RESEARCH ARTICLES

Comparative Genomics of Accessory Gland Protein Genes in Drosophila melanogaster and D. pseudoobscura

Bradley J. Wagstaff*† and David J. Begun†

*Section of Integrative Biology, University of Texas, Austin; and †Section of Evolution and Ecology, University of California, Davis

Male accessory gland protein genes (Acps) evolve rapidly in the melanogaster species subgroup of Drosophila. However, conservation of Acps in more diverged lineages is poorly understood. We used comparisons of the D. melanogaster and D. pseudoobscura genome sequences, along with empirical investigation of D. pseudoobscura transcription, to assay the D. pseudoobscura genome for orthologs of 13 D. melanogaster Acps (Acp26Aa, Acp26Ab, Acp29AB, Acp32CD, Acp33A, Acp36DE, Acp53EA, Acp62F, Acp63F, Acp70A, Acp76A, Acp95EF, and Acp98AB). We find that Acp26Aa, Acp26Ab, Acp32CD, and Acp53EA are present at the expected microsyntenic locations of D. pseudoobscura. Acp62F and Acp70A are also present, although they are located in nontsyntenic regions. For six of the remaining seven Acps, computational and molecular biological evidence suggests they are D. melanogaster orphans. The weighted average of interspecific amino acid identity for alignable residues across the six orthologous Acps is 35.6%. Population genetic data for D. pseudoobscura Acp26Aa show that this gene has been evolving under directional selection, as it has been in D. melanogaster/D. simulans.

All four D. melanogaster Acps we analyze from chromosome arm 3L are absent from the homologous D. pseudoobscura XR chromosome arm, which was autosomal before an X chromosome–autosome fusion event in the D. pseudoobscura lineage. This observation is consistent with the hypothesis that male-advantage genes on the Drosophila X chromosome are disfavored by natural selection.

Introduction

Much of comparative genomics research seeks to detect putative functional elements (e.g., genes) by virtue of sequence conservation (e.g., Batzoglou et al. 2000; Wiehe et al. 2001; Jaillon et al. 2003). However, genes that respond to persistent directional selection are also functionally important and can be overlooked in comparative analyses that focus on sequence conservation. Thus, an understanding of biological diversity and adaptation will require evolutionary and functional analysis of rapidly evolving genes. The gain or loss of genes over time must also be explained. For example, microorganisms that take on an obligate intracellular lifestyle often lose genes (e.g., Moran 2003). Over long time periods, even conserved proteins can be lost in certain lineages (Korschak et al. 2003; Krylov et al. 2003). Nonetheless, our general understanding of gene loss is likely plagued by ascertainment bias. For example, genes that are prone to loss over relatively shorter time scales may tend to evolve quickly and, therefore, are more likely to be unannotated in model system genomes. Gain and loss of genes is intriguing because it suggests the possibility that “homologous” functions can be partially (or even mostly) coded for by nonhomologous proteins. The population genetic mechanisms of gene loss are also interesting. For example, gene loss could represent decay of a “nonessential” gene under mutation pressure, a change of the biology in a lineage that renders a previously essential gene dispensable, or removal of a gene by selection (Olson 1999; Galvani and Slatkin 2003; Olson and Varki 2003). We would like to distinguish among these possibilities.

Drosophila is an attractive model system for addressing these questions. Flies have relatively compact genomes for animals, and the deep annotation and experimental tractability of the model fly, D. melanogaster, provide an excellent starting point for investigating the functional and evolutionary biology of rapidly evolving proteins. D. pseudoobscura is currently the only Drosophila species other than D. melanogaster with a high quality genome sequence (Richards et al. 2005). D. pseudoobscura diverged from the melanogaster group approximately 21 to 46 MYA (Beckenbach, Wei, and Liu 1993). Comparative analyses of these species have shown that the majority of D. melanogaster release 3 gene models are highly conserved in D. pseudoobscura and that microsynteny is largely maintained (Bergman et al. 2002; Richards et al. 2005).

Data from animals suggest that the portion of the genome coding for reproduction-related function may be unusually dynamic. For example, an interesting generality emerging from studies of molecular evolution is the relatively rapid evolution of proteins associated with male reproduction (e.g., Swanson and Vacquier 2002). In Drosophila, testis and accessory gland proteins (Acps) show rapid divergence (Coulthart and Singh 1988; Begun et al. 2000; Swanson et al. 2001; Kern, Jones, and Begun 2004) compared with other proteins. Three known genes contributing to reproductive isolation in flies (Ting et al. 1998; Barbash et al. 2003; Presgraves et al. 2003) evolve extremely quickly, suggesting that rapidly evolving genes may play an important role in speciation.

Drosophila Acps have probably received more population genetic attention than any other class of reproduction-related gene in flies. Males transfer Acps to females during mating. Acps have been implicated in induction of oviposition, in rendering females recalcitrant to remating, and in mediating sperm displacement and sperm storage in females (Neubaum and Wolfner 1999; Tram and Wolfner 1999) (reviewed in Wolfner [2002] and Heifetz and Wolfner [2004]). As noted previously, Acps evolve quickly compared with other Drosophila proteins. Some of this rapid evolution...
is likely the result of directional selection (Aguadé 1998; Tsaur, Ting, and Wu 1998; Begun et al. 2000; Holloway and Begun 2004).

These previous observations of Drosophila molecular evolution motivate the work reported here, which addresses three main questions regarding molecular evolution and gain/loss of Acps in the D. melanogaster versus D. pseudoobscura comparison. First, how does one identify orthologous, rapidly evolving genes that may be sufficiently diverged so as to preclude identification through simple Blast comparisons between genomes? Second, what are the patterns of protein evolution for highly diverged genes? Third, and perhaps most interesting, to what extent are rapidly evolving proteins likely to be lineage restricted (i.e., absent in at least some lineages)? This last question is especially interesting to us because gene presence/absence variation could be an important aspect of the unique biology of particular lineages, and reproduction-related genes may be more likely than other types of genes to show lineage-restricted distributions. Here, we use computational and molecular approaches to investigate these questions by comparison of 13 annotated Acp genes from the D. melanogaster reference sequence to the D. pseudoobscura genome sequence.

Materials and Methods

Computational Analysis

The D. pseudoobscura genome (August 2003, Freeze 1 Assembly; BGM-HGSC, http://www.hgsc.bcm.tmc.edu/projects/Drosophila/) was screened through extensive Blast version 2.2.9 analysis (Altschul et al. 1997) for the presence of 13 D. melanogaster Acps (Acp26Aa, Acp26Ab, Acp29AB, Acp32CD, Acp33A, Acp36DE, Acp53Ea, Acp62F, Acp63F, Acp70A, Acp76A, Acp93EF, and Acp98AB). These particular Acps were among the first identified and have the strongest empirical support (Wolfner 1997; Wolfner et al. 1997). A combination of Blast methods was used to investigate presence/absence of D. pseudoobscura orthologs. tBlastN (peptide sequence query to all six possible reading frames of a nucleotide database) searches of all D. melanogaster Acps were performed. D. melanogaster Acp flanking sequence was also analyzed to establish larger scale homology and microsynteny (or lack thereof) between species. Depending on the immediate genomic neighborhood of individual Acps, this either involved tBlastN analysis of flanking genes, BlastN (nucleotide to nucleotide query) analysis of noncoding intergenic sequence, or some combination.

