Recently evolved genes identified from Drosophila yakuba and Drosophila erecta accessory gland expressed sequence tags

Permalink
https://escholarship.org/uc/item/9hb7461t

Journal
Genetics, 172(3)

ISSN
0016-6731

Authors
Begun, D J
Lindfors, H A
Thompson, M E
et al.

Publication Date
2006-03-01

Peer reviewed
Recently Evolved Genes Identified From Drosophila yakuba and D. erecta Accessory Gland Expressed Sequence Tags

David J. Begun,1 Heather A. Lindfors, Melissa E. Thompson and Alisha K. Holloway

Section of Evolution and Ecology, University of California, Davis, California 95616

Manuscript received August 31, 2005
Accepted for publication November 22, 2005

ABSTRACT

The fraction of the genome associated with male reproduction in Drosophila may be unusually dynamic. For example, male reproduction-related genes show higher-than-average rates of protein divergence and gene expression evolution compared to most Drosophila genes. Drosophila male reproduction may also be enriched for novel genetic functions. Our earlier work, based on accessory gland protein genes (Acp’s) in D. simulans and D. melanogaster, suggested that the melanogaster subgroup Acp’s may be lost and/or gained on a relatively rapid timescale. Here we investigate this possibility more thoroughly through description of the accessory gland transcriptome in two melanogaster subgroup species, D. yakuba and D. erecta. A genomic analysis of previously unknown genes isolated from cDNA libraries of these species revealed several cases of genes present in one or both species, yet absent from ingroup and outgroup species. We found no evidence that these novel genes are attributable primarily to duplication and divergence, which suggests the possibility that Acp’s or other genes coding for small proteins may originate from ancestrally noncoding DNA.

A

An extensive literature documenting the unusually rapid evolution of reproductive traits in many taxa suggests that sexual selection may be a primary agent of evolution in natural animal populations (e.g., Eberhard 1985; Andersson 1994; Birkhead and Moller 1998). Although most data bearing on evolution of reproductive traits are morphological or behavioral in nature, directional selection on reproductive function should be manifest in patterns of genome evolution. For example, a genomic approach for identifying biological functions that may be under directional selection is to use sequence divergence in concert with gene annotation to identify functions enriched for rapidly evolving proteins (e.g., Nielsen et al. 2005; Richards et al. 2005). Such analyses support the idea that proteins functioning in male reproduction in Drosophila, mice, and primates evolve unusually quickly (Zhang et al. 2004; Good and Nachman 2005; Nielsen et al. 2005; Richards et al. 2005). Such data do not prove that rapid evolution results from directional selection. However, the repeatability across taxa of the pattern of rapid protein evolution is certainly consistent with this idea.

Drosophila ACPs (seminal fluid proteins) have been the subject of several evolutionary and functional investigations. These proteins elicit manifold physiological and behavioral changes in females (reviewed in Chapman and Davies 2004) and play an important role in sperm storage (Neubaum and Wolfer 1999; Tram and Wolfer 1999). They evolve quite rapidly compared to most proteins (Begun et al. 2000; Swanson et al. 2001; Holloway and Begun 2004; Kern et al. 2004; Mueller et al. 2005; Wagstaff and Begun 2005a,b). Population genetic evidence for directional selection on Acp’s has been found in the melanogaster subgroup, the repleta group, and the obscura group of Drosophila (Tsaur and Wu 1997; Aguadé 1999; Begun et al. 2000; Holloway and Begun 2004; Kern et al. 2004; Wagstaff and Begun 2005a,b; Begun and Lindfors 2005), perhaps due to male–male, male–female, or fly–pathogen interactions.

As noted previously, genomic surveys of divergence of male reproduction-related genes have demonstrated that they evolve rapidly compared to most other protein classes. Indeed, many testis-expressed Drosophila melanogaster genes have no obvious homolog in D. pseudoobscura (Richards et al. 2005), which is consistent with either very rapid evolution or gene presence/absence variation (i.e., lineage-restricted genes). The notion that genes coding for male reproductive functions may be enriched for lineage-restricted genes in Drosophila is supported by reports of recently evolved, novel genes that are expressed in Drosophila testes (Long and Langley 1993; Nurminsky et al. 1998; Betran and Long 2003).

