# CHAPTER TWO
# CROSS-SPECIES COMPARISON OF DROSOPHILA MALE ACCESSORY GLAND PROTEIN GENES[1]


## Abstract

*Drosophila melanogaster* males transfer seminal fluid proteins along with sperm during mating. Among these proteins, Acps (Accessory gland proteins) from the male's accessory gland induce behavioral, physiological and lifespan reduction in mated females and mediate sperm storage and utilization. A previous evolutionary EST screen in *D. simulans* identified partial cDNAs for 57 new candidate Acps. Here we report the annotation and confirmation of the corresponding Acp genes in *D. melanogaster*. Of 57 new candidate Acp genes previously reported in *D. melanogaster*, 34 conform to our more stringent criteria for encoding putative male accessory gland extracellular proteins, thus bringing the total number of Acps identified to 52 (34 + 18 previously identified). This comprehensive set of Acps genes allows us to dissect the patterns of evolutionary change in a suite of proteins from a single male-specific reproductive tissue. We used sequence based analysis to examine codon bias, gene duplications and levels of divergence (via dN/dS values and ortholog detection) of the 52 *D. melanogaster* Acps in *D. simulans*, *D. yakuba*, and *D. pseudoobscura*. We show that 58% of the 52 *D. melanogaster* Acp genes are detectable in *D. pseudoobscura*. Sequence comparisons of Acps shared and not shared between *D. melanogaster* and *D. pseudoobscura* show that there are separate classes undergoing distinctly dissimilar evolutionary dynamics.


[1] The work in this chapter has been accepted in the journal Genetics and is in press; Mueller, J.L., Ravi Ram, K., McGraw, L.A., Bloch Qazi, M.C., Siggia, E.D., Clark, A.G., Aquadro, C.F., Wolfner, M.F., Cross-species comparison of *Drosophila* male accessory gland protein genes. My experimental contribution to this work includes the annotation and cloning of 30 and 29 *D. melanogaster* Acps, respectively. I have also performed all the evolutionary sequence analysis of the *D. melanogaster* Acps and their orthologs in *D. simulans* and *D. yakuba*. Lastly, I performed the X chromosome Acp sequence analysis in *D. melanogaster*. 

## Introduction

Acps induce a variety of physiological, behavioral and reproductive changes when transferred to the female. Between 25-150 Acps were initially thought to be transferred to the female during mating (Coulthart and Singh, 1988; Ingman-Baker and Candido, 1980; Schmidt, 1985; Whalen and Wilson, 1986; Wolfner et al., 1997). Males lacking Acps have impaired fertility, indicating that Acps perform important reproductive functions (Kalb et al., 1993; Xue and Noll, 2000). Specifically, Acps cause females to increase their egg-production, egg-laying and ovulation rates, decrease their propensity to remate, store and utilize sperm (reviewed in Chapman and Davies, 2004; Wolfner, 2002). Acps also participate in formation of the mating plug (Lung and Wolfner, 2001) and mediate a decrease in the mated female's lifespan (Chapman et al., 1995). Genetic analyses have revealed the functions of four Acps thus far. Acp26Aa (ovulin) is a prohormone that triggers an increase in ovulation rate (Heifetz et al., 2000; Herndon and Wolfner, 1995). Acp36DE is a glycoprotein that is essential for sperm storage (Neubaum and Wolfner, 1999), by regulating sperm accumulation into storage (Bloch Qazi and Wolfner, 2003). Acp70A (sex peptide) induces egg-laying and decreases females' receptivity to remating; it also contributes to the cost of mating to females (Aigaki et al., 1991; Chapman et al., 2003; Chen et al., 1988; Liu and Kubli, 2003; Wigby and Chapman, 2005). Acp62F is a trypsin protease inhibitor that localizes to the sperm storage organs of mated females and has been suggested to preserve sperm viability (Lung et al., 2002). Acp62F also enters the female's circulation and is toxic to flies upon repeated ectopic expression, suggesting a possible role in the lifespan cost-of-mating (Lung et al., 2002). In addition, the transfer of antimicrobial Acps to the female (Lung et al., 2001) and the Acp-induced up-regulation of antimicrobial peptides in mated females (Lawniczak and Begun, 2004; McGraw et al., 2004) suggests that Acps may contribute to a female's immune

defense. Altogether, Acps appear to participate in a complex set of interactions by competing/cooperating with seminal fluid proteins of other males (Clark et al., 1995; Clark et al., 1999; Prout and Clark, 1996; Snook and Hosken, 2004), receptors present in the female or on sperm, and pathogens. To better understand this diverse set of interactions of Acps it is important to fully characterize the Acps involved and examine their evolutionary dynamics.

Initially, 18 *Drosophila melanogaster* Acps had been identified from multiple screens (Chen et al., 1988; Simmerl et al., 1995; Wolfner et al., 1997); however this was far below the predicted 25-150 Acps (Civetta and Singh, 1995; Ingman-Baker and Candido, 1980; Schmidt, 1985; Whalen and Wilson, 1986; Wolfner et al., 1997). In an extensive screen (Swanson et al., 2001a), 57 new candidate Acps were identified from partial gene sequencing of ESTs obtained from a *D. simulans* accessory gland cDNA library. These 57 candidate Acps, plus the 18 previously identified, led to 75 putative Acps. Statistical analysis of the frequency of multiple isolates predicted that these genes represented approximately 90% of the total number of Acp genes (Swanson et al., 2001a). The SWANSON *et al*. (2001a) EST screen identified Acps from partial gene sequencing and from a species in which genetic analysis is not routine, *D. simulans*. Because it is important to obtain the complete sequence of these genes in a species in which genetic analysis is possible, we obtained and report here the *D. melanogaster* orthologs of the 57 *D. simulans* Acp candidates. Our RT-PCR and bioinformatic analyses determined that 34 of the candidate 57 Acps identified by SWANSON *et al*. (2001a) have sequences suggestive of encoding extracellular proteins and expression patterns suggestive of encoding Acps. This resets the total number of *D. melanogaster* Acps identified to 52 (34 + 18 previously identified).

An unusually high fraction of the genes encoding Acps show signs of positive selection (Aguadé, 1999; Aguadé et al., 1992; Begun et al., 2000; Cirera and Aguadé,

1997; Kern et al., 2004; Kohn et al., 2004; Panhuis et al., 2003; Stevison et al., 2004; Tsaur and Wu, 1997). Acps, as a class, evolve at about twice the rate of non-reproductive proteins (Civetta and Singh, 1995; Swanson et al., 2001a; Whalen and Wilson, 1986). Swanson *et al*. (2001a) found that approximately 11% of the partially sequenced ESTs they identified have an excess of non-synonymous over synonymous nucleotide changes, suggesting divergence of these genes is being accelerated by positive selection. Three selective forces are predicted to drive the generation of sequence diversity of Acps: female sperm preference (Eberhard and Cordero, 1995), sperm competition (Clark et al., 1995), and sexual conflict (Rice, 1984). Previous evolutionary analyses of Acps focused on some of the initially identified 18 Acps (Aguadé, 1999; Aguadé et al., 1992; Begun et al., 2000; Cirera and Aguadé, 1997; Kern et al., 2004; Tsaur and Wu, 1997). Here we present a detailed examination of the molecular evolution of the entire set of stringently selected and annotated 52 Acps. We performed sequence based comparisons of these *D. melanogaster* Acps with their orthologs in three *Drosophila* species (*D. simulans*, *D. yakuba*, and *D. pseudoobscura*). This allowed us to determine levels of codon bias, rates of gene duplication, and levels of sequence divergence among three members of the *D. melanogaster* subgroup (*D. melanogaster*, *D. simulans*, *and D. yakuba*) and, via ortholog detection, which Acps are conserved between *D. melanogaster* and *D. pseudoobscura*. These evolutionary analyses demonstrate that Acps represent a combination of divergent and conserved proteins which undergo different patterns of sequence evolution.

