sequence clusters) of known function (Higgs & Attwood 2005). When the sequences being annotated are rapidly evolving it is reasonable to expect difficulty with annotations using this approach. An alternative is to base annotations on similarities in predicted protein folding and tertiary structure rather than primary sequence (*e.g.* comparative structural modeling) (Bennett-Lovsey *et al.* 2008). This approach allowed the

functional annotation of many *Drosophila* SFPs which could not be annotated on the basis of primary sequence comparisons (Mueller *et al.* 2004). In our case, however, results from the PHYRE protein fold recognition meta-server were highly consistent with other methods based on primary sequence similarity. In only two cases (HACP043 & HACP054) and did the comparative structural modeling method show significant similarities to proteins where primary sequence comparisons did not. This suggests that many of these unannotated *Heliconius* SFPs are completely novel in structure and function. More exhaustive structural modeling efforts as well as empirical characterization of function will be needed to verify and expand these annotations.

The second major pattern among the functional annotations is the prevalence of proteins predicted to regulate proteolysis. Ten are serine proteases (chymotrypsins) and three are protease inhibitors (Table 2.2). This abundance of protease and protease inhibitors is common among SFPs and is consistent with observations from many different taxa ranging from mammals to insects (Gillott 2003, Mueller *et al.* 2004, Sirot *et al.* 2008). Other functional classes generally common among SFPs and present in *Heliconius* are CRISPs (cystein rich secreted proteins) and oxidoreductases.

Beyond the persistence or expansion of functional classes in SFPs across taxa, another important issue is whether direct homology – even orthology – can be detected between SFPs from distantly related taxa. Such observations would be compelling given the widespread observation of rapid evolution among reproductive proteins. To address this issue we cross-referenced published lists of SFPs with the results from BLASTing *Heliconius* SFPs against the complete proteome of fruit fly (*D. melanogaster*), mosquito (*A. aegypti*), and honey bee (*A. mellifera*).

Only eight *Heliconius* SFPs showed any BLASTp similarity ($E<10^{-4}$) to SFPs from the other three insects, a pattern indicating relatively little detectable homology

among SFPs between insect orders (Table 2.4). Given the lack of a complete genome sequence for *Heliconius* it is impossible to robustly interpret these BLAST results in the context orthology and paralogy. Nonetheless, it seems clear that for six of the eight *Heliconius* SFPs, BLAST similarity to SFPs in the other species is most parsimoniously explained by the two proteins being paralogous (*i.e.* members of a larger protein family) and not directly orthologous. These are the cases where the BLAST hit ranks relatively low among the 30 recorded hits and there is a large score differential between the SFP hit and the top ranking hit.

It is worth noting that in the BLAST of *H. erato* SFPs against fruit fly, HACP059 returned a top hit to an SFP and that this might be interpreted as an argument for close homology. However, at this locus the sequence from *H. erato* is incomplete relative to *H. melpomene*. A similar result was not obtained from the analogous BLAST search in *H. melpomene* and in both species this protein shows strong similarity to many fruit fly proteins (several of which also happen to be SFPs). We thus consider this top hit from *H. erato* to be a spurious result and not strong evidence for close homology.

The strongest argument for orthology between SFPs based on BLAST results can be made for HACP054 and HACP061. In both cases the BLASTs to fruit fly returned only a single hit to an SFP. The putative *D. melanogaster* ortholog to HACP054 is not functionally annotated in FlyBase but was found to be transferred to females at mating (Findlay *et al.* 2008, Tweedie *et al.* 2009). However, comparative structural modeling (via PHYRE) of both the *Heliconius* and *D. melanogaster* protein sequences indicates a significant similarity to a juvenile hormone binding protein from Wax moth (*Galleria mellonella*). Curiously, no sequences similar to HACP054 were found in the *B. mori* (silk moth) genome; neither BLASTp searches to protein predictions nor tBLASTx search to the complete nucleotide assembly returned

significant hits. However, the current *B. mori* genome assembly is only 80% complete, covering only 432 Mb of the 530 Mb genome, so there is a good chance the homologous sequence is lacking in the current assembly (Yamamoto *et al.* 2008). Searches (tBLASTn) against several other insect genomes yielded moderate alignments (bit scores ≈ 40) with a single sequence each in *A. gambiae* and *T. castaneum* (data not shown).

HACP061 and its sole BLAST hit in fruit fly are both predicted to be CRISPs; (Ram & Wolfner 2007) list the fruit fly protein as functioning in 'defense response'. Protein BLAST in *B. mori* yields a single highly significant hit (BGIBMGA000027; $E=2x10^{-31}$) and tBLASTn returns significant alignments in the same genomic location and nowhere else. No similar sequences were found in other insect genomes outside of *Drosophila*.

While further work will be needed to verify the hypotheses of homology suggested here by BLAST results, the observation of a few SFPs conserved across several orders of insects would be a striking result given that these proteins are typically considered to be rapidly evolving. It would also present an excellent opportunity to investigate factors influencing the tempo and mode of molecular evolution among insect SFPs. There is already good evidence of functional conservation between *Drosophila* and Lepidoptera for one well-studied SFP, sex peptide. In *Drosophila,* sex peptide elicits several post-mating response in females, including increased oviposition and egg production thought to be mediated by upregulation juvenile hormone. Injecting synthetic sex peptide into virgin *Helicoverpa armigera* (Bollworm moth) reduces sex pheromone production and upregulates juvenile hormone; these similar effects strongly indicate a conserved function (Fan *et al.* 1999, Fan *et al.* 2000, Wedell 2005, Ram & Wolfner 2007). While an obvious homolog has not been identified in any Lepidopteran species (and is

not apparent among *Heliconius* SFPs), a putative homolog can be identified in honey bee (Ram & Wolfner 2007), providing further evidence for evolutionary constraint among at least few insect SFPs. Similarly, homologs of the *Drosophila* receptor for sex peptide can be identified in many different insects and even in nematode worms (Yapici *et al.* 2008). Such widely conserved SFPs should be of particular interest in context of applied entomology since they presumably play a critical role in reproductive success. Disrupting their function might provide promising new approaches for pest management in many taxa.

Molecular Evolution of Heliconius SFPs

Given the expectation of rapid and adaptive evolution among reproductive proteins (Swanson & Vacquier 2002a, Swanson & Vacquier 2002b, Clark *et al.* 2006, Turner & Hoekstra 2008), we sought to test for these patterns among *Heliconius* SFPs. The issue of rapid evolution is best formulated as a relative question: Are *Heliconius* SFPs evolving more rapidly on average than other proteins in the genome? To address this question we compared the maximum likelihood estimates of evolutionary rates from pairwise alignments between *H. erato* and *H. melpomene* for 30 SFPs and 365 other 'control' proteins. These control proteins are primarily sampled from developing wing tissue but also include putative orthologs found in the accessory gland libraries which were not determined to be SFPs. Biologically, we assume these control proteins have no direct role in reproductive processes and we note that misclassifying reproductive proteins as 'controls' would make this test more conservative.

The rate of molecular evolution is distinctly elevated among SFPs relative to the other proteins sampled (Figures 2.2 and 2.3 and Table 2.5). All three measures of evolutionary rate (dN, dS, and ω) are significantly greater among SFPs , a result which

extends the taxonomic breadth and further confirms the widespread observation of rapid evolution among reproductive proteins. This observation of rapidly evolving reproductive proteins is also fundamentally important in laying a foundation for developing Lepidoptera, and *Heliconius* butterflies in particular, as a model system to investigate the causes and consequences of rapid evolution of reproductive proteins.

Nonetheless, it would be naïve to attribute the observation of rapid evolution of *Heliconius* SFPs solely to the effect of differential selection pressures arising from reproductive processes. Several factors correlate with and may influence the evolutionary rate of proteins, including: mutation rate, nucleotide composition, codon bias, recombination, genomic location, expression level, *etc*. (Li 1997, Lynch 2007). Many of these factors cannot be addressed with our data due to the lack of a complete genomic sequence in *Heliconius* or any closely related species. However, we did examine patterns of nucleotide composition and codon bias in SFP and control genes. Both GC content and codon bias are significantly lower among SFPs (Figure 2.4); this pattern holds using estimates generated from either *H. melpomene* or *H. erato*.

The evolutionary and functional significance of variation in codon bias and compositional bias has been debated at length without a clear consensus yet emerging *c.f.* (Drummond *et al.* 2005, Plotkin *et al.* 2006, Drummond *et al.* 2006, Plotkin & Fraser 2007, Drummond & Wilke 2008, Hershberg & Petrov 2008). While our data cannot address this debate, we note that that differences in GC content and codon bias potentially indicate that variation in evolutionary rates arise from mechanisms other than selection on protein function. We therefore explicitly tested whether the pattern of rapid evolution among *Heliconius* SFPs holds when nucleotide composition and codon bias are taken into account. We fit a linear regression model to our data which predicted the evolutionary rate of proteins as a function of GC3, ENc, and functional class (SFP or control). The 'full' model fit the data significantly better than a model

which excluded the functional class as a factor. This result strengthens the argument that selection on protein function plays a prominent role in explaining the rapid evolution of reproductive proteins. We hasten to add that our formulation of this test was extremely conservative because we excluded from this analysis 98 (of 363) control proteins which lacked any observed nonsynonymous substitutions. This had the dual effect of reducing our power to detect a difference while also distinctly upwardly biasing the mean evolutionary rate of control loci. Nonetheless, the full model fit significantly better no matter how the test was implemented: using parameter estimates from *H. erato* or *H. melpomene* and using ω or dN as the response variable (in the latter case, dS was added to the model as a predictor).

In contrast to the issue of *relatively* rapid evolution, inferences of positive directional selection at the molecular level are typically formulated in an absolute sense by testing gene by gene for $\omega > 1$. Maximum likelihood models of codon-site evolution allow this inference to be made in a variety of ways; some focus on the average evolutionary rates across the entire molecule (such as in pairwise comparisons) while others focus on variation in evolutionary rate between codon sites within a molecule (site models) (Yang & Bielawski 2000, Anisimova & Kosiol 2009). Our analysis of evolutionary rates based on pairwise models yielded no loci with $\omega > 1$. However, it is well known that pairwise estimates of evolutionary rates offer only a very conservative test for adaptive molecular evolution because averaging evolutionary rates across codon sites potentially masks a signal of positive selection at specific codon sites (Anisimova *et al.* 2001, Anisimova & Kosiol 2009). In contrast, site models offer a more powerful and sensitive test for adaptive molecular evolution, though these methods are limited by the need for data from multiple species. It has been shown that $\omega \geq 0.5$ from a pairwise estimate is a good predictor of observing $\omega > 1$ at specific codon sites when data from additional taxa are available (Swanson *et al.*

2004). In our set of pairwise estimates two SFP loci showed $\omega \geq 0.5$, HACP004 and HACP018. Thus these loci were therefore obvious candidates for analyzing with site models in order to better address the question of whether adaptive evolution can be invoked as a cause of rapid evolution among *Heliconius* SFPs.

None of the site model tests gave significant results for HACP018, so while this protein is evolving relatively rapidly, this elevated rate does not appear to be the result of recent positive selection. In contrast, HACP004 yielded highly significant results for all three tests (M1a-M2a, M7-M8, and M8a-M8), strongly implicating positive selection as a major factor underlying the rapid evolution of this protein. For both loci the results reported here assume gene genealogies are consistent with the most recently published species phylogeny (Beltran *et al.* 2007); using genealogies reconstructed from the data via maximum likelihood and neighbor-joining gives comparable results (Table 2.6). Considered broadly, this result confirms the principle that adaptive evolution is a reasonable explanation for the rapid evolution of at least some *Heliconius* SFPs (though clearly not all, *e.g.* HACP018). Unfortunately, we were unable to functionally annotate either of these proteins so it is difficult to discuss or speculate on these results in any functional context.

Conclusion

Post-mating sexual selection persists in the literature as a widely-invoked explanation for the frequent observation of rapid and adaptive evolution among reproductive proteins, yet current empirical support for this hypothesis is still tentative. Thoroughly evaluating this hypothesis requires expanding the sampling of reproductive proteins into taxa providing informative comparisons between mating systems. The work presented here substantially contributes towards this goal. We have combined proteomic, expression, and bioinformatic criteria to identify several

dozen novel and rapidly evolving SFPs in *Heliconius* butterflies, an emerging genomic model system with a strong precedent of ecological, ethological, and evolutionary research. The Lepidoptera have a rich history as model systems for studying sexual selection and mating systems at the organismal level. In particular, the divergent mating systems in *Heliconius* offer an informative contrast for evaluating the effects of sexual selection on reproductive characters. Our work provides the foundation for extending this research to the molecular genetic level.

REFERENCES

Aagaard, J. E., X. H. Yi, M. J. MacCoss, and W. J. Swanson. 2006. Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. Proceedings of the National Academy of Sciences of the United States of America **103**:17302-17307.

