doi:10.1111/j.1558-5646.2008.00493.x

AN EVOLUTIONARY EXPRESSED SEQUENCE TAG ANALYSIS OF DROSOPHILA SPERMATHECA GENES

Adrianne Prokupek,[1,2] Federico Hoffmann,[1,3] Seong-il Eyun,[1] Etsuko Moriyama,[1] Min Zhou,[1,4] and Lawrence Harshman[1,5]

[1]*School of Biological Sciences, University of Nebraska, Lincoln, Nebraska 68588*

[2]*E-mail: aprokup1@bigred.unl.edu*

[4]*Department of Food Science and Technology, Shanghai Jiao Tong University, Shanghai, China*

[5]*E-mail: lharsh@unlserve.unl.edu*

Received September 10, 2007
Accepted July 10, 2008

This study investigates genes enriched for expression in the spermatheca, the long-term sperm storage organ (SSO) of female *Drosophila*. SSO genes are likely to play an important role in processes of sexual selection such as sperm competition and cryptic female choice. Although there is keen interest in the mechanisms of sexual selection at the molecular level, very little is known about the female genes that are involved. In the present study, a high proportion of genes enriched for expression in the spermatheca are evolving rapidly. Most of the rapidly evolving genes are proteases and genes of unknown function that could play a specialized role in the spermatheca. A high percentage of the rapidly evolving genes have secretion signals and thus could encode proteins that directly interact with ejaculate proteins and coevolve with them. In addition to identifying rapidly evolving genes, the present study documents categories of genes that could play a role in spermatheca function such as storing, maintaining, and utilizing sperm. In general, candidate genes discovered in this study could play a key role in sperm competition, cryptic female choice of sperm, and sexually antagonistic coevolution, and ultimately speciation.

KEY WORDS: **Coevolution, molecular evolution, sexual selection, speciation, sperm storage.**

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 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 of deficit in these studies is in the identification and classification of female reproductive proteins, especially those proteins interacting directly with male seminal products (e.g., sperm storage proteins). Important female reproductive proteins are likely to be found in organs dedicated to the storage of sperm. Sperm storage organs (SSOs) are found in the females of a variety of animal taxa and function in the retention, maintenance, and use of sperm after mating has occurred. The present study focuses on one of the two types of SSOs present in *Drosophila*, the spermathecae, which is the long-term SSO.

Although spermathecae are commonly found in a 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 (Wheeler and Krutzsch 1994; Weirich et al. 2002; Collins et al. 2004; Collins et al. 2006). Reproductive proteins associated with female SSOs (such as spermathecae) are candidates to play important roles in evolutionary phenomena such as

[3]Current address: Istituto Carlos Chagas—ICC—Fiocruz, Rua Prof. Algacyr Munhos Mader 3775-CIC, 81350-010, Curitiba, Brazil.

sperm competition, female sperm choice, sexually antagonistic coevolution, and consequently in speciation. In spite of their potential importance, little is known about how they function and even less about the evolution of genes associated with these organs. Identification of female proteins expressed within the spermathecae 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 *Drosophila melanogaster* second male sperm precedence (P2) is due, in part, to a nonsperm component of the ejaculate (Harshman and Prout 1994). It is now established that male accessory gland proteins play a role in sperm competition in *Drosophila* (Ravi-Ram and Wolfner 2007). Allelic variation in male accessory gland protein genes has been associated with the differential fertilization success of both the first male to mate with a female and the second male to mate with the female (Clark et al. 1995; Fiumera et al. 2005, 2007). Genetic studies reveal that female processes have a major effect on the outcome of sperm competition in *D. melanogaster* (Clark and Begun 1998; Clark et al. 1999), but the specific female genes that have these effects are not known.

Studies of conspecific sperm precedence have provided clues about the importance of male and female reproductive proteins in sperm competition (Coyne and Orr 2004). 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). 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 1999; Howard et al. 2008). 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 displaced from the female SSOs by the seminal fluid of the second male to mate. Thus, sperm precedence may be due to interactions between male seminal fluid and female 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 could play a major role in conspecific sperm precedence, or preclude fertilization by heterospecific sperm.

Rapid evolution of reproductive proteins has been documented in protistans, fungi, plants, and animals (Clark et al. 2006) and in both male and female gametic proteins (Swanson and Vacquier 2002; Galindo et al. 2003). For example, in sea urchins male sperm bindin evolves rapidly (Palumbi 1999) as does the bindin receptor 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, 2003). Rapidly evolving *Drosophila* male accessory gland proteins (Acps) have been a focus of molecular population genetic and molecular evolution studies. The average rate of sequence divergence of *D. melanogaster* Acps is approximately twice that of nonreproductive proteins (Begun et al. 2000; Swanson et al. 2001a; Wagstaff and Begun 2004; Mueller et al. 2005). 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. 2004) 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. The present study is the first molecular evolutionary study of genes sampled from a specific 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 and Jonsson 1962; Neubaum and Wolfner 1999), whereas a pair of spermathecae are the site of long-term storage. Sperm are stored in the spermathecal lumen, which receives proteins of unknown function from surrounding secretory epithelial cells (Filosi and 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 and Pitnick 2002, 2003). One rationale for investigating genes in the spermatheca is that rapidly evolving genes in this SSO might coevolve with rapidly evolving *Drosophila* Acps.

