The Evolutionary Analysis of "Orphans" From the Drosophila Genome Identifies Rapidly Diverging and Incorrectly Annotated Genes

Karl J. Schmid and Charles F. Aquadro

Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853

Manuscript received November 14, 2000
Accepted for publication July 6, 2001

ABSTRACT

In genome projects of eukaryotic model organisms, a large number of novel genes of unknown function and evolutionary history ("orphans") are being identified. Since many orphans have no known homologs in distant species, it is unclear whether they are restricted to certain taxa or evolve rapidly, either because of a lack of constraints or positive Darwinian selection. Here we use three criteria for the selection of putatively rapidly evolving genes from a single sequence of *Drosophila melanogaster*. Thirteen candidate genes were chosen from the *Adh* region on the second chromosome and 1 from the tip of the X chromosome. We succeeded in obtaining sequence from 6 of these in the closely related species *D. simulans* and *D. yakuba*. Only 1 of the 6 genes showed a large number of amino acid replacements and in-frame insertions/deletions. A population survey of this gene suggests that its rapid evolution is due to the fixation of many neutral or nearly neutral mutations. Two other genes showed "normal" levels of divergence between species. Four genes had insertions/deletions that destroy the putative reading frame within exons, suggesting that these exons have been incorrectly annotated. The evolutionary analysis of orphan genes in closely related species is useful for the identification of both rapidly evolving and incorrectly annotated genes.

GENOME projects aim at correctly identifying all genes encoded by a genome (*e.g.*, Bork and Koonin 1998; Brenner 1999; Adams *et al.* 2000) and understanding their genetic, biochemical, and cellular functions (Hieter and Boguski 1997; Bork *et al.* 1998). Achieving these goals is a considerable challenge because all genomes studied so far harbor many proteins with no or little similarity to proteins of known function. A comparison of publications describing partial or complete genome sequences from eukaroytic model organisms over the past 5 years reveals that about one-third of all predicted protein-coding genes fall into this class, despite the exponential growth of sequence databases. Such genes have been called "orphans" and their function needs to be determined by genetic or biochemical approaches (Oliver 1996).

There are two major explanations for the large number of orphans. Both need to take into account that most model organisms whose genomes are currently being sequenced are separated by large evolutionary distances. First, many orphans might consist of genes whose phylogenetic distribution is restricted to certain evolutionary lineages; *e.g.*, they are specific to plants or vertebrates. Second, orphan genes may diverge rapidly between closely related species because the proteins they encode are unconstrained in their sequence evolution or subjected to directed Darwinian selection, whereas their structure and function might be conserved even between distantly related organisms. Such a hypothesis is supported by estimates of only a few thousand naturally occurring protein superfamilies (Chothia 1992; Brenner *et al.* 1997). Orphans might therefore consist of highly divergent, rapidly evolving members of this limited set of superfamilies.

Evolutionary comparisons of closely related genomes will help to differentiate between the two hypotheses. For example, by a hybridization and sequencing approach it was estimated that about one-third of all expressed Drosophila genes diverge rapidly within the genus Drosophila (Schmid and Tautz 1997). These data support the rapid evolution hypothesis for a large number of orphan genes. Surveys of nucleotide polymorphism of some of these rapidly diverging orphan genes in populations of *D. melanogaster* and *D. simulans* revealed that a lack of constraints may be responsible for their evolution because the majority of the numerous amino acid substitutions are neutral or nearly neutral (Schmid *et al.* 1999).

There appears to be a relationship between the function and evolutionary conservation of genes. For example, the genetic and sequence analysis of 3 Mb of the *Adh* region of *D. melanogaster* revealed strong functional differences between conserved and nonconserved genes (Ashburner *et al.* 1999). Sequence analysis predicted 220 protein-coding genes, of which only 79 had a detectable phenotype (lethality, sterility, or morphological deformations). The lack of sequence conservation was very different between the two classes of genes: About two-

Corresponding author: Karl Schmid, Department of Genetics and Evolution, Max Planck Institute for Chemical Ecology, Carl Zeiss Promenade 10, D-07745 Jena, Germany. E-mail: schmid@ice.mpg.de

thirds of the 79 genes with a phenotype had a homolog in distantly related species (yeast, vertebrates, *C. elegans*, and prokaryotes) in contrast to only 14% of the genes without a phenotype. Clearly, the latter class is less conserved and both its function and evolution remain largely obscure. Additionally, genes with a mutant phenotype are more highly expressed as evaluated by comparisons to >80,000 expressed sequence tags (ESTs) from Drosophila (Ashburner *et al.* 1999).

The goals of this study were to test whether rapidly evolving candidate genes can be reliably identified in single genomic sequences, to verify by comparative sequencing that candidate genes evolve rapidly, and to distinguish between low constraint and positive Darwinian selection as causes for the sequence divergence of the proteins encoded by such genes. By combining data on the long-term evolutionary conservation in distant species and matches to Drosophila ESTs with sequence features like codon usage, we found 13 rapidly evolving candidate genes from the *Adh* region on the second chromosome (Ashburner *et al.* 1999) and the tip of the X chromosome (Benos *et al.* 2000). Homologs of 6 genes were sequenced from the closely related species *D. simulans* and *D. yakuba*. We discovered 1 very rapidly evolving and several incorrectly annotated genes.

MATERIALS AND METHODS

Analysis of annotated Drosophila sequences: Sequences from the European Drosophila Genome Project (EDGP) were downloaded from the EGDP FTP site (ftp.ebi.ac.uk) and coding sequences were extracted using the annotation in the GenBank format. The 3-Mb region of the *Adh* region and a gff-formatted file containing the annotation information were downloaded from the Berkeley Drosophila Genome Project (BDGP) website (http://www.fruitfly.org) and the coding sequences were extracted. The coding sequences were searched against the collection of 86,000 Drosophila ESTs and the nonredundant GenBank database at the National Center for Biotechnology Information using BLAST with standard settings (Altschul *et al.* 1997). The effective number of codons (ENC) and the GC content at silent sites were calculated according to Wright (1990). Sequence extractions, database searches, and analyses were performed with Perl scripts written by K. J. Schmid.

