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

Erin S. Kelleher1*, 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

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 of information suggests that the female side of reproductive tract interactions has been underappreciated, and that the female reproductive tract provides a fascinating model system for exploring a novel coevolutionary mechanism of ejaculate–female interaction.

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 reproductive 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.

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 indicates 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_{ns}/d_{sy}) between our D. arizonae ESTs and their orthologs in the D. mojavensis genome. Candidate female reproductive proteins exhibit significantly larger d_{ns}/d_{sy} values than intracellular proteins in our dataset (median test, p >
of D. mojavensis, 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 female reproductive tract proteins. Specifically, 47% (16) of all secreted proteases in D. arizonae are 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 was 50%. Of 227 pairwise comparisons, 31 (14%) were identified with dN/dS > 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.

**Table 1. Candidate Female Reproductive Proteins**

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Signal</th>
<th>Transmembrane</th>
<th>dbN</th>
<th>dbS</th>
<th>dbN/dbS</th>
<th>D. melanogaster</th>
<th>D. melanogaster</th>
<th>Conserved</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. mojavensis CDS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GLEANR_9982</td>
<td>S</td>
<td>0</td>
<td>0.03</td>
<td>0.02</td>
<td>1.83</td>
<td>CG7443-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>anon-EST:Kelleher15</td>
<td>S</td>
<td>1</td>
<td>0.03</td>
<td>0.02</td>
<td>1.38</td>
<td></td>
<td>No hits</td>
<td>NA</td>
</tr>
<tr>
<td>GLEANR_5396</td>
<td>S</td>
<td>1</td>
<td>0.02</td>
<td>0.02</td>
<td>1.30</td>
<td>Gp150-PD</td>
<td>Cell adhesion/signal transduction</td>
<td>LRR_1</td>
</tr>
<tr>
<td>GLEANR_4627</td>
<td>Q</td>
<td>10</td>
<td>0.03</td>
<td>0.03</td>
<td>1.28</td>
<td>CG30344-PA</td>
<td>Unknown function</td>
<td>MFS_1</td>
</tr>
<tr>
<td>anon-EST:Kelleher5</td>
<td>S</td>
<td>0</td>
<td>0.05</td>
<td>0.04</td>
<td>1.20</td>
<td>CG10472</td>
<td>Proteolysis</td>
<td>Trypsin</td>
</tr>
<tr>
<td>GLEANR_17128</td>
<td>S</td>
<td>2</td>
<td>0.06</td>
<td>0.06</td>
<td>1.08</td>
<td>No hits</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_5358</td>
<td>Q</td>
<td>1</td>
<td>0.01</td>
<td>0.01</td>
<td>1.04</td>
<td>CG30415-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_1967</td>
<td>S</td>
<td>1</td>
<td>0.05</td>
<td>0.05</td>
<td>0.96</td>
<td>CG7778-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_896</td>
<td>S</td>
<td>0</td>
<td>0.11</td>
<td>0.12</td>
<td>0.89</td>
<td>CG31954-PA</td>
<td>Proteolysis</td>
<td>Trypsin</td>
</tr>
<tr>
<td>GLEANR_17617</td>
<td>S</td>
<td>4</td>
<td>0.02</td>
<td>0.03</td>
<td>0.82</td>
<td>CG4729-PA</td>
<td>Metabolism</td>
<td>Acrystase</td>
</tr>
<tr>
<td>GLEANR_3367</td>
<td>A</td>
<td>0</td>
<td>0.02</td>
<td>0.03</td>
<td>0.82</td>
<td>Lectin-146cb-PA</td>
<td>Sugar binding</td>
<td>Lectin_C</td>
</tr>
<tr>
<td>GLEANR_5037</td>
<td>A</td>
<td>4</td>
<td>0.01</td>
<td>0.02</td>
<td>0.80</td>
<td>CG15098-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_9209</td>
<td>A</td>
<td>2</td>
<td>0.01</td>
<td>0.01</td>
<td>0.80</td>
<td>Nsp2-P</td>
<td>Proteolysis</td>
<td>Peptidase_M13_N, Peptidase_M13</td>
</tr>
<tr>
<td>GLEANR_13559</td>
<td>Q</td>
<td>12</td>
<td>0.02</td>
<td>0.03</td>
<td>0.79</td>
<td>CG10960-PB</td>
<td>Metabolism/transport</td>
<td>Sugar_tr, MFS_1</td>
</tr>
<tr>
<td>GLEANR_2703</td>
<td>S</td>
<td>0</td>
<td>0.09</td>
<td>0.12</td>
<td>0.77</td>
<td>CG15254-PA</td>
<td>Proteolysis</td>
<td>Astacin</td>
</tr>
<tr>
<td>GLEANR_9617</td>
<td>S</td>
<td>0</td>
<td>0.07</td>
<td>0.09</td>
<td>0.74</td>
<td>CG3739-PA</td>
<td>Proteolysis</td>
<td>Peptidase_S28</td>
</tr>
<tr>
<td>GLEANR_12094</td>
<td>S</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
<td>0.73</td>
<td>CG6409-PA</td>
<td>Gpl anchor biosynthesis</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_10002</td>
<td>S</td>
<td>0</td>
<td>0.04</td>
<td>0.05</td>
<td>0.70</td>
<td>CG9418-PA</td>
<td>DNA binding</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_1209</td>
<td>A</td>
<td>1</td>
<td>0.04</td>
<td>0.06</td>
<td>0.69</td>
<td>No hits</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_6649</td>
<td>S</td>
<td>0</td>
<td>0.01</td>
<td>0.02</td>
<td>0.67</td>
<td>LIG1913-PB</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_7394</td>
<td>S</td>
<td>0</td>
<td>0.02</td>
<td>0.03</td>
<td>0.66</td>
<td>CG17271-PA</td>
<td>Ca2+ Binding</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_10683</td>
<td>S</td>
<td>0</td>
<td>0.03</td>
<td>0.04</td>
<td>0.65</td>
<td>CG5630-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_6054</td>
<td>Q</td>
<td>2</td>
<td>0.01</td>
<td>0.02</td>
<td>0.62</td>
<td>CG4627-PA</td>
<td>Unknown function</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_2775</td>
<td>Q</td>
<td>10</td>
<td>0.03</td>
<td>0.05</td>
<td>0.59</td>
<td>CG4726-PA</td>
<td>Metabolism/transport</td>
<td>MFS_1</td>
</tr>
<tr>
<td>anon-EST:Kelleher9</td>
<td>S</td>
<td>0</td>
<td>0.04</td>
<td>0.06</td>
<td>0.59</td>
<td>CG10472</td>
<td>Proteolysis</td>
<td>Trypsin</td>
</tr>
<tr>
<td>GLEANR_1245</td>
<td>Q</td>
<td>1</td>
<td>0.03</td>
<td>0.05</td>
<td>0.57</td>
<td>No hits</td>
<td>NA</td>
<td>Lamp</td>
</tr>
<tr>
<td>GLEANR_16396</td>
<td>S</td>
<td>0</td>
<td>0.05</td>
<td>0.09</td>
<td>0.55</td>
<td>No hits</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>GLEANR_2881</td>
<td>Q</td>
<td>1</td>
<td>0.03</td>
<td>0.05</td>
<td>0.54</td>
<td>CG14536-PB</td>
<td>Unknown function</td>
<td>Ubiquitin</td>
</tr>
<tr>
<td>GLEANR_8258</td>
<td>S</td>
<td>0</td>
<td>0.03</td>
<td>0.06</td>
<td>0.53</td>
<td>CG3734-PA</td>
<td>Proteolysis</td>
<td>Peptidase_S28</td>
</tr>
<tr>
<td>GLEANR_7051</td>
<td>S</td>
<td>0</td>
<td>0.05</td>
<td>0.09</td>
<td>0.52</td>
<td>CG6283-PA</td>
<td>Lipid metabolism</td>
<td>Lipase</td>
</tr>
<tr>
<td>GLEANR_3613</td>
<td>S</td>
<td>0</td>
<td>0.02</td>
<td>0.04</td>
<td>0.51</td>
<td>CG18067-PB</td>
<td>Unknown function</td>
<td>No</td>
</tr>
</tbody>
</table>