## Materials and Methods

*Annotation of* D. melanogaster *orthologs of D. simulans Acp-ESTs*

We sequenced *D. simulans* Acp ESTs (Swanson et al., 2001a) from their 3′ ends to determine the translational stop position. This in combination with previously sequenced 5′ end sequences (Swanson et al., 2001a), provided each candidate Acp's complete ORF. The complete EST sequences can be found under Genbank Accession numbers DQ088689-DQ088699 and DQ079991-DQ079998. These *D. simulans* EST sequences were subsequently aligned using Sequencher 4.0.5 (Gene Codes) to the *D. melanogaster* genome (Release 4.0) (Celniker et al., 2002) to identify their *D. melanogaster* orthologs. Each translational start was located by the presence of sequences encoding a predicted signal peptide, either from the BDGP *D. melanogaster* annotation, or via manual inspection if the BDGP annotation did not match the *D. simulans* EST. Manual searches for predicted signal peptides constituted scanning ~1.5kb of noncoding upstream *D. melanogaster* sequence from the 5′ end of the *D. simulans* EST or BDGP predicted translational start site. Predicted signal peptides were identified using SignalP (Richards et al., 2005).

Candidate Acp genes were then examined for accessory gland-specific expression. Eighteen previously identified Acps (Chen et al., 1988; Simmerl et al., 1995; Wolfner et al., 1997) were already known to show accessory gland-predominant or -exclusive expression. We searched each of the 57 new candidate Acps identified by Swanson *et al*. (2001a) against the *D. melanogaster* BDGP EST database (http://www.fruitfly.org/EST/EST.shtml) to see if the gene was expressed in other tissues (e.g. head, embryo, tissue culture). Occasionally adult testis ESTs (Rubin et al., 2000a) included our Acp candidates. For example, Acp36DE (a highly expressed Acp (Wolfner et al., 1997)) has 32 testes EST hits, but has been shown by Western blots to be an accessory gland-specific protein (Bertram et al., 1996; Wolfner et al.,

1997). These differences may result from low-level contamination by accessory gland fragments or cells in the large scale testes preparations for the EST project or may indicate that Acp36DE is transcribed, but not translated, in the testes. Consistent with such models, all 15 Acp antibodies thus far generated detect exclusively accessory gland-specific proteins, even though nine of fifteen genes (CG8982 (Acp26Aa), CG4605 (Acp32CD), CG7157 (Acp36DE), CG6289, CG8137, CG9334, CG17575, CG1656, CG9029) (Bertram et al., 1996; Coleman et al., 1995; Lung et al., 2002; Monsma et al., 1990; Ravi Ram et al., 2005) have testis EST hits (Andrews et al., 2000; Parisi et al., 2004; Rubin et al., 2000a).

Of the 57 Acp candidates previously selected by Swanson *et al*. (2001a), seven were eliminated from further study because mutational analysis (per FlyBase (http://flybase.net/)) indicates their phenotypes affect non-reproductive processes. Sixteen further candidates were removed because they either had EST hits in multiple non-reproductive tissue types or could not be annotated, thus leaving 34 candidate Acps (see http:// www.genetics.org/supplemental/) for list of Acps removed from the previous 57 candidates). It is important to note that secreted proteins found in other tissues, or whose mutants have additional non-reproductive phenotypes may be present in accessory gland secretions. However, we focused on accessory gland-specific candidates since the evolutionary pressures and functions of these genes should be more comprehensible than those of genes expressed in multiple tissues and thus likely having multiple functions.

This selection process resulted in a collection of 52 Acps (Table 2.1). Twelve (CG1262 (Acp62F), CG4986, CG6069, CG10284, CG10956, CG11598, CG14034, CG17097, BG642378, BG642312, BG642167, and BG642163) were either not identified in, or have different ORFs from those predicted in, the *D. melanogaster*

Table 2.1
Cross-species comparisons of sequence divergence levels and ortholog detection analyses for individual *D. melanogaster* Acps

| Gene | Functional Class | dN sim | dS sim | dN/dS sim | dN yak | dS yak | dN/dS yak |
|---|---|---|---|---|---|---|---|
| ***Conserved in* D. pseudoobscura** | | | | | | | |
| CG1262 (Acp62F) | Trypsin Protease Inhibitor | 0.050 | 0.126 | 0.399 | 0.219 | 0.359 | 0.611 |
| CG1462* | Alkaline Phosphatase | 0.013 | 0.125 | 0.107 | 0.034 | 0.348 | 0.097 |
| CG1652* | C-type Lectin | 0.021 | 0.148 | 0.140 | 0.052 | 0.281 | 0.186 |
| CG1656* | C-type Lectin | 0.016 | 0.096 | 0.171 | 0.071 | 0.262 | 0.273 |
| CG3359* | Fasciclin | 0.010 | 0.091 | 0.104 | 0.006 | 0.200 | 0.028 |
| CG4605 (Acp32CD) | | 0.014 | 0.012 | 1.176 | 0.069 | 0.220 | 0.312 |
| CG4847* | Cysteine Protease | 0.020 | 0.114 | 0.177 | 0.043 | 0.282 | 0.151 |
| CG6069* | Serine Protease | 0.016 | 0.131 | 0.119 | 0.120 | 0.404 | 0.297 |
| CG6168* | Serine Protease | 0.036 | 0.176 | 0.203 | 0.085 | 0.365 | 0.232 |
| CG8093* | Acid Lipase | 0.005 | 0.119 | 0.039 | 0.016 | 0.399 | 0.040 |
| CG8194* | RNase | 0.011 | 0.121 | 0.094 | 0.020 | 0.276 | 0.072 |
| CG8622 (Acp53Ea) | | 0.039 | 0.143 | 0.275 | 0.120 | 0.282 | 0.425 |
| CG9024 (Acp26Ab) | | 0.018 | 0.059 | 0.305 | 0.150 | 0.356 | 0.422 |
| CG9029* | | 0.077 | 0.156 | 0.494 | 0.309 | 0.444 | 0.695 |
| CG9997* | Serine Protease | 0.031 | 0.108 | 0.291 | 0.100 | 0.393 | 0.254 |
| CG10284* | CRISP | 0.044 | 0.106 | 0.413 | 0.135 | 0.348 | 0.387 |
| CG10363* | Alpha-macroglobulin | 0.015 | 0.071 | 0.210 | 0.036 | 0.291 | 0.123 |
| CG10433* | Defensin | 0.009 | 0.027 | 0.317 | 0.037 | 0.105 | 0.347 |
| CG11598* | Acid Lipase | 0.026 | 0.217 | 0.121 | 0.466 | 1.421 | 0.328 |
| CG11864* | Metalloprotease | 0.020 | 0.090 | 0.227 | 0.099 | 0.339 | 0.293 |
| CG13309* | | 0.029 | 0.117 | 0.246 | 0.085 | 0.301 | 0.282 |
| CG16707* | | 0.052 | 0.108 | 0.483 | 0.064 | 0.211 | 0.304 |
| CG17097* | Acid Lipase | 0.010 | 0.125 | 0.082 | 0.064 | 0.208 | 0.309 |
| CG17575* | CRISP | 0.007 | 0.165 | 0.039 | 0.016 | 0.172 | 0.093 |
| CG17673 (Acp70A) | | 0.028 | 0.124 | 0.227 | 0.146 | 0.294 | 0.497 |
| CG17843* | Thioredoxin | 0.019 | 0.110 | 0.175 | 0.059 | 0.403 | 0.147 |
| CG17924 (Acp95EF) | | 0.037 | 0.223 | 0.164 | 0.259 | 0.411 | 0.630 |
| CG18284* | Acid Lipase | 0.034 | 0.190 | 0.179 | 0.104 | 0.424 | 0.244 |
| CG32952-A (Acp33A) | | 0.007 | 0.081 | 0.085 | 0.129 | 0.531 | 0.243 |
| ***Without a* D. pseudoobscura true *ortholog*** | | | | | | | |
| CG3801 (Acp76A) | Serpin | 0.025 | 0.142 | 0.178 | 0.169 | 0.467 | 0.361 |
| CG4986 | | 0.158 | 0.161 | 0.978 | 0.528 | 0.589 | 0.897 |
| CG5016 | | 0.000 | 0.000 | 0.000 | 0.156 | 0.368 | 0.423 |
| CG6289* | Serpin | 0.077 | 0.125 | 0.616 | 0.359 | 0.414 | 0.867 |
| CG7157 (Acp36DE) | | 0.049 | 0.132 | 0.371 | 0.292 | 0.633 | 0.461 |
| CG8137* | Serpin | 0.083 | 0.094 | 0.882 | 0.169 | 0.390 | 0.433 |
| CG8982 (Acp26Aa) | | 0.156 | 0.167 | 0.934 | 0.484 | 0.465 | 1.040 |
| CG9074 | | 0.040 | 0.253 | 0.157 | 0.174 | 0.619 | 0.282 |