PCR and sequencing: Primers were designed with the Primer3 program (Rozen and Skaletsky 1998). Polymerase chain reactions were carried out using standard conditions (*e.g.*, Schmid *et al.* 1999). The primer sequences can be found in the supplementary information on our website (address below). PCR products were sequenced on an ABI377 automated sequencer using BIG-DYE Terminator chemistry and both the PCR and internal primers. Base-calling, sequence assembly, and sequence alignment were performed with *Phred* (Ewing *et al.* 1998), *Phrap* (P. Green, unpublished data), and *Consed* (Gordon *et al.* 1998).

Lines: The lines from *D. melanogaster* and *D. simulans* used for the population survey were collected in Harare, Zimbabwe and established as inbred isofemale lines. The DNA from these lines was prepared by standard protocols and purified with CsCl centrifugation. *D. yakuba* and *D. erecta* lines were obtained from the Drosophila Species Stock Center at Bowling Green. DNA was isolated from ~50 flies each using phenol/chloroform extraction and phenol precipitation.

Sequence analysis: The values of d_n and d_s were estimated with the maximum-likelihood method of Yang and Nielsen (1998) using the $F3 \times 4$ model (Yang 1999). Homologous sequences from *D. melanogaster* and *D. simulans* were downloaded from GenBank, and the coding sequences were extracted and semiautomatically aligned, using ClustalW (Thompson *et al.* 1994) and Perl scripts. DnaSP3.0 (Rozas and Rozas 1999) was used to calculate two estimates of nucleotide diversity, the average number of pairwise differences, π (Nei 1987), and an estimate of the mutation parameter $4N_e\mu$, θ (Watterson 1975), and to perform tests of neutrality. The following tests were used: Tajima's *D* (Tajima 1989), Fu and Li's *D* with an outgroup (Fu and Li 1993), the HKA test (Hudson *et al.* 1987), and the MK test (McDonald and Kreitman 1991). Further details about the tests can be found in the references.

Supplementary information: Sequences were submitted to GenBank under accession nos. AF264913–AF 264947. Further information is available from our website at http://www.mbg.cornell.edu/aquadro/sequences.html.

RESULTS

Identification of candidate genes: We analyzed genes from the annotated genome sequences located on the tip of the X chromosome from the EDGP (Benos *et al.* 2000) and the annotated 3-Mb *Adh* region on the second chromosome (Ashburner *et al.* 1999). The three criteria for the selection of putative rapidly evolving genes were (i) no or little (<25%) sequence identity to genes from distant organisms, (ii) no or few matches to ESTs, and (iii) a low codon bias. The use of codon bias as an indicator of rapid amino acid sequence evolution was based on the following rationale: An analysis of codon usage patterns in Drosophila genes revealed that amino acids encoded by unpreferred codons tend to be less conserved (Akashi 1996). This is probably due to a lack of selection of translational accuracy on functionally less important amino acid residues. Thus, proteins encoded by a large number of unpreferred codons should have many unconstrained amino acids and evolve rapidly. This hypothesis is confirmed by the codon usage patterns in rapidly evolving Drosophila genes (Schmid *et al.* 1999). Additionally, a comparison of proteins from several species known to evolve under strong Darwinian selection also revealed that many of them show little codon bias (K. J. Schmid, unpublished observation). It is important to note that codon usage is influenced by several factors (*e.g.*, expression level and length of coding sequence) and may not be strongly correlated with rapid evolution of the amino acid sequence. Finally, nucleotide composition and patterns of codon usage are important criteria for gene prediction algorithms like GENEFINDER and GENSCAN (Green 1995; Burge and Karlin 1997). Genes with unusual patterns of codon usage should have, on average, lower scores in the prediction and might be incorrectly annotated genes.

To identify genes with a low codon usage bias, the ENC (Wright 1990) was plotted against the GC content

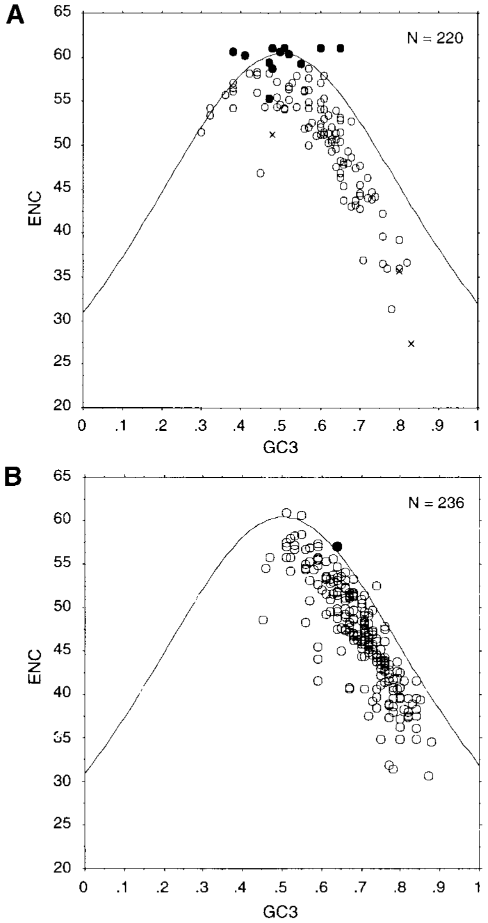

Figure 1.—Relationship between GC content at synonymous sites and the effective number of codons (ENC; Wright 1990) for all genes of the *Adh* region (A) and the tip of the X chromosome of *D. melanogaster* (B). Solid circles represent candidates for rapidly evolving genes. The × symbol shows the three "control" genes (see results). The line gives the expected relationship of GC content at synonymous sites and ENC of random sequences (Wright 1990). ENC values of 61 indicate indiscriminate use of synonymous codons.