Although there has been little systematic investigation regarding the question of whether reproductive functions are characteristic of lineage-restricted genes, we previously reported that in Drosophila, an Acp in a given species is sometimes absent from a related species (Begun and Lindfors 2005; Wagstaff and Begun 2005a). For example, 6 of 13 D. melanogaster Acp’s investigated were absent from D. pseudoobscura (Wagstaff and Begun 2005a). A subsequent analysis of additional
**D. melanogaster Acp**'s vs. D. pseudoobscura yielded comparable results (Mueller et al. 2005). A subset of the D. melanogaster Acp's that are absent from D. pseudoobscura have loss-of-function phenotypes or show evidence of directional selection in D. melanogaster/D. simulans, which suggests that invoking “functional redundancy” and gene loss is overly simplistic. In fact, these analyses of D. melanogaster vs. D. pseudoobscura could not broach the issue of whether the lineage distribution of Acp's in these two species is explained by gene loss in D. pseudoobscura, gene gain in D. melanogaster, or some combination. We also found putative cases of recent loss of Acp's in the melanogaster/simulans subgroup (Begun and Lindfors 2005). For example, D. melanogaster is missing an Acp that was present in the common ancestor of D. melanogaster and D. simulans and that is present as a single-copy gene in D. simulans, indicating that this gene was lost within the last 2–3 million years. Begun and Lindfors (2005) did not find unambiguous evidence for gains of Acp's in the melanogaster subgroup. Nevertheless, loss of Acp's implies either that compensatory gains maintain melanogaster subgroup seminal fluid protein-coding capacity or that the melanogaster subgroup is evolving toward a lower equilibrium number of Acp's per genome.

The gain and/or loss of Acp's over time will result in the gradual functional divergence of seminal fluid function between Drosophila lineages, presumably under the influence of natural selection. One possible mechanism for gene gain is duplication followed by functional divergence (Ohno 1970). However, computational analysis of the D. melanogaster genome suggested that most duplicated Acp's are ancient (Holloway and Begun 2004; Mueller et al. 2005), which does not support the idea that recent losses of the melanogaster subgroup Acp's are entirely compensated for by recent duplication and divergence. The purpose of the work presented here was to systematically investigate potential gains of Acp's in the melanogaster subgroup of Drosophila. This was accomplished by description of the accessory gland transcriptome in D. yakuba and D. erecta, followed by computational analysis of melanogaster group species genome assemblies. We have assumed that D. yakuba and D. erecta are sister species (Ko et al. 2003; Parsch 2003); D. ananassae served as the outgroup.

**MATERIALS AND METHODS**

**D. yakuba and D. erecta accessory gland cDNA libraries and ESTs:** Accessory glands from 100 D. yakuba males (line Tai8E2) and 45 D. erecta males (line 14021-0224.0) were dissected in RNA-Later (Ambion, Austin, TX). Total accessory gland RNA was isolated using the Ambion mirVana miRNA kit and RNaseA (Ambion DNA-Free kit). RACE-ready cDNA was synthesized from 2 μg of each prep [Invitrogen (San Diego) GeneRacer kit; the SSII module and oligo(dT) primer were used for the RT step]. The resulting cDNA was amplified (eight cycles for D. erecta; five cycles for D. yakuba) using the Roche Expand High Fidelity PCR System. Amplified libraries were purified [QIAGEN (Chatsworth, CA) QIAquick PCR purification kit], incubated in Promega (Madison, WI) Taq polymerase, and ligated into PCR4 TOPO vector (Invitrogen). Ligations were transformed and plated, with the resulting colonies subjected to PCR using vector primers. Colony PCR products were sequenced at the University of California at Davis College of Agricultural and Environmental Sciences Genomics Facility. For D. yakuba, 415 clones were sequenced. They yielded 360 high-quality sequences, which assembled (Lasergene) into 119 unique contigs. For D. erecta, 335 clones were sequenced. They yielded 252 high-quality sequences and 114 unique contigs. Unique D. yakuba and D. erecta accessory gland ESTs can be found under GenBank accession nos. DV998435–DV998658.

The complexity of these libraries appears to be considerably greater than that estimated from random sequencing of a D. mojavensis accessory gland cDNA library (Wagstaff and Begun 2005b; 26 transcripts from 139 random clones). This suggests that Drosophila species vary in the complexity of the accessory gland transcriptome, but more quantitative data would be required to address this issue.