The search for homologous D. pseudoobscura sequence began with tBlastN analysis of D. melanogaster Acps. We used E < e−4 as our typical significance threshold. However, sequences with marginally significant E scores (e−4 < E < e−2) were scrutinized if they represented the best opportunity for orthology (e.g., analysis of Acp70A). All potential D. pseudoobscura ortholog candidates were BlastP analyzed back to D. melanogaster predicted proteins. To eliminate nonorthologous genes with shared domains or from gene families, only candidates that hit the original D. melanogaster Acp at the lowest E score were considered further (there were no ambiguous cases in which a D. melanogaster Acp E score was close to the score from another gene). Proximal and distal flanking sequence was then analyzed for all 13 Acps. Starting from immediate flanking sequence and moving out in both directions, noncoding intergenic sequence and neighboring genes were Blast analyzed. Flanking sequences were typically queried in 2-kb to 4-kb intervals, but exact lengths depended on the genetic neighborhood of individual Acps. Flanking genes were analyzed in the same manner as the Acps described above. The same E score threshold (E < e−4) was used for intergenic sequence BlastN analysis, but additional hits (E < 0.05) to D. pseudoobscura microsyntenic sequence were also noted, once homology was already established. For every D. melanogaster Acp, the amount of flanking sequence analysis was dictated based on certainty of homology. For example, if 2 kb of flanking sequence produced five intergenic BlastN hits of E < e−10 each, we did not necessarily analyze additional sequence from that flank.

D. pseudoobscura Acp ortholog candidate regions, as defined by patterns of microsynteny, were further analyzed for the presence of open reading frames (ORFs) and evidence of transcription. Computational analysis of D. pseudoobscura Acp ortholog candidate regions consisted of identifying potential ORFs that showed similarity to D. melanogaster counterparts in amino acid similarity, ORF length, intron/exon structure, protein domains, or presence/absence of putative signal sequences. The SignalP version 3.0 server (hidden Markov method) was used to detect putative signal peptides (Nielsen and Krogh 1998; Bendtsen et al. 2004). NCBI CD-Search was used to identify conserved domains (Marchler-Bauer et al. 2003). Protein sequences were aligned using the default Clustal parameters of MegAlign in the DNASTAR software package (Lasergene, Madison, Wis.). Protein similarity was calculated as the number of identical residues divided by the total number of alignable residues.

Empirical Methods

Two approaches, RACE and reverse Northern, were used to empirically investigate transcription in D. pseudoobscura genomic regions that are homologous to regions containing Acps in D. melanogaster. RACE templates were separately produced from sexually mature male and female D. pseudoobscura flies from a stock that combined two isofemale lines originally collected by M. Noor. mRNA from each sex was isolated using the MicroPolyA-Pure kit (Ambion, Austin, Tex.). RACE-ready cDNA was prepared, and target molecules were PCR-amplified and isolated using the GeneRacer (Invitrogen) kit according to the manufacturer’s instructions. The protocol separates the truncated from the complete and mature mRNA products, preferentially selecting the full-length transcripts for first-strand cDNA synthesis. Target-specific primers were paired with either 3’ or 5’ RACE primers to amplify candidate transcripts. In many cases, multiple target primers were used. RACE was performed on pooled aliquots of male and female RACE-ready cDNA. Amplified products were cloned into the TOPO vector (Invitrogen) and used for bacterial transformations according to manufacturer’s instructions. Direct sequencing of colony PCR products was carried out on an Applied Biosystems 3700 sequencer (ABI).
Although RACE should be sensitive to low transcript abundance, failure of RACE to amplify a transcript could be a result of suboptimal gene-specific primers. This problem is a particular concern for small putative transcripts, for which primer design options can be limited. Therefore, regions providing no evidence of transcription from RACE reactions were subjected to reverse Northern analysis. Unlike RACE, this approach has the virtue of requiring no specific inferences regarding details of putative protein-coding genomic regions. Candidate and control D. pseudoobscura cDNA by 32P-labeling using the Prime-It II kit (Stratagene). These probes were hybridized overnight to the replicate filters at 65°C in a buffer consisting of 0.5 M NaPi (pH 7.2), 7% SDS, and 1 mM EDTA. Filters were washed at 60°C in 40 mM NaPi, 1% SDS, and 1 mM EDTA. The resulting membranes were exposed to X-ray film to infer evidence of transcription in male and female D. pseudoobscura.

Population Genetics

Isofemale lines derived from flies collected by M. Noor were used for population genetics analysis. The sample consisted of five D. pseudoobscura lines, one D. persimilis line, and one D. miranda line. The sequenced D. pseudoobscura genome was used to add one additional allele to the analysis. The Expand High-Fidelity Polymerase System (Roche Molecular Biochemicals) was used for PCR amplification. To isolate single alleles for sequencing, PCR products were directly cloned into the TOPO vector (Invitrogen) and used for bacterial transformations according to manufacturer’s guidelines. Amplified colony PCR products and their associated sequences were obtained using M13 reverse and T7 primers. All sequencing was performed using M13 reverse and T7 primers. All sequencing was performed using M13 reverse and T7 primers. All sequencing was performed using M13 reverse and T7 primers.

Evidence of Gene Presence

Acp26Aa and Acp26Ab

Figure 1A shows an illustration of the putative homology between D. melanogaster and D. pseudoobscura in the Acp26Aa and Acp26Ab region. Acp26Aa showed no Blast similarity to any D. pseudoobscura sequence, and Acp26Ab generated only a marginally significant (E = 0.045) tBlastN hit. Nevertheless, investigation of nearby flanking sequences revealed strong evidence for a D. pseudoobscura region of homology on chromosome 4, the correct arm given the homology of D. pseudoobscura 4 and D. melanogaster 2L (Lakovaara and Saura 1982; Steineman, Pinsker, and Sperlich 1984).

The first 2 kb immediately proximal to D. melanogaster Acp26Aa generated five highly significant and contiguous BlastN hits, averaging 41 bp (from E = 3e–10 to E = 6e–5), to a portion of D. pseudoobscura chromosome 4 (region a, figure 1A). The 4.5-kb region immediately distal to Acp26Ab was similarly characterized by four BlastN hits, averaging 70 bp (from E = 3e–19 to E = 2e–9; partially depicted by region b, figure 1A). Given the contiguous physical organization of the flanking regions in the two species and given the fact that the marginally significant Acp26Ab tBlastN hit fell within the hypothesized microsyntenic 5.1-kb region in D. pseudoobscura spanning BlastN hits in regions a and b (fig. 1A), it is highly likely that we have identified the homologous region in D. pseudoobscura.

RACE analysis of the D. pseudoobscura 5.1-kb candidate sequence was used to identify the putative transcripts corresponding to Acp26Aa and Acp26Ab. One gene-specific primer for 5' RACE was designed from sequence corresponding to the D. pseudoobscura tBlastN hit for Acp26Ab. Six additional 5' RACE primers were designed from the 3 kb of D. pseudoobscura candidate sequence immediately upstream of the tBlastN hit to Acp26Ab. The rationale for this was that at least one of these six primers should amplify a portion of a D. pseudoobscura Acp26Aa ortholog if it exists within this homologous region. DNA sequences of the resulting successful RACE reactions on D. pseudoobscura–derived cDNA and comparison of these RACE products to genomic sequence clearly revealed both genes. Conservation of intron/exon structure and evidence of predicted signal peptides support
an inference of orthology (table 1). Male-specific transcription within the *D. pseudoobscura* *Acp26Aa* candidate region (fig. 2) provides additional support for orthology. Interestingly, despite the compelling evidence for orthology, the predicted proteins are extraordinarily diverged, especially *Acp26Aa* (table 1).

Of the *Acps* that have been subjected to evolutionary analysis in the *melanogaster* subgroup species, *Acp26Aa* shows the strongest evidence for directional selection, including Ka/Ks > 1 (Tsaur and Wu 1997), significant McDonald-Kreitman tests (Aguadé 1998; Tsaur, Ting, and Wu 1998), and overdispersed amino acid substitution (Kern, Jones, and Begun 2004). We were interested in determining whether the *D. pseudoobscura* *Acp26Aa* ortholog showed patterns of molecular polymorphism and divergence similar to those observed in the *melanogaster* subgroup. We collected population genetic data for *Acp26Aa* from *D. pseudoobscura* (six alleles) and its sister species, *D. persimilis* (one allele), along with a single outgroup species allele from *D. miranda*. There is evidence of gene flow between *D. pseudoobscura* and *D. persimilis* (Hey and Nielsen 2004). Our single *Acp26Aa* *D. persimilis* allele clusters with the six *D. pseudoobscura* alleles. Thus, we report polymorphism and diverged data with the *D. persimilis* allele both included and removed from the *D. pseudoobscura* data set (tables 3 and 4).