at synonymous codon positions (*GC*3) for the 220 genes of the *Adh* region and 236 genes from the tip of the X chromosome (Figure 1). We also compared the codon usage of predicted genes with their GENEFINDER and GENSCAN scores as obtained from Ashburner *et al.* (1999), but they did not reveal a simple relationship (results not shown). Genes with very low and very high ENC values tend to have lower GENEFINDER or GENSCAN scores. Preferred codons in *D. melanogaster* end in C or G (Akashi 1995), and genes under selection for optimal codon usage should have a *GC*3 > 0.5. We selected genes with ENC > 55 and/or *GC*3 < 0.5 and no or weak similarity to genes from distant species (Table 1). Our sample also included several biased and conserved genes as controls. For example, to evaluate the relationship between codon usage and amino acid evolution, two members of a gene cluster encoding hypothetical metalloproteases were compared. *BACR44L22.4* has an ENC value of 60.6 and is the most divergent member in pairwise comparisons of the six paralogs (Ashburner *et al.* 1999). The codon usage of *BACR44L22.3* is more biased (ENC = 51.3) and it is the most conserved paralog in the cluster. Two additional "controls" were the highly biased genes *DS01068.5* (ENC = 35.6) and *DS00810.3* (ENC = 27.0), which are not conserved outside insects.

Sequence comparisons: To obtain homologs from *D. simulans* and *D. yakuba*, primers were designed from the *D. melanogaster* sequence using GC-rich regions in or around exons. Among 16 primer pairs tested, 10 resulted in a PCR product in *D. simulans* and 6 in *D. yakuba* (Table 1). We expected that only a subset of primers would work in the other species, because the average divergence at silent sites is ~11% between *D. melanogaster* and *D. simulans* (Bauer and Aquadro 1997; Powell and Moriyama 1997) and 23% between *D. melanogaster* and *D. yakuba* (Schmid and Tautz 1997). Out of 10 PCR products obtained from *D. simulans*, 7 could be sequenced successfully, and 5 could be sequenced from *D. yakuba.* Only one of the three high codon bias control genes could be amplified and sequenced successfully (*BACR44L22.3*) in both *D. simulans* and *D. yakuba.*

An alignment of the sequences revealed many nucleotide substitutions and insertions/deletions (indels). Among the six genes with low codon bias, the putative coding region of four genes showed out-of-frame indels in a comparison between *D. melanogaster* and *D. simulans*, *D. yakuba*, or *D. erecta.* These genes include *DS01514.3*, *DS03192.3*, and exon 1 of *DS07721.6*, which are probably incorrectly annotated exons. In two genes, we observed several in-frame indels (*DS06283.4* and exon 3 of *DS07721.6*). In the comparison between *D. melanogaster* and *D. erecta* homologs of *EG0007.10*, an out-of-frame indel in the 3′ region of the coding sequence leads to a longer protein in *D. erecta.* It is unclear whether this gene encodes a functional protein. Estimates of d_n and d_s are given in Table 1. Only *DS07721.6* can be considered to be a rapidly evolving protein (d_n = 0.0494) between *D. melanogaster* and *D. simulans*, whereas all other genes are more conserved and exhibit d_n values that are not significantly different from the control gene (*BACR44L22.3*).

DS07721.6 is predicted to encode a large protein of 1585 amino acids of unknown function (Ashburner *et al.* 1999). Secondary structure analysis of the protein sequence suggests that it is a transmembrane protein (data not shown). Because of its length, we focused our sequencing on the extracellular domain (Figure 2). The nonsynonymous divergence of *DS07721.6* is similar to *anon1G5*, the most rapidly evolving gene from an earlier screen for rapidly evolving genes (Schmid and Tautz 1997), which also exhibits in-frame indels in comparisons between *D. melanogaster*, *D. simulans*, and *D. yakuba.* Although the first exon of *DS07721.6* contains two out-of-frame indels, we consider it to be a functional gene

TABLE 1

List of candidate genes chosen for sequencing in *D. simulans* and *D. yakuba*

						Codons sequenced		Nucleotide divergence[b]				
								MEL-SIM		MEL-YAK		
Gene	Codons	ENC	gf/gs scores[a]	BLAST hit	EST match	SIM	YAK	d_n	d_s	d_n	d_s	Indel[c]
				Candidate genes for rapid evolution								
DS07660.1	454	61.0	—, 101	Phosphate cotransporter	None	0	0					
DS06238.4	219	61.0	30, 63	Cuticular protein	None	206	213	0.0103	0.1610	0.0129	0.2967	IF
DS03192.4	46	61.0	—, —	No match	Embryo	0	0					
DS03192.3	57	61.0	—, —	No match	Head	39	39	(0.0438)	(0.0484)	(0.1960)	(0.3945)	OOF
BACR44L22.4	241	60.6	—, —	Zn^{2+} metalloprotease	None	240	0	0.0260	0.1538			
DS01514.3	396	60.6	—, 49	No match	None	83	0	(0.0280)	(0.0160)			OOF
Mst35Ba	147	60.4	—, —	No match	Head	0	0					
DS07108.5	295	60.2	—, 25	Antibacterial protease	None	0	0					
DS07721.6	1585	59.4	—, 271	REJ-domain proteins	None	1004	455	0.0494	0.1374	0.1687	0.4302	OOF[d], IF[e]
DS04095.1	303	59.3	—, 64	No match	Yes	0	0					
DS04095.2	191	58.8	—, 53	No match	None	0	0					
EG0007.10	169	57.0	—, 26	No match	Embryo	168	168	0.0119	0.0895	0.0512	0.3043	OOF[f]
Mst35Bb	147	55.4	—, —	No match	Head	0	0					
				Control genes								
BACR44L22.3	254	51.3	—, 53	Zn^{2+} metalloprotrease	None	241	241	0.0119	0.0895	0.0321	0.2586	
DS01068.5	160	35.6	35, 382	Immune-related protein	None	0	0					
DS00810.3	69	27.0	16, 14	No match	Many	0	0					