**Analysis of ESTs:** Each unique EST was compared by BLAST to predicted D. melanogaster genes and proteins. ESTs returning E-values <1e-15 were considered to be candidate unannotated homologous Acp’s or candidate Acp’s absent from the D. melanogaster genome. Each candidate was then compared (BLASTn) to D. melanogaster chromosome arms to determine if there was evidence for an unannotated D. melanogaster gene corresponding to the D. yakuba or D. erecta EST. ESTs that failed to show convincing BLAST hits to D. melanogaster were candidate lineage-restricted genes (although they could also be highly diverged orthologs). RACE was used to isolate the entire transcript associated with each putative lineage-restricted gene. These genes were investigated in terms of splicing, predicted protein sequence, and whether they were present as putative single-copy genes in D. yakuba or D. erecta on the basis of BLAST or BLAT analyses to genome assemblies. Finally, given that most ACPs have strongly predicted signal sequences (Swanson et al. 2001), which are required for secretion, the predicted proteins were analyzed by SignalP to determine the likely presence/absence of a signal peptide (Bendtsen et al. 2004). Candidate lineage-restricted genes were subjected to additional investigation, as described in the next section.

**Search for orthologs based on syntenic alignments:** Syntenic regions of variable size (generally several kilobases) encompassing each candidate gene were isolated from the D. yakuba or D. erecta genome assemblies (BLAT via the UCSC genome browser [Kent et al. 2002; http://genome.ucsc.edu]) to D. yakuba (Release 1.0; Washington University Medical Genome Sequencing Center) or BLAT to D. erecta contigs (October 2004 assembly; sequencing by Agencourt) at http://rana.lbl.gov/drosophila/. These regions were then analyzed by BLAT to identify putative orthologous regions of the D. melanogaster genome. This resulted in a putative orthologous region from D. melanogaster, D. yakuba, and D. erecta for each candidate, along with the gene annotation derived from our EST/RACE data and computational analysis for either D. yakuba or D. erecta. Finally, we attempted to isolate a syntenic region from D. ananassae (July 2004 assembly; sequencing by Agencourt) for each candidate. Generally, this was more difficult (and not always successful), probably because of greater sequence divergence, and often required investigation of larger genomic regions, occasionally up to 10–15 kb. Each gene region identified from a D. yakuba or a D. erecta accessory gland EST was investigated in detail in the corresponding region of the other species. This entailed pairwise alignments using the Martinez/Needleman-Wunsch algorithm as implemented in DNASTAR and/or multispecies alignments using
ClustalW v. 1.82. In many cases, there was no DNA in other species corresponding to the gene of interest. In other cases, there was apparently a homologous sequence, but no obvious conserved open reading frame (ORF). For the latter, we computationally investigated the genomic sequence in the homologous region to determine protein-coding capacity and whether any putative proteins showed sequence similarity or similar protein lengths relative to the candidate, or whether a predicted protein had a predicted signal sequence. In a few cases, these investigations revealed evidence for highly diverged orthologous genes, likely Acp’s, which would have gone undetected on the basis of the alignment of DNA sequences.

**Population genetic analysis:** Molecular population genetic data were collected for several D. yakuba- and/or D. erecta-specific genes. High-fidelity PCR was used to amplify Acp’s from multiple D. yakuba isofemale lines and a single D. teissieri isofemale line (provided by P. Andolfatto and M. Long, respectively). These PCR products were cloned and subjected to colony PCR. A single allele was isolated and sequenced from each line. Summary statistics and tests of neutral evolution were generated by use of DnaSP (ROZAS et al. 2003). Sequence data for the population genetics analysis can be found under GenBank accession nos. DQ318145–319181.

**Signal sequence potential of D. melanogaster intergenic and intronic sequences:** Intergenic sequences (defined as sequences between two adjacent genes, independent of a strand) and introns were obtained from release 4.1 of the D. melanogaster genome. Intron sequences were parsed to mark known exons embedded within them. RepeatMasker (SMIR et al. 1996–2004) was used to mask repetitive elements of intergenic and intronic sequences. A Perl script was used to identify single-exon ORFs within them. RepeatMasker (Smit et al. 1996–2004) was used to mask repetitive elements of intergenic and intronic sequences. A Perl script was used to identify single-exon ORFs within them.