Almeida, F. C., and R. DeSalle. 2008. Evidence of adaptive evolution of accessory gland proteins in closely related species of the Drosophila repleta group. Molecular Biology and Evolution **25**:2043-2053.

Almeida, F. C., and R. DeSalle. 2009. Orthology, Function and Evolution of Accessory Gland Proteins in the Drosophila repleta Group. Genetics **181**:235-245.

Andersson, J., A. K. Borg-Karlson, and C. Wiklund. 2004. Sexual conflict and anti-aphrodisiac titre in a polyandrous butterfly: male ejaculate tailoring and absence of female control. Proceedings of the Royal Society of London Series B-Biological Sciences **271**:1765-1770.

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Molecular Biology and Evolution **23**:1574-1584.

Andres, J. A., L. S. Maroja, and R. G. Harrison. 2008. Searching for candidate speciation genes using a proteomic approach: seminal proteins in field crickets. Proceedings of the Royal Society B-Biological Sciences **275**:1975-1983.

Anisimova, M., J. P. Bielawski, and Z. H. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution **18**:1585-1592.

Anisimova, M., and C. Kosiol. 2009. Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. Molecular Biology and Evolution **26**:255-271.

Bebas, P., J. Kotwica, E. Joachimiak, and J. M. Giebultowicz. 2008. Yolk protein is expressed in the insect testis and interacts with sperm. Bmc Developmental Biology **8**.

Beltran, M., C. D. Jiggins, A. V. Z. Brower, E. Bermingham, and J. Mallet. 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.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology **340**:783-795.

Bennett-Lovsey, R. M., A. D. Herbert, M. J. E. Sternberg, and L. A. Kelley. 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. Proteins-Structure Function and Bioinformatics **70**:611-625.

Bergstrom, J., and C. Wiklund. 2002. Effects of size and nuptial gifts on butterfly reproduction: can females compensate for a smaller size through male-derived nutrients? Behavioral Ecology and Sociobiology **52**:296-302.

Bissoondath, C. J., and C. Wiklund. 1995. Protein-Content of Spermatophores in Relation to Monandry Polyandry in Butterflies. Behavioral Ecology and Sociobiology **37**:365-371.

Bissoondath, C. J., and C. Wiklund. 1996. Male butterfly investment in successive ejaculates in relation to mating system. Behavioral Ecology and Sociobiology **39**:285-292.

Boggs, C. L. 1979. Resource Allocation and Reproductive Strategies in Several Heliconine Butterfly Species. Thesis. University of Texas, Austin.

Boggs, C. L. 1981. Selection Pressures Affecting Male Nutrient Investment at Mating in Heliconiine Butterflies. Evolution **35**:931-940.

Boggs, C. L., J. T. Smiley, and L. E. Gilbert. 1981. Patterns of pollen exploitation by *Heliconius* butterflies. Oecologia **48**:284-289.

Boggs C. L., W. B. Watt, and P. R. Ehrlich. 2003. Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago.

Braswell, W. E., J. A. Andres, L. S. Maroja, R. G. Harrison, D. J. Howard, and W. J. Swanson. 2006. Identification and comparative analysis of accessory gland proteins in Orthoptera. Genome **49**:1069-1080.

Brautigam, A., R. P. Shrestha, D. Whitten, C. G. Wilkerson, K. M. Carr, J. E. Froehlich, and A. P. M. Weber. 2008. Low-coverage massively parallel pyrosequencing of cDNAs enables proteomics in non-model species: Comparison of a species-specific database generated by pyrosequencing with databases from related species for proteome analysis of pea chloroplast envelopes. Journal of Biotechnology **136**:44-53.

Brower, A. V. Z. 1997. The evolution of ecologically important characters in Heliconius butterflies (Lepidoptera: Nymphalidae): A cladistic review. Zoological Journal of the Linnean Society **119**:457-472.

Brown, K. S. 1981. The Biology of *Heliconius* and related genera. Annual Review of Entomology **26**:427-456.

Bustamante, C. D., A. Fledel-Alon, S. Williamson, R. Nielsen, M. T. Hubisz, S. Glanowski, D. M. Tanenbaum, T. J. White, J. J. Sninsky, R. D. Hernandez, D. Civello, M. D. Adams, M. Cargill, and A. G. Clark. 2005. Natural selection on protein-coding genes in the human genome. Nature **437**:1153-1157.

Cardoso, M. Z., and L. E. Gilbert. 2007. A male gift to its partner? Cyanogenic glycosides in the spermatophore of longwing butterflies (Heliconius). Naturwissenschaften **94**:39-42.

Cardoso, M. Z., J. J. Roper, and L. E. Gilbert. 2009. Prenuptial agreements: mating frequency predicts gift-giving in Heliconius species. Entomologia Experimentalis et Applicata **131**:109-114.

Civetta, A., and R. S. Singh. 1995. High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in Drosophila melanogaster and Drosophila virilis group species. Journal of Molecular Evolution **41**:1085-1095.

Civetta, A., and R. S. Singh. 1998. Sex-related genes, directional sexual selection, and speciation. Molecular Biology and Evolution **15**:901-909.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

Clark, N. L., G. D. Findlay, X. H. Yi, M. J. MacCoss, and W. J. Swanson. 2007. Duplication and selection on abalone sperm lysin in an allopatric population. Molecular Biology and Evolution **24**:2081-2090.

Clark, N. L., and W. J. Swanson. 2005. Pervasive Adaptive Evolution in Primate Seminal Proteins. PLoS Genet. **1**:e35.

Collins, A. M., T. J. Caperna, V. Williams, W. M. Garrett, and J. D. Evans. 2006. Proteomic analyses of male contributions to honey bee sperm storage and mating. Insect Molecular Biology **15**:541-549.

Davies, S. J., and T. Chapman. 2006. Identification of genes expressed in the accessory glands of male Mediterranean Fruit Flies (Ceratitis capitata). Insect Biochemistry and Molecular Biology **36**:846-856.

Deinert, E. I., J. T. LONGINO, and L. E. Gilbert. 1994. Mate Competition in Butterflies. Nature **370**:23-24.

Deinert, E. L. 2003. Mate location and competition for mates in a pupal mating butterfly. Pages 91-108 *in* Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago.

Dorus, S., S. A. Busby, U. Gerike, J. Shabanowitz, D. F. Hunt, and T. L. Karr. 2006. Genomic and functional evolution of the Drosophila melanogaster sperm proteome. Nat.Genet. **38**:1440-1445.

Dorus, S., P. D. Evans, G. J. Wyckoff, S. S. Choi, and B. T. Lahn. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nature Genetics **36**:1326-1329.

Drummond, B. A. 1984. Multiple mating and sperm competition in the Lepidoptera. Pages 291-370 *in* Smith R.L. editor. Sperm Competition and the Evolution of Animal Mating Systems. Academic Press, New York.

Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold. 2005. Why highly expressed proteins evolve slowly. Proceedings of the National Academy of Sciences of the United States of America **102**:14338-14343.

Drummond, D. A., A. Raval, and C. O. Wilke. 2006. A single determinant dominates the rate of yeast protein evolution. Molecular Biology and Evolution **23**:327-337.

Drummond, D. A., and C. O. Wilke. 2008. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell **134**:341-352.

Fan, Y. L., A. Rafaeli, C. Gileadi, E. Kubli, and S. W. Applebaum. 1999. Drosophila melanogaster sex peptide stimulates juvenile hormone synthesis and depresses sex pheromone production in Helicoverpa armigera. Journal of Insect Physiology **45**:127-133.

Fan, Y. L., A. Rafaeli, P. Moshitzky, E. Kubli, Y. Choffat, and S. W. Applebaum. 2000. Common functional elements of Drosophila melanogaster seminal peptides involved in reproduction of Drosophila melanogaster and Helicoverpa armigera females. Insect Biochemistry and Molecular Biology **30**:805-812.

Fay, J. C., and C. I. Wu. 2003. Sequence divergence, functional constraint, and selection in protein evolution. Annu.Rev.Genomics Hum.Genet. **4**:213-235.

Felsenstein J. PHYLIP (Phylogeny Inference Package). [3.6]. 2005. University of Washington, Seattle, Department of Genome Sciences.
Ref Type: Computer Program

Findlay, G. D., X. H. Yi, M. J. MacCoss, and W. J. Swanson. 2008. Proteomics reveals novel Drosophila seminal fluid proteins transferred at mating. Plos Biology **6**:1417-1426.

Fung, K. Y. C., L. M. Glode, S. Green, and M. W. Duncan. 2004. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. Prostate **61**:171-181.

Galicia, I., V. Sanchez, and C. Cordero. 2008. On the function of signa, a genital trait of female Lepidoptera. Annals of the Entomological Society of America **101**:786-793.

Galindo, B. E., V. D. Vacquier, and W. J. Swanson. 2003. Positive selection in the egg receptor for abalone sperm lysin. Proceedings of the National Academy of Sciences of the United States of America **100**:4639-4643.

Gilbert, L. E. 2003. Adaptive noverlty through introgression in Heliconius Wing Patters: Evidence for a shared Genetic "Toolbox" from Synthetic hybrid zones and a theory of diversification. Pages 281-318 *in* Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago.

Gilbert, L. E. 1991. Biodiversity of a Central American Heliconius community: pattern, process, and problems. Pages 403-427 *in* P. W. Price, T. M. Lewinsohn, G. W. Fernandes, and W. W. Benson editors. Plant-Animal Interactions Evolutionary Ecology in Tropical and Temperate Regions. John Wiley and Sons.

Gilbert, L. E. 1976. Postmating Female Odor in Heliconius Butterflies - Male-Contributed Anti-Aphrodisiac. Science **193**:419-420.

Gillott, C. 2003. Male accessory gland secretions: Modulators of female reproductive physiology and behavior. Annual Review of Entomology **48**:163-184.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. R. Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh. 2007. Evolution in the fast lane: Rapidly evolving sex-related genes in drosophila. Genetics **177**:1321-1335.

Hall, T. A. 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series **41**:95-98.

Herlyn, H., and H. Zischler. 2007. Sequence evolution of the sperm ligand zonadhesin correlates negatively with body weight dimorphism in primates. Evolution Int.J.Org.Evolution **61**:289-298.

Hershberg, R., and D. A. Petrov. 2008. Selection on Codon Bias. Annual Review of Genetics **42**:287-299.

Higgs P. G., and T. K. Attwood. 2005. Bioinformatics and molecular evolution. Blackwell, Malden, MA.

Holmes, E. C. 2004. Adaptation and immunity. Plos Biology **2**:1267-1269.

Jensen-Seaman, M. I., and W. H. Li. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. Journal of Molecular Evolution **57**:261-270.

Jiggins, C. D., S. Baxter, W. O. McMillan, N. Chamberlain, and R. ffrench-Constant. 2008. Prospects for locating genes of interest in lepidopteran genomes: A case study of butterfly colour patterns. *in* M. R. Goldsmith, and F. Marec editors. Lepidopteran Molecular Biology and Genetics. CRC.

Jiggins, C. D., J. Mavarez, M. Beltran, W. O. McMillan, J. S. Johnston, and E. Bermingham. 2005. A Genetic Linkage Map of the Mimetic Butterfly Heliconius melpomene. Genetics **171**:557-570.

Jin, Z. Y., and H. Gong. 2001. Male accessory gland derived factors can stimulate oogenesis and enhance oviposition in Helicoverpa armigera (Lepidoptera : Noctuidae). Archives of Insect Biochemistry and Physiology **46**:175-185.

Joron, M., C. D. Jiggins, A. Papanicolaou, and W. O. McMillan. 2006. Heliconius wing patterns: an evo-devo model for understanding phenotypic diversity. Heredity **97**:157-167.

Kapan, D. D., N. S. Flanagan, A. Tobler, R. Papa, R. D. Reed, J. A. Gonzalez, M. R. Restrepo, L. Martinez, K. Maldonado, C. Ritschoff, D. G. Heckel, and W. O. McMillan. 2006. Localization of Mullerian mimicry genes on a dense linkage map of Heliconius erato. Genetics **173**:735-757.

Karlsson, B. 1995. Resource-Allocation and Mating Systems in Butterflies. Evolution **49**:955-961.

Karr, T. L. 2008. Application of proteomics to ecology and population biology. Heredity **100**:200-206.

Karr, T. L. 2007. Fruit flies and the sperm proteome. Human Molecular Genetics **16**:R124-R133.

Kelly, V. C., S. Kuy, D. J. Palmer, Z. Z. Xu, S. R. Davis, and G. J. Cooper. 2006. Characterization of bovine seminal plasma by proteomics. Proteomics **6**:5826-5833.

Kersey, P. J., J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney, and R. Apweiler. 2004. The International Protein Index: An integrated database for proteomics experiments. Proteomics **4**:1985-1988.

