To investigate the potential importance of gene duplication in *D. melanogaster* accessory gland protein (*Acp*) gene evolution we carried out a computational analysis comparing annotated *D. melanogaster* *Acp* genes to the entire *D. melanogaster* genome. We found that two known *Acp* genes are actually members of small multigene families. Polymorphism and divergence data from these duplicated genes suggest that in at least four cases, protein divergence between *D. melanogaster* and *D. simulans* is a result of directional selection. One putative *Acp* revealed by our computational analysis shows evidence of a recent selective sweep in a non-African population (but not in an African population). These data support the idea that selection on reproduction-related genes may drive divergence of populations within species, and strengthen the conclusion that *Acps* may often be under directional selection in *Drosophila*.

At least three classes of models have been proposed to explain the evolutionary processes for the retention and subsequent divergence of gene duplicates. Lynch and Force (2000) suggest that ancestral genes with multiple functions in different tissues or developmental stages may have high rates of retention of duplicates under mutation-selection balance. In this model, degenerative mutations result in subfunctionalization, which favors retention and subsequent evolution of duplicates. A second class of models invokes fixation of duplications by genetic drift (e.g., Lynch and Conery 2003; Walsh 2003). Finally, a third class of models relies on, new, beneficial mutations driving adaptive divergence (and thus retention) of duplicates (Hughes 1994). One would expect new duplicates from classes of proteins under chronic directional selection to have unusually high fixation probabilities because a higher proportion of new mutations may be beneficial in such genes. For example, if reproduction-related proteins experience directional selection more frequently than other proteins (Civetta and Singh 1998; Nurminsky et al. 1998; Swanson and Conery 2003; Walsh 2003), then perhaps a large number of duplicate reproduction-related genes spread through populations and diverge under directional selection.

We investigated duplication and divergence in reproduction-related accessory gland proteins genes (*Acps*) in *Drosophila*. *Acps* are male-specific seminal fluid proteins that affect multiple aspects of female physiology and behavior (for review see Wolfner 1997). We carried out Blast comparisons of the 13 annotated *Acps* (see Methods) to the *D. melanogaster* reference sequence (Flybase Consortium 2003). These Blast analyses suggested that two genes, *Acp29AB* and *Acp53Ea*, are members of small multigene families.

**E-values** returned from the tBlastN search (default parameters) with *Acp29AB* as the query sequence were $1.5 \times 10^{-47}$ and $2.6 \times 10^{-35}$ for *Lectin29Ca* and *Lectin30A*, respectively. Intraspecific paralogous protein divergence was, on average, 31% between *Acp29AB* and *Lectin29Ca*, 35.5% between *Acp29AB* and *Lectin30A*, and 38% between *Lectin29Ca* and *Lectin30A*. *Lectin29Ca* is 356 bases distal to *Acp29AB* and *Lectin30A* is approximately 1 Mbase distal to these tandem duplicates. *Acp29AB* is 234 amino acids, while *Lectin29Ca* and *Lectin30A* are 236 and 223 amino acids long, respectively. Each gene is composed of a single exon. Our analysis of *Lectin30A* and comparison to its paralogs suggested that the 5′ end was incorrectly annotated. We confirmed this hypothesis by RACE, and we used our annotation in all analyses. The three members of the *Acp29AB* family are predicted to be lectin galactose binding proteins (Theopold et al. 1999) and to have a signal sequence (SignalP v2.0, Nielsen et al. 1997). The tBlastN search returned several other more distantly related putative *Acp29AB* paralogs, primarily lectins (*Lectin21Ch*, *Lectin24Db*, *Lectin22C*, *Lectin21Ca*, *Lectin24A*, *Lectin28C*, and CG15818). However, we will not present data from these genes in this report.

E-values returned from the tBlastN search with *Acp53Ea* as the query sequence were $2.1 \times 10^{-5}$ for CG8626 and $9.4 \times 10^{-4}$ for CG15616. CG8626 and CG15616 will hereafter be referred to as *Acp53C14a* and *Acp53C14b*, respectively, based on putative function, genomic location, and gene structure. Another more highly diverged putative duplicate identified by B. Wagstaff (personal communication) did not appear in our Blast results. However, this gene (*Acp53C14c*) appears to be another tandem duplicate and shows male-limited expression (B. Wagstaff, personal communication). Intraspecific paralogous protein divergence was 48.5% between *Acp53Ea* and *Acp53C14a*, 42.5% between *Acp53Ea* and *Acp53C14b*, and 45% between *Acp53C14a* and *Acp53C14b*. The divergence of *Acp53C14c* from other putative *Acp53Ea* duplicates was >65%. These genes are tandem duplicates, with *Acp53C14a* located 423 bp proximal to *Acp53C14b*, *Acp53Ea* 487 bases distal to *Acp53C14b*, and *Acp53C14c* 519 bp distal to *Acp53Ea*. *Acp53Ea*, *Acp53C14a*, *Acp53C14b*, and *Acp53C14c* are predicted to be 120, 121, 132, and 124 amino acids long, respectively. Each is composed of two exons with a 50–60 nt intron roughly 40 bases from the initiation codon. All genes are predicted to be peptide hormones and to have a signal sequence (SignalP v2.0, Nielsen et al. 1997).