**RESULTS**

Many of our D. yakuba/D. erecta accessory gland ESTs returned highly significant BLAST hits to annotated D. melanogaster genes or proteins. These were not considered further. Several ESTs had highly significant BLAST hits to unannotated D. melanogaster sequence (as well as to D. yakuba and D. erecta genomic sequence). On the basis of the conserved location and organization of an open reading frame and the presence of a strongly predicted signal sequence in either D. yakuba or D. erecta and D. melanogaster, we consider 20 genes to be candidates for previously unknown Acp’s that are shared among melanogaster subgroup species [supplemental Data A at http://www.genetics.org/supplemental/ presents the putative D. melanogaster protein-coding sequence (CDS) for each gene]. However, additional empirical work would be required to solidify their status as such.

Accessory gland ESTs for which we failed to find putative orthologs in other species are presented in more detail below. None are associated with repetitive sequences; all show male-specific expression as determined by RT–PCR on templates generated from RNA isolated from whole adult males or females. Syntenic alignments of these putative lineage-restricted genes and orthologous regions can be found in Supplemental Data B at http://www.genetics.org/supplemental/; putative CDS regions are in boldface type with the exception of Gene144, for which the transcript is in boldface type; introns are underlined. Table 1 summarizes inferred phylogenetic distributions of putative lineage-restricted genes and some physical properties of the gene/protein, including the probability that the predicted amino acid has a signal sequence, which is frequently found in Acp’s (SWANSON et al. 2001). Table 2 presents the results of BLAST analysis of several D. yakuba accessory gland ESTs corresponding to putative novel genes compared to the genomes of D. yakuba (April 2004 assembly), D. melanogaster (release 4.2.1), D. erecta (August 2005 assembly), and D. ananassae (August 2005 assembly). Table 3 provides summary statistics of D. yakuba polymorphism and divergence to D. teissieri for five genes.

**Putative lineage-restricted genes identified from D. yakuba accessory gland ESTs:** Acp134 codes for a predicted protein of 35 residues. This gene is represented in the D. yakuba testis EST collection (CV785591, CV785729, CV786139), probably as a result of low-level contamination of the testis dissection with accessory gland tissue. Acp134 returns no significant BLAST results vs. D. melanogaster, D. erecta, or D. ananassae. The putative syntenic alignments for the D. yakuba Acp134 region with D. melanogaster, D. erecta, and D. ananassae suggest that there are no plausible orthologous protein-coding regions in D. melanogaster, D. erecta, or D. ananassae that correspond

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>SignalP</th>
<th>Length (aa)</th>
<th>No. of exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acp134</td>
<td>yakura-derived</td>
<td>1.000 35 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp225</td>
<td>yakura-derived</td>
<td>0.991 121 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp223</td>
<td>yakura-derived</td>
<td>1.000 116 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp224</td>
<td>yakura-derived</td>
<td>1.000 231 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp158</td>
<td>yakura-derived</td>
<td>1.000 71 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene144</td>
<td>yakura-derived</td>
<td>1.000 231 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp157a</td>
<td>yakura-derived</td>
<td>0.981 112 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 1**

Summary of inferred phylogenetic distributions of genes identified from D. yakuba and D. erecta accessory gland ESTs

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>SignalP</th>
<th>Length (aa)</th>
<th>No. of exons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acp15</td>
<td>erecta-derived</td>
<td>0.998 114 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Acp100</td>
<td>erecta-derived</td>
<td>0.999 190 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gene37</td>
<td>erecta-derived</td>
<td>0.010 80 2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Moreover, a computational analysis of these orthologous regions also revealed no potential genes that were plausible orthologs. These data strongly suggest that \textit{Acp134} is present only in \textit{D. yakuba}.

\textit{Acp225} codes for a predicted protein of 121 residues. The syntenic alignment strongly suggests that there is no ortholog of \textit{Acp225} in \textit{D. melanogaster} or \textit{D. erecta}. A small ORF (36 bp) in \textit{D. erecta} in the region near the first exon of \textit{D. yakuba} \textit{Acp225} is clearly not orthologous. A putative syntenic alignment between \textit{D. yakuba} and \textit{D. ananassae} is presented in supplemental data at http://www.genetics.org/supplemental/. However, the quality of this alignment leads us to consider the status of the gene in \textit{D. ananassae} as ambiguous.

