DOI: 10.1534/genetics.108.096263

Orthology, Function and Evolution of Accessory Gland Proteins in the *Drosophila repleta* Group

Francisca C. Almeida[1] and Rob DeSalle

Department of Biology, New York University, New York, New York 10003 and Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York 10024

Manuscript received September 15, 2008
Accepted for publication November 10, 2008

ABSTRACT

The accessory gland proteins (Acps) of Drosophila have become a model for the study of reproductive protein evolution. A major step in the study of Acps is to identify biological causes and consequences of the observed patterns of molecular evolution by comparing species groups with different biology. Here we characterize the Acp complement of *Drosophila mayaguana*, a *repleta* group representative. Species of this group show important differences in ecology and reproduction as compared to other Drosophila. Our results show that the extremely high rates of Acp evolution previously found are likely to be ubiquitous among species of the *repleta* group. These evolutionary rates are considerably higher than the ones observed in other Drosophila groups' Acps. This disparity, however, is not accompanied by major differences in the estimated number of Acps or in the functional categories represented as previously suggested. Among the genes expressed in accessory glands of *D. mayaguana* almost half are likely products of recent duplications. This allowed us to test predictions of the neofunctionalization model for gene duplication and paralog evolution in a more or less constrained timescale. We found that positive selection is a strong force in the early divergence of these gene pairs.

ACCESSORY gland proteins (Acps) are secreted by the accessory glands of Drosophila males during insemination and perform fundamental roles in reproduction, being essential for egg fertilization (for reviews, see Wolfner 2002 and Chapman and Davies 2004). Comparisons between *Drosophila simulans* and *D. melanogaster* orthologs showed that Acps on average have two times more replacement substitutions than non-Acp genes (Swanson *et al.* 2001). Rapid evolution and high turnover rates of Acps result in the observation that the more phylogenetically distant two species are, the more difficult it is to identify Acp orthologs in their genomes (Haerty *et al.* 2007). For example, of 52 Acps identified in *D. melanogaster*, only 29 had detectable orthologs in the *D. pseudoobscura* genome (Mueller *et al.* 2005), while all 52 Acps are present in *D. simulans*, a species closely related to *D. melanogaster* (but see also Begun and Lindfors 2005). These and other studies on the *melanogaster* and *pseudoobscura* groups showed that Acps are frequently subject to gene duplication and gene loss (Begun and Lindfors 2005; Wagstaff and Begun 2005a).

The *melanogaster* group, on which most of the Acp studies have been focused, represents only a small sample of the Drosophila genus. Drosophila encompasses a large number of species with a great diversity of ecologies, reproductive strategies, and developmental pathways. More recently, an *e*xpressed *s*equence *t*ag (EST) study was done to identify the Acps of *D. mojavensis* (Wagstaff and Begun 2005b), a species of the *repleta* group. This group belongs in a different subgenus of Drosophila than the other species studied thus far (Throckmorton 1975). The *repleta* group represents one of the biggest radiations in the genus Drosophila (Durando *et al.* 2000). The species in this group have a very different reproductive biology as compared to the *melanogaster* group flies. Higher remating rates, frequent formation of an insemination reaction that prevents remating for a few hours, and high levels of seminal fluid absorption by the female are some of these differences (Markow and Ankney 1988; Pitnick *et al.* 1997; Knowles and Markow 2001). One question that arises is whether the Acp complement can account for these differences. Wagstaff and Begun (2005b) results confirm some of the previous findings in other Drosophila species, such as a high evolutionary rate of Acps as compared to non-Acp genes (in this case, testis expressed genes). Nevertheless, since accessory glands were not dissected separately in that study, it was not possible to make a thorough comparison of the *D. mojavensis* Acp complement with that of other Drosophila species.

To further understand the nature and evolution of Acps in the *repleta* group, we developed a cDNA library of accessory glands for *D. mayaguana*, a species in the

Sequence data from this article have been deposited with the EMBL/GenBank Data Libraries under accession nos. FE040599–FE040997.

[1]*Corresponding author:* Department of Mammalogy, American Museum of Natural History, 79th St. at Central Park W., New York, NY 10024. E-mail: falmeida@amnh.org

same cluster (*mulleri*) as *D. mojavensis*. This allowed us to make not only nucleotide sequence comparisons between the two species but also comparisons of expression data. In particular, the availability of these new data allowed us to test hypotheses concerning phylogenetic relationships of orthologs and paralogs of Acps, neofunctionlization of duplicated Acps, and the general role of reproductive biology in molding the molecular evolution of these proteins. We demonstrate here that very general statements about the evolution of these proteins can be made and that the potential for more specific tests of hypotheses to clarify and explore the molecular evolution of these proteins can be realized by changing the scale of the targets of the analysis.

Our results indicate that Acp functional categories in the *repleta* group are not as divergent from the *melanogaster* group Acps as suggested by previous studies. Acps in the *repleta* group are subjected to the same molecular mechanisms and evolutionary processes as in other Drosophila groups. Nevertheless, molecular evolutionary analysis of *D. mayaguana* Acps confirmed and extended to other species the finding that evolutionary rates of Acps are considerably faster in the *repleta* group than in the *melanogaster* group (Wagstaff and Begun 2005b).

A large number of *D. mayaguana* Acps were found to be the product of recent gene duplications. We used the data generated on gene families together with sequence data available for other members of the *repleta* group to address hypotheses concerning the evolution of duplicated Acp genes. We tested predictions of the neofunctionalization hypothesis (Ohno 1970) for the maintenance of duplicated genes and found that duplicated Acps conform to many of the predictions made by this hypothesis.

MATERIALS AND METHODS

Cloning *D. mayaguana* Acps: A cDNA library from accessory glands of *D. mayaguana* was made using a subtraction protocol to enrich the library for male-specific genes. Expression specificity of ESTs was further checked with a dot-blot procedure using whole-female cDNA as probe. All molecular methods (RNA extraction, cDNA synthesis, library construction, dot blotting, and DNA sequencing) are described in detail in Almeida (2007) and in supplemental text 1.

Library characterization: Estimates of the total number of genes expressed in the accessory glands of *D. mayaguana* were obtained by fitting a Poisson model to the distribution of the number of clones in the library per unique sequence. This analysis was done with the program ESTstat 1.0 (Wang *et al.* 2004). The same method was used to estimate the total number of Acps.

Gene characterization: cDNA sequences were translated in all six frames and all possible open reading frames (ORFs) were characterized. The presence of a signal peptide, a marker of secreted proteins, was verified by evaluating the parameters *S-mean* and *D-score* in the neural-network model, and the probability of having a signal peptide as estimated using a hidden Markov model, in all cases as estimated with the program SignalP 3.0 (Bendtsen *et al.* 2004). Functional protein domains were identified by using the CD-search online tool and the Conserved Domain Database [(CDD), National Center for Biotechnology Information (NCBI)], which includes domains from other databases such as Pfam, SMART, and COGs.