Kimura, M. 1968. Evolutionary Rate at the Molecular Level. Nature **217**:624-&.

Kingan, S. B., M. Tatar, and D. M. Rand. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. Journal of Molecular Evolution **57**:159-169.

Lewontin, R. C., and J. L. Hubby. 1966. A Moleuclar Approach to Study of Genic Heterozygosity in Natural Populations .2. Amount of Variation and Degree of Heterozygosity in Natural Populations of Drosophila Pseudoobscura. Genetics **54**:595-&.

Li W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Lynch M. 2007. The origins of genome architecture. Sinauer Associates, Sunderland, Mass.

Mallet, J., W. O. McMillan, and C. D. Jiggins. 1998. Estimating the mating behavior of a pair of hybridizing Heliconius species in the wild. Evolution **52**:503-510.

Markow, T. A. 2002. Perspective: Female remating, operational sex ratio, and the arena of sexual selection in Drosophila species. Evolution **56**:1725-1734.

Matz, M., N. Alieva, A. Chenchik, and S. Lukyanov. 2003. Amplification of cDNA ends using PCR suppresion effect and step-out PCR. Pages 41-50 *in* S. Y. Ying editor. Generation of cDNA libraries. Humana Press, Totowa.

Matz, M., D. Shagin, E. Bogdanova, O. Britanova, S. Lukyanov, L. Diatchenko, and A. Chenchik. 1999. Amplification of cDNA ends based on template-switching effect and step-out PCR. Nucleic Acids Research **27**:1558-1560.

Mueller, J. L., K. R. Ram, L. A. Mcgraw, M. C. B. Qazi, E. D. Siggia, A. G. Clark, C. F. Aquadro, and M. F. Wolfner. 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics **171**:131-143.

Mueller, J. L., D. R. Ripoll, C. F. Aquadro, and M. F. Wolfner. 2004. Comparative structural modeling and inference of conserved protein classes in Drosophila seminal fluid. Proceedings of the National Academy of Sciences of the United States of America **101**:13542-13547.

Nei, M. 2005. Selectionism and neutralism in molecular evolution. Mol.Biol.Evol. **22**:2318-2342.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. vonHeijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering **10**:1-6.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. J. Sninsky, M. D. Adams, and M. Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. Plos Biology **3**:976-985.

Papanicolaou, A., M. Joron, W. O. McMillan, M. L. Blaxter, and C. D. Jiggins. 2005. Genomic tools and cDNA derived markers for butterflies. Molecular Ecology **14**:2883-2897.

Papanicolaou, A., S. Gebauer-Jung, M. L. Blaxter, W. Owen McMillan, and C. D. Jiggins. 2008. ButterflyBase: a platform for lepidopteran genomics. Nucleic Acids Research **36**:D582-D587.

Plotkin, J. B., J. Dushoff, M. M. Desai, and H. B. Fraser. 2006. Codon usage and selection on proteins. Journal of Molecular Evolution **63**:635-653.

Plotkin, J. B., and H. B. Fraser. 2007. Assessing the determinants of evolutionary rates in the presence of noise. Molecular Biology and Evolution **24**:1113-1121.

R Development Core Team. R: A language and environment for statistical computing. 2005. Vienna, Austria, R Foundation for Statistical Computing.
Ref Type: Computer Program

Ram, K. R., and M. F. Wolfner. 2007. Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology **47**:427-445.

Ramm, S. A., P. L. Oliver, C. P. Ponting, P. Stockley, and R. D. Emes. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Molecular Biology and Evolution **25**:207-219.

Reinhardt, K., C. H. Wong, and A. S. Georgiou. 2009. Detection of seminal fluid proteins in the bed bug, Cimex lectularius, using two-dimensional gel electrophoresis and mass spectrometry. Parasitology **136**:283-292.

Roth, C., M. J. Betts, P. Steffansson, G. Saelensminde, and D. A. Liberles. 2005. The Adaptive Evolution Database (TAED): a phylogeny based tool for comparative genomics. Nucleic Acids Research **33**:D495-D497.

Rozen, S., and Skaletsky H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. Pages 365-386 *in* S. M. S. Krawetz editor. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Scott, J. A. 1972. Mating of Butterflies. Journal of Resarch on the Lepidoptera **11**:99-127.

Sirot, L. K., R. L. Poulson, M. C. McKenna, H. Girnary, M. F. Wolfner, and L. C. Harrington. 2008. Identity and transfer of male reproductive gland proteins of the dengue vector mosquito, Aedes aegypti: Potential tools for control of female feeding and reproduction. Insect Biochemistry and Molecular Biology **38**:176-189.

Smedley, S. R., and T. Eisner. 1996. Sodium: a male moth's gift to its offspring. Proc.Natl.Acad.Sci.U.S.A **93**:809-813.

Solensky, M. J., and K. S. Oberhauser. 2009. Male monarch butterflies, Danaus plexippus, adjust ejaculates in response to intensity of sperm competition. Animal Behaviour **77**:465-472.

Steen, H., and M. Mann. 2004. The ABC's (and XYZ's) of peptide sequencing. Nature Reviews Molecular Cell Biology **5**:699-711.

Svard, L., and C. Wiklund. 1989. Mass and production-rate of ejaculates in relatino to monandry and polyandry in butterflies. Behavioral Ecology and Sociobiology **24**:395-402.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proceedings of the National Academy of Sciences of the United States of America **98**:7375-7379.

Swanson, W. J., and V. D. Vacquier. 2002a. Reproductive protein evolution. Annual Review of Ecology and Systematics **33**:161-179.

Swanson, W. J., and V. D. Vacquier. 2002b. The rapid evolution of reproductive proteins. Nature Reviews Genetics **3**:137-144.

Swanson, W. J., A. Wong, M. F. Wolfner, and C. F. Aquadro. 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics **168**:1457-1465.

Thomas, S., and R. S. Singh. 1992. A Comprehensive Study of Genic Variation in Natural-Populations of Drosophila-Melanogaster .7. Varying Rates of Genic Divergence As Revealed by 2-Dimensional Electrophoresis. Molecular Biology and Evolution **9**:507-525.

Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research **22**:4673-4680.

Turner, L. M., and H. E. Hoekstra. 2008. Causes and consequences of the evolution of reproductive proteins. International Journal of Developmental Biology **52**:769-780.

Tweedie, S., M. Ashburner, K. Falls, P. Leyland, P. McQuilton, S. Marygold, G. Millburn, D. Osumi-Sutherland, A. Schroeder, R. Seal, and H. Y. Zhang. 2009. FlyBase: enhancing Drosophila Gene Ontology annotations. Nucleic Acids Research **37**:D555-D559.

Wagstaff, B. J., and D. J. Begun. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert drosophila. Genetics **177**:1023-1030.

Wagstaff, B. J., and D. J. Begun. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in Drosophila mojavensis and D-arizonae. Genetics **171**:1083-1101.

Walters, J. R., and R. G. Harrison. 2008. EST analysis of male accessory glands from Heliconius butterflies with divergent mating systems. BMC Genomics **9**.

Wedell, N. 2005. Female receptivity in butterflies and moths. Journal of Experimental Biology **208**:3433-3440.

Wedell, N., and P. A. Cook. 1999. Butterflies tailor their ejaculate in response to sperm competition risk and intensity. Proceedings of the Royal Society of London Series B-Biological Sciences **266**:1033-1039.

Yamamoto, K., J. Nohata, K. Kadono-Okuda, J. Narukawa, M. Sasanuma, S. Sasanuma, H. Minami, M. Shimomura, Y. Suetsugu, Y. Banno, K. Osoegawa, P. J. de Jong, M. R. Goldsmith, and K. Mita. 2008. A BAC-based integrated linkage map of the silkworm Bombyx mori. Genome Biology **9**.

Yang, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences **13**:555-556.

Yang, Z. H., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution **15**:496-503.

Yapici, N., Y. J. Kim, C. Ribeiro, and B. J. Dickson. 2008. A receptor that mediates the post-mating switch in Drosophila reproductive behaviour. Nature **451**:33-37.

Zdobnov, E. M., and R. Apweiler. 2001. InterProScan - an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17**:847-848.

DECOUPLING OF RAPID AND ADAPTIVE EVOLUTION AMONG REPRODUCTIVE PROTEINS IN *HELICONIUS* BUTTERFLIES WITH DIVERGENT MATING SYSTEMS

Introduction

The rapid evolution of reproductive proteins is a widespread but enigmatic phenomenon. Many proteins which have a direct role in reproductive processes, such as those mediating interactions between sperm and egg or those found in seminal fluid, tend to diverge rapidly between related species and often evolve via positive Darwinian selection (Swanson & Vacquier 2002b, Clark *et al.* 2006, Turner & Hoekstra 2008). This pattern has been observed across many animals and plants. However, the underlying processes are not well understood. It is a particularly compelling issue because mutations affecting reproductive processes should have substantial effects on fitness. Given the large potential for deleterious consequences, why do we not observe a pattern of extreme constraint among reproductive proteins? Several hypothesis have been proposed, but arguably the most promising– and certainly the most widely invoked – is that post-mating sexual selection is the primary factor driving the rapid and adaptive evolution of reproductive proteins (Swanson & Vacquier 2002b, Jensen-Seaman & Li 2003, Dorus *et al.* 2004, Andres *et al.* 2006, Clark *et al.* 2006, Herlyn & Zischler 2007, Haerty *et al.* 2007, Nadeau *et al.* 2007, Turner & Hoekstra 2008, Ramm *et al.* 2008, Herlyn & Zischler 2008).

The hypothesis that post-mating sexual selection increases the incidence of positive selection and, as a result, the overall evolutionary rates among reproductive proteins (henceforth the sexual selection hypothesis) offers at least one clear route for evaluation. Sperm competition, cryptic female choice, sexual conflict, and any other component of post-mating sexual selection should differ between divergent mating

systems, particularly in response to female polyandry. The more mates a female has, the more intense the post-mating sexual selection will be. Therefore, the sexual selection hypothesis predicts that evolutionary rate and frequency of adaptive evolution among reproductive proteins will correlate positively across mating systems with degree of female polyandry (Dorus *et al.* 2004, Herlyn & Zischler 2007, Nadeau *et al.* 2007, Ramm *et al.* 2008, Herlyn & Zischler 2008).

Several researchers have reported results suggesting such a relationship (Kingan *et al.* 2003, Clark & Swanson 2005, Wagstaff & Begun 2005, Wagstaff & Begun 2007, Almeida & DeSalle 2008, Almeida & DeSalle 2009, Martin-Coello *et al.* 2009). However, currently only a few studies have explicitly tested for an association between mating system and the evolutionary rate of reproductive proteins (Dorus *et al.* 2004, Herlyn & Zischler 2007, Nadeau *et al.* 2007, Ramm *et al.* 2008, Herlyn & Zischler 2008, Ramm *et al.* 2009). These studies generally support the sexual selection hypothesis, but most are quite limited in scope because each focuses on only one or a few proteins. These studies therefore demonstrate the potential for using a comparative approach to address this issue, but the accumulated data are not yet sufficient to clearly establish the role that post-mating sexual selection plays in the rapid evolution of reproductive proteins. Moreover, one of the most recent and ambitious studies showed only very limited support for the sexual selection hypothesis, despite using data from many proteins across many taxa. Of seven rodent seminal fluid proteins (SFPs) examined for a correlation between adaptive evolution and intensity of post-mating sexual selection, only one yielded a significant result (Ramm *et al.* 2008). Overall, relevant empirical results are thus far consistent with the sexual selection hypothesis, but far from definitive.

In this paper we extend this comparative approach both taxonomically and methodologically. Taxonomically, we present the first comparative analysis of

reproductive protein evolution in the Lepidoptera (moths and butterflies). Lepidopterans provide a rich framework for pursuing research involving comparisons between mating systems because: 1) they exhibit a wide diversity of mating systems between species, 2) courtship behavior and mating are often conspicuous and easily observed, and 3) in most species the male-derived spermatophore persists indefinitely in the female reproductive tract after mating, providing a reliable life-long record of female mating history (Scott 1972, Drummond 1984, Bissoondath & Wiklund 1995, Bissoondath & Wiklund 1996, Bergstrom & Wiklund 2002). Our work focuses on *Heliconius* butterflies, a neotropical genus containing about 40 species. These butterflies exhibit a striking dichotomy in mating systems. About half of *Heliconius* species display an unusual pupal mating behavior: females are mated before or during eclosion and typically mate only once (i.e. females are monandrous). Otherwise *Heliconius* butterflies mate as fully developed adults and typically mate multiple times (i.e. females are polyandrous) (Gilbert 1976, Deinert *et al.* 1994, Mallet *et al.* 1998, Deinert 2003, Cardoso *et al.* 2009). *Heliconius* species fall evenly into two major clades which correspond perfectly with mating system (Figure 3.1) (Beltran *et al.* 2007), providing an informative context for assaying differences in evolutionary patterns among reproductive proteins.