The present study is an evolutionary EST investigation of genes enriched for expression in *Drosophila* spermathecae. The focus is on spermathecae because they secrete proteins into the sperm storage lumen that 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.

Methods and Materials

cDNA LIBRARY PREPARATION AND DNA SEQUENCE GENERATION

RNA was isolated from both 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 h after mating was observed. Total RNA was isolated from spermathecae

using the TRIzol reagent (Invitrogen, Carlsbad, CA). 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 (Simple Modular Architecture Research Tool) approach (Zhu et al. 2001). Subtractive hybridization was performed and a cDNA library was generated (Evrogen, Moscow, Russia) using the suppressive subtraction hybridization (SSH) method in both directions (tester vs. driver and driver vs. tester) (Diatchenko et al. 1996, 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, Piscataway, NJ). Three hundred and eighty-three DNA sequences were generated using the MegaBACE 400 automated DNA sequencer (Amersham Biosciences). Vector sequences were masked using the CAP3 program (Huang and Madan 1999).

IDENTIFICATION OF GENES EXPRESSED IN SPERMATHECAE

To ensure that we found all the genes represented by the *D. simulans* spermathecae EST (expressed sequence tags) sequences, we queried both the *D. simulans* and *D. melanogaster* genomes as the latter is more complete. 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*. We excluded those sequences with similarities of 80% or lower, and those with expected (E) values greater than 0.01 (139 sequences were excluded).

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*, and *D. pseudoobscura*). The Comparative Analysis Freeze 1 (CAF1) genomic sequences of the *Drosophila* species were downloaded from the AAAWiki website (http://rana.lbl.gov/drosophila) of 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 a 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.

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 (d_N) and synonymous (d_S) changes to the patterns of nucleotide variation was compared using the codon-based maximum-likelihood framework described by Goldman and Yang (1994) implemented in PAML version 3.15 (Yang et al. 2000). A pairwise comparison was performed between *D. simulans* and *D. melanogaster.* The likelihood of d_N being higher than d_S was evaluated by comparing a model in which d_N and d_S were estimated as free parameters (L_1) to a model in which d_N equals d_S (L_0). The two models were compared in a likelihood-ratio test with one degree of freedom. Historically, d_N/d_S ratios that exceeded 1.0 were considered to be indicative of positive selection, but recent studies, such as Swanson et al. (2004), argue that this ratio is conservative, especially for the identification of candidate genes. Lowering the d_N/d_S ratio to 0.5 was found to be reasonable for the identification of candidate genes to undergo further investigation into the forces of selection (Swanson et al. 2004). The d_N/d_S ratios were also estimated by using the branch model of PAML. This model allows the d_N/d_S ratios to vary among branches in a given phylogeny and is useful in detecting positive selection acting on particular lineages (Yang 1998).

Variation in the d_N/d_S ratio among sites was also explored using the tree-based models described by 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. 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 inclusion of *D. pseudoobscura* could lead to an overestimation of d_N/d_S due to a saturation of d_S obtained between *D. pseudoobscura* and other species. Therefore, the analysis was done both including and excluding *D. pseudoobscura*. Additionally, to ensure that possible incomplete lineage sorting reported in the melanogaster subgroup (Pollard et al. 2006; Wong et al. 2008) did not affect the outcome of our analysis, the analysis was done using three six-species trees varying the phylogenetic placement of *D. erecta* and *D. yakuba*. The assumptions of the models and test statistics are briefly described in Results, for a full description see Yang et al. (2000).

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 (Tusnády and 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 to identify signal peptides from N-terminal regions, often misidentified as a TM region by these 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

All genes were subject to conserved domain searches by CD-Search at National Center for Biotechnology Information (Marchler-Bauer and Bryant 2004). Function was inferred from a combination of information gained from the conserved domain searches, FlyBase classification, Gene Ontology database classification, and literature searches.

Results

CODING SEQUENCES IN THE cDNA LIBRARY

Of the 383 EST library clone sequences, 244 matched coding sequences (CDSs) in the *D. simulans* genome representing 44 unique CDSs. The remaining 139 sequences were short with typically less than 20 base pairs matched against any *D. melanogaster* or *D. simulans* CDSs, and had E values greater than 0.01 (accepted sequences had an average E value of 4.5×10^{-6}). These sequences were excluded from further analysis because they did not meet the criteria described in Materials and Methods.