Relative rates of replacement to silent site evolution in the *D. pseudoobscura*/*D. persimilis* versus *D. miranda* comparison are comparable to the rates of evolution in the *melanogaster* subgroup (table 3). Replacement polymorphism in *D. pseudoobscura*/*D. persimilis* is similar to both African and American populations of *D. melanogaster*, whereas...
silent sites are more than twice as variable in *D. pseudoobscura*/*D. persimilis* and African *D. melanogaster* than American *D. melanogaster* (table 3). Our McDonald-Kreitman test of *D. pseudoobscura*/*D. persimilis* versus *D. miranda* sequences showed convincing evidence for adaptive protein evolution (\(G = 5.76, P = 0.016\) [table 4]). African *D. melanogaster* populations likewise show significant evidence of adaptive protein evolution (\(P = 0.002\)), whereas American *D. melanogaster* populations show a nonsignificant trend toward excess replacement fixations (\(P = 0.109\)), probably as a consequence of lower levels of polymorphism in this population. Thus, our data suggest *Acp26Aa* is evolving at comparable rates in both the *D. melanogaster* and the *D. pseudoobscura* lineages and that adaptive protein evolution occurs in both lineages.

**Acp32CD**

*D. melanogaster* *Acp32CD* and its two nearest neighbors generated clear tBlastN hits to a single, small contiguous region of *D. pseudoobscura*, chromosome 4 (fig. 1B). Of the three genes, *CG14913* is the most highly conserved (\(E = 2e^{-79}\)), followed by the last exon of *CG31868* (\(E = 1e^{-27}\)), and *Acp32CD* (\(E = 9e^{-12}\)). *D. pseudoobscura* *Acp32CD*, like its *D. melanogaster* ortholog, is a single-exon gene with a predicted signal peptide sequence (table 1). The *D. pseudoobscura* *Acp32CD* protein contains 299 residues, compared with 252 residues in *D. melanogaster*. The difference in size is largely because of the middle section of the *D. pseudoobscura* protein, which contains a section of several glycine residue repeats. Even so, the orthologs show 43.7% similarity.

**Acp53Ea and Duplicates**

*Acp53Ea* is one of four tandemly duplicated genes in *D. melanogaster* found in a region just over 3 kb in length (fig. 1C). Paralogous *D. melanogaster* protein divergence is 48.5% between *Acp53Ea* and *Acp53C14a*, 42.5% between *Acp53Ea* and *Acp53C14b*, and 45% between *Acp53C14a* and *Acp53C14b* (Holloway and Begun 2004). *Acp53C14c* was previously unannotated and was discovered as a secondary tBlastN hit to *Acp53C14b*.
It is the most diverged of the duplicates, at greater than 65% divergence from the other three. Similar gene structures, predicted protein lengths, and strongly predicted signal peptides for all four genes (table 1) support the hypothesis that they are related through repeated tandem duplication.

BLASTN comparisons of each of the four duplicates to the D. pseudoobscura genome revealed corresponding orthologs on chromosome 3, thereby suggesting that these duplications predate the D. melanogaster/D. pseudoobscura split (E scores for Acp53C14c, Acp53Ea, Acp53C14b, and Acp53C14a are 3e–15, 9e–13, 1e–28, and 4e–26, respectively).

Acp53C14c was found near the endpoint of one D. pseudoobscura chromosome 3 contig, but the other three were located contiguously on another chromosome 3 contig. However, further scrutiny of the Acp53C14c contig strongly suggests that Acp53C14c is likely just upstream of the other Acp53 genes, just as it is in D. melanogaster. This inference comes from the observation that in D. pseudoobscura, CG8566 (tBLASTN, E = 0.0) is just under 3 kb to the left of Acp53C14c (orientation as in figure 1C), whereas in D. melanogaster, CG8566 is about 2.2 kb to the left of (distal to) Acp53C14c. Protein similarity leaves little doubt as to the true orthology of these duplicates, as the most similar interspecific pairings is consistent with conserved microsynteny between species (40.5%, 41.7%, 48.5%, and 55% similarity for Acp53C14c, Acp53Ea, Acp53C14b, and Acp53C14a, respectively).

A major difference between these species in this region is that D. pseudoobscura has three additional tandem duplications.

### Table 1

<table>
<thead>
<tr>
<th>Gene</th>
<th>Amino Acid Residues</th>
<th>Exons</th>
<th>Introns</th>
<th>Signal Peptide</th>
<th>% Similar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acp26Aa</td>
<td>264</td>
<td>34, 761</td>
<td>56</td>
<td>1.00</td>
<td>18.5</td>
</tr>
<tr>
<td>pse-Acp26Aa</td>
<td>250</td>
<td>37, 716</td>
<td>68</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp26Ab</td>
<td>90</td>
<td>31, 242</td>
<td>61</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>pse-Acp26Ab</td>
<td>92</td>
<td>31, 248</td>
<td>65</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp29AB</td>
<td>234</td>
<td>705</td>
<td>—</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Acp32CD</td>
<td>252</td>
<td>759</td>
<td>99</td>
<td>43.7</td>
<td></td>
</tr>
<tr>
<td>pse-Acp32CD</td>
<td>299</td>
<td>900</td>
<td>—</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp33A</td>
<td>47</td>
<td>144</td>
<td>—</td>
<td>0.97</td>
<td></td>
</tr>
<tr>
<td>Acp36DE</td>
<td>912</td>
<td>208, 2531</td>
<td>59</td>
<td>0.98</td>
<td></td>
</tr>
<tr>
<td>Acp53Ea</td>
<td>120</td>
<td>42, 321</td>
<td>65</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>pse-Acp53Ea</td>
<td>120</td>
<td>42, 321</td>
<td>72</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp53C14a</td>
<td>121</td>
<td>42, 324</td>
<td>52</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>pse-Acp53C14a</td>
<td>120</td>
<td>42, 321</td>
<td>71</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp53C14b</td>
<td>132</td>
<td>42, 357</td>
<td>56</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>pse-Acp53C14b</td>
<td>132</td>
<td>42, 357</td>
<td>65</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp53C14c</td>
<td>124</td>
<td>42, 333</td>
<td>57</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>pse-Acp53C14c</td>
<td>121</td>
<td>42, 324</td>
<td>56</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>pse-Acp53C14d</td>
<td>129</td>
<td>42, 348</td>
<td>70</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp53C14e</td>
<td>127</td>
<td>42, 342</td>
<td>51</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Acp53C14f</td>
<td>127</td>
<td>33, 351</td>
<td>60</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp62F</td>
<td>115</td>
<td>348</td>
<td>—</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>pse-Acp62F</td>
<td>135</td>
<td>408</td>
<td>—</td>
<td>0.99</td>
<td></td>
</tr>
<tr>
<td>Acp63F</td>
<td>81</td>
<td>28, 156, 62</td>
<td>61</td>
<td>54.0</td>
<td></td>
</tr>
<tr>
<td>Acp70A</td>
<td>55</td>
<td>115, 53</td>
<td>65</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp63F</td>
<td>57</td>
<td>118, 56</td>
<td>74</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp95EF</td>
<td>52</td>
<td>18, 141</td>
<td>62</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>Acp98AB</td>
<td>28–31</td>
<td>87–96</td>
<td>—</td>
<td>0.00</td>
<td></td>
</tr>
</tbody>
</table>

* Number of nucleotides per exon/intron, starting from the initiation codon and going through the stop codon.

