THE ANALYSIS OF GENES EXPRESSED IN THE SPERM STORAGE ORGANS OF DROSOPHILA MELANOGASTER: PATTERNS OF EVOLUTION AND EXPRESSION

by

Adrianne Marie Prokupek-Pickett

A DISSERTATION

Presented to the Faculty of
The Graduate College at the University of Nebraska
In Partial Fulfillment of Requirements
For the Degree of Doctor of Philosophy

Major: Biological Sciences

Under the Supervision of Professor Lawrence G. Harshman

Lincoln, NE
May, 2008
The molecular evolutionary rate of genes associated with sexual traits and reproduction is (on average) faster than that of genes coding for non-sex-related traits. In Drosophila, this trend has been observed in genes expressing male accessory gland products, as well as in genes expressed in the entire reproductive tract of the females. Though a general trend of rapid evolution has been observed in the Drosophila female reproductive tract, relatively little is known about the evolutionary patterns of the genes expressed in specific organs within the female reproductive system. Sperm storage organs are important to the reproductive success of both males and females, as sperm in storage must remain viable and be properly released for successful fertilization. Further more, sperm storage organs provide an arena for important evolutionary processes such as sperm competition and female sperm choice. This study investigated the evolutionary and expression patterns of genes found in the sperm storage organs of Drosophila. A high percentage of genes found in sperm storage organs were found to be evolving at rapid rates. Polymorphism data confirmed that for five of the genes discovered, the rapid evolution can be attributed to positive selection acting on a subset of the codons of these genes. The function of
many of the genes identified suggests that these genes could interact with seminal products, and this interaction could be a factor in the elevation of evolutionary rates. Regions upstream of genes typically contain a number of non-coding regulatory regions; changes in these regions could result in changes in gene expression. Regions upstream of five of the sperm storage genes were found to be over twice as polymorphic as neutral regions of non-coding DNA. Using microarrays to investigate expression patterns, it was found that the expression of genes associated with sperm storage showed time and organ dependent patterns, presumably associated with different sperm storage events. The identification of genes involved in sperm storage and the characterization of their expression and evolutionary patterns is a new and important development in the understanding of the mechanics and timing of fertilization. This understanding has broader impacts such as insight into important processes (entrance, maintenance and release) of sperm storage for a wide range of species. This study is one of the first to describe important female proteins that may be interacting (possibly co-evolving), either in concert or antagonistically with male seminal products.
TABLE OF CONTENTS

CONTENTS .................................................................................................................. PAGE

TITLE PAGE .................................................................................................................. i
ABSTRACT ................................................................................................................... ii
TABLE OF CONTENTS ................................................................................................. iv
LIST OF FIGURES ........................................................................................................ vi
LIST OF TABLES .......................................................................................................... vii
LIST OF ABBREVIATIONS ........................................................................................... viii
ACKNOWLEDGEMENTS ............................................................................................... x
CHAPTER I: INTRODUCTION ......................................................................................... 1

CHAPTER II: An evolutionary expressed sequence tag analysis of Drosophila spermatheca genes

Abstract...................................................................................................................... 23
Introduction................................................................................................................. 24
Materials and Methods ............................................................................................ 28
Results ....................................................................................................................... 34
Discussion ................................................................................................................. 39
Acknowledgements.................................................................................................. 48
References............................................................................................................... 49


Abstract...................................................................................................................... 73
Introduction................................................................................................................. 74
CHAPTER IV: Patterns of gene expression in the sperm storage organs of *Drosophila melanogaster*

Abstract.................................................................123

Introduction..............................................................124

Materials and Methods ................................................127

Results/Discussion .....................................................131

Acknowledgements.....................................................140

References.....................................................................141
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>FIGURE</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>2.1</td>
<td></td>
<td>59</td>
</tr>
<tr>
<td>2.2</td>
<td></td>
<td>61</td>
</tr>
<tr>
<td>3.1</td>
<td></td>
<td>96</td>
</tr>
<tr>
<td>3.2</td>
<td></td>
<td>98</td>
</tr>
<tr>
<td>3.3</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>3.4</td>
<td></td>
<td>102</td>
</tr>
<tr>
<td>3.5</td>
<td></td>
<td>104</td>
</tr>
<tr>
<td>3.6</td>
<td></td>
<td>106</td>
</tr>
<tr>
<td>3.7</td>
<td></td>
<td>108</td>
</tr>
<tr>
<td>4.1</td>
<td></td>
<td>165</td>
</tr>
<tr>
<td>4.1 (Supplementary)</td>
<td></td>
<td>167</td>
</tr>
<tr>
<td>4.2 (Supplementary)</td>
<td></td>
<td>169</td>
</tr>
</tbody>
</table>
LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>63</td>
</tr>
<tr>
<td>2.2</td>
<td>65</td>
</tr>
<tr>
<td>2.1 (Supplementary)</td>
<td>67</td>
</tr>
<tr>
<td>3.1</td>
<td>110</td>
</tr>
<tr>
<td>3.2</td>
<td>112</td>
</tr>
<tr>
<td>3.3</td>
<td>114</td>
</tr>
<tr>
<td>3.4</td>
<td>116</td>
</tr>
<tr>
<td>3.5</td>
<td>118</td>
</tr>
<tr>
<td>3.6</td>
<td>120</td>
</tr>
<tr>
<td>4.1</td>
<td>163</td>
</tr>
<tr>
<td>4.1 (Supplementary)</td>
<td>182</td>
</tr>
<tr>
<td>4.2 (Supplementary)</td>
<td>183</td>
</tr>
<tr>
<td>4.3 (Supplementary)</td>
<td>184</td>
</tr>
<tr>
<td>4.4 (Supplementary)</td>
<td>185</td>
</tr>
<tr>
<td>4.5 (Supplementary)</td>
<td>186</td>
</tr>
<tr>
<td>4.6 (Supplementary)</td>
<td>187</td>
</tr>
<tr>
<td>4.7 (Supplementary)</td>
<td>188</td>
</tr>
<tr>
<td>4.8 (Supplementary)</td>
<td>189</td>
</tr>
</tbody>
</table>
List of Abbreviations

Acp .......................................................... male accessory gland protein
bp ............................................................... base pair
cDNA ......................................................... complementary Deoxyribonucleic acid
cRNA ........................................................ complementary Ribonucleic acid
D. ana ......................................................... *Drosophila ananassae*
D. ere ......................................................... *Drosophila erecta*
D. grim ...................................................... *Drosophila grimshawi*
D. mel ......................................................... *Drosophila melanogaster*
D. moj ......................................................... *Drosophila mojavensis*
D. per ......................................................... *Drosophila persimilis*
D. pse ....................................................... *Drosophila pseudoobscura*
D. sec ......................................................... *Drosophila sechellia*
D. sim ......................................................... *Drosophila simulans*
D. vir ......................................................... *Drosophila virilis*
D. wil ......................................................... *Drosophila willistoni*
D. yak ....................................................... *Drosophila yakuba*
dN ......................................................... nonsynonymous substitutions per nonsynonymous site
dS ......................................................... synonymous substitutions per synonymous site
hLRT ......................................................... hierarchial likehood ratio test
Ka ......................................................... nonsynonymous substitutions per nonsynonymous site
Ks ......................................................... synonymous substitutions per synonymous site
Kb ................................................................. Kilobases
mRNA .......................................................... messenger Ribonucleic Acid
PAML ........................................................... phylogenetic analysis by maximum likelihood
PCR .............................................................. polymerase chain reaction
RNA ............................................................. ribonucleic acid
SP ................................................................. sex peptide
SP ................................................................. signal peptide
SR ................................................................. seminal receptacle
SSO ............................................................. sperm storage organ
ST ................................................................. spermathecae
TM ............................................................... transmembrane
Acknowledgments

I am deeply grateful and eternally indebted to all of my family, friends and colleagues who helped me both academically and emotionally, through all stages of this dissertation.

To my advisor, Dr. Lawrence Harshman, thank you for all of the time you spent helping to guide me through the rough terrains of graduate school. Thank you for providing me with the beginnings of a project, and for you support and patience in allowing me to take that project make it my own. During the six years I spent in your lab I was allowed to develop my research, critical thinking, writing and teaching skills to levels I never imagined I could reach. You truly are a great teacher, scientist and friend.

I am grateful to my committee members for all of your comments and suggestions. I have benefited greatly from all of your advice and insight. Thank you for taking time out of your lives to make sure that I stayed focused and continued on in the right direction.

I do not posses the words to fully express the depths of my graditute to my entire family. To come into and grow up in a world filled with so much love and support is more than I could wish for, and I am grateful everyday for every opportunity I was provided. To my parents, your unwavering belief in me allowed me to fully believe in myself. My big brother, you gave me the courage to take chances, whether I wanted to or not. My little sister, you helped teach me patience, and kept my life silly and young. And to my husband, thank you for always being my
“biggest fan”. You gave me a safe place to run to when I felt like things were getting too big for me, but also gave me the strength to go back out and keep trying.

To all of the faculty and staff members in School of Biological Sciences who were always kind and willing to help. I came in with requests that were sometimes outrageous, and almost always last minute, but you always came through for me. I am especially thankful to all of the members of the Manter lunch table, no matter how stressed out I was you always made me laugh and have fun.

I could not have succeeded without the friendship and understanding of all of my lab mates, both past and present. You made this more than a place to come and work, you made it like a home. You opened up my small Nebraska world to a wide diversity of cultures and experiences, and forever changed me. I hope while we all pursue our many differing interests we never lose track of each other.

To my friends who have stuck with me through all of this. As I look back on all of the adventures we have had together, I am sometimes so surprised we have all made it so far. Thank you for filling my life with surprises, laughter and love.

A final thank you to all of my gracious sources of for providing teaching and research assistantships, travel and research funding throughout my graduate career: School of Biological Sciences, Initiative for Ecology and Evolutionary Analysis (IEEA), Nebraska Idea Networks of Biomedical Research Excellence (Ne-INBRE), U.S. Department of Education (GAANN), and NSF grant DEB-ESP0346476.

Prokupek 2008
Chapter I.

Introduction
An emerging theme in evolutionary biology is that reproductive proteins evolve rapidly in animals and plants (Clark et al., 2006). One example is genes in the accessory glands of male *Drosophila melanogaster*, which produce proteins transferred to females at the time of mating (Swanson et al., 2001a, Wolfner, 2002). Some of these genes exhibit evidence for positive selection based elevated rates of non-synonymous substitutions inferred from comparative DNA sequence analysis (Swanson et al., 2001a). Genes expressed in the female reproductive tract of *D. melanogaster* also show evidence for positive selection (Swanson et al., 2004b). However, in *D. melanogaster* almost nothing is known about what genes are expressed specifically in sperm storage organs (SSOs), how these genes contribute to SSO function, and how these genes evolve. The absence of data about the evolution of these genes is a major gap in knowledge because genes that are expressed specifically in SSOs potentially play important roles in evolutionary processes such as sexual antagonistic coevolution, sperm competition, and speciation. The second chapter of this dissertation addresses this problem using *Drosophila* as a model organism and a unique and powerful technique, hybrid selection, to create a cDNA library enriched for genes that are specifically expressed in the spermathecae, the long term SSO of *Drosophila*. This procedure allowed for the focus to be on a relatively small set of genes that were hypothesized to evolve rapidly and to play an important role in sperm storage. The hypothesis for this study was that a relatively high proportion of genes identified by enriched expression in the spermathecae evolve rapidly and exhibit the molecular signatures of positive Darwinian selection. The third chapter of this dissertation describes a polymorphism study conducted to help define the evolutionary patterns of genes expressed in the spermathecae. Polymorphism studies are helpful in

Prokupek 2008
determining the rate and pattern of evolutionary change at the population level, versus the long-term evolution inferred from divergence studies alone. The hypothesis was that there would be various indicators of selection based on the pattern of variation within and between populations, and between related species. The fourth chapter of this dissertation uses another powerful technique, microarrays, to conduct a transcriptome analysis of the two types of SSOs of D. melanogaster (spermathecae and seminal receptacle). This analysis provided a means to investigate gene expression and identify proteins in these organs. The hypothesis was that expression patterns of the genes are distinctive between the two SSOs based on their differing roles after sperm acquisition.

A. Sperm Storage

Sperm storage is an important reproductive strategy utilized by females in species with internal fertilization. Sperm storage allows for the female acquisition of sperm to take place days or even months before it is needed to fertilize a mature oocyte. The functional reproductive benefits of this process are great as contact with a suitable male does not necessary correspond with egg availability. For multiply mated females, sperm storage can offer an additional selection step via sperm competition and/or cryptic female sperm choice, as well as give the female a mechanism to compensate for infertile/genetically incompatible males. The duration of time sperm spend in storage is highly variable, extending from a few hours (e.g., mouse), up to decades (e.g., honey bee). This complex and dynamic process includes components such as female and male anatomical structures, muscular contraction of the female genital tract, site of ejaculate deposition, molecules in the seminal fluid, female genital tract secretions, sperm

Prokupek 2008
mobility, and proteins associated with the sperm or other seminal products (Wolfner, 1997, Bloch-Qazi et al., 1998, Neubaum & Wolfner, 1999c, Simmons & Kotiaho, 2002). Specialized SSOs have evolved in a number of species including sperm storage tubules in birds, spermathecae in amphibians and arachnids and in insects SSOs typically appear as sac-like structures (spermathecae) or long tubules (seminal receptacles). These organs are responsible for maintaining sperm viability, organizing the sperm in storage, and facilitating the proper release of sperm from storage (reviewed in (Neubaum & Wolfner, 1999c, Bloch-Qazi et al., 2003).

*Drosophila melanogaster* females have evolved two types of specialized SSOs, the seminal receptacle and spermathecae (Figure 1.1). The seminal receptacle (SR) is a coiled structure which is innervated and partially surrounded by muscle (Fowler, 1973). In addition, spermathecae (ST), a pair of sperm storage pouches, are the site of longer-term sperm storage (upwards of two weeks) (Figure 1.1). Aging of sperm in storage is an issue (Snook & Hosken, 2004) implying that one of the important functions sperm storage, especially in the long term sperm (the ST) is sperm maintenance. In the course of a single mating, 4,000-6,000 sperm are transferred to a female but only approximately 1,000 of the transferred sperm are stored (Gilbert, 1981). More sperm are stored in the SR (65-80%) than in the ST (Gilbert, 1981, Tram & Wolfner, 1999, Bloch-Qazi et al., 2003). The number of sperm in storage is correlated with net fecundity and the timing of female remating (Gromko et al., 1984, Newport & Gromko, 1984, Harshman & Clark, 1998).

When a female mates with two males in succession, most of the offspring are fertilized by sperm from the second male (Lefevre & Jonsson, 1962), this process is referred to as sperm precedence. Sperm precedence could be attributed to a variety of
processes which include sperm displacement (Lefevre & Jonsson, 1962), sperm stratification (layering of stored sperm based on time of entrance into storage) (Waage, 1986, Schlager, 1960), or female sperm selection (Eberhard, 1996, Birkhead & Moller, 1998). The relative number of sperm transferred is a determinant of the proportion of progeny sired by the second male (Gilchrist & Partridge, 2000), possibly due to the displacement of the previous males sperm. Direct observation of fluorescently labeled sperm indicates that the first male’s sperm can be physically displaced from storage by sperm from the second male to mate (Civetta, 1999). Male accessory gland fluid can also act to reduce the amount of progeny sired by previously stored sperm (Scott & Richmond, 1990, Scott & Williams, 1993, Harshman & Prout, 1994). Male accessory gland products could impact sperm in storage by incapacitation of the first male’s sperm by the second male’s seminal products (Price et al., 1999). Sperm storage and sperm competition are related issues and many studies using D. melanogaster have played an important role in investigating sperm competition (Prout & Bundgaard, 1977, Gilchrist & Partridge, 2000, Harshman & Prout, 1994, Clark et al., 1995, Clark & Begun, 1998).

The nature of the relationship between sperm competition and sperm storage remain undefined for any species. The major area of deficit lies in the understanding of the processes involved in sperm storage. How do female SSOs function and what selective forces act on genes specifically expressed in these structures? SSOs are black boxes and this lack of knowledge is surprising given that the SSOs could play a major role in evolutionary phenomena such as sperm competition, sexually antagonistic coevolution, and speciation.
B. Sexual Conflict and Male - Female Coevolution

Sexual conflict is defined by Parker (1979) as “a conflict between the evolutionary interests of individuals of the two sexes” (Parker, 1979). This conflict stems from anisogamy, the large difference in gamete size in the two sexes. Anisogamy leads to different parental investment, and therefore different optimal values of reproductive traits in the two sexes (Trivers, 1972, Parker, 1979). The reciprocal selection pressures on reproductive traits associated with sexual conflict can lead to sexually antagonistic coevolution in which males and females are locked in an evolutionary race (Rice & Holland, 1997, Rice, 1996, Gavrilets, 2000, Rice, 2000, Rowe & Day, 2006). Sexually antagonistic coevolution can drive rapid evolution, even over a short time frame, and have important ramifications for a wide range of traits and phenomena including gamete interactions, sperm use, mate choice and speciation (e.g. (Rowe et al., 1994, Chapman et al., 1995, Pitnick et al., 1999, Civetta & Clark, 2000, Czesak & Fox, 2003, Lessells, 2006). In a key study using *D. melanogaster*, male traits were allowed to evolve while the evolution of female traits was arrested. In as little as 40 generations, dramatic changes were seen in the ability of adapted males to both secure mates and induce female mating refractoriness. A marked increase in the toxic effects of products in the seminal fluid was observed in terms of shortened female lifespan (Rice, 1996). A subsequent study showed that females can evolve resistance to the toxic effects of males (Holland & Rice, 1999). It is important to identify the molecules and understand the mechanisms that underlie this remarkably strong selection as it is likely to provide insight into a range of evolutionary processes. Male accessory gland protein gene products (Acps) could be a proximate cause of sexual antagonism. Acps induce post-mating responses in females which may
have a negative impact on female fitness. For example, experiments using males from a transgenic line, in which the main cells (the source of accessory gland proteins) of the male accessory gland had been ablated by transgenic expression of a toxin, showed that Acps cause the cost of mating (Chapman et al., 1995). In general, sexual conflict could be found widely as a source of strong selection (Rowe & Arnqvist, 2002, Arnqvist & Rowe, 1995). Antagonistic interactions between male seminal products and female proteins in the SSOs could be responsible for the rapid evolution of SSO genes.

C. Male Accessory Gland Proteins in D. melanogaster

Male accessory gland proteins (Acps) are transferred to the female at the time of mating in D. melanogaster. Thereafter they exert a range of effects on female reproduction and physiology including the induction of sperm storage, ovulation and egg-laying, and decreasing female receptivity to future mates and female lifespan (Wolfner, 1997, Swanson et al., 2001a). Interaction of Acps with proteins of the SSOs could be a strong driver of selection. Genes that encode these proteins show evidence for positive selection measured by the analysis of $d_N$ (nonsynonymous substitutions per nonsynonymous site) to $d_S$ (synonymous substitutions per synonymous site) ratios (Swanson et al., 2001a). A lower proportion of the Acps show evidence for positive selection compared to the proportion of genes identified by enriched expression in spermathecae (Chapter II).