Criteria for Acp status: The term "Acp" is employed to designate genes that are expressed exclusively (or mostly) in the accessory glands and are secreted to be part of the seminal fluid. Some cDNA sequences obtained in a tissue-specific library may represent genes that are expressed in other organs as well or housekeeping genes, such as ribosomal proteins, that cannot be classified as Acps. Taking into account several possible methodological biases (Swanson *et al.* 2001), we relied on the following set of criteria to decide if a *D. mayaguana* transcript should be classified as an Acp: (1) a high probability of a transcript being homologous to a gene identified as Acp in other Drosophila species on the basis of the results of BLAST searches; (2) negative dot-blot results; (3) the presence of a signal peptide; and (4) CDD-predicted protein function usually associated with Acps, such as lipases and serine protease inhibitors (Mueller *et al.* 2004). Criterion 3, the presence of a signal peptide, was relaxed when the beginning of the coding region could not be accurately identified (*e.g.*, no stop codon was detected upstream from the first methionine).

Nomenclature: *D. mayaguana* transcripts that matched an EST identified in the male reproductive tract *D. mojavensis* library were named with the same number but with the prefix *mayAcp*. In this way, the ortholog of *Acp1* in *D. mojavensis* was named *mayAcp1* and so on. *D. mayaguana* transcripts without a *D. mojavensis* ortholog were numbered consecutively from 54 (the number of unique ESTs found in the *D. mojavensis* reproductive tract library) on. *D. mayaguana* transcripts that were not classified as an Acp according to the criteria listed above received only the prefix *may* (not followed by *Acp*) and a number. When two or more *D. mayaguana* transcripts seemed to be orthologous to the same *D. mojavensis* genomic region (paralogs resulting from a duplication that probably happened after the split of the two lineages), they received the same number and a letter starting from *a*.

Homolog searches: Drosophila ortholog searches were carried out using tblastx and the newest assembly of the genomes of 12 Drosophila species available at the DroSpeGe website (http://insects.eugenes.org/species/blast/). Only sequences with an $E < 1e^{-05}$ were considered potential orthologs. All ESTs were also compared to the NCBI database using blastn and tblastx (Altschul *et al.* 1990) in searching for similar annotated genes in Drosophila and other organisms, also using a threshold of $E = 1e^{-05}$. To look more specifically at similarity with Acp genes of other Drosophila species, three databases were assembled using sequences available from GenBank (NCBI). The first one contained 40 *D. melanogaster* mRNA sequences, ranging in size from 80 to 2855 bp and representing 32 different genes that had been classified as Acps. Some of these sequences were obtained in a cDNA library (Mueller *et al.* 2005) and the remaining were annotated from the genome on the basis of previous knowledge of the gene or characterization in other studies (Begun and Lindfors 2005). The second Acp database included all the 239 cDNA sequences obtained in an accessory gland cDNA library of *D. simulans*, with lengths ranging between 28 and 746 bp (Swanson *et al.* 2001). The third similarly included all other 361 sequences obtained by Wagstaff and Begun (2005b) in two *D. mojavensis* cDNA libraries: one from the entire male reproductive tract and the other from testis only. These 361 *D. mojavensis* sequences represent 172 unique ESTs, ranging from 81 to 936 bp, among which 118 were found in the testis library. Among the reproductive-tract unique ESTs, 24 were classified

as Acp on the basis of tissue expression profiles by Wagstaff and Begun (2005b). Since this database is limited and the best hit in it may not be the best hit in the *D. mojavensis* genome, before suggesting orthology we checked whether the *D. mayaguana* query had the same best hit in the genome as compared to its corresponding *D. mojavensis* Acp best hit. Finally, to determine whether there were novel gene families within *D. mayaguana*, each sequence obtained in the accessory gland library was used as query in searches against all other sequences similarly identified. All the searches using Acp databases were locally run using both blastn and tblastx.

Molecular evolutionary analyses: Alignments of potential homologs were done using MAFFT v5.861 (Katoh *et al.* 2002; Katoh *et al.* 2005), followed by inspection and manual adjustment of codon alignment of gaps using MacClade v4.08 (Maddison and Maddison 2000). For clusters including four or more homologs, maximum-likelihood phylogenetic analyses were run on Paup* 4.0b10 (Swofford 2003), using the GTR + Γ model. An initial tree was generated using maximum parsimony and likelihood parameters were estimated from the data. Maximum-likelihood trees were obtained with five replicates of random stepwise addition. Pairwise d_N/d_S estimates were obtained with both the Nei–Gojobori (NG; Nei and Gojobori 1986) and the maximum-likelihood estimate (MLE; Yang *et al.* 1998) methods, using the program codeml included in the PAML v3.15 package (Yang 1997), by choosing M0 (Nsites = 0). The latter model (MLE) estimates ω, an estimator of d_N/d_S that accounts for transition/transversion rates and codon bias. For gene families including four or more sequences, the site models M1, M2, M7, M8, and M8a were used in tests for positive selection using codeml (Yang and Bielawski 2000; Swanson *et al.* 2003; Wong *et al.* 2004; Yang *et al.* 2005). In all tests we used the default initial values of ω (0.4), since it seems not to affect the results of the tests employed (Schully and Hellberg 2006). Statistical tests were performed by comparing the model likelihoods using the likelihood-ratio test (LRT).

RESULTS

Characterization *of D. mayaguana* Acps: Of the 600 clones that we sequenced, 91 potential unique Acp sequences (excluding alternative splicing forms) were found with an average length of 467 bp and ranging from 133 to 1229 bp (supplemental Table 1). The range and average of transcript length were similar to those obtained in accessory gland cDNA libraries of other Drosophila species (Swanson *et al.* 2001; Wagstaff and Begun 2005a), suggesting that the protocols used here did not affect library results in these matters. By fitting a Poisson model to the distribution of the number of clones per transcript, we estimated the total number of genes expressed in the accessory glands of *D. mayaguana* males to be 133. Considering the available information on all four Acp criteria, the final tally was 54 Acp candidates (supplemental text 2 and supplemental Table 1). Using this number of candidates, we estimated that the total number of Acps in this species is ~70 genes.

The results of tblastx searches using the 12 Drosophila genomes available (supplemental Figure 1) showed, as expected, that cross-species sequence conservation of Acps is highly influenced by phylogenetic relationships. The best hit was almost always in *D. mojavensis* and the second best in *D. virilis*, with a significant increase in *E*-values in the remaining species that did not differ much from one another in general (supplemental Table 2). Considering a cutoff value of $E = 1e^{-05}$ in tblastx and blastn searches, 87 of the 91 *D. mayaguana* transcripts had hits in *D. mojavensis*. This number drops to 39 in *D. virilis* and, in the remaining species, it ranged from 30 to 36 (supplemental Figure 1). About 35 (38%) transcripts did not have any hits in Drosophila genomes other than that of *D. mojavensis*. Most transcripts that were similar to annotated genes in the NCBI database had their best hits in Drosophila, many of which with unknown function (supplemental Table 2).