*b Probability of signal peptide as predicted by the hidden Markov method of SignalP version 3.0 (Nielsen and Krogh 1998; Bendtsen et al. 2004).

c Percent amino acid identities, calculated as the number of identical residues/total number of alignable residues.
E scores of 2e–06 (genome revealed the last two duplicates through obscura identified Acp53C14d duplicates (E = 0.001). Additional tBlastN analysis of Acp53C14d to the D. pseudoobscura genome revealed the last two duplicates through E scores of 2e–06 (Acp53C14f) and 5e–04 (Acp53C14e). None of these additional duplicates appear to have D. melanogaster orthologs. tBlastN analysis of all three back to the D. melanogaster genome only produced one significant hit for Acp53C14d to D. melanogaster Acp53C14d (E = 2e–05) and two nonsignificant hits for Acp53C14d to D. melanogaster Acp53C14a (E = 0.13) and Acp53C14b (E = 0.28). Neither Acp53C14e nor Acp53C14f Blasts registered even weak hits to D. melanogaster. Therefore, these additional D. pseudoobscura duplicates either originated in the D. pseudoobscura lineage or were lost from the D. melanogaster lineage.

Evidence of Gene Presence Associated with Genomic Rearrangement Acp62F

D. melanogaster Acp62F is an intronless gene that codes for a 115-residue protein with a trypsin inhibitor domain and a predicted signal peptide sequence. The nearest distal gene, CG32296, is 11 kb away. CG1240 is the nearest proximal gene, at about 20 kb away. Nevertheless, BlastN analysis of 3 kb of intergenic sequence along each genomic flank revealed a microsyntenic region to D. pseudoobscura chromosome XR (fig. 1D). The 5’ flank is characterized by five highly significant BlastN matches (from E = 2e–18 to E = 2e–8) that average 52 bp in length (region a, figure 1D). The 3’ flank is similarly characterized by four BlastN matches that average 54 bp (E values ranging from 6e–18 to 2e–11 [region b, figure 1D]).

An Acp62F ortholog could not be identified in the D. pseudoobscura candidate microsyntenic region (between BlastN matches of regions a and b in figure 1D). Computational analysis of this 3.4-kb region revealed six candidate ORFs, ranging from 62 to 155 residues in length. None of these candidates showed good evidence of a signal peptide sequence (SignalP probabilities ranged from 0 to 0.35) or a trypsin inhibitor domain. RACE analysis of all six possible candidates also failed to detect any evidence of D. pseudoobscura transcription. Finally, a PCR product spanning the complete D. pseudoobscura candidate region failed to hybridize to male-derived and female-derived 32P-labeled cDNA (fig. 2).

Despite the lack of evidence for a putative D. pseudoobscura Acp62F homolog in the expected D. pseudoobscura microsyntenic region, tBlastN analysis of D. melanogaster Acp62F revealed three highly significant ortholog candidates (E = 8e–17, 2e–11, and 4e–10 for candidates 1 to 3, respectively) at different positions of D. pseudoobscura chromosome 3 (not tandemly arranged). All three D. pseudoobscura ortholog candidates were then

### Table 2
Accession Numbers and Initiation Codon Positions for D. pseudoobscura Acp Orthologs and Microsyntenic Contigs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Accession Numbers</th>
<th>Position</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>pse-Acp26Aa</td>
<td>AADE01000040</td>
<td>9279</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp26Ab</td>
<td>AADE01000040</td>
<td>7192</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp32CD</td>
<td>AADE01000037</td>
<td>19118</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp53Ea</td>
<td>AADE01000143</td>
<td>121107</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp53F14a</td>
<td>AADE01000143</td>
<td>119222</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp53F14b</td>
<td>AADE01000143</td>
<td>120132</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp53F14c</td>
<td>AADE01000146</td>
<td>25072</td>
<td>+</td>
</tr>
<tr>
<td>pse-Acp53F14d</td>
<td>AADE01000143</td>
<td>121785</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp53F14e</td>
<td>AADE01000143</td>
<td>122365</td>
<td>–</td>
</tr>
<tr>
<td>pse-Acp62F</td>
<td>AADE01003187</td>
<td>3724</td>
<td>+</td>
</tr>
<tr>
<td>pse-Acp70A</td>
<td>AADE01000040</td>
<td>4090</td>
<td>+</td>
</tr>
</tbody>
</table>

### Table 3
Silent and Replacement Polymorphism and Divergence for Acp26Aa in D. melanogaster and D. pseudoobscura

<table>
<thead>
<tr>
<th>Sample</th>
<th>Number of Sites</th>
<th>0_syn</th>
<th>0_0p</th>
<th>KS^a</th>
<th>KA^a</th>
<th>Ka/Ks</th>
</tr>
</thead>
<tbody>
<tr>
<td>pse^b</td>
<td>154</td>
<td>524</td>
<td>0.034</td>
<td>0.008</td>
<td>0.096</td>
<td>0.100</td>
</tr>
<tr>
<td>pse + pev^c</td>
<td>154</td>
<td>524</td>
<td>0.037</td>
<td>0.100</td>
<td>0.097</td>
<td>0.101</td>
</tr>
<tr>
<td>mel (USA)^d</td>
<td>174</td>
<td>615</td>
<td>0.014</td>
<td>0.006</td>
<td>0.167</td>
<td>0.156</td>
</tr>
<tr>
<td>mel (Malawi)^d</td>
<td>174</td>
<td>615</td>
<td>0.033</td>
<td>0.008</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

^a Divergence estimates pertain to D. miranda and D. simulans for D. pseudoobscura/D. persimilis and D. melanogaster, respectively.

^b Population genetic data are restricted to the six D. pseudoobscura alleles.

^c Population genetic data includes the six D. pseudoobscura alleles as well as a single D. persimilis allele.

BlastP analyzed back to *D. melanogaster* predicted proteins. Candidate 3 was eliminated from consideration, as its strongest match was another *D. melanogaster* trypsin inhibitor domain protein, CG5267. The two remaining candidates returned *D. melanogaster* Acp62F at the lowest E score (2e–18 and 1e–13 for candidates 1 and 2, respectively). Both *D. pseudoobscura* Acp62F ortholog candidates hit the *D. melanogaster* chromosome 3L gene CG33259 secondarily (E = 8e–17 and E = 1e–11 for candidates 1 and 2, respectively). tBlastN of *D. melanogaster* CG33259 back to *D. pseudoobscura* sequences hits candidates 1 and 2 at the lowest E scores (8e–17 and 2e–11 for candidates 1 and 2, respectively). As is the case for *D. melanogaster* Acp62F, both *D. pseudoobscura* ortholog candidates and *D. melanogaster* CG33259 have predicted signal peptides (*P* = 0.985, 0.955, and 0.999 for candidates 1, 2, and CG33259, respectively) and contain trypsin inhibitor domains. Gene organization is also similar to Acp62F, as *D. pseudoobscura* candidates 1 and 2 and *D. melanogaster* CG33259 are single-exon genes (135, 120, and 119 residues for candidates 1, 2, and CG33259, respectively). Intergenic flanking sequence analysis of the *D. pseudoobscura* candidates clearly identified microsyntenic tBlastN homology (from E = 5e–28 to E = 4e–15 for each of the four flanks) to different portions of *D. melanogaster* chromosome 2R, the correct arm given the homology of *D. melanogaster* 2R and *D. pseudoobscura* chromosome 3 (Steinemann, Pinser, and Sperlich 1984). In both cases, there were no gene annotations in the corresponding *D. melanogaster* microsyntenic region and no evidence of ORFs containing signal peptide sequences or trypsin inhibitor domains. Thus, there is no evidence that any of these trypsin inhibitor domain genes have orthologs within the appropriate microsyntenic regions.