Three D. melanogaster Acps will be briefly discussed based on the fact that after mating all three are found in the female SSOs. For decades it has been known that mating increases female egg production and reduces female receptivity to additional mates.
Acp70A (sex peptide, (SP)) is largely responsible for these effects (Kubli, 1996) as demonstrated by a mutation in the SP gene and RNAi suppression of the expression of this gene in males (Liu & Kubli, 2003, Chapman et al., 2003b). SP enters SSOs attached to sperm and, when cleaved, stimulates the biosynthesis of juvenile hormone which induces ovulation (Moshitzky et al., 1996). Two of the genes identified in the spermathecal library (Chapter II) are associated with juvenile hormone function. Another male accessory gland protein, Acp36DE, is required for normal levels of sperm storage (Neubaum & Wolfner, 1999b, Chapman et al., 2000, Bertram et al., 1996). It moves in loose association with the sperm and accumulates in both SSOs (Tram & Wolfner, 1999, Bloch-Qazi & Wolfner, 2003). Acp62F is a protease inhibitor shown to be toxic to females when ectopically expressed (Lung et al., 2002a). Importantly, Acp62F is one of a substantial number of protease inhibitors among the protein products encoded by Acps. This is important because a high proportion of the rapidly evolving genes in the spermathecae are serine proteases (Chapter II). Approximately 20 percent of the predicted Acps are proteases or protease inhibitors (Swanson et al., 2001a, Wolfner, 2002).

D. Evolution of Reproductive Proteins

Elevated rates of evolution for reproductive proteins have been observed in various taxa (Clark et al., 2006). Rapid evolutionary divergence has been documented in protistans, fungi, plants, and animals in gamete surface proteins (Swanson & Vacquier, 2002). This divergence is most often attributed to positive selection, indicating that changes in the structure of the protein are adaptive (Yang & Bielawski, 2000).
High rates of evolution have been observed in both egg proteins and sperm receptors. In abalone, egg vitelline envelope receptor (VERL) and sperm lysin are a cognate pair of gamete recognition proteins that regulate fertilization. These proteins mediate species-specific sperm binding to eggs. The amino-terminal end of the egg VERL, as well as sperm lysin, has undergone very rapid positive Darwinian selection (Galindo et al., 2003b). In mammals, egg coat (zona pellucida) glycoproteins and several sperm proteins evolve rapidly due to positive selection (Swanson et al., 2001b, Swanson et al., 2003). Positive selection has also been shown for Drosophila Acps (Begun et al., 2000b, Swanson et al., 2001a) as well as proteins within the Drosophila female reproductive tract, including proteases (Swanson et al., 2004b). Protease inhibitors have been found to be rapidly evolving among the male accessory gland proteins (Swanson et al., 2001a) perhaps due to interactions with rapidly evolving female proteases. The selective pressures driving the high rate of evolution are unknown, but they could be due to a conflict in the reproductive interests. It is possible that the two sexes are locked in a “coevolutionary chase” driven by sperm competition, sexual conflict, or sexual selection (Galindo et al., 2003b, Swanson et al., 2003).

E. The Effect of Mating on Gene Expression in Female D. melanogaster

The identification of changes triggered in females by mating is important for the understanding of the female role in post-copulatory processes such as sperm storage. McGraw et al. (2004) used microarrays to compare gene expression of D. melanogaster females which were virgin, mated to normal males, mated to spermless males or mated to males lacking Acps. It was determined that 1,783 genes showed a change in expression
level due to mating, though only 46 showed a two fold or greater change. A genome-wide comparative microarray analysis done on virgin, courted, and two-hour post-mated females found that 23 genes were differentially expressed in virgin courted versus uncourtled females, and 38 differentially expressed genes were observed in the comparison of virgin and mated females (Lawniczak & Begun, 2004). In both studies, immune-related genes and serine proteases were found to be effected by mating. The number of serine proteases influenced by mating was determined to be higher than expected by chance (Lawniczak & Begun, 2004). The role of such proteases within the female is unknown. It is interesting to note that serine proteases, as well as serine protease inhibitors (serpins), are also present in seminal fluid (Coleman et al., 1995, Swanson et al., 2001a). Proteases observed to be differentially expressed in microarray studies (Arbeitman et al., 2004, McGraw et al., 2004, Lawniczak & Begun, 2004) were found to partially overlap with protease genes identified in the spermathecal evolutionary EST study (Chapter II). Mack et al. (2006) investigated the effect of mating on gene expression in a region of the female reproductive tract that included ST and the SR using microarrays and two-dimensional protein electrophoresis (Mack et al., 2006b). The study by Mack et al. showed a peak in gene expression at 6 hours post-mating which helped to define one of the time points for the microarray study in this dissertation (Chapter III).

A study investigating a nuclear hormone receptor HR39 of Drosophila found that female mutants for Hr39 have defective or absent spermathecae. Hr39 was found to function similarly to the mammalian splicing factor 1 (SF1), which is necessary to produce both androgens and Mullerian-inhibiting substances to prevent oviduct degeneration (Allen & Spradling, 2008). The spermathecae and parovaria (Figure 1.1)
were found to secrete proteins which function in sperm maturation and in storage, similar to those found in the male epididymis in mammals. The Allen and Spradling study helped to reveal close connections between Dipteran and mammalian reproductive biology, suggesting that there is conservation in specific steps of reproduction including sperm storage. The results of this dissertation could provide relevant ties to mammalian (human) reproduction biology.
LITERATURE


Prokupek 2008


Figure 1.1: Diagram of the female reproductive system in *Drosophila melanogaster*. SR, seminal receptacle; Sp, spermatheca; Ovary; Parovaria (female accessory gland).

(Reprinted from Patterson and Stone 1952)
Chapter II.

Expressed Sequence Tag Analysis of *Drosophila* Spermatheca Genes: Protein Identity and Evolution

[Based on the manuscript submitted to *Evolution* by Adrianne M. Prokupek, Federico Hoffmann, Seong-il Eyun, Etsuko N. Moriyama, Min Zhou, Lawrence G. Harshman]
ABSTRACT

Sperm competition, cryptic female choice of sperm, and sexually antagonistic co-evolution are forms of sexual selection that can drive strong selection and play an important role in speciation. Although there is keen interest in molecular mechanisms of sperm competition and female choice of sperm, very little is known in any species about female genes which affect sexual selection. Sperm storage organ genes are likely to play an important role in sexual selection and speciation. In this study, genes enriched for expression in the long term sperm storage organ of *Drosophila*, the spermatheca, were investigated. A high proportion of genes enriched for expression in the spermatheca have evolved rapidly. Especially notable was the high incidence of genes that exhibit the molecular signature of positive selection even when compared to male gland protein genes, which are a paradigm for this pattern of evolution. Genes that encode serine proteases were a prominent category of genes associated with the spermatheca. They tend to evolve rapidly and all have secretion signals. Genes with secretion signals and transmembrane domains encode proteins that might directly interact with ejaculate proteins and coevolve with them. This study identified spermatheca genes which are candidates to affect sexual selection and speciation.
INTRODUCTION

Darwin (1871) defined sexual selection as the advantage some individuals have over others of the same sex in relation to reproduction. This type of selection can result in the evolution of conspicuous traits such as extravagant secondary sexual characteristics in some species or, more subtly, may be working on a molecular level through the evolution of proteins involved in reproductive processes such as sperm competition, female sperm storage or female sperm choice. Rapidly evolving reproductive proteins are likely involved in sexual selection, playing specific roles in inter (between), intra (within) sex competitions, or a combination of both. One major area where there is a deficit in information is in the identification and classification of female reproductive proteins, especially those proteins interacting directly with male seminal products (e.g. sperm storage proteins).

Reproductive proteins associated with female sperm storage organs (SSOs) are candidates to play important roles in evolutionary phenomena such as sperm competition, female sperm choice, sexual selection and consequently in speciation. However, SSOs are poorly understood in terms of the evolution of proteins associated with these organs, how the organs function at the molecular level, and the relationship between evolution and function. Identification of female proteins expressed within SSOs is a vital step in the development of a comprehensive understanding of the role of SSOs in evolution.

Specific genes and proteins known to play a role in sperm competition have, thus far, only been identified in males. In *D. melanogaster*, second male sperm precedence (P2) is due, in part, to a non-sperm component of the ejaculate (Harshman & Prout, 1994). It is now established that male accessory gland proteins play a role in sperm
competition in *Drosophila* (Ravi-Ram & Wolfner, 2007). Allelic variation in male accessory gland protein genes has been associated with the fertilization success of both the first male to mate with a female (first male sperm precedence (P1)), and the second male to mate with the female (second male sperm precedence (P2)) in this species (Clark et al., 1995, Fiumera et al., 2005, Fiumera et al., 2007). Genetic studies reveal that female processes have a major effect of the outcome of sperm competition in *D. melanogaster* (Clark & Begun, 1998, Clark et al., 1999), but the specific female genes that have these effects have not been identified. Female SSOs provide an arena for sperm competition, making female SSOs a likely place to find female reproductive molecules that affect sperm competition.

Studies on conspecific sperm precedence have provided clues about the importance of male and female reproductive proteins in sperm competition. Conspecific sperm precedence occurs when a female mated to both conspecific and heterospecific males, regardless of the mating order, preferentially produces conspecific rather than hybrid offspring (Howard, 1999). Conspecific sperm precedence is likely to play a major role in speciation by the reproductive isolation of closely related species (Coyne & Orr, 2004). This phenomenon has been observed in a diverse range of taxa including flour beetles, sea urchins, *Drosophila*, rabbits, and several plant species (reviewed in Howard et al., in prep, Howard, 1999). In *D. melanogaster*, conspecific sperm precedence can involve the incapacitation of sperm of the first male to mate by the seminal fluid of the second male to mate (Price, 1997, Price et al., 1999). In this species, heterospecific sperm are not typically displaced by the second male, therefore, the heterospecific sperm in the female SSOs are exposed to the seminal fluid of the second male to mate. Thus, sperm
precedence may be due to interactions between the seminal fluid and reproductive proteins in SSOs. In *D. mauritiana*, stored heterospecific sperm are rapidly lost from SSOs (Price et al., 2001), presumably due to improper storage. Female SSOs and the genes they express could play a major role in conspecific sperm precedence, or precluding fertilization by heterospecific sperm. Direct interactions between female proteins and sperm, or male seminal products may be necessary for the identification of conspecific versus heterospecific sperm, providing another scenario for important male-female interaction.

Rapid evolution of reproductive proteins has been documented in protistans, fungi, plants, and animals (Clark et al., 2006) in both male and female gametic proteins (Swanson & Vacquier, 2002a, Galindo et al., 2003a). For example, in sea urchins male sperm bindin evolves rapidly (Palumbi, 1999) as does the corresponding receptor for bindin on the egg (Palumbi, 1999, Kamei et al., 2000). In mammals, egg coat (zona pellucida) glycoproteins and several sperm proteins evolve rapidly and exhibit the molecular signature of positive (adaptive) selection (Swanson et al., 2001b, Swanson et al., 2003b).

Rapidly evolving *Drosophila* male accessory gland proteins (Acps) have been foci of molecular population genetic and molecular evolution studies. The average rate of sequence divergence of *D. melanogaster* Acps is approximately twice that of non-reproductive proteins (Begun et al., 2000a, Swanson et al., 2001a-b, Mueller et al., 2005a, Wagstaff & Begun, 2004a). At least seven Acps are transferred to the SSOs after mating (Ravi-Ram & Wolfner, 2005), including a rapidly evolving protease inhibitor. By contrast, female reproductive genes are understudied in *Drosophila*, but the signature of
positive selection has been revealed by evolutionary expressed sequence tag (EST) studies using the lower reproductive tract of both *D. simulans* (Swanson et al., 2004a) and *D. arizonae* (Kelleher et al., 2007). These studies did not investigate the rate of evolution of a broad sample of genes from a specific organ as was done in the present study; only a small proportion of the genes identified in the present study overlap with the previously identified proteins in the lower reproductive tract of the same species (Swanson et al., 2004a). The study presented here is the first molecular evolutionary study of genes sampled from a female SSO in any species.

*Drosophila* species typically have two types of organs dedicated to sperm storage (Fowler, 1973, Pitnick et al., 1999). The seminal receptacle contains the majority (65 - 80%) of the sperm (Lefevre & Jonsson, 1962, Neubaum & Wolfner, 1999a), while a pair of spermathecae are the site of long term storage. Sperm are stored in the lumen of the spermatheca which receives proteins of unknown function from surrounding secretory epithelial cells (Filosi & Perotti, 1975). Evolutionary interactions have been identified between sperm and SSOs. For example, evolutionary changes in sperm length resulted in corresponding changes in the length of the seminal receptacle (Miller & Pitnick, 2002, Miller & Pitnick, 2003). One rationale for investigating genes in the spermatheca was that rapidly evolving genes in this SSO might coevolve with rapidly evolving *Drosophila* Acps.

The present study is an evolutionary investigation into genes that are enriched for expression in the long term sperm storage organ of *Drosophila*, the spermathecae. This study uses RNA isolated from the spermathecae of *Drosophila simulans* to identify expressed sequence tagged sites (ESTs), allowing for the identification of gene expressed
in the organ at the time of collection. The focus is on spermathecae because, not only are these organs in extended contact with male seminal products during storage, they also secrete proteins into the sperm storage lumen which could interact with male proteins in the female (Acps and sperm proteins). Rapidly evolving proteins in the spermatheca are prime candidates to play an important role in female-ejaculate interactions. The results suggest that a high proportion of spermathecal proteins evolve rapidly. Such proteins include those with secretion signals and thus are capable of directly interacting with male reproductive proteins in this SSO. Female-ejaculate interactions are thought to mediate key features of sperm storage and important evolutionary phenomena.

**MATERIALS and METHODS**

**cDNA library preparation and DNA sequence generation**

RNA was isolated from both of the spermathecae, including the spermathecal ducts, dissected from 250 *D. simulans* females. The females were held as virgins until the fourth day of adult life when each was paired with a single male. Dissection occurred 3 hours after mating was observed. Total RNA was isolated using the TRIzol reagent (Invitrogen). Total RNA was also purified from female whole bodies minus spermathecae to be used as the driver in subtractive hybridization. cDNA was generated from total RNA using the SMART approach (Zhu et al., 2001). A cDNA library was generated (Evrogen, Moscow) using the suppressive subtraction hybridization (SSH) method in both directions (tester vs. driver and driver vs. tester) (Diatchenko et al., 1996, Diatchenko et al., 1999). An aliquot of the library was plated and 384 colonies were used for DNA template generation by rolling circle amplification using TempliPhi (Amersham
Biosciences). 383 DNA sequences were generated using the MegaBACE 400 automated DNA sequencer (Amersham Biosciences). Vector sequences were masked and the sequences assembled into 383 contigs (expressed sequence tags, ESTs) using the CAP3 program (Huang & Madan, 1999).

**Identification of genes expressed in spermathecae**

To ensure that all of the genes represented by the *D. simulans* spermathecae EST sequences and been found, we queried both the *D. simulans* and *D. melanogaster* genomes. The *D. melanogaster* genome provides a more complete genome, making it less likely to contain annotation errors. Each of the 383 *D. simulans* ESTs was used as a query in a blastn DNA similarity search (Altschul et al., 1990) conducted against the entire CDS sets of *D. simulans* and *D. melanogaster*. Sequences were excluded if the calculated similarity values were 80% or lower or if their expected (E) values were greater than 0.01.

**Ortholog identification**

Using each of the *D. simulans* CDSs obtained above as a query, a blastp protein similarity search (Altschul et al., 1990) was performed to identify ortholog candidates from five additional *Drosophila* genomes (*D. sechellia, D. yakuba, D. erecta, D. ananassae, D. pseudoobscura*). This similarity search was also performed with *D. melanogaster* CDSs as the query to ensure that all identifiable ortholog candidates were found. The Comparative Analysis Freeze 1 (CAF1) genomic sequences of the *Drosophila* species were downloaded from the AAAWiki web site.
(http://rana.lbl.gov/drosophila) for the 12 Drosophila genome project. The entire set of coding sequences of D. melanogaster was obtained from FlyBase (Release 5.1; http://flybase.org). The top hit from each species was then used as a query and reciprocal blastp search was performed against the entire D. simulans CDS set to confirm the orthologous relationships. When multiple sequences were identified with almost identical lowest E-values, all were used as the queries for the reciprocal blastp search. After examining the results of the reciprocal search, ortholog candidates from each species were identified for each of the D. simulans genes. To determine the presence or absence of possible distant orthologs in other species, reciprocal blast was performed against an additional five (more-distantly related) Drosophila genomes (D. persimilis, D. willistoni, D. mojavensis, D. virilis, and D. grimshawi). In addition to blastp, tblastn against these DNA scaffolds was also used.

**Species-specific duplications**

Our orthologous gene set was compared to the list of homologs provided by the 12 Drosophila genome project (http://rana.lbl.gov/~venky/AAA/freeze_20061030/protein_coding_gene). As the list provided by the genome project identified only homolog candidates regardless of whether a gene was an ortholog or paralog, we confirmed that all of our ortholog candidates were included among their homolog candidates. If two or more genes were identified as the top hits with almost identical E-values, then these genes were analyzed as possible duplicates by further investigation including DNA and protein phylogenetic analysis to identify paralog/ortholog relationships.

Prokupek 2008
Reconstruction of multiple alignments from orthologous gene sets

We first reconstructed protein alignments using MUSCLE (Edgar, 2004), and each alignment was adjusted manually. Protein alignments were reverse translated to nucleotide alignments based on their nucleotide sequences using the protal2dna web server (http://bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html). Each nucleotide alignment was again adjusted manually. Finally, the nucleotide alignments were translated to protein alignments for confirmation. The final nucleotide alignments were used for molecular evolutionary analyses.

Evolutionary analyses

The relative contribution of nonsynonymous (dN) and synonymous (dS) changes to the patterns of nucleotide variation was compared using the codon-based maximum-likelihood framework described by Goldman and Yang (Goldman & Yang, 1994) which is implemented in the program PAML (phylogenetic analysis of maximum likelihood) (version 3.15) (Yang et al., 2000). A pairwise comparison was performed between the 42 orthologs of D. simulans and D. melanogaster. The likelihood of dN being higher than dS was evaluated by comparing a model where dN and dS were estimated as free parameters (L1) to a model where dN equals dS (L0). The two models were compared in a likelihood ratio test with one degree of freedom. Historically the signature of adaptive evolution is defined as the gene having a dN/dS ratio > 1. Ratios equal to one are indicative of natural evolution; equal numbers of relative synonymous and nonsynonymous substitutions are presumably due to a lack of functional constraints. Ratios less than one indicate fewer relative nonsynonymous substitutions due to the processes of purifying (negative
selection). In a study by Swanson et al. 2004, it was found that when looking along the entire length of a gene finding $d_N/d_S > 1$ is rare since, presumably only a subset of the codons are subjected to positive selection. Using a survey of literature, Swanson (2004) found that when lowering the $d_N/d_S$ threshold to 0.5, 95% of genes had statistical evidence for positive selection in a subset of their codons. Given that the purpose of the current study is to identify candidate genes which are subject to positive selection, possibly in only a subset of their codons, we adapted the lowered $d_N/d_S$ threshold of 0.5 to be indicative of positive selection.

In addition to the pairwise comparison, we explored variation in the $d_N/d_S$ ratio among sites using the tree based models described by Yang et al. (Yang et al., 2000) based on the alignment of $D. simulans$, $D. melanogaster$, $D. sechellia$, $D. yakuba$, $D. erecta$, $D. ananassae$ and $D. pseudoobscura$ orthologs. The accuracy and power of PAML models increase with more sequences and longer length (Anisimova et al. 2001). PAML was run using the maximum number of orthologs possible; a minimum of four orthologs were used to circumvent problems caused by model convergence. The assumptions of the models and test statistics are briefly described, for a full description see Yang et al. (Yang et al., 2000). PAML was used to determine estimates of $d_N/d_S$ for models of varying complexity; six models were used in this study (M0, M1a, M2a, M7 and M8). M0 and M3 are discrete models in that they assign codons to populations of distinct $d_N/d_S$ values; $d_N/d_S$ can be any value greater than 0. M0 assigns all sites to a single $d_N/d_S$ ratio whereas M3 assigns codons to three categories of sites with $d_N/d_S$ free to vary for each site. M1 assigns codons to one of two site classes, $d_N/d_S > 1$ and $d_N/d_S = 1$. M2a is similar to M1a, but adds an additional class of sites with $d_N/d_S > 1$. M7 assigns
codons using 10 categories of sites with 10 $d_N/d_S$ ratios in the range of 0 – 1 from a
discrete approximation of the $\beta$ distribution. M8 is similar to M7 with the addition of a
category of sites with a $d_N/d_S$ ratio that is free to vary from 0 to >1. Each PAML model
generates a log likelihood indicating how well the models fit the input data. Since PAML
models are nested within each other (M0 with M3; M1 within M2; M7 within M8) twice
the log likelihood difference between the two models is compared with a $\chi^2$ distribution
with degrees of freedom equal to the difference in degrees of freedom between the two
models. $P$ values for sites potentially under positive selection are obtained using a
Baysian approach in PAML.