\textit{Acp223} codes for a predicted protein of 116 residues. It is located between the \textit{D. yakuba} orthologs of \textit{Obp56f} and \textit{Obp56e}. Indeed, the organization of the three genes is similar, which together with their physical location, suggests that they are paralogous. \textit{D. erecta} also has a copy of \textit{Acp233}. \textit{D. yakuba} \textit{Acp223} is more highly diverged from the \textit{D. yakuba} \textit{Obp56e} and \textit{Obp56f} genes than these genes are from one another. A partial, homologous \textit{D. melanogaster} ORF appears to be present; however, it codes for a predicted protein of only 44 residues, which leaves it with questionable status in \textit{D. melanogaster} (Supplemental Data B at http://www.genetics.org/supplemental/). A syntenic alignment of the putative \textit{D. ananassae} orthologous region with \textit{D. yakuba} provides no evidence for a \textit{D. ananassae} copy of \textit{Acp223}.

\textit{Acp224} codes for a predicted protein of 231 residues in \textit{D. yakuba} and is located within an intron of \textit{CG31757}. An alignment of the orthologous region from \textit{D. erecta} reveals that the reading frame starting with the \textit{D. yakuba} initiation codon codes for a predicted protein of 75 residues. However, the fact that the \textit{D. yakuba} gene and the putative \textit{D. erecta} ortholog are extremely divergent in terms of length and sequence casts some doubt on the status of the \textit{D. erecta} gene. To address this uncertainty, we used RACE on accessory gland cDNA to isolate the ends of the \textit{D. erecta} gene. The RACE results revealed that there is an apparently orthologous \textit{D. erecta} transcript, which codes for two potential ORFs (89 codons

<table>
<thead>
<tr>
<th>Gene</th>
<th>Species</th>
<th>melanogaster annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yak</td>
<td>ere</td>
</tr>
<tr>
<td>\textit{Acp134}</td>
<td>2e30</td>
<td>4e05</td>
</tr>
<tr>
<td></td>
<td>3e13</td>
<td></td>
</tr>
<tr>
<td>\textit{Acp225}</td>
<td>0.0</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>\textit{Acp223}</td>
<td>0.0</td>
<td>5e36</td>
</tr>
<tr>
<td></td>
<td>0.013</td>
<td></td>
</tr>
<tr>
<td>\textit{Acp224}</td>
<td>0.0</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>\textit{Acp158}</td>
<td>0.0</td>
<td>0.20</td>
</tr>
<tr>
<td></td>
<td>1.4</td>
<td></td>
</tr>
<tr>
<td>\textit{Gene144}</td>
<td>8e66</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>0.058</td>
<td></td>
</tr>
<tr>
<td>\textit{Acp157a}</td>
<td>e169</td>
<td>0.056</td>
</tr>
<tr>
<td></td>
<td>0.22</td>
<td></td>
</tr>
</tbody>
</table>

\textit{Acp134} from \textit{D. yakuba} and \textit{D. ananassae} is presented in supplemental data at http://www.genetics.org/supplemental/. However, the quality of this alignment leads us to consider the status of the gene in \textit{D. ananassae} as ambiguous.

\textit{Acp223} codes for a predicted protein of 116 residues. It is located between the \textit{D. yakuba} orthologs of \textit{Obp56f} and \textit{Obp56e}. Indeed, the organization of the three genes is similar, which together with their physical location, suggests that they are paralogous. \textit{D. erecta} also has a copy of \textit{Acp223}. \textit{D. yakuba} \textit{Acp223} is more highly diverged from the \textit{D. yakuba} \textit{Obp56e} and \textit{Obp56f} genes than these genes are from one another. A partial, homologous \textit{D. melanogaster} ORF appears to be present; however, it codes for a predicted protein of only 44 residues, which leaves it with questionable status in \textit{D. melanogaster} (Supplemental Data B at http://www.genetics.org/supplemental/). A syntenic alignment of the putative \textit{D. ananassae} orthologous region with \textit{D. yakuba} provides no evidence for a \textit{D. ananassae} copy of \textit{Acp223}.