Table 2.1 (Continued)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **CG9334*** | **Serpin** | **0.087** | **0.118** | **0.737** | **0.160** | **0.399** | **0.402** |
| **CG10852 (Acp63F)** | | **0.132** | **0.176** | **0.752** | **0.421** | **0.552** | **0.763** |
| **CG10956*** | **Serpin** | **0.031** | **0.132** | **0.238** | **0.067** | **0.346** | **0.193** |
| **CG11664*** | **Serine Protease** | **0.029** | **0.160** | **0.183** | **0.103** | **0.403** | **0.255** |
| **CG14034*** | **Phospholipase** | **0.022** | **0.161** | **0.138** | **0.121** | **0.385** | **0.315** |
| **CG14560*** | | **0.071** | **0.145** | **0.492** | **0.207** | **0.446** | **0.463** |
| **CG17797 (Acp29AB)** | **C-type Lectin** | **0.078** | **0.253** | **0.308** | **0.434** | **0.973** | **0.446** |
| **CG31056 (Acp98AB)** | | **0.119** | **0.000** | **N/A** | **0.067** | **0.261** | **0.257** |
| **CG31872*** | **Acid Lipase** | **0.032** | **0.234** | **0.136** | **0.183** | **0.368** | **0.497** |
| **CG32952-B (Acp33A)** | | **0.007** | **0.081** | **0.085** | **n/a** | **n/a** | **n/a** |
| **BG642378(6h1)*** | **Serpin** | **0.063** | **0.150** | **0.421** | **0.169** | **0.360** | **0.469** |
| **BG642167(1a8)*** | | **0.154** | **0.311** | **0.494** | **0.148** | **0.322** | **0.458** |
| **BG642312(4h1)*** | | **0.084** | **0.160** | **0.521** | **0.145** | **0.182** | **0.798** |
| **BG642163(1a3)*** | | **0.078** | **0.050** | **1.568** | **0.401** | **0.564** | **0.711** |
| | | | | | | | |
| **All Acp Averages** | | **0.045** | **0.131** | **0.473** | **0.161** | **0.397** | **0.407** |

Table 2.1 (Continued)

| **Gene** | **Functional Class** | **WGA** | **TBN** |
|---|---|---|---|
| ***Conserved in* D. pseudoobscura** | | | |
| **CG1262 (Acp62F)** | **Trypsin Protease Inhibitor** | **+** | **N/A** |
| **CG1462*** | **Alkaline Phosphatase** | **+** | **N/A** |
| **CG1652*** | **C-type Lectin** | **+** | **N/A** |
| **CG1656*** | **C-type Lectin** | **+** | **N/A** |
| **CG3359*** | **Fasciclin** | **+** | **N/A** |
| **CG4605 (Acp32CD)** | | **+** | **N/A** |
| **CG4847*** | **Cysteine Protease** | **+** | **N/A** |
| **CG6069*** | **Serine Protease** | **+** | **N/A** |
| **CG6168*** | **Serine Protease** | **+** | **N/A** |
| **CG8093*** | **Acid Lipase** | **+** | **N/A** |
| **CG8194*** | **RNase** | **+** | **N/A** |
| **CG8622 (Acp53Ea)** | | **+** | **N/A** |
| **CG9024 (Acp26Ab)** | | **+** | **N/A** |
| **CG9029*** | | **+** | **N/A** |
| **CG9997*** | **Serine Protease** | **+** | **N/A** |
| **CG10284*** | **CRISP** | **+** | **N/A** |
| **CG10363*** | **Alpha-macroglobulin** | **+** | **N/A** |
| **CG10433*** | **Defensin** | **+** | **N/A** |
| **CG11598*** | **Acid Lipase** | **+** | **N/A** |
| **CG11864*** | **Metalloprotease** | **+** | **N/A** |
| **CG13309*** | | **+** | **N/A** |
| **CG16707*** | | **+** | **N/A** |
| **CG17097*** | **Acid Lipase** | **+** | **N/A** |
| **CG17575*** | **CRISP** | **+** | **N/A** |

Table 2.1 (Continued)

| | | | |
|---|---|---|---|
| **CG17673 (Acp70A)** | | **+** | **N/A** |
| **CG17843*** | **Thioredoxin** | **+** | **N/A** |
| **CG17924 (Acp95EF)** | | **+** | **N/A** |
| **CG18284*** | **Acid Lipase** | **+** | **N/A** |
| **CG32952-A (Acp33A)** | | **+** | **N/A** |
| | | | |
| ***Without a* D. pseudoobscura true *ortholog*** | | | |
| **CG3801 (Acp76A)** | **Serpin** | **-** | **+** |
| **CG4986** | | **-** | **-** |
| **CG5016** | | **-** | **-** |
| **CG6289*** | **Serpin** | **-** | **+** |
| **CG7157 (Acp36DE)** | | **-** | **-** |
| **CG8137*** | **Serpin** | **-** | **+** |
| **CG8982 (Acp26Aa)** | | **-** | **-** |
| **CG9074** | | **-** | **+** |
| **CG9334*** | **Serpin** | **-** | **+** |
| **CG10852 (Acp63F)** | | **-** | **-** |
| **CG10956*** | **Serpin** | **-** | **+** |
| **CG11664*** | **Serine Protease** | **-** | **+** |
| **CG14034*** | **Phospholipase** | **-** | **+** |
| **CG14560*** | | **-** | **+** |
| **CG17797 (Acp29AB)** | **C-type Lectin** | **-** | **+** |
| **CG31056 (Acp98AB)** | | **-** | **-** |
| **CG31872*** | **Acid Lipase** | **c/b** | **N/A** |
| **CG32952-B (Acp33A)** | | **-** | **N/A** |
| **BG642378(6h1)*** | **Serpin** | **-** | **+** |
| **BG642167(1a8)*** | | **-** | **-** |
| **BG642312(4h1)*** | | **-** | **-** |
| **BG642163(1a3)*** | | **-** | **-** |
| | | | |
| **All Acp Averages** | | | |

*D. melanogaster* Acps conserved in *D. pseudoobscura* are listed first and *D. melanogaster* Acps not identifiable via our WGA detection methods are listed second.
* = the 34 newly Acps selected from Swanson *et al.* (2001a) EST which fit our newly defined criteria.
c/b = Contig breakpoint at site of an Acp, presence of ortholog undeterminable
TBN = TBLASTN hits against the *D. pseudoobscura* genome
WGA = SMASH blocks based whole-genome alignment identification of true Acp orthologs in *D. pseudoobscura*
sim = *D. simulans*
yak = *D. yakuba*
dN = non-synonymous nucleotide substitution value
dS = synonymous nucleotide substitution value

genome sequence (Release 4.0) (Celniker et al., 2002). We may have identified alternative splice forms of the predicted genes. An example is CG10956, whose Release 4.0 annotation predicts a single exon, while our annotation has identified a second exon at the 3′ end. Our annotation may have also revealed species-specific differences, since the EST library was constructed from *D. simulans*, or differences with the *D. melanogaster* annotation (Release 4.0) (Celniker et al., 2002).

We revised the current *D. melanogaster* (Release 4.0) annotation (Celniker et al., 2002) of the translational start sites for both CG4986 and CG10956; the splicing patterns for CG1262 (Acp62F), CG11598, CG6069, CG10284, and CG17097; and the translational start, translational stop and splicing pattern for CG14034. Four Acp *D. simulans* ESTs (BG642378, BG642312, BG642167, and BG642163, Swanson et al., 2001a) likely represent real genes but remain un-annotated in the current *D. melanogaster* genome Release 4.0 (Celniker et al., 2002). All genes unidentified and/or misannotated in *D. melanogaster* (Release 4.0) genome annotation were submitted to Genbank under Accession numbers: BK005692-BK005702.