The tBlastN evidence suggests *D. pseudoobscura* candidate 1 is most likely orthologous to *D. melanogaster* Acp62F if a true ortholog exists. Our RACE analysis of this putative ortholog proves that it is transcribed and intronless as expected. A protein-distance tree puts *D. melanogaster* Acp62F and CG33259 as the most closely related pair, followed by *D. pseudoobscura* candidate 1 and then *D. pseudoobscura* candidate 2. Given the possibility that the shared trypsin inhibitor domains obscure the evolutionary relationships as a result of convergent or parallel evolution, we also carried out a distance analysis with the shared domains removed (the domain covers 54 to 55 residues in all four genes). Although similarities decreased as expected, the structure of the distance tree remained the same. *D. melanogaster* Acp62F and CG33259 are 51.9% similar across the complete proteins. *D. pseudoobscura* candidate 1 is 41.6% similar to Acp62F. The other pairwise comparisons are below 38% similar. With domains removed, *D. melanogaster* Acp62F and CG33259 are 32.7% similar, and *D. pseudoobscura* candidate 1 is 30.9% similar to Acp62F. Remaining pairwise comparisons drop below 25%.

We conclude that *D. pseudoobscura* candidate 1 is orthologous to *D. melanogaster* Acp62F and that microsynteny has been disrupted as a result of genomic rearrangement in one or both lineages. Given that the gene is on different Muller elements in the two species, a transposition event is likely. We also propose that *D. melanogaster* Acp62F and CG33259 are related through a duplication event that occurred subsequent to the *D. melanogaster*/ *D. pseudoobscura* split. *D. pseudoobscura* candidate 2 is likely either related through a more ancient duplication (and lost in *D. melanogaster*) or is similar through parallel or convergent evolution. However, the shared trypsin inhibitor domain and lack of microsyntenic conservation between species precludes a definitive assessment of orthology from our data.

**Acp70A**

The Blast analysis of Acp70A provided no clear evidence of a *D. pseudoobscura* ortholog. However, analysis of 4 kb of the 5’ flank and 2 kb of the 3’ flank indicated that this portion of map region 70A is homologous to a portion of *D. pseudoobscura* chromosome XR through seven small BlastN matches averaging 55 bp (from E = 4e–35 to E = 9e–7 [regions a to d, figure 1E]). The regions of similarity are contiguous between species, with the exception of a pair that indicate a likely microinversion event (region b, figure 1E). Accounting for this apparent microinversion, a pair of *D. pseudoobscura* ortholog were present in this microsyntenic region, it could be on the plus strand between regions b and c or on the minus strand between regions a and b.

Given a small first exon (115 bp of the ORF [table 1]), there were approximately nine candidate *D. pseudoobscura* first exons within regions a to c. However, only one of the nine carried the signature of a signal peptide sequence (SignalP, *P* = 0.969). Neither 5’ nor 3’ RACE reactions using primers designed from this first exon candidate successfully amplified *D. pseudoobscura* cDNA. Furthermore, hybridization of *D. pseudoobscura* cDNA to a PCR fragment

### Table 4: McDonald-Kreitman Tests of Neutral Molecular Evolution at Acp26Aa in *D. melanogaster* and *D. pseudoobscura*

<table>
<thead>
<tr>
<th>Sample</th>
<th>Polymorphic</th>
<th>Fixed*</th>
<th>Synonymous</th>
<th>Replacement</th>
<th>P*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Synonymous</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Replacement</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>pse</em></td>
<td>10</td>
<td>9</td>
<td>12</td>
<td>39</td>
<td>0.022</td>
</tr>
<tr>
<td><em>pse</em> + <em>per</em></td>
<td>12</td>
<td>11</td>
<td>12</td>
<td>39</td>
<td>0.016</td>
</tr>
<tr>
<td><em>mel</em> (USA)</td>
<td>7</td>
<td>9</td>
<td>24</td>
<td>78</td>
<td>0.109</td>
</tr>
<tr>
<td><em>mel</em> (Malawi)</td>
<td>19</td>
<td>15</td>
<td>20</td>
<td>77</td>
<td>0.002</td>
</tr>
</tbody>
</table>

*Fixations pertain to *D. miranda* and *D. simulans* for *D. pseudoobscura*/ *D. persimilis* and *D. melanogaster*, respectively.

*Probability determined by G-test.*

*Polymorphism and fixation data as calculated by excluding (*pse*) and including (*pse* + *per*) the single *D. persimilis* allele.*

*a Fixation data are from Aguade (1998).*
spanning gene regions a to c provided no evidence of a transcribed gene (fig. 2), suggesting that a microsyntenic ortholog is unlikely.

The most significant tBlastN result from comparison of *D. melanogaster* Acp70A to the *D. pseudoobscura* genome was **E = 0.002**, a value sufficiently large to be ignored in most cases. However, closer analysis provided additional support for orthology. The hit was to chromosome 4 and was identical at 13 of 14 residues from the second exon. Successful 5’ RACE amplification of the corresponding region of *D. pseudoobscura* revealed a potential gene with the same intron/exon structure as *D. melanogaster* Acp70A with a strongly predicted signal peptide (SignalP, *P = 1.0*). The candidate protein is 57 residues, two residues longer that the *D. melanogaster* Acp70A protein, with one additional residue in each of the two *D. pseudoobscura* exons (table 1). BlastP analysis of the predicted *D. pseudoobscura* Acp70A protein to predicted *D. melanogaster* proteins hit only one, Acp70A (**E = 2e–05**), supporting the hypothesis of orthology. Protein alignment of the putative orthologs shows 54.7% similarity.

Analysis of the flanking regions of the putative *D. pseudoobscura* Acp70A ortholog suggested that the gene is located in a region homologous to region 35F in *D. melanogaster*, between CG31819 and CG12455. BlastN analysis of this gene in *D. pseudoobscura*, including 4 kb of each genomic flank, generated 13 highly significant and contiguous results to this region, averaging 91 bp in length (E scores from **E = 5e–56** to **E = 8e–7** for five 5’ flank matches and eight 3’ matches). There is no computational evidence for a microsyntenic *D. melanogaster* gene within the space between 3’ and 5’ flank BlastN hits. In fact, this region comprises 4.6 kb in *D. pseudoobscura*, compared with only 590 bp in *D. melanogaster*. We conclude that both species possess a copy of Acp70A, although they are in nonsyntenic locations as a result of genome rearrangement, probably transposition between Muller elements.

**Acps with Assembly Gaps**

**Acp33A**

The only *Acp* near incompletely assembled *D. pseudoobscura* microsyntenic sequence is Acp33A. tBlastN analysis returns no significant hits for either of two potential isoforms of Acp33A. The nearest gene, CG6541, is almost 5 kb distal to Acp33A. BlastN comparison of 3 kb of 5’ flank sequence to *D. pseudoobscura* generated no significant results. However, BlastN comparison of the next 2.5 kb of 5’ flank sequence did return a highly significant result to a *D. pseudoobscura* chromosome 4 contig, consisting of 10 contiguous nucleotide segments and averaging 73 bp each (E scores from **E = 4e–31** to **E = 3e–10** [region a, figure 1F]). BlastN of 2 kb of 3’ flank sequence reveals a second highly significant set (E scores from **E = 4e–15** to **E = 3e–10** [region b, figure 1F]) of seven contiguous hits averaging 63 bp in length to the beginning of another *D. pseudoobscura* chromosome 4 contig. If there has been no major evolutionary change in the organization of this region, the two *D. pseudoobscura* contigs would be about 3.5 kb apart. However, our long PCR attempts to span the putative *D. pseudoobscura* genome sequence gap were unsuccessful. Although our evidence provides no support for an Acp33A ortholog in *D. pseudoobscura*, assembly of the homologous *D. pseudoobscura* contigs is necessary before any conclusions can be reached.

**Evidence of Gene Absence**

Acp29AB and lectin-29Ca

Acp29AB and lectin-29Ca are highly diverged, tandem duplicates in *D. melanogaster* (Holloway and Begun 2004). Our tBlastN analysis of both genes was complicated by the lectin domain they share with many fly genes and resulted in several significant hits (**E < 1e–10** threshold yields eight Acp29AB hits and seven lectin-29Ca hits). However, the most significant Blast results for each of the predicted *D. pseudoobscura* proteins back to *D. melanogaster* predicted proteins were to several lectin domain-containing genes other than Acp29AB or lectin-29Ca, ruling out orthology. tBlastN analysis of three neighboring genes allowed us to identify the *D. pseudoobscura* region that is homologous to the *D. melanogaster* Acp29AB/lectin-29Ca region (fig. 1G). These three genes returned highly significant tBlastN results (CG17814, CG1893, and CG13394 returned E scores of **5e–17**, 5e–28, and 1e–111, respectively) to a single contiguous region of *D. pseudoobscura* chromosome 4.