**Transmembrane and signal peptide prediction and functional domain detection**

Protein sequences from *D. melanogaster* orthologs were used for motif prediction.
Transmembrane (TM) region prediction was conducted using two programs: HMMTOP
(version 2.0; (Tusnady & Simon, 2001) and Phobius (Käll et al., 2004). Both methods use
hidden Markov models for predicting the transmembrane topology. Phobius combines
TM prediction and signal peptide prediction in order to identify signal peptides from N-
terminal regions, often misidentified as a TM region by TM prediction methods. We list a
protein as having a transmembrane domain if both HMMTOP and Phobius predicted TM
regions, or if one program predicted more than one TM region. For signal peptide
prediction, we used TargetP version 1.1 (Emanuelsson et al., 2007) in addition to
Phobius. The TargetP program ranks support for the signal peptides. Only the genes in
the highest class of support, which were also identified as having signal peptides by
Phobius, were listed as having signal peptides.
**Functional categories**

The function of each gene was inferred from a combination of information gained from methods including: conserved domain searches, FlyBase classification, Gene Ontology database classification, and searches of relevant literature. Conserved domain searches by CD-Search at National Center for Biotechnology Information (Marchler-Bauer & Bryant, 2004) ([http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml](http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml)), compared the genes discovered in the EST library against a database of protein modules and domains in order to identify regions of conservation. Conserved domain searches are useful for identifying regions of the genes which may be functionally important, as well as identifying protein domains specific to classes of proteins in order to identify punitive functions of unclassified genes. Gene Ontology (GO) database ([http://www.geneontology.org](http://www.geneontology.org)) is a database of genes described by a controlled vocabulary in terms relating to cellular components, biological processes and molecular functions. GO allows for a way to provide consistent descriptions of gene products. GO terminology is utilized by FlyBase ([http://flybase.bio.indiana.edu/](http://flybase.bio.indiana.edu/)), a website dedicated to gene and genomic information for Drosophila.

**RESULTS**

*Coding sequences in the cDNA library*

Of the 383 EST library clone sequences, 244 matched coding domain sequences (CDSs) in the *D. simulans* genome representing 44 unique CDSs. This number is consistent with to the number of genes typically discovered in EST studies of male accessory gland genes (see Discussion for details). The remaining 139 EST sequences
had no significant matches in either the *D. simulans* or *D. melanogaster* genome and were excluded from further analysis because they did not meet the criteria described in Materials and Methods. The matched sequences found against excluded ESTs were very short, typically less than 20 base pairs, and had E values greater than 0.01 (accepted sequences had an average E value of $4.5 \times 10^{-6}$).

**Orthologs and functional categories**

Sequence similarity was used to identify orthologs in seven *Drosophila* genomes: *D. simulans, D. melanogaster, D. sechellia, D. yakuba, D. erecta, D. ananassae, and D. pseudoobscura*. Of the 44 genes, 30 had identifiable orthologous genes in all seven species. For the remaining 14 genes, orthologs were found in some, but not all of the seven species. For two of 44 genes (dsim_GLEANR_6594 and dsim_GLEANR_15604) there were no orthologous genes shared by *D. simulans* and *D. melanogaster*. These genes were not used for molecular evolutionary analysis because of the lack of identifiable orthologs. The method used to identify orthologs was comprehensive and it sometimes revealed annotation errors. For example, two *D. simulans* genes (dsim_GLEANR_16297 and 17130) appeared to be an incorrect annotation of two exons from a larger gene based on the gene structure of the *D. melanogaster* ortholog (CG32702). Therefore, these two exons were combined into one gene for *D. simulans* as well as in *D. erecta* and *D. ananassae* (dere_GLEANR_3759 and 3760, and dana_20818 and 20819).

Gene function was investigated for all genes having orthologs present in both *D. simulans* and *D. melanogaster* (42 genes in total; Table 2.1; Supplementary Table 2.1).
The most likely function was determined by the Gene Ontology database, conserved domains, and relevant literature. Of the 42 genes, 11 are putative serine proteases each of which has the catalytic triad of amino acids typical of active serine proteases. Expression Analysis Systematic Explorer (EASE) integrated into the DAVID bioinformatics database (http://david.abcc.ncifcrf.gov/home.jsp) indicated that proteases were significantly (P-value $4.5 \times 10^{-9}$) overrepresented in the spermathecae (26%) when compared to the percentage of such genes in the entire *D. melanogaster* genome (5%) (Ross et al., 2003).

**Evolutionary analysis**

Non-synonymous and synonymous substitution rates were determined for each of the 42 genes. One analysis was a pairwise comparison between *D. simulans* and *D. melanogaster*. The rates of synonymous and nonsynonymous substitutions were calculated by a maximum-likelihood method using PAML (Yang, 2007, Yang, 1997). The average $d_{SN}/d_S$ ratio from the pairwise comparisons between the *D. simulans* and *D. melanogaster* sequences is 0.269 ±0.2932, with an average $d_N$ of 0.032 and an average $d_S$ of 0.119. In this comparison, 10 out of the 42 genes (Figure 1; Supplementary Table 2.1) have $d_{SN}/d_S$ higher than the 0.5 threshold adopted by Swanson *et al* (Swanson et al., 2004a).

Another analysis compared the fit of the data to different models of codon evolution (Yang & Nielsen, 2000). PAML models were used to explore heterogeneity in $d_{SN}/d_S$ along the gene, and to test for positive selection (Table 2.2; Supplementary Table 2.1). These comparisons were restricted to the 37 genes for which sequences were
available from at least four species. The first comparison examined heterogeneity along
the length of the gene by comparing the fit of data to the one-ratio model (M0) against a
model that classifies sites into 3 classes (M3). For 27 of the 38 genes, the fit of data to
M3 was significantly better than M0, indicating heterogeneity of evolutionary rates along
the gene for a high percentage of the genes analyzed. Direct tests of positive selection
were also performed. Briefly, two likelihood ratio tests (LRTs) were used to compare null
models that do not allow $d_N/d_S > 1$, M1a and M7, with alternative models that allow a
class of sites to have $d_N/d_S = 1$, M2a and M8 (Yang & Nielsen, 2002, Yang & Swanson,
2002). In the first test the null model (M1a) assumes two site classes, the first with $d_N/d_S < 1$, and the second with $d_N/d_S = 1$; this is compared with the alternative model (M2a)
which adds a class of sites with $d_N/d_S > 1$. The second test uses M7 as the null model,
where $d_N/d_S$ estimates are drawn from a beta distribution with $0 \leq d_N/d_S \leq 1$, with the
alternative model M8, which adds a class of sites with $d_N/d_S > 1$. If the LRTs were
significant, positive selection was inferred (Yang & Nielsen, 2002, Yang & Swanson,
2002). Comparisons of M7 to M8 and M1a to M2a provided evidence for positive
selection in 14 of 37 genes. Genes having elevated $d_N/d_S$ from the pairwise comparison,
and/or support for positive selection from PAML models, are listed in Table 2.2 (for
complete data on all genes see Supplementary Table 2.1).

In order to obtain extended taxonomic insight into the evolution of spermathecal
genes, the presence or absence of homologous genes was examined by comparing D.
melanogaster to 11 sequenced genomes (D. simulans, D. sechellia, D. yakuba, D. erecta,
D. ananassae, D. pseudoobscura, D. persimilis, D. willistoni, D. mojavensis, D. virilis,
and D. grimshawi) (Supplementary Table 2.1). A number of genes, including predicted
proteases, a peptidase, an actin biosynthesis gene, Drosomycin, and a gene of unknown function, were not detectable in taxa distant from the *melanogaster* subgroup (*D. mojavensis*, *D. virilis*, and *D. grimshawi*). A high percentage of the genes with undetectable orthologs in more distantly related species encode serine proteases (see Discussion). One of the two *D.simulans* genes (dsim_GLEANR_6594) discovered in the library, but having no *D. melanogaster* ortholog, is predicted to be a serine protease by the use of conserved domain searches. Orthologs to this gene were also found in *D. sechellia*, and *D. yakuba*.

Possible species-specific duplications were found in *D. erecta*, *D. ananassae*, and *D. pseudoobscura*. Duplicated copies in *D. erecta* (dere_GLEANR_9251, dere_GLEANR_13114, dere_GLEANR_16834) and in *D. ananassae* (dana_GLEANR_20091 and dana_GLEANR_20093) share 80 and 90% similarities, respectively, against their corresponding *D. simulans* genes. A pair of *D. ananassae* genes (dana_GLEANR_9014 and dana_GLEANR_10165) was almost identical (only one nucleotide difference) to a pair of *D. pseudoobscura* genes (dpse_GLEANR_6308 and dpse_GLEANR_6306). Further investigation would be needed to determine if these apparent duplications are due to artifacts such as assembly mistakes.

**Secretion signal sequence and transmembrane region prediction**

Of the 42 genes examined, 20 (47%) are predicted to have signal peptides and 12 (29%) predicted to have transmembrane regions (Table 2.1; Supplementary Table 2.1). All of the proteases have predicted secretion signal sequences.
DISCUSSION

Rapidly evolving reproductive proteins are candidates to play an important role in sexual selection and speciation. In order to identify candidate genes that could play a role in these evolutionary processes, the molecular evolution of genes expressed in the spermatheca was analyzed and likely gene function characterized. A high proportion of spermatheca genes are predicted to encode serine proteases. Many of the serine proteases discovered in this study evolve rapidly, and all have secretion signals. Serine proteases expressed in the spermatheca are prime candidates to participate in evolutionarily dynamic interactions with male seminal products. Overall, a very high percentage of the genes in the spermatheca are rapidly evolving and a very high proportion exhibit the molecular signal of positive selection based on $d_N/d_S$ ratios and PAML analysis. Insight into the function of the spermatheca was obtained from the identity of genes expressed in this SSO.

Although spermathecae are found in a wide range of invertebrate and vertebrate taxa (Eberhard, 1996), the function of proteins and other macromolecules associated with this organ are understudied. An exception is social insects - bees and ants - in which the spermatheca are known to be important for long term sperm storage. Reproductively capable females (queens) mate with several males early in life and rely on stored sperm to fertilize eggs over many years. Since the reproductive success of the queen is directly dependent on the availability and viability of stored sperm, there is strong selection for both sperm viability as well as efficient utilization of sperm stores (Baer et al., 2006). In a social ant (*Crematogaster opuntiae*), as well as in honey bees, the secretory cells surrounding the spermatheca have ample glycogen, serving as a possible energy reserve.
for sperm in storage (Wheeler & Krutzsch, 1994). In honeybees the spermathecal fluid contains sugars including glucose, trehalose, and fructose, as well as a high level of trehalase activity (Alumot et al., 1969a). Enzymes with antioxidant activity have been found in the spermathecae of honeybees, presumably acting to protect stored sperm from oxidative damage (Weirich et al., 2002, Collins et al., 2004c).

In the present study, 44 unique genes were identified in the hybrid-selected cDNA library. The small number of genes found in this study is quite similar to the number of genes found in comparable studies in Drosophila and other species. Multiple studies, using cDNA hybridization or EST sequences, were needed to raise the level of Acp genes Drosophila melanogaster to a total of 57 genes (DiBenedetto et al., 1987, Monsma & Wolfner, 1988, Wolfner et al., 1997, Swanson et al., 2001a-b). EST studies of the accessory glands of mosquito Anopheles gambiae yielded 46 genes, and similar studies in crickets (Gryllus firmis and Gryllus pennsylvanicus) yielded 30 genes (Andres et al., 2006). Using Mediterranean fruit flies (Ceratitis capitata), DNA sequence analysis of a subtractive hybridization cDNA library from RNA isolated from the male accessory gland resulted in identification of 13 unique genes (Davies & Chapman, 2006).

It is informative to compare the evolution of the 42 spermatheca genes identified in this study with relevant previous studies. Based on the pairwise comparison between D. simulans and D. melanogaster, 24% of the spermatheca genes have an overall dS/dS > 0.5 (Figure 1). Two especially relevant previous studies (Swanson et al. 2001a; Swanson et al. 2004) have made pairwise comparisons of D. simulans and D. melanogaster. The incidence of genes with dS/dS > 0.5 in the spermatheca (present study) is appreciably higher than that observed in the female reproductive tract of D. melanogaster, minus...
ovaries, in which 6% of the genes had a $d_N/d_S > 0.5$ (Swanson et al., 2004a) and at least as high as that observed for male accessory gland genes among which 19% of genes had a $d_N/d_S > 0.5$ (Swanson et al., 2001a-b). The number of genes in the present study that overlapped with the most similar study (Swanson et al. 2004) was only 5, even though the Swanson study included spermathecae in the mix of tissue investigated. This suggests the possibility that a large number of genes (especially those of lower expression) were missed when using the entire lower reproductive tract. Of the 10 spermathecal genes with elevated $d_N/d_S (> 0.5)$, the most rapidly evolving was a gene of unknown function ($d_N/d_S = 0.95$), followed by 5 serine proteases ($d_N/d_S = 0.72 – 0.89$), and two more genes of unknown function ($d_N/d_S = 0.64 – 0.66$). The average rate of sequence divergence of *D. melanogaster* Acps is approximately twice that of non-reproductive proteins (Begun et al., 2000a, Swanson et al., 2001a-b, Mueller et al., 2005a, Wagstaff & Begun, 2004a). Acp male genes are notable for their rapid rates of evolution and it appears that female spermatheca genes are similarly evolutionarily dynamic.

The analysis of the pattern of molecular evolution among a larger set of related species is also informative. Molecular evolutionary analyses showed that 14 of 37 (37.8%) spermatheca genes contain at least one region that conforms to a model of positive selection (Table 2.2; Supplementary Table 2.1). Categories of genes showing evidence for positive selection include serine proteases, cell communication, translation, sugar metabolism, peptidase activity, protein-protein interaction and genes of unknown function (Table 2.2; Supplementary Table 2.1). The proportion of positively selected spermathecae genes can be compared to the molecular evolution rates of *Drosophila melanogaster* seminal fluid proteins (Haerty et al., 2007). Twenty five seminal fluid
genes had orthologs in all of the *melanogaster* subgroup species and 4 of these genes (16%) exhibited positive selection by the criteria of acceptance of M8 over M7. Acceptance of M8 over M7 indicates that the data better fit a models which includes a class of sites that allow $d_N/d_S > 1$ (positively selected). Using the same criteria in female *Drosophila*, of 679 genes in the reproductive tract, and of 9921 genes not related to sex or reproduction, 6.2% and 6.0% respectively were consistent with the hypothesis of positive selection by acceptance of model 8 (Haerty et al., 2007). Overall, a much higher proportion of genes enriched for expression in the spermatheca showed evidence for positive selection (37%) compared to seminal fluid proteins (16%) or a collection of genes in the female reproductive tract (6.2%). The incidence of directional selection among spermathecal genes is striking.

In the present study, serine proteases are the predominant category of genes which are lack detectable orthologs as a function of evolutionary distance. Among spermatheca proteases, four have lost orthologs in species belonging to the *melanogaster* subgroup and five protease genes have no detectable ortholog in the obscura subgroup. At the level of differentiation between *D. melanogaster/D. simulans* and the repleta group, seven protease genes have no orthologs. Between the *melanogaster* subgroup and a Hawaiian *Drosophila*, 10 protease genes have no orthologs. In other studies using *Drosophila* species, reproductive system proteases show evidence of accelerated and positive evolution (Kelleher et al., 2007, Swanson et al., 2004a, Kern et al., 2004, Panhuis & Swanson, 2006, Lawniczak & Begun, 2007, Wong et al., 2007), indicating that the evolutionary patters for this group of proteins is dynamic.
All of the spermatheca serine proteases have secretion signals (Table 2.1; Supplementary Table 2.1) and are possibly secreted into the lumen of this SSO. Potential roles for male and female proteases are discussed in Ravi-Ram and Wolfner (Ravi-Ram & Wolfner, 2007). Spermathecal proteases may be involved in interactions with male reproductive proteins, or play roles functionally analogous to male reproductive proteins. Previous studies have described at least two *Drosophila* male proteins that are transferred to females and undergo cleavage within the female reproductive tract, perhaps as a mechanism to control activity levels of the proteins (Monsma et al., 1990, Bertram et al., 1996, Ravi-Ram & Wolfner, 2007). Female proteases might act to control the viscosity of the internal milieu of the lumen of the spermatheca analogous to the semen coagulation role played by the primate prostate specific antigen (PSA) in males (Malm et al., 2000). PSA is a serine protease and its role in humans suggests an analogous function for spermathecal proteases in *Drosophila*.

An Acp protein that plays a key role in sperm storage (Acp36DE) is found in the spermatheca after mating and it is rapidly evolving. Moreover a protease Acp associated with regulation of sperm use is also evolving rapidly (Wong et al., 2007) and it also is a candidate for a coevolutionary interaction with spermatheca proteases based on direct interaction. As an exciting possibility, male derived protease inhibitors might inhibit female proteases secreted into the lumen spermatheca in a specific male-female (ejaculate-female) molecular interaction. Seven protease inhibitors have been reported among Acps of *D. melanogaster*. Acp 62F, which is able to transverse the female reproductive tract and enter the hemolymph, is toxic upon ectopic expression; this Acp is present in the spermatheca after mating (Lung et al., 2002b). The evolutionary
importance of the interactions which occur between females proteins and male ejaculates are being increasingly recognized (Pitnick et al., in prep.). Spermathecal proteases are prime candidates to be involved in evolutionarily dynamic interactions with male reproductive proteins such as protease inhibitors and proteases.

Four of the protease genes identified in the present study are found in a cluster on chromosome arm 2L. These genes exhibit approximately 30% sequence similarity to each other in *D. melanogaster* and each gene is approximately 90% similar to its ortholog in *D. simulans*. They have no introns and they encode proteins with the canonical serine protease catalytic triad of amino acids. The cluster of proteases has been found to be transcriptionally activated by mating (Lawniczak & Begun, 2007). These proteases, and several others, are rapidly evolving between populations of *D. melanogaster* and *D. simulans* (Lawniczak & Begun, 2007). Five of the proteases found in the present study (CG18125, and the cluster on chromosome II) have been foci for previous molecular population genetic and molecular evolution studies. These studies showed that the sites of molecular changes in these proteases were associated with the active site, suggesting the evolution of functional changes related to catalysis (Panhuis & Swanson, 2006, Lawniczak & Begun, 2007). This rapid divergence can further be exemplified by the two spermatheca genes found in *simulans* without a *melanogaster* ortholog (see Results). One of these genes, dsim_GLEANR_6594, is found in the middle of the four clustered proteases in the *D. simulans* genome, and is predicted to be a serine protease based on conserved domains. A large corresponding portion of this region is missing from the *D. melanogaster* genome, which provides an explanation for
the lack of ortholog found in *D. melanogaster* and yields a picture of rapid change between the two genomes.