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 identified from the library, 29 had identifiable orthologous genes in all seven species (31 genes had identifiable orthologs in all six species when excluding *D. pseudoobscura*). For the remaining 15 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*. Orthologs to these genes were found in other *Drosophila* species (dsim_15604 was found in *D. erecta* and *D. yakuba*; dsim_6594 was found in *D. sechellia*, *D. erecta*, and *D. yakuba*), but were not used for molecular evolutionary analysis. 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 1, Supporting Table 1). The most likely function was determined by the Gene Ontology database, conserved domains, and relevant literature. Active serine proteases are indicated by the presence of three residues (Ser195, Asp102, and His57) termed the catalytic triad. Identification of residues of the catalytic triad was done using the SMART (Schultz et al. 1998; Letunic et al. 2006). Of the 42 genes, 11 are putative serine proteases 10 of which have the catalytic triad of amino acids. The substrate specificity of serine proteases is due to residues surrounding Ser195 (Perona and Craik 1995; Hedstrom 2002), these areas were identified in the proteases of this study. Positive selection was predicted to be operating in regions adjacent to the substrate specificity sites (Fig. 1; Supporting Figure S1). 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 4.5×10^{-9}) overrepresented in the spermathecae (26%) when compared to the percentage of such genes in the *D. melanogaster* genome (5%)

Table 1. Functional annotation of 42 genes enriched for expression in the spermatheca

Function[1]	Number[2]	SP[3]	TM[4]
Serine protease	11	11	
Cell communication	3		2
Peptidase	3	2	2
Translation	2		
Actin formation/biosynthesis	2		
Amino acid transport	1	1	1
Antimicrobial	1	1	
Apoptosis	1		1
Cation transport	1		1
Dehydrogenase	1	1	
Helicase	1		
Juvenile hormone catabolism	1	1	1
Nerve signaling	1		1
Phospholipid metabolism	1	1	1
Secondary metabolism	1		
Protein–protein interaction	1		
Sugar metabolism	1	1	
Unknown	9	4	3

[1]Predicted function of encoded proteins.
[2]Number of genes.
[3]Number of genes predicted to encode proteins that have secretion signal peptides.
[4]Number of genes predicted to encode proteins with transmembrane regions.

(Ross et al. 2003). We do not assume that an exhaustive sampling of the spermathecae library was performed; nevertheless, there is no reason to assume a bias in the choice of clones for sequencing. The relative proportions of genes in the categories listed are expected to be similar to the actual proportions. One of the genes (dsim_GLEANR_6594) discovered in the library, but having no *D. melanogaster* ortholog, is also predicted to be a serine protease.

EVOLUTIONARY ANALYSES

Nonsynonymous 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 1997, 2007). The average d_N/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 of the 42 genes (Fig. 2) have d_N/d_S higher than the 0.5 threshold adopted by Swanson et al. (2004). Results of the branch-model PAML analysis supported the *D. melanogaster/D. simulans/D. sechellia* lines subject to positive selection for five of the 10 genes identified above (Supporting Table S2). Only one gene (CG15098) constantly showed positive selection both in the *D. melanogaster* and *D. sechellia* lineages regardless of the placement of *D. yakuba* and *D. erecta*. On the other hand, six genes (including five not identified by pairwise comparisons) showed possible positive selection in the *D. simulans / D. sechellia* lines.

Another analysis compared the fit of the data to different models of codon evolution (Yang and Nielsen 2000). The "site-model" analysis in PAML was used to explore heterogeneity in d_N/d_S along the gene, and to test for positive Darwinian selection (Table 2, Supporting Table 1). These comparisons were restricted to the 40 genes for which sequences were available from at least four species. The first comparison examined the fit of data to the one-ratio model (M0) against the model that classifies sites into three classes (M3). For 37 of the 40 genes, the fit of data to M3

Figure 1. Serine proteases identified by PAML as having positively selected codons flanking the serine active site (bold and underlined). Amino acid residues in bold are integral for serine protease substrate specificity. * indicates amino acid residues that were significant for positive selection using PAML. See Supporting Table S3 for a complete list the serine protease active site residues of the serine proteases discovered in the present study.

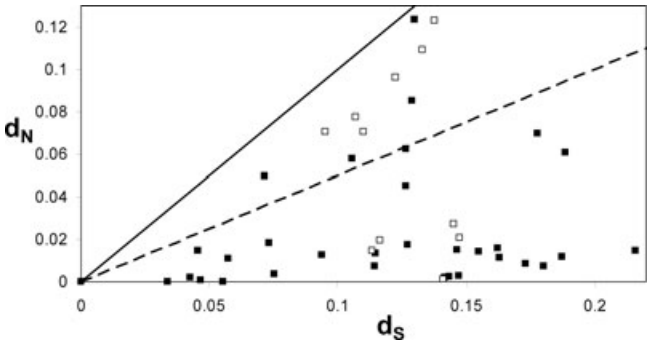

Figure 2. The number of nonsynonymous substitutions per site (d_N) plotted against the number of synonymous substitutions per site (d_S) for the *D. simulans* spermathecae EST library. The solid line represents d_N/d_S of 1.0, the historical threshold for positive selection. The dashed line represents d_N/d_S of 0.5, the threshold for the identification of candidate gene for positive selection. All squares on the graph represent genes found in the spermathecae EST library and open squares correspond to the 11 serine proteases.