Previously we identified several dozen SFPs in two species of *Heliconius* butterflies, one pupal mater (*H. erato*) and one adult mater (*H. melpomene*). We also demonstrated that these SFPs were rapidly evolving in *Heliconius* and that at least one protein showed evidence for adaptive evolution. Given this precedent, testing for differences in evolutionary patterns and pressures among SFPs between adult mating and pupal mating *Heliconius* provides an important test of the sexual selection hypothesis.

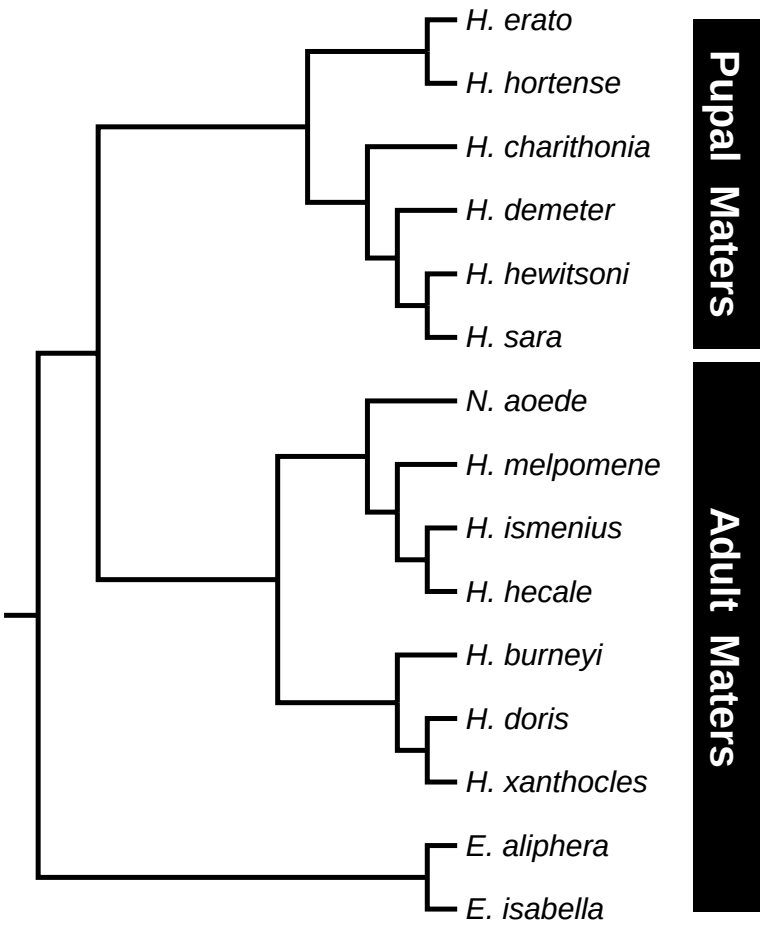

Figure 3.1. Phylogenetic relationships between species sampled in this study (adapted from Beltran *et al.* 2007).

Methodologically we have extended the comparison of reproductive protein evolution across mating systems by testing for directional differences in overall evolutionary rate separately from biases in the incidence of adaptive evolution. Contrary to what is predicted under the sexual selection hypothesis, we do not find concordance between these two assays of evolutionary patterns among reproductive proteins.

Methods

Samples and sequencing

Our analysis focuses on 18 SFPs that were proteomically detected in the spermatophores of one or both species of *Heliconius* reported by Walters & Harrison (2009). For comparison to the SFPs we selected 11 control loci from the complete set of control loci reported by Walters & Harrison (2009). Here the criteria for selecting control loci included having at least 150 aligned amino acids between *H. erato* and *H. melpomene* and having at least a few observed amino acid substitutions between these two species in the aligned regions. For each locus we manually designed degenerate PCR primers based on non-degenerate primers suggested by the Primer3 software (Rozen & Skaletsky H.J. 2000). Further details for all loci sampled are given in Table 3.1.

We used RT-PCR to amplify each locus from a panel of 21 male butterflies from 14 different species (Table 3.2). Typically two individuals per species were included in the panel, though limited availability of samples meant some species were represented by only a single individual. Taxa were sampled broadly across the *Heliconius* phylogeny, including a species from the genus *Neruda,* which renders *Heliconius* paraphyletic (Brower & Egan 1997, Beltran *et al.* 2007). To obtain outgroup sequences we also included two species from the genus *Eueides*, which is the sister taxon to *Heliconius*.

Total RNA was extracted from male abdomens using TRIzol reagent (Invitrogen, Carlsbad, CA, USA) and subsequently purified using RNeasy columns (Qiagen, Valencia, CA, USA). Between one and three μg total RNA was used to generate an amplified pool of double-stranded cDNA (Matz *et al.* 2003). A 50-fold

Table 3.1. Details of species sampling and putative function (if known) for each seminal fluid protein and control locus.

Locus	No. taxa sampled	Codons sampled	Annotations or putative function	*Heliconius ismenius* is	*Heliconius hecale* hl	*Neruda aoede* ao	*Heliconius burneyi* bu	*Heliconius doris* do	*Heliconiu xanthocles* xa	*Heliconius hortense* ht	*Heliconius hewitsoni* hw	*Heliconius sara* sr	*Heliconius demeter* dm	*Heliconius charitonia* ch	*Eueides isabella* ib	*Eueides aliphera* al
HACP001	13	271	chymotrypsin	99	105	427	426	108	NA	192	28	101	357	32	NA	423
HACP002	15	228	chymotrypsin	102	105	427	426	108	347	192	28	101	357	32	158	423
HACP003	14	95	chymotrypsin	99	105	427	426	108	347	192	28	101	357	32	NA	423
HACP004	13	85	NA	99	105	427	426	108	NA	192	28	101	357	32	158	NA
HACP006	13	203	NA	102	105	427	NA	98	347	191	28	101	357	32	172	NA
HACP010	13	319	chymotrypsin	99	105	427	426	NA	347	191	28	80	357	32	158	NA
HACP011	14	163	NA	102	105	427	426	108	347	192	28	101	357	32	NA	423
HACP012	13	220	NA	102	105	427	426	108	347	191	28	80	357	NA	172	NA
HACP016	14	114	NA	99	106	427	425	108	347	191	24	80	357	30	172	NA

Table 3.1 (continued)

Locus	No. taxa sampled	Codons sampled	Annotations or putative function	*Heliconius ismenius* is	*Heliconius hecale* hl	*Neruda aoede* ao	*Heliconius burneyi* bu	*Heliconius doris* do	*Heliconiu xanthocles* xa	*Heliconius hortense* ht	*Heliconius hewitsoni* hw	*Heliconius sara* sr	*Heliconius demeter* dm	*Heliconius charitonia* ch	*Eueides isabella* ib	*Eueides aliphera* al
HACP018	14	38	NA	99	106	424	425	98	347	191	24	80	357	32	172	NA
HACP020	14	139	NA	99	106	424	425	98	347	191	24	80	357	32	172	NA
HACP026	14	210	chymotrypsin	102	106	424	425	98	347	191	24	80	357	32	172	NA
HACP027	13	250	chymotrypsin	102	105	427	426	98	347	191	28	80	357	32	172	NA
HACP030	13	224	NA	99	106	NA	425	98	347	191	24	101	357	32	172	NA
HACP037	14	180	chymotrypsin	99	106	424	425	98	347	191	24	80	357	32	172	NA
HACP038	13	277	chymotrypsin	99	105	427	426	108	NA	192	28	101	357	32	NA	NA
HACP058	14	193	Aldo/keto reductase	102	106	427	425	108	347	191	28	80	357	30	158	NA
HACP059	14	182	proteinase inhibitor	99	105	427	426	108	347	192	28	101	357	32	NA	423
HCTL021	13	181	Protein disulphide isomerase	99	105	427	426	98	NA	191	28	101	357	32	158	NA
HCTL023	14	194	Chitin Binding Peritrophin-A	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL024	14	185	Porin (Voltage-dependent anion selective channel)	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL025	14	201	Ribosomal protein L6E	99	105	427	426	98	347	191	28	101	357	32	NA	423

Table 3.1 (continued)

Locus	No. taxa sampled	Codons sampled	Annotations or putative function	*Heliconius ismenius* is	*Heliconius hecale* hl	*Neruda aoede* ao	*Heliconius burneyi* bu	*Heliconius doris* do	*Heliconiu xanthocles* xa	*Heliconius hortense* ht	*Heliconius hewitsoni* hw	*Heliconius sara* sr	*Heliconius demeter* dm	*Heliconius charitonia* ch	*Eueides isabella* ib	*Eueides aliphera* al
HCTL026	12	204	Polyprenyl synthetase	99	105	NA	426	98	NA	191	28	101	357	32	158	NA
HCTL028	14	174	ATP synthase gamma subunit	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL029	14	196	G protein-coupled receptor associated sorting protein 1	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL033	14	184	Protein disulfide isomerase	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL034	14	150	Obstractor B (Tribolium)	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL035	14	207	EN10 protein (Eukaryotic translation initiation factor 3 subunit H)	99	105	427	426	98	347	191	28	101	357	32	158	NA
HCTL036	14	217	60S ribosomal protein L3	99	105	427	426	98	347	191	28	101	357	32	158	NA

Table 3. 2. Field collection information for tissue samples.

Taxon Code	Individual Identifier	Species		Source
al	423	*Eueides*	*aliphera*	Collected December 2008 near kilometer marker 24 on the road between Yurimaguas and Tarapoto, San Martin, Peru
ib	158	*Eueides*	*isabella*	Obtained from the Niagara Butterfly Conservatory, August, 2007
ib	172	*Eueides*	*isabella*	Obtained from the Niagara Butterfly Conservatory, August, 2007
hl	106	*Heliconius*	*hecale*	Collected March 2007 near Canazas, Panama
hl	105	*Heliconius*	*hecale*	Collected March 2007 near Canazas, Panama
is	102	*Heliconius*	*ismenius*	Collected March 2007 near Canazas, Panama
is	99	*Heliconius*	*ismenius*	Collected March 2007 near Canazas, Panama
ao	427	*Neruda*	*aoede*	Collected December 2008 near kilometer marker 24 on the road between Yurimaguas and Tarapoto, San Martin, Peru
ao	424	*Neruda*	*aoede*	Collected December 2008 near kilometer marker 24 on the road between Yurimaguas and Tarapoto, San Martin, Peru
bu	426	*Heliconius*	*burneyi*	Collected December 2008 near kilometer marker 24 on the road between Yurimaguas and Tarapoto, San Martin, Peru
bu	425	*Heliconius*	*burneyi*	Collected December 2008 near kilometer marker 24 on the road between Yurimaguas and Tarapoto, San Martin, Peru

Table 3.2 (continued)

Taxon Code	Individual Identifier	Species		Source
do	108	*Heliconius*	*doris*	Collected March 2007 near Gamboa, Panama
do	98	*Heliconius*	*doris*	Collected March 2007 near Canazas, Panama
xa	347	*Heliconius*	*xanthocles*	Collected Dec 2008 near Tarapoto, Peru
ht	192	*Heliconius*	*hortense*	Obtained from Houston Butterfly Gardens, October 2006
ht	192	*Heliconius*	*hortense*	Obtained from Houston Butterfly Gardens, October 2006
hw	28	*Heliconius*	*hewitsoni*	Obtained from culture matained by L. Gilbert, Univ. of Texas Austin, April 2006
hw	24	*Heliconius*	*hewitsoni*	Obtained from culture matained by L. Gilbert, Univ. of Texas Austin, April 2006
sr	101	*Heliconius*	*sara*	Collected March 2007 near Canazas, Panama
sr	101	*Heliconius*	*sara*	Collected March 2007 near Gamboa, Panama
dm	357	*Heliconius*	*demeter*	Collected Dec 2008 near Tarapoto, Peru

dilution of this amplified cDNA served as the template for all RT-PCRs. RT-PCR was performed using a touch-down thermocycling protocol, with an initial denaturation of 95°C (2 min), 10 cycles of 95°C (30 sec) then 60-51°C (30 sec, decreasing one degree per cycle) then 72°C (2 min), 25 cycles of 95°C (30 sec) then 50°C (30 sec) then 72°C (2 min), and a final extension of 72°C (4 min). PCR products were electrophoresed on 1.5% agarose gels stained with ethidium bromide and visualized under UV light. When loci amplified from two conspecific individuals, the reaction with a brighter band (higher concentrations of amplicon) was chosen for sequencing. Selected amplicons were enzymatically cleaned with EXOSAP, sequenced directly in both directions with Big Dye chemistry, and analyzed on an ABI 3730 automated sequencer.