Conserved protein domains: More than two-thirds (61/91) of the sequences retrieved in the cDNA library of *D. mayaguana* accessory glands did not match any conserved protein domain in the databases when translated (supplemental Table 1). The remaining sequences matched 16 different domains. Among these matches, there were some domains normally found in Acps, domains that are expected in proteins of the accessory glands (although not directly related to reproduction), and domains that had never been reported among Acps. Among the first group, the most common domains matched were protease inhibitors (12 ESTs) belonging to two different families: Kazal-type serine protease inhibitor (9 ESTs) and BPT1/Kunitz-type serine protease inhibitor (3 ESTs). In addition to these, other domains commonly found in Acps included proteases (one serine protease and one metalloprotease), C-lectins (one), lipases (two), and cysteine-rich secretory proteins (CRISP, four). One transcript had a lysosomal thiol reductase domain. Although this domain has not been found in Acps of other species, it is related to the thioredoxins, a family that includes two *D. melanogaster* Acps (Mueller *et al.* 2004).

Domains that would be expected in Acps, although not directly related to reproduction, include a signal peptidase (SPC12) and an endoplasmatic reticulum protein domain (PDIa) usually found in organs with high secretory activity. Although ribosomal proteins are quite ubiquitous and have often been found in accessory gland libraries (Swanson *et al.* 2001; Wagstaff and Begun 2005b), only one sequence in the *D. mayaguana* library had a fragment encoding a domain of this type. This lack of ribosomal proteins is probably a result of the subtraction procedure and evidence of its success in reducing the number of housekeeping genes in the cDNA pool. One domain that has never been found in other Drosophila Acps, a fibrinogen-related domain (FReD), was found to be encoded by two transcripts that were among the most common in the library (*mayAcp64* and *mayAcp75*).

Comparison to other Drosophila Acps: In the searches using *D. simulans* and *D. melanogaster* Acps as the database, 11 *D. mayaguana* sequences had a hit. These hits

TABLE 1

Orthology of *D. mayaguana* ESTs to *D. mojavensis* Acps

D. mojavensis Acp	*E*-value[a]	No. of hits	*D. melanogaster* Acp
Acp1	$1e^{-107}$	1	*Acp53C14c*
Acp2	$8e^{-50}$	2	*Acp53C14c*
Acp3	$4e^{-14}$	2	—
Acp4	No hit	—	—
Acp5a	$3e^{-29}$	4	—
Acp7	$3e^{-62}$	1	—
Acp8	$5e^{-16}$	2	—
Acp11	$2e^{-22}$	1	—
Acp15	No hit	—	—
Acp16a	$1e^{-13}$	1	—
Acp17	No hit	—	—
Acp19	0.0	2	CG10284
Acp21a	No hit	—	—
Acp22	$7e^{-17}$	2	—
Acp23	$7e^{-27}$	6	—
Acp24	$8e^{-12}$	1	—
Acp25	$5e^{-81}$	2	*Acp53C14c*
Acp27a	No hit	—	—
Acp27b	No hit	—	—
Acp36	No hit	—	—
Acp42	$4e^{-40}$	1	—
Acp45	$2e^{-60}$	1	—
Acp48	No hit	—	—
Acp54	No hit	—	—
moj44	$1e^{-102}$	1	*lectin46b*

The first column shows all the ESTs classified as Acps by their authors (Wagstaff and Begun 2005b) plus the only other EST hit by a *D. mayaguana* transcript in the database (*moj44*).

[a] Smallest *E*-value when there were several hits.

included the Acps *lectin46Cb*, CG10284, CG14034, CG17097, *Acp53C14c*, and *Acp24Aa*. Some of these Acps have very conserved sequences and likely orthologs in all 12 Drosophila species with genome sequences available. For others, with intermediate *E*-values, the matching was probably due to the presence of gene regions encoding for the same protein domain and may not necessarily indicate orthology (supplemental Table 2).

While the genomic searches showed that 87 of 91 *D. mayaguana* ESTs had a hit in *D. mojavensis*, only 28 showed likely orthology (see materials and methods) to a *D. mojavensis* Acps (supplemental Table 3). Some of these 28 ESTs had the same hit in the *D. mojavensis* Acp database, suggesting gene duplications in the *D. mayaguana* lineage (Table 1). Most *D. mayaguana* transcripts with a *D. mojavensis* Acp ortholog had other evidence suggestive of an Acp status (Table 1; supplemental Table 3).

The *D. mojavensis* Acp database also included sequences that were not considered Acps by their authors' criteria, although they were obtained in the same reproductive-tract cDNA library as the Acps (Wagstaff and Begun 2005b). Only one *D. mayaguana* transcript showed likely orthology to a gene in this category, *moj44* (Table 1). However, the exclusion of *moj44* from the Acp set by Wagstaff and Begun (2005b) was based on lack evidence for inclusion according to their Acp criteria. Their main criterion for Acp status was an at least fivefold-higher expression in the accessory glands as opposed to testis and nonreproductive tissues. These data, obtained by quantitative PCR, were not available for *moj44*. Nevertheless, *moj44* has a predicted signal peptide (as determined by all three statistics used in program SignalP; supplemental text 2) and is similar to a *D. melanogaster* Acp (supplemental Table 2). Therefore, we suggest that *moj44* as its *D. mayaguana* ortholog *mayAcp44* should be classified as Acps. Similarly, we suggest an Acp status for *moj37*, a very abundant transcript in the Wagstaff and Begun (2005b) library that contains a lipase domain, commonly found among Acps.

No clear correlation was detected between the number of clones and expression levels as measured by quantitative PCR by Wagstaff and Begun (2005b) in *D. mojavensis*. Since we do not have quantitative PCR data for *D. mayaguana*, direct comparisons of expression are not possible. Looking at the quantitative PCR data, we observe that eight of the *D. mojavensis* Acps with the highest expression levels in the accessory glands had orthologs in *D. mayaguana* (in order: *Acp3*, *Acp7*, *Acp11*, *Acp5a*, *Acp1*, *Acp19*, *Acp45*, and *Acp2*). On the other hand, *Acp48* was also highly expressed in *D. mojavensis* but did not have an ortholog in the *D. mayaguana* library. Among the *D. mayaguana* ESTs with many clones, only two, *mayAcp64* and *may75* (39 and 10 copies, respectively), did not have hits in the *D. mojavensis* library.

Molecular evolution of *D. mayaguana* Acps: To investigate whether *D. mayaguana* Acps show rapid evolution and divergence patterns as a result of evolution by positive selection, we estimated the ratio of replacement to synonymous substitution rates in comparisons with *D. mojavensis* orthologs. ORF alignment was possible for 13 *D. mayaguana* Acps with an Acp ortholog in *D. mojavensis* (Table 2). The results obtained with the two different methods used (NG and MLE) are in general agreement, although MLE estimates tended to be a little larger (Table 2).