*Confirmation of* D. melanogaster *Annotations*

RT-PCR of full coding regions in *D. melanogaster* was performed from RNA isolated from whole, 3-day old adult virgin Canton-S males. Approximately 30 flies were homogenized in Trizol according to the manufacturer's instructions (Gibco/Bethesda), and total RNA was prepared for RT-PCR as in Carninci et al. (2002). Full-length coding regions were amplified using primers designed from our annotations, which verified the annotations and expression. All amplified products were PCR-purified, cloned into pENTR-DTopo or pDONR-201 vectors (Invitrogen), and sequenced by the Biotechnology Resource Center at Cornell using the vector's internal primers. Acps that could not be RT-PCR'd from whole adult male Canton-S cDNA were amplified from available EST clones (Rubin et al., 2000a) and

subsequently cloned as above. Incomplete sequence information for Acp53Eb and a very short coding sequence for CG31056 (Acp98AB) (Wolfner et al., 1997) did not allow cloning into pENTR-DTopo or pDONR-201 vectors. Complete coding, amino acid, and primer sequences for each of the 34 new Acps can be found at see http:// www.genetics.org/supplemental/.

D. melanogaster *Acp sequence analysis*

Codon Bias: Codon bias was measured by both the frequency of optimal codons (Fop) and the percent G/C content in the third codon position (G/C3$^{rd}$) (Moriyama and Powell, 1997). Fop values range from 0.33 to 1, where 0.33 indicates homogeneous codon usage and 1 indicates that only optimal codons are used. Fop, G/C3$^{rd}$ and gene GC content calculations were performed using the codonw program (http://www.molbiol.ox.ac.uk/cu/). Codon bias values (see web supplement) were calculated using the *D. melanogaster* codon frequency table settings of the codonw program. Previous codon bias analysis of CG32952 (Acp33A) in *D. melanogaster* to *D. simulans* comparisons have combined its two ORFs (CG32952-A and CG32952-B) (Begun et al., 2000), however since each ORF contains its own predicted signal sequence we performed our analysis as two separate genes.

For comparison, we generated a random sample of 100 *D. melanogaster* genes showing 2-fold higher expression in testes versus ovaries from the Parisi *et al.* (2004) microarray dataset. Additionally, a random sample of 150 *D. melanogaster* genes with approximately the same gene lengths as Acps (Acp mean gene nucleotide length = 994.7; random gene nucleotide length= 957.6) was obtained from BDGP (http://www.fruitfly.org/sequence/dlMisc.shtml).

Gene Duplications: Sequence comparisons and chromosomal location were used together to identify gene duplicates. Individual Acp protein sequences were compared to the *D. melanogaster* genome using BlastP. Acp gene duplicate

candidates were considered if they had a conservative E-value of $10^{-10}$ and a minimum of 30% sequence identity across at least 80% of the protein (Gu et al., 2002). Because many gene duplicates often are found in tandem (Parks et al., 2004) we extended our search to locate significant matches falling within neighboring Acp genes which did not meet the above 30% sequence identity cutoff. If such a hit was present we checked for a similar protein domain prediction (http://www-cryst.bioc.cam.ac.uk/~fugue/prfsearch.html) and conserved splicing pattern to support its being a possible duplicate. Candidates which had both a BlastP E-value of $10^{-10}$ or smaller and matched all three sequence search criteria were also considered gene duplicates, even though their sequence identity may be less than 30%. Gene duplication conservation in *D. simulans* and *D. yakuba* were searched via tBlastN to their whole genome alignments (WashU-GSC http://genome.wustl.edu/tools/blast/).

*Calculation of the expected number of Acps in the* D. melanogaster *genome*

Two estimates of the total number of Acp genes in the *D. melanogaster* genome were performed as in Swanson *et al*. (2001a) by using maximum likelihood fits to a truncated Poisson distribution. A third estimate was obtained by nonparametric maximum likelihood. The first two predictions differ with respect to how they deal with five Acps (Acp26Aa, Acp26Ab, Acp32CD, Acp33A, Acp36DE) which were not adequately prescreened by Swanson *et al*. (and hence appeared in the post-screening library). In the first estimate, we ignore the five Acps pre-screened by SWANSON *et al*. (2001a) and fit a truncated Poisson distribution to the frequency spectrum (counts of singleton hits, doubleton hits, etc.). This gives a maximum likelihood count of 52 Acps in addition to the 18 that were prescreened by SWANSON *et al*. (2001a), for a total of 70. For the second estimate, we include the five Acp hit counts as though they were not prescreened at all, we obtain a maximum likelihood count of 59 Acps. If the 13 Acps that were successfully pre-screened (or at

least were not observed among the sequenced clones) are added back to the estimate of 59 Acps, this yields a prediction of 72 Acp genes in the *D. melanogaster* genome. The third method was designed for an unscreened library, and fits the data to a Poisson mixture model by nonparametric maximum likelihood (Ji-Ping Wang, personal communication). The perl script eststat.pl (available at http://www.floralgenome.org /cgi-bin/eststat/eststat.cgi) took the frequency spectrum of EST hits and produced an estimate of the total count of distinct Acps in the library at 106. This figure may be considered as an upper bound because of the prescreening that was applied to the library, leaving a more uniform frequency distribution than would be found in an unscreened library.

D. simulans *and* D. yakuba *sequence comparisons to* D. melanogaster *Acps*

Nonsynonymous substitutions per nonsynonymous site (dN) and synonymous substitutions per synonymous site (dS) values for some previously characterized Acps (Aguadé, 1999; Aguadé et al., 1992; Begun et al., 2000; Cirera and Aguadé, 1997; Kohn et al., 2004) were incorporated into this analysis. *D. yakuba* sequences were retrieved via BlastN alignment outputs of the *D. melanogaster* Acps to the *D. yakuba* genome (WashU-GSC http://genome.wustl.edu/tools/blast/). *D. simulans* and *D. yakuba* coding regions (see http:// www.genetics.org/supplemental/) were aligned to the *D. melanogaster* coding regions with ClustalX (Thompson et al., 1997). dN and dS values were calculated using DNASP 4.0 (Rozas et al., 2003). In a few cases, partial gene sequences were used. In a single *D. yakuba* case, CG32952-B, an adenine to cytosine change disrupted the apparent start codon. No other plausible ATG could be identified upstream of CG32952-B to compensate for this difference and CG32952-B was thus omitted from *D. melanogaster* to *D. yakuba* comparisons, though rare CUG start codons do exist (Prats et al., 1989). *D. simulans* and *D. yakuba* codon bias values (see http:// www.genetics.org/supplemental/) were calculated as above. The *D.*

*yakuba* non-Acp dataset was obtained from a set of non-sex specific transcripts (Domazet-Loso and Tautz, 2003). The StatView statistical program (version 5.0.1; SAS Institute Inc.) was used for statistical analyses.

*Detection of Acp orthologs in* D. pseudoobscura

The whole-genome alignment (WGA) of the *D. melanogaster* and *D. pseudoobscura* genome (Richards et al., 2005) was taken from (Emberly et al., 2003). The SMASH program (Zavolan et al., 2003) was used to find the strongest set of syntenic anchors between the *D. pseudoobscura* contigs and the *D. melanogaster* genome. Anchors were high-similarity regions from 10's to 100's of base pairs, and covered about 30% of the genome. The LAGAN program (Brudno et al., 2003) gave similar alignments. Since the size of syntenic domains between the two species generally exceeds 10kb (i.e. much larger than most repeat elements within the sequenced euchromatin), using synteny eliminated almost all ambiguities due to repeats. The SMASH blocks along with the contigs they matched, were displayed on top of the Release 3 annotation (Celniker et al., 2002) using GBROWSE (http://www.gmod.org/ggb/index.shtml).