Base-calling and assembly of chromatograms were performed using the phred-phrap algorithm as implemented in the CodonCode Aligner software (CodonCode Corp, Dedham, MA). All assembled contigs were trimmed of primer sequence, individually inspected, and (when necessary) edited manually. Sequence data from each locus were supplemented with previously determined sequences from *H. erato* and *H. melpomene* (Walters and Harrison 2009). For each locus, amino acid translations were aligned using clustalW and back-translated to the original DNA sequence using the BioEdit software.

Evolutionary Analyses

Maximum likelihood estimates of evolutionary rates, fitting of null and alternative models, and simulations were all performed using codon models implemented in the PAML v 4.2 software package (Yang 1997). Statistical tests were performed using either Microsoft Excel (Microsoft Corp., Redmond, WA) or the R statistical computing package (R Development Core Team 2005). All sets of analyses

were performed three times, each time with an independently determined genealogical relationship between taxa. First, we used a topology based on a previously published multi-locus, species-level molecular phylogeny of the Heliconiini tribe (Beltran *et al.* 2007). We also performed these tests with topologies inferred independently for each locus via maximum likelihood (ML) and Bayesian methodologies. ML trees were reconstructed using the DNAml application in the PHYLIP software package (Felsenstein 2005). Support values for each node were generated from 500 bootstrapped data sets. Bayesian trees were reconstructed using Mr. Bayes implementing the GTR + Γ + I nucleotide substitution model (Huelsenbeck & Ronquist 2001, Ronquist & Huelsenbeck 2003). We used the 50% majority rule consensus tree generated from sampling every 100 generations over 500,000 generation run with a burn-in of 1,250 generations. The fit of the codon models to the data under the one-ratio (M0) model using the different topologies was compared using the Akaike Information Criterion (AIC).

We focused on the ratio of the nonsynonymous substitution rate (dN) to the synonymous substitution rate (dS) as a measure of evolutionary change in our comparisons of adult versus pupal maters (Yang & Bielawski 2000, Anisimova & Kosiol 2009). The dN/dS ratio (symbolized ω) can be interpreted as a normalized measure of amino acid substitution. It can also be interpreted as an indicator of selective pressure experience by a protein coding sequence. Values of $\omega < 1$ correspond to evolutionary constraint while $\omega > 1$ is evidence of positive selection acting at that locus; $\omega = 1$ indicates neutral evolution.

With this focus on ω as a measure of evolutionary rate and pressures, we used *branch* codon models to test a null hypothesis with a single ω value (ω_0) for the entire phylogeny versus an alternative hypothesis with two ω values, ω_p for pupal mating clade and ω_a for the adult-mating clade including the outgroup; adult-mating is the

ancestral state (Yang 1997, Yang 1998, Anisimova & Kosiol 2009). Two versions of the alternative model were implemented such that the branch rooting the pupal mating clade was assigned either to ω_p or ω_a (Figure 3.2). The fit of null versus alternative models was statistically evaluated using a likelihood ratio test (LRT) where twice the difference in log-likelihood ($2\Delta\ln L$) between models is compared to a χ^2 distribution with degrees of freedom (d.f.) equal to the difference in number of model parameters. Such tests are typically administered individually to each locus. However, in this case we combined individual locus analyses into a single global test by summing $2\Delta\ln L$ across loci and comparing this to a χ^2 with d.f. equal to the number of combined loci. This global analysis provides a single test for an overall difference in evolutionary rate between the two mating systems (clades) for the SFP and control loci.

Unfortunately, this global analysis is a two-tailed test and the biological interpretation of a significant result (i.e. $\omega_p \neq \omega_a$) is problematic when the direction of the difference in evolutionary rates differs between loci (*e.g.* $\omega_p > \omega_a$ at one locus, but $\omega_p < \omega_a$ at another). The appropriate one-tailed test would involve an alternative two-ratio model where one estimate of ω is constrained to be greater than or equal to the other ω (i.e. $\omega_p \leq \omega_a$). PAML does not provide for such a constrained two-ratio model

to be evaluated directly, but it is possible to 'force' such a test by setting $2\Delta\ln L$ equal to zero when the difference in evolutionary rates runs counter to the assumed constraint (i.e. to test for more rapid evolution among adult maters, $2\Delta\ln L=0$ when $\omega_p > \omega_a$). We evaluated the significance of the observed post-hoc one-tailed test statistic by comparison to 1000 comparable tests performed on data simulated under the null model using PAML's *evolver* application (Yang 1997). One complication that arises in analyzing so many data sets is that occasionally the maximum likelihood algorithm fails to converge, producing a negative value for $2\Delta\ln L$. For this reason, we removed simulated data sets from the null distribution when any constituent locus produced a

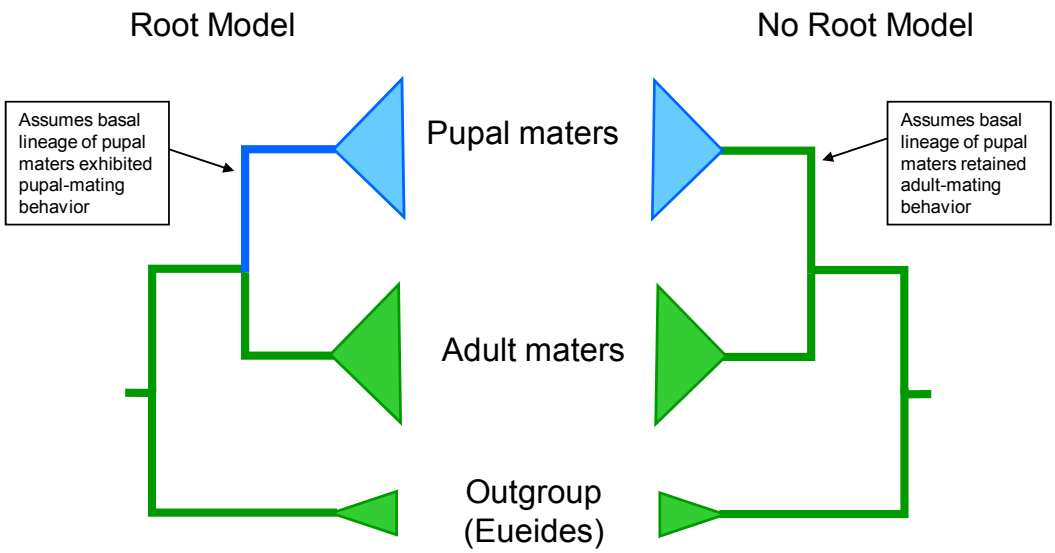

Figure 3.2. Graphical depiction of the difference between *root* and *no root* models implemented in codon models and tests for differences in molecular evolutionary rate between pupal-mating and adult-mating *Heliconius* butterflies.

negative value for 2ΔlnL. For example, if the alternative model failed to converge for the simulated data at the second locus of the 29 total loci in the 50th simulated data set, then all 29 loci in that 50th simulation were excluded. In no case were more than 60 of the 1000 simulated data sets removed from the focal analysis.

We further analyzed these data using the *site* models implemented in PAML, which provide sensitive tests for adaptive evolution occurring at just one or a few codon sites across the protein (Yang *et al.* 2000). We report results only for the M8a – M8 comparison, though we note that all loci produced comparable results under the M1-M2 and M7-M8 model comparisons as well. In this case each locus was tested independently via an LRT and a Bonferroni correction for 29 multiple tests was applied, requiring a nominal p-value of .0017 or less for statistical significance. Finally, we also used *branch-site* models to test for adaptive evolution at particular codon sites occurring only along particular lineages in the phylogeny (Yang & Nielsen 2002, Zhang *et al.* 2005). We applied the same branch partitioning as in the branch models implemented above, alternatively assigning pupal-maters or adult-maters (including the outgroup taxa) as the 'foreground' branches where $\omega > 1$ is allowed. Again each locus was tested independently via LRTs and a Bonferroni correction for 29 multiple tests was applied.

Results

Taxonomic sampling, phylogenetic analyses, and evolutionary models

With one exception, we obtained sequences at every locus from at least five adult mating species, five pupal mating species, and one outgroup species. For the control locus HCTL026, sequences from only four adult mating species were obtained. Including sequences from *H. melpomene* and *H. erato* yielded between 13 and 15

species sampled for each locus (12 for HCTL026) and a well-balanced representation between mating systems. Details for the taxa and sequence length sampled for each locus are given in Table 3.1.

The ML and MB phylogenies often differed slightly from the published phylogeny (Beltran *et al.* 2007), but were never grossly discordant with that phylogeny or with each other. Both analyses showed the pupal mating species as being monophyletic at all loci. However, this was not true for the adult mating *Heliconius* (including *N. aoede*). Quite often the *Eueides* outgroup species fell among the adult-mating species, usually with *N. aoede* or *H. burneyi* occupying the outgroup position in the ML and MB phylogenies. Nodes separating the adult mating *Heliconius* from the *Eueides* branch typically had low support values, indicating these sequences provide relatively little power to resolve the relationships among the basal nodes in the tree. This lack of resolution could, in itself, explain the discrepancies between the published species phylogeny and the ML and MB phylogenies. In other words, these discrepancies may reflect methodological artifacts arising from having insufficient data to accurately reconstruct the evolutionary history of these taxa. It can often be difficult to accurately estimate the basal relationships among rapidly radiating groups of species like *Heliconius* butterflies (Edwards *et al.* 2007). However, it is also expected that the random sorting of ancestral alleles during cladogenesis will produce discordant genealogies across loci and that this phenomenon will be more prominent among rapidly radiating lineages like *Heliconius* (Maddison & Knowles 2006). Thus it is possible that the phylogenetic discrepancies observed here reflect biological reality, and not methodological artifacts.

No matter what the causes, differences between the three phylogenies did not substantially affect the overall results obtained from tests for differential patterns of evolution between the two mating systems, which are discussed in detail below. In

particular, the site and branch-site analyses at individual loci gave consistent results across the three phylogenies. The branch analyses were slightly more variable between assumed topologies. There were two loci (HACP004 and HCTL024) where the estimate of ω_p being greater or less than ω_a differed between topologies. However, in both cases the inferred topology appeared to fit the data substantially better than the established species phylogeny ($\Delta AIC > 20$). Indeed, in most cases the ML and MB phylogenies provided a substantial improvement in the fit of the M0 model ($\Delta AIC>10$) over the published species phylogeny.

Because of this difference in model fit, and because the overall results are robust to the differences in topology tested here, we will henceforth assume results from the ML phylogenies as the default analysis. We prefer the ML over MB topologies because the Bayesian consensus tree frequently included polytomies at some nodes, which we *a priori* assume are uncommon in reality and therefore are unrealistic in an evolutionary model.

In addition to assaying different models regarding the evolutionary relationships between taxa, we also implemented two different models accounting for differences in the origin of the pupal mating behavior (Figure 3.2). Pupal mating is an obvious synapomorphy for the pupal-mating clade (Gilbert 1991, Beltran *et al.* 2007), but it is uncertain when this behavior arose in the ancestral pupal-mating lineage relative to that lineage's divergence from adult-mating *Heliconius*. To clarify this point, consider two possible scenarios which represent extremes of a continuum. In one case, the origin of the pupal mating behavior was concomitant with the origin of the pupal mating lineage. In the opposite case, the pupal mating behavior arose only when the first cladogenetic event within the pupal-mating lineage occurred. In the latter case, the basal lineage giving rise to pupal-mating species would have retained the adult mating habit. The reality probably falls somewhere between these

two, with the pupal mating behavior arising sometime after the divergence from adult-maters but substantially preceding subsequent diversification of more recent pupal mating lineages. However, these two extreme examples illustrate the scenarios that can be accommodated using branch models in PAML, where ω for the branch rooting the pupal clade is either lumped with the pupal maters or with the adult maters. We refer to these different model implementations as the *root* and *no-root* models, respectively.

We performed all tests using both the root and no-root models. As with the topologies, results from individual loci for the site and branch-site tests were robust to this difference in models. This was less true for the branch analyses where for five loci the estimate of whether ω_p was greater than ω_a differed between the root and no-root models. A few other loci showed a difference in significance between root and no-root models at a nominal p-value of 0.05, but none of these remained significant after a Bonferroni correction. Importantly, these different model implementations did not qualitatively affect the global analysis applied using the branch models. We therefore assume the root model as the focus of our results and discussion for the remainder of this manuscript.

Branch models and tests for differences in evolutionary rate.