\textit{Acp224} codes for a predicted protein of 231 residues in \textit{D. yakuba} and is located within an intron of \textit{CG31757}. An alignment of the orthologous region from \textit{D. erecta} reveals that the reading frame starting with the \textit{D. yakuba} initiation codon codes for a predicted protein of 75 residues. However, the fact that the \textit{D. yakuba} gene and the putative \textit{D. erecta} ortholog are extremely divergent in terms of length and sequence casts some doubt on the status of the \textit{D. erecta} gene. To address this uncertainty, we used RACE on accessory gland cDNA to isolate the ends of the \textit{D. erecta} gene. The RACE results revealed that there is an apparently orthologous \textit{D. erecta} transcript, which codes for two potential ORFs (89 codons

<table>
<thead>
<tr>
<th>Gene</th>
<th>(\pi_s)</th>
<th>(\pi_A)</th>
<th>(K_s)</th>
<th>(K_A)</th>
<th>Fixed</th>
<th>Polymorphic</th>
<th>G-test, P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>\textit{Acp134} (n = 9)</td>
<td>0.087</td>
<td>0.061</td>
<td>0.137</td>
<td>0.174</td>
<td>Silent</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replacement</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>\textit{Acp157a} (n = 7)</td>
<td>0.025</td>
<td>0.009</td>
<td>0.300</td>
<td>0.370</td>
<td>Silent</td>
<td>19</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replacement</td>
<td>62</td>
<td>6</td>
</tr>
<tr>
<td>\textit{Acp158} (n = 7)</td>
<td>0.052</td>
<td>0.002</td>
<td>0.138</td>
<td>0.076</td>
<td>Silent</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replacement</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>\textit{Acp223} (n = 10)</td>
<td>0.009</td>
<td>0.003</td>
<td>0.224</td>
<td>0.089</td>
<td>Silent</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replacement</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>\textit{Acp225} (n = 9)</td>
<td>0.028</td>
<td>0.043</td>
<td>0.135</td>
<td>0.044</td>
<td>Silent</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Replacement</td>
<td>9</td>
<td>7</td>
</tr>
</tbody>
</table>

\(n\) is the number of \textit{D. yakuba} alleles sampled. For \textit{D. teissieri}, \(n = 1\) for all loci. Genes are on chromosome arm 2R with the exception of \textit{Acp225}, which is on 3R. Divergence estimates are Jukes–Cantor corrected.
and the aforementioned 75 codons) that share the same reading frame (but different initiation codons). The shorter ORF has a more strongly predicted signal sequence, which suggests that it is the more likely candidate. Acp224 is the only putative Acp from our study that has a recognizable functional domain based on an NCBI conserved domain search (Marchler-Bauer et al. 2003). The D. yakuba copy has three predicted Kazal-type serpin domains, while the D. erecta copy has one such predicted domain. Serpin domains have previously been observed in Drosophila Acp’s (Swanson et al. 2001; Mueller et al. 2004). Syntenic alignments of D. yakuba Acp224 region vs. D. melanogaster and D. ananassae (Supplemental Data B at http://www.genetics.org/supplemental/) strongly suggest that the gene is absent from these species. Thus, Acp224 is likely a very rapidly evolving D. yakuba/D. erecta-lineage gene.

Acp158 codes for a predicted protein of 71 residues. Syntenic alignments of orthologous regions in D. melanogaster and D. erecta provide no evidence of an orthologous gene in these species. This gene is located within an intron of Phe53E. Another putative Acp, Acp133, which is likely shared in D. melanogaster, D. yakuba, and D. erecta, is located ~1.2 kb 5′ of Acp158 in D. yakuba, also in a Phe53E intron. Acp133 and Acp158 code for proteins of roughly equal length (62 and 71 residues, respectively) and both are composed of two small exons and one small intron. These similarities, along with their physical proximity, suggest the possibility that the two genes are related by duplication. However, their predicted protein sequences are too highly diverged to provide strong evidence of homology. The data are consistent with the idea that Acp158 is a highly diverged duplication of Acp133 that is present only in D. yakuba. This implies either that Acp158 is a recent duplication that has diverged incredibly rapidly or that Acp158 is an old duplication that has been lost multiple times in the melanogaster subgroup. Alternatively, it is possible that these two genes are not paralogous. The alignment of the D. yakuba Acp158 region with the putative orthologous region of D. ananassae suggests that neither it nor Acp133 is present in this species, although some uncertainty regarding the alignment means that this conclusion should be considered provisional.