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 (nearly neutral) and M7 (beta), with alternative models that allow a class of sites to have $d_N/d_S > 1$, M2a (positive selection) and M8 (beta and omega) (Yang and Nielsen 2002; Yang and Swanson 2002). In the first test, the null model (M1a) assumes two site classes, the first with d_N/d_S close to 0, and the second with $d_N/d_S = 1$; this is compared with the alternative model (M2a) that 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 $0 \leq d_N/d_S \geq 1$, 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 and Nielsen 2002; Yang and Swanson 2002). Comparisons of M7 to M8 and M1 to M2 provided evidence for positive selection in 17 of 40 genes. Genes having elevated d_N/d_S from the pairwise comparison, and/or support for positive selection from PAML models using trees excluding *D. pseudoobscura*, are listed in Table 2. All values are listed in Supporting Table S1, including the results of analysis including *D. pseudoobscura*.

Results of the branch model analysis show seven genes in the *D. simulans*/*D. sechellia* lines subject to positive selection. Only one gene (CG15098) constantly showed positive selection in all three topologies in the *D. melanogaster* lineage. All data from PAML branch model analysis are reported in Supporting Table S2.

To obtain extended taxonomic insight into the evolution of spermathecal genes, the presence or absence of homologous genes was examined 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)* (Supporting Table S1). 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).

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, 23 (55%) are predicted to have signal peptides and 13 (31%) are predicted to have transmembrane regions (Table 1). Three of the genes predicted to have transmembrane regions and 11 of the genes predicted to have signal peptides show evidence of positive selection (Supporting Table S1). 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. 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 was characterized. A high proportion of spermatheca genes are predicted to encode serine proteases many of which 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 high percentage of the genes exhibit the molecular signal of positive selection. Insight into the function of the spermatheca was obtained from the identity of genes expressed in this SSO.

Table 2. Genes with elevated pairwise d_N/d_S (>0.5) and genes identified as having regions that provide evidence for evolution by positive selection (PAML analysis)

				M0 vs. M3		M1 vs. M2		M7 vs. M8		
Genes	Function[1]	d_N/d_S[2]	Species[3]	p_s[4]	d_N/d_S[5]	p_s[6]	d_N/d_S[7]	p_s[8]	d_N/d_S[9]	pos. sel[10]
CG8331	cell comm.	0.19	M,S,Sc,Y,E,A	0.01	30.56[14]	0.08	1.00	0.01	30.61[12]	1
CG10650	peptidase	0.55	M,S,Sc,Y,A	0.11	2.94[14]	0.03	4.93	0.06	3.67[12]	
CG32702	protein int.	0.14	M,S,Sc,Y,E	0.02	2.55[14]	0.00	6.09	0.01	3.58[13]	
CG3066	ser. protease	0.17	M,S,Sc,Y,E,A	0.03	2.56[14]	0.04	1.00	0.02	2.83[12]	
ηTry	ser. protease	0.14	M,S,Sc,Y,E,A	0.03	3.02[14]	0.01	3.74	0.03	3.26[12]	3
Ser12	ser. protease	0.64	M,S,Y,E	0.09	6.27[14]	0.09	6.25[14]	0.09	6.25[14]	9
CG17012	ser. protease	0.89	M,S,Sc,A	0.07	6.08[14]	0.05	8.21[14]	0.06	6.96[14]	
CG17234	ser. protease	0.74	M,S,Sc,Y,E	0.05	7.23[14]	0.09	5.25[14]	0.10	4.86[14]	
CG17239	ser. protease	0.73	M,S,Sc,Y,E,A	0.16	2.71[14]	0.09	3.41[14]	0.13	2.94[14]	6
CG18125	ser. protease	0.79	N/A							
CG31681	ser. protease	0.82	M,S,Y,E	0.09	5.57[14]	0.08	6.05[14]	0.09	5.84[14]	6
Treh	sugar metab.	0.12	M,S,Sc,Y,E,A	0.04	3.36[14]	0.02	5.51[14]	0.03	3.75[14]	5
Ef1α48D	translation	0.02	M,S,Sc,Y,E,A	0.01	2.11[14]	0.01	2.11	0.01	2.11[12]	4
Qm	translation	0.02	M,S,Sc,Y,E,A	0.005	3.25[14]	0.01	3.25	0.01	3.25[13]	1
CG2233	unknown	0.95	M,S,Sc,Y,E	0.02	33.31[14]	0.18	4.16[14]	0.17	4.37[14]	
CG11137	unknown	0.1	M,S,Sc,Y,E,A	0.01	1.80[12]	0.00	25.23	0.01	1.80[12]	1
CG15098	unknown	0.69	M,S,Sc,Y,E,A	0.09	2.77[14]	0.04	3.66	0.07	3.00[12]	1
CG30197	unknown	0.05	M,S,Sc,Y,E	0.02	4.37[13]	0.02	4.37	0.02	4.37[12]	
CG31686	unknown	0.66	N/A							

[1]Predicted protein function.

[2]d_N/d_S: pairwise comparison of *D. melanogaster* and *D. simulans* sequences, estimated assuming no rate heterogeneity.

[3]Species: 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*.

N/A indicates that PAML analysis was not done due to too few species. The following statistics are all derived from PAML analysis.

[4]p_S: the proportion of sites estimated to belong to the class that has the highest d_N/d_S in M3.

[5]d_N/d_S: for the highest class in M3.