The major difference in the organization of the microsyntenic region in the two species is that the sequence between the termination codon of CG13394 and the initiation codon of CG13394, which contains Acp29AB and lectin-29Ca, is 2.2 kb in *D. melanogaster* (fig. 1G). The same region in *D. pseudoobscura* is only 145 bp, clearly ruling out the possibility of microsyntenic orthologs. We also found no evidence from tBlastN analysis for a chromosomal rearrangement, as we observed for Acp62F and Acp70A. Therefore, we conclude that Acp29AB and lectin-29Ca could only be present in *D. pseudoobscura* given a model of extreme sequence divergence and genomic rearrangement.

**Acp36DE**

Acp36DE is located between distantly separated exons of CG5803 in a gene-poor region of the *D. melanogaster* genome. It is 35 kb proximal to the first exon of CG5803 and 24 kb distal to the second exon. There are no other annotated genes in this region. tBlastN comparison of *D. melanogaster* Acp36DE to the *D. pseudoobscura* genome revealed no evidence for a *D. pseudoobscura* Acp36DE homolog. However, BlastN analysis using 5’ and 3’ flanking *D. melanogaster* sequences revealed clear evidence for a region of microsynteny in the two species. Analysis of 3.5 kb of 5’ flanking sequence to Acp36DE returned four BlastN matches (from **E = 2e–30** to **E = 6e–6** [region a, figure 1H]), averaging 57 bp in length. Similarly, BlastN analysis of 1.5 kb of 3’ flanking sequence revealed hits for six small DNA segments averaging 42 bp in length and which had E-values ranging from **E = 5e–14** to **E = 2e–4** (region b, figure 1H). The highly similar proximal-to-distal linear organizations of these small regions in the two species provide strong evidence of microsynteny.

However, two pieces of evidence suggest that there is no *D. pseudoobscura* ortholog of Acp36DE. First, the
physical scale of the homologous region in the two species suggests that the size of the *D. pseudoobscura* region is insufficient to harbor Acp36DE. The *D. melanogaster* Acp36DE CDS covers 2,739 bp and includes two exons. The second exon is considerably larger, coding for 843 of the 912 protein residues. Nevertheless, the homologous region of *D. pseudoobscura* spans only 1,471 bp (fig. 1H). The largest possible ORF (including those not starting with methionine) in this region of *D. melanogaster* second exon (309 bp in *D. pseudoobscura* compared with 2,531 bp in *D. melanogaster*). Finally, our molecular data provide no evidence in *D. pseudoobscura* for transcripts in the region corresponding to the Acp36DE transcript region of *D. melanogaster* (fig. 2).

**Acp63F**

Proximal to Acp63F, CG1065 exons 2 to 4 generate significant tBlastN homology to *D. pseudoobscura* chromosome XR (E = 4e–67, 2e–74, and 2e–74 for exons 2 to 4, respectively [fig. 1I]). Distally, the small first exon of CG1065 also generates a microsyntenic BlastN hit (E = 2e–14; BlastN only because of small exon size of 13 residues). tBlastN analysis of Acp63F produced no significant or even marginal hits to the *D. pseudoobscura* genome.

The intron–exon organization of CG1065 is conserved between the two species. However, there is a major difference between *D. melanogaster* and *D. pseudoobscura* in the size of the first intron, which defines the boundaries of the Acp63F gene region in *D. melanogaster*. The intron is almost five times larger in *D. melanogaster* than in *D. pseudoobscura* (2.3 kb versus 470 bp, respectively). The candidate region that would contain the *D. pseudoobscura* Acp63F ortholog can be further refined by noting a small stretch of apparently conserved first-intron nucleotides (26/27 identical to *D. melanogaster*) within 61 bp of the *D. pseudoobscura* CG1065 first exon. Thus, the *D. pseudoobscura* genomic region that would contain Acp63F (start to stop codon) is 383 bp. The *D. melanogaster* Acp63F genomic sequence from start to stop codon (including introns) is 361 bp. Including putative 5' and 3' flanking UTRs, the *D. melanogaster* region is 432 bp. Therefore, it seems rather unlikely that the *D. pseudoobscura* Acp63F gene would fit within this much smaller piece of DNA. Finally, and most importantly, our molecular experiments provide no evidence for *D. pseudoobscura* transcripts associated with the region that would contain Acp63F based on patterns of microsynteny in the two species (fig. 2).

**Acp76A**

*D. melanogaster* Acp76A is a relatively large accessory gland gene, consisting of a 994-bp first exon, a 69-bp intron, and a 173-bp second exon. The Acp76A protein contains a serpin domain. Figure 1J illustrates Blast results comparing the *D. melanogaster* Acp76A gene region with the *D. pseudoobscura* genome sequence. BlastN analysis of a 2-kb region of 5' flanking DNA revealed three contiguous matches (E ranging from 1e–28 to 2e–08) averaging 80 bp. BlastN comparison of 2 kb of 3' flanking DNA returned a highly significant result (E ranging from 8e–26 to 2e–10) of five contiguous nucleotide sequences averaging 83 bp each. These regions correspond to *D. pseudoobscura* chromosome XR. The amount of genomic DNA defined by these regions of sequence similarity is about 2.3 kb in *D. melanogaster* but only 1,031 bp in *D. pseudoobscura*. Thus, given the size of the *D. melanogaster* transcript (1,235 bp from start to stop, intron included), it seems unlikely that there would be sufficient genomic sequence to harbor a similarly structured *D. pseudoobscura* homolog. Furthermore, this candidate *D. pseudoobscura* region shows no Blast similarity to *D. melanogaster* Acp76A; its largest possible ORF is only 61 residues or 183 bp, which is considerably shorter than the 994-bp first exon of *D. melanogaster* Acp76A. Finally, we found no evidence of a *D. pseudoobscura* transcript associated with the 1,235-bp candidate region of DNA (fig. 2).

Although the microsyntenic region does not appear to contain a *D. pseudoobscura* Acp76A ortholog, we observed two weakly significant tBlastN hits to Acp76A from other parts of the *D. pseudoobscura* genome. The strongest hit was to chromosome 3 (E = 2e–06) but was ruled out as a true ortholog based on the fact that a BlastN search of its predicted peptide sequence back to *D. melanogaster* genes returned more than 20 serpin domain–containing genes with considerably lower E scores than the Acp76A score (E = 3e–9 for Acp76A, compared with a low of E = 3e–63 for CG9456). The other weakly significant BlastN hit to this gene in *D. pseudoobscura* comprised two contiguous stretches of peptide sequence to a nontyntenic portion of chromosome XR (E = 7e–04). When compared with *D. melanogaster* predicted proteins, the candidate peptide sequences only returned Acp76A as a significant BlastP hit (E = 7e–7). However, the corresponding *D. pseudoobscura* genomic sequence does not appear to contain a viable candidate ortholog. The putative peptide sequences correspond to residues 199 to 239 and 271 to 298, both from the first exon of *D. melanogaster* Acp76A. The similar sequences in *D. pseudoobscura* are in the proper order but are separated by 65 bp, negating the possibility of a single continuous reading frame covering both matches. Moreover, the largest possible ORF that includes either of these putative peptide sequences is only 60 residues, less than one fifth of the amino acid sequence coded for by the first exon in *D. melanogaster*. Additionally, several attempts to amplify RACE products associated with this candidate sequence failed, suggesting that transcription within this region is unlikely.