Gene144 has a single exon. The protein-coding potential of this gene is unclear. Transcript data from our original cDNA clone and RACE experiments suggest the possibility of three open reading frames, two of which start with methionine and code for predicted proteins of 14 residues and one of which starts with isoleucine and codes for a predicted protein of 39 residues (which is not predicted to have a signal sequence). None of the three open reading frames is conserved in D. melanogaster, although there is apparently orthologous genomic sequence. This is likely not an Acp, and may not be a protein-coding gene (e.g., Tupy et al. 2005). However, the fact that we isolated this putative transcript twice (cDNA clone and RACE), along with the absence of a genomic poly(A) sequence downstream of the transcript, suggests that it is not the result of genomic contamination. We unsuccessfully attempted to amplify the homologous region by RT–PCR using RNA isolated from whole D. melanogaster males. This failure is consistent with the idea that this gene is not present in each of the melanogaster subgroup species.

Acp157a codes for a 112-residue-long predicted protein. An alignment of the D. yakuba Acp157a region to orthologous regions of the D. erecta and D. melanogaster genomes shows that D. erecta contains an ortholog, while D. melanogaster does not. A similar alignment to the putative orthologous region of the D. ananassae assembly strongly suggests that the gene is not in this species. Thus, Acp157a is likely a D. yakuba/D. erecta-specific gene. D. yakuba, but not other species, harbors a nearby, recent duplication (~4 kb 5′) of Acp157a. However, this duplication has no long open reading frame, suggesting that it is a D. yakuba-specific pseudogene.

Putative lineage-restricted genes identified from D. erecta accessory gland ESTs: Acp100 codes for a predicted protein of 190 residues. A potential highly diverged D. yakuba ortholog is present. This D. yakuba gene shares the putative D. erecta initiation codon, but with a predicted length of 263 residues, is significantly longer than the predicted D. erecta protein. Both species share a canonical polyadenylation signal downstream of their putative stop codons. A syntenic alignment between D. erecta and D. melanogaster suggests that the gene is absent from the latter. We were unable to generate a convincing syntenic alignment with D. ananassae.

Gene 37 codes for a predicted protein of 80 residues. This protein does not have a predicted signal sequence, casting some doubt on its status as an Acp. Syntenic alignments to D. yakuba, D. melanogaster, and D. ananassae suggest that this gene is D. erecta specific. We computationally discovered a second putative open reading frame (single exon, 210 residues) that is 3′ of gene 37 and coded on the opposite strand (the putative CDS is annotated by left-facing arrows in the supplemental data alignment at http://www.genetics.org/supplemental/). This second putative gene, which contains a strongly predicted signal sequence and a predicted fibrinogen domain, overlaps gene 37 (their putative 3′-ends overlap). The best hit in a BLASTp analysis of this second gene to D. melanogaster proteins is to CG30281 (6e36, 40% identity). CG30281 is associated with the gene ontology terms “receptor binding” and “defense response.” It appears to be D. erecta specific. However, we were unable to generate a D. erecta RT–PCR product, which casts doubt on its status.

Population genetics of lineage-restricted Acp’s: We collected polymorphism and divergence data from several D. yakuba/D. erecta-specific putative Acp’s to investigate mechanisms of protein evolution between D. yakuba and D. teissieri (Table 2). The data, pooled across genes, reject the null (neutral) model (Kimura 1983) in the direction of adaptive protein divergence (McDonald
and Kreitman 1991); however, only one gene, Acp158, is individually significant. Removing the data from Acp158 yields a nonsignificant test on data from the remaining genes ($P = 0.17$). Thus, although the rates of protein divergence reported here are high compared to most Drosophila genes (e.g., Begun 2002; Richards et al. 2005), there is no strong support for recent, recurrent directional selection on these genes overall.

**DISCUSSION**

We discovered several genes, many of which are likely Acp's, that have a lineage-restricted distribution in the melanogaster subgroup. Each lineage-restricted gene described here could be explained in two ways: (i) as a novel gene gained in D. yakuba, D. erecta, or their common ancestor or (ii) as multiple losses of a gene. One’s intuition is that gains of novel genetic functions are much less likely than losses. The problem with this formulation is that it raises the question, How many losses must one invoke before entertaining the hypothesis of gene gain as equally (or more) parsimonious? Regardless of the conclusion for any particular Acp, it seems unreasonable to repeatedly invoke multiple losses and disallow occasional gains, as this would imply that ancestral seminal fluid function is being lost from Drosophila, which seems unlikely. Thus, we favor the interpretation that some of the orphan genes described here are newly evolved.