[6]p_S: the proportion of sites estimated to belong to the class that has $d_N/d_S>1$ in M2.

[7]d_N/d_S: the estimate for the class with the ratio>1 in M2.

[8]p_S: the proportion of sites estimated to belong to the class that has $d_N/d_S>1$ in M8.

[9]d_N/d_S: the estimate for the class with the ratio>1 in M8.

[10]pos.sel.: the number of codons significantly ($P<0.05$) recognized by PAML as being positively selected; [12]$P<.05$; [13]$P<.01$; [14]$P<.001$.

In the present study, 44 unique *D. simulans* genes were identified in the hybrid-selected cDNA library. Orthologs were discovered in *D. melanogaster* for 42 of the 44 genes. 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. (DiBenedetto et al. 1987; Monsma and Wolfner 1988; Wolfner et al. 1997; Swanson et al. 2001a; Andres et al. 2006; Davies and 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 $d_N/d_S > 0.5$ (Fig. 1). Two especially relevant previous studies (Swanson et al. 2001a, 2004) have made pairwise comparisons of *D. simulans* and *D. melanogaster*. The incidence of genes with $d_N/d_S > 0.5$ in the spermatheca (present study) is at least twice as high as 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. 2004) 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). The number of genes in the present study that overlapped with the most similar study (Swanson et al. 2004) was only five, even though their study included spermathecae in the mix of tissue investigated. The small overlap between Swanson (2004) and the present study could be attributed to the small proportion of tissue mass contributed by the spermathecae to the lower reproductive tract. Genes expressed in the spermathecae might be represented in such low levels compared to genes from larger tissues that they were undetected in the lower reproductive tract cDNA library. Alternately, a number of technical differences could be responsible for the small overlap in genes between these two studies, including the processes used

to screen and select clones or the use of females of different ages. 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 five serine proteases ($d_N/d_S = 0.72 - 0.89$), and two more genes of unknown function ($d_N/d_S = 0.64 - 0.66$). 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 evolution analyses showed that 17 of 40 (42.5 %) spermatheca genes contain at least one region that conforms to a model of positive selection (Table 2). It has been suggested that tests, such as PAML, which involved the fitting of a distribution of substitution rates across sites as a method for inferring individually evolving sites may be prone to type I (Suzuki and Nei 2004) or type II Kosakovsy and Frost 2005) errors. This is especially a problem when the test species are highly similar, or if too few species are used (Anisimova et al. 2001). The current analysis uses species of *Drosophila*, which are expected to be sufficiently divergent to minimize the amount of type II error. The simulation study performed by Kosakovsky and Frost (2005) also showed that with their eight-sequence datasets, PAML (M8 model) performed as well as all other approaches. Analysis performed in the current study using five, six, or seven species showed consistent results (Supporting Table S1). 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). The proportion of positively selected genes can be compared to molecular evolution of *Drosophila* seminal fluid proteins using the *melanogaster* species subgroup (Haerty et al. 2007). Twenty-five seminal fluid genes had orthologs in all of the *melanogaster* subgroup species and four of these genes (16%) exhibited positive selection by the criteria of acceptance of M8 over M7. By the same criteria, of 679 genes expressed in the reproductive tract of *D. melanogaster*, and of 9921 nonsex/reproduction-related genes, 6.2% and 6.0%, respectively, were consistent with the hypothesis of positive selection by acceptance of model 8 (Haerty et al. 2007). It is important to note that direct comparisons made between studies are tempered by the technical differences of the individual studies. For example, a hybrid selection study such as the present study is expected to be based on a more restricted set of genes than other approaches. Nevertheless, the incidence of directional selection among spermathecal genes is striking.

In the present study, serine proteases are the predominant category of genes lost from *D. melanogaster* and *D. simulans* as a function of evolutionary distance. Among spermatheca proteases, four have no detectable 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. Rapid evolution of protease genes in the *melanogaster* subgroup (and related species) continues until most spermathecal protease genes are lost, or no longer recognizable as an ortholog, in more distantly related taxa. In other studies using *Drosophila* species, reproductive system proteases show evidence of accelerated and positive evolution (Kern et al. 2004; Swanson et al. 2004; Panhuis and Swanson 2006; Kelleher et al. 2007; Lawniczak and Begun 2007; Wong et al. 2008) the proportion of such genes in the spermatheca is the highest recorded considering these relevant studies.

All of the spermatheca serine proteases have secretion signals (Table 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 (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 and 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*.

The evolutionary importance of the interactions that occur between females and male ejaculate are being increasingly recognized. 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. 2002). 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. 2008) and it also is a candidate for a coevolutionary interaction with spermatheca proteases based on direct interaction.

Four of the protease genes identified in the present study are found in a cluster on the chromosomal 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 and Begun 2007). These proteases, and several others, are rapidly evolving between populations of *D. melanogaster* and diverging between *D. melanogaster* and *D. simulans.* (Lawniczak and 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.