One class of spermathecal proteins identified in this study contains at least one protein-protein interaction motif called a CUB domain. CUB domains, which consist of approximately 110 amino acids with four positionally conserved cysteines, (Bork & Beckman, 1993) play a variety of roles including interaction with sperm in both vertebrate and invertebrate taxa (Haley & Wessel, 2004, Kamei & Glabe, 2003). CUB domains bind other proteins with high specificity (Song et al., 2006) and tend to exist as a cluster of multiple repeats along the length of a single gene. A gene identified in the present study, CG32702, contains approximately 20 CUB domains in one region of the protein, along with a repeat of five EGF-CA like domains at the C-terminal end (Figure 2.2). A second gene (CG30371) encodes a trypsin-like serine protease domain and a motif that is 67% similar to a CUB domain. CG32702 (the gene with many CUB domains) exhibits evolutionary stasis in much of the gene, but relatively rapid evolution in some regions of the gene. Having multiple CUB domains potentially allows for a relaxation of selective constraints. Changes could be tolerated in a subset of the domains because the original specificity may be retained by the remaining (unchanged) domains. A general argument about redundancy and relaxation of selective constraints when repeated motifs are present in a protein has been made by Metz and Palumbi (Metz & Palumbi, 1996) and is used to interpret the evolution of VERL domains in reproductive proteins (Swanson & Vacquier, 1998). There is evidence for positive selection in regions of the CUB protein even though the protein is sufficiently conserved to be found in all 12 sequenced *Drosophila* genomes.
Other genes encoding proteins with potentially important roles associated with sperm storage and maintenance were identified. Trehalase activity (sugar metabolism gene in Table 2.1) could play a role in sperm nutrition. The *Drosophila* trehalase RNA encodes a predicted secretion signal and thus its protein could be active in the lumen of the spermatheca. A gene encoding an antifungal defense peptide (*Drs*) also was identified in the present study. This gene is not spermatheca-specific, it is constitutively expressed in both types of SSOs of *D. melanogaster* (Ferrandon et al., 1998). The SSOs are apparently the only site of constitutive expression whereas the gene is expressed in many locations after induction with a pathogen (Ferrandon et al., 1998). Juvenile hormone epoxide hydrolase 3 (JHEH3) is an example of a gene that could play an interesting role in evolution. This enzyme catabolizes juvenile hormone (JH) to an inactive metabolite. Its protein product is predicted to have a secretion signal and six transmembrane domains suggesting it could be a receptor. It is possible that JHEH3 could be acting to control JH levels in the spermatheca and as a systemic hormone regulator if it is secreted into the hemolymph. In *D. melanogaster*, Acp70 is transferred to females at the time of mating stimulates juvenile hormone synthesis (Peng et al., 2005a). An enzyme that produces a precursor of juvenile hormone is elevated in the lower female reproductive tract after mating (Mack et al., 2006a). The presence of enzyme activity that degrades juvenile hormone (JHEH3) in the long-term SSO is intriguing because it might oppose the male effect of stimulating the synthesis of JH. Genes with unknown function in *Drosophila* or other species (Table 2.1) might be quite interesting in terms of having spermatheca-specific roles because functions for such genes have not been identified in other tissues or taxa. Four of these genes have transmembrane domains which could be receptors having
spermatheca-specific function. These receptors could potentially interact with male accessory gland proteins or other proteins found on sperm. Identification of such receptors would be important for understanding the evolution of Acps and how they function in females.

CONCLUSIONS

This study has produced insight into the evolution and function of genes enriched for expression in *Drosophila* spermathecae. We find that genes expressed in the spermatheca evolve as rapidly as genes in the male accessory gland. Importantly, the proportion of genes with the overall signature of positive selection is higher than Acp genes which are a paradigm for rapid evolution. Rapidly evolving spermatheca proteins of established and novel function could participate in female reproductive molecule-ejaculate interactions which are increasingly recognized as evolutionarily important (Pitnick et al., in prep., Ravi-Ram & Wolfner, 2007).
ACKNOWLEDGMENTS

We thank Mariana Wolfner, Alex Wong, Tony Zera and Andy Clark for comments on the manuscript and other contributions to the research. This research was supported by a National Science Foundation grant (DEB-ESP0346476).
REFERENCES


Clark, A. G. & Begun, D. J. 1998. Female genotypes affect sperm displacement in

Clark, A. G., Begun, D. J. & Prout, T. 1999. Female x male interactions in Drosophila
sperm competition Science 283: 217-220.

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

Collins, A. M., Williams, V. & Evans, J. D. 2004. Sperm storage and antioxidative
enzyme expression in the honey bee, Apis mellifera. Insect Mol Biol 13: 141-146.

Massachusetts.


Davies, S. J. & Chapman, T. 2006. Identification of genes expressed in the accessory
glands of male Mediterranean Fruit Flies (Ceratitis capitata). Insect Biochem Mol

Diatatchenko, L., Lau, Y., Campbell, A., Chenchik, A., Mogadam, F., Huang, B.,
Suppression subtractive hybridization: a method for generating differentially
regulated or tissue-specific cDNA probes and libraries. Proc Natl Acad Sci USA
93: 6025-30.

hybridization: a versatile method for identifying differentially expressed genes.


Prokupekk 2008


**Figure 2.1.** The number of nonsynonymous substitutions per site ($d_N$) plotted against the number of synonymous substitutions per site ($d_S$) for the *D. simulans* spermathecae EST library. The solid line represents $d_N/d_S$ of 1, and the dashed line represents $d_N/d_S$ of 0.5. All points on the graph represent genes found in the spermathecae EST library and open squares correspond to the eleven serine proteases.
**Figure 2.2.** Schematic representation of the CUB domains (putative sperm binding domains) of gene CG32702. This gene is located on the X chromosome. CG32702 is 12,484 bp in length, has twelve introns and a transcript size of 11,275 nucleotides (http://flybase.org). The gene is relatively conserved throughout the twelve *Drosophila* genomes although there are sites that have evolved rapidly. Image obtained from the National Center for Biotechnology Information (Marchler-Bauer & Bryant, 2004).
Table 2.1: Functional annotation of 42 genes enriched for expression in the spermatheca
<table>
<thead>
<tr>
<th>Function</th>
<th>Number</th>
<th>SP</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine protease</td>
<td>11</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Cell communication</td>
<td>3</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>Peptidase</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Translation</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actin formation/biosynthesis</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amino acid transport</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Antimicrobial</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Apoptosis</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Cation transport</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Dehydrogenase</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Helicase</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Juvenile Hormone Catabolism</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Nerve signaling</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Phospholipid metabolism</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Secondary metabolism</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Protein-Protein interaction</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sugar metabolism</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Unknown</td>
<td>9</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>

1Predicted function of encoded proteins; 2Number of genes; 3Number of genes predicted to encode proteins which have secretion signal peptides; 4Number of genes predicted to encode proteins with transmembrane regions
Table 2.2: Genes with elevated pairwise $d_N/d_S$ (>0.5) and those identified as having regions which provide evidence for evolution by positive selection (PAML analysis).
<table>
<thead>
<tr>
<th>Genes</th>
<th>Function</th>
<th>d_s/d_s</th>
<th>Species</th>
<th>M0 vs. M3</th>
<th>M1 vs. M2</th>
<th>M7 vs. M8</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>p_s^4</td>
<td>d_s/d_s^5</td>
<td>p_s^6</td>
</tr>
<tr>
<td>CG8331</td>
<td>cell comm.</td>
<td>0.19</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.21***</td>
<td>0.59</td>
<td>0.01</td>
</tr>
<tr>
<td>me31B</td>
<td>helicase</td>
<td>0.05</td>
<td>M,S,Sc,Y,E,A</td>
<td>0.02***</td>
<td>0.82</td>
<td>0.01</td>
</tr>
<tr>
<td>CG10650</td>
<td>peptidase</td>
<td>0.55</td>
<td>M,S,Sc,Y,E,A</td>
<td>0.09***</td>
<td>2.97</td>
<td>0.03</td>
</tr>
<tr>
<td>ηTry</td>
<td>ser. protease</td>
<td>0.14</td>
<td>M,S,Sc,Y,E</td>
<td>0.03***</td>
<td>2.69</td>
<td>0.07</td>
</tr>
<tr>
<td>Ser12</td>
<td>ser. protease</td>
<td>0.64</td>
<td>M,S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG17012</td>
<td>ser. protease</td>
<td>0.89</td>
<td>M,S,Sc,A</td>
<td>0.07***</td>
<td>6.08</td>
<td>0.05***</td>
</tr>
<tr>
<td>CG17234</td>
<td>ser. protease</td>
<td>0.74</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.06***</td>
<td>5.94</td>
<td>0.10***</td>
</tr>
<tr>
<td>CG17239</td>
<td>ser. protease</td>
<td>0.73</td>
<td>M,S,Sc,Y,E,A</td>
<td>0.16***</td>
<td>2.70</td>
<td>0.09**</td>
</tr>
<tr>
<td>CG18125</td>
<td>ser. protease</td>
<td>0.79</td>
<td>M,S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG31681</td>
<td>ser. protease</td>
<td>0.82</td>
<td>M,S</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG32834</td>
<td>ser. protease</td>
<td>0.19</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.09***</td>
<td>2.74</td>
<td>0.02</td>
</tr>
<tr>
<td>CG32702</td>
<td>protein int.</td>
<td>0.14</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.06***</td>
<td>1.41</td>
<td>0.05</td>
</tr>
<tr>
<td>Treh</td>
<td>sugar metab.</td>
<td>0.12</td>
<td>M,S,Sc,Y,E</td>
<td>0.05***</td>
<td>2.95</td>
<td>0.04**</td>
</tr>
<tr>
<td>Ef1α48D</td>
<td>translation</td>
<td>0.02</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.01**</td>
<td>3.44</td>
<td>0.01*</td>
</tr>
<tr>
<td>Qm</td>
<td>translation</td>
<td>0.02</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.01**</td>
<td>1.76</td>
<td>0.0003</td>
</tr>
<tr>
<td>CG2233</td>
<td>unknown</td>
<td>0.95</td>
<td>M,S,Sc,Y,E,A</td>
<td>0.23***</td>
<td>2.64</td>
<td>0.10***</td>
</tr>
<tr>
<td>CG11137</td>
<td>unknown</td>
<td>0.1</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.02***</td>
<td>0.84</td>
<td>0.01</td>
</tr>
<tr>
<td>CG15098</td>
<td>unknown</td>
<td>0.69</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.09***</td>
<td>2.72</td>
<td>0.05</td>
</tr>
<tr>
<td>CG30197</td>
<td>unknown</td>
<td>0.05</td>
<td>M,S,Sc,Y,E</td>
<td>0.02**</td>
<td>4.37</td>
<td>0.02</td>
</tr>
<tr>
<td>CG31686</td>
<td>unknown</td>
<td>0.66</td>
<td>M,S</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1Predicted protein function; 2d_s/d_s based on pairwise comparison of *D. melanogaster* and *D. simulans* sequences, estimated assuming no rate heterogeneity; 3Species: refers to the species of *Drosophila* from which sequences were obtained for PAML analysis M= *melanogaster*, S= *simulans*, Sc= *sechellia*, Y= *yakuba*, E= *erecta*; A= *ananassae*; P= *pseudoobscura*. The following statistics are all derived from PAML analysis: 4p_s: the proportion of sites estimated to belong to the class that has the highest d_s/d_s in M3; 5d_s/d_s: for the highest class in M3; 6p_s: the proportion of sites estimated to belong to the class that has d_s/d_s > 1 in M2; 7d_s/d_s: the estimate for the class with the ratio > 1 in M2; 8p_s: the proportion of sites estimated to belong to the class that has d_s/d_s > 1 in M8; 9d_s/d_s: the estimate for the class with the ratio > 1 in M8; Statistical significance for the LRT: *P <.05; **P<.01; ***P<.001.
**Supplementary Table 2.1**: Gene identification numbers for *D. simulans* and identification numbers for *D. melanogaster* corresponding to the 44 genes identified by the spermathecae EST library; Function of predicted proteins; Secretion signal prediction; Transmembrane region prediction; Pairwise $d_S/d_S$; PAML data for all genes with sufficient number of orthologs;
<table>
<thead>
<tr>
<th>D.sim ID</th>
<th>D.mel CG</th>
<th>Function</th>
<th>SP</th>
<th>TM</th>
<th>(d_s/d_i)</th>
<th>PAML Species</th>
<th>(p_6)</th>
<th>(d_s/d_i)</th>
<th>(p_8)</th>
<th>(d_s/d_i)</th>
<th>(p_{10})</th>
<th>Genome Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>dsim_GLEANR_16371</td>
<td>CG2233</td>
<td>Unknown</td>
<td>X</td>
<td>0.95</td>
<td>0.02***</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.02***</td>
<td>33.31</td>
<td>0.18***</td>
<td>4.16</td>
<td>0.17***</td>
<td>3.10</td>
</tr>
<tr>
<td>dsim_GLEANR_3734</td>
<td>CG3066</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.17</td>
<td>0.09***</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.09***</td>
<td>1.45</td>
<td>0.09</td>
<td>1.45</td>
<td>0.09</td>
<td>1.46</td>
</tr>
<tr>
<td>dsim_GLEANR_9093</td>
<td>CG3831</td>
<td>Unknown</td>
<td>X</td>
<td>0.32</td>
<td>0.14***</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.14***</td>
<td>1.54</td>
<td>0.14</td>
<td>1.54</td>
<td>0.14</td>
<td>1.54</td>
</tr>
<tr>
<td>dsim_GLEANR_7778</td>
<td>CG4214</td>
<td>Nerve signaling</td>
<td>X</td>
<td>0.36</td>
<td>0.31**</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.31**</td>
<td>0.54</td>
<td>0.07</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>dsim_GLEANR_8979</td>
<td>CG4254</td>
<td>Actin metabolism</td>
<td>0.00</td>
<td>0.25</td>
<td>0.00</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.00</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>1.52</td>
</tr>
<tr>
<td>dsim_GLEANR_2206</td>
<td>CG4370</td>
<td>Cation transport</td>
<td>X</td>
<td>0.01</td>
<td>0.09***</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.09***</td>
<td>1.56</td>
<td>0.09</td>
<td>1.56</td>
<td>0.09</td>
<td>1.56</td>
</tr>
<tr>
<td>dsim_GLEANR_7402</td>
<td>CG4916</td>
<td>Helicase</td>
<td>0.05</td>
<td>0.01</td>
<td>2.92</td>
<td>M,S,Sc,Y,E,A</td>
<td>0.01</td>
<td>2.92</td>
<td>0.01</td>
<td>2.92</td>
<td>0.01</td>
<td>2.92</td>
</tr>
<tr>
<td>dsim_GLEANR_2360</td>
<td>CG7415</td>
<td>Peptidase</td>
<td>X</td>
<td>0.01</td>
<td>0.02</td>
<td>M,S,Sc,Y,E,P</td>
<td>0.02</td>
<td>2.92</td>
<td>0.02</td>
<td>2.92</td>
<td>0.02</td>
<td>2.92</td>
</tr>
<tr>
<td>dsim_GLEANR_10797</td>
<td>CG8280</td>
<td>Translation</td>
<td>0.02</td>
<td>0.003**</td>
<td>11.54</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.003**</td>
<td>11.54</td>
<td>0.002*</td>
<td>11.54</td>
<td>0.002*</td>
<td>11.54</td>
</tr>
<tr>
<td>dsim_GLEANR_11000</td>
<td>CG8331</td>
<td>Cell communication</td>
<td>X</td>
<td>0.19</td>
<td>0.01***</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.01***</td>
<td>33.67</td>
<td>0.01</td>
<td>34.35</td>
<td>0.01</td>
<td>33.74</td>
</tr>
<tr>
<td>dsim_GLEANR_11526</td>
<td>CG9364</td>
<td>Sugar metabolism</td>
<td>X</td>
<td>0.12</td>
<td>0.05***</td>
<td>M,S,Sc,Y,E,A,P</td>
<td>0.05***</td>
<td>2.95</td>
<td>0.04**</td>
<td>3.11</td>
<td>0.05***</td>
<td>2.93</td>
</tr>
<tr>
<td>DSM</td>
<td>Gene ID</td>
<td>Description</td>
<td>p</td>
<td>M, S, Y, E, A, P</td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>---------</td>
<td>-------------------------</td>
<td>---</td>
<td>----------------</td>
<td>----------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_11514</td>
<td>CG10067</td>
<td>Actin formation</td>
<td>0.00</td>
<td>M, S, Y, E, A, P</td>
<td>0.37 0.00 0.00 1.00 0.00 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_13207</td>
<td>CG10469</td>
<td>Serine Protease</td>
<td>X 0.07</td>
<td>M, S, Y, E, A, P</td>
<td>0.11 0.97 0.04 1.00 0.10 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_5555</td>
<td>CG10650</td>
<td>Peptidase</td>
<td>X 0.55</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.09*** 2.97 0.03 5.36 0.05** 3.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_13789</td>
<td>CG10810</td>
<td>Antimicrobial defense</td>
<td>X 0.00</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.46 0.03 0.00 1.00 0.00 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_12120</td>
<td>CG11137</td>
<td>Unknown</td>
<td>X 0.10</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.15 0.24 0.01 1.00 0.00 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_9282</td>
<td>CG11200</td>
<td>Dehydrogenase</td>
<td>X 0.02</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.18 0.41 0.04 1.00 0.00 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_12517</td>
<td>CG12232</td>
<td>Cell communication</td>
<td>0.05</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.09 0.47 0.02 1.00 0.00 2.32</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_9897</td>
<td>CG12386</td>
<td>Serine Protease</td>
<td>X 0.14</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.05*** 2.35 0.06 2.09 0.05 2.31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_10438</td>
<td>CG12840</td>
<td>Cell communication</td>
<td>X 0.25</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.005* 14.33 0.00 14.19 0.00 14.22</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_9414</td>
<td>CG14495</td>
<td>Unknown</td>
<td>0.09</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.09*** 1.56 0.09 1.56 0.09 1.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_9351</td>
<td>CG15098</td>
<td>Unknown</td>
<td>X 0.69</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.02*** 4.63 0.04 3.84 0.05 3.23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_9344</td>
<td>CG15106</td>
<td>Juv. Horm. Catabolism</td>
<td>X X 0.06</td>
<td>M, S, Sc, Y, E, A, P</td>
<td>0.15*** 0.48 0.03 1.00 0.00 1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Pr, W, Mj, V, G</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prokupek 2008
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>dsim_GLEANR_5780</td>
<td>Unknown</td>
<td>X</td>
<td>0.49</td>
<td>0.20***</td>
<td>1.42</td>
<td>0.07</td>
<td>1.69</td>
<td>0.49</td>
<td>1.15</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_5271</td>
<td>Phospholipid metabolism</td>
<td>X</td>
<td>0.13</td>
<td>0.19***</td>
<td>0.47</td>
<td>0.05</td>
<td>1.00</td>
<td>0.00</td>
<td>1.00</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_6593</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.89</td>
<td>0.07***</td>
<td>6.08</td>
<td>0.05***</td>
<td>8.21</td>
<td>0.06***</td>
<td>6.96</td>
<td>M.S,Sc,A,V</td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_2235</td>
<td>Amino acid transport</td>
<td>X</td>
<td>0.07</td>
<td>0.04***</td>
<td>1.54</td>
<td>0.04</td>
<td>1.54</td>
<td>0.04</td>
<td>1.55</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_6882</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.74</td>
<td>0.05***</td>
<td>7.23</td>
<td>0.09***</td>
<td>5.25</td>
<td>0.10***</td>
<td>4.86</td>
<td>M.S,Sc,Y,E,P,Pr</td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_6595</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.73</td>
<td>0.22***</td>
<td>2.40</td>
<td>0.22*</td>
<td>2.41</td>
<td>0.22**</td>
<td>2.40</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,V</td>
</tr>
<tr>
<td>dsim_GLEANR_6596</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M.S,Y,E</td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_12116</td>
<td>Translation</td>
<td></td>
<td>0.02</td>
<td>0.005**</td>
<td>2.77</td>
<td>0.01</td>
<td>2.77</td>
<td>0.001*</td>
<td>2.77</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_5775</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.79</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M.S,Sc</td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_11034</td>
<td>Unknown</td>
<td>X</td>
<td>0.05</td>
<td>0.02**</td>
<td>4.37</td>
<td>0.02</td>
<td>4.37</td>
<td>0.002*</td>
<td>4.37</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_15401</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.13</td>
<td>0.36*</td>
<td>0.48</td>
<td>0.05</td>
<td>1.00</td>
<td>1.05</td>
<td>1.00</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_4658</td>
<td>Peptidase</td>
<td>X</td>
<td>0.14</td>
<td>0.13***</td>
<td>1.11</td>
<td>0.13</td>
<td>1.11</td>
<td>0.12</td>
<td>1.12</td>
<td>M.S,Sc,Y,E,A,P</td>
<td>Pr,W,Mj,V,G</td>
</tr>
<tr>
<td>dsim_GLEANR_18113</td>
<td>Serine Protease</td>
<td>X</td>
<td>0.82</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>M.S,Sc,Y,E</td>
<td></td>
</tr>
</tbody>
</table>