EG0007.10 is X-linked; all other genes are located on the second chromosome. MEL, *D. melanogaster*; SIM, *D. simulans*; YAK, *D. yakuba*.

[a] gf, GENEFINDER score; gs, GENSCAN score.

[b] Values in parentheses are for the longest conserved reading frame in coding sequences with out-of-frame indels.

[c] OOF, out-of-frame indel; IF, in-frame indel.

[d] Exon 1 of annotated sequence; probably an incorrectly annotated exon of an otherwise functional protein.

[e] Exon 3 of annotated sequence.

[f] Extending 3′ end of coding sequence in *D. erecta*.

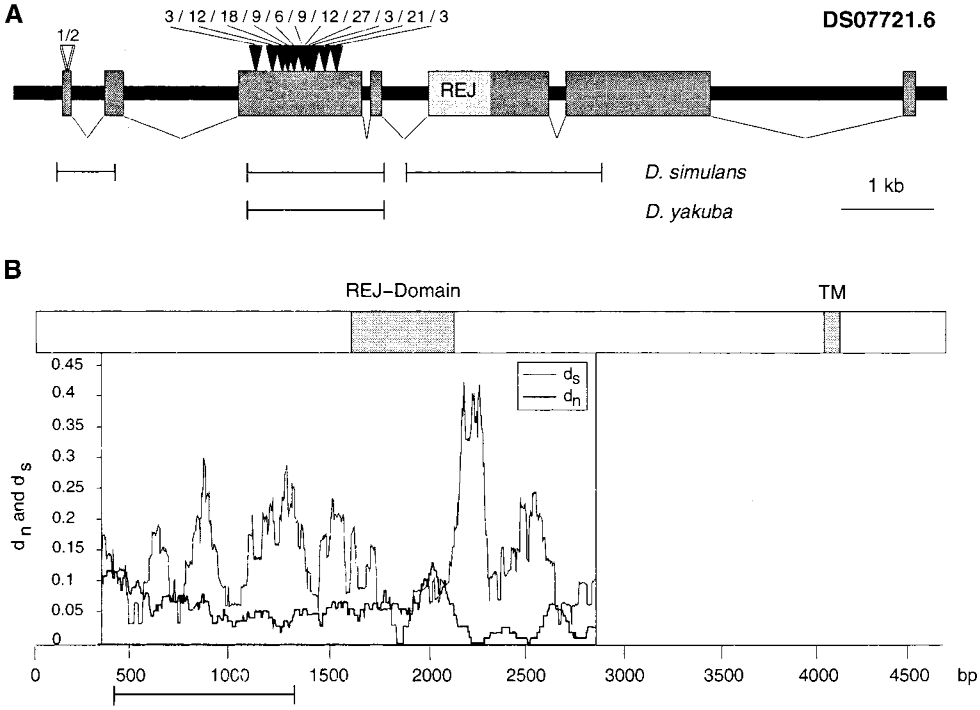

Figure 2.—(A) Schematic structure of gene *DS07721.6.* Shaded boxes designate exons, solid arrowheads show in-frame indels (multiples of three), and open arrowheads show out-of-frame indels (no multiples of three). Numbers above the arrowheads are the lengths of insertions in base pairs. Bars show the regions sequenced in *D. simulans* or *D. yakuba.* (B) Schematic structure of the predicted protein sequence of *DS07721.6.* The locations of the moderately conserved REJ module and the single transmembrane helix (TM) are shown as shaded boxes. The graph shows a sliding window analysis of the d_n and d_s values in the *D. melanogaster-D. simulans* comparison using a window size of 90 codons and a step size of one. The sliding window analysis of d_n and d_s was performed with the program *wina* (Endo *et al.* 1996). The bar shows the region surveyed in populations of *D. melanogaster* and *D. simulans.*

because the numerous indels in exon 3 are in frame and there is a weak but significant sequence similarity to REJ-domain-containing proteins from other animal phyla (data not shown). The rapid evolution of parts of *DS07721.6* raises the question of whether this is due to a high rate of neutral evolution or positive Darwinian selection. A sliding window analysis of the d_n and d_s values along the coding sequence of exons 3–6 shows that the nonsynonymous sequence divergence is relatively constant whereas the synonymous sequence divergence is highly variable between different regions of the coding sequence (Figure 2B). Interestingly, in the fragment encoding the region C-terminal of the REJ module, d_n drops to zero and d_s increases up to 0.4, which is much higher than the expected neutral sequence divergence. This fragment may consist of a mutational hotspot combined with strong constraints on nonsynonymous substitutions.

DNA polymorphism in *DS07721.6*: Because of the high rate of amino acid evolution and silent divergence, we obtained sequences of exons 3 and 4 from 10 lines each of African populations of *D. melanogaster* and *D. simulans.* A comparison of intraspecific polymorphism and interspecific divergence can be used to discriminate between neutral evolution and positive Darwinian selection as the causes for the rapid evolution of these genes. We chose the African lines because they represent ancestral populations of both species and are probably close to a mutation-selection-drift equilibrium (Begun and Aquadro 1993).