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 and Swanson 2006; Lawniczak and Begun 2007). Three sites surrounding Ser195, one of the three identified positions of the serine protease catalytic triad, have been identified as responsible for substrate specificity (Perona and Craik 1995; Hedstrom 2002) were examined in the serine proteases of the present study. A variety of changes were seen in and around these regions, suggesting that the changes affected catalysis and substrate specificity (Fig. 1; Supporting Figure S1). CG18125 was also found to be significantly induced by mating (McGraw et al. 2004). The expression of CG18125 in mated females was over twice that of virgin females.

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 and Beckman 1993) play a variety of roles including interaction with sperm in both vertebrate and invertebrate taxa (Kamei and Glabe 2003; Haley and Wessel 2004). 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. 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 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 (1996) and is used to interpret the evolution of VERL domains in reproductive proteins (Swanson and 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 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. In honey bees the spermathecal fluid contains sugars including glucose, trehalose, and fructose, as well as a high level of trehalase activity (Alumot et al. 1969). 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, which stimulates JH synthesis (Peng et al. 2005). An enzyme that produces a precursor of JH is elevated in the lower female reproductive tract after mating (Mack et al. 2006). 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 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 that 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 that of Acp genes that are a paradigm for rapid evolution. Rapidly evolving spermatheca proteins of established and novel function could participate in female reproductive molecule-ejaculate interactions that are increasingly recognized as evolutionarily important (Ravi-Ram and Wolfner 2007).

ACKNOWLEDGMENTS

We thank M. Wolfner, A. Wong, W. Wagner, M. Noor, C. Aquadro, and A. Clark for comments on the manuscript and other contributions to the research. This research was supported by a National Science Foundation grant (DEB-ESP0346476).

LITERATURE CITED

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Alumot, E., Y. Lensky, and P. Holstein. 1969. Sugars and trehalase in the reproductive organs and hemolymph of the queen and drone honey bees (*Apis mellifera L.Var. Ligustica Spi.*). Comp. Biochem. Physiol. B 28:1419–1425.

Andres, J. A., L. S. Maroja, S. M. Bogdanowicz, W. J. Swanson, and R. G. Harrison. 2006. Molecular evolution of seminal proteins in field crickets. Mol. Biol. Evol. 23:1574–1584.

Anisimova, M., J. P. Bielawski, and Z. Yang. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18:1585–1592.

Begun, D. J., P. Whitley, B. L. Todd, W.-D. H.M., and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in *Drosophila*. Genetics 156:1879–1888.

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

Bork, P., and G. Beckman. 1993. The CUB domain. A widespread module in develomentally regulated proteins. J. Mol. Biol. 231:539–545.

Clark, A. G., and D. J. Begun. 1998. Female genotypes affect sperm displacement in *Drosophila*. Genetics 149:1487–1493.

Clark, A. G., M. Aguade, T. Prout, L. G. Harshman, and C. H. Langley. 1995. Variation in sperm displacement and its association with accessory gland protein loci in *Drosophila melanogaster*. Genetics 139:189–201.

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

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

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

Collins, A. M., T. J. Caperna, V. Williams, W. M. Garrett, and J. D. Evans. 2006. Proteomic analysis of male contributions to honeybee sperm storage and mating. Insect Mol. Biol. 15:541–549.

Coyne, J. A., and H. A. Orr. 2004. Speciation. Sinauer Associates, Sunderland, MA.

Darwin, C. 1871. The descent of man and selection in relation to sex. J. Murray, London.

Davies, S. J., and T. Chapman. 2006. Identification of genes expressed in the accessory glands of male Mediterranean Fruit Flies (*Ceratitis capitata*). Insect Biochem. Mol. Biol. 36:846–856.

Diatchenko, L., Y. Lau, A. Campbell, A. Chenchik, F. Mogadam, B. Huang, S. Lukyanov, K. Lukyanov, N. Gurskaya, E. Sverdlov, and P. Siebert. 1996. Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries. Proc. Natl. Acad. Sci. USA 93:6025–6030.

Diatchenko, L., K. Lukyanov, Y. Lau, and P. Siebert. 1999. Suppression subtractive hybridization: a versatile method for identifying differentially expressed genes. Method Enzym. 303:349–380.

DiBenedetto, A. J., D. M. Lakich, W. D. Kruger, J. M. Belote, B. S. Baker, and M. F. Wolfner. 1987. Sequences expressed sex-specifically in *Drosophila melanogaster* adults. Dev. Biol. 119:242–251.

Eberhard, W. G. 1996. Female control: sexual selection by cryptic female choice. Princeton Univ. Press, Princeton, NJ.

Edgar, R. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

Emanuelsson, O., S. Ren-Brunak, G. vonHeijne, and H. Nielsen. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nature Prot. 2:953–971.

Ferrandon, D., A. Jung, M. Criqui, B. Lemaitre, S. Uttenweiler-Joseph, L. Michaut, J. Reichhart, and J. Hoffman. 1998. A drosomycin-GFP reporter transgene reveals a local immune response in *Drosophila* that is not dependent on the Toll pathway. EMBO 17:1217–1227.

Filosi, M., and M. Perotti. 1975. Fine structure of spermatheca of *Drosophila melanogaster* Meig. J. Submicr. Cytol. 7:259–270.