We then examined the syntenic regions for each Acp individually at the sequence level. In 48 out of 51 cases (51 Acps instead of 52 were compared because Acp53Eb's sequence information has yet to be determined) there were SMASH blocks from a single contig that either bracketed or 'hit' the annotated gene in *D. melanogaster*. SMASH blocks from a single *D. pseudoobscura* contig, that spans a given Acp locus, indicates that the Acp genomic region in question can be aligned at the sequence level. For CG31872, a contiguous *D. pseudoobscura* sequence could not be aligned because the Acp fell into a gap between two contigs. In two other cases, CG14560 and CG9074, contained SMASH block hits to multiple contigs which differed from the contig spanning this region. The coding sequence of CG14560 and

CG9074 were then submitted to Repeatmasker (http://www.repeatmasker.org/) which indicated both Acps contained repetitive regions, thus explaining the multiple SMASH block contig hits. After filtering out the repetitive regions for CG14560 and CG9074 we could generate a single contig that bracketed each gene. Upon verification of the *D. melanogaster* to *D. pseudoobscura* contig alignments of the Acps, we retrieved the corresponding *D. pseudoobscura* sequence within the aligned contig and searched the *D. pseudoobscura* contig sequence via tBlastN using the *D. melanogster* protein sequence. If coding sequence alignments could not be identified we used GENSCAN (Burge and Karlin, 1997) and Genie (Adams et al., 2000) to locate possible ORFs. All Acps for which coding sequence alignments could be generated with the corresponding *D. pseudoobscura* contig region are considered true orthologs (Table 2.1). The SMASH block based coding sequence alignments were confirmed using another more recent *D. pseudoobscura* WGA (Karolchik et al., 2003). *D. pseudoobscura* coding sequences of conserved Acps and *D. melanogaster* to *D. pseudoobscura* contig alignments for absent or undetectable Acps can be found in the at http:// www.genetics.org/supplemental/. It is important to note that even though we define conserved Acps between *D. melanogaster* and *D. pseudoobscura* as true orthologs, we have not determined whether these Acps have maintained their accessory gland expression in *D. pseudoobscura*.

*D. melanogaster* Acps that could not be detected within the retrieved *D. pseudoobscura* contig were searched via tBlastN to the *D. pseudoobscura* genome, via the Baylor of College Medicine Drosophila Genome project web site (http://www.hgsc.bcm.tmc.edu/projects/drosophila/). For tBlastN searches, only hits with a E value of 1e-04 (Zdobnov et al., 2002) or smaller were considered significant. Whenever a significant tBlastN hit in *D. pseudoobscura* was identified, the corresponding *D. pseudoobscura* sequence was then return searched against the *D.*

*melanogaster* genome (www.flybase.net) via BlastP to determine whether it hit the Acp in question or a protein within a similar sequence/structure function class. In all cases significant *D. pseudoobscura* tBlastN hits were false positives (ex. *D. melanogaster* Acps CG8137 (serpin) and CG9334 (serpin) both hit the *D. pseudoobscura* ortholog of CG9456 (serpin)). Alignments and "false-positive *D. melanogaster* genes" for Acps whose true ortholog could not be detected via WGA, yet have a significant tBlastN hit in *D. pseudoobscura* whose return *D. melanogaster* BlastP does not match an Acp can be found at http:// www.genetics.org/supplemental/.

## Results and Discussion

### D. melanogaster *Acp genes*

Secreted proteins synthesized by the *D. melanogaster* male accessory gland have important functions in reproduction (reviewed in Chapman and Davies, 2004; Kubli, 2003; Wolfner, 2002). To address more thoroughly the functions and evolution of these Acp proteins, we carried out a comprehensive identification and annotation of *D. melanogaster* Acp genes. Prior to 2001, 18 Acp genes had been reported in *D. melanogaster* (Chen et al., 1988; Simmerl et al., 1995; Wolfner et al., 1997). In 2001, SWANSON *et al*. (2001a) identified 57 additional candidate Acp genes in *D. simulans* via an evolutionary EST approach that was done to permit a rapid scan to identify genes with features suggesting rapid evolution. However, the ESTs identified by SWANSON *et al*. were partial cDNAs, and from a species, *D. simulans*, which is presently less amenable to genetic analyses than is *D. melanogaster*. We therefore full-length sequenced a select set of the *D. simulans* EST sequences identified by SWANSON *et al*. (2001a). The full-length *D. simulans* Acp EST sequences allowed us to identify the complete ORF of their *D. melanogaster* orthologs. We then applied a more stringent set of criteria to identify those genes on which to focus, based on

what is known of the initial 18 Acps. We define bona fide Acps here as genes that: (a) encode a protein with a predicted secretion signal sequence, (b) have a pattern of EST hits in other tissue- or cell-type-specific EST screens consistent with accessory gland predominant expression (c) have no previously characterized non-Acp function (d) show male and/or accessory gland predominant expression in *D. melanogaster*. Using these stringent criteria we utilized secretion signal prediction programs, EST databases, reports of mutant phenotypes and RT-PCR to screen through the 57 candidate Acps identified by SWANSON *et al*. (2001a) (see Materials and Methods for details). Thirty-four Acps fit these new stringent criteria (see Table 2.1). The other 23 genes identified by SWANSON *et al*. could encode proteins made in accessory glands and potentially also transferred to females, but their additional tissues of expression and/or non-reproductive functions complicate genetic and functional analyses and evolutionary interpretations; thus we do not consider them further. It is also formally possible that the expression characteristics of some of these 23 Acps differ in *D. melanogaster* and *D. simulans*, resulting in their exclusion from the stringently-selected *D. melanogaster* Acps on which we focus. The 34 stringently selected *D. melanogaster* Acps that fit the above criteria, in combination with the 18 previously known Acps, makes a total of 52 *D. melanogaster* Acps (Table 2.1) whose gene boundaries and expression have been confirmed. This comprehensive and characterized set of 52 Acps has also allowed a recalculation of the predicted number of Acps in the genome. Fitting the frequency spectrum of the 52 Acps with EST hits from the SWANSON *et al*. (2001a) screen to a truncated Poisson distribution and to a Poisson mixture model, gave maximum likelihood estimates in the range of 70 to 106 Acps in the *D. melanogaster* genome (see Materials and Methods), respectively. Additionally, recently-identified Acps CG8626, CG15616 and CG17799 (Holloway

and Begun, 2004) suggest that the field is steadily approaching a complete list of Acps in *D. melanogaster*.

These 52 *D. melanogaster* Acps are expected to be extracellular, and thus transferred to the female upon mating and to be produced primarily or exclusively in the male's accessory gland. Indeed, all 16 Acp genes tested so far encode seminal proteins detectable only in the male's accessory gland and transferred to the female during mating (Albright, 2003; Bertram et al., 1996; Chen et al., 1988; Coleman et al., 1995; Lung et al., 2002; Monsma et al., 1990; Ravi Ram et al.). Additional support that this set of 52 *D. melanogaster* Acps truly represents accessory-gland predominant genes stems from the finding that 29 out of 46 tested Acp genes showed 2-fold or higher expression values in germlineless males versus germlineless females comparisons (6 Acps were not present on the microarrays) (Parisi et al., 2004; Parisi et al., 2003).

*Presence of multiple Acp gene duplicates across the D. melanogaster genome*

About 40% of the *D. melanogaster* genome (5536 of 13601 genes) appears to be gene duplicates (Rubin et al., 2000b). Similarly, 16 (31%) of the 52 Acps appear to have gene duplicates (Table 2.2) within the *D. melanogaster* genome. CG8137 and CG9334 are the only gene duplicates not in tandem, though they share the same intron splice positions. Percent identities of the Acp gene duplicates range from 25% (CG3801 (Acp76A) and BG642378) to 92% (CG6289 and CG6663) indicating that a range of recent and ancient gene duplicates have been identified. Nine of these cases of gene duplication are within the 52 Acp collection (3 duplicate pairs plus one triplicate) (Table 2.2, part B), indicating that these duplicates have similar expression profiles. This is consistent with the observation that gene duplication events often lead to co-expressed genes that cluster together (Boutanaev et al., 2002). These nine Acp gene duplicates are found in tandem clusters of paired (or triplicate) genes, and

Table 2.2
List of Acp Gene Duplicates in *D. melanogaster*

**Part A**

Acps and their gene duplicates which are not expressed in the male accessory gland

| Acp | Non-Acp Gene Duplicate(s) | % Protein Identity |
|---|---|---|
| CG17797 | CG17799 | 45 |
| CG17843 | CG6690 | 39 |
| CG6289 | CG6663 | 92 |
| CG13309 | CG13308 CG13312 | 50 37 |
| CG17575 | CG30486 | 30 |
| CG11864 | CG15254 | 48 |
| CG11598 | CG11600 | 46 |

**Part B**

Acps whose gene duplicates retain male accessory gland expression

| Acp | Acp Duplicate(s) | % Protein Identity |
|---|---|---|
| CG1652 | CG1656 | 46 |
| CG3801 | BG642378 | 25 |
| CG31872 | CG17097 CG18284 | 39 88 |
| CG8137 | CG9334 | 72 |

they share the same relative splice site positions, which are also conserved in *D. simulans* and *D. yakuba*.