Of 13 ortholog pairs, we found $d_N/d_S > 1$ in 7 cases with an overall average of 1.25 (MLE). The theoretical threshold for determining whether a gene is under positive selection is $d_N/d_S > 1$. However, it is known that, due to substitution rate heterogeneity across sites, even genes with $d_N/d_S < 1$ may have some sites with adaptive evolution. As a matter of fact, $d_N/d_S > 0.5$ has been suggested as a more realistic cutoff on the basis of a compilation of published results of site model tests (Swanson *et al.* 2004). Among the Acps analyzed here, only *mayAcp44* had a $d_N/d_S < 0.5$. Our results suggest a trend of positive selection in *D. mayaguana* Acps, although for some of them ($d_N/d_S < 1$) only site model tests could confirm the existence of positively selected sites.

TABLE 2

d_N/d_S estimates between *D. mojavensis* and *D. mayaguana* orthologs

Acp	Size	NG (d_N, d_S)	MLE (d_N, d_S)	moj/ari/mul	Paralog
mayAcp1	366 (342)	0.850 (0.139, 0.163)	1.106 (0.154, 0.140)	1.374	Yes
mayAcp2a	366 (342)	0.500 (0.187, 0.375)	0.578 (0.204, 0.353)	0.971	Yes
mayAcp3a	171 (78)	1.865 (0.403, 0.216)	4.558 (0.566, 0.124)	0.932	Yes
mayAcp3b	171 (78)	1.422 (0.299, 0.210)	1.587 (0.343, 0.216)	0.932	Yes
mayAcp16d	276 (144)	0.747 (0.432, 0.578)	1.321 (0.595, 0.450)	0.808	Yes
mayAcp16e	276 (144)	0.492 (0.313, 0.637)	0.792 (0.436, 0.550)	0.808	Yes
mayAcp25	366 (342)	0.651 (0.158, 0.242)	0.744 (0.170, 0.229)	0.539	Yes
mayAcp7	402 (339)	0.682 (0.170, 0.250)	1.084 (0.205, 0.189)	0.808	No
mayAcp11	249 (210)	0.648 (0.282, 0.436)	0.790 (0.309, 0.391)	0.245	No
mayAcp19	648 (645)	0.904 (0.104, 0.115)	0.977 (0.108, 0.111)	1.242	No
mayAcp42	678 (546)	0.890 (0.336, 0.378)	1.127 (0.386, 0.343)	0.615	No
mayAcp44	558 (552)	0.301 (0.06, 0.230)	0.259 (0.066, 0.255)	NE	No
mayAcp45	540 (414)	1.321 (0.241, 0.182)	1.367 (0.253, 0.185)	0.915	No
Average		0.867	1.253	0.849	

Size, length of alignment used in base pairs before and after gap deletion (in parentheses); NG, Nei–Gojobori method; MLE, maximum-likelihood estimate; moj/ari/mul, estimates obtained by MLE in comparisons between *D. mojavensis*, *D. arizonae*, and *D. mulleri* as estimated by Wagstaff and Begun (2005b); NE, estimate not available.

Wagstaff and Begun (2005b) obtained similar values when comparing sequences of *D. mojavensis*, *D. arizonae*, and *D. mulleri* for 19 Acps (average $d_N/d_S = 0.93$). These authors noted that *D. mojavensis* had significantly faster protein evolution rates when compared to *D. arizonae* and *D. mulleri*. To check whether the high d_N/d_S values obtained in *D. mayaguana* comparisons were inflated because *D. mojavensis* orthologs were used, we estimated the ratios comparing *D. mayaguana* and *D. arizonae* sequences. The average was 1.08, not significantly different from the average in the comparisons with *D. mojavensis* (1.13) using the same alignment (Wilcoxon rank-sum test, $W = 45$, $P = 0.931$). For comparison, the average d_N/d_S of testis genes in *D. mojavensis* and *D. arizonae* comparisons was 0.19 (Wagstaff and Begun 2005b). The d_N/d_S averages for Acps of the *repleta* group are considerably high even if compared to other Acp sets. Estimates between *D. melanogaster* and *D. yakuba* Acps had an average of 0.41 (Mueller *et al.* 2005).

Gene families: In the BLAST searches for identifying gene families among the unique *D. mayaguana* accessory gland ESTs, we found that 46 ESTs had at least one hit ($E < 1e^{-10}$) to other *D. mayaguana*'s EST. The average length of these duplicated transcripts was significantly smaller than the average of nonduplicated sequences (*t*-test, $t = -2.825$, $P = 0.006$), in accordance with previous findings that shorter genes are more often duplicated (Nembaware *et al.* 2002). On the basis of the results of the BLAST searches, these 46 sequences can be arranged into 13 clusters or gene families (cluster 1–13 in Table 3), in which each member hit all or almost all the other members in its cluster.

Most paralogs in a same cluster had their best BLAST hit to the same locus in the *D. mojavensis* genome. This

TABLE 3

Clusters of similar genes coexpressed in the accessory glands of *D. mayaguana*

Cluster	*N*	Members	*E*-value	*D. mojavensis* genomic region	*D. mojavensis* Acp
1	2	*mayAcp64*, *mayAcp75*	e^{-37}	Adjacent	—
2	2	*mayAcp68a*, *mayAcp68b*	e^{-83}	Same	—
3	2	*mayAcp69a*, *mayAcp69b*	e^{-76}	Same	—
4	2	*mayAcp3a*, *mayAcp3b*	e^{-81}	Same	*Acp3*
5	2	*mayAcp67a*, *may67b*	e^{-113}	Same	—
6	2	*may79*, *may102*	e^{-61}	Different	—
7	2	*mayAcp16b1*, *mayAcp16b2b*	e^{-51}	Same	*Acp16b*, *Acp24*
8	3	*mayAcp78a*, *may78b*, *c*	e^{-50}–e^{-115}	Same	—
9	3	*mayAcp8a*, *may8b*, *may83*	e^{-16}–e^{-56}	Same/adjacent	*Acp8*
10	3	*mayAcp22a*, *mayAcp66a*, *may66b*	e^{-26}–e^{-53}	No hit	—
11	4	*mayAcp1*, *mayAcp2a*, *mayAcp2b*, *mayAcp25*	e^{-23}–e^{-55}	Adjacent	*Acp1*, *Acp2*, *Acp25*
12	9	*mayAcp5a–c*, *may5d*, *mayAcp23a–e*	e^{-10}–e^{-84}	Different	*Acp5a*, *Acp23*
13	10	*mayAcp59a–c*, *mayAcp60a–e*, *mayAcp61*	e^{-10}–e^{-146}	Same/adjacent	—

TABLE 4

d_N/d_S ratios in pairwise comparisons between members of gene clusters with two to three sequences