The high levels of silent and replacement divergence among putative paralogs suggest that the duplication events...
predate the split of D. yakuba from the D. melanogaster/D. simulans lineage. Nevertheless, the conserved structures, expression patterns, presence of predicted signal peptides, and, for most cases, tandem organization all indicate that we have correctly identified paralogous genes.

Acp29AB and Acp53Ea are expressed only in male accessory glands (Wolfer et al. 1997). Our RT-PCR experiments showed that the only detectable expression of Lectin29Ca, Acp53C14a, and Acp53C14b is in accessory glands (fig. 1). We were unable to detect an RT-PCR product from Lectin30A. However, given that our RACE products were derived from male cDNA, we are certain the gene is expressed in males (perhaps at low levels). The fact that the Acp duplicates identified here share accessory-gland enriched expression further supports the inference of paralogy and suggests that subfunctionalization, at least with respect to gene expression (sensu Lynch and Force 2000), cannot explain fixation of Acp duplicates. Levels of protein polymorphism and divergence for these Acp genes and putative duplicates (table 1) were higher than those typically seen in D. simulans and D. melanogaster genes, as was the case for previous surveys of Acp variation (Begun et al. 2000; Swanson et al. 2001). There was, however, a major exception in the Acp29AB family.

Lectin29Ca had no silent polymorphisms and only a single replacement polymorphism in our U.S. D. melanogaster sample (table 2). This is highly unusual given the relatively high levels of variation in D. melanogaster generally and in Acp genes specifically. Low levels of heterozygosity are even more surprising given high levels of silent and replacement divergence at this gene. We used the HKA test (Hudson, Kreitman, and Aguade 1987) to compare polymorphism and divergence data for both U.S. and African samples and because there is no evidence that vermilion has been subject to directional or balancing selection [Begun and Aquadro 1995]. Lectin29Ca versus the coding region of vermilion for the U.S. sample showed a highly significant departure from neutrality \((P = 0.0022)\), consistent with a hitchhiking event reducing variation in Lectin29Ca. In contrast, the African sample showed considerably higher levels of polymorphism but similar levels of divergence (table 2). The HKA test of African variation at Lectin29Ca and vermilion showed no deviation from the neutral model \((P = 0.7875)\). Given that the North American populations are thought to be recently derived from ancestral African populations (David and Capy 1988), the data support the idea that a selective sweep at Lectin29Ca occurred in the very recent past. Note that Acp29AB is only 356 bp upstream of Lectin29Ca, yet it shows normal levels of heterozygosity in the California sample. This suggests that the window of reduced heterozygosity associated with Lectin29Ca may be quite small, though additional population genetic data from the other flanking region

![Figure 1](https://academic.oup.com/mbe/article-abstract/21/9/1625/992141/fig1a.png)

**Fig. 1.**—RT-PCR analysis of tissue-specific expression of putative duplicates. (A) Lectin29Ca, (B) Acp53C14a, and (C) Acp53C14b. Lane assignments for each gel: (1) 1 kB ladder, (2) whole females, (3) males without reproductive tracts, (4) testes, (5) accessory glands, and (6) negative control.

### Table 1

Silent and Replacement Site Heterozygosity and Divergence for Acp29AB and Acp53Ea Gene Families in D. melanogaster and D. simulans

<table>
<thead>
<tr>
<th>Gene</th>
<th>Sil.</th>
<th>Repl.</th>
<th>Sample</th>
<th>(\theta_{sil})</th>
<th>(\theta_{repl})</th>
<th>(\text{Div}_{sil})</th>
<th>(\text{Div}_{repl})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acp29AB</td>
<td>148</td>
<td>254</td>
<td>mel</td>
<td>0.0209</td>
<td>0.0028</td>
<td>0.2368</td>
<td>0.0772</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0359</td>
<td>0.0053</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lectin29Ca</td>
<td>143</td>
<td>559</td>
<td>mel</td>
<td>0.0000</td>
<td>0.1745</td>
<td>0.1745</td>
<td>0.0807</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>A/P</td>
<td>0.0142</td>
<td>0.0116</td>
<td>0.1841</td>
<td>0.0848</td>
</tr>
<tr>
<td>Lectin30A</td>
<td>137</td>
<td>532</td>
<td>mel</td>
<td>0.0112</td>
<td>0.0029</td>
<td>0.1262</td>
<td>0.0579</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0286</td>
<td>0.0703</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp53Ea</td>
<td>88</td>
<td>266</td>
<td>mel</td>
<td>0.0399</td>
<td>0.0066</td>
<td>0.1574</td>
<td>0.0390</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0046</td>
<td>0.0061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp53C14a</td>
<td>85</td>
<td>278</td>
<td>mel</td>
<td>0.0174</td>
<td>0.0013</td>
<td>0.1255</td>
<td>0.0068</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0289</td>
<td>0.0029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp53C14b</td>
<td>93</td>
<td>303</td>
<td>mel</td>
<td>0.0435</td>
<td>0.0024</td>
<td>0.1689</td>
<td>0.0352</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0307</td>
<td>0.0027</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp53C14c</td>
<td>92</td>
<td>280</td>
<td>mel</td>
<td>0.0401</td>
<td>0.0058</td>
<td>0.1387</td>
<td>0.0823</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>sin</td>
<td>0.0595</td>
<td>0.0156</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 other</td>
<td>286</td>
<td>935</td>
<td>sim</td>
<td>0.0349</td>
<td>0.0013</td>
<td>0.1084</td>
<td>0.0107</td>
</tr>
<tr>
<td>Genes 3Rc</td>
<td>(19)</td>
<td>(65)</td>
<td></td>
<td>(0.0044)</td>
<td>(0.0003)</td>
<td>(0.0097)</td>
<td>(0.0032)</td>
</tr>
</tbody>
</table>