Nucleotide diversity is low in the 858 bases in *D. melanogaster* (π = 0.0018; Table 2). Only 6 polymorphisms were discovered; 3 of them are synonymous and 3 nonsynonymous. In *D. simulans,* 27 polymorphisms are segregating in the sample (π = 0.0124), of which 10 are synonymous and 17 nonsynonymous. The level of DNA diversity (π) is about seven times higher than in *D. melanogaster,* which is within the range observed for other genes that were surveyed in both species (Moriyama and Powell 1996). The large number of replacement polymorphisms is consistent with the rapid evolution of this region of *DS07721.6.* Most other surveyed genes have much smaller numbers of nonsynonymous polymorphisms (Moriyama and Powell 1996). Several tests of neutrality were applied to the data and none of them rejected the null hypothesis of neutral evolution (Tables 2 and 3). The HKA test was not significant in comparisons of *DS07721.6* to various neutrally evolving reference loci (*anon1A3*, *anon1E9*, and *anon1G5*; Schmid *et al.* 1999), although it was marginally significant in *D. melanogaster* when the *Adh*-5′ region of Kreitman and Hudson (1991) was used for comparison (χ^2 = 3.824, P = 0.0505).

We also looked at lineage-specific substitutions to analyze the effect of different species-level effective population sizes on the evolution and polymorphism of this region (see Schmid *et al.* 1999 for a more detailed discussion). Using *D. yakuba* as an outgroup, 37 out of 41 fixed differences could be assigned to either the *D. melanogaster* or *D. simulans* lineages. Thirteen nonsynonymous and 6 synonymous substitutions occurred in the *D. melanogaster* lineage and 14 nonsynonymous and 4

TABLE 2

Nucleotide diversity in exon 4 of *DS07721.6* in *D. melanogaster* and *D. simulans*

		Polymorphic sites								Lineage-specific fixed differences	
Species	Sites	Total	Replacement	Silent	π	θ	Tajima's *D*	Fu and Li's *D*	HKA (χ^2)[a]	Replacement	Silent
D. melanogaster	858	6	3	3	0.0018	0.0025	−1.15	−1.27	1.47	13	6
D. simulans	858	27	17	10	0.0124	0.0112	0.48	0.31	0.15	14	4

None of the tests for neutrality is significant ($P < 0.05$).

[a] The *anon1A3* gene (Schmid *et al.* 1999) was used as a reference locus.

TABLE 3

McDonald-Kreitman test for *DS07721.6*

	Substitutions		
Class	Replacement	Synonymous	*G*-value
Fixed between species	38	19	0.326
Polymorphic within species	20	13	

The *G*-value is not significant ($P > 0.05$).

synonymous substitutions in the *D. simulans* lineage. The numbers for the two types of substitutions differ little between the two lineages. This suggests that the nonsynonymous substitutions are either completely neutral or have been fixed by relatively strong positive selection that occurred in both lineages.

DISCUSSION

Identifying rapidly evolving genes: Our motivation for this study was to test whether orphans in the Drosophila genome are rapidly evolving genes and, if so, whether they evolve neutrally because of relaxed constraints or positive selection. Rapidly evolving genes have recently attracted considerable interest, because they might play a role in the adaptive evolution of phenotypic traits (*e.g.*, Murphy 1993; Swanson and Vacquier 1995; Pamilo and O'Neill 1997; Civetta and Singh 1998; Michaelmore and Meyers 1998; Duda and Palumbi 1999; Yokoyama *et al.* 1999; Wyckoff *et al.* 2000). An understanding of the evolution and function of such "adaptive trait loci" may be highly relevant to the study of species differences (Tautz and Schmid 1998). Thus, after rapidly evolving genes have been identified, it is of interest to test whether they diverge neutrally or are responding to positive Darwinian selection.

Since no extensive genomic sequence from a closely related species of *D. melanogaster* is available, we identified candidate genes by a synopsis of data on sequence features, function, and evolutionary conservation in distantly related organisms. The genes of the *Adh* region and the tip of the X chromosome region are good candidates for testing such an approach because they are among the best-characterized regions of the Drosophila genome and much information on sequence conservation, expression, and genetic function is available. Our criteria for selecting putative rapidly evolving genes were codon usage, a low level of expression, and no or weak similarity to distant organisms. Among four surveyed genes without codon bias that retained an intact reading frame in *D. simulans* or *D. yakuba*, only one (*DS07721.6*) was rapidly evolving at the amino acid level, suggesting that a lack of codon usage alone may not be a good indicator for the discovery of rapidly evolving