			NG			MLE		
Cluster	*N*	bp	d_N/d_S	d_N	d_S	ω	d_N	d_S
1	2	609 (417)	0.61	0.45	0.73	0.61	0.38	0.62
2	2	546	1.96	0.18	0.09	2.14	0.19	0.09
3	2	336 (177)	1.55	0.08	0.05	1.14	0.07	0.06
4	2	190 (129)	2.82	0.10	0.04	3.39	0.11	0.03
5	2	549 (432)	0.71	0.14	0.20	1.01	0.16	0.16
7	2	213 (210)	1.25	0.19	0.15	1.93	0.21	0.11
8	3	270	3.06, 2.97, ∞	0.03–0.27	0.00–0.09	6.21, 5.98, ∞	0.30–0.03	0.05–0.00
9	3	225 (147)	0.82	0.16	0.19	1.04	0.21	0.21
10	3	264	0.30, 0.61, 0.90	0.04–0.29	0.13–0.38	0.39, 0.75, 1.21	0.04–0.33	0.11–0.34
Acp16a/b	2	276 (150)	1.29	0.33	0.25	2.20	0.61	0.28
Acp27a/b	2	351 (303)	2.25	0.11	0.05	3.60	0.12	0.03

Cluster numbers correspond to *D. mayaguana* clusters in Table 3. *Acp16a/b* and *Acp27a/b* are *D. mojavensis* gene clusters. bp, number of sites in the alignment before and after gap deletion (in parentheses). *N*, number of sequences in the cluster; NG, Nei–Gojobori method; MLE, maximum-likelihood estimate. For cluster 9, the analysis included only the two most similar sequences because the third sequence was too divergent and alignment of ORFs did not seem reliable. The symbol ∞ means that $d_S = 0$ and therefore it is not possible to calculate d_N/d_S.

result suggests extensive gene duplication in the lineage leading to *D. mayaguana* after its split from the *D. mojavensis* lineage. It is interesting to note that all clusters but one included sequences that were classified as Acp. Three of the clusters detected here had also been detected in *D. mojavensis*, including those of *Acp16* (cluster 7), *Acp5* (cluster 12), and *Acp1/Acp2/Acp25* (cluster 11) (Wagstaff and Begun 2007). Here we observed, in addition, that *Acp16* has similarity to *Acp24* and that *Acp5* has similarity to *Acp23*. Further details on these and other *D. mayaguana* gene families are described in supplemental text 3.

Evolution of duplicated Acps in the *D. repleta* group: One pattern that often appears in theoretical models and empirical analyses of gene families is that, at some point after duplication, paralogs experience relaxed or positive selection (Castillo-Davis *et al.* 2004; Jordan *et al.* 2004; Lynch and Katju 2004). This post-duplication selection contrasts with a more constrained regime of purifying selection that acts on most sites of coding genes (*e.g.*, *D. melanogaster* × *D. yakuba* non-Acp genes have an average $d_N/d_S = 0.082$; Mueller *et al.* 2005). To investigate whether the post-duplication pattern is also observed in duplicated Acp genes in the *repleta* group, we estimated d_N/d_S between paralogs. We expected that if duplicated Acps follow the pattern, d_N/d_S should be ≥1 in most of the paralog clusters. This analysis was carried out for 14 gene families, including 12 described here (cluster 6 was not included in the analysis because it was the only one that did not have any members classified as Acp) and two *D. mojavensis* Acp paralog pairs (Table 4 and Table 5; alignments used are available upon request).

Of the 11 clusters with up to three members, 10 showed MLE results compatible with relaxed or positive selection ($d_N/d_S \geq 1$) for at least some of the paralog comparisons within the cluster (Table 4). The NG method corroborated these results for seven clusters. Positive selection is very likely acting on clusters 4 and 8, which showed about three times more nonsynonymous changes as compared to synonymous changes between paralogs. In almost all *D. mayaguana* Acp clusters with $d_N/d_S > 1$ (Table 4), duplication is likely to be recent as suggested by the coincident location of their best hits (from the BLAST searches) in the *D. mojavensis* genome (Table 3). This would imply that relaxed/positive selection occurs soon after Acp gene duplication. Accordingly, among the clusters with up to three members, the cluster with lower d_N/d_S (cluster 1) was the only one in which duplication most probably occurred before the split of the *D. mojavensis* and *D. mayaguana* lineages.

We also obtained d_N/d_S for genes in clusters with more than three members. For these clusters, an increased number of sequences allowed for statistical tests for positive selection on the basis of site models (Yang *et al.* 1998; Yang and Swanson 2002). Although cluster 7 composes only two genes in *D. mayaguana*, it was also tested because orthologs were available for *D. mojavensis* (*Acp16a*, *Acp16b*, and *Acp24*) and for *D. arizonae* (*Acp16b*). The results of the tests are shown in Table 5. Strong statistical support was obtained for the presence of positively selected sites in genes in clusters 7, 12, and 13. Cluster 13 was the only group included in the site model tests that had a region encoding a conserved protein domain. Of its 15 sites in this sequence, with a posterior probability of $P > 0.95$ of being under selection, 7 are within the region encoding a conserved protease inhibitor domain.

Cluster 11, which includes four *D. mayaguana* Acps, showed contrasting results with the remaining clusters.

TABLE 5

Results of tests for positive selection for clusters with more than three sequences

Models	lnL	lnL	LRT	*P*	BEB	ω_2	p_2
			Cluster 7 ($\omega = 1.28$, 267 bp)				
M1 × M2	−806.894	−797.251	19.287	<0.001	5 (3)	6.65	0.299
M7 × M8	−807.199	−797.251	19.897	<0.001	6 (4)	6.66	0.299
M8a × M8	−806.894	−797.251	19.287	<0.001			
			Cluster 11 ($\omega = 0.73$, 363 bp)				
M1 × M2	−2665.003	−2662.747	4.512	NS	NA	NA	NA
M7 × M8	−2664.758	−2662.136	5.245	NS	NA	NA	NA
			Cluster 12 ($\omega = 1.44$, 147 bp)				
M1 × M2	−898.693	−880.911	35.564	<0.0001	9 (7)	4.39	0.502
M7 × M8	−899.766	−880.802	37.929	<0.0001	13 (8)	4.28	0.522
M8a × M8	−898.698	−880.802	35.793	<0.0001			
			Cluster 13 ($\omega = 1.26$, 261 bp)				
M1 × M2	−1226.579	−1191.778	69.602	<0.0001	14 (8)	10.65	0.234
M7 × M8	−1228.047	−1191.789	72.515	<0.0001	15 (4)	10.82	0.234
M8a × M8	−1226.594	−1191.789	69.611	<0.0001			

lnL, log likelihood of null and alternative hypothesis; LRT, likelihood-ratio test; BEB, number of sites with $P > 95\%$ to be under positive selection by the Bayes empirical Bayes estimate (in parentheses, number of sites with $P > 99\%$); ω_2, average ω of the class of sites with $\omega > 1$ (class 2); p_2, proportion of sites in class 2.

The overall ω among paralogs in *D. mayaguana* was 0.689, among the lowest in *D. mayaguana* paralog comparisons. To obtain a more accurate test for positive selection in genes of cluster 11, we included the sequences of homologs in *D. mojavensis*, *D. arizonae*, and *D. mulleri* in these analyses. Both the comparisons of M1 × M2 and M7 × M8 did not support the presence of sites evolving under positive selection (Table 5). This gene family has Acp orthologs in all Drosophila species that are always found in duplicates (Wagstaff and Begun 2005a; see supplemental text 3). This pattern of conservation indicates strong evolutionary constraints. Contrary to the general pattern (see below), orthologs in this cluster (mean ω ± standard deviation = 0.846 ± 0.232) seem to be diverging faster than paralogs (ω = 0.409 ± 0.310).