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

———. 2007. Associations between sperm competition and natural variation in male reproductive genes on the third chromosome of *Drosophila melanogaster*. Genetics 176:1245–1260.

Fowler, G. 1973. Some aspects of the reproductive biology of *Drosophila*: sperm transfer, sperm storage and sperm utilization. Adv. Genet. 17:293–360.

Galindo, B., V. Vacquier, and W. Swanson. 2003. Positive selection in the egg receptor for abalone sperm lysine. Proc. Natl. Acad. Sci. USA 100:4639–4643.

Goldman, N., and Z. Yang. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11:725–736.

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

Haley, S., and G. Wessel. 2004. Proteolytic cleavage of the cell surface protein p160 is required for detachment of the fertilization envelope in the sea urchin. Dev. Biol. 272:191–202.

Harshman, L. G., and T. Prout. 1994. Sperm displacement without sperm transfer in *Drosophila melanogaster*. Evolution 48:758–766.

Hedstrom, L. 2002. Serine protease mechanism and specificity. Chem. Rev. 102:4501–4524.

Howard, D. J. 1999. Conspecific sperm and pollen precedence and speciation. Ann. Rev. Ecol. Syst. 30:109–132.

Howard, D. J., S. R. Palumbi, L. Birge, and M. K. Manier. 2008 Sperm and speciation. Chapter 9 *in* T. R. Birkhead, D. J. Hosken, and S. Pitnick, eds. Sperm biology: an evolutionary approach. Elsevier Press, Amsterdam.

Huang, X., and A. Madan. 1999. CAP3: A DNA sequence assembly program. Genome Res. 9:868–877.

Käll, L., A. Krogh, and E. L. L. Sonnhammer. 2004. A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. 338:1027–1036.

Kamei, N., and C. Glabe. 2003. The species-specific egg receptor for sea urchin sperm adhesion is EBR1, a novel ADAMTS protein. Genes Devel. 17:2501–2507.

Kamei, N., W. Swanson, and C. Glabe. 2000. A rapidly diverging EGF protein regulates species-specific signal transduction in early sea urchin development. Dev. Biol. 225:267–276.

Kelleher, E. S., W. J. Swanson, and T. A. Markow. 2007. Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet. 3:e148.

Kern, A. D., C. D. Jones, and D. J. Begun. 2004. Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 167:725–735.

Kosakovsy Pond, S. L., and S. D. Frost. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol. Biol. Evol. 22:1208–1222.

Lawniczak, M. K., and D. J. Begun. 2007. Molecular population genetics of female-expressed mating-induced serine proteases in *Drosophila melanogaster*. Mol. Biol. Evol. 24:1944–1951.

Lefevre, G., and U. B. Jonsson. 1962. Sperm transfer, storage, displacement, and utilization in *Drosophila melanogaster*. Genetics 47:1719–1736.

Letunic, I., R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork. 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res. 34:D257–260.

Lung, O., U. Tram, M. Finnerty, M. Eipper-Mains, J. Kalb, and M. Wolfner. 2002. The *Drosophila melanogaster* seminal fluid protein Acp62F is a protease inhibitor that is toxic upon ectopic expression. Genetics 160:211–214.

Mack, P., A. Kapelnikov, Y. Heifetz, and M. Bender. 2006. Mating-responsive genes in reproductive tissues of female *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA 103:10358–10363.

Malm, J., J. Hellman, P. Hog, and H. Lilja. 2000. Enzymatic action of prostate-specific antigen (PSA or hK3): substrate specificity and regulation by Zn (2+). Prostate 45:132–139.

Marchler-Bauer, A., and S. Bryant. 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 32:327–331.

McGraw, L. A., G. Gibson, A. G. Clark, and M. F. Wolfner. 2004. Genes regulated by mating, sperm, or seminal proteins in mated female *Drosophila melanogaster*. Curr. Biol. 14:1509–1514.

Metz, E., and S. Palumbi. 1996. Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. 13:397–406.

Miller, G., and S. Pitnick. 2002. Sperm-female co-evolution in *Drosophila*. Science 298:1230–1233.

———. 2003. Functional significance of seminal receptacle length in *Drosophila melanogaster*. J. Evol. Biol. 16:114–116.

Monsma, S. A., and M. F. Wolfner. 1988. Structure and expression of a *Drosophila* male accessory gland gene whose product resembles a peptide pheromone precursor. Genes Dev. 2:1063–1073.

Monsma, S. A., H. A. Harada, and M. F. Wolfner. 1990. Synthesis of two *Drosophila* male accessory gland proteins and their fate after transfer to the female during mating. Dev. Biol. 142:465–475.

Mueller, J., K. Ravi-Ram, M. McGraw, M. Bloch-Qazi, E. Siggia, A. Clark, C. Aquadro, and M. Wolfner. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. Genetics 171:131–143.

Neubaum, D. M., and M. F. Wolfner. 1999. Mated *Drosophila melanogaster* females require a seminal fluid protein, Acp36DE, to store sperm efficiently. Genetics 153:857–869.

Palumbi, S. 1999. All males are not created equal: fertility difference depend on gamete recognition polymorphisms in sea urchins. Proc. Natl. Acad. Sci. USA 99:12632–12637.