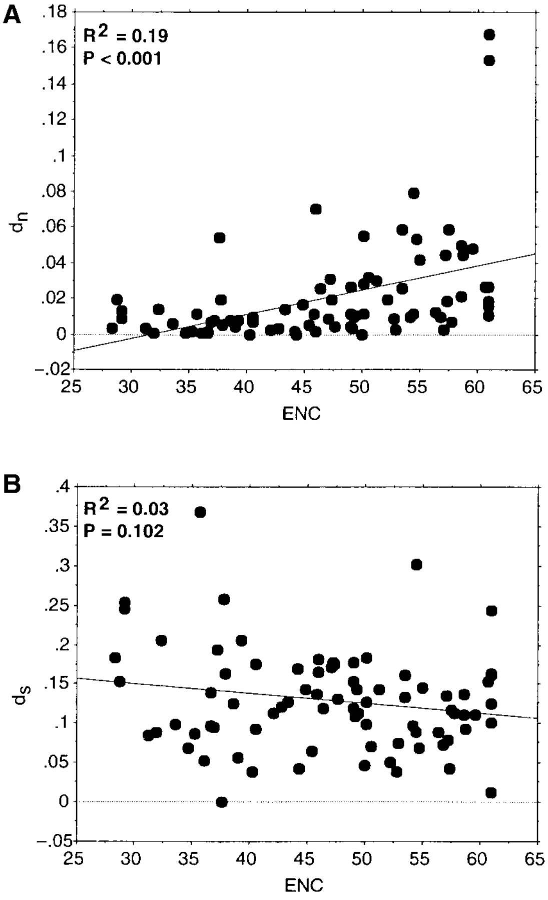

Figure 3.—Correlation between the effective number of codons, ENC (calculated from the *D. melanogaster* sequences), the number of nonsynonymous, d_n (A), and synonymous, d_s (B), substitutions calculated from alignments of homologous sequences from *D. melanogaster* and *D. simulans* (n = 85) that were retrieved from GenBank or sequenced in this study.

genes. This notion is further supported by a comparison of ENC with d_n for genes (including those of this study) where partial or complete coding sequences were available from *D. melanogaster* and *D. simulans* (n = 85). Although we find a highly significant negative correlation between codon usage bias and nonsynonymous divergence (Figure 3A), it is not very strong. This suggests that, although there is evidence for selection on translational accuracy, additional factors such as gene length (Comeron *et al.* 1999), expression level (Shields *et al.* 1988; Powell and Moriyama 1997; Duret and Mouchiroud 1999), mutation bias (Kliman and Hey 1994), and local rates of recombination (Kliman and Hey 1993) also influence codon usage patterns in the genome of Drosophila. These additional factors may blur the relationship between codon usage and nonsynonymous divergence. Furthermore, under selection for translational accuracy, a positive relationship between ENC and d_s is also expected, as has been found in several studies (*e.g.*, Sharp and Li 1989). In a more recent study, however, such a relationship was not obtained, and simulations suggested that such a relationship represents different assumptions in the estimation of nucleotide divergence (Dunn *et al.* 2001). Using the same maximum-likelihood method for estimating nucleotide divergence as Dunn *et al.* (2001) and with a larger number of genes, we also do not find a significant correlation between ENC and synonymous divergence in comparisons between *D. melanogaster* and *D. simulans* (Figure 3B). One explanation for the absence of such a correlation may be variable mutational pressures in different evolutionary lineages, which can lead to a negative correlation between ENC and d_s (Bielawski *et al.* 2000). In addition, our data do not show a positive correlation between d_n and d_s (R^2 = 0.02, P = 0.25), which is in contrast to earlier studies (Akashi 1994; Comeron and Kreitman 1998; Dunn *et al.* 2001). However, since we are mainly interested in the relationship between ENC and d_n, the lack of such a relationship has no consequences for our study. In conclusion, it can be stated that, although there is a positive correlation between d_n and ENC, the use of codon bias alone is not sufficient for a reliable identification of rapidly evolving genes.

Therefore, additional information about gene function needs to be taken into account for generating better predictions of rapidly evolving genes from single genome sequences. Such information can be the type and strength of mutant phenotypes (Ashburner *et al.* 1999), the tissue where genes are expressed (Hurst and Smith 1999), or the type of protein that is encoded by a gene (*e.g.*, subcellular location). For example, *DS06238.4* is probably identical to the *pupal* gene, which has a lethal phenotype. Under the assumption that functionally important genes should be more conserved (Wilson *et al.* 1977), rapid sequence divergence is not expected in this gene. In fact, its sequence is highly conserved in distant insects but not in other phyla, suggesting that its occurrence is restricted to insects, where it may have acquired an essential function. The lack of codon bias in this gene could be related to its repetitive amino acid sequence.

Causes of rapid evolution: Only one of the candidate genes we examined was apparently both functional and rapidly evolving. Neither the sequence comparisons between Drosophila species nor the population variation analysis of the rapidly evolving gene *DS07721.6* revealed evidence for positive selection being important for its evolution. The levels of nonsynonymous divergence and replacement polymorphisms are very similar to other rapidly evolving orphan genes (Schmid *et al.* 1999). These results together suggest that the primary sequence of numerous (correctly annotated) orphan genes may evolve relatively unconstrained at the amino acid level. Whereas the criteria we used are expected to be compatible with the identification of genes evolving under relaxed selective constraints, low levels of expression (indicated by the absence of EST matches) and low codon bias may not necessarily be a characteristic of genes evolving under positive Darwinian selection. However, one can expect that many genes evolving under positive selection have specialized functions with a restricted expression (Tautz and Schmid 1998) and therefore may not be represented in current EST collections. This notion is supported by a recent EST sequencing study of genes expressed in the testis of *D. melanogaster*, which found that about one-half of 1560 cDNA sets fail to align with existing Drosophila ESTs (Andrews *et al.* 2000). This suggests that many tissue-specific genes have not yet been discovered, although they may be expressed at a high level within a tissue. As EST collections grow in size, information about the number of tissues in which genes are expressed can be used to identify rapidly evolving genes. There is little theoretical support for the notion that genes evolving under positive selection can be expected to have low codon bias, but one can assume that translational accuracy may not be very strong in such genes. This hypothesis is consistent with the observation of several genes encoding male accessory gland proteins that evolve under positive Darwinian selection and are characterized by low codon bias (Begun *et al.* 2000).

It should be noted that most tests employed for detecting positive selection are not very powerful in detecting weak or episodic selection (for a more detailed discussion, see Schmid *et al.* 1999). More powerful tests need to be developed for detecting these types of adaptive molecular evolution. Generation of data for such genes from additional species may allow codon-specific models to be used such as those developed by Z. Yang and R. Nielsen (*e.g.*, Yang *et al.* 2000). Such a study of *DS07721.6* may reveal that positive selection at a subset of the amino acids, coupled with selective constraint at others, may account for its rapid evolution shown here.