Trends in Acp gene family evolution: Two main models have been proposed to explain the retention of paralogs after a gene duplication event: neofunctionalization and subfunctionalization (reviewed in Lynch and Katju 2004). These two models have very different predictions concerning evolutionary forces acting on the duplicated gene copies. In the subfunctionalization model, the duplicates accumulate degenerative mutations in different parts of the gene (Force *et al.* 1999; Prince and Pickett 2002). In this way, different independent subfunctions acquired by the ancestral gene will be divided between the two new copies. This can lead to differential expression patterns (a case that does not apply to Acps) or to complementary rescue of the ancestral function by the combined action of the gene duplicates. The subfunctionalization model does not require positive selection for the maintenance of the two gene copies, but rather predicts that these two copies, or at least part of them, will be under purifying selection (Lynch and Katju 2004).

The neofunctionalization model (Ohno 1970) suggests the gain of a new function by one of the paralogs as a consequence of accumulated mutations, while the other copy retains the ancestral function. The evolutionary forces involved in the gain of a new function by one of the duplicates lead to specific predictions of nucleotide substitution patterns. One of these predictions is asymmetric evolution of paralogs, assuming that one paralog will have evolutionary constraints related to the maintenance of the original function of the duplicated gene. To test this prediction, we compared the d_N/d_S ratios of five *D. mayaguana* and one *D. mojavensis* paralog pair in relation to an ortholog (Table 6). Except for one comparison (*mayAcp67*), in all other pairs the evolutionary rate differences were >60%.

Another prediction of the neofunctionalization model is that, while paralogs in the same genome will have relaxed or positive evolution, orthologs will evolve more slowly due to functional constraints. In this way it is expected that d_N/d_S between paralogs will be higher than the same ratio between orthologs. We compared between-paralog with between-ortholog ratios using each of the duplicates (paralog 1 and paralog 2) and the average of these ratios (Table 6). These comparisons showed that d_N/d_S was significantly higher between paralogs than between orthologs in comparisons including the slower-evolving paralog or the average of the paralogs' d_N/d_S (Wilcoxon signed-rank test, one tail, $P = 0.015$). This result is in agreement with differential evolutionary constraints among paralogs.

TABLE 6

d_N/d_S (ω) in ortholog and paralog comparisons of *D. mayaguana* and *D. mojavensis* Acp gene clusters

Acp cluster	OP1	OP2	Average OP	P1P2	Ortholog	P1P2 d_S
mayAcp2	0.960	0.578	0.769	0.741	*mojAcp2*	0.275
mayAcp3	3.644	2.088	2.866	3.392	*mojAcp3*	0.032
mayAcp16	1.321	0.792	1.057	2.086	*mojAcp16b*	0.116
mayAcp67	0.512	0.494	0.503	1.214	moj pred[a]	0.127
mayAcp69	1.579	0.785	1.182	1.567	moj pred[a]	0.134
mojAcp16	1.411	0.873	1.142	3.662	*mojAcp24*[b]	0.107
Average	1.571	0.935	1.253	2.110		

O, ortholog; P1, paralog 1; P2, paralog 2.

[a] Predicted genes based on ORFs identified in the highly similar (likely orthologous) *D. mojavensis* genomic region; the second BLAST hit had a much higher *E*-value.

[b] In the absence of orthologs equally distant from *mojAcp16a/b*, we used comparisons with another, more distant paralog.

We observed a trend of decreasing d_N/d_S in ortholog comparisons (OP1 and OP2 in Table 6) with increasing d_S between paralogs, a rough estimate of paralog age (P1P2 d_S in Table 6). This correlation was nonsignificant ($P = 0.15$), probably as a consequence of the small number of comparisons. Nevertheless, the trend is in accordance with the adaptive evolution of paralogs soon after duplication that affects ortholog divergence. However, this effect on ortholog divergence seems to be transitory. Accordingly, we did not find a difference in ortholog d_N/d_S ratio between genes with and without a paralog (Wilcoxon rank-sum test: $P = 0.792$; genes used in the analysis are the ones in Table 2).

DISCUSSION

Drosophila Acps have become a model for molecular evolutionary studies. The interest in these proteins in large part is due to their rapid evolutionary rates in terms of nucleotide substitution, gene duplication, and gene turnover (Swanson *et al.* 2001; Holloway and Begun 2004; Begun and Lindfors 2005; Mueller *et al.* 2005; Haerty *et al.* 2007). The availability of several Drosophila genome sequences and the ability to selectively add critical taxa to a study, such as *D. mayaguana*, allow for more precise testing of hypotheses about the evolutionary dynamics of these diverse proteins in both a phylogenetic and a functional context.

Here we address three main questions on the *repleta* group Acps. First, we checked whether the high rates of protein evolution previously observed in *D. mojavensis* and its sister species could be extended to other, more distantly related members of the *repleta* group. This is an important question since it may help in identifying the biological causes for the observed evolutionary pattern. Second, we addressed the possibility of functional divergence of the Acp set between the *repleta* and *melanogaster* groups and, therefore, whether functional divergence could be related to the reproductive differences between the two species groups. Finally, we took advantage of the considerable number of recently duplicated *D. mayaguana* Acps to determine whether the maintenance of these genes could be explained by the neofunctionalization hypothesis.

Reproductive biology of flies and the evolution of Acps: One of the motivations for studying Acps in a species of the *repleta* group was to provide data for comparison between the Drosophila groups with different biologies. In this case, a pertinent question is whether the reproductive strategy differences observed between the *melanogaster* and the *repleta* group could be associated with particular characteristics of the Acp set. Theoretical explanations for the evolutionary patterns observed for Acps rely on selective forces that originate in the interactions between male and female reproductive molecules after copulation. Higher remating frequency increases the chance of occurrence and the intensity of these interactions. Therefore, it was expected that higher remating rates would lead to faster evolutionary rates in Acps. Here we provide support for this hypothesis by revealing further evidence of (at least twofold) higher d_N/d_S ratios in Acps of species of the *repleta* group as compared to the *melanogaster* group (Mueller *et al.* 2005). Although more difficult to compare between groups, our data also suggest high turnover rates and differences in gene expression level between *repleta* species that diverged relatively recently, *i.e.*, ~10 million years ago (Russo *et al.* 1995).