**Acp95EF**

*D. melanogaster* Acp95EF contains two exons and has a strongly predicted signal sequence (table 1). Based on tBlastN analysis, neighboring genes are present in *D. pseudoobscura* (fig. 1K). The proximal neighbor, CG13609, generated a highly significant tBlastN hit to a portion of *D. pseudoobscura* chromosome 4 (E = 3e–42). CG5677 is also highly conserved in the same relative position in *D. pseudoobscura* (E = 3e–96). tBlastN analysis of Acp95EF, however, did not produce even a weak hit to any portion of the *D. pseudoobscura* genome. Conservation of Muller elements within *Drosophila* suggests *D. melanogaster* chromosome...
3R is homologous to *D. pseudoobscura* chromosome 2 (Lakovaara and Saura 1982; Steinemann, Pinkser, and Sperlich 1984). Whether this apparent 3R-to-4 homology is real or an error in the *D. pseudoobscura* genome assembly is unclear. Regardless, the microsynteny of Acp95EF flanking genes clearly defines a candidate region for a *D. pseudoobscura* ortholog.

The region of microsynteny defined by CG13609/CG5677, which would contain *D. pseudoobscura* Acp95EF, is only 204 bp, compared with 1.2 kb in *D. melanogaster*. The genomic sequence from start to stop codon of *D. melanogaster* Acp95EF spans 221 bp. Given the requirements for 5' and 3' UTRs, it seems highly improbable that a *D. pseudoobscura* Acp95EF homolog is located within this 204-bp *D. pseudoobscura* genomic sequence. The small size of the candidate region coupled with encroaching 3' UTRs of CG13609/CG5677 made reverse Northern analysis superfluous. Computational analysis is enough to dismiss the hypothesis of a microsyntenic *D. pseudoobscura* ortholog. There is only one possible initiation codon in this region. Unlike *D. melanogaster* Acp95EF (SignalP, *P* = 1.0), an intronless *D. pseudoobscura* peptide sequence originating from this codon is not strongly predicted to have a signal peptide (SignalP, *P* = 0.71) and could not exceed 23 residues. Furthermore, an ortholog of comparable length would be impossible within this region, even assuming intron loss in *D. pseudoobscura*. Given the requirements for intron splicing sites and conservatively assuming a minimum intron size of 40 bp, the longest possible *D. pseudoobscura* ortholog could still only consist of 30 residues, less than 58% of the size of the relatively small *D. melanogaster* Acp95EF protein. A signal sequence for this candidate is also not strongly predicted (SignalP, *P* = 0.64). Thus, our computational evidence leads us to conclude that a *D. pseudoobscura* Acp95EF ortholog is not present within this microsyntenic region and that Acp95EF is likely a *D. melanogaster* orphan.

**Acp98AB**

Acp98AB is in a gene-rich portion of chromosome 3R in *D. melanogaster*. It is located within the 757 bp intron of CG12879. The Acp98AB ORF does not contain any easily detected signature sequences for computational analysis. There is no evidence of a typical methionine initiation codon and predicted peptide lengths vary from 28 to 31 residues, depending on the assumed first codon. There are no conserved domains and no evidence for a signal peptide sequence (SignalP, *P* = 0.0 [table 1]). There are no tBlastN hits in *D. pseudoobscura* to suggest an ortholog to Acp98AB. The neighboring genes, however, reveal the homologous region in *D. pseudoobscura*. tBlastN scores for the second exon of CG12879 (*E* = 1e–162), as well as two distal neighbors, CG12876 and CG12878 (*E* = 0.0 and 1e–111, respectively) clearly indicate this homologous region as a portion of *D. pseudoobscura* chromosome 2 (fig. 1L). This homology is also reinforced by BlastN analysis of 2 kb of noncoding DNA proximal to CG12879 in *D. melanogaster*. A total of seven small nucleotide sequences, averaging 58 bp in length, are microsyntenous between the two species (E values from *E* = 5e–24 to *E* = 3e–4; partially depicted by homologous region a [figure 1L]). One additional gene, CG12880, is immediately proximal to these matching nucleotide sequences. tBlastN analysis shows that this gene is also in a microsyntenic position in *D. pseudoobscura* (*E* = 2e–62, not shown in figure 1L). Just 5' of CG12878 CDS, BlastN analysis identified one additional microsyntenic nucleotide sequence, depicted as region c in figure 1L (*E* = 2e–12, 51/55 identical).

Comparison of the relative positions of these genes shows an inversion event between *D. melanogaster* and *D. pseudoobscura*. Based on clear regions of orthology, this inversion covers at least the second exon of CG12879 and the entire CG12876 gene. The regions labeled a and c in figure 1L are the closest conserved markers clearly outside of the inversion breakpoints. The unknown location of the first CG12879 exon in *D. pseudoobscura* (no tBlastN or BlastN identity was detected) complicates efforts to determine whether or not Acp98AB might have been included in the inversion. In fact, our RACE data show CG12879 to be an intronless gene in *D. pseudoobscura*. There are no intron gaps in the consensus 5' *D. pseudoobscura* RACE sequence and a single ORF possibility (moving upstream from the putative initiation codon, a stop codon comes into frame before an alternative initiation codon is reached). The protein alignment between species is very robust beyond the missing *D. pseudoobscura* first exon, with the first *D. pseudoobscura* residue matching residue 61 in *D. melanogaster* and high levels of conservation continuing to the end of the protein for an overall 69.8% level of similarity. We should note that there is no empirical support from full-length cDNAs or expressed sequence tags (ESTs) for the annotated *D. melanogaster* first exon. In fact, an alternate initiation codon exists in *D. melanogaster* that leads to a 398-residue, single-exon protein that is the exact same size as its *D. pseudoobscura* counterpart. Thus, we proceeded to target candidate regions in *D. pseudoobscura* under the conservative assumption that the first exon of *D. melanogaster* CG12879 may not be real.

If Acp98AB were included in the inversion, we would expect the *D. pseudoobscura* ortholog to be on the minus strand between CG12879 and conserved region c in figure 1L. Alternatively, if Acp98AB were outside of the inversion breakpoints, we would expect the *D. pseudoobscura* ortholog to be on the plus strand between conserved region a and CG12876 in figure 1L. These possibilities lead to candidate regions of 352 bp and 2 kb, respectively. BlastN analysis of the 2-kb sequence to all *D. melanogaster* sequences revealed a highly significant match to Jonah99C (four separate matches averaging 116 bp, E scores from 2e–55 to 1e–9 [region b, figure 1L]), a member of a gene family that includes multiple repetitive sequences (Carlson and Hogness 1985). Excising the sequence spanning Jonah99C BlastN matches, two *D. pseudoobscura* candidate regions of 797 bp and 407 bp exist between microsyntenic region a and CG12876. The 407-bp candidate region can be further condensed to approximately 360 bp, considering the requirements for a CG12876 5' UTR. Thus, through our analyses of *D. melanogaster*/*D. pseudoobscura* microsynteny, we have narrowed the *D. pseudoobscura* Acp98AB candidate space to three sequences of *D. pseudoobscura* chromosome 2, covering approximately 1.5 kb and spanning less than 7 kb.
Because of the fragmented nature of the candidate regions and the uncertainty about transcription boundaries of the tightly arranged adjacent genes, reverse Northern and RACE analyses were impractical. The power of our computational analyses was compromised by the short Acp98AB gene sequence, the lack of a traditional methionine start codon, and the absence of signature sequences such as a conserved domain or predicted signal sequence. A total of 19 ORFs are possible within the three D. pseudoobscura candidate sequences (13, 3, and 3 for the three candidate sequences from left to right [fig. 1L]). However, none show any resemblance to D. melanogaster Acp98AB. Thus, we propose that Acp98AB is a D. melanogaster orphan, though a highly diverged D. pseudoobscura ortholog would be very difficult to detect.