Improving the annotation: A surprising result of our survey is the large proportion of incorrectly annotated genes. In four out of six candidate genes, the putative open reading frame contained out-of-frame indels in either *D. simulans*, *D. yakuba*, or *D. erecta*. Two of these sequences may not be protein-coding genes at all. Furthermore, there are at least two additional paralogs of gene *DS07721.6* in the Drosophila genome that were not recognized and annotated by the gene prediction algorithms used for the annotation (data not shown). These observations confirm the conclusions of the Drosophila Genome Annotation Assessment Project (GASP; Reese *et al.* 2000) that, even in the relatively compact Drosophila genome, purely computer-based gene annotations (*ab initio* predictions) both over- and underpredict genes. Many predictions contain errors (*e.g.*, the incorrect identification of the 5′ end of open reading frames), particularly for genes with a lack of sequence conservation or with unusual patterns of codon usage like the candidate genes of this study (Guigó *et al.* 2000). Gene predictions need additional experimental verification such as full-length cDNA sequencing, sequencing of ESTs from tissue-specific libraries (Andrews *et al.* 2000), or, as described in this study, sequencing of homologous genes from closely related species. It should be noted that our small sample of genes does not allow an estimation of how many predicted genes contain annotation errors. However, we expect that a substantial proportion of nonconserved genes may be overpredicted and that many genes not recognized by prediction algorithms may consist of rapidly evolving genes.

Comparative sequencing of related species: The fact that only one of six candidate genes evolves rapidly suggests that the identification of such genes in single genomic sequences is difficult, in particular because of the requirement of a correctly annotated sequence. In addition, the PCR approach used in this pilot study is not practical for analyzing a large number of candidate genes because about one-half of the primer pairs designed using the *D. melanogaster* sequence did not work in *D. simulans* or *D. yakuba*. However, because numerous rapidly evolving genes can be expected in the genome of *D. melanogaster* and other model organisms (Schmid and Tautz 1997), alternative approaches might be taken to identify such genes on a large scale. Possible approaches include the sequencing of the complete genome (at low coverage) or of ESTs from cDNA libraries of closely related "satellite" species. Suitable species for comparisons to *D. melanogaster* are *D. simulans* or *D. yakuba*. Values of d_s range from 0.05 to 0.18 between *D. melanogaster* and *D. simulans* (Bauer and Aquadro 1997; Powell and Moriyama 1997) and from 0.11 to 0.35 between *D. melanogaster* and *D. yakuba* (Schmid and Tautz 1997). d_n and d_s values from such comparisons give good estimates of sequence divergence and facilitate the genome-wide identification of rapidly evolving genes like *DS07721.6* that are candidates for positively selected genes. In addition, comparisons between *D. melanogaster* and *D. simulans* or *D. yakuba* are sufficiently divergent to detect incorrectly annotated exons because of the large number of point and indel mutations that are being fixed by chance in noncoding sequences. Such approaches would not only lead to the identification of rapidly evolving genes with potential roles in the phenotypic divergence of species and enhance our understanding of genome-wide patterns of protein evolution but also assist in the correct annotation of "difficult" genes for which currently available gene prediction methods are not reliable.

We are grateful to B. Haubold, W. Swanson, T. Wiehe, M. Aguadé, and two anonymous reviewers for helpful comments on the manuscript. This work was funded by a postdoctoral fellowship of the German Academic Exchange Service (DAAD) to K.J.S. and a National Institutes of Health (NIH) grant to C.F.A.

LITERATURE CITED

Adams, M., S. Celniker, R. Holt, C. Evans, J. Gocayne *et al.*, 2000 The genome sequence of *Drosophila melanogaster*. Science **287:** 2185–2195.

Akashi, H., 1994 Synonymous codon usage in *Drosophila melanogaster*: Natural selection and translational accuracy. Genetics **136:** 927–935.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics **139:** 1067–1076.

Akashi, H., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297–1307.

Altschul, S., T. Madden, A. Schäffer, J. Zhang, Z. Zhang *et al.*, 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Andrews, J., G. Bouffard, C. Cheadle, J. Lü, K. Becker *et al.*, 2000 Gene discovery using computational and microarray analysis of transcription in the *Drosophila melanogaster* testis. Genome Res. **10:** 2030–2043.

Ashburner, M., S. Misra, J. Roote, S. Lewis, R. Blazej *et al.*, 1999 An exploration of the sequence of a 2.9-Mb region of the genome of *Drosophila melanogaster*—The *Adh* region. Genetics **153:** 179–219.

Bauer, V., and C. Aquadro, 1997 Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **14:** 1252–1257.

Begun, D., and C. Aquadro, 1993 African and North American populations of *Drosophila melanogaster* are very different at the DNA level. Nature **365:** 548–550.

Begun, D., P. Whitley, B. Todd, H. Waldrip-Dail and A. Clark, 2000 Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156:** 1879–1888.

Benos, P., M. Gatt, M. Ashburner, L. Murphy, D. Harris *et al.*, 2000 From sequence to chromosome: the tip of the X chromosome of *D. melanogaster*. Science **287:** 2220–2222.

Bielawski, J., K. Dunn and Z. Yang, 2000 Rates of nucleotide substitution and mammalian nuclear gene evolution: approximate and maximum-likelihood methods lead to different conclusions. Genetics **156:** 1299–1308.

Bork, P., and E. Koonin, 1998 Predicting functions from protein sequences: where are the bottlenecks? Nat. Genet. **18:** 313–318.

Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen *et al.*, 1998 Predicting function: from genes to genomes and back. J. Mol. Biol. **283:** 707–725.

