Large Number of Replacement Polymorphisms in Rapidly Evolving Genes of Drosophila: Implications for Genome-Wide Surveys of DNA Polymorphism

Karl J. Schmid,*,‡ Loredana Nigro,† Charles F. Aquadro‡ and Diethard Tautz*,[1]

**Zoologisches Institut, Universität München, 80333 München, Germany, †Dipartimento di Biologia, University of Padua, 35122 Padua, Italy and ‡Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853*

Manuscript received May 13, 1999
Accepted for publication August 3, 1999

ABSTRACT

We present a survey of nucleotide polymorphism of three novel, rapidly evolving genes in populations of *Drosophila melanogaster* and *D. simulans.* Levels of silent polymorphism are comparable to other loci, but the number of replacement polymorphisms is higher than that in most other genes surveyed in *D. melanogaster* and *D. simulans.* Tests of neutrality fail to reject neutral evolution with one exception. This concerns a gene located in a region of high recombination rate in *D. simulans* and in a region of low recombination rate in *D. melanogaster*, due to an inversion. In the latter case it shows a very low number of polymorphisms, presumably due to selective sweeps in the region. Patterns of nucleotide polymorphism suggest that most substitutions are neutral or nearly neutral and that weak (positive and purifying) selection plays a significant role in the evolution of these genes. At all three loci, purifying selection of slightly deleterious replacement mutations appears to be more efficient in *D. simulans* than in *D. melanogaster*, presumably due to different effective population sizes. Our analysis suggests that current knowledge about genome-wide patterns of nucleotide polymorphism is far from complete with respect to the types and range of nucleotide substitutions and that further analysis of differences between local populations will be required to understand the forces more completely. We note that rapidly diverging and nearly neutrally evolving genes cannot be expected only in the genome of Drosophila, but are likely to occur in large numbers also in other organisms and that their function and evolution are little understood so far.

THE question of which evolutionary forces are responsible for the evolution of genes and proteins has been a contentious issue among molecular evolutionists. Many sequence comparisons of homologous proteins seem to confirm that the sequence evolution of proteins results mainly from the random fixation of neutral sequence variants, because the overwhelming majority of proteins exhibits fewer replacements than silent substitutions. According to the neutral theory of molecular evolution, functional and structural constraints determine what proportions of new variants are deleterious, thereby causing rate differences between different proteins. The rapidly growing database of DNA sequences provides evidence for both