For seven additional Acps we detect duplicates in the genome (Table 2.2, part A). Again, tandem arrangements are seen in *D. simulans* and *D. yakuba*, and the *D. melanogaster* duplicates share the same splice site positions. However, in these seven cases, only one member of each duplicate pair is a member of our 52-Acp collection. This could be because the collection is incomplete (only 52 of the predicted 70 to106 Acps are described here), or because a given duplicate's expression might not fit our stringent criteria of accessory gland-predominant expression, or because a given duplicate has an entirely different expression pattern. An example of the first is CG17799. This gene duplicate of CG17797 (Acp29AB), has recently been shown to also be expressed in the *D. melanogaster* accessory gland (Holloway and Begun, 2004), but is not among the 52 genes we focused on here, simply because it was not detected in the Swanson *et al*. (2001a) EST screen or previous screens. It is likely that other gene duplicates of Acps whose expression profile have yet to be determined may later be identified as Acps. The identification of the Acp gene duplicates will have an impact on future genetic analysis since duplication may introduce genetic redundancy. Additionally, since many Acps are rapidly evolving, Acps provide a good example to define which evolutionary processes drive the divergence of gene duplicates.

*Comparative sequence analysis of the* D. melanogaster *Acps and their* D. simulans *and* D. yakuba *orthologs*

Several Acps have features indicative of rapid evolution (Aguadé, 1999; Aguadé et al., 1992; Begun et al., 2000; Cirera and Aguadé, 1997; Kern et al., 2004; Kohn et al., 2004; Panhuis et al., 2003; Stevison et al., 2004; Tsaur and Wu, 1997) and Swanson *et al.*'s (2001a) data suggested that rapidly evolving genes are represented at a high level among Acps. With our larger collection of fully-annotated Acp genes,

and the recent release of *Drosophila* genomic sequences, we could examine this question in detail. We investigated the patterns of codon bias and rates of evolution (by examining rates of nonsynonymous (dN) and synonymous (dS) nucleotide substitutions) for the 52 stringently-defined Acp genes and compared those results to those with a control set of genes that are not expressed in the accessory gland.

Codon Bias: Levels of codon bias have been used as a criterion for detecting rapidly evolving genes in *Drosophila* (Schmid and Aquadro, 2001). Although codon bias alone cannot conclusively prove rapid evolution, genes that are rapidly evolving tend to have low codon bias (Schmid et al., 1999). A previous study of 10 Acp genes (Begun et al., 2000) found that Acp genes tend to have lower levels of codon bias relative to the rest of the *Drosophila* genome. The 52 *D. melanogaster* Acp genes defined here as a class, have significantly lower levels of codon bias (Mann Whitney test, $P < 0.001$ for both Fop and G/C3$^{rd}$ calculations, Table 2.3) than the control random sample of *D. melanogaster* genes of approximately the same length. *D. melanogaster* Acp genes do not exhibit significant differences (Fop, Mann Whitney test P=0.612, G/C3$^{rd}$ Mann Whitney test P=0.302, Table 2.3) in codon bias from the majority of genes expressed in the testis. Comparing levels of codon bias in the *D. simulans* Acp gene orthologs to non-Acp genes we also find that Acp genes exhibit lower levels of codon bias (data not shown). We also determined whether this phenomenon is found in a more distantly related species, *D. yakuba*. Levels of codon bias in *D. yakuba* Acps were also significantly lower (Fop, Mann Whitney test, P<0.001, G/C3$^{rd}$ Mann Whitney test, P<0.001, Table 2.3) than a collection of *D. yakuba* non-Acps (Domazet-Loso and Tautz, 2003).

Our findings with the extended set of 52 Acps agree with the findings by Begun *et al*. (2000) – on average the 52 Acps exhibited lower than average levels of codon bias in *D. melanogaster*, *D. simulans* and *D. yakuba*. It is possible that these

Table 2.3
*D. melanogaster* and *D.yakuba* codon bias comparisons between different gene classes

| Gene class averages | Fop [a] | % G/C 3rd [a] |
|---|---|---|
| *D. melanogaster* Acps | 0.498 | 0.512 |
| *D. melanogaster* non-Acps | 0.543 | 0.666 |
| *D. melanogaster* Testes-specific | 0.366 | 0.516 |
| | | |
| *D. yakuba* Acps | 0.432 | 0.545 |
| *D. yakuba* non-Acps | 0.544 | 0.653 |

| Side by Side Comparisons | Mean difference between classes | |
|---|---|---|
| *D. melanogaster* Acps vs. non-Acps | P<0.001 | P<0.001 |
| *D. melanogaster* Acps vs. Testes-biased genes | P=0.612 | P=0.302 |
| *D. yakuba* Acps vs. non-Acps | P<0.001 | P<0.001 |

[a] High values are associated with codon bias for both the frequency of optimal codons (Fop) and the percentage of GC bases in the third position (%G/C 3rd). Mann Whitney test used to test for significant differences

low levels of codon bias could be due to rapid rates of protein evolution of Acps (Akashi, 1994). *Drosophila* codon bias can also be influenced by sequence length (Duret and Mouchiroud, 1999), expression level and local GC content. Because short *Drosophila* genes tend to exhibit high levels of codon bias (Duret and Mouchiroud, 1999), and because Acp genes also tend to be short, our control set was selected to be genes of similar length to avoid the contribution of gene length. The unusual levels of codon bias seen for both Acps and testis-genes (Table 2.3) suggest that male-reproductive proteins in general may exhibit lower levels of codon bias. Low levels of codon bias for *D. melanogaster* testis genes is consistent with their poorly conserved sequence and sex-specific expression pattern when compared to *A. gambiae* (Parisi et al., 2004) or *D. simulans* (Ranz et al., 2003), respectively. That male-biased genes evolve more rapidly at the sequence (Singh and Kulathinal, 2000) and expression pattern levels (Meiklejohn et al., 2003) suggests their rapid evolution may not allow adaptation to high levels of codon bias.

Levels of divergence: A high dN/dS ratio can identify genes for which amino acid replacement is being driven by a selective pressure. Acp genes have already been reported to demonstrate higher levels of sequence divergence than non-Acp genes between *D. simulans* and *D. melanogaster* (Kern et al., 2004; Stevison et al., 2004; Swanson et al., 2001a). However, those analyses used only partial sequences or included genes that our present analyses have shown not to fit the stringent definition of Acps in *D. melanogaster* and thus could be subject to additional or different selection pressures.

Here we compare our complete sequences of a set of stringently-selected *D. melanogaster* Acps with their *D. simulans* and *D. yakuba* orthologs. Acps exhibit high levels of sequence divergence with average dN values for *D. simulans* of 0.045 (Table 2.1), similar to previously reported dN values for *D. simulans* Acps of 0.052

(Swanson et al., 2001a) and 0.050 (Begun et al., 2000). The average level of dS for this set of Acps in *D. simulans* is 0.13 (Table 2.1), similar to the known average *D. simulans* dS value of 0.11 (Bauer and Aquadro, 1997; Begun and Whitley, 2000; Betancourt et al., 2002; Moriyama and Powell, 1997). We also compared Acp to non-Acp levels of sequence divergence between *D. melanogaster* and *D. yakuba*. In this comparison as well, Acps have significantly higher dN (0.161) and dN/dS (0.407) values than non-Acps (dN and dN/dS values 0.026 and 0.082, respectively) (Table 2.4, Mann Whitney test, both dN and dN/dS, $P<0.001$).