Brenner, S., 1999 Errors in genome annotation. Trends Genet. **15:** 132–133.

Brenner, S., C. Chothia and T. Hubbard, 1997 Population statistics of protein structures: lessons from structural classifications. Curr. Opin. Struct. Biol. **7:** 369–376.

Burge, C., and S. Karlin, 1997 Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. **268:** 78–94.

Chothia, C., 1992 One thousand families for the molecular biologist. Nature **357:** 543–544.

Civetta, A., and R. S. Singh, 1998 Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15:** 901–909.

Comeron, J., and M. Kreitman, 1998 The correlation between synonymous and nonsynonymous substitution in Drosophila: mutation, selection, or relaxed constraints? Genetics **150:** 767–775.

Comeron, J., M. Kreitman and M. Aguadé, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Duda, T. F., and S. R. Palumbi, 1999 Molecular genetics of evolutionary diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. Proc. Natl. Acad. Sci. USA **96:** 6820–6823.

Dunn, K., J. Bielawski and Z. Yang, 2001 Substitution rates in Drosophila nuclear genes: implications for translational selection. Genetics **157:** 295–305.

Duret, L., and D. Mouchiroud, 1999 Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila* and *Arabidopsis*. Proc. Natl. Acad. Sci. USA **96:** 4482–4487.

Endo, T., K. Ikeo and T. Gojobori, 1996 Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13:** 685–690.

Ewing, B., L. Hillier, M. Wendl and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:** 175–185.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gordon, D., C. Abajian and P. Green, 1998 Consed: a graphical tool for sequence finishing. Genome Res. **8:** 195–202.

Green, P., 1995 *GENEFINDER Documentation* (http://genetics.mgh.harvard.edu/doc/genefinder.doc.html).

Guigó, R., P. Agarwal, J. Abril, M. Burset and J. Fickett, 2000 An assessment of gene prediction accuracy in large DNA sequences. Genome Res. **10:** 1631–1642.

Hieter, P., and M. Boguski, 1997 Functional genomics: its all how you read it. Science **278:** 601–602.

Hudson, R. R., M. Kreitman and M. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Hurst, L., and N. Smith, 1999 Do essential genes evolve slowly? Curr. Biol. **9:** 747–750.

Kliman, R., and J. Hey, 1993 Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 1239–1258.

Kliman, R., and J. Hey, 1994 The effects of mutation and natural selection on codon bias in the genes of Drosophila. Genetics **137:** 1049–1056.

Kreitman, M., and R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

McDonald, J., and M. Kreitman, 1991 Adaptive evolution at the *Adh* locus in Drosophila. Nature **351:** 652–654.

Michaelmore, R. W., and B. C. Meyers, 1998 Clusters of resistance genes in plants evolve by divergent selection and a birth-and-death process. Genome Res. **8:** 1113–1130.

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol. Biol. Evol. **13:** 261–277.

Murphy, P. M., 1993 Molecular mimicry and the generation of host defense protein diversity. Cell **42:** 823–826.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Oliver, S., 1996 From DNA sequence to biological function. Nature **379:** 597–600.

Pamilo, P., and R. J. W. O'Neill, 1997 Evolution of the *Sry* genes. Mol. Biol. Evol. **14:** 49–55.

Powell, J., and E. Moriyama, 1997 Evolution of codon usage bias in Drosophila. Proc. Natl. Acad. Sci. USA **94:** 7784–7790.

Reese, M., G. Hartzell, N. Harris, U. Ohler, J. Abril *et al.*, 2000 Genome annotation assessment in *Drosophila melanogaster*. Genome Res. **10:** 483–501.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Rozen, S., and H. Skaletsky, 1998 Primer3 (code available at http://www-genome.wi.mit.edu).

Schmid, K. J., and D. Tautz, 1997 A screen for fast evolving genes from Drosophila. Proc. Natl. Acad. Sci. USA **94:** 9746–9750.

Schmid, K. J., L. Nigro, C. F. Aquadro and D. Tautz, 1999 Large number of replacement polymorphisms in rapidly evolving genes of Drosophila: implications for genome-wide surveys of DNA polymorphism. Genetics **153:** 1717–1729.

Sharp, P., and W.-H. Li, 1989 On the rate of DNA sequence evolution in Drosophila. J. Mol. Evol. **28:** 398–402.

Shields, D., P. Sharp, D. Higgins and F. Wright, 1988 "Silent" sites in Drosophila genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

Swanson, W. J., and V. D. Vacquier, 1995 Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc. Natl. Acad. Sci. USA **92:** 4957–4961.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Tautz, D., and K. J. Schmid, 1998 From genes to individuals—developmental genes and the generation of the phenotype. Proc. R. Soc. London Ser. B **353:** 231–240.

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Wilson, A., S. Carlson and T. White, 1977 Biochemical evolution. Annu. Rev. Biochem. **46:** 573–639.

Wright, F., 1990 The 'effective number of codons' used in a gene. Gene **87:** 23–29.

Wyckoff, G., W. Wang and C. Wu, 2000 Rapid evolution of male reproductive genes in the descent of man. Nature **403:** 304–309.

Yang, Z., 1999 *Phylogenetic Analysis Using Maximum Likelihood (PAML), Version 2.* University College, London.

Yang, Z., and R. Nielsen, 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J. Mol. Evol. **46:** 409–418.

Yang, Z., R. Nielsen, N. Goldman and A. Pedersen, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155:** 431–449.

Yokoyama, S., H. Zhang, F. B. Radlwimmer and N. S. Blow, 1999 Adaptive evolution of color vision of the Comoron coelacanth (*Latimeria chalumnae*). Proc. Natl. Acad. Sci. USA **96:** 6279–6284.

Communicating editor: M. Aguadé