What causes the extremely fast evolutionary rates observed in Acps of the *repleta* group? A likely hypothesis is that higher remating rates increase selective pressure on sperm and accompanying substances through sperm competition and male × female antagonistic coevolution (Wagstaff and Begun 2005b). Alternative explanations include ecological and demographic differences between the two groups. While differences in demographics are difficult to quantify and there are no reliable estimates for the species discussed here, it could be expected that the desert Drosophila of the *repleta* group have smaller populations and are more prone to

population fluctuations due to their ecological specialization (Wasserman 1982). Population size does not exert a large influence on the probability and time of fixation of new advantageous mutations, but it significantly affects the chances of fixation of neutral or quasi-neutral mutations due to genetic drift (Ohta 1973, 1993). Therefore, smaller population sizes in the *repleta* group could alternatively explain the higher d_N/d_S estimates observed. Demographic factors, however, would affect the genome as a whole and not only Acps. Further studies on non-Acp genes of the *repleta* group will shed light on these hypotheses.

Acps in the *repleta* group: Wagstaff and Begun (2005b), on the basis of a subset of the *D. mojavensis* Acps, suggested that Acps in the *repleta* group were functionally divergent from Acp genes in the *melanogaster* group. Their suggestion was based on the low incidence in the *D. mojavensis* male reproductive tract library of ESTs carrying encoding regions of protein domains usually found in Acps of other Drosophila groups, such as protease inhibitors, lipases, and proteases (Mueller *et al.* 2004). The more comprehensive accessory gland library obtained for *D. mayaguana* does not support a major functional divergence of the *repleta* group Acps. Among the *D. mayaguana* ESTs, 30% of the Acp genes had a region encoding for one of these three domains, in addition to other domains that were also found in the *melanogaster* group Acps, including C-type lectin and CRISP. Our results, however, are in agreement with those of Wagstaff and Begun (2005b) in that many of the transcripts identified in the *D. mayaguana* accessory glands have unknown function. In fact, this pattern is a general characteristic of the Acp sets of all the Drosophila species studied so far.

Although some differences in the number of genes in each functional class were observed (*e.g.*, *D. mayaguana* has more protease inhibitors and fewer lipases as compared to *D. melanogaster*), it is difficult to assess their significance and whether they could be associated with differences in reproductive biology between the two groups. The most striking result of functional difference is the presence among the *D. mayaguana* Acps of two transcripts, both with a high number of clone copies, carrying a gene region encoding for a FReD domain. This domain, which has never been found in Acps, is involved in the formation of blood clots in mammals. It is possible that these proteins are involved in the formation of the vaginal mass, common to species of the *repleta* group, but absent from the *melanogaster* group. Nevertheless, the vaginal mass is also observed in *D. mojavensis*, but no Acp containing this motif has been detected in this species.

An interesting result was the low similarity between the *D. mayaguana* and the *D. mojavensis* Acp sets despite the close relationship of the two species. Since 95% of the *D. mayaguana* transcripts seemed to have an ortholog in the *D. mojavensis* genome, these library differences would be mostly due to differences in expression. According to the results shown here, possibly >50% of the genes expressed in the accessory gland of one species may not be expressed in the same organ of the other species. Nine of 24 (37.5%) *D. mojavensis* Acps did not have an Acp homolog in *D. mayaguana*, some of which were among the ones with the highest expression levels in the accessory glands of the former species. These expression differences could be another consequence of the increased selective pressure on Acps of the *repleta* group.

Alternatively, library differences can be attributed to methodological random bias. Considering the difference between unique sequences in the *D. mayaguana* library and total gene number estimates, it is possible that by chance some orthologs were not represented in one of the libraries. Another likely source of bias is the difference in the developmental stage of flies used in the two libraries. The *D. mojavensis* library was done with 5-day-old flies, which are not reproductively mature, while we used a pool of flies ranging from 7- to 12-day-old flies. Therefore, the degree of cross-species conservation of the Acp transcriptome in the *repleta* group is still unknown, although the results of the comparison between *D. mojavensis* and *D. mayaguana* suggest differences at both the gene identity and the gene expression level.

Paralog evolution and neofunctionalization: Fifty percent of the *D. mayaguana* transcripts are members of gene families. This high incidence of duplicated Acps may, in part, be related to a mutation bias toward duplication of short genes (Nembaware *et al.* 2002; Lynch and Katju 2004). Gene duplication is believed to be the main mechanism of emergence of new genes and biological functions. Only recently, with the availability of genomic sequences, has it been possible to study the evolution of duplicated genes in a broader perspective (reviewed in Jordan *et al.* 2004). The patterns of paralog evolution, however, are still under debate and analyses of empirical data have given contradictory results (Dermitzakis and Clark 2001; Kondrashov *et al.* 2002; Castillo-Davis *et al.* 2004; Jordan *et al.* 2004). One problem of genomewide analyses is that they do not usually distinguish recent and old duplicates but rather pool together very different functional categories that may have very different evolutionary constraints. A better scenario for study of the evolution of recently duplicated genes would be one where the paralogs have only one ortholog in a closely related species (Lynch and Katju 2004). This is the case of most duplicated *D. mayaguana* Acps.

Our analyses revealed evolutionary patterns compatible with the neofunctionalization hypothesis for the maintenance of duplicated genes (Ohno 1970) for most duplicated Acps, such as adaptive sequence divergence and asymmetrical evolutionary rates. Since, the Acp paralogs are expressed in the same organ and seem to have retained the same biological function, we

suggest that the "neofunctionalization" in this case would be more of a "neospecificity" to female or other male sperm proteins.

Another pattern expected by the neofunctionalization hypothesis is that duplicated genes evolve more slowly in relation to their orthologs than genes in single copy. While some empirical studies corroborate this prediction (Davis and Petrov 2004; Jordan *et al.* 2004), others show that duplicated genes actually have faster evolution than singletons (Kondrashov *et al.* 2002; Nembaware *et al.* 2002). These studies, however, are based on comparisons between species much more divergent than *D. mayaguana* and *D. mojavensis* and/or on paralogs with a wider range of ages. Among the recently duplicated paralogs studied here, we found that amino acid evolution between paralogs is in fact faster than between these duplicated genes and their orthologs. Yet we found no significant difference in ortholog d_N/d_S between genes with and without paralogs. Our results suggest that evolutionary rates of duplicated genes show an increase after duplication that persists a short period of time. This is in contrast with results by Nembaware *et al.* (2002), who found increased d_N/d_S for genes that have intermediate paralogs ($0.34 \leq d_S \leq 0.74$), but not for genes with recent paralogs.

The gene family cluster that includes *Acp1*, *Acp2*, and *Acp25* is apparently under different evolutionary forces when compared to the remaining clusters. The site model tests used here show no support for adaptive evolution in the divergence of these paralogs. What selective pressure might account for the maintenance of duplicated copies in this gene family that has orthologs in all the Drosophila species studied so far? It is possible that this cluster represents a case of gene duplicate retention due to subfunctionalization. An alternative explanation comes from the observation that highly expressed genes have a higher chance of being duplicated and maintained in yeast (Seoighe and Wolfe 1999; Davis and Petrov 2004). The benefit in the dosage level may account for their maintenance in the genome in the absence of neo- or subfunctionalization. This could be the case of this gene family, since *Acp1* seems to have high expression levels in both *D. mayaguana* and *D. mojavensis*. The other genes of this family were found in intermediate numbers of clones in the *D. mayaguana* accessory gland library.