Prokupek 2008
<table>
<thead>
<tr>
<th>Genomic Accession</th>
<th>GenBank Accession</th>
<th>Protein Function</th>
<th>dN/dS</th>
<th>pS</th>
<th>dN/dS</th>
<th>pS</th>
<th>dN/dS</th>
<th>pS</th>
</tr>
</thead>
<tbody>
<tr>
<td>dsim_GLEANR_6881</td>
<td>CG31686</td>
<td>Unknown</td>
<td>0.66</td>
<td></td>
<td>1.18</td>
<td></td>
<td>1.18</td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_14267</td>
<td>CG32068</td>
<td>Secondary</td>
<td>0.39</td>
<td>0.08***</td>
<td>2.55</td>
<td>0.00</td>
<td>6.09</td>
<td>0.01**</td>
</tr>
<tr>
<td></td>
<td></td>
<td>metabolism</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_17130</td>
<td>CG32072</td>
<td>Protein-protein</td>
<td>0.10</td>
<td>0.02***</td>
<td>1.76</td>
<td>0.21</td>
<td>1.76</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>interact.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dsim_GLEANR_15288</td>
<td>CG32834</td>
<td>Serine Protease</td>
<td>0.19</td>
<td>0.21***</td>
<td>2.74</td>
<td>0.02</td>
<td>5.15</td>
<td>0.06*</td>
</tr>
<tr>
<td>dsim_GLEANR_10398</td>
<td>CG33134</td>
<td>Apoptosis</td>
<td>0.06</td>
<td>0.06</td>
<td>0.97</td>
<td>0.03</td>
<td>1.00</td>
<td>0.06</td>
</tr>
<tr>
<td>dsim_GLEANR_5795</td>
<td>CG33306</td>
<td>Unknown</td>
<td>0.04</td>
<td>0.27</td>
<td>0.47</td>
<td>0.03</td>
<td>1.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

1 Predicted protein function; 2"X" indicates that the proteins has a predicted secretion signal; 3"X" indicates the proteins has a predicted transmembrane region; 4 dN/dS: pairwise comparison of D. melanogaster and D. simulans sequences, estimated assuming no rate heterogeneity; 5 Species of Drosophila from which sequences were obtained for PAML analysis M= melanogaster, S= simulans, Sc= sechellia, Y= yakuba, E= erecta; A= ananassae; P= pseudoobscura. The following statistics are all derived from PAML analysis: 6 pS: the proportion of sites estimated to belong to the class that has the highest dN/dS in M3; 7 dN/dS: for the highest class in M3; 8 pS: the proportion of sites estimated to belong to the class that has dN/dS > 1 in M2; 9 dN/dS: the estimate for the class with the ratio > 1 in M2; 10 pS: the proportion of sites estimated to belong to the class that has dN/dS > 1 in M8; 11 dN/dS: the estimate for the class with the ratio > 1 in M8; Statistical significance for LRT: * P <.05; ** P<.01; *** P<.001. 12 The genomes in which orthologs were found using the 12 Drosophila genomes. Pr = persimilis, W = willistoni, Mj = mojavensis, V = virilis, and G = grimshawi.
Chapter III.

Resequencing study of spermathecal genes: a cluster of proteases and a gene of unknown function.
ABSTRACT

Genes expressed within the *Drosophila* spermathecae (a female sperm storage organ) are diverging at a rate exceeding that of *Drosophila* male accessory gland proteins, which are a paradigm for rapid evolution. Extensive studies of male reproductive genes and their products, such as accessory gland proteins in male *Drosophila*, have alluded to the importance of male-female interactions in the rapid evolution of reproductive proteins. Some of the most rapidly evolving genes within the spermathecae are serine-type proteases. Female proteases are candidate genes for such interactions; they have secretion signals and could be functioning in cascades, processing of male proteins or interactions with male serine protease inhibitors. Genes of unknown classification identified in the spermathecae are candidates because they may be playing a role specific to spermathecae function. In this study, we sequenced 5 genes, (4 proteases and 1 gene of unknown classification) located on the second chromosome in *Drosophila melanogaster* previously identified as expressed in the spermathecae. Polymorphism data reveals significant polymorphism-to-divergence heterogeneity in specific regions of the genes and a significant departure from equilibrium-neutral expectations in 3 genes. This data supports the hypothesis of positive selection occurring in these genes within *Drosophila* populations.
INTRODUCTION

Female sperm storage is an essential reproductive process utilized by most animals with internal fertilization (reviewed in (Neubaum & Wolfner, 1999c, Bloch-Qazi et al., 2003). Sperm storage is thought to protect sperm from displacement, extend the amount of time sperm are available for fertilization, and provide an area for competition between the ejaculates of males and for female sperm choice (Parker, 1970, Thornhill, 1983, Birkhead & Moller, 1998, Eberhard, 1996). Despite the importance of sperm storage, the identity and function of reproductive proteins associated with female sperm storage organs (SSOs) are poorly understood. These proteins, which are likely to mediate SSO processes such as storing, maintaining and releasing sperm, could also play key roles in female interactions with male reproductive proteins such as male accessory gland proteins (Acps), possibly mediating sperm competition or facilitating female cryptic choice of sperm or counteracting the detrimental effects of these proteins. An understanding of sperm storage is essential to defining important co-evolutionary processes such as sexual selection and sexual conflict.

Insect male accessory gland proteins (Acps) have been foci for molecular population genetic and molecular evolutionary studies. The average rate of sequence divergence of *D. melanogaster* Acps is approximately twice that of non-reproductive proteins (Begun et al., 2000a, Swanson et al., 2001a-a, Mueller et al., 2005a, Wagstaff & Begun, 2004b). At least seven Acps are transferred to the SSOs after mating (Ravi-Ram & Wolfner, 2005), including rapidly evolving Acps such as the protease inhibitor Acp62F (Lung et al., 2002b) and a protein that plays a key role in sperm storage (Acp36DE). Comparatively, the female reproductive genes are understudied, but the signature of positive selection has been revealed by a *D. simulans* evolutionary expressed sequence tag (EST) study (Swanson et al., 2004a), *D. melanogaster* molecular population genetic studies of female reproductive tract genes (Lawniczak & Begun, 2007), and an evolutionary EST study using *D. arizonae* reproductive tracts (Kelleher et al., 2007). A study focusing directly on Drosophila spermathecae, the long-term sperm storage organ in female, found that genes enhanced for spermathecae-specific expression were evolving under positive selection (Prokupek et al.). The proportion of positively selected genes in
the spermathecae (38%) (Prokupek et al.) detected by maximum likelihood analysis is greater than that seen in male seminal fluid genes (16%) (Haerty et al., 2007).

In *D. melanogaster*, second male sperm precedence (P2) can be partially attributed to a non-sperm component of the ejaculate (Harshman & Prout, 1994). Male Acps play a major role in sperm competition in *Drosophila* (Ravi-Ram & Wolfner, 2007). Allelic variation in male Acp genes has been associated with first male sperm precedence (P1) and P2 in this species (Clark et al., 1995, Fiumera et al., 2005, Fiumera et al., 2007). Although the specific genes remain unknown, genetic studies reveal that females play a major role in sperm competition in *D. melanogaster* (Clark & Begun, 1998, Clark et al., 1999). SSOs are a likely place to find specific female reproductive molecules that affect sperm competition.

*Drosophila* species typically have two types of organs dedicated to sperm storage (Fowler, 1973, Pitnick et al., 1999). The seminal receptacle contains the majority (65 - 80%) of the sperm (Lefevre & Jonsson, 1962, Neubaum & Wolfner, 1999a), while a pair of spermathecae are the site of long term storage. *Drosophila* spermatheca are cuticle-lined organs derived from ectodermal tissue partially surrounded by secretory cells. Sperm are stored in the spermathecal lumen, which receives proteins of unknown function from surrounding secretory epithelial cells (Filosi & Perotti, 1975). The interaction between sperm and SSOs are known to be evolutionarily important. For example, changes in sperm length result in the evolution of changes in the length of the seminal receptacle (Miller & Pitnick, 2002, Miller & Pitnick, 2003). Dynamic evolutionary interactions could exist between female reproductive proteins and male reproductive proteins, particularly in the spermathecae due to the extended period of time sperm spend in this organ.

A high proportion of genes expressed in the spermathecae show rapid rates of evolution, and an especially high level of positive selection, characteristic of reproductive proteins (i.e., sperm-egg interaction proteins, Acps) (Prokupek et al.). Serine-type proteases were overrepresented in the spermathecae, and were among the most rapidly evolving genes identified. Serine proteases could interact with male seminal products in a number of ways, such as the digestion of male proteins, proteolytic cleavage of male proteins for activation or in cascades (immune response, semen coagulation). Genes of
unknown classification were also among the most rapidly evolving genes expressed in the spermathecae. Genes with unknown classification might be quite interesting in terms of having spermatheca-specific roles because functions for such genes have not been identified in other tissues or taxa.

A proper investigation into the genetic evolution within a species requires an accurate description of variation both within and between populations, as well as the ability to distinguish between the potential causes of the observed variation distribution (Begun & Aquadro, 1993). There are two types of measurement for genetic variation at the DNA level: the average number of pairwise nucleotide differences and the number of segregating (polymorphic) sites among a sample of DNA sequences (Takahata & Clark, 1993). Comparative approaches (i.e., between species analysis by a maximum likelihood such as that conducted by the PAML program) use a comparison of the ratio of nonsynonymous changes per nonsynonymous site to synonymous changes per synonymous site in order to test for past selection. These tests, while informative, can not sufficiently differentiate between the forces of selection, nor do they allow inference into present day selection. Population genetic approaches aim at detecting ongoing selection in a population by analyzing current levels of polymorphisms. In addition, polymorphic approaches are much more sensitive than interspecific approaches for the identification and classification of selective pressures. Measures of the level of polymorphism can be revealing in terms of discovering forces acting on particular genomic loci. Low levels of polymorphism indicate either directional selection or genetic drift while high levels of polymorphism are reflective of high mutation rates, large historic population size, neutral evolution or balancing selection. The combination of multiple test statistics can be used to reveal more about the patterns of selection (Otto, 2000).

Relatively little research has been done to detect positive selection in noncoding regions based on comparative data. Methods similar to those used to detect elevated $K_{a}/K_{s}$ have been devised as a way of estimating selection in these regions (Wong & Nielsen, 2004, Haygood et al., 2007). The presence of highly variable sites in noncoding regions may be interesting in terms of the evolutionary changes which occur in noncoding regulatory regions (typically upstream regions) which result in phenotypic differences due to changes in gene regulation. A large portion of the non-translated
genome of *Drosophila* has been suggested as functionally important, and therefore subject to both purifying and adaptive evolution (Andolfatto, 2005).

Studies on genetic variation provide a powerful means for elucidating the genetic, evolutionary and demographic factors shaping the *Drosophila* genome. In the present study we examined genetic variation for 5 genes (and regions associated with these genes) on the chromosome 2L in 35 lines from 2 populations of *Drosophila melanogaster* (19 North American, and 16 African). Four of the genes investigated were serine-proteases, and one was a gene of unknown classification. These genes all have enhanced expression in the spermathecae as well as signal peptides (proteases) or transmembrane domains (gene of unknown classification) and thus may be involved in interactions with male seminal products.

This study addressed two major questions: (1) Is the pattern of variation in these genes incompatible with neutral evolution? (2) What is the pattern of variation in non-coding regions immediately upstream and downstream of coding regions (likely containing promoter and regulatory regions) compared to a non-coding region not associated with a gene?

**MATERIALS and METHODS**

*Experimental line origins and DNA sequencing*

*D. melanogaster* flies used in this study came from second chromosome extracted lines from two populations (Uganda and Pennsylvania). Nineteen lines were used from the Pennsylvania (PA) population, and 16 lines from the Uganda (UG) population. DNA was extracted from single males from each line.

DNA sequences of the complete coding and non-coding (upstream, downstream, and intronic) regions for each of the proteases genes, and the gene of unknown function were produced. Eleven primer pairs were designed along the entire regional cluster of four protease genes (CG17012, CG17234, CG17239, CG17240 (Ser12)). The resulting sequence included the coding regions, all intergenic sequences between the protease genes and stretches of DNA (~500bp) immediately upstream and downstream of the cluster (genomic sequence location 2L: 2250003 – 2255500). The genes of the cluster contain no introns. Five primer pairs were designed for a gene of unknown function
(CG15098) located on the second chromosome which included upstream and downstream
regions (~500bp) (genomic sequence location 2R: 14720524 - 147722760). One primer
pair was designed for a randomly selected region (~600bp) of non-coding DNA from
chromosome 2 (genomic sequence location 2L: 2257882 – 2258882). The closest gene to
this region is 2832bp upstream. DNA samples were sent to High Throughput Sequencing
Solutions (Seattle, WA) for primer design, PCR amplification and 2-way sequencing.

**Clean-up and Alignment**

The Contig Express program (Invitrogen) was used to clean-up the DNA
sequences, as well as to create contiguous sequences out of the individual sequence reads.
Both forward and reverse strand sequences were used to create the consensus sequence
for each sample set. Regional sequences from Flybase (Release 5.1; [http://flybase.org](http://flybase.org))
were used as a reference for proper alignment of the contigs. Regions where sequences
were ambiguous between the two strands, or missing in sequence data, were replaced
with dummy (N) sequences. Alignments of each region were generated with the AlignX
program of Vector NTI Advance. Sequences from Flybase were used as a reference for
the coding frame; all sequences were adjusted by hand when necessary to keep the proper
coding frame.

**DNA sequence analysis**

**Polymorphism**

Polymorphism analysis tests were done using the DnaSP version 4.20 software
package (Rozas et al., 2003). A sliding window analysis of the ratio of polymorphism to
divergence using a window of 50bp with a step size of 10 was also conducted using
DnaSP. To test for departures from neutrality we calculated Tajima’s D (Tajima, 1989),
Fu and Li’s D (Fu & Li, 1993), and Fay and Wu’s H (Fay & Wu, 2000). Significance for
Fay and Wu’s H statistic was determined by coalescent simulations using a
recombination parameter, R, estimated by DnaSP for each region. Tests based on the
frequency spectrum (Fay and Wu’s H, Tajima’s D, and Fu and Li’s D) detect levels of
variation that are inconsistent with the expectation of neutral equilibrium model, but are
influenced by demographic effects. Fu and Li’s D compares the number of mutations

Prokupek 2008
which occur in external versus internal branches of the genealogy (Fu & Li, 1993).

Tajima’s D test compares a standardized measure of the total number of polymorphic sites in the sample with the average number of mutations between pairs in the sample (Tajima, 1989). If D (either Tajima’s or Fu and Li’s) is significantly different from 0, we reject the neutral mutation hypothesis. A D significantly smaller than 0 is suggestive of directional selection, greater than 0 is suggests balancing selection. One problem in using the statistics from these tests alone is that selective sweeps, bottlenecks and population expansion/reduction can result in similar effects at particular loci. Fay and Wu’s H compares the proportion of alleles at intermediate versus high frequency. Positive selection is indicated by a negative value of H (Fay & Wu, 2000). This test is useful in the detection of past hitchhiking events because H is sensitive to an excess of new, high frequency alleles.

Divergence

The McDonald – Kreitman (MK) (McDonald & Kreitman, 1991) test was conducted to compare levels of synonymous and nonsynonymous variation within and between species. The neutral prediction is that the ratio of nonsynonymous to synonymous substitutions is equal to the ratio of nonsynonymous to synonymous polymorphisms.

The Hudson-Kreitman-Aguade (HKA) (Hudson et al., 1987) test, used in the present study, predicts a correlation between the levels of polymorphism within a species, and divergence between two closely related species. HKA is a goodness of fit test, comparing estimates of parameters describing within-species diversity and between-species divergence using data from two or more loci (Hudson et al., 1987). If functional constraints limit evolution, then levels of polymorphism and divergence are equally constrained. This test is typically applied to sites of silent variation (untranslated or synonymous sites), testing if the pattern of silent variation within and between species is different for two regions. The HKA test uses a neutral locus for comparison; a random, non-coding region of the 2nd chromosome was used in the present study as a neutral loci. This locus was relatively far from any up or downstream coding regions, and thus is unlikely to contain promoter or regulatory regions. Tests of neutrality were also
performed on this region to ensure that it conforms to the hypothesis of neutral evolution. Homogeneity between the loci of interest and the assigned neutral loci would support a hypothesis of neutrality. Failure to reject this model implies that the differences among locus polymorphisms arise from differences in neutral mutation rates.

In tests requiring between species comparisons (HKA and MK), the corresponding chromosome region from D. simulans (a sister species of D. melanogaster) was used. The orthologous region of D. simulans was determined from NCBI BLAST (Altschul et al., 1990), and aligned using Clustal W. All sequences were manually adjusted when necessary in order to keep the coding frame. The use of a sister species prevents the outgroup species from being too distantly related. If the species are too distant, the tests will be biased in the direction of excess changes within a species by violating the assumption of a single mutation per site. If the species are too close, the test could be biased with a finding of excess changes between species, because shared polymorphisms will be confounded with fixed differences (reviewed in Wayne & Simonsen, 1998).

RESULTS

The evolutionary patterns of genes expressed within the spermathecae of Drosophila melanogaster were examined. The intention was to expand knowledge of the forces driving the evolution of female reproductive traits, in this case, sperm storage. Six loci on chromosome 2L, five gene regions including upstream, downstream, and intronic regions, as well as a single non-coding region in 35 lines from 2 populations of Drosophila melanogaster (19 North American, and 16 African) and one outgroup (D. simulans) were used for analyses. Analyses were performed when it was determined that data would not be skewed by limited sequence information.

Numerous indels were detected in the intron and non-coding region alignments and were ignored in the analysis. A large indel (~1000bp) was present in the D. simulans genome between the ortholog to CG17239 and CG17234; this region was missing in the D.melanogaster genome. This region in the D. simulans genome was ignored for analysis. This region encodes the D. simulans gene dsim_GLEANR_6594. Using a conserved domain search, this gene is predicted to be a serine protease (E-value $6^{-46}$).
Polymorphism analyses

Polymorphism studies allow for more comprehensive inferences about the nature of selective forces driving the evolution of genes compared to divergence studies alone. The data from polymorphism studies are used for testing deviations for neutrality. A summary of polymorphisms is presented in Table 3.1. The genome average of heterozygosity calculated from a random sample of autosomal genes is .001 (Andolfatto, 2001). The average replacement heterozygosity calculated for the genes in this study is .007. This number is much greater than the genome average and similar to that seen in the Acps of *D. melanogaster* (Begun et al., 2000b).