A two-tailed global analysis of the 18 SFP genes significantly rejected the null hypothesis of equal evolutionary rates among adult and pupal mating *Heliconius* ($\omega_p \neq \omega_a$, $p < 0.001$, χ^2 d.f. = 18). A comparable test for the 11 control genes did not indicate any difference in evolutionary rates between mating system ($\omega_p = \omega_a$, $p > 0.05$, χ^2 d.f. = 11). Taken together these two tests indicate a significant difference in evolutionary rates between mating systems that is exclusive to reproductive proteins. However, this result is problematic to interpret biologically because whether ω_p or ω_a

is greater varies across loci. Still, considered individually, most loci fit two-ratio models which estimate $\omega_p > \omega_a$. This group includes HACP026, the only locus showing a significantly better fit of the two-ratio model after a Bonferroni correction, as well as HACP003 and HACP027, which both show a better fit to the two-ratio model that is marginally significant (nominal $p < 0.01$). A further complication was the fact that the estimate of ω_p being greater or less than ω_a differed between the root and no-root models at four SFP loci, though none of the individual LRTs were significant and under both models a substantial majority of loci still showed $\omega_p > \omega_a$.

Despite the variability across loci, the general trend was clearly towards more rapid evolution among pupal maters. We sought to confirm this result by conducting post-hoc one-tailed global tests. As with the global two-tailed test, this directional one-tailed test combines results across loci by summing $2\Delta\ln L$ as individually determined at each locus. However, unlike the two-tailed test, loci are counted towards the value of the final test statistic only when the difference between estimates of ω_p and ω_a agrees with the direction of the test (i.e. adult-maters faster or pupal-maters faster). Thus the test takes into account how well the two-ratio model improves fit to the data as well as whether the estimates of ω are consistent with an *a priori* directional prediction. Significance of the observed test statistic can be determined by comparison to the distribution of test statistics derived from applying the same one-tailed test to data simulated under the null hypothesis of a single ω value shared between mating systems.

This post-hoc test strongly indicates more rapid evolution of SFPs among pupal maters, but does not indicate any differences in rates at control loci (Table 3.3 and Figure 3.3). There is a concern that this result could arise primarily because of a single outlier locus. In particular, the value of $2\Delta\ln L$ for HACP026 is nearly double the next largest value of $2\Delta\ln L$, so this locus contributes disproportionately to the final

value of the test statistic for $\omega_p > \omega_a$. However, reanalyzing the data excluding this locus still gives a highly significant results (p =0.001), so the overall results do not depend on including this locus.

Table 3.3. Results of post-hoc one-tailed tests for differences in overall evolutionary rate among control and seminal fluid proteins (SFPs) between pupal-mating and adult-mating *Heliconius* butterflies.

Topology	Protein Class	Model	Test	p-value	Number sims
Maximum Likelihood	SFPs	Root	$\omega_p > \omega_a$	0	986
			$\omega_p < \omega_a$	0.985	
		No Root	$\omega_p > \omega_a$	0	962
			$\omega_p < \omega_a$	0.928	
	Control	Root	$\omega_p > \omega_a$	0.231	989
			$\omega_p < \omega_a$	0.311	
		No Root	$\omega_p > \omega_a$	.053	944
			$\omega_p < \omega_a$	.561	
Species	SFPs	Root	$\omega_p > \omega_a$	0	983
			$\omega_p < \omega_a$	0.994	
		No Root	$\omega_p > \omega_a$	0	968
			$\omega_p < \omega_a$	0.911	
	Control	Root	$\omega_p > \omega_a$	0.207	978
			$\omega_p < \omega_a$	0.312	
		No Root	$\omega_p > \omega_a$	0.042	949
			$\omega_p < \omega_a$	0.555	
Bayesian (Mr. Bayes)	SFPs	Root	$\omega_p > \omega_a$	0	968
			$\omega_p < \omega_a$	0.916	
		No Root	$\omega_p > \omega_a$	0	988
			$\omega_p < \omega_a$	0.980	
	Control	Root	$\omega_p > \omega_a$	0.252	995
			$\omega_p < \omega_a$	0.350	
		No Root	$\omega_p > \omega_a$	0.057	947
			$\omega_p < \omega_a$	0.594	

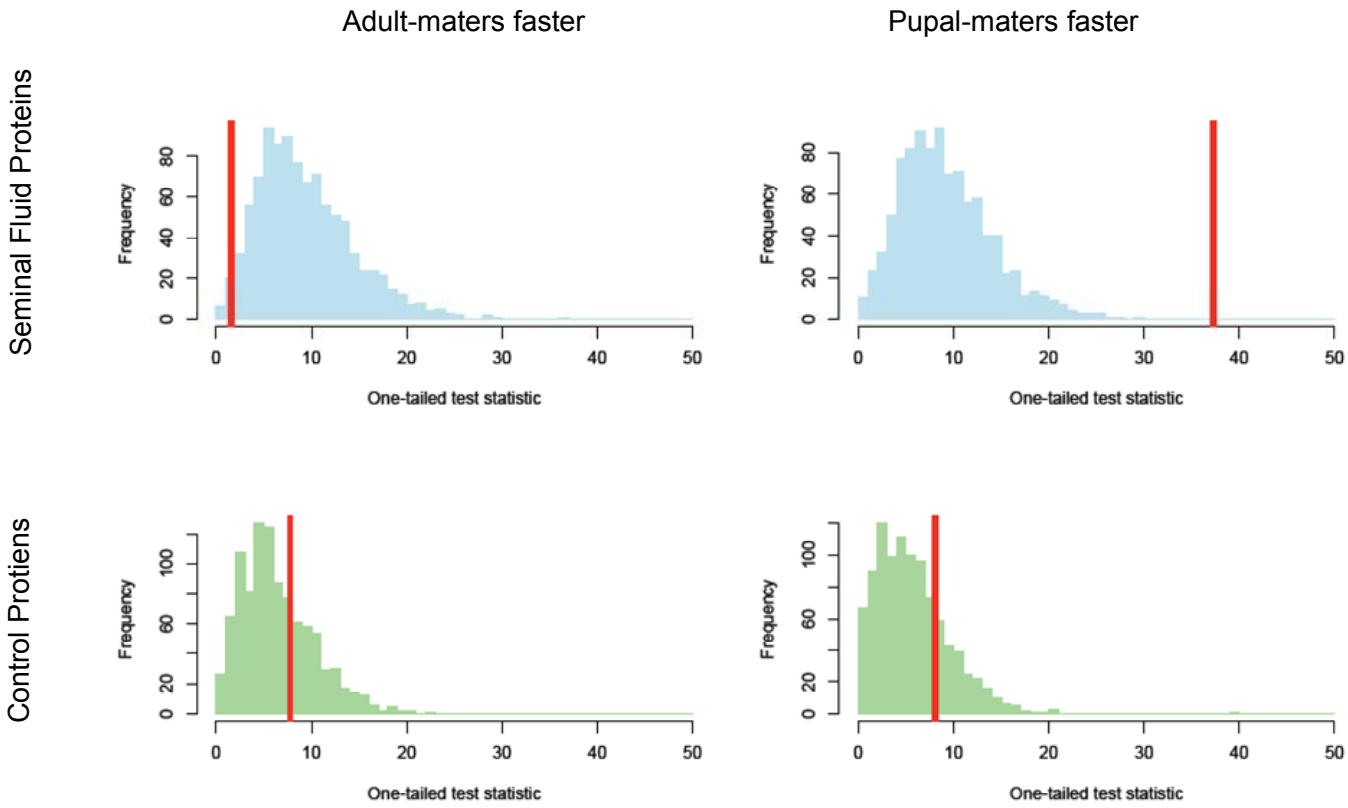

Figure 3.3. Comparison of observed values (red line) and simulated null distributions for the directional post-hoc one tailed tests for differential evolutionary rates between adult and pupal mating *Heliconius* butterflies based on maximum likelihood estimates of phylogeny. Results are comparable using Bayesian and the published species phylogenies.

Tests for adaptive evolution

We tested for positive selection ($\omega > 1$) at each locus by comparing the M8a vs. M8 site models implemented in PAML (Swanson *et al.* 2003). Only one locus, HACP004, showed unequivocal evidence of adaptive evolution after a Bonferroni correction (nominal $p < 0.0001$, $2\Delta\ln L = 22.75$, χ^2 d.f. = 1). Bayes Empirical Bayes (BEB) analysis of codon sites indicated three sites with > 95% posterior probability of being positively selected (Yang *et al.* 2005). Another locus, HACP020, showed moderate evidence of recent adaptive evolution. In this case the M8a-M8 LRT was only marginally significant after a Bonferroni correction (nominal $p < 0.004$, $2\Delta\ln L = 8.4$, χ^2 d.f. = 1), but we note that both a Bonferroni correction and using a χ^2_1 test for this model comparison are both known to be conservative statistical procedures. Moreover, the M7-M8 comparison was significant for this locus after a Bonferroni correction (nominal $p < 0.001$, $2\Delta\ln L = 14.99$, χ^2 d.f. = 2). We therefore interpret these results as implicating adaptive evolution in this protein's recent history; the BEB analysis indicates only one codon with > 95% probability of being positively selected.

Although these site model comparisons provide a powerful means to detect adaptive evolution, they provide little indication as to where on the phylogeny positive selection has acted. In order to better visualize this and identify differences between the two mating systems in the incidence of adaptive evolution, we mapped nonsynonymous substitutions onto the corresponding phylogenies and noted on which branches these substitutions occurred at positively selected sites (Figures 3.4 and 3.5). Both loci showed the same pattern. There were twice as many branches with adaptive changes in the adult mating clade than in the pupal mating clade. Moreover, in HACP004, the adult mating clade had several branches with changes at two or three adaptively evolving sites while among pupal maters there was never more than one adaptive substitution per branch. Thus both loci distinctly indicate more frequent or

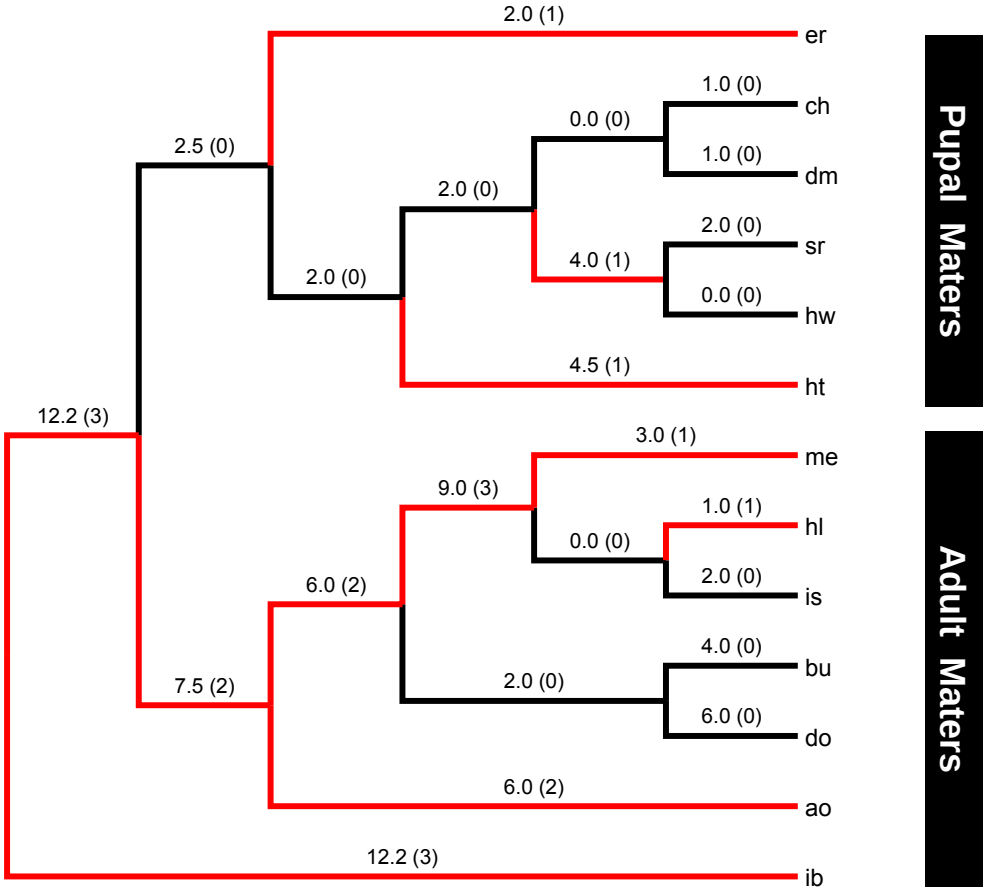

Figure 3.4. Cladogram of HACP004 (based on ML phylogeny) with counts of estimated nonsynonymous codon substitutions per branch. Substitutions at sites >95% probability $\omega > 1$ are in parentheses.