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
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 2. Distribution of Three Protease Gene Families in D. mojavensis and D. virilis Genomes](http://rana.lbl.gov/drosophila/).](http://rana.lbl.gov/drosophila/)
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 \textit{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 \textit{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
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 endoproteases, 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

<table>
<thead>
<tr>
<th>Gene Family Name</th>
<th>Enzyme Class</th>
<th>Sequences Analyzed</th>
<th>LRT(M1vsM2)</th>
<th>$\omega$</th>
<th>$p$</th>
<th>LRT(M8vsM8)</th>
<th>$\omega$</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protease gene family 1</td>
<td>Elastase/chymotrypsin</td>
<td><em>mojavensis</em> (5), <em>arizonae</em> (6)</td>
<td>103.96***</td>
<td>5.39</td>
<td>0.08</td>
<td>91.6**</td>
<td>4.63</td>
<td>0.09</td>
</tr>
<tr>
<td>Protease gene family 2</td>
<td>Trypsin</td>
<td><em>mojavensis</em> (5), <em>arizonae</em> (4), <em>virilis</em> (1), <em>grimshawi</em> (1)</td>
<td>12.41***</td>
<td>2.86</td>
<td>0.04</td>
<td>11.52**</td>
<td>2.00</td>
<td>0.11</td>
</tr>
<tr>
<td>Protease gene family 3</td>
<td>Astacin</td>
<td><em>mojavensis</em> (4), <em>arizonae</em> (2), <em>virilis</em> (1), <em>grimshawi</em> (1)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.11</td>
<td>15.39**</td>
<td>11.96</td>
<td>0.01</td>
</tr>
<tr>
<td>Dmoj\GLEANR_12324\12325</td>
<td>Serine carboxypeptidase</td>
<td><em>mojavensis</em> (2), <em>arizonae</em> (2), <em>virilis</em> (1)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.05</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Dmoj\GLEANR_8258\8259</td>
<td>Serine-type peptidase/lipase</td>
<td><em>mojavensis</em> (2), <em>arizonae</em> (2), <em>virilis</em> (1)</td>
<td>0.00</td>
<td>1.00</td>
<td>0.05</td>
<td>0.01</td>
<td>1.00</td>
<td>0.02</td>
</tr>
</tbody>
</table>

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.  $\omega$ corresponds to the estimated highest estimated dN/dS of all site classes, and  $p$ corresponds to the proportion of sites in this class.