Polymorphism tests show that three of the genes significantly deviate from the neutral expectations (CG17012, CG17234, and CG17240) (Table 3.2). In all three cases Tajima’s D values were negative. Negative values can be due to the presence of an excess of low-frequency variants usually caused by a sudden population expansion or by the selection of one specific allele over alternative alleles. The same three genes had significantly negative Fu and Li’s D which is consistent with positive selection.

Divergence

*Ka* (nonsynonymous substitutions per nonsynonymous site) and *Ks* (synonymous substitutions per synonymous site) estimates for the comparisons between *D. melanogaster* and *D. simulans* are presented in Table 3.3. The genome average synonymous site divergence (*Ks*) between these two species is estimated to be ~0.12 (Begun & Whitley, 2000) and the average amino acid divergence (*Ka*) average is estimated to be ~.011 (Begun, 2002). The average *Ka* for the protease genes of this study is 0.07, over six times the calculated genome wide average, whereas the *Ks* average for the protease genes of this study is similar to the genome wide average (0.11). We also compared our values to the average *Kd*, and *Ks* values for 166 randomly chosen serine proteases (*Ks* =0.1279, *Kd* = 0.0258, *Kd/Ks* = 0.203) (Lawniczak & Begun, 2007). On average, the *Kd* (0.07) and *Kd/Ks* (0.66) of the spermathecae serine proteases was 2 to 3-fold higher than that of a random selection of serine proteases.
When comparing the gene of unknown function to the genomic average, we see that although the $K_a$ value is markedly higher than average, the $K_s$ value is lower (Table 3.4). This indicates that the elevated $K_a/K_s$ ratios are affected by changes in both the nonsynonymous and synonymous substitution rates. Decreases in synonymous substitution rates may be due to selection for optimal codon usage (Moriyama & Powell, 1997), or an overall decrease in the mutation rate at this locus. The $K_a/K_s$ ratio of the gene of unknown function is as high, or higher than the protease ratios due to the low $K_s$ value.

_Hudson, Kreitman, Aguade Test_

The HKA test was used to assess the neutral prediction that the ratio of polymorphism to divergence is the same for different loci for a series of pairwise comparisons with _D. simulans_ as the outgroup. This test is used for detecting a deficit or excess of polymorphism either due to directional selection or balancing selection, respectively. A single non-coding region of ~600 base pairs (over 2kb away from the closest coding region) was used as a ‘neutral’ locus for comparison. We also used the upstream region of each gene as an additional locus for each comparison. Additionally for CG15098 we concatenated the introns and used them as an additional neutral locus. All five genes conformed to the neutral prediction in all comparisons.

_McDonald-Kreitman Test._

The MK test is analogous to the HKA test, but by tracking synonymous versus nonsynonymous substitution it is more sensitive to the detection of positive selection. The MK test was not significant for any of the genes tested. The ratio of nonsynonymous to synonymous fixed substitutions was not statistically different from the ratio of nonsynonymous to synonymous polymorphic sites.

For all of the protease genes, there are a number of replacement polymorphisms, but no amino acid fixations between the PA and UG populations. The excess of replacement polymorphisms is consistent with both neutral processes (relaxation of functional constraints) and selective processes (segregation of deleterious mutation or balancing selection).

Prokupek 2008
Non-coding DNA

Evolutionary changes which occur in non-coding regulatory regions (typically upstream regions) may result in phenotypic differences due to changes in gene regulation. A large portion of the non-translated genome of Drosophila, including regulatory regions, is potentially functionally important, and therefore subject to both purifying and adaptive evolution (Andolfatto, 2005). Promoter regions were predicted using the Neural Network Promoter Prediction from the Berkley Drosophila Genome Project (http://www.fruitfly.org/seq_tools/promoter.html). Significant negative values for Tajima’s D and Fu and Li’s D were seen in upstream regions of CG17012, CG17234, CG17239 and CG17240, consistent with positive selection.

The average divergence between D.melanogaster and D.simulans in untranslated regions (UTRs) is estimated to be 0.045 (Andolfatto, 2005). We chose to use the randomly selected genomic region of the second chromosome for comparison with the upstream and intronic regions associated with the genes in our study (divergence ($K_g$) = ~0.064). To get an idea of the evolutionary rates of the upstream and intronic regions of our study genes we calculated the ratio of substitution rates in the upstream/intronic regions compared to those in the random genomic regions (Tables 3.2 and 3.3). This ratio is analogous to the ratio of substitution rates at nonsynonymous sites to synonymous sites in coding regions, and allows a way to estimate the forces of positive selection acting on non-coding regions (Wong & Nielsen, 2004, Haygood et al., 2007). The average value for the ratio of upstream regions to genomic region substitutions was 2.5 ± 0.75. The average value for the intronic regions of CG15098 was 1.56, and for the downstream region of CG15098 was 1.25. The upstream regions are on average twice as diverse as the random genomic, intronic and downstream regions.

Sliding Window Analysis

Oftentimes different regions of a single gene can be exposed to different selective pressures. In these cases, calculating $K_a/K_s$ over the entire length of the gene does not provide an accurate picture of how selection is acting on the gene. Both the MK and the HKA tests use the average polymorphism and divergence rate for each gene. Failure to
reject the neutral predictions for these tests may be explained by the over-all gene averages masking selection on specific regions. Sliding window analysis of both polymorphism and divergence can help determine if specific regions and codons in a gene are under selection, or if they are highly conserved. The sliding widow analysis (Figures 3.1-5) shows inflated $K_a/K_s$ (divergence) values relative to $\Pi_a/\Pi_s$ (polymorphism) values for all five of the genes analyzed. This result is strongly indicative of these regions are evolving under positive selection. Moreover, the sliding window analysis was helpful in finding non-coding regions which showed higher than average levels of polymorphism (Figures 3.6 and 3.7).

DISCUSSION

Female reproductive proteins are evolving rapidly in *Drosophila* (Civetta & Singh, 1995, Panhuis & Swanson, 2006). Sexual selection has been implicated to explain the evolution of reproductive proteins (e.g. gamete recognition proteins) (Swanson & Vacquier, 2002). The exact nature of selection driving the evolution of spermatheca genes is not known, but it is likely to be a form of sexual selection possibly driven by interactions with male seminal products.

Sperm storage is a dynamic process in which both males and females play active roles (Adams & Wolfner, 2007). In *D. melanogaster*, a feminized nervous system is required for accumulation of sperm in storage (Arthur et al., 1998). Female secretions, as seen in the honeybee *Aphis mellifera*, are necessary to support sperm storage (Weirich et al., 2002, Collins et al., 2004). Male seminal proteins are also required for proper sperm storage and accumulation (Kalb et al., 1993, Tram & Wolfner, 1999). Roles of both the male and female have been identified in sperm competition (Clark et al., 1995, Clark et al., 1999, Clark & Begun, 1998, Hosken, 1999, Hosken & Ward, 2000, Lawniczak & Begun, 2005). The interactions between male and female reproductive proteins have been recognized as important in evolutionary processes such as male-female co-evolution, sexual selection and speciation (Pitnick et al., in prep.).

Very little is known about reproductive proteins present in SSOs. A previous evolutionary expressed sequence tag (EST) study on genes expressed within the spermathecae identified this organ as a region of high rates of evolution (Prokupek et al.).
PAML analysis conducted in the EST study predicted that a high proportion of genes in the spermathecae are rapidly evolving. The five genes in the present study were previously analyzed using overall $K_a/K_s$ ratios as well as in a PAML analysis using multiple species. It was determined that all five of these genes showed signs of positive selection on a subset of their codons, but without polymorphism data the previous study lacked information about evolution within species and it lacked the comprehensive battery of tests used in the present study.

Examination of both divergence and polymorphism provided insight into the evolutionary history of a set of proteases expressed in the spermathecae. The failure to reject Fay and Wu’s $H$ indicates that the reduced level of variation observed is not due to hitchhiking (Otto, 2000). A non-significant $H$, combined with the results of other neutrality tests (Tajima’s $D$, Fu and Li’s $D$), is suggestive that three of the proteases are undergoing positive selection. Positive Fu and Li’s $D$ test values were found in the upstream region of CG17239 as well as the downstream region of CG16098, indicating that balancing selection may be operating in these regions. Positive strong peaks (greater than 1) of $K_a/K_s$, along with low levels of relative polymorphism at the same point, were observed in the sliding window analysis. These peaks suggest that positive selection is operating on a subset of the codons in all of the protease genes.

Serine –type endopeptidases are recognized as important reproductive proteins both in the female reproductive tract, and in seminal fluids. Female reproductive tract secretion of digestive enzymes alludes to a form of ejaculate-female interaction in which females actively degrade or activate protein components of the male ejaculate. Serine proteases, found in the spermathecae may be interacting with Acps which enter the sperm storage organs. Four of the Acps found in the sperm storage organs are serine protease inhibitors. Serine proteases inhibitors (serpins) bind irreversibly to serine proteases, functionally inactivating them (Potempa et al., 1994). While no direct interaction between the serine proteases of the female reproductive tract and serpin Acps has been found thus far, it does present an interesting scenario for these two classes of proteins. Serine proteases could also be acting in a non-antagonistic manner with male proteins in coordinating immune response or semen coagulation cascades.
The four protease genes investigated in the present study are found in a cluster on chromosome arm 2L. They have no introns and they encode proteins with the canonical serine protease catalytic triad of amino acids. The cluster is transcriptionally activated by mating (Lawniczak & Begun, 2007). In *Drosophila arizonae*, reproductive tract-specific digestive proteases show evidence of directional selection, with certain amino acids experiencing strong positive selection (Kelleher et al., 2007).

Tests of CG15098 (gene of unknown classification) failed to reject the hypothesis of neutrality. However, this gene showed higher than average levels of heterogeneity in polymorphism along the length of the gene. Sliding window analysis revealed two peaks which could indicate that positive selection is operating in these specific regions of the gene. Genes of unknown function could play roles that are spermathecae-specific. The gene analyzed in this study has a predicted transmembrane region, and could be interacting with male proteins in the spermathecae. The two regions identified by sliding window analysis could possibly be the regions of male-female interaction.

Polymorphism tests on the non-coding regions of the genes revealed that regions upstream of the genes are two to three times more polymorphic than the designated neutral regions. Analysis of specific predicted promoter regions showed that these regions, present in the upstream regions, had relatively low levels of polymorphism. Upstream regions, though not translated, could be undergoing positive selection which affects expression levels of the genes.

Patterns of nucleotide variation can reveal rapid evolution by positive selection as seen in *Drosophila* male reproductive proteins (Acps) (Swanson et al., 2001a). Acps, which are transferred to the female at the time of mating, play important roles in sperm storage, sperm competition/defense, female mate receptivity, egg laying rate, and female lifespan (reviewed in (Wolfner, 1997). Some Acps have been found to localize in specific regions of the female reproductive tract including the genital opening, the distal portion of the uterus and in sperm storage structures (Bertram et al., 1996, Heifetz et al., 2000, Ravi-Ram & Wolfner, 2005). Due to the important role Acps play in reproduction it is not surprising that these proteins have been found to evolve at twice the rate of non-reproductive proteins. What is perhaps more intriguing, is that female counter-parts to
These proteins remain undiscovered, although secreted proteins encoded by spermathecae genes are promising candidates to play this role.

This study provided an in depth look into the evolutionary patterns shaping five genes expressed within the spermatheca, the long term sperm storage organ of *Drosophila*. The results of this study indicate that the evolutionary patterns of female reproductive proteins present in the spermathecae is consistent with positive selection operating in a subset of the codons. The evolutionary pressures shaping these genes could be generated by male-female interactions. Further investigation is needed to identify specific proteins interacting with these candidate genes.

**Acknowledgements**

This research was supported by a National Science Foundation grant (DEB-ESP0346476).
REFERENCES


Figure 3.1 Sliding window analysis of divergence $K_a/K_s$ ratios and polymorphism $\pi_a/\pi_s$ in the gene CG15098. The solid line indicates a $K_a/K_s$ ratio of one, ratios larger than one are indicative of positive selection.
Figure 3.2  Sliding window analysis of divergence $K_a/K_s$ ratios and polymorphism $\pi_a/\pi_s$ in the gene CG17012. The solid line indicates a $K_a/K_s$ ratio of one, ratios larger than one are indicative of positive selection.
CG17012 CDS

Nucleotide Position

$P_i(a)/P_i(s)$

$K_a/K_s$
Figure 3.3 Sliding window analysis of divergence $K_a/K_s$ ratios and polymorphism $\pi_a/\pi_s$ in the gene CG17234. The solid line indicates a $K_a/K_s$ ratio of one, ratios larger than one are indicative of positive selection.
Figure 3.4 Sliding window analysis of divergence $K_a/K_s$ ratios and polymorphism $\pi_a/\pi_s$ in the gene CG17239. The solid line indicates a $K_a/K_s$ ratio of one, ratios larger than one are indicative of positive selection.
Figure 3.5  Sliding window analysis of divergence $K_a/K_s$ ratios and polymorphism $\pi_a/\pi_s$ in the gene CG17240 (Ser12). The solid line indicates a $K_a/K_s$ ratio of one, ratios larger than one are indicative of positive selection.
Figure 3.6 Sliding window analysis of substitutions ($K_{nc}$) and polymorphism ($\pi_{nc}$) in the non-coding regions (upstream, downstream and introns) of gene CG15098. Coding regions were excluded from analysis.
Noncoding regions of CG15098

- Pi(nc)
- Knc

Exon 1 & 2
Exon 3

Nucleotide Position
Figure 3.7 Sliding window analysis of substitutions ($K_{nc}$) and polymorphism ($\pi_{nc}$) in the non-coding regions (upstream) of the genes in the protease cluster (CG17012, CG17234, CG17239 and CG17240 (Ser12)). Coding regions were excluded from analysis.
Table 3.1 Polymorphism summary statistics for the coding regions in the cluster of proteases and the gene of unknown classification.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Pop</th>
<th>size</th>
<th>n</th>
<th>S</th>
<th>πₐ</th>
<th>πₛ</th>
<th>πₐ/πₛ</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG17012</td>
<td>all</td>
<td>240</td>
<td>35</td>
<td>21</td>
<td>0.006</td>
<td>0.017</td>
<td>0.368</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>766</td>
<td>19</td>
<td>27</td>
<td>0.008</td>
<td>0.016</td>
<td>0.480</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>240</td>
<td>16</td>
<td>19</td>
<td>0.008</td>
<td>0.017</td>
<td>0.445</td>
</tr>
<tr>
<td>CG17234</td>
<td>all</td>
<td>179</td>
<td>35</td>
<td>26</td>
<td>0.007</td>
<td>0.020</td>
<td>0.367</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>572</td>
<td>19</td>
<td>6</td>
<td>0.0003</td>
<td>0.002</td>
<td>0.162</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>202</td>
<td>16</td>
<td>29</td>
<td>0.014</td>
<td>0.032</td>
<td>0.424</td>
</tr>
<tr>
<td>CG17239</td>
<td>all</td>
<td>724</td>
<td>35</td>
<td>22</td>
<td>0.004</td>
<td>0.011</td>
<td>0.377</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>747</td>
<td>19</td>
<td>16</td>
<td>0.004</td>
<td>0.008</td>
<td>0.454</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>724</td>
<td>16</td>
<td>18</td>
<td>0.005</td>
<td>0.013</td>
<td>0.387</td>
</tr>
<tr>
<td>CG17240</td>
<td>all</td>
<td>n/a</td>
<td></td>
<td></td>
<td>0.019</td>
<td>0.017</td>
<td>1.112</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>161</td>
<td>19</td>
<td>23</td>
<td>0.019</td>
<td>0.017</td>
<td>1.112</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG15098</td>
<td>all</td>
<td>980</td>
<td>35</td>
<td>25</td>
<td>0.006</td>
<td>0.007</td>
<td>0.808</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>980</td>
<td>19</td>
<td>16</td>
<td>0.005</td>
<td>0.005</td>
<td>1.086</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>981</td>
<td>16</td>
<td>22</td>
<td>0.005</td>
<td>0.009</td>
<td>0.533</td>
</tr>
<tr>
<td>Gene average</td>
<td>all</td>
<td></td>
<td></td>
<td></td>
<td>0.006</td>
<td>0.009</td>
<td>1.306</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td></td>
<td></td>
<td></td>
<td>0.007</td>
<td>0.010</td>
<td>0.659</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td></td>
<td></td>
<td></td>
<td>0.008</td>
<td>0.018</td>
<td>0.447</td>
</tr>
</tbody>
</table>

Genomic average from literature 0.001 0.009 0.111

All: combined data analysis from both the Pennsylvania and Uganda lines; PA: Pennsylvania lines; UG: Uganda lines; Size: the length of sequence data used for each gene; n: the number of lines used for analysis; S: the number of segregating sites; πₐ: a measure of nonsynonymous nucleotide diversity; πₛ: a measure of synonymous nucleotide diversity.
Table 3.2 Neutrality test summary statistics for the coding and non-coding regions associated with the cluster of genes and gene of unknown function, and the non-coding random genomic region.
<table>
<thead>
<tr>
<th>Location</th>
<th>Fu and Li’s D</th>
<th>p</th>
<th>Tajima’s D</th>
<th>p</th>
<th>Fay and Wu’s H</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomic</td>
<td>PA</td>
<td>0.48</td>
<td>ns</td>
<td>0.09</td>
<td>ns</td>
<td>-2.86</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-1.70</td>
<td>ns</td>
<td>-1.27</td>
<td>ns</td>
<td>-8.38</td>
</tr>
<tr>
<td>Upstream 17012</td>
<td>PA</td>
<td>0.38</td>
<td>ns</td>
<td>0.81</td>
<td>ns</td>
<td>-0.02</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-2.18</td>
<td>0.05</td>
<td>-1.58</td>
<td>0.05</td>
<td>-0.29</td>
</tr>
<tr>
<td>CG17012 CDS</td>
<td>PA</td>
<td>-0.59</td>
<td>ns</td>
<td>-0.42</td>
<td>ns</td>
<td>-2.65</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-2.44</td>
<td>0.05</td>
<td>-1.94</td>
<td>0.05</td>
<td>0.94</td>
</tr>
<tr>
<td>Upstream 17234</td>
<td>PA</td>
<td>-3.05</td>
<td>0.02</td>
<td>-2.03</td>
<td>0.05</td>
<td>-0.88</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-2.59</td>
<td>0.02</td>
<td>-2.03</td>
<td>0.05</td>
<td>-0.05</td>
</tr>
<tr>
<td>CG17234 CDS</td>
<td>PA</td>
<td>-2.76</td>
<td>0.02</td>
<td>-2.02</td>
<td>0.05</td>
<td>0.66</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-3.31</td>
<td>0.02</td>
<td>-2.21</td>
<td>0.01</td>
<td>2.80</td>
</tr>
<tr>
<td>Upstream 17239</td>
<td>PA</td>
<td>1.51</td>
<td>0.02</td>
<td>1.56</td>
<td>ns</td>
<td>-1.14</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-2.81</td>
<td>0.02</td>
<td>-1.01</td>
<td>ns</td>
<td>0.73</td>
</tr>
<tr>
<td>CG17239 CDS</td>
<td>PA</td>
<td>-0.60</td>
<td>ns</td>
<td>-0.90</td>
<td>ns</td>
<td>-5.86</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-0.69</td>
<td>ns</td>
<td>-0.06</td>
<td>ns</td>
<td>-0.16</td>
</tr>
<tr>
<td>Upstream 17240</td>
<td>PA</td>
<td>-2.66</td>
<td>0.05</td>
<td>-2.05</td>
<td>0.05</td>
<td>-3.90</td>
</tr>
<tr>
<td>CG17240 CDS</td>
<td>PA</td>
<td>-3.25</td>
<td>0.02</td>
<td>-2.35</td>
<td>0.01</td>
<td>0.74</td>
</tr>
<tr>
<td>Upstream 15098</td>
<td>PA</td>
<td>-0.47</td>
<td>ns</td>
<td>-0.63</td>
<td>ns</td>
<td>-1.31</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.03</td>
<td>ns</td>
<td>0.51</td>
<td>ns</td>
<td>-2.90</td>
</tr>
<tr>
<td>15098 CDS</td>
<td>PA</td>
<td>-0.12</td>
<td>ns</td>
<td>-0.34</td>
<td>ns</td>
<td>-2.83</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-0.99</td>
<td>ns</td>
<td>-0.65</td>
<td>ns</td>
<td>-3.40</td>
</tr>
<tr>
<td>15098 introns</td>
<td>PA</td>
<td>0.40</td>
<td>ns</td>
<td>0.55</td>
<td>ns</td>
<td>-0.97</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.43</td>
<td>ns</td>
<td>-0.20</td>
<td>ns</td>
<td>-1.92</td>
</tr>
<tr>
<td>Downstream 15098</td>
<td>PA</td>
<td>2.17</td>
<td>0.05</td>
<td>-1.50</td>
<td>0.01</td>
<td>1.46</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>-0.88</td>
<td>ns</td>
<td>-0.35</td>
<td>ns</td>
<td>2.83</td>
</tr>
</tbody>
</table>