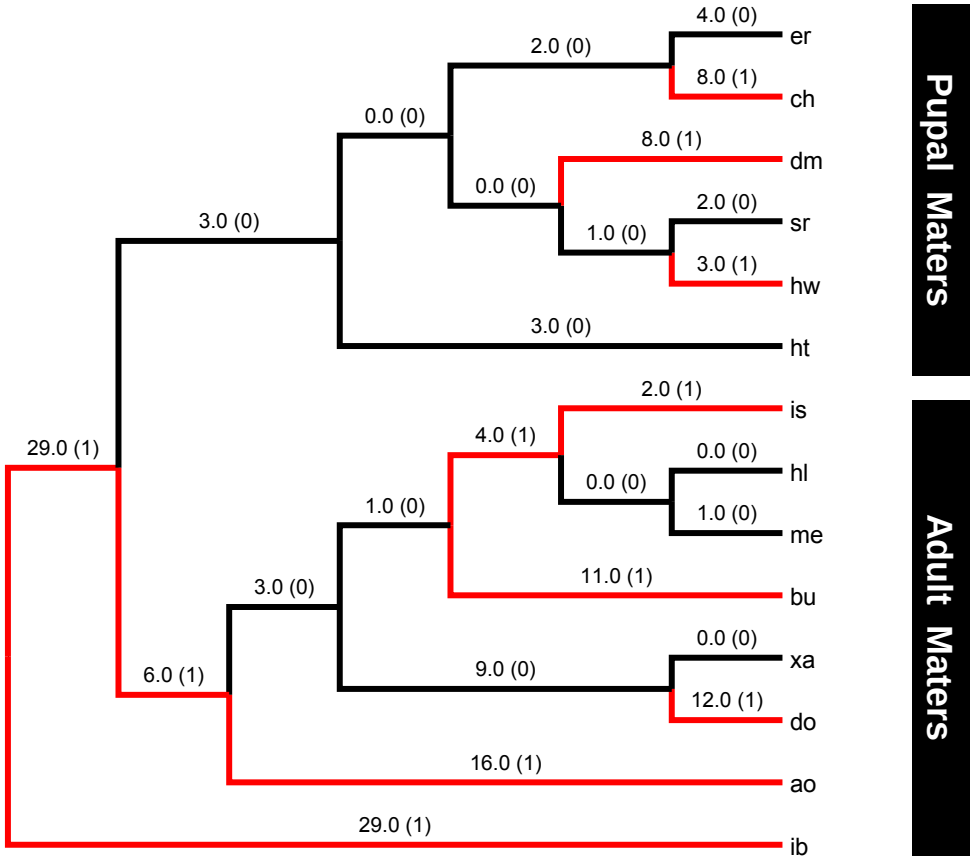

Figure 3.5. Cladogram of HACP020 (based on ML phylogeny) with counts of estimated nonsynonymous codon substitutions per branch. Substitutions at sites >95% probability $\omega > 1$ are in parentheses.

intense positive selection among adult mating *Heliconius*.

This interpretation was supported, at least for HACP004, by results from branch-site tests. We implemented the "branch-site test for adaptive evolution" (Zhang *et al.* 2005) twice for each locus, once designating the adult maters (including outgroup) as foreground branches and once designating the pupal maters as foreground branches. With this test none of the loci showed evidence for adaptive evolution among pupal mating lineages. In contrast, the test strongly indicated adaptive evolution among adult-mating lineages for HACP004 (nominal $p < 0.001$, $2\Delta\ln L = 17.86$, χ^2 d.f. = 1). No other sites, including HACP020, showed even marginally significant results after a Bonferroni correction.

Discussion

The observation of a few adaptively evolving proteins among many rapidly evolving proteins is now a well established phenomenon in many different taxa (Swanson & Vacquier 2002b, Clark *et al.* 2006, Turner & Hoekstra 2008). Previously we demonstrated that, like in other taxa, reproductive proteins (or at least SFPs) are rapidly evolving in *Heliconius* butterflies (Walters and Harrison 2009). Similarly, the results presented here implicate a role for adaptive evolution in explaining this widespread pattern of rapid evolution among reproductive proteins. While the process underlying these patterns has not yet been confidently determined, it is certainly true that the most widely invoked hypothesis is that post-mating sexual selection causes both the few incidents of adaptive evolution as well as the more widespread observation of rapid evolution (Swanson *et al.* 2003, Clark & Swanson 2005, Andres *et al.* 2006, Haerty *et al.* 2007, Ramm *et al.* 2009). Here we have tested this hypothesis by contrasting the molecular evolution of SFPs between adult mating and pupal mating *Heliconius* butterflies. If the sexual selection hypothesis were true, we

would expect to see both more rapid evolution and a higher incidence of adaptive evolution among the polyandrous adult-maters relative to the monandrous pupal-maters. Yet this is not what we observe and our results do not support a strict interpretation of the hypothesis that the rapid evolution of reproductive proteins results solely from adaptive molecular evolution arising from post-mating sexual selection.

To be clear, our results are at least partially consistent with the sexual selection hypothesis. When adaptive evolution occurred it was distinctly more prevalent among the adult-maters. This is exemplified by HACP004, which provided not only the strongest evidence for adaptive evolution, but also the clearest signal of positive selection being biased towards adult-maters in both the mapping of substitutions onto the phylogeny and in the branch-site tests. HACP020 also trended in this direction, as indicated by mapping of substitutions onto the phylogeny, though the branch-site test did not statistically confirm this trend. However, our results also indicate that the adaptive evolution of reproductive proteins can be decoupled from an overall pattern of rapid evolution. The global analyses testing for an overall difference in evolutionary rates show that it is the pupal mating lineage where SFPs, on average, evolve more rapidly. This is an unexpected result and is not consistent with the hypothesis that the rapid evolution of reproductive proteins results from post-mating sexual selection. In this case, evolutionary rates of reproductive proteins appear to be elevated in the monandrous taxa relative to the polyandrous taxa and are therefore associated with a reduction in the intensity of post-mating sexual selection.

This result is robust to different assumptions regarding the geneaologies of the surveyed loci. It is also does not depend on whether we assume the basal lineage giving rise to pupal mating species was adult or pupal mating. Nonetheless, it is worth considering how misclassifying the ancestral pupal lineage would affect our tests for a difference in evolutionary rates or adaptive evolution between the mating systems.

We argue that in either case it makes the test more conservative (i.e. less likely to reject the null hypothesis of no difference when the alternative is true). Consider the scenario where the sexual selection hypothesis is correct and adaptive evolution drives the rapid evolution of reproductive proteins, but the root-model is inaccurate and the pupal mating ancestor was actually an adult-mating species. This would cause a polyandrous lineage with adaptively evolving reproductive proteins to be erroneously included with otherwise monandrous lineages and would act to inflate estimates of the evolutionary rates of that group, making it more difficult to reject the null hypothesis of no difference in rates. Alternatively, if the no-root model were wrong, the ancestral pupal mating species would have been monandrous, but assumed to be polyandrous. This scenario would also make it more difficult to reject the null hypothesis because it would lump a more slowly evolving monandrous lineage with otherwise adaptively evolving polyandrous taxa, reducing the estimated difference between the two groups. It is not clear which of these two scenarios is the more conservative when simply testing for differences in evolutionary rates. This would depend on the strength of sexual selection in the root scenario and, in contrast, on the extent of evolutionary constraint in the no-root scenario. However, we have chosen the root model as the default for our analyses and consider it slightly more conservative, at least in regard to detecting adaptive evolution, because including the ancestral lineage with the pupal maters increases the sum of branch lengths for that group. Increased branch length would increase power to detect adaptive evolution in the pupal-mating lineage using branch-site models (Anisimova *et al.* 2001), a result which is not expected under the sexual selection hypothesis. Ultimately branch-site analyses did not indicate any adaptive evolution among pupal maters, despite the associated higher overall evolutionary rate of SFPs.

This decoupling of rapid evolution from adaptive evolution among reproductive proteins does not fit well with any single current hypothesis explaining the rapid evolution of reproductive proteins (reviewed in (Swanson & Vacquier 2002a, Swanson & Vacquier 2002b, Clark *et al.* 2006, Turner & Hoekstra 2008). In particular, an observation of elevated evolutionary rates among monandrous species relative to polyandrous species distinctly contradicts the sexual selection hypothesis. If these results are replicated in other taxa, it will mean that a comprehensive explanation for the widespread observation of rapid and frequently adaptive evolution among reproductive proteins will require invoking either a novel hypothesis or some combination of current hypotheses.

One possible explanation for the patterns observed here would be that SFPs experience relaxed constraint in pupal maters. Perhaps the focus on adaptive evolution that has heretofore accompanied explanations for the rapid evolution of reproductive proteins has caused an underestimation of the role that reduced constraint can play in broadly elevating the evolutionary rates of reproductive proteins. Relaxed constraint has been previously proposed as a hypothesis for the rapid evolution of reproductive proteins, though only in the context of concerted evolution in highly repetitive proteins (Swanson & Vacquier 1998, Swanson & Vacquier 2002a). We suggest that transitions between mating systems can alter the functionality of reproductive proteins in a way that broadly reduces constraint and allows more rapid evolution.

The transition to pupal mating in *Heliconius* is accompanied by other striking changes in post-mating reproductive phenotypes. For example, unlike nearly all other butterflies, the spermatophores in pupal-mating *Heliconius* completely degrade in a relatively short period of time (Boggs 1979, Boggs 1981, Deinert 2003). While the molecular-genetic basis underlying this transition is unknown, it seems likely to result

from loss-of-function mutations at one or a few loci. Such mutations should reduce constraint on these causal loci as well as on interacting proteins. In many insects the protein structural components of spermatophores are secreted by the male accessory gland along with other SFPs, and this is likely to be true in *Heliconius* as well (Walters & Harrison 2008). It therefore seems plausible that reduced constraint on proteins associated with the transition to rapidly degrading spermatophores could, in part, be directly responsible for the overall increase in the evolutionary rate of SFPs in pupal maters.

The role of relaxed constraint in elevating the evolutionary rates of reproductive proteins might be considerable if post-mating reproductive phenotypes have genetic architectures that tend to be epistatic or polygenic. Such genetic architectures might constrain evolutionary rates such that a loss-of-function mutation at one locus would have a cascading effect, reducing constraint and increasing evolutionary rate on many others (Fraser & Hirsh 2004, Weinreich *et al.* 2005, Schlosser & Wagner 2008). Consider another morphological transition corresponding to the origin of pupal mating in *Heliconius*: the loss of signa. Signa are scleritized rasp-like protrusions on the interior of the bursa copulatrix, the organ where the spermatophore is received from the male. Signa are believed to function in the emptying and collapse of spermatophores, the remnants of which typically persist indefinitely in the bursa copulatrix of the female (Gilbert 2003, Galicia *et al.* 2008). Presumably signa were lost in pupal maters because rapidly degrading spermatophores rendered them unnecessary. Even if the loss of signa is adaptive, such a substantial morphological transition will necessarily reduce constraint on loci which formally functioned to produce signa in the ancestral lineage.

In contrast to the case of the spermatophores, we do not mean to suggest there could be a direct link between the formation (or loss) of signa and the observed

difference in evolutionary rates among *Heliconis* SFPs. Rather, we cite the signa as an example of a rapid transition in reproductive phenotype which could reduce constraint and accelerate evolution among associated proteins. Such substantial transitions of post-mating reproductive phenotypes associated with shifts in mating system are not unique to *Heliconius*. A similar phenomenon appears in eusocial bees, where polyandrous honey bees (genus A*pis*) have large accessory glands while monandrous stingless bees (Tribe: Meliponini) have lost their accessory glands (Colonello & Hartfelder 2005). Curiously, the opposite trend occurs in attine fungus-gardening ants, where the evolutionary transition from monandry to polyandry coincides perfectly with the loss of male accessory glands (Baer & Boomsma 2004, Mikheyev 2004). It therefore seems likely that this phenomenon also occurs among reproductive phenotypes that manifest only at the molecular level. There is growing evidence, particularly from *D. melanogaster* that the function of SFPs are highly interdependent. One study reports a protease found in seminal fluid which is necessary to functionally activate two other SFPs which, in turn, directly influence sperm storage and ovulation in females (Ravi *et al.* 2006). Other recent work implicates interactions between no less than five SFPs in generating the so-called 'long term post-mating response' in females, which includes increased oviposition, effective sperm storage, and a reluctance to remate (Ram & Wolfner 2007) (Ram and Wolfner, unpublished results). If SFPs do primarily function in networks, the potential then exists for the functional disruption of a single SFP to reduce constraint and accelerate evolution among many other SFPs.

In particular, a shift in mating system could potentially have opposing effects on the selective regime experienced by SFPs depending on what selective pressures existed in the ancestor. This potentially confounds the simple expectation under the sexual selection hypothesis of more rapid and adaptive evolution of reproductive

proteins in polyandrous mating systems relative to monandrous ones. For *Heliconius* the evolutionary transition in question was from a polyandrous to a monandrous mating system. While an emphasis on sexual selection leads to the expectation of slower evolution among pupal maters due to less frequent adaptive evolution, other proteins may have experienced relaxed constraint as a result of the shift in mating system. This reduced constraint could produce a general pattern of more rapid evolution among pupal maters even in the absence of adaptive evolution due to post-mating sexual selection. Contrasting the incidence of adaptive evolution versus the overall evolutionary rate between sister taxa with divergent mating systems, as has been done here, offers a way to test this combinatorial hypothesis of sexual selection and reduced constraint in future studies. When monandry is derived, then the decoupling of adaptive evolution (due to sexual selection) from rapid evolution (due to relaxed constraint) would be predicted. But when monandry is ancestral, then the shift to a polyandrous mating system should make more rapid and adaptive evolution of reproductive proteins coincident in the derived lineage.