We acknowledge the insightful comments of three anonymous reviewers on an earlier version of the manuscript. Funds for this research were provided by the Sackler Institute for Comparative Genomics (American Museum of Natural History), the Cullman Program in Molecular Systematics, and the National Science Foundation (DEB 0129105 to R.D.). F.C.A. was supported by the Henry McCraken Fellowship (New York University).

LITERATURE CITED

Almeida, F., 2007 Molecular evolution of accessory gland proteins in the *Drosophila repleta* group and their potential role in speciation. Ph.D. Thesis, New York University, New York.

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

Begun, D. J., and H. A. Lindfors, 2005 Rapid evolution of genomic Acp complement in the *melanogaster* subgroup of *Drosophila*. Mol. Biol. Evol. **22:** 2010–2021.

Bendtsen, J. D., H. Nielsen, G. von Heijne and S. Brunak, 2004 Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. **340:** 783–795.

Castillo-Davis, C. I., D. L. Hartl and G. Achaz, 2004 cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res. **14:** 1530–1536.

Chapman, T., and S. J. Davies, 2004 Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. Peptides **25:** 1477–1490.

Davis, J. C., and D. A. Petrov, 2004 Preferential duplication of conserved proteins in eukaryotic genomes. PLoS Biol. **2:** E55.

Dermitzakis, E. T., and A. G. Clark, 2001 Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. **18:** 557–562.

Durando, C. M., R. H. Baker, W. J. Etges, W. B. Heed, M. Wasserman *et al.*, 2000 Phylogenetic analysis of the repleta species group of the genus *Drosophila* using multiple sources of characters. Mol. Phylogenet. Evol. **16:** 296–307.

Force, A., M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan *et al.*, 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics **151:** 1531–1545.

Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram *et al.*, 2007 Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics **177:** 1321–1335.

Holloway, A. K., and D. J. Begun, 2004 Molecular evolution and population genetics of duplicated accessory gland protein genes in *Drosophila*. Mol. Biol. Evol. **21:** 1625–1628.

Jordan, I. K., Y. I. Wolf and E. V. Koonin, 2004 Duplicated genes evolve slower than singletons despite the initial rate increase. BMC Evol. Biol. **4:** 22.

Katoh, K., K. Misawa, K. Kuma and T. Miyata, 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:** 3059–3066.

Katoh, K., K. Misawa, H. Toh and T. Miyata, 2005 MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. **33:** 511–518.

Knowles, L. L., and T. A. Markow, 2001 Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert *Drosophila*. Proc. Natl. Acad. Sci. USA **98:** 8692–8696.

Kondrashov, F. A., I. B. Rogozin, Y. I. Wolf and E. V. Koonin, 2002 Selection in the evolution of gene duplications. Genome Biol. **3:** research0008.1—0008.9

Lynch, M., and V. Katju, 2004 The altered evolutionary trajectories of gene duplicates. Trends Genet. **20:** 544–549.

Maddison, D., and W. Maddison, 2000 *MacClade 4: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.

Markow, T. A., and P. F. Ankney, 1988 Insemination reaction in Drosophila: found in species whose males contribute material to oocyte before fertilization. Evolution **42:** 1097–1101.

Mueller, J. L., D. R. Ripoll, C. F. Aquadro and M. F. Wolfner, 2004 Comparative structural modeling and inference of conserved protein classes in *Drosophila* seminal fluid. Proc. Natl. Acad. Sci. USA **101:** 13542–13547.

Mueller, J. L., K. R. Ram, L. A. McGraw, M. C. Bloch Qazi, E. D. Siggia *et al.*, 2005 Cross-species comparison of Drosophila male accessory gland protein genes. Genetics **171:** 131–143.

Nei, M., and T. Gojobori, 1986 Simple method for estimating the numbers of synonymous and non-synonymous substitutions. Mol. Biol. Evol. **17:** 73–118.

Nembaware, V., K. Crum, J. Kelso and C. Seoighe, 2002 Impact of the presence of paralogs on sequence divergence in a set of mouse-human orthologs. Genome Res. **12:** 1370–1376.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Berlin/Heidelberg, Germany/New York.

Ohta, T., 1973 Slightly deleterious mutant substitutions in evolution. Nature **246:** 96–98.

Ohta, T., 1993 Amino acid substitution at the Adh locus of *Drosophila* is facilitated by small population size. Proc. Natl. Acad. Sci. USA **90:** 4548–4551.

Pitnick, S., G. S. Spicer and T. Markow, 1997 Phylogenetic examination of female incorporation of ejaculate in *Drosophila*. Evolution **51:** 833–845.

Prince, V. E., and F. B. Pickett, 2002 Splitting pairs: the diverging fates of duplicated genes. Nat. Rev. Genet. **3:** 827–837.

Russo, C. A., N. Takezaki and M. Nei, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol. Biol. Evol. **12:** 391–404.

Schully, S. D., and M. E. Hellberg, 2006 Positive selection on nucleotide substitutions and indels in accessory gland proteins of the *Drosophila pseudoobscura* subgroup. J. Mol. Evol. **62:** 793–802.

Seoighe, C., and K. H. Wolfe, 1999 Yeast genome evolution in the post-genome era. Curr. Opin. Microbiol. **2:** 548–554.

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

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

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

Swofford, D. L., 2003 *PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods)*. Sinauer Associates, Sunderland, MA.

Throckmorton, L., 1975 The phylogeny, ecology, and geography of *Drosophila*, pp. 421–469 in *Handbook of Genetics*, edited by R. King. Plenum, New York.

Wagstaff, B. J., and D. J. Begun, 2005a Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol. Biol. Evol. **22:** 818–832.

Wagstaff, B. J., and D. J. Begun, 2005b Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. Genetics **171:** 1083–1101.

Wagstaff, B. J., and D. J. Begun, 2007 Adaptive evolution of recently duplicated accessory gland protein genes in desert Drosophila. Genetics **177:** 1023–1030.

Wang, J. P., B. G. Lindsay, J. Leebens-Mack, L. Cui, K. Wall *et al.*, 2004 EST clustering error evaluation and correction. Bioinformatics **20:** 2973–2984.

Wasserman, M., 1982 Evolution and speciation in selected species groups: evolution of the *repleta* group, pp. 61–139 in *The Genetics and Biology of Drosophila*, edited by M. Ashburner, H. L. Carson and J. N. Thompson, Jr. Academic Press, London.

Wolfner, M. F., 2002 The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. Heredity **88:** 85–93.

Wong, W. S., Z. Yang, N. Goldman and R. Nielsen, 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics **168:** 1041–1051.

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

Yang, Z., and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:** 496–503.

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

Yang, Z., R. Nielsen and M. Hasegawa, 1998 Models of amino acid substitution and applications to mitochondrial protein evolution. Mol. Biol. Evol. **15:** 1600–1611.

Yang, Z., W. S. Wong and R. Nielsen, 2005 Bayes empirical Bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22:** 1107–1118.

Communicating editor: M. Aguadé