**Note.**—Sil. = silent sites in coding regions; Repl. = replacement.

* a Divergence \((\text{Div.})\) is between all pairs of D. melanogaster and D. simulans genes (Sujes-Cantor corrected).

* b Malawi, Africa.

* c Data are means with standard error values below in parentheses for 19 unrelated proteins on chromosome arm 3R from Begun and Whitley (2000) and Begun (2002).
would be necessary to determine if this is indeed the case. It seems notable that of these original candidates with evidence for recent selection in non-African D. melanogaster samples (desat, Takahashi et al. 2001; Acp36DE, Begun et al. 2000; Lectin29Ca, this report), two are Acps. Such observations support the notion that selection on sexual traits can cause rapid divergence of Drosophila populations (Knowles and Markow 2001; Miller, Starmer, and Pitnick 2003; Pitnick et al. 2003). Additional work will be required to precisely determine the extent of the “swept” region of Lectin29Ca in non-African populations and to identify putative mutations that might be targets of selection.

Numbers of polymorphic and fixed, silent, and replacement mutations were compared to predictions of the neutral model (McDonald and Kreitman 1991). Of the six genes in our study, four (Acp29AB, Lectin30A, Acp53C14b, and Acp53C14c) reject the null hypothesis in a direction consistent with adaptive protein evolution (table 2). Moreover, our population genetic data support the notion that all three members of the Acp29AB family have been influenced by recent directional selection. Overall, our results put firmer ground the conclusion that adaptive protein evolution is a major cause of divergence of Acp proteins in D. melanogaster and D. simulans.

Methods

Amino acid sequences of 13 annotated Acps from D. melanogaster (Acp26AA, Acp26AB, Acp29AB, Acp32CD, Acp33A, Acp36DE, Acp53EA, Acp62F, Acp63F, Acp70A, Acp76A, Acp95EF, and Acp 98AB) were compared to the D. melanogaster reference sequence (Genome Release 3.0, Flybase Consortium 2003) by tBlastN searches using default parameters. Sequences were aligned by manual curation and molecular population genetics of putative duplicates with <50% nucleotide divergence from a known Acp were investigated. Nucleotide divergence (or uncorrected pairwise-distance) was calculated as a measure of the number of mismatched nucleotides/total number of nucleotides at 1st and 2nd codon positions. Duplicate genes are and considered putative Acps but are referred to as Acps for convenience.

Candidate Acp duplicates were subjected to RT-PCR to determine whether their expression was restricted to accessory glands. mRNA was extracted from four tissues of D. melanogaster: male accessory glands, testes, male carcasses, and whole females. First strand synthesis was carried out using an oligo-dT primer and Superscript Reverse Transcriptase (Invitrogen, San Diego, Calif.). RT-PCR was carried out using gene-specific primers on RNA/DNA heteroduplex isolated from each tissue.

D. simulans population genetic data are from inbred lines established from flies collected at the Wolfskill Orchard in Winters, Calif. (Begun and Whitley 2000). D. melanogaster population genetic data are from isochromosomal lines derived from the Wolfskill Orchard or from isofemale lines from Malawi, Africa. D. yakuba sequences are from an isofemale line. Some D. simulans Acp29AB sequences are from Begun et al. (2000). In most cases DNA sequencing was carried out directly on PCR products. For cases in which inbred lines were not available, PCR products were cloned prior to sequencing.

Sequences were assembled using SeqMan (DNASTar, Inc., Madison, WI) and manually curated in MacClade 4.0 (Maddison and Maddison 2000). Alignments are available upon request from the authors. Summary statistics and tests of the neutral equilibrium model were carried out using DnaSP version 3.53 (Rozas and Rozas 1999). SignalP version 2.0 was used to predict presence/absence of signal peptides characteristic of Acps and other secreted proteins (Nielsen et al. 1997). Sequences were submitted to GenBank under accession numbers AY635196-AY635290.

Acknowledgments

We thank A. Kern and S. Joseph for useful comments on the manuscript. Grants from the University of Texas Graduate Program in Evolution, Ecology, and Behavior to A.K.H. and the NIH (GM55298) and NSF (DEB-0327049) to D.J.B supported this work.

Literature Cited


Lauren McIntyre, Associate Editor

Accepted June 15, 2004