What are plausible mechanisms for the origin of novel Acp's? One possibility is duplication and divergence (Holloway and Begun 2004; Mueller et al. 2005). For example, Acp158, which appears to be present only in D. yakuba, may be a highly diverged duplicate of Acp133, which is present in D. melanogaster, D. yakuba, and D. erecta. However, most of our orphans cannot be explained this way (Table 2), as BLAST results support the idea that they are unique. This is consistent with previous analyses of the D. melanogaster genome suggesting the presence of few recent Acp duplications (Holloway and Begun 2004; Mueller et al. 2005). An alternative possibility is that novel genetic functions can be co-opted from previously noncoding sequence. Such phenomena have been observed before. For example, the recently evolved D. melanogaster gene, Sdic, is partially derived from an intron of a cytoplasmic dynein gene (Nurminsky et al. 1998). In nototheneoid fishes, intronic sequence from an ancestral trypsinogen gene has been co-opted into protein-coding function in a descendant antifreeze protein (Chen et al. 1997). Such examples support the plausibility of the recruitment of ancestral noncoding sequence into coding function. For the genes described here, however, there is neither evidence for partial derivation from ancestral protein-coding sequence nor evidence of association with transposable elements or other repetitive sequences.

These observations raise the question of the plausibility of the birth of novel Acp's entirely from small open reading frames present in ancestrally noncoding sequence. Acp's have several features that make this suggestion worth considering. First, they tend to have short open reading frames, of which there are huge numbers in noncoding genomic sequence. Second, as secreted proteins, a signal sequence is the primary functional element. Although signal sequences tend to be hydrophobic and α-helical (Doudna and Battey 2004), the amino acid sequences are not always highly conserved (Nielsen et al. 1997). Third, Acp’s frequently have no known functional domains apart from their signal sequences (Swanson et al. 2001; Mueller et al. 2005; Wagstaff and Begun 2005b), which is consistent with the potential for a large degree of functional and evolutionary lability. Finally, seminal fluid function may be under stronger or more frequent directional selection than many other biological functions, which may make it more likely for novel Acp's to invade populations.

Unannotated portions of eukaryotic genomes (and, indeed, random DNA sequences) contain many short (e.g., 30–100 residues) open reading frames. A fraction of new mutations, most of which are likely deleterious (Hahn et al. 2003), may create promoters near such ORFs, thereby driving their expression, even if at a low level. Moreover, the consensus, highly conserved animal polyadenylation signal AATAAA (Zhao et al. 1999) is short, simple, and, therefore, common. Thus, at mutation-selection balance there is likely a large pool of small open reading frames (many of which possess signal sequences) that are a short mutational distance from deleterious expression and translation. Occasionally, however, a “spuriously” expressed ORF coding for a small, secreted peptide could be recruited into a novel function by natural selection.

To investigate the plausibility of this scenario, we carried out an analysis of the signal peptide-coding potential of the intergenic and intronic portions of the D. melanogaster reference sequence. We found that Repeat-Masked D. melanogaster intergenic sequence harbors 174,779 open reading frames of ≥ 40 residues. Of these, we conservatively estimate that ~6071 (3.5%) have a strongly predicted signal sequence (SignalP, hidden Markov model $P > 0.95$ and positive neutral network prediction). The corresponding numbers for introns are 53,003 ORFs and 1963 strongly predicted signal sequences (3.7%). Although a small fraction of these ORFs may be previously undescribed genes or exons, it seems more likely that we should conclude that the coding potential for novel, small, secreted peptides in Drosophila noncoding DNA is impressively large. Recent reports that a surprisingly high fraction of eukaryotic genomes is transcribed (Bertone et al. 2004; Stolec et al. 2004, 2005) would favor the mutation-selection-recruitment model for the origin of small peptides. Direct support for this model could be best obtained through the discovery of small, novel, polymorphic proteins in populations.

It seems clear that Acp's are much more likely than most other genes to have lineage-restricted distributions.
The proximate and ultimate explanations for this pattern are unclear, although, in principle, the small size of Acp’s and the fact that they may be under unusually strong directional selection may contribute to a rapid gain of seminal fluid proteins. Comparative functional analysis of Acp’s, including the lineage-restricted genes described here, could greatly illuminate their evolutionary explanation.

M. Levine, S. Schaeffer, and two anonymous reviewers provided useful comments. This work was supported by National Science Foundation grant DEB-0327049 and National Institutes of Health grant GM071926.

LITERATURE CITED


Communicating editor: S. W. Schaeffer

Novel Drosophila Acp’s

1681