Population summary statistics for neutrality tests. PA: Pennsylvania lines; UG: Uganda lines; P-values for Fu and Li’s D and Tajima’s D were obtained using a 2-tailed Fisher’s exact test. P-values for Fay and Wu’s H were calculated using coalescent simulations; ns: non-significant.
Table 3.3 Analysis of the divergence in the cluster of proteases between D.melanogaster and D. simulans, including non-coding, promoter and coding regions.
<table>
<thead>
<tr>
<th>Gene</th>
<th>Coding regions</th>
<th>Non-coding regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>K_a</td>
<td>K_s</td>
</tr>
<tr>
<td>CG17012</td>
<td>all</td>
<td>0.09</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>0.11</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.09</td>
</tr>
<tr>
<td>CG17234</td>
<td>all</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.07</td>
</tr>
<tr>
<td>CG17239</td>
<td>all</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.07</td>
</tr>
<tr>
<td>CG17240</td>
<td>all</td>
<td>n/a</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>n/a</td>
</tr>
<tr>
<td>Gene average</td>
<td>all</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>PA</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>0.08</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Genomic region</th>
<th>K_a</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PA</td>
</tr>
<tr>
<td>all</td>
<td>0.064</td>
</tr>
</tbody>
</table>

K_a - nonsynonymous changes per nonsynonymous site; K_s - synonymous changes per synonymous site; K_u - silent changes in immediately upstream regions; K_g - silent changes in the randomly selected non-coding genomic region of the 2nd chromosome; K_p - changes in predicted promotor regions.
Table 3.4  Analyses of the divergence in CG15098 between *D.melanogaster* and *D. simulans*, including non-coding, promoter and coding regions.
<table>
<thead>
<tr>
<th>Gene name</th>
<th>Coding regions</th>
<th>Non-coding regions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$K_s$</td>
<td>$K_s$</td>
</tr>
<tr>
<td>CG15098</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>PA</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>UG</td>
<td>0.06</td>
<td>0.07</td>
</tr>
</tbody>
</table>

Genomic region

<table>
<thead>
<tr>
<th></th>
<th>$K_u$</th>
<th>$K_d$</th>
<th>$K_i$</th>
<th>$K_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>all</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
<td></td>
</tr>
</tbody>
</table>

$K_s$: nonsynonymous changes per nonsynonymous site; $K_s$: synonymous changes per synonymous site; $K_u$: silent changes in immediately upstream regions; $K_i$: silent changes in intronic regions; $K_d$: silent changes in regions immediately downstream; $K_p$: changes in predicted promotor region
Table 3.5: HKA summary statistics for the cluster of proteases and the gene of unknown function. HKA tests results are calculated from a comparison of neutral genomic regions compared to the coding regions of the genes.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CG17012</td>
<td>PA</td>
<td>28</td>
<td>6.43</td>
<td>-0.13</td>
<td>0.93</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>19</td>
<td>2.65</td>
<td>-0.43</td>
<td>0.90</td>
</tr>
<tr>
<td>CG17234</td>
<td>PA</td>
<td>6</td>
<td>0.71</td>
<td>-0.16</td>
<td>0.70</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>29</td>
<td>3.22</td>
<td>-0.49</td>
<td>0.84</td>
</tr>
<tr>
<td>CG17239</td>
<td>PA</td>
<td>16</td>
<td>3.63</td>
<td>-0.11</td>
<td>0.92</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>18</td>
<td>5.16</td>
<td>-0.26</td>
<td>0.86</td>
</tr>
<tr>
<td>CG17240</td>
<td>PA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG15098</td>
<td>PA</td>
<td>12</td>
<td>2.59</td>
<td>-0.12</td>
<td>0.90</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>15</td>
<td>3.37</td>
<td>-0.33</td>
<td>0.94</td>
</tr>
</tbody>
</table>

Gene: *D. melanogaster* CG number; PA: Pennsylvania lines; UG: Uganda lines; Seg. Sites: segregating sites; Ave. diff: average number of differences between *D. melanogaster* and *D. simulans*; P: P-value obtained from a $\chi^2$ test (df =1) of the difference between null and selection models.
Table 3.6: McDonal d-Kreittman test statistics for the coding regions of the protease cluster and the gene of unknown function
<table>
<thead>
<tr>
<th>Gene</th>
<th>Population</th>
<th>Polymorphic</th>
<th></th>
<th>Fixed</th>
<th></th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Silent</td>
<td>Replacement</td>
<td>Silent</td>
<td>Replacement</td>
<td></td>
</tr>
<tr>
<td>CG17012</td>
<td>PA</td>
<td>9</td>
<td>15</td>
<td>25</td>
<td>56</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>12</td>
<td>0.69</td>
</tr>
<tr>
<td>CG17234</td>
<td>PA</td>
<td>2</td>
<td>1</td>
<td>16</td>
<td>28</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>CG17239</td>
<td>PA</td>
<td>7</td>
<td>9</td>
<td>18</td>
<td>37</td>
<td>0.55</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>8</td>
<td>10</td>
<td>17</td>
<td>35</td>
<td>0.37</td>
</tr>
<tr>
<td>CG17240</td>
<td>PA</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG15098</td>
<td>PA</td>
<td>13</td>
<td>25</td>
<td>2</td>
<td>11</td>
<td>0.30</td>
</tr>
<tr>
<td></td>
<td>UG</td>
<td>12</td>
<td>26</td>
<td>5</td>
<td>12</td>
<td>1</td>
</tr>
</tbody>
</table>

Gene: D. melanogaster CG number; PA: Pennsylvania; UG: Uganda; Prob: P-values were obtained using a 2-tailed Fisher’s exact test
Chapter IV.

Patterns of gene expression in the sperm storage organs of *Drosophila melanogaster*

[Based on the manuscript submitted to *Insect Molecular Biology* by Adrianne M. Prokupek, Stephen D. Kachman, Lawrence G. Harshman]
ABSTRACT

Sperm storage is a dynamic and complex process utilized by female animals across a wide range of taxa. Sperm storage organs play an important role in reproduction and are important in evolutionary processes such as speciation. The function of these organs is poorly understood, especially at the protein level. This study investigated the transcriptome of the two *Drosophila melanogaster* sperm storage organs and identified genes likely to impact sperm storage, maintenance and use. Functional categorization of the genes expressed in the SSOs indicates that each organ play a unique roles in the process of sperm storage. The spermathecae was enriched for proteases, metabolism genes and antioxidants. The seminal receptacle exhibited a number of genes involved in localization, signaling, ion transport and immunity/defense.
INTRODUCTION

Transcriptome analysis using microarrays provides a means to investigate gene expression and identify proteins in organs for which little functional information is available. Genes expressed in these tissues/organs can be grouped by function, allowing for the identification of proteins that play significant functional roles. The female sperm storage organs (SSOs) are important in reproduction and various evolutionary processes, but little is currently known about the genes expressed in these organs for any species. There is a general lack of functional knowledge of SSOs creating a major deficit in the understanding of sperm storage, an important process of reproduction.

Female sperm storage is a reproductive strategy utilized by most internally fertilizing animals. The structure and function of SSOs appear to be both dynamic and complicated (Bloch-Qazi & Wolfner, 2003, Adams & Wolfner, 2007). Sperm storage allows females to acquire sperm to take place days or even months before it is needed to fertilize mature oocytes. The reproductive benefits of this process are potentially great as contact with a suitable male does not necessarily correspond with egg availability. For species which engage in multiple matings, sperm storage can offer an additional selection step via sperm competition and/or cryptic female sperm choice, as well as gives the female a mechanism to compensate for infertile/genetically incompatible males. The duration of time sperm spend in storage is highly variable, extending from a few hours (e.g., mouse) up to decades (e.g., honey bee). Specialized SSOs have evolved in a number of species; in insects SSOs typically appear as sac-like structures (spermathecae) or long tubules (seminal receptacles). These organs are responsible for maintaining sperm
viability, organizing the sperm in storage, and facilitating the proper release of sperm from storage (reviewed in (Neubaum & Wolfner, 1999c, Bloch-Qazi et al., 2003).

*Drosophila melanogaster* females mate multiply and store sperm in specialized SSOs (Lefevre & Jonsson, 1962, DeVries, 1964, Pitnick et al., 1999). Copulation in *D. melanogaster* results in the transfer of approximately 4000-6000 sperm (Kaplan et al., 1962), approximately a quarter of the transferred sperm are stored, and an estimated 30-80% of stored sperm are used for fertilization (Kaplan et al., 1962, Fowler, 1973). The process of sperm storage in *D. melanogaster* begins before copulation has ended (Fowler, 1973, Gilbert et al., 1981, Lefevre & Jonsson, 1962), and is complete within 6 hours after mating (Lefevre & Jonsson, 1962, Gilbert, 1981, Neubaum & Wolfner, 1999b, Tram & Wolfner, 1999, Bloch-Qazi & Wolfner, 2003). *D. melanogaster* have two types of SSOs; the seminal receptacle (SR) and paired spermathecae (ST). The SR stores the majority of the sperm (65-80%) and is the first SSO to store sperm as well as the first to release sperm (Gilbert, 1981, Neubaum & Wolfner, 1999c). The ST is considered to be the long term SSO of *Drosophila*. Sperm accumulate slower in the ST and are utilized after the sperm in the SR are depleted. Within 1.5 hours of mating ovulation begins and by 3 hours the female is laying fertilized eggs (Heifetz et al., 2000), indicating that the SR is actively storing and releasing sperm simultaneously. By 10 hours post-mating a noticeable decline is seen in the sperm stores of the SR, and by 48 hours the SR sperm stores are ~50% used, whereas the ST sperm stores are only ~15% depleted (Neubaum & Wolfner, 1999b).

It is unclear why *Drosophila* species require both the ST and the SR for sperm storage. It has been suggested that the SR is more efficient at storing sperm, but the
sperm within the SR are more susceptible to displacement by the sperm of other males (Pitnick et al., 1999, Civetta, 1999, Price et al., 1999). The ST may protect sperm from displacement. Secretory cells line the walls of the ST and secrete proteins into the lumen where sperm are stored; these proteins could function to keep sperm viable (Filosì & Perotti, 1975, Anderson, 1945, Pitnick et al., 1999).

Parallels between reproductive proteins of distantly related taxa have been identified, indicating that there is some degree of conservation. Genes encoding proteins have been identified in Dipteran species which closely resemble mammalian reproductive proteins both in sequence and putative function. A study by Allen and Spradling (2008) investigated proteins expressed within the ST in D. melanogaster. They discovered that products secreted by the ST are necessary for sperm maturation and function, paralleling the role of the mammalian epididymis and female reproductive tract. The gene hormone receptor 39 (Hr39) in Drosophila is a putative hormone receptor closely related to the steroidogenic factor 1 (Sf1) nuclear hormone receptor of mammals. Sf1 is active during embryogenesis and is necessary for proper reproductive tract development (Allen & Spradling, 2008). The Allen and Spradling study revealed a closer connection between Dipteran and mammalian reproductive biology than previously believed and sets the stage for use of D. melanogaster as a model for human reproduction.

The current study utilized a transcriptome analysis was conducted for each of the SSOs of D. melanogaster, before and after mating. The gene expression data from this project provides insight into the identity and expression patterns of genes expressed in the sperm storage organs. The information is potentially applicable to a wide range of species. This study identified genes differentially expressed within each SSO and
between the ST and SR of *Drosophila melanogaster* at two time points post-mating, providing the most detailed information on sperm storage gene expression patterns available for any species.

**MATERIALS and METHODS**

*Stock maintenance*

The Canton-S (CS-C) stock of *D. melanogaster* were used in the present study. Flies were reared in larval density controlled vials, on a standard *Drosophila* diet in a temperature controlled environment (25 degree Celsius) with a standard 12/12 light/dark cycle. Adults were collected as virgins. Virginity was confirmed by observing that no larvae appeared in the vials.

*Mating and sample collection*

Virgin females between four and seven days of adult age were used for virgin dissection of SSOs and for matings. For mated female collections, a single virgin female was placed with a single male. The female was separated from the male immediately after the observed mating had ended naturally and the timing of post-mating began. Tissues were collected from virgin, 3 hour post-mating, and 6 hour post-mating females. Females were dissected on a cold plate in RNase later (Ambion).

Approximately 80 females were used for each dissection event. The ST and SR were removed from each female and placed into microcentrifuge tubes containing TRIzol (Ambion) for RNA extraction. The ST sample included the spermathecae from each
female and the attached spermathecal duct. From each dissection period 2 samples were obtained, one of ~160 ST and the other of ~ 80 SRs.

**RNA extraction**

Freshly dissected tissues were ground in TRIzol and liquid nitrogen using mortar and pestle. RNA was extracted and the samples were then cleaned using Qiagen micro RNA clean-up columns and quantified using a NanoDrop (ND-1000) spectrophotometer (NanoDrop Technologies, Inc). To obtain sufficient quantities of RNA from ST and SR, the RNA for each sample was pooled from 10 independent dissection periods and extractions. Each ST sample used for one microarray chip contained the RNA from ~ 1600 ST (80 females X 2 ST per female X 10). Each SR sample used for one microarray chip contained the RNA from ~ 800 SR (80 females X 1 SR per female X 10).

**Microarray assays**

ST and SR microarray cRNA sample preparation used a two-cycle target and labeling kit (Affymetrix, Santa Clara, CA), which is designed for starting RNA levels of $\leq 100$ng for each microarray chip. For both the ST and SR 100ng of total RNA were used for amplification prior to cDNA synthesis followed by cRNA synthesis. The Gene Chip Drosophila genome 2.0 array (Affymetrix), which includes probe sets to 18880 *D. melanogaster* genes, was used to measure gene expression. Hybridization of cRNA to the Drosophila Gene Chips was performed by the Microarray Core Facility at the University of Nebraska-Lincoln. All microarray hybridizations were performed using three replicate chips for each treatment (virgin ST and SR, 3 hours post-mating ST and SR).
ST, 6 hours post-mating ST and SR) using 18 independent samples of RNA extracted from different sets of flies.

**Microarray analysis**

The Robust Multichip Averaging (RMA) procedure (Irizarry et al. 2003) implemented in the Affy module of the BioConductor microarray analysis package was used to background correct and normalize the microarray data. The resulting measurements of hybridization intensity were used as a proxy for gene expression levels. Average hybridization intensity was calculated for each gene using the three arrays for each sample. A cut-off value of 100 for average hybridization intensity was selected to help prevent the inclusion of genes which are not truly expressed in the sample. Using a cut-off value of 100 selected for approximately one-third of the genes.

To detect temporal patterns of female-expressed genes we performed pairwise comparisons of the hybridization data from virgins, 3 hour post-mated females, and 6 hour post-mated females using both SR and ST. To detect patterns of gene expression changes between organs, pairwise comparisons were made between the ST and SR at each time point. Differential expression of genes was determined using both fold change (FC), and false discovery rate (FDR). FDR is defined as the expected proportion of false positives among the declared significant results, and provides a more direct interpretation of microarray data in comparison with standard p-values (Pawitan et al., 2005, Benjamini & Hochberg, 1995) The cut-offs used for differentially expressed genes were FC ≥ 2, and a FDR q-value ≤ 0.1.
**Functional analysis**

The Database for Annotation, Visualization and Integrated Discovery (DAVID) (http://david.abcc.ncifcrf.gov) was used to classify gene function for the genes present in all lists using the most relevant gene ontology (GO) associated terms. The associated fourth level molecular function (MF) and biological process (BP) is listed for each of the expressed genes. DAVID was used to determine significant enrichment of known functional annotations within our differentially expressed gene lists. DAVID calculates the probability for representation of genes within a given category for an input list based on the observed representation number of genes, within the same category for a pre-defined background list. The predefined *D. melanogaster* background list consisted of GenBank accession numbers for each unique transcript probe on our arrays. All significantly (*p* < 0.05) enriched functional groups within all levels of biological process (BP) and molecular function (MF) was reported in the Supporting Information tables.

The Protein ANalysis THrough Evolutionary Relationships (PANTHER) (http://www.pantherdb.org) classification system was also used to classify genes into broad functional categories. Genes are classified into families and subfamilies using hidden Markov models (HMMs) based on shared functions, then further classified by molecular function and biological process ontology terms. PANTHER categorized the genes into more generalized categories compared to DAVID. The ontology of PANTHER uses a controlled vocabulary of molecular function and biological process arranged as directed acyclic graphs, similar to the Gene Ontology (GO), but abbreviated and simplified to facilitate high-throughput analysis. The PANTHER database contains annotation data for 14,115 *Drosophila melanogaster* genes.
RESULTS and DISCUSSION

In this study, we investigated gene expression changes in the SSOs of *D. melanogaster* at two time points (3 and 6 hours) after mating. A pre-mating (virgin) state was used as a baseline for the determination of gene expression patterns after mating. Three hour post-mating was used because this is the time needed for the complete entrance of sperm into storage, as well as the start of egg deposition and fertilization (Heifetz et al., 2000, Heifetz et al., 2001). Six hours post-mating was used based on a previous study which showed a spike in gene expression in the lower reproductive tract six hours after the completion of mating (Mack et al., 2006b). We found distinctly different patterns of gene expression for the two types of organs, and for pre- versus post-mating time points. As expected based on the fact that both of these organs store sperm, a substantial overlap in the genes expressed was also found. The distinct patterns produced for the two sperm-storage organs will be valuable for future functional research of sperm storage and studies on the role of SSOs in evolution.

Average expression

Genes having a hybridization intensity of greater than 100 after background correction were considered to be expressed. Average hybridization intensities as well as DAVID gene ontology (GO) categories for fourth level biological processes (BP) and molecular functions (MF) is listed for genes in the ST and SR at virgin, 3 hour and 6 hour time points (Supplementary table 1, 2). The highest expressed gene, based on hybridization intensity, at all three time points in the SR is a gene of unknown function.
Genes of unknown function can be interesting in terms of their potential to play organ-specific roles. Genes classified as unknown could play roles that are unique to sperm storage, or may represent new classes of genes which have not been identified in other, organs of *Drosophila* or other taxonomic groups.