Using levels of dN and dS as a metric to identify rapidly evolving genes, which have a dN/dS value greater than 1, Swanson *et al*. (2001a) identified 19 genes whose partial sequence had dN/dS > 1 in *D. melanogaster*/*D. simulans* comparisons. However, our reanalysis of the 52 Acps using complete gene sequences yields only three Acps from both *D. melanogaster*/*D. simulans* and *D. melanogaster*/*D. yakuba* comparisons with dN/dS > 1 (Table 2.1). We believe this discrepancy between Swanson *et al*. (2001a) results and those reported here is because we analyzed full-length coding regions from an accurately annotated list of genes instead of partially sequenced cDNAs which in some cases were misaligned. In addition, for many rapidly evolving genes often only part of the gene is under positive selection (Hughes and Nei, 1988; Swanson et al., 2001b). Thus some partial cDNAs analyzed by Swanson *et al*. (2001a) may have fortuitously contained regions under positive selection giving a higher dN/dS than seen when the entire gene is tested. For this reason a dN/dS > 0.5 was recently proposed as a more practical cutoff when using full-length sequences, to identify candidate genes which may be driven by positive selection (Swanson et al., 2004). Applying this cutoff value of 0.5 to the 52 Acps we find that nine Acps (but not the same nine as in Swanson *et al*. 2001a) in both the *D. melanogaster*/*D. simulans* and *D. melanogaster*/*D. yakuba* have dN/dS > 0.5. This

Table 2.4
Divergence levels of *D. melanogaster* Acps versus non-Acps when compared to their *D. yakuba* orthologs

| Divergence level averages | Acps | non-Acps | Mann-Whitney test P-value |
|---|---|---|---|
| *D. melanogaster*/*D. yakuba* dN | 0.161 | 0.026 | P<0.001 |
| *D. melanogaster*/*D. yakuba* dS | 0.397 | 0.306 | P=0.002 |
| *D. melanogaster*/*D. yakuba* dN/dS | 0.407 | 0.082 | P<0.001 |

Non-synonymous (dN) and synonymous (dS) nucleotide substitution rates.

proportion of Acps (9/52 = 17%) is similar to the percentage of Acps identified in the Swanson *et al.* (2001a) male accessory gland EST screen (19%) with a dN/dS > 0.5. Comparable percentages of Acps with a dN/dS > 0.5 described here to those Acps identified in Swanson *et al*. (2001a) supports the idea that dN/dS > 0.5 may serve as a good indicator for candidate rapidly evolving genes (Swanson et al., 2004). Further analysis of the role of natural selection in shaping Acp sequence evolution using codon-substitution models will be presented elsewhere.

*Detection of* D. melanogaster *Acp orthologs in* D. pseudoobscura

The complete genome sequence of *D. pseudoobscura* (Richards et al., 2005) allowed us to search for conserved *D. melanogaster* Acps in a distantly related species outside of the *D. melanogaster* subgroup. A whole genome alignment (WGA) approach was used to determine which of the 52 *D. melanogaster* Acps can be identified in *D. pseudoobscura*. Syntenic regions covering each Acp were generated for 50 Acps. Limited sequence information for the other two Acps (Acp53Eb and CG31872) prevented generation of accurate comparative genome sequence alignments. We verified all the *D. melanogaster* to *D. pseudoobscura* contig alignments and identified the corresponding *D. pseudoobscura* Acp, to generate coding sequence alignments between the two species. All *D. melanogaster* Acps for which coding sequence alignments could be generated with the corresponding *D. pseudoobscura* contig are considered true orthologs (see Table 2.1). We found that, via WGA, 58% (29/50) of the *D. melanogaster* Acps have true orthologs in *D. pseudoobscura* (Table 2.1). For the 21 *D. melanogaster* Acps for which true orthologs could not be identified in *D. pseudoobscura* we used tBlastN against all *D. pseudoobscura* contigs and orphan sequences to ensure we had not missed *D. melanogaster* Acps that had moved to non-syntenic chromosomal locations in *D. pseudoobscura*. In ten cases, tBlastN comparisons gave significant *D. pseudoobscura*

hits. However, each hit was interpreted as a false positive because it either matched repetitive sequence in the Acp or a different *D. pseudoobscura* gene with a respective non-Acp *D. melanogaster* ortholog (see Materials and Methods). Our inability to detect a *D. pseudoobscura* ortholog for a *D. melanogaster* Acp gene via this method does not mean that a *D. pseudoobscura* ortholog does not exist, but only that our searches were negative. *D. melanogaster* Acps undetectable in *D. pseudoobscura* via our methods, could be highly diverged, located in an unsequenced region of the *D. pseudoobscura* genome, or potential *D. melanogaster* lineage–specific proteins. A recent study (Wagstaff and Begun, 2005) uncovered a *D. pseudoobscura* gene with 18.5% amino acid sequence identity to *D. melanogaster* Acp26Aa. This is below the similarity level detectable in our search for *D. pseudoobscura* orthologs. For another gene, Acp95EF, our analysis revealed its *D. pseudoobscura* ortholog which was undetected by Wagstaff and Begun (2005). Differences in methodologies and the limited alignability of the *D. pseudoobscura* genome (only ~48%; Richards *et al*. 2005) likely account for these two differences in Acp ortholog detection.

Of the 29 Acps we found conserved between *D. melanogaster* and *D. pseudoobscura* it had been possible to generate comparative structural models to known protein classes for 20 Acps (Table 2.1) (Mueller et al., 2004). This represents a greater fraction (20/29 = 69%) than is seen for those *D. melanogaster* Acps that do not have *D. pseudoobscura* counterparts (9/21 = 43%). That more proteins within predicted protein functional classes are conserved between *D. melanogaster* and *D. pseudoobscura* suggests that these proteins may mediate reproductive strategies that are conserved across *Drosophila*. Interestingly, the protease inhibitor class is not well conserved between the two species (Table 2.1): only one (Acp62F) of seven predicted or known Acp protease inhibitors is identifiable between the two species (Table 2.1). The lack of conservation of protease inhibitors between *D. melanogaster* and *D.*

*pseudoobscura* is significantly greater than the percentage of Acps not shared in all other protein classes (Chi-square= 12.28, df= 1, P<0.001). Acps that are predicted protease inhibitors have been suggested to participate in sperm storage, cost-of-mating (specifically Acp62F (Lung et al., 2002)) and/or immune regulation (Khush and Lemaitre, 2000; McGraw et al., 2004) which may contribute to their evolution between *D. melanogaster* and *D. pseudoobscura* lineages.

*Comparative sequence analysis within the* D. melanogaster *subgroup of Acps shared or not shared with* D. pseudoobscura

Within the set of Acps conserved between *D. melanogaster* and *D. pseudoobscura*, we examined levels of codon bias and dN/dS with two other species in the *D. melanogaster* subgroup. We tested whether codon bias and dN/dS values could distinguish those *D. melanogaster* Acps which share or do not share true orthologs in *D. pseudoobscura*. We find that *D. melanogaster* Acps without detectable *D. pseudoobscura* true orthologs have significantly lower levels of codon bias in *D. melanogaster* (Fop and G/C3$^{rd}$ Mann Whitney test, P=0.001 and P<0.001, respectively) and *D. yakuba* than Acps conserved between *D. melanogaster* and *D. pseudoobscura* (Fop and G/C3$^{rd}$ Mann Whitney test, P<0.001 and P<0.001, respectively, Table 2.5, Part A). Additionally, levels of dN/dS are significantly higher for *D. melanogaster*/*D. simulans* and *D. melanogaster*/*D. yakuba* comparisons of Acps without true orthologs in *D. pseudoobscura* compared to Acps conserved between *D. melanogaster* and *D. pseudoobscura* (*D. simulans* and *D. yakuba*, Mann Whitney test, P=0.002 and P<0.001, respectively, Table 2.5, Part A). This subgroup divergence analysis can be extended to the case of the *D. melanogaster* predicted protease inhibitor Acps that do not have counterparts in *D. pseudoobscura* (Table 2.1). We find that the seven predicted or known Acp protease inhibitors have both significantly lower levels of codon bias and higher levels of sequence divergence

(dN/dS) than Acps in other predicted functional classes (Table 2.5, Part B). Together, these results suggest that *D. melanogaster* Acps without a true *D. pseudoobscura* ortholog have greater levels of sequence divergence (dN/dS) within the *D. melanogaster* subgroup, than *D. melanogaster* Acps with a detectable *D. pseudoobscura* ortholog. Those *D. melanogaster* Acps with higher sequence divergence levels which do not share a true ortholog in *D. pseudoobscura* thus serve as good candidates for mediating reproductive functions in close relatives of *D. melanogaster*.