Panhuis, T. M., and W. J. Swanson. 2006. Molecular evolution and population genetic analysis of candidate female reproductive genes in *Drosophila*. Genetics 173:2039–2047.

Peng, J., S. Chen, S. Busser, H. Liu, T. Honegger, and E. Kubli. 2005. Gradual release of sperm bound sex-peptide controls female postmating behavior in *Drosophila*. Curr. Biol. 15:207–213.

Perona, J. J., and C. S. Craik. 1995. Structural basis of substrate specificity in the serine proteases. Protein Sci. 4:337–360.

Pitnick, S., T. A. Markow, and G. S. Spicer. 1999. Evolution of multiple kinds of female sperm-storage organs in *Drosophila*. Evolution 53:1804–1822.

Pollard, D. A., V. N. Iyer, A. M. Moses, and M. B. Eisen. 2006. Widespread discordance of gene trees with species tree in *Drosophila*: evidence for incomplete lineage sorting. PLoS Genet. 2:e173.

Price, C. S. 1997. Conspecific sperm precedence in *Drosophila*. Nature 388:663–666.

Price, C. S., C. H. Kim, C. J. Gronlund, and J. A. Coyne. 2001. Cryptic reproductive isolation in the *Drosophila simulans* species complex. Evolution 55:81–92.

Price, C. S. C., K. A. Dyer, and J. A. Coyne. 1999. Sperm competition between *Drosophila* males involves both displacement and incapacitation. Nature 400:449–452.

Ravi-Ram, K., and M. F. Wolfner. 2007. Seminal influences: *Drosophila* Acps and the molecular interplay between males and females during reproduction. Int. Comp. Biol. 47: 427–445.

Ross, J., H. Jiang, M. Kanost, and Y. Wang. 2003. Serine proteases and their homologs in *Drosophila melanogaster* genome: an initial analysis of sequence conservation and phylogenetic relationships. Gene 304:117–131.

Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc. Natl. Acad. Sci. USA 95:5857–5864.

Song, J., J. Wong, and G. Wessel. 2006. Oogenesis: single cell development and differentiation. Dev. Biol. 300:385–405.

Suzuki, Y., and M. Nei. 2004. False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. Mol. Biol. Evol. 21:914–921.

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

Swanson, W., and V. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. 3:137–144.

Swanson, W. J., A. G. Clark, H. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001a. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. Proc. Natl. Acad. Sci. USA 98:7375–7379.

Swanson, W. J., Z. Yang, M. F. Wolfner, and C. F. Aquadro. 2001b. Positive selection droves the evolution of several female reproductive proteins in mammals. Proc. Natl. Acad. Sci. USA 98:2509–5214.

Swanson, W., R. Nielsen, and Z. Yang. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. Mol. Biol. Evol. 20:18–20.

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

Tusnády, G. E., and I. Simon. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics 17:849–850.

Wagstaff, B., and D. J. Begun. 2004. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol. Biol. Evol. 22:818–832.

Weirich, G. G., A. M. Collins, and V. P. Williams. 2002. Antioxidant enzymes in the honey bee, *Apis mellifera*. Apidologie 33:3–14.

Wheeler, D. E., and P. H. Krutzsch. 1994. Ultrastructure of the spermathecae and its associated gland in the and *Crematogaster opuntiae* (Hymenoptera: Formicidae). Zoomorphology 114:203–214.

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

Wong, A., M. C. Turchin, M. F. Wolfner, and C. F. Aquadro. 2008. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. Mol. Biol. Evol. 25:497–506.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13:555–556.

———. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol. Biol. Evol. 15:568–573.

———. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol. Biol. Evol. 24:1586–1591.

Yang, Z., and R. Nielsen. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. J. Mol. Evol. 46:409–418.

———. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol. Biol. Evol. 19:908–917.

Yang, Z., and W. J. Swanson. 2002. Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol. Biol. Evol. 19:49–57.

Yang, Z., W. J. Swanson, and V. D. Vacquier. 2000. Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. 17:1446–1454.

Zhu, Y., E. Machleder, A. Chenchik, and P. Siebert. 2001. Reverse transcriptase template switching: a SMART approach for full-length cDNA library construction. Biotechniques 30:892–897.

Associate Editor: T. Crease

Supporting Information

The following supporting information is available for this article:

Figure S1. Amino acid sites surrounding the serine active site (bold and underlined) of the catalytic triad for the serine proteases discovered in the spermathecae.

Table S1. Gene identification numbers for *D. simulans* and *D. melanogaster*. *D. melanogaster* CG and FB gene identification numbers; Function of predicted proteins; Secretion Signal prediction; Transmembrane region prediction; Pairwise dN/dS; PAML data; Presence of orthologs in 12 genomes

Table S2. Gene identification numbers for *D. simulans* and *D. melanogaster*; Pairwise dN/dS; PAML branch model data.

This material is available as part of the online article from:
http://www.blackwell-synergy.com/doi/abs/10.1111/j.1558-5646.2008.00493.x
(This link will take you to the article abstract).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.