At all three time points the gene in the ST with the highest expression was a serine protease (CG9897 in virgin and CG17239 at both 3 and 6 hours post-mating). Serine proteases have been identified as important molecules in reproduction, both in the seminal fluid and in the female reproductive tract (Ravi-Ram & Wolfner, 2007, Lung et al., 2002a, Lawniczak & Begun, 2004, Swanson et al., 2001a). The hybridization intensity of protease CG17239 remains highest in the ST 3 days after mating (Allen & Spradling, 2008), alluding to the importance of this particular protein for ST function. CG17239 is found on the long arm of chromosome 2, in a cluster of 4 proteases (CG17239, CG17012, CG17234, and CG17240), all of which have secretion signals (Prokupek et al.). ST proteins with secretion signals could directly interact with male seminal fluid proteins or sperm in the lumen of this organ. Three of the four proteases found in this cluster had high expression in the ST at all time points (Table 4.1), suggesting they might be necessary for storing sperm. The four proteases comprising the cluster were discovered in a previous EST study of the ST at three hours post-mating (Prokupek et al) (CG17240 was not found in the current microarray study), and found that these genes are enhanced for expression in the ST. The overlap in the ST of the genes of these two studies, as well as the low level of expression of these genes in the SR, suggests that the proteases in this cluster play roles important to the ST.
All three of the yolk proteins in *D. melanogaster* (YP1-YP3) were highly expressed in the ST at all time points, but only YP1 had high expression in the SR. In a study by Allen and Spradling (2008), YP1, 2 and 3 were found to be highly expressed in the spermathecae three days post-mating. Yolk proteins, found in tissues such as fat bodies and follicle cells, are an integral component of yolk in eggs (Brennan et al., 1982, Barnett et al., 1980). The presence of these proteins in the SSOs may indicate that these organs contribute to yolk production for eggs as suggested by Allen and Spradling (2008). Alternatively, YPs might play a novel role in the SSOs by protecting sperm from oxidative damage. In honeybees, the antioxidant activity of vitellogenins (similar to yolk proteins) apparently promotes individual survival (Seehuus et al., 2006), suggesting YPs might similarly promote the survival of sperm.

**Molecular function**

PANTHER was used to group the top 100 genes, based on hybridization intensity, into categories of molecular function. Proteases made up a high proportion (almost 10 percent in most cases) of the top expressed genes of the ST. Structural proteins, proteins involved in transport, muscular contraction and ion channels, synthetases and hydrolases were highly expressed in the SR at all time points. Perhaps transport and ion channel activities act to generate a suitable environment for stored sperm before and after mating. Both the ST and SR showed a high number of genes involved in nucleic acid binding (MF) and protein metabolism (BP); most of the nucleic acid binding genes encode ribosomal proteins (Supplementary figure 1).
**Differential gene expression**

Differential gene expression was measured by pairwise comparisons of virgins to the two time points post-mating for ST and SR. A larger number of genes were differentially expressed within the SR compared to the ST at both time points (Figure 4.1). In comparison with our study, far fewer genes were differentially expressed in the whole body of *D. melanogaster*, or lower reproductive tract after mating, compared to virgins (McGraw et al., 2004, Lawniczak & Begun, 2004). A possible explanation for the paucity of genes found to be differentially expressed in the whole body after mating is that opposing gene expression patterns in different organs negate each other (McGraw et al., 2004). The high number of differentially expressed genes in the SR compared to the ST (Figure 4.1) can possibly be explained by the fact that the SR is both actively storing and releasing sperm at 3 hours post-mating, and by 6 hours is presumably releasing sperm at a steady rate; the spermathecae is presumably only storing sperm at both of these time points.

**Overrepresented gene categories**

Up and down regulated genes were categorized using DAVID. Statistically overrepresented gene categories (\(p < .05\)) are presented for ST and SR in Supplementary table 6 and 7 respectively. Among overrepresented genes upregulated in the ST (at each time point), the functional categories that stand out include genes involved in metabolic pathways, such lipid, carbohydrate metabolism, and genes with catalytic activity such as juvenile hormone catabolism. Downregulated, overrepresented categories of genes of the
ST included metabolism and catabolism pathways, electron transport, and catalytic activity.

Upregulated functional categories of genes overrepresented in the SR included genes involved in immunity and defense, RNA modification, metabolism (peptidoglycan), female gamete generation, and catalytic activity. Downregulated overrepresented gene categories included metabolism (carbohydrate, glycogen), oxoreductase activity, and defense response. An overrepresentation of serine type peptidases in the SR was seen in upregulated genes at three hours ($p = 0.01$), and in down regulated genes at six hours ($p = 0.03$), suggesting a peak in proteolysis at the three hour time point.

The abundance of immunity/defense genes upregulated in the SR is not mirrored in the ST samples. This indicates that while the ST may be the primary SSO for protein modification (high level of proteases), the role of the SR may be more defense from introduced pathogens.

**Organ-specific changes in gene expression**

To detect patterns of gene expression, we compared hybridization data from the SR and ST. Genes were considered differentially expressed if the resulting fold change was $>2$, and $q$-value $< 0.1$. Comparing virgin SR and ST, a total of 1,794 genes were differentially expressed; 687 genes upregulated and 1,107 downregulated in the ST compared to the SR. For 3 hour post-mating SR and ST, a total of 1,589 genes were differentially expressed; 621 genes upregulated and 968 downregulated in the ST compared to the SR. At 6 hour post-mating SR and ST, a total of 2,538 genes were
differentially expressed; 1,054 genes upregulated and 1,484 downregulated in the ST compared to the SR (Supplementary table 5).

Analysis of the data in the preceding paragraph revealed interesting patterns. Overrepresented categories of genes, those statistically enriched compared to a random expectation, upregulated in the ST in comparison to the SR in virgin and both times post-mating include metabolism, defense/immunity, transport, binding, and proteolysis. Genes upregulated in the SR (downregulated in the ST by comparison) fell into the following categories: cell communication, ion transport, development, defense/immunity (different genes from those relatively upregulated in the ST), localization, and homeostasis (Supplementary Figure 4.2). Apparently, there are few defense/immunity genes held in common in the two different sperm storage structures. The proteins involved in cell communication, ion transport, and localization within the SR may be facilitating the proper timing of release of sperm from this storage organ. A complete list of overrepresented categories is presented in Supplementary table 8. The large number of differentially expressed genes is surprising, as one might assume that the organs have fundamentally similar functions. At each time point, the organs may have very different functions; the SR both storing and releasing sperm, and the ST preparing sperm for long term storage.

**General discussion**

During the time of mating male *Drosophila* pass to females a number of proteins via their seminal fluid. Some of these proteins are transferred to the SSOs where they may elicit responses that are common to both types of organ. These proteins, especially...
male accessory gland proteins (Acps), are responsible for eliciting a suite of post-mating responses in the female. Female responses include increases in ovulation, ovipositioning, and mating refractoriness, as well as sperm storage and decreases in female lifespan (Wolfner, 1997, Wolfner, 2002). At least one Acp (Acp36DE) has been associated with proper sperm storage. Acp36DE associates with the sperm mass and localizes at the openings of the SSOs as well as within them (Bertram et al., 1996, Neubaum & Wolfner, 1999b). Another Acp, sex peptide (SP) enters the SSOs bound to the sperm tail, and is released over time. This protein stimulates juvenile hormone synthesis. (Liu & Kubli, 2003, Chapman et al., 2003a). There may be a connection between juvenile hormone catabolism in the ST and juvenile hormone synthesis induced by SP. Specifically, females may be acting to counter a detrimental physiological effect induced by males; juvenile hormone increases eggs production which shortens the life span of females.

Female genes found to be differentially expressed in the present study might directly interact with Acps. For example, the cleavage of SP is accomplished by a trypsin member of the serine protease family (Peng et al., 2005b). Most Acps are no longer detected after 6 hours postmating, but SP continues to be cleaved from sperm for the entire time sperm are in storage (Peng et al., 2005b). The identity of the specific serine protease involved in the release of SP is not known, but it is quite possibly female derived. Four of the seven Acps which enter the SSOs are serine protease inhibitors (serpins) (Mueller et al., 2005b). Serpins bind to serine proteases, blocking their proteolytic function, and they might interact with proteases in the ST. Proteases found within SSOs may also be interacting with male proteins in cascades, similar to those seen
in immune function and blood or sperm coagulation (Overduin & de Beer, 2000, Malm et al., 2000, Kim et al., 2008).

Although ST are found in a wide range of invertebrate and vertebrate taxa (Eberhard, 1996) the function of proteins and other macromolecules associated with this organ are scarcely understood. An exception is social insects - bees and ants - in which the function of ST proteins have been best studied. Reproductively capable females (queens) mate with several males early in life and rely on stored sperm to fertilize eggs over many years. The reproductive success of the queen is directly dependent on quality of sperm in storage, which creates strong selection for both sperm viability as well as efficient utilization of sperm stores (Baer et al., 2006). In the social ant (Crematogaster opuntiae), as well as in honey bees, the secretory cells surrounding the ST have ample glycogen, serving as a possible energy reserve for sperm in storage (Wheeler & Krutzsch, 1994). In honeybees the ST fluid contains sugars including glucose, trehalose and fructose, and a high level of trehalase activity (Alumot et al., 1969b). Enzymes with antioxidant activity have been found in the ST of honeybees, presumably acting to protect stored sperm from oxidative damage (Weirich et al., 2002, Collins et al., 2004b). Perhaps YPs found in the SSOs of D. melanogaster play a similar antioxidative function.

This study found that differentially expressed (up and down) genes in diverse metabolic pathways were overrepresented. The differential expression of genes related to metabolism may correspond to the changing nutritional/environmental needs of sperm as they are going through steps of sperm storage and release. Genes for carbohydrate/lipid metabolism are expressed at higher levels in the ST when compared to the SR at all time points, indicating a role in long term sperm storage (Supplementary Figure 4.1).
McGraw et al. (2004) used microarrays to compare gene expression of *D. melanogaster* females which were virgin, mated to wild-type males, mated to spermless males, or mated to males lacking Acps. A genome-wide comparative microarray analysis was also done on virgin, courted, and two-hour post-mated females (Lawniczak & Begun, 2004). In both studies, immune-related genes and serine proteases were found to be affected by mating. The number of serine proteases influenced by mating was determined to be higher than expected by chance (Lawniczak & Begun, 2004). The role of such proteases within the female is unknown. Induction of both serine proteases and immune/defense related genes was observed in the present study. Comparisons between the SR and the ST showed similar numbers of immune/defense genes, but higher levels of proteases in the ST.

**CONCLUSIONS**

Microarrays were used to investigate the transcriptome of *Drosophila melanogaster* sperm storage organs in virgin females and at two time points post-mating. A high number of SSO genes were differentially expressed following mating. At each time point high numbers of genes were differentially expressed between the two sperm storage organs, indicating that each SSO plays a unique role in the process of sperm storage. The spermathecae was enriched for genes involved in proteolysis and metabolism. These genes could be interacting with male proteins such as protease inhibitors or encode proteins that provide lipids/carbohydrates for sperm maintenance. Other spermatheca proteins, such as yolk proteins, could protect sperm perhaps by acting as an antioxidant. The seminal receptacle exhibited a number of overrepresented genes.
involved in localization, signaling, and ion transport. These genes could be working to
maintain a homeostatic environment, as well as serving communication roles between the
SR and oocytes to ensure the proper timing of fertilization. More genes were
differentially expressed in the seminal receptacle perhaps due to the dual role of sperm
storage and sperm release played by this organ at the time points used in the present
study. Further investigation of the roles of proteins in these two organs will allow for a
more complete understanding of genes playing essential roles in the process of sperm
storage, maintenance, and use.

ACKNOWLEDGEMENTS

This research was supported by a National Science Foundation grant (DEB-ESP0346476).
REFERENCES


Bork, P. & Beckman, G. 1993. The CUB domain. A widespread module in

are a major site of vitellogenin synthesis in *Drosophila melanogaster*. *Dev Biol* 89: 225-36.


& Partridge, L. 2003b. The sex peptide of *Drosophila melanogaster*: female post-

mating in *Drosophila melanogaster* females is mediated by male accessory gland


Civetta, A. 1999. Direct visualization of sperm competition and sperm storage in
*Drosophila*. *Current Biology* 9: 841-44.

Civetta, A. & Clark, A. G. 2000. Correlated effects of sperm competition and postmating


Prokupek 2008


Prokupek 2008

Princeton University Press, New Jersey.


Prokupek 2008
Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila.

*Genetics* **177**: 1321-35.


Prokupek 2008


Moshitzky, P., Fleischmann, I., Chaimov, N., Saudan, P., Klauzer, S., Kubli, E. & Applebaum, S. W. 1996. Sex-peptide activates juvenile hormone biosynthesis in


Prokupek 2008


Prokupek 2008


Figure 4.1: Differential regulation of genes expressed in the spermathecae and seminal receptacle at two time points post-mating. Three comparisons are made in this graph for each organ 3h – V, 6h – V and 6h – 3hr.
3h - V: difference in expression between in the 3 hour time point compared to virgin. 6h – V: difference in expression between the 6 hour time point compared to virgin. 6h – 3h: difference in expression between the 6 hour time point and 3 hour time point. up: number genes with higher expression in the first comparative category; down: number of genes with lower expression in the first comparative category. For example in 3h-V :up refers to the genes that have higher expression at 3 hours compared to virgin; down refers to genes that have lower expression at 3 hours compared to virgin; total refers to the total number of genes differentially expressed between 3 hours and virgin.
Table 4.1. Rank expression of serine proteases and yolk proteins expressed in the spermathecae and seminal receptacle
Proteases and yolk proteins ranked by average hybridization intensity. Genes are listed which are in the top 100 intensities in at least one experimental category (note that the rank number can be much lower than 100 in other categories). Numbers indicate their ranking by hybridization intensity. ST = spermatheca, SR = seminal receptacle. *The ranking of the expression level 3 days post-mating as reported in Allen and Spradling, 2008.

<table>
<thead>
<tr>
<th>Gene symbol</th>
<th>ST V</th>
<th>ST 3</th>
<th>ST 6</th>
<th>ST 3*</th>
<th>SR V</th>
<th>SR 3</th>
<th>SR 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proteases</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CG17239</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>553</td>
<td>53</td>
<td>1838</td>
</tr>
<tr>
<td>CG32834</td>
<td>20</td>
<td>26</td>
<td>12</td>
<td>6</td>
<td>2300</td>
<td>1833</td>
<td>6483</td>
</tr>
<tr>
<td>CG31681</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>7</td>
<td>354</td>
<td>100</td>
<td>1433</td>
</tr>
<tr>
<td>CG32277</td>
<td>27</td>
<td>35</td>
<td>34</td>
<td>11</td>
<td>2027</td>
<td>1026</td>
<td>5507</td>
</tr>
<tr>
<td>CG17012</td>
<td>5</td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>448</td>
<td>161</td>
<td>1930</td>
</tr>
<tr>
<td>CG18125</td>
<td>137</td>
<td>5</td>
<td>9</td>
<td>18</td>
<td>1204</td>
<td>104</td>
<td>2889</td>
</tr>
<tr>
<td>CG9897</td>
<td>1</td>
<td>14</td>
<td>11</td>
<td>19</td>
<td>1552</td>
<td>1095</td>
<td>5793</td>
</tr>
<tr>
<td>CG17234</td>
<td>73</td>
<td>3</td>
<td>3</td>
<td>25</td>
<td>747</td>
<td>28</td>
<td>1143</td>
</tr>
<tr>
<td>CG30371</td>
<td>36</td>
<td>10</td>
<td>16</td>
<td></td>
<td>1144</td>
<td>287</td>
<td>2507</td>
</tr>
<tr>
<td>CG10469</td>
<td>89</td>
<td>138</td>
<td>93</td>
<td></td>
<td>104</td>
<td>297</td>
<td>214</td>
</tr>
<tr>
<td>CG13318</td>
<td>508</td>
<td>590</td>
<td>459</td>
<td></td>
<td>50</td>
<td>15</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Yolk Proteins</th>
<th>YP1</th>
<th>YP2</th>
<th>YP3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>7</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>41</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>14</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>256</td>
<td>286</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>119</td>
<td>172</td>
</tr>
<tr>
<td></td>
<td>44</td>
<td>121</td>
<td>179</td>
</tr>
</tbody>
</table>
Supplementary Figure 4.1: Classification of the top 100 genes based on average hybridization intensity in the ST and SR. Classification of genes by biological process and molecular function was independently performed.
Supplementary Figure 4.2: PANTHER classification of genes upregulated in the SR and ST. Upregulation refers to a higher expression in one SSO organ when compared to the other SSO at each time point. Data is represented as the percentage of genes in each category. Classification of genes by biological process and molecular function was independently performed.
<table>
<thead>
<tr>
<th>Biological process</th>
<th>Molecular Function</th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal transduction</td>
<td>Receptor</td>
</tr>
<tr>
<td>Protein metabolism</td>
<td>Transcription factor</td>
</tr>
<tr>
<td>Developmental processes</td>
<td>Oxidoreductase</td>
</tr>
<tr>
<td>Transport</td>
<td>Hydrolase</td>
</tr>
<tr>
<td>Lipid metabolism</td>
<td>Cytoskeletal</td>
</tr>
<tr>
<td>Immunity and defense</td>
<td>Nucleic acid binding</td>
</tr>
<tr>
<td>Carbohydrate metabolism</td>
<td>Protease</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>Transferase</td>
</tr>
<tr>
<td>Cell proliferation</td>
<td>Kinase</td>
</tr>
<tr>
<td>Muscle contraction</td>
<td>Calcium binding protein</td>
</tr>
<tr>
<td>Intracellular protein traffic</td>
<td>Ion channel</td>
</tr>
<tr>
<td>Electron transport</td>
<td>Cell adhesion</td>
</tr>
<tr>
<td>Amino acid metabolism</td>
<td>Extracellular matrix</td>
</tr>
<tr>
<td>Homeostasis</td>
<td>Lyase</td>
</tr>
<tr>
<td>Sulfur metabolism</td>
<td>Synthase and synthetase</td>
</tr>
<tr>
<td>Protein targeting</td>
<td>Membrane traffic protein</td>
</tr>
<tr>
<td></td>
<td>Ligase</td>
</tr>
<tr>
<td></td>
<td>Isomerase</td>
</tr>
</tbody>
</table>
Supplementary Table 4.1: Average expression (hybridization intensity) of genes in the spermathecae (transcript level > 100). Genes are classified by DAVID into level 4 biological process and level 4 molecular function gene ontology categories.
**Supplementary Table 4.2:** Average expression (hybridization intensity) of genes in the seminal receptacle (transcript level > 100). Genes are classified by DAVID into level 4 biological process and level 4 molecular function gene ontology categories.
Supplementary Table 4.3: Differential expression of spermathecae genes at two time points (3 and 6 hours) post-mating.
Supplementary Table 4.4: Differential expression of seminal receptacle genes at two time points (3 and 6 hours) post-mating.
Supplementary Table 4.5: Genes differentially expressed in the spermathecae compared to the seminal receptacle in virgin, 3h and 6h post-mating samples. In this comparison, upregulated (up) refers to genes which are upregulated in the spermathecae in comparison to the seminal receptacle. Downregulated (down) refers to genes which are downregulated (down) in the spermathecae compared to the seminal receptacle.
Supplementary Table 4.6: DAVID analysis of over-represented gene categories of differentially expressed genes in the spermathecae at two time points (3 and 6 hours) post-mating.
Supplementary Table 4.7: DAVID analysis of over-represented gene categories of differentially expressed genes in the seminal receptacle at two time points (3 and 6 hours) post-mating.
**Supplementary Table 4.8:** DAVID analysis of over-represented categories of differentially expressed genes from a comparison of the spermathecae and seminal receptacle samples at virgin, 3 and 6 hour post-mating time points.