*Underrepresentation of Acps on the* D. melanogaster *X chromosome*

As previously reported (Swanson et al., 2001a; Wolfner et al., 1997), Acps' chromosomal locations are biased to autosomes in *D. melanogaster*. Only one of the 52 Acps, CG11664, falls on the X chromosome at cytological band 1D2 in *D. melanogaster*. The remaining 51 Acps are evenly distributed across the 2nd (27 Acps) and 3rd (24 Acps) chromosomes. Given that the X chromosome contains ~17% of the total *D. melanogaster* genome (Celniker et al., 2002), if the 52 Acps were randomly distributed across the genome we would expect ~9 Acps of the 52 Acps to fall on the X chromosome and 43 on autosomes. The presence of only a single X-linked Acp is highly unlikely to have occurred by chance ($G_{corr}$= 7.908, df=1, P=0.005), supporting reports that the *D. melanogaster* X chromosome is deficient in male-biased genes (Andrews et al., 2000; Parisi et al., 2004; Ranz et al., 2003; Swanson et al., 2001a; Wolfner et al., 1997).

An alternative approach to understanding the chromosomal bias of sex-specific genes is to focus on the region that contains the single X-linked *D. melanogaster* Acp. The 50kb region flanking CG11664 is unusual in several respects. First, CG11664 lies in an apparently gene-poor region, with only six other genes within the

Table 2.5
Distinct sequence evolution patterns for *D. melanogaster* Acps present and undetectable in *D. pseudoobscura*

**Part A**

| Averages | *D. pseudoobscura* orthologs present | *D. pseudoobscura* orthologs undetectable | Mann-Whitney test P-value |
|---|---|---|---|
| Fop [a] *D. melanogaster* | 0.439 | 0.351 | P= 0.001 |
| G/C3$^{rd}$ [a] *D. melanogaster* | 0.559 | 0.456 | P< 0.001 |
| Fop [a] *D. yakuba* | 0.467 | 0.371 | P< 0.001 |
| G/C3$^{rd}$ [a] *D. yakuba* | 0.598 | 0.474 | P< 0.001 |
| dN/dS [b] *D. simulans* | 0.240 | 0.525 | P= 0.002 |
| dN/dS [b] *D. yakuba* | 0.287 | 0.515 | P< 0.001 |

**Part B**

| Averages | All Other Predicted Functional Classes | Protease Inhibitors | Mann-Whitney test P-value |
|---|---|---|---|
| Fop [a] *D. melanogaster* | 0.476 | 0.365 | P= 0.002 |
| G/C3$^{rd}$ [a] *D. melanogaster* | 0.596 | 0.477 | P= 0.009 |
| Fop [a] *D. yakuba* | 0.503 | 0.395 | P= 0.007 |
| G/C3$^{rd}$ [a] *D. yakuba* | 0.602 | 0.498 | P= 0.005 |
| dN/dS [b] *D. simulans* | 0.173 | 0.496 | P= 0.001 |
| dN/dS [b] *D. yakuba* | 0.235 | 0.477 | P= 0.004 |

[a] High values are associated with codon bias for both the frequency of optimal codons (Fop) and the percentage of GC bases in the third position (%G/C 3$^{rd}$).
[b] Non-synonymous (dN) to synonymous (dS) nucleotide substitution ratios.

surrounding 100kb. On average there are ~ 11 genes/100kb in the *D. melanogaster* genome (= 13792 genes/120Mb) (Adams et al., 2000; Celniker et al., 2002). Second, of the six neighboring genes, four (CG3713, CG11663, CG14634, and CG14635) appear to be testis-biased in their expression (Andrews et al., 2000; Parisi et al., 2004, no expression data could be found for CG14632 and CG14633); thus perhaps this region is a "hotspot" for harboring male-biased genes on the X chromosome. Third, more than half of the genes in this region do not appear to be conserved between *D. pseudoobscura* and *D. melanogaster*, consistent with the report that male-biased genes tend to evolve more rapidly at both expression (Ranz et al., 2003) and sequence (Parisi et al., 2004) levels. Fourth, five of the six neighboring ORFs, in addition to CG11664, are intronless, suggesting they may be retrogenes. Additionally, this region appears to also be a hotspot for transposable elements. In the recent transposable element (*piggyBac* and P-element) insertion mutagenesis collection release of 16,500 fly lines (Thibault et al., 2004), the 100kb region surrounding CG11664 contained 34 insertions, which is more than the average of ~ 14 transposable elements per 100kb (= 16500 elements/120Mb). Altogether, the region surrounding CG11664 contains a number of unique features which may help determine what pressures are driving the evolution of sex-specific genes on the X chromosome in *D. melanogaster*.

Multiple hypotheses including sexual antagonism, dosage compensation and X-inactivation may explain the paucity of male-biased genes on the *D. melanogaster* X chromosome (reviewed in Parisi et al., 2004). The ability to help distinguish the importance of these phenomena could be assisted by looking at *D. pseudoobscura*. In *D. pseudoobscura*, the X chromosome consists primarily of a region largely syntenic to the left arm of the 3$^{rd}$ chromosome in *D. melanogaster* (3L) that fused more recently in the *D. pseudoobscura* lineage to a region syntenic to the X chromosome of *D. melanogaster* (Segarra and Aguadé, 1992). Thus all Acps with *D. pseudoobscura*

orthologs that are located on 3L in *D. melanogaster* (CG1262 (Acp62F), CG10852 (Acp63F), CG17673 (Acp70A), CG3801 (Acp76A), CG6289, CG13309, CG14560, BG642312, CG16707, CG8194, BG642378, CG6168), would now be on the right arm of the *D. pseudoobscura* X chromosome (XR). If there is selection against X-linkage for Acps, we would expect a higher "loss" of Acps from the "new" (*D. melanogaster* 3L homologue) X-linked genes in the *D. pseudoobscura* lineage than for Acps on autosomes in *D. pseudoobscura*. We find that a larger proportion of "new" Acps on the *D. pseudoobscura* X chromosome are not shared between the two species (as compared to autosomal Acps in *D. pseudoobscura*), although this difference is not statistically significant (*D. pseudoobscura* X chromosome (7/13= 54% absent or undetected) versus autosomes (13/36= 36% absent or undetected); Chi-square=1.01, df=1, P=0.322). That fewer X-linked *D. pseudoobscura* Acps are conserved than autosomal Acps is consistent with selection against X-linked Acps. However, the *D. melanogaster* 3L chromosome and its *D. pseudoobscura* XR counterpart show the second lowest level of genome sequence alignability between species: 46.5% of *D. melanogaster* 3L's base pairs are alignable with *D. pseudoobscura* XR as compared to an average across all chromosomes of 48%. Therefore, the relatively low sequence conservation of the *D. pseudoobscura* XR arm suggests that loss or translocation of Acps from this arm may have resulted from the particular X-chromosomal evolutionary dynamics in the *D. pseudoobscura* lineage, rather than any sex-specific selection acting differentially on the X chromosomes versus autosomes.

## Conclusions

Genes with increased rates of evolution increase the frequency with which incompatibilities evolve between closely related species. Since some Acps in *Drosophila* evolve faster than other genes, these rapidly evolving Acps serve as good

candidates for examining the selection pressures associated with reproductive functions. We have characterized here such divergent Acps, whose divergence may be attributable to sexually antagonistic evolution with proteins from the female or male (Swanson et al., 2001a; Swanson and Vacquier, 2002). The female's genotype has been shown to play an active role in sperm displacement (Clark and Begun, 1998) and a recent EST screen identified a number of candidate receptors/sexually antagonistic genes for Acps (Swanson et al., 2004). Candidate receptors would likely serve as the most upstream female genes in signaling pathways for the numerous biological processes/pathways regulated by Acps, sperm, and the act of mating (McGraw et al., 2004). The comprehensive set of Acps described here thus provides a basis for understanding both the evolutionary dynamics and function of specific Acps. This, in turn, may help tease apart the functional importance of male-female interactions during the evolution of reproductive isolation.