It should not be overlooked that other possible explanations exist for the rapid and adaptive evolution of reproductive proteins that have not yet been carefully examined empirically (Swanson & Vacquier 2002a). These include the hypothesis that avoiding or deterring pathogens drives reproductive protein evolution. A surprising number of SFPs in *Drosophila* are predicted to have antimicrobial functions and it is well known that the evolutionary antagonism between host and pathogen often results in rapid and adaptive protein evolution (Tennessen 2005, Lazzaro 2008). Other suggestions include an enhanced role for gene duplication as well as reinforcement during speciation (Swanson & Vacquier 2002a)

There is a strong and sustained interest among evolutionary researchers in detecting positive selection and adaptive evolution at the molecular level (Ellegren

2008). Currently genomic and proteomic technologies are advancing with unprecedented speed, making novel surveys of reproductive protein evolution in many new taxa increasingly tractable (Mitchell-Olds *et al.* 2008). Taken together, these two facets of contemporary evolutionary genetic research leave little doubt many such studies will be designed and executed in the near future. We strongly advocate that these future studies focus on taxa where informative contrasts between mating systems or other characters will allow the critical evaluation of competing hypotheses potentially explaining the rapid and adaptive evolution of reproductive proteins.

REFERENCES

Almeida, F. C., and R. DeSalle. 2008. Evidence of adaptive evolution of accessory gland proteins in closely related species of the Drosophila repleta group. Molecular Biology and Evolution **25**:2043-2053.

Almeida, F. C., and R. DeSalle. 2009. Orthology, Function and Evolution of Accessory Gland Proteins in the Drosophila repleta Group. Genetics **181**:235-245.

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Molecular Biology and Evolution **23**:1574-1584.

Anisimova, M., J. P. Bielawski, and Z. H. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Molecular Biology and Evolution **18**:1585-1592.

Anisimova, M., and C. Kosiol. 2009. Investigating Protein-Coding Sequence Evolution with Probabilistic Codon Substitution Models. Molecular Biology and Evolution **26**:255-271.

Baer, B., and J. J. Boomsma. 2004. Male reproductive investment and queen mating-frequency in fungus-growing ants. Behavioral Ecology **15**:426-432.

Beltran, M., C. D. Jiggins, A. V. Z. Brower, E. Bermingham, and J. Mallet. 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.

Bergstrom, J., and C. Wiklund. 2002. Effects of size and nuptial gifts on butterfly reproduction: can females compensate for a smaller size through male-derived nutrients? Behavioral Ecology and Sociobiology **52**:296-302.

Bissoondath, C. J., and C. Wiklund. 1995. Protein-Content of Spermatophores in Relation to Monandry Polyandry in Butterflies. Behavioral Ecology and Sociobiology **37**:365-371.

Bissoondath, C. J., and C. Wiklund. 1996. Male butterfly investment in successive ejaculates in relation to mating system. Behavioral Ecology and Sociobiology **39**:285-292.

Boggs, C. L. 1979. Resource Allocation and Reproductive Strategies in Several Heliconine Butterfly Species. Thesis. University of Texas, Austin.

Boggs, C. L. 1981. Selection Pressures Affecting Male Nutrient Investment at Mating in Heliconiine Butterflies. Evolution **35**:931-940.

Brower, A. V. Z., and M. G. Egan. 1997. Cladistic analysis of Heliconius butterflies and relatives (Nymphalidae: Heliconiiti): a revised phylogenetic position for Eueides based on sequences from mtDNA and a nuclear gene. Proceedings of the Royal Society of London Series B-Biological Sciences **264**:969-977.

Cardoso, M. Z., J. J. Roper, and L. E. Gilbert. 2009. Prenuptial agreements: mating frequency predicts gift-giving in Heliconius species. Entomologia Experimentalis et Applicata **131**:109-114.

Clark, N. L., J. E. Aagaard, and W. J. Swanson. 2006. Evolution of reproductive proteins from animals and plants. Reproduction **131**:11-22.

Clark, N. L., and W. J. Swanson. 2005. Pervasive Adaptive Evolution in Primate Seminal Proteins. PLoS Genet. **1**:e35.

Colonello, N. A., and K. Hartfelder. 2005. She's my girl - male accessory gland products and their function in the reproductive biology of social bees. Apidologie **36**:231-244.

Deinert, E. I., J. T. LONGINO, and L. E. Gilbert. 1994. Mate Competition in Butterflies. Nature **370**:23-24.

Deinert, E. L. 2003. Mate location and competition for mates in a pupal mating butterfly. Pages 91-108 *in* Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago.

Dorus, S., P. D. Evans, G. J. Wyckoff, S. S. Choi, and B. T. Lahn. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nature Genetics **36**:1326-1329.

Drummond, B. A. 1984. Multiple mating and sperm competition in the Lepidoptera. Pages 291-370 *in* Smith R.L. editor. Sperm Competition and the Evolution of Animal Mating Systems. Academic Press, New York.

Edwards, S. V., L. Liu, and D. K. Pearl. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences of the United States of America **104**:5936-5941.

Ellegren, H. 2008. Comparative genomics and the study of evolution by natural selection. Molecular Ecology **17**:4586-4596.

Felsenstein J. PHYLIP (Phylogeny Inference Package). [3.6]. 2005. University of Washington, Seattle, Department of Genome Sciences.
Ref Type: Computer Program

Fraser, H. B., and A. E. Hirsh. 2004. Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. Bmc Evolutionary Biology **4**.

Galicia, I., V. Sanchez, and C. Cordero. 2008. On the function of signa, a genital trait of female Lepidoptera. Annals of the Entomological Society of America **101**:786-793.

Gilbert, L. E. 2003. Adaptive noverlty through introgression in Heliconius Wing Patters: Evidence for a shared Genetic "Toolbox" from Synthetic hybrid zones and a theory of diversification. Pages 281-318 *in* Butterflies: ecology and evolution taking flight. University of Chicago Press, Chicago.

Gilbert, L. E. 1991. Biodiversity of a Central American Heliconius community: pattern, process, and problems. Pages 403-427 *in* P. W. Price, T. M. Lewinsohn, G. W. Fernandes, and W. W. Benson editors. Plant-Animal Interactions Evolutionary Ecology in Tropical and Temperate Regions. John Wiley and Sons.

Gilbert, L. E. 1976. Postmating Female Odor in Heliconius Butterflies - Male-Contributed Anti-Aphrodisiac. Science **193**:419-420.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. R. Ram, L. K. Sirot, L. Levesque, C. G. Artieri, M. F. Wolfner, A. Civetta, and R. S. Singh. 2007. Evolution in the fast lane: Rapidly evolving sex-related genes in drosophila. Genetics **177**:1321-1335.

Herlyn, H., and H. Zischler. 2007. Sequence evolution of the sperm ligand zonadhesin correlates negatively with body weight dimorphism in primates. Evolution Int.J.Org.Evolution **61**:289-298.

Herlyn, H., and H. Zischler. 2008. The molecular evolution of sperm zonadhesin. International Journal of Developmental Biology **52**:781-790.

Huelsenbeck, J. P., and F. Ronquist. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics **17**:754-755.

Jensen-Seaman, M. I., and W. H. Li. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. Journal of Molecular Evolution **57**:261-270.

Kingan, S. B., M. Tatar, and D. M. Rand. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. Journal of Molecular Evolution **57**:159-169.

Lazzaro, B. P. 2008. Natural selection on the Drosophila antimicrobial immune system. Curr.Opin.Microbiol. **11**:284-289.

Maddison, W. P., and L. L. Knowles. 2006. Inferring phylogeny despite incomplete lineage sorting. Systematic Biology **55**:21-30.

Mallet, J., W. O. McMillan, and C. D. Jiggins. 1998. Estimating the mating behavior of a pair of hybridizing Heliconius species in the wild. Evolution **52**:503-510.

Martin-Coello, J., H. Dopazo, L. Arbiza, J. Ausio, E. R. Roldan, and M. Gomendio. 2009. Sexual selection drives weak positive selection in protamine genes and high promoter divergence, enhancing sperm competitiveness. Proc.Biol.Sci.

Matz, M., N. Alieva, A. Chenchik, and S. Lukyanov. 2003. Amplification of cDNA ends using PCR suppresion effect and step-out PCR. Pages 41-50 *in* S. Y. Ying editor. Generation of cDNA libraries. Humana Press, Totowa.

Mikheyev, A. S. 2004. Male accessory gland size and the evolutionary transition from single to multiple mating in the fungus-gardening ants. Journal of Insect Science **4**.

Mitchell-Olds, T., M. Feder, and G. Wray. 2008. Evolutionary and ecological functional genomics. Heredity **100**:101-102.

Nadeau, N. J., T. Burke, and N. I. Mundy. 2007. Evolution of an avian pigmentation gene correlates with a measure of sexual selection. Proceedings of the Royal Society B-Biological Sciences **274**:1807-1813.

R Development Core Team. R: A language and environment for statistical computing. 2005. Vienna, Austria, R Foundation for Statistical Computing.
Ref Type: Computer Program

Ram, K. R., and M. F. Wolfner. 2007. Seminal influences: Drosophila Acps and the molecular interplay between males and females during reproduction. Integrative and Comparative Biology **47**:427-445.

Ramm, S. A., L. McDonald, J. L. Hurst, R. J. Beynon, and P. Stockley. 2009. Comparative Proteomics Reveals Evidence for Evolutionary Diversification of Rodent Seminal Fluid and Its Functional Significance in Sperm Competition. Molecular Biology and Evolution **26**:189-198.

Ramm, S. A., P. L. Oliver, C. P. Ponting, P. Stockley, and R. D. Emes. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Molecular Biology and Evolution **25**:207-219.

Ravi, R. K., L. K. Sirot, and M. F. Wolfner. 2006. Predicted seminal astacin-like protease is required for processing of reproductive proteins in Drosophila melanogaster. Proc.Natl.Acad.Sci.U.S.A **103**:18674-18679.

Ronquist, F., and J. P. Huelsenbeck. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19**:1572-1574.

Rozen, S., and Skaletsky H.J. 2000. Primer3 on the WWW for general users and for biologist programmers. Pages 365-386 *in* S. M. S. Krawetz editor. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press, Totowa, NJ.

Schlosser, G., and G. P. Wagner. 2008. A simple model of co-evolutionary dynamics caused by epistatic selection. Journal of Theoretical Biology **250**:48-65.

Scott, J. A. 1972. Mating of Butterflies. Journal of Resarch on the Lepidoptera **11**:99-127.

Swanson, W. J., R. Nielsen, and Q. F. Yang. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Molecular Biology and Evolution **20**:18-20.

Swanson, W. J., and V. D. Vacquier. 2002a. Reproductive protein evolution. Annual Review of Ecology and Systematics **33**:161-179.

Swanson, W. J., and V. D. Vacquier. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science **281**:710-712.

Swanson, W. J., and V. D. Vacquier. 2002b. The rapid evolution of reproductive proteins. Nature Reviews Genetics **3**:137-144.

Tennessen, J. A. 2005. Molecular evolution of animal antimicrobial peptides: widespread moderate positive selection. Journal of Evolutionary Biology **18**:1387-1394.

Turner, L. M., and H. E. Hoekstra. 2008. Causes and consequences of the evolution of reproductive proteins. International Journal of Developmental Biology **52**:769-780.

Wagstaff, B. J., and D. J. Begun. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert drosophila. Genetics **177**:1023-1030.

Wagstaff, B. J., and D. J. Begun. 2005. Molecular population genetics of accessory gland protein genes and testis-expressed genes in Drosophila mojavensis and D-arizonae. Genetics **171**:1083-1101.

Weinreich, D. M., R. A. Watson, and L. Chao. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution **59**:1165-1174.

Yang, Z. H. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer Applications in the Biosciences **13**:555-556.

Yang, Z. H. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Molecular Biology and Evolution **15**:568-573.

Yang, Z. H., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends in Ecology & Evolution **15**:496-503.

Yang, Z. H., and R. Nielsen. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Molecular Biology and Evolution **19**:908-917.

Yang, Z. H., R. Nielsen, N. Goldman, and A. M. K. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yang, Z. H., W. S. W. Wong, and R. Nielsen. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Molecular Biology and Evolution **22**:1107-1118.

Zhang, J. Z., R. Nielsen, and Z. H. Yang. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution **22**:2472-2479.