

Gene Duplication and Adaptive Evolution of Digestive Proteases in *Drosophila arizonae* Female Reproductive Tracts

Erin S. Kelleher^{1*}, Willie J. Swanson², Therese A. Markow¹

1 Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona, United States of America, **2** Department of Genome Sciences, University of Washington, Seattle, Washington, United States of America

It frequently has been postulated that intersexual coevolution between the male ejaculate and the female reproductive tract is a driving force in the rapid evolution of reproductive proteins. The dearth of research on female tracts, however, presents a major obstacle to empirical tests of this hypothesis. Here, we employ a comparative EST approach to identify 241 candidate female reproductive proteins in *Drosophila arizonae*, a *repleta* group species in which physiological ejaculate–female coevolution has been documented. Thirty-one of these proteins exhibit elevated amino acid substitution rates, making them candidates for molecular coevolution with the male ejaculate. Strikingly, we also discovered 12 unique digestive proteases whose expression is specific to the *D. arizonae* lower female reproductive tract. These enzymes belong to classes most commonly found in the gastrointestinal tracts of a diverse array of organisms. We show that these proteases are associated with recent, lineage-specific gene duplications in the *Drosophila repleta* species group, and exhibit strong signatures of positive selection. Observation of adaptive evolution in several female reproductive tract proteins indicates they are active players in the evolution of reproductive tract interactions. Additionally, pervasive gene duplication, adaptive evolution, and rapid acquisition of a novel digestive function by the female reproductive tract points to a novel coevolutionary mechanism of ejaculate–female interaction.

Citation: Kelleher ES, Swanson WJ, Markow TA (2007) Gene duplication and adaptive evolution of digestive proteases in *Drosophila arizonae* female reproductive tracts. PLoS Genet 3(8): e148. doi:10.1371/journal.pgen.0030148

Introduction

Extensive research across a broad range of taxa has revealed that the proteins involved in sexual reproduction often evolve rapidly due to positive selection (reviewed in [1–3]). Although the selective forces that underlie this pattern remain unclear, it frequently has been postulated that adaptive evolution of reproductive proteins may result from intersexual coevolution [1–3]. Indeed, this has been demonstrated in the fertilization proteins of the free-spawning marine gastropod abalone, in which the male protein lysin and its female receptor, vitelline envelope receptor for lysin (VERL), both exhibit signatures of adaptive evolution [4–7]. In internally fertilizing organisms, however, such as mammals or insects, the biochemical interactions between male and female reproductive proteins may be vastly more complex. Reproductive outcomes depend not only on interactions between male and female gamete proteins, but additionally on interactions between male seminal proteins and proteins in the lumen of a female's reproductive tract [8–11].

Fruit flies of the genus *Drosophila* provide an important model system for exploring the function and evolution of reproductive tract interactions (reviewed in [9–12]). In *Drosophila melanogaster*, the male ejaculate comprises just under 100 proteins, several of which are known to stimulate important processes in mated females such as ovulation, oogenesis, and sperm storage (reviewed in [9–11]). Several male proteins either undergo proteolytic cleavage in mated females [13–15], or localize to specific portions of the female reproductive tract [16–18], indicating that ejaculate–female interactions are mediated biochemically by females. Between

species, rapid changes in ejaculate composition frequently have resulted in lineage-specific seminal proteins [19–21], many of which may be novel coding sequences [22]. Additionally, molecular evolutionary studies indicate that a significant portion of this ejaculate is subject to positive selection in the *melanogaster* [23–25], *obscura* [26], and *repleta* species groups [27].

By comparison, the female side of reproductive tract interactions has received little attention. Female reproductive tract proteins have been identified transcriptionally only in *D. simulans* [28], and their functions remain entirely unknown. Furthermore, although several female reproductive tract proteins [28–30] and egg membrane proteins [31] show evidence of positive selection, these analyses largely have been confined to the *melanogaster* species group. It is unclear, therefore, how diversity in female reproductive physiology and mating system across the genus [reviewed in 12,32] is reflected in their reproductive proteins. This overall paucity

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received: May 23, 2007; **Accepted:** July 13, 2007; **Published:** August 31, 2007

A previous version of this article appeared as an Early Online Release on July 18, 2007 (doi:10.1371/journal.pgen.0030148.eor).

Copyright: © 2007 Kelleher et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abbreviations: EST, expressed sequence tag; GO, gene ontology; MYA, million years ago; RT-PCR, reverse transcriptase PCR

* To whom correspondence should be addressed. E-mail: kelleher@email.arizona.edu

Author Summary

In a broad range of organisms, including humans, molecular interactions between the male ejaculate and the female reproductive tract play integral roles in sexual reproduction. Although these interactions are essential, the biochemical composition of the male ejaculate can change rapidly over short evolutionary time periods. It is often hypothesized that this rapid evolution reflects a coevolutionary relationship with the female reproductive tract. The paucity of research on females, however, presents a formidable challenge to empirical tests of this hypothesis. In this study, we sought to identify proteins in the female reproductive tracts of *D. arizonae* that may be interacting or coevolving with the male ejaculate. Unexpectedly, we discovered that *D. arizonae* females produce an array of “digestive” enzymes in their reproductive tracts. These classes of enzymes are normally found in the gut, where they degrade ingested food for nutritional uptake. In *D. arizonae*, these enzymes have resulted from recent gene duplications, and natural selection has caused rapid and radical changes in their amino acid sequences. We propose that this pattern of duplication and diversification reflects the “female side” of a coevolutionary relationship with the male ejaculate. Exploring the “male side” of this relationship is an important avenue for future research.

of research on females presents a major obstacle to understanding the evolution of ejaculate–female interactions and the role of intersexual dynamics in the divergence of reproductive proteins.

Here we use a comparative expressed sequence tag (EST) approach to characterize candidate female reproductive tract proteins in *D. arizonae*. *D. arizonae* is a *repleta* group species that exhibits important differences from the *melanogaster* group in mating system and female physiology. *D. arizonae* females remate daily, while *D. simulans* females wait several days before remating [12]. Female promiscuity may affect the evolution of reproductive proteins by increasing the number of competing male ejaculates [33]. Females of *D. arizonae* additionally exhibit two remarkable post-mating physiological processes not seen in the *melanogaster* group. First, they incorporate peptide components of the male ejaculate into somatic tissues and oocytes [34], an adaptation which may help defray the cost of egg production during periods of resource limitation [35]. Second, they exhibit an insemination reaction, an opaque white mass of unknown biochemical composition that forms in the female uterus after copulation [36].