**Table 2. PAML Analysis of Positive Selection in Duplicated Proteases**

**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.t002
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 MFS_1, 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.

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 receptacles, and uterus) were dissected from mature 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 sets, the result of two replications of two *D. arizonae* colonies (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 μg) 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 1 kb. 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 cloning was done at the Arizona Genomics Institute on an ABI 3700 DNA analyzer (http://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 GLEAN 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 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 hit 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 (dN), synonymous substitution rate (dS), and the ratio of nonsynonymous substitutions per nonsynonymous site to synonymous substitutions per synonymous site (dN/dS) 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://trana.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 MrBayes v3.0b4. Nested maximum-likelihood models of codon evolution were implemented in the codem 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 selective model, in which a class of sites is permitted to exhibit dN/dS \(>1\) (M2). In the second test, a beta distribution of site classes in which the most rapidly evolving is fixed to \(n=1\) (M8a) is compared to a similar model in which the most rapidly evolving site class is
permitted to exhibit $\omega > 1$ (MB) [62]. Multiple initial values of $\omega$ 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 $d_{xy}$, were determined by implementing branch-specific models [64].

**Determination of duplication rate.** A total of 34 secreted proteases were selected from *D. arizonae* female reproductive tracts. Using BLASTN homology and maximum-likelihood phylogenetic reconstruction implemented in PAUP*, we determined these 34 proteases correspond to 37 orthologs in the genome of *D. mojavensis*, and 23 orthologs in the genome of *D. viridis* (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. viridis* lineages. Duplication rate can therefore be estimated by the following exponential growth equation:

$$C_M = C_A e^{r t}$$

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. viridis* ($t = 23$ MYA [42]), and $r$ is the estimated rate of duplication per million years.

**RT-PCR.** *D. arizonae* RNA was extracted from 20 whole males, 70 reproducitively 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) transcribed with the Superscript II RNA synthesis kit (Bio-Rad, http://www.bio-rad.com/). Resultant cDNA was diluted to 10 ng/µl, 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'-AAGAACGG-CACCAAAGCTTCTGTCG-3'; DmojGLEANR_8528-R 5'-TCTGTTGTCGATACCCTTGGGCTT-3'; DmojGLEANR_8528-F1 5'-TCTGTTGTCGATACCCTTGGGCTT-3'; DmojGLEANR_8528-F2 5'-ATGTGGAATCTAAGCCCAGCCAA-3'; DmojGLEANR_8528-F3 5'-CTCAAACCGCARTAGYTRTCCT-3'; protease gene family 1-F1 5'-ATGAAACCGATCCGAGTTGATAG-3'; protease gene family 1-R1 5'-GATYGAATACCACTACGRGTCT-3'; protease gene family 1-R2 5'-ACGATRCAATCAGCRGTGAGA-3'; protease gene family 2-F1 5'-CITCAACCGCARTAGYTRTCCT-3'; protease gene family 2-F2 5'-CITCAACCGCARTAGYTRTCCT-3'; protease gene family 2-R1 5'-CACRCCRTGTYTTCRCRTTACCTC-3'; protease gene family 2-R2 5'-CACCAGCGTGTCTYTYTTATCATT-3'; protease gene family 3-F1 5'-TGAAACCGATCCGAGTTGATAG-3'; protease gene family 3-F2 5'-ATGAACCCCATCAGCGATTTAG-3'; protease gene family 3-R1 5'-ATGAAACCGATCCGAGTTGATAG-3'; protease gene family 3-R2 5'-ATGAAACCGATCCGAGTTGATAG-3'; protease gene family 4-F1 5'-ATGAACCCCATCAGCGATTTAG-3'; protease gene family 4-R1 5'-ATGAACCCCATCAGCGATTTAG-3'; protease gene family 4-R2 5'-ATGAACCCCATCAGCGATTTAG-3'; protease gene family 4-R3 5'-ATGAACCCCATCAGCGATTTAG-3'.

**Structural modeling.** Three dimensional structure was predicted by SWISS-MODEL [65], and visualized by Deep View. Selected sites were determined using the M8 in PAML [47].

**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

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

**Acknowledgments**

The authors would like to acknowledge Luciano Matzkin and James Pennington for helpful discussion, and Jeff Good, Matt Dean, Gabriela Wasiuk, 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.

**References**


**PloS Genetics | www.plosgenetics.org August 2007 | Volume 3 | Issue 8 | e148**