Discussion
Evidence of Gene Presence Versus Absence

Comparative analyses of D. melanogaster and D. pseudoobscura have shown that most annotated genes are conserved between lineages (Bergman et al. 2002; Richards et al. 2005). This stands in stark contrast to the results reported here for Acps. Ignoring Acp33A (because of the incomplete genome assembly), we demonstrate likely orthology for only half (6/12) of the analyzed incomplete genome assembly), we demonstrate, reported here for D. melanogaster Acp98AB. Thus, we propose that Acp98AB is a D. melanogaster orphan, though a highly diverged D. pseudoobscura ortholog would be very difficult to detect.

Comparison of Orthologous Acps

Varying levels of protein conservation were observed for the six genes for which homologs were identified in the two species (table 1). The weighted average of amino acid identity across the alignable portions of these six orthologs is 35.6% (or 39.3%, including Acp33Ea duplicates). This level of conservation is much lower than the reported modal similarity of 85% for all orthologous pairs across the D. melanogaster/D. pseudoobscura genomes (Richards et al. 2005). Our Acp protein similarity translates to a conservative Ka estimate of about 0.28 (assuming only one replacement mutation per diverged residue and 2.3 replacement sites per codon). In contrast, Bergman et al. (2002) estimate 0.146 replacement divergence between D. melanogaster/D. pseudoobscura across a semi-random set of 41 genes. Thus, the subset of Acps for which we were able to identify D. pseudoobscura orthologs evolve at a much faster rate than other genes, as expected based on previous observations from the melanogaster subgroup (e.g., Begun et al. 2000; Swanson et al. 2001).

Of particular interest are proteins that are clearly orthologous based on genomic location, gene organization and length, and gene expression but for which divergence is so great that protein sequences provide no support for orthology. A good example is Acp26AA, which is not detectable through tBlastN analysis but is clearly orthologous in the two species. In D. melanogaster, Acp26AA transferred during mating is processed by the female and has effects on oviposition during the first 24 hours postmating (Herndon and Wolfner 1995; Heifetz et al. 2000). Whether Acp26AA has similar functions in the two species despite the lack of sequence similarity is an interesting question. The finding that Acp26AA protein evolves rapidly in two distantly related Drosophila lineages as a result of directional selection suggests that a history of directional selection at this gene will be widely shared among species from this genus. It remains to be seen what other Acps or other types of proteins tend to be under directional selection during most of their evolutionary history. Given the long history of adaptive evolution between D. melanogaster and D. pseudoobscura Acp26AA, a comparative functional analysis would be most interesting and could potentially reveal whether the underlying mechanisms of natural selection are similar in the two lineages.

Implications for Functional Biology

Previous population genetic data from Acp29AB and Acp36DE support the idea that both have been under directional selection in D. melanogaster/D. simulans (Aguadé 1999, Begun et al. 2000). Thus, the fact that our analysis suggests that both are absent from the D. pseudoobscura genome is particularly interesting. There are two possible explanations for the presence/absence data. Either both genes were present in the D. melanogaster/D. pseudoobscura ancestor and then lost in the D. pseudoobscura lineage or both genes were gained in the D. melanogaster lineage. The approaches used here, when applied to other Drosophila species, are likely to provide a clear answer to this question. Still, from an evolutionary perspective, either
A scenario is interesting. If the genes originated in the *D. melanogaster* lineage and are also under directional selection in that lineage, one might speculate that this is a common feature of lineage-specific new genes, consistent with data from other such genes in *Drosophila* (reviewed in Long et al. [2003]). Alternatively, if the genes were lost in the *D. pseudoobscura* lineage but were under directional selection in *D. melanogaster/D. simulans*, the interpretation would be that radically different selection regimes had been operating in these two lineages.

Of course, the evolutionary questions have a parallel in issues relating to the functional biology of these two genes and these two species. For example, the evidence for directional selection of *Acp29AB* and *Acp36DE* in *D. melanogaster/D. simulans* certainly suggests they are functionally important. Although the function of *Acp29AB* is unknown, flies that are mutant for *Acp36DE* in *D. melanogaster* have major defects. *Acp36DE* protein is required for proper sperm storage. Females mated to mutant males lacking *Acp36DE* store only 15% as many sperm as females mated to wild-type males (Neubaum and Wolfnier 1999). This protein binds to sperm heads and also localizes to the opening of the sperm storage organs (Bertram, Neubaum, and Wolfnier 1996). The loss of sperm from seminal receptacles occurs rapidly on the second day after mating, thus affecting female patterns of remating as continued female resistance to male mating attempts requires stored sperm (Neubaum and Wolfnier 1999). It would be fair to say that the *Acp36DE* protein plays an important role in *D. melanogaster* fertility. Given these data and our presence/absence data, there are two possible interpretations. Either the function of *Acp36DE* is required in both lineages, yet is fulfilled by another protein in *D. pseudoobscura*, or the functional biology of male–female interactions are sufficiently diverged such that not all functions are represented in all *Drosophila* lineages. Genetic analysis should allow these alternatives to be distinguished.

X Chromosome Versus Autosomal Linkage of *D. pseudoobscura* Acps

The ancestral *Drosophila* karyotype is five acrocentric rods (Ashburner 1989). In the *D. pseudoobscura* lineage, a relatively recent X chromosome–autosome fusion has resulted in a large X chromosome that contains roughly 40% of the genome, rather than the typical 20% for most species, including *D. melanogaster* (Powell and DeSalle 1995). In *D. melanogaster*, Acps and other genes associated with male reproduction appear to be underrepresented on the X chromosome (Wolfnier et al. 1997; Parisi et al. 2003; Ranz et al. 2003). Conservation of *Drosophila* Muller elements strongly predicts that some Acps that were on the chromosome corresponding to *D. melanogaster* 3L became X-linked in the lineage leading to *D. pseudoobscura* as a result of fusion of Muller elements (corresponding to X and 3L of *D. melanogaster*). If selection disfavors X-linked Acps, genes corresponding to 3L Acps in *D. melanogaster* should have been under strong selection for loss or transposition to an autosome in *D. pseudoobscura*. In fact, our two examples of Acp-related rearrangements leading to nonhomologous locations for orthologs (*Acp62F* and *Acp70A*) were 3L-located *D. melanogaster* genes that have avoided XR-linkage in *D. pseudoobscura* (but see Stevison, Counterman, and Noor [2004] for XR-linked Acps). Moreover, two other Acps, *Acp63F* and *Acp76A*, which should be on XR in *D. pseudoobscura*, appear to be entirely absent from the *D. pseudoobscura* genome. Thus, none of the four Acps that should be X-linked in *D. pseudoobscura* as a result of an X chromosome–autosome fusion actually are X-linked. This supports the idea that X chromosome versus autosome location can have major roles in the evolution of genome content and organization (Betran, Thornton, and Long 2002).

One hypothesis for this pattern is that natural selection disfavors X-linked locations for male-advantage genes that are deleterious to females (Parisi et al. 2003). Our data are consistent with this hypothesis. Acps have been implicated as the likely components of seminal fluid that confer a cost of mating to females (Chapman et al. 1995). Little is known about the specific phenotypes associated with *Acp62F* and *Acp76A*. However, *Acp62F* is a protease inhibitor that is known to be toxic upon ectopic expression in females (Lung et al. 2002). *Acp76A*, although not shown to be deleterious to females, is a protein that serves a male agenda by increasing egg laying rate and reducing female receptivity to remating (Chen et al. 1988; Chapman et al. 2003; Liu and Kubli 2003). Further analysis of comparative genomic data and elucidation of additional Acp phenotypes will help explain the X chromosome versus autosome disparity in male-biased genes.

Acknowledgments

This work was supported by grants NIH GM55298 and NSF DEB-0327049 and an NSF Doctoral Dissertation Improvement Grant. We thank M. Nachman and two anonymous reviewers for useful comments.

Literature Cited


Beckenbach A., Y. Wei, and H. Liu. 1993. Relationships in the *Drosophila obscura* species group, inferred from


Michael Nachman, Associate Editor

Accepted November 25, 2004