By comparing post-mating outcomes in inter- and intrapopulation crosses, several studies have presented evidence for ejaculate–female coevolution in natural populations of *D. arizonae* and its sister species *D. mojavensis* (most recent common ancestor, ~1.5 million years ago [MYA]) [37–41]. Intrapopulation crosses of both species produce larger eggs than interpopulation crosses [38], a process known to be stimulated by several components of the male ejaculate in *D. melanogaster* (reviewed in [9–11]). Additionally, the insemination reaction exhibits a larger size and duration in interpopulation crosses relative to intrapopulation crosses, suggesting this trait is subject to sexually antagonistic coevolution [39]. Finally, desiccation resistance is higher in mated than unmated females [40], and the magnitude of this effect differs between inter- and intrapopulation crosses [41]. Such extensive evidence for physiological coevolution in-

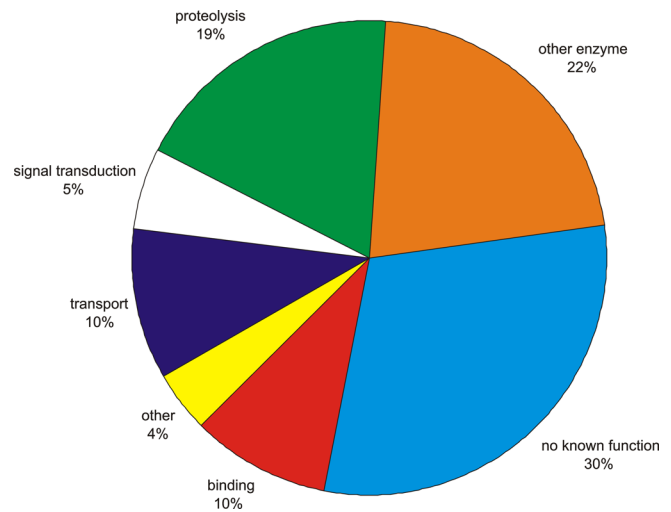


Figure 1. Functional Composition of Candidate Female Reproductive Proteins

Functional composition of 241 secreted and transmembrane proteins in *D. arizonae* female reproductive tracts based on GO terms [59]. doi:10.1371/journal.pgen.0030148.g001

icates this will be an exciting system to explore the molecular basis of reproductive tract interactions.

Our study identifies 241 candidate female reproductive proteins in *D. arizonae*, of which 31 show elevated rates of amino acid substitution suggestive of adaptive evolution. Unexpectedly, we also discovered three lineage-specific gene families of digestive proteases whose expression is specific to the lower female reproductive tract. These proteins exhibit strong signatures of adaptive evolution, and selected sites cluster near functionally important amino acids. The implications of these findings for ejaculate–female interactions and intersexual coevolution are discussed.

Results/Discussion

Functional Classes of Female Reproductive Proteins

We sequenced a total of 2,304 ESTs derived from the *D. arizonae* lower female reproductive tract (parovaria, oviduct, spermathecae, seminal receptacle, and uterus) representing 649 unique proteins (for a complete list see Table S1). Of particular interest are proteins found on cell surfaces or in the lumen of this tissue, which interact directly with the male ejaculate and likely play an integral role in reproductive tract interactions [28]. We therefore designate candidate female reproductive proteins as those that exhibit secreted signal sequences, or transmembrane domains. The gross functional composition of the 241 candidate female reproductive proteins identified in this study (Figure 1) are similar to those of *D. simulans* [28], and include transport, signal transduction, and proteolysis.

Rapid Evolution of Female Reproductive Proteins

To explore the evolutionary histories our candidate female reproductive proteins, we calculated the ratio of replacement to silent substitutions (d_N/d_S) between our *D. arizonae* ESTs and their orthologs in the *D. mojavensis* genome. Candidate female reproductive proteins exhibit significantly larger d_N/d_S values than intracellular proteins in our dataset (median test, $p >$

Table 1. Candidate Female Reproductive Proteins

<i>D. mojavensis</i> CDS	Signal Sequence	Transmembrane Domains	d_N	d_S	d_N/d_S	<i>D. melanogaster</i> CDS	<i>D. melanogaster</i> Function	Conserved Domain
GLEANR_9982	S	0	0.03	0.02	1.83	CG7443-PA	Unknown function	No
anon-EST:Kelleher15	S	1	0.03	0.02	1.38	No hits	NA	No
GLEANR_5396	S	1	0.02	0.02	1.30	Gp150-PD	Cell adhesion/ signal transduction	LRR_1
GLEANR_4627	Q	10	0.03	0.03	1.28	CG30344-PA	Unknown function	MFS_1
anon-EST:Kelleher5	S	0	0.05	0.04	1.20	CG10472	Proteolysis	Trypsin
GLEANR_17128	S	2	0.06	0.06	1.08	No hits	NA	No
GLEANR_5358	Q	1	0.01	0.01	1.04	CG30415-PA	Unknown function	No
GLEANR_1967	S	1	0.05	0.05	0.96	CG7778-PA	Unknown function	No
GLEANR_896	S	0	0.11	0.12	0.89	CG31954-PA	Proteolysis	Trypsin
GLEANR_17617	S	4	0.02	0.03	0.82	CG4729-PA	Metabolism	Acyltransferase
GLEANR_3367	A	0	0.02	0.03	0.82	lectin-46Cb-PA	Sugar binding	Lectin_C
GLEANR_5037	A	4	0.01	0.02	0.80	CG15098-PA	Unknown function	No
GLEANR_9029	A	2	0.01	0.01	0.80	Nep2-PA	Proteolysis	Peptidase_M13_N, Peptidase_M13
GLEANR_13559	Q	12	0.02	0.03	0.79	CG10960-PB	Metabolism/transport	Sugar_tr, MFS_1
GLEANR_2703	S	0	0.09	0.12	0.77	CG15254-PA	Proteolysis	Astacin
GLEANR_9617	S	0	0.07	0.09	0.74	CG3739-PA	Proteolysis	Peptidase_S28
GLEANR_12094	S	0	0.01	0.02	0.73	CG6409-PA	GPI anchor biosynthesis	No
GLEANR_10002	S	0	0.04	0.05	0.70	CG9418-PA	DNA binding	No
GLEANR_1209	A	1	0.04	0.06	0.69	No hits	NA	No
GLEANR_6649	S	0	0.01	0.02	0.67	I(1)G0193-PB	Unknown function	No
GLEANR_7394	S	0	0.02	0.03	0.66	CG17271-PA	Ca ²⁺ Binding	No
GLEANR_10683	S	0	0.03	0.04	0.65	CG5630-PA	Unknown function	No
GLEANR_6054	Q	2	0.01	0.02	0.62	CG4627-PA	Unknown function	No
GLEANR_2775	Q	10	0.03	0.05	0.59	CG4726-PA	Metabolism/transport	MFS_1
anon-EST:Kelleher9	S	0	0.04	0.06	0.59	CG10472	Proteolysis	Trypsin
GLEANR_1245	Q	1	0.03	0.05	0.57	No hits	NA	Lamp
GLEANR_16396	S	0	0.05	0.09	0.55	No hits	NA	No
GLEANR_2881	Q	1	0.03	0.05	0.54	CG14536-PB	Unknown function	Ubiquitin
GLEANR_8258	S	0	0.03	0.06	0.53	CG3734-PA	Proteolysis	Peptidase_S28
GLEANR_7051	S	0	0.05	0.09	0.52	CG6283-PA	Lipid metabolism	Lipase
GLEANR_3613	S	0	0.02	0.04	0.51	CG18067-PA	Unknown function	No

Signal sequence: S, secreted; A, anchor; and Q, quiescent as predicted by SignalP 3.0 [56].

TM, number of identified transmembrane domains [57] Ka, estimated nonsynonymous substituted ratio nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site, calculated in PAML [47]; *D. melanogaster* CG, best tblastx hit in the *D. melanogaster* genome; *D. melanogaster* function, from flybase annotations; and conserved domain, pfam conserved domain predicted from hmmpfam [58].

doi:10.1371/journal.pgen.0030148.t001

0.0001), suggesting that these proteins evolve more rapidly than their intracellular counterparts. This elevated rate of amino acid substitution is predicted if adaptive evolution of secreted and transmembrane proteins is a frequent consequence of molecular coevolution with components of the male ejaculate.

Under strict neutrality, only $d_N/d_S \gg 1$ can be considered robust evidence of adaptive evolution. While several of our candidate genes show $d_N/d_S > 1$, none of these tests is statistically significant (Table 1). A literature survey has shown, however, that 95% genes that exhibit a pairwise $d_N/d_S > 0.5$ contain a class of sites with $d_N/d_S \gg 1$ [28]. Of 227 pairwise comparisons, 31 (14%) were identified with $d_N/d_S > 0.5$, indicating they are likely experiencing positive selection (Table 1). This result is largely independent of gene duplication, as the estimated frequency of adaptive evolution it is still 13% when recent duplicates are excluded from the dataset.

On a functional level, several protein classes that commonly occur in seminal and fertilization proteins, including lipases, lectins, glycoproteins and proteases, are found in our candidates for adaptive evolution (Table 1). Roughly half of

these 31 candidates, however, have no known function, and several others belong to functional classes that are not commonly represented among reproductive proteins. Proteins with unusual or unknown functions make excellent candidates for discovering genes which have acquired novel functions in a biochemical network which likely evolves rapidly. Future studies of these 31 candidates will yield significant insight into the function and evolution of reproductive tract interactions in the *repleta* species group.

Gene Duplication in Female Reproductive Proteins

Gene duplication plays an integral role in the evolution of *D. arizonae* female reproductive tract proteins. Specifically, 47% (16) of all secreted proteases in *D. arizonae* female reproductive tracts have at least one closely related paralog that also is expressed in these same tissues. Duplication events have been extremely recent; as multiple, tandemly-duplicated paralogs in the *D. mojavensis* genome correspond to only a single gene in *D. virilis*, the most closely related fully sequenced outgroup (most recent common ancestor, ~23 MYA; reviewed in [42]). We therefore estimate that the duplication rate of secreted proteases expressed in *D. arizonae*

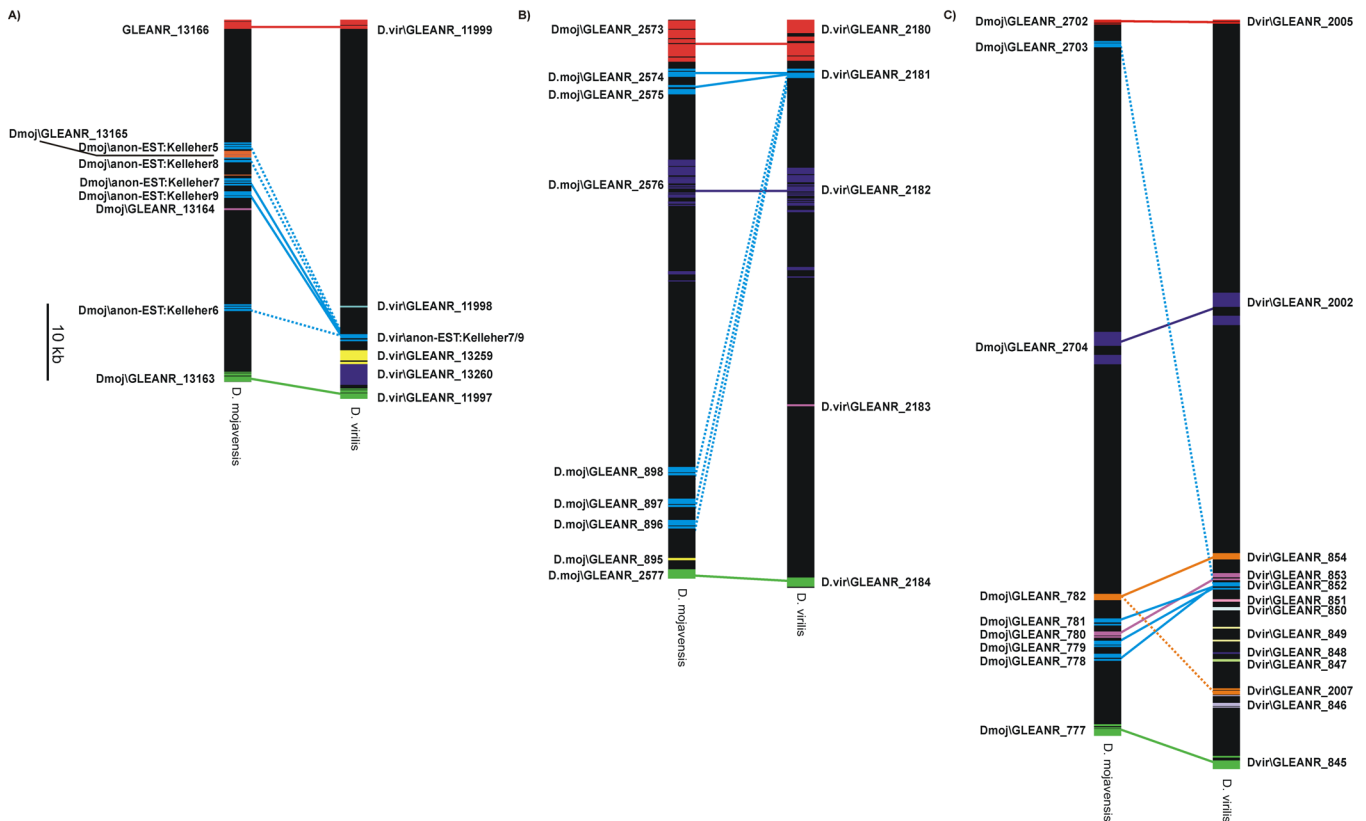


Figure 2. Distribution of Three Protease Gene Families in *D. mojavensis* and *D. virilis* Genomes

(A) Syntenic regions of protease gene family 1: *D. mojavensis* Chromosome 4 (scaffold_6680, bp 10216565–10169309) and *D. virilis* Chromosome 3 (scaffold_13049, bp 10558802–10608251).

(B) protease gene family 2: *D. mojavensis* Chromosome 3 (scaffold_6500, bp 18241557–18296199) and *D. virilis* Chromosome 4 (scaffold_12963, bp 15263878–15319561).

(C) protease gene family 3: *D. mojavensis* Chromosome 3 (scaffold_6500, bp 20970182–21063420) and *D. virilis* chromosome 4 (scaffold_12963 bp 12250368–12347919).

Colored blocks indicate individual exons, where each gene is indicated by a different color. Orthologous genes are the same color in both species, and connected by colored lines. Solid lines indicate orthologs with the same orientation, while dotted lines indicate inverted orthologs. Multiple, tandemly duplicated copies in the genome of *D. mojavensis* correspond to a single gene in the genome of *D. virilis*. Annotation and assembly obtained from unpublished *Drosophila* genomes (<http://rana.lbl.gov/drosophila/>). doi:10.1371/journal.pgen.0030148.g002

tracts is 0.0298 (duplications per gene per million years, see Materials and Methods), which is 21-fold higher than the genome wide estimate for *D. melanogaster* (0.0014, [43]). Although the selective forces involved are yet obscure, such recent and pervasive gene duplication has not been seen in any class of reproductive protein yet studied, including *D. simulans* female reproductive proteins [28].

Four (of 16) duplicated proteases have resulted from two single gene duplication events. The remaining 12 duplicated proteases, however, are associated with small lineage-specific gene families. Each family contains four to six tandemly duplicated paralogs in the genome of *D. mojavensis* that are syntenic to a single ortholog in the genome of *D. virilis* (Figure 2). For brevity, we hereafter refer to these three families of tandem duplicates as protease gene family 1, 2, and 3. Phylogenetic analysis of *D. arizonae* ESTs, and coding sequences from the genomes of *D. mojavensis*, *D. virilis*, and *D. grimshawi* (<http://rana.lbl.gov/drosophila/>), reveals the majority of these tandem duplicates in the *D. mojavensis* genome have a *D. arizonae* ortholog that is expressed in the lower female reproductive tract (Figure 3). This strongly suggests

that the gene duplication events relate in some way to the reproductive function of these proteases. Indeed, reverse transcriptase PCR (RT-PCR) of all three gene families reveals that in adult *D. arizonae* these genes are exclusively expressed in the lower female reproductive tract (Figure 4). Gene copies present in the *D. mojavensis* genome that do not correspond to *D. arizonae* ESTs are likely not highly expressed.

While the function of these duplicated proteins in *D. arizonae* female reproductive tracts is unknown, they are often similar or identical in their key amino acid residues to several families of digestive proteases found almost exclusively in gastrointestinal tracts (Table 2). Specifically, protease gene families 1 and 2 share appreciable homology with trypsin, chymotrypsin, and elastase, serine endopeptidases commonly found in digestive tracts of both insects and mammals [reviewed in 44]. While, serine endopeptidases can also function in immune signaling cascades across a broad array of organisms, such proteases generally have secondary protein–protein interaction domains that allow for localized regulation of physiological responses [45]. No such domains are seen in either protease gene family 1 or 2, suggesting

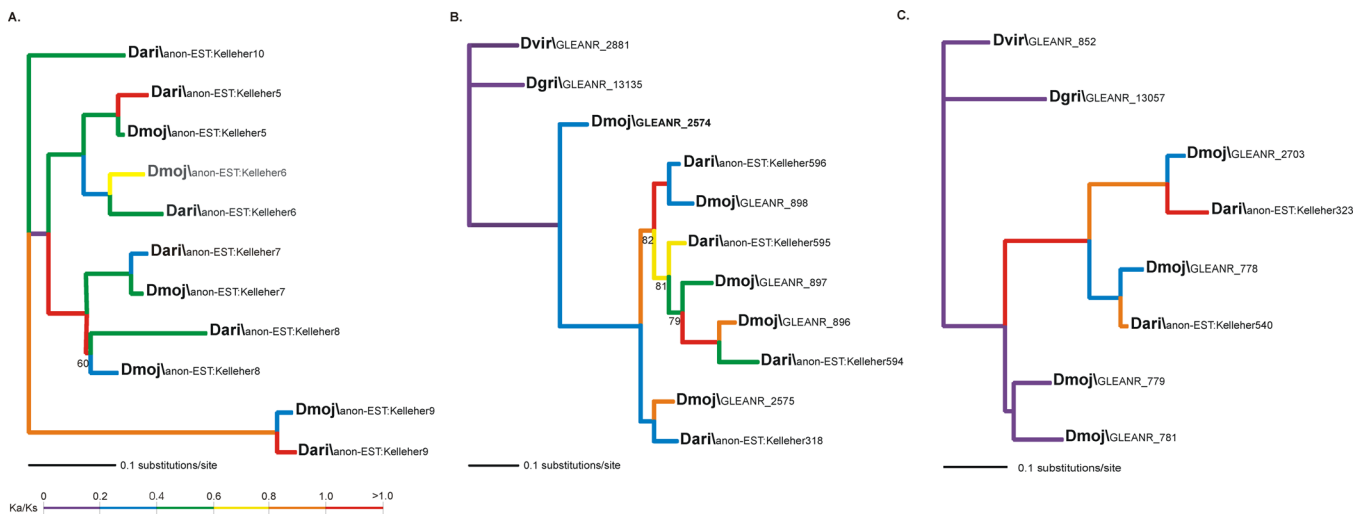


Figure 3. Bayesian Phylogenies of (A) Protease Gene Family 1, (B) Protease Gene Family 2, and (C) Protease Gene Family 3

(A) is midpoint rooted, as *D. virilis* sequence was too divergent to make an appropriate outgroup. Grey taxon name denotes a pseudogene. Branch colors indicate Ka/Ks values calculated in the codeml package of PAML [47]. Posterior probabilities < 90 are noted. doi:10.1371/journal.pgen.0030148.g003

these proteases exhibit a primarily digestive function. Similar to the two families of serine endopeptidases, protease gene family 3 contains zinc metalloendoproteases very similar to astacin, a prominent digestive enzyme in the crayfish midgut [reviewed in 46]. The reproductive tract-specific expression of these proteases, coupled with recent, lineage-specific gene duplications, suggest that *D. arizonae* female reproductive tracts recently have acquired a novel digestive function. Digestive enzymes in female reproductive tracts likely have important implications for male reproductive success, and therefore, the evolution of the male ejaculate.

Adaptive Evolution of Digestive Proteases

There is compelling evidence that directional selection has played an important role in the evolution of reproductive tract-specific secreted digestive proteases in *D. arizonae* females. All three families of digestive proteases exhibit a class of sites whose ratio of nonsynonymous to synonymous substitutions (d_N/d_S) is significantly greater than the neutral expectation of 1 (Table 2). d_N/d_S values for these selected sites range from 2 to 11.96, indicating certain amino acids in these proteins have experienced strong positive selection. Notably, the two single gene duplication events show no evidence of adaptive evolution (Table 2), indicating that directional selection has been exclusive to the lineage-specific families of digestive proteases.

In order to interpret selection in terms of both duplication and speciation events, we used the PAML free ratios model [47] to estimate d_N/d_S along every branch in each of the three phylogenies (Figure 3). Positive selection associated with three different speciation events suggests that ongoing changes in the biochemical environment of the female reproductive tract, including possible male contributions to this environment, have resulted in adaptive evolution in some of these proteins. A total of five gene duplication events are also immediately followed by a period of positive selection in one of the paralogous branches ($d_N/d_S > 1$), indicating neofunctionalization of a duplicate gene copy. The other seven duplication events however, are followed by elevated amino acid substitution rates ($d_N/d_S = 0.2-1$) but no evidence of adaptive evolution. This suggests that relaxed constraint created by functional redundancy between paralogs has also played an important role in the evolution of these gene families.

Evidence for adaptive amino acid evolution in duplicated genes implies that selection has acted to diversify the paralogs functionally. Indeed, in all three of the protease gene families, polar, nonpolar, and charged amino acids are seen to inhabit the same selected site in different paralogs. This indicates

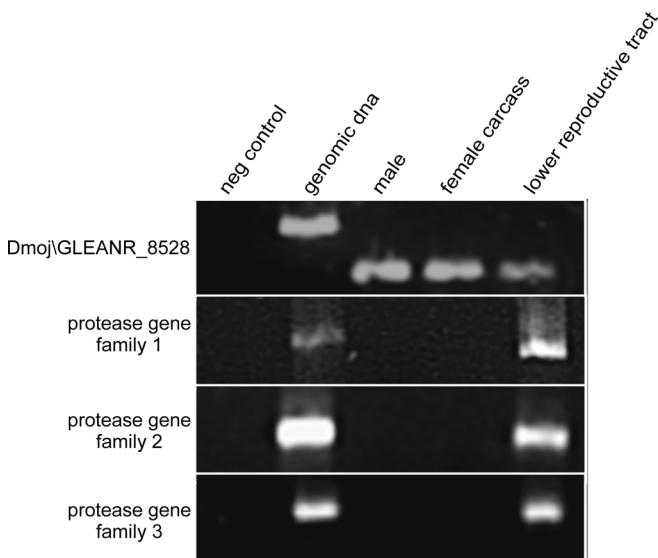


Figure 4. RT-PCR of Three Gene Families

Universal primers for each gene family were used to amplify genomic DNA, and cDNA from males, female carcasses (no lower reproductive tract), and lower reproductive tracts (for complete gels see Figure S1). doi:10.1371/journal.pgen.0030148.g004

Table 2. PAML Analysis of Positive Selection in Duplicated Proteases

Gene Family Name	Enzyme Class	Sequences Analyzed	LRT(M1vsM2)	ω	p	LRT(M8avsM8)	ω	p
Protease gene family 1	Elastase/chymotrypsin	<i>mojavensis</i> (5), <i>arizonae</i> (6)	103.96***	5.39	0.08	91.6***	4.63	0.09
Protease gene family 2	Trypsin	<i>mojavensis</i> (5), <i>arizonae</i> (4), <i>virilis</i> (1), <i>grimshawi</i> (1)	12.41***	2.86	0.04	11.52**	2.00	0.11
Protease gene family 3	Astacin	<i>mojavensis</i> (4), <i>arizonae</i> (2), <i>virilis</i> (1), <i>grimshawi</i> (1)	0.00	1.00	0.11	15.39**	11.96	0.01
Dmoj\GLEANR_12324\12325	Serine carboxypeptidase	<i>mojavensis</i> (2), <i>arizonae</i> (2), <i>virilis</i> (1)	0.00	1.00	0.05	0.00	1.00	0.00
Dmoj\GLEANR_8258\8259	Serine-type peptidase/lipase	<i>mojavensis</i> (2), <i>arizonae</i> (2), <i>virilis</i> (1)	0.00	1.00	0.05	0.01	1.00	0.02

Gene families are identified by their assigned name. Enzyme class is determined from hmmpfam [58] and SWISS-MODEL [65]. Species analyzed are indicated, followed by number of paralogs per species in parentheses. *D. mojavensis*, *D. virilis*, and *D. grimshawi* sequences were obtained from their published genomes (<http://rana.lbl.gov/drosophila>), while all *D. arizonae* sequences were found in the library. LRT denotes the value of the likelihood ratio test between the two models, followed by an indication of the statistical significance of the test. ω corresponds to the estimated highest estimated dN/dS of all site classes, and p corresponds to the proportion of sites in this class.

doi:10.1371/journal.pgen.0030148.t002

that directional selection has resulted in recurrent and radical amino acid substitutions, likely affecting the structure and function of the encoded proteins. By mapping selected sites onto predicted molecular structures, it is possible to make more specific inferences about how the biochemical function of these enzymes has been impacted by adaptive evolution. In the two families of serine endopeptidases (protease gene families 1 and 2), positive selection clusters near the catalytic triad: the three amino acids essential for proteolytic function (reviewed in [44]) (Figure 5). Furthermore, in protease gene family 1, positive selection is found adjacent to, and in one case synonymous with, three amino acid sites known to effect substrate specificity (reviewed in [48]). Collectively, these data indicate that directional selection has acted to diversify the catalytic activity of both families of serine endopeptidases, and that protease gene family 1 has concomitantly undergone adaptive evolution for increased breadth in substrate specificity. Future functional studies of these enzymes, particularly in terms of how they interact with the male ejaculate, will yield significant insight into the selective pressures that underlie diversification of these extraordinary gene families.

Implications

Our most striking result was the observation of three lineage-specific radiations of secreted digestive proteases in *D. arizonae* female reproductive tracts. Although the biological significance of these gene duplications is yet unclear, they may relate to two unusual physiologies exhibited by both *D. arizonae* and *D. mojavensis* females. First, the insemination reaction must be degraded by females prior to oviposition or remating [36], a process that could require specialized digestive machinery. Second, female incorporation of ejaculate-derived protein, as observed in *D. arizonae* and *D. mojavensis*, could be facilitated by degrading seminal proteins and/or sperm into smaller fragments that are more easily absorbed.

Regardless of their physiological function, lower female reproductive-tract specific expression of digestive enzymes points to a novel form of ejaculate-female interaction, in which females may actively degrade, rather than process or activate [13–15], protein components of the male ejaculate. Digestion of seminal proteins or sperm would undoubtedly have important implications for male reproductive success, predicting an evolutionary response from males. Indeed, the

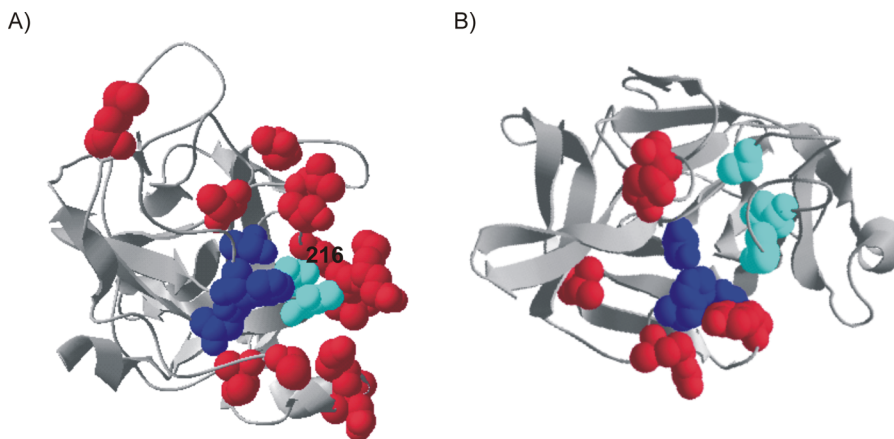


Figure 5. Structural Models Generated in SWISS-MODEL

(A) protease gene family 1 and (B) protease gene family 2. The blue amino acids comprise the catalytic triad of the active site. The aquamarine amino acids are determinants of substrate specificity [48]. The red amino acids indicate positively selected sites. The labeled amino acid in (A), 216, is a positively selected amino acid that is also a determinant of substrate specificity.

doi:10.1371/journal.pgen.0030148.g005

association of these proteases with recent gene duplications and strong signatures of adaptive evolution suggests they are involved in an intersexual arms race. Exploring the male side of this interaction, therefore, is an important avenue of future research.

The 31 candidates for adaptive evolution also have important implications for reproductive tract interactions and intersexual coevolution. Roughly half of these proteins have no known function or conserved domain, suggesting they are enriched for novel biochemical functions. Additionally, the candidates include several classes of proteins that have not been implicated previously in reproductive tract interactions. Particularly intriguing are three transmembrane proteins with the conserved transporter domain MFS₁, for inorganic solutes (Table 1). Although the biochemical composition of the *Drosophila* ejaculate is largely unknown outside of its protein constituents, females of several species incorporate ejaculate-derived phosphorus into somatic tissues and oocytes [49]. It is unclear if these transporters underlie such a process in *D. arizonae*. Their presence and evolutionary history point, however, to nonpeptide biochemical interactions in female reproductive tracts which also may evolve rapidly.

If divergence of reproductive proteins is driven by intersexual dynamics, particularly sexually antagonistic coevolution [50–52], species with more promiscuous mating systems are predicted to exhibit comparatively more adaptive evolution in their reproductive proteins. *D. arizonae* is significantly more promiscuous than its previously examined congener *D. simulans* [28], and, consistent with the prediction, we find evidence that this difference in mating system may be reflected in the evolution of their female reproductive proteins. Specifically, we observed that candidate female reproductive proteins in our dataset exhibit higher d_N/d_S values than intracellular proteins, while this effect was not seen in similar comparisons between *D. simulans* and *D. melanogaster* [28]. Additionally, the estimated frequency of adaptive evolution in *D. arizonae* female reproductive tract proteins (14%) is significantly higher (Fisher's Exact Test $p = 0.003$) than that of *D. simulans* (5%) [28]. Although the experimental approach for these two studies was quite similar, differences in divergence times between *D. arizonae* and *D. mojavensis* (~1.5 MYA, [37]), and *D. simulans* and *D. melanogaster* (~3 MYA, [53]), could result in more stochastic influence on our measures of d_N/d_S . Firm conclusions about the effect of mating system on the evolution of female reproductive proteins therefore requires further empirical testing across a broader array of taxa.

Although the function and evolution of male seminal proteins have been researched extensively in both insects and mammals, our understanding of the female reproductive tract proteins with which they interact remains sparse. Our data, as well as previous research in the *melanogaster* group [28–30], indicate that rapid evolution is common among female reproductive tract proteins. We furthermore present compelling evidence that differences in female physiology and possibly mating system between *Drosophila* species are reflected in their reproductive tract proteins. Our research indicates that female reproductive proteins are active players in reproductive tract interactions, and that rapid evolution of seminal proteins must be considered in terms of their relationship with female counterparts.

Materials and Methods

Tissue harvesting. *D. arizonae* used in this study were collected in December 2005 in Tucson, Arizona by E. S. K. A total of 873 lower reproductive tracts (parovaria, oviduct, spermathecae, seminal receptacle, and uterus) were dissected from mature adult females 9 d or older. In order to maximize transcriptional diversity obtained, dissected females were sampled from a diverse array of mating states. Of the females, 662 were from population bottles, while approximately 40 females were dissected from each of the following treatments: virgin, homospecifically mated 4–8 h postcopulation, homospecifically mated 24 h postcopulation, heterospecifically (to *D. mojavensis*) mated 4–8 h postcopulation, and heterospecifically mated 24 h postcopulation.

Library construction. The harvested tracts were pooled into four separate aliquots of TRIZOL reagent (Invitrogen, <http://www.invitrogen.com>) and total RNA was extracted according to manufacturer instructions. Quality of these samples was verified with an Agilent 2100 bioanalyzer (<http://www.home.agilent.com/>), at which point they were pooled. mRNA enrichment was achieved by binding poly-A tails on Oligotex (Qiagen, <http://www.qiagen.com/>) spin columns. Quality of enriched mRNA was verified with an Agilent 2100 bioanalyzer, and the total yield (1.5 μ g) was used for library construction with the Cloneminer cDNA library construction kit (Invitrogen). Approximately 300,000 colony-forming units were obtained with an estimated insert size of 1kb. Of these clones, 10,000 were picked with a QBOT (Genetix, <http://www.genetix.com/>) operated by the Arizona Genomics Institute (<http://www.genome.arizona.edu/>). Of these clones, 1,920 were sequenced bidirectionally, and an additional 384 were sequenced exclusively from their 5' ends. All sequencing was done on at the Arizona Genomics Institute on an ABI 3700 DNA analyzer (<https://products.appliedbiosystems.com/>) with big-dye terminator chemistry.

Sequence data analysis. Base calling and assembly were implemented in Phred and Phrap [54]. All bases with a Phred quality score below 20 (99% accurate) were excluded from further analysis. The estimated frequency of sequencing errors in included bases was 0.04%. BLASTN [55] (e -value = 0.01) against the GLEANR coding sequence annotations (from CAF1 assembly <http://rana.lbl.gov/drosophila/>) of the *D. mojavensis* genome was used to identify orthologs of *D. arizonae* ESTs. For ESTs with no good BLASTN hit to annotated coding sequence, BLASTN (e -value = 0.01) was implemented against the complete CAF1 assembly of the *D. mojavensis* genome. ESTs with BLAST hits in the *D. mojavensis* genome that contained long open reading frames were used to annotate additional genes in *D. mojavensis* by eye. No examples of ESTs with long open reading frames but no good BLASTN hit in the *D. mojavensis* genome were identified.

Translations of these coding sequences were used to identify secreted proteins and cell surface receptors using SignalP [56], and transmembrane proteins using TMHMM [57]. Conserved protein family (Pfam) domains were identified with hmmpfam [58]. Gene Ontology (GO) terms [59] were obtained from FlyBase (<http://flybase.bio.indiana.edu/>) for *D. melanogaster* homologs, or based on conserved Pfam domains if no *D. melanogaster* homolog was found. For explicit definitions of GO terms see <http://www.geneontology.org/>.

In total, the *D. arizonae* ESTs corresponded to 649 unique proteins in the *D. mojavensis* genome. The orthologous genes were aligned using CLUSTALW [60] and alignment accuracy was verified by eye. Maximum-likelihood estimates of nonsynonymous substitutions rate (d_N), synonymous substitution rate (d_S), and the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (d_N/d_S), were obtained from PAML [47]. For duplicated genes, only reciprocally monophyletic homologs were compared in pairwise analyses.

Sequence analysis of multigene families. Sequence data for *D. arizonae* was obtained from the EST library, while sequences from *D. mojavensis*, *D. virilis*, and *D. grimshawi* were obtained from their unpublished, publicly available genomes (<http://rana.lbl.gov/drosophila/>). GENECONV was used to test for gene conversion between paralogs, using the method of Sawyer [61]. Phylogenetic reconstruction of multigene families was implemented in Mr. Bayes v3.0b4. Nested maximum-likelihood models of codon evolution were implemented in the codeml program of PAML [47] and compared using likelihood ratio tests. Two tests of positive selection were performed. In the first test, the neutral model (M1) is compared with the selection model, in which a class of sites is permitted to exhibit d_N/d_S (ω) > 1 (M2). In the second test, a beta distribution of site classes in which the most rapidly evolving is fixed to $\omega = 1$ (M8a) is compared to a similar model in which the most rapidly evolving site class is

permitted to exhibit $\omega > 1$ (M8) [62]. Multiple initial values of ω were used to ensure convergence on the likelihood optima. For the second test, critical values of the test statistic are determined from Wong et al [63]. Lineage-specific selection patterns of d_N/d_S were determined by implementing branch-specific models [64].

Determination of duplication rate. A total of 34 secreted proteases were identified in *D. arizonae* female reproductive tracts. Using BLASTN homology and maximum-likelihood phylogenetic reconstruction implemented in PAUP*, we determined these 34 proteins correspond to 37 orthologs in the genome of *D. mojavensis*, and 23 orthologs in the genome of *D. virilis* (<http://rana.lbl.gov/drosophila/>). Assuming no gene conversion or gene loss, the total copy number of these genes was 23 at the divergence of the *D. mojavensis* and *D. virilis* lineages. Duplication rate can therefore be estimated by the following exponential growth equation:

$$C_M = C_A 2^{rt}$$

Where C_M is copy number of *D. mojavensis* (37), C_A is the ancestral copy number (23), t is the divergence time between *D. mojavensis* and *D. virilis* ($t = 23$ MYA [42]), and r is the estimated rate of duplication per gene per million years.

RT-PCR. *D. arizonae* RNA was extracted from 20 whole males, 70 reproductively mature females from population bottles lacking their lower reproductive tracts, and 70 lower reproductive tracts preserved in TRIZOL (Invitrogen) according to manufacturer instructions. Purified RNA was treated with DNaseI (Gibco, <http://www.invitrogen.com/>), and reverse transcribed with the iScript cDNA synthesis kit (Bio-Rad, <http://www.bio-rad.com/>). Resultant cDNA was diluted to 10 ng/ μ l, and used as a template for standard PCR using universal primers, with *D. arizonae* genomic DNA as a positive control. Primer sequences are as follows: DmojGLEANR_8528-F, 5'-AAGAAGCG-CACCAAGCACTTCATC-3'; DmojGLEANR_8528-R 5'-TCTGTTGTCGATACCTTGGGCTT-3'; protease gene family 1 -F1 5'-ATGTGGAATCTAAGCCAGCCAA-3'; protease gene family 1 -F2 5'-RTAGATGGCAGTTGCTYCTGTG-3'; protease gene family 1 -R1 5'-GATGYGATACCAATCACRGTGCT-3'; protease gene family 1 -R2 5'-ACGATRCCAATCACRGTGCTYAGA-3'; protease gene family 2 -F1 5'-CTCAAACCCGARTAGYTRCTCT-3'; protease gene family 2 -F2 5'-CTTCAAGCCGCMGTWGCCTGCTC-3'; protease gene family 2 -R1 5'-CACCRCTGTGYTCCTRATCCATTC-3'; protease gene family 2 -R2 5'-CACCGCWGTGCTCYTATCCATT-3'; protease gene family 3 -F1 5'-TGAAACCCGACTCCAGACTTATAGC-3'; protease gene family 3 -F2 5'-ATGAAACCCGACTCCGAGTTGATAG-3'; protease gene family 3 -R1 5'-ATCAGCCATGCTCAATTCTGTGCG-3'; and protease gene family 3 -R2 5'-ATCAGCCAGCTTAATTC-TAGTCG-3'.

Structural modeling. Three dimensional structure was predicted by SWISS-MODEL [65], and visualized by Deep View. Selected sites were determined from Bayes Empirical Bayes calculation [66] implemented under M8 in PAML [47].

References

- Swanson WJ, Vacquier VD (2002) The rapid evolution of reproductive proteins. *Nat Rev Genet* 3: 137–144.
- Panhuis TM, Clark NL, Swanson WJ (2006) Rapid evolution of reproductive proteins in abalone and *Drosophila*. *Philos Trans R Soc Lond B Biol Sci* 361: 261–8.
- Clark NL, Aagaard JE, Swanson WJ (2006) Evolution of reproductive proteins from animals and plants. *Reproduction* 131: 11–22.
- Lee YH, Ota T, Vacquier VD (1995) Positive selection is a general phenomenon in the evolution of abalone sperm lysin. *Mol Biol Evol* 12: 231–238.
- Yang Z, Swanson WJ, Vacquier VD (2000) Maximum-likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. *Mol Biol Evol* 17: 1446–55.
- Galindo BE, Moy GW, Swanson WJ, Vacquier VD (2002) Full-length sequence of VERL, the egg vitelline envelope receptor for abalone sperm lysin. *Gene* 288: 111–117.
- Galindo BE, Vacquier VD, Swanson WJ (2003) Positive selection in the egg receptor for abalone sperm lysin. *Proc Natl Acad Sci U S A* 100: 4639–4643.
- Roberston SA (2007) Seminal fluid signaling in the female reproductive tract: Lessons from rodents and pigs. *J Anim Sci* 85: E36–E44.
- Wolfner MF (2002) The gifts that keep on giving: Physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. *Heredity* 88: 85–93.
- Kubli E (2003) Sex-peptides: Seminal peptides of the *Drosophila* male. *Cell Mol Life Sci* 60: 1689–1704.

Supporting Information

Figure S1. RT-PCR of (A) DmojGLEANR_8528, (B) Protease Gene Family 1, (C) Protease Gene Family 2, and (D) Protease Gene Family 3. One-kilobase markers are indicated. L, DNA ladder; N, negative control; M, whole male cDNA; C, female carcass (no lower reproductive tract) cDNA; R, lower female reproductive tract cDNA. Found at doi:10.1371/journal.pgen.0030148.sg001 (7.8 MB TIF).

Table S1. *D. arizonae* Female Reproductive Tract ESTs

D. mojavensis CDS: coding sequence from GLEANR annotations (<http://rana.lbl.gov/>), *D. melanogaster* homolog identified by BLAST. SignalP: S, secreted; A, anchor; Q, quiescent as predicted by SignalP 3.0 [56]; TMHMM, number of identified transmembrane domains [57]; Ka, estimated nonsynonymous substitutions per nonsynonymous site; Ks, estimated synonymous substitutions per synonymous site; Ka/Ks, estimated ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site; PROT %ID, Protein % identity; CDS %ID, coding sequence % identity calculated in PAML [47]; Conserved domain, pfam conserved domain predicted from hmmpfam [58].

Found at doi:10.1371/journal.pgen.0030148.st001 (171 KB XLS).

Accession Numbers

All sequences for this study are available from the National Institute for Biotechnology Information (NCBI) GenBank (<http://www.ncbi.nlm.nih.gov/Entrez/index.html>) accession numbers EV41299147751410-EV41383447752253

Acknowledgments

The authors would like to acknowledge Luciano Matzkin and James Pennington for helpful discussion, and Jeff Good, Matt Dean, Gabriela Wlasiuk, and four anonymous reviewers for generous comments on this manuscript.

Author contributions. ESK, WJS, and TAM conceived and designed the experiments. ESK performed the experiments, analyzed the data, and wrote the paper. ESK and TAM contributed reagents/materials/analysis tools.

Funding. This research was funded by the University of Arizona, and the NSF-IGERT program in Evolutionary, Functional and Computational Genomics at the University of Arizona. ESK was supported by a National Science Foundation Integrative Graduate Education and Research Traineeship in evolutionary, functional, and computational genomics at the University of Arizona. WJS was supported by NIH grant HD42563.

Competing interests. The authors have declared that no competing interests exist.

- Chapman T, Davies SJ (2004) Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. *Peptides* 25: 1477–1490.
- Markow TA (1996) Evolution of *Drosophila* mating systems. *Evol Biol* 29: 73–106.
- Monsma SA, Harada HA, Wolfner MF (1990) Synthesis of two *Drosophila* male accessory gland proteins and their fate after transfer to the female during mating. *Dev Biol* 142: 465–475.
- Park M, Wolfner MF (1995) Male and female cooperate in the prohormone-like processing of a *Drosophila melanogaster* seminal fluid protein. *Dev Biol* 171: 694–702.
- Peng J, Chen S, Busser S, Liu H, Honegger T, Kubli E (2005) Gradual release of sperm bound sex-peptide controls female postmating behavior in *Drosophila*. *Curr Biol* 15: 207–213.
- Bertram MJ, Neubaum DM, Wolfner MF (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.
- Heifetz Y, Lung O, Frongillo EA Jr, Wolfner MF (2000) The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. *Curr Biol* 10: 99–102.
- Ravi Ram K, Ji S, Wolfner MF (2005) Fates and targets of male accessory gland proteins in mated female *Drosophila melanogaster*. *Insect Biochem Mol Biol* 35: 1059–1071.
- Civetta A, Singh RS (1995). High divergence of reproductive tract proteins and their association with postzygotic reproductive isolation in *Drosophila melanogaster* and *Drosophila virilis* group species. *J Mol Evol* 41: 1085–1095.
- Begun DJ, Lindfors HA (2005) Rapid evolution of genomic Acp complement in the *melanogaster* subgroup of *Drosophila*. *Mol Biol Evol* 22: 2010–2021.

21. Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, Wolfner MF (2005) Cross-species comparison of *Drosophila* male accessory gland protein genes. *Genetics* 171: 131–143.
22. Begun DJ, Lindfors HA, Thompson ME, Holloway AK (2006) Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172: 1675–1681.
23. Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG (2000) Molecular population genetics of male accessory gland proteins in *Drosophila*. *Genetics* 156: 1879–1888.
24. Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF (2001) Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proc Nat Acad Sci U S A* 13: 7375–7379.
25. Kern AD, Jones CD, Begun DJ (2004) Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. *Genetics* 167: 725–735.
26. Schully SD, Hellberg ME (2006) Positive Selection on Nucleotide Substitutions and Indels in Accessory Gland Proteins of the *Drosophila pseudoobscura* Subgroup. *J Mol Evol* 62: 793–802.
27. Wagstaff BJ, Begun DJ (2005) Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. *Genetics* 171: 1083–101.
28. Swanson WJ, Wong A, Wolfner MF, Aquadro CF (2004) Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168: 1457–1465.
29. Panhuis T, Swanson WJ (2006) Molecular evolution and population genetics of candidate female reproductive genes in *Drosophila*. *Genetics* 173: 2039–2047.
30. Lawniczak MK, Begun DJ (2007) Molecular population genetics of female-expressed mating-induced serine proteases in *Drosophila melanogaster*. *Mol Biol Evol*. E-pub 14 June 2007.
31. Jagadeeshan S, Singh RS (2007) Rapid evolution of outer egg membrane proteins in the *Drosophila melanogaster* subgroup: A case of ecologically driven evolution of female reproductive traits. *Mol Biol Evol* 24: 929–938.
32. Markow TA (2002) Female remating, operational sex ratio, and the arena of sexual selection in *Drosophila*. *Evolution* 56: 1725–1734.
33. Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nat Genet* 36: 1326–1329.
34. Markow TA, Ankney P (1988) Insemination reaction in *Drosophila*: A copulatory plug in species showing male contribution to offspring. *Evolution* 42: 1097–1100.
35. Markow TA, Gallagher PD, Krebs RA (1990) Ejaculate-derived nutritional contribution and female reproductive success in *Drosophila mojavensis* (Patterson and Crow. *Func Ecol* 4: 67–73.
36. Patterson JT (1946) A new type of isolating mechanism in *Drosophila*. *Proc Nat Acad Sci U S A* 32: 202–208.
37. Matzkin LM (2004) Population genetics and geographic variation of alcohol dehydrogenase (Adh) paralogs and glucose-6-phosphate dehydrogenase (G6pd) in *Drosophila mojavensis*. *Mol Biol Evol* 21: 276–285.
38. Pitnick S, Miller GT, Schneider K, Markow TA (2003) Ejaculate-female coevolution in *Drosophila mojavensis*. *Proc Nat Acad Sci U S A* 270: 507–1512.
39. Knowles LL, Markow TA (2001) Sexually antagonistic coevolution of a postmating prezygotic reproductive character in desert *Drosophila*. *Proc Nat Acad Sci U S A* 98: 8692–8696.
40. Knowles LL, Hernandez BB, Markow TA (2005) Exploring the consequences of postmating-prezygotic interactions between the sexes. *Proc Biol Sci* 271 Suppl 5: S357–S359.
41. Knowles LL, Hernandez BB, Markow TA (2005) Non-antagonistic interactions between the sexes revealed by the ecological consequences of reproductive traits. *J Evol Biol* 18: 156–161.
42. Powell JR (1997) Progress and prospects in evolutionary biology: The *Drosophila* model. New York: Oxford University Press. 576 p.
43. Gu Z, Cavalcanti A, Chen FC, Bouman P, Li WH (2002) Extent of gene duplication in the genomes of *Drosophila*, nematode, and yeast. *Mol Biol Evol* 19: 256–62.
44. Neurath H (1984) Evolution of proteolytic enzymes. *Science* 224: 350–357.
45. Ross J, Jiang H, Kanost MR, Wang Y (2003) Serine proteases and their homologs in the *Drosophila melanogaster* genome: An initial analysis of sequence conservation and phylogenetic relationships. *Gene* 304: 117–31.
46. Stocker W, Zwilling R (1995) Astacin. *Methods Enzymol* 248: 305–25.
47. Yang Z (1997) PAML: A program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13: 555–556.
48. Sprang SR, Fletterick RJ, Graf L, Rutter WJ, Craik CS (1988) Studies of specificity and catalysis in trypsin by structural analysis of site-directed mutants. *Crit Rev Biotechnol* 8: 225–36.
49. Markow TA, Coppola A, Watts TD (2001) How *Drosophila* males make eggs: It is elemental. *Proc Biol Sci* 268: 1527–1532.
50. Rice WR (1996) Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* 381: 232–4.
51. Gavrilets S (2000) Rapid evolution of reproductive barriers driven by sexual conflict. *Nature* 403: 886–889.
52. Hayashi TI, Vose M, Gavrilets S (2007) Genetic differentiation by sexual conflict. *Evolution*. 61: 516–29.
53. Hey J, Kliman RM (1993) Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. *Mol Biol Evol* 10: 804–822.
54. Ewing B, Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* 8: 186–194.
55. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215: 403–410.
56. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340: 783–795.
57. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6: 175–182.
58. Eddy S. R. (1998) Profile hidden Markov models. *Bioinformatics* 14: 755–763.
59. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25: 25–29.
60. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680.
61. Sawyer SA (1989) Statistical tests for detecting gene conversion. *Mol Biol Evol* 6: 526–538.
62. Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20: 18–20.
63. Wong WS, Yang Z, Goldman N, Nielsen R (2004) Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168: 1041–1051.
64. Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* 15: 568–573.
65. Schwede T, Kopp J, Guex N, Peitsch MC (2003) SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res* 31: 3381–3385.
66. Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol Biol Evol* 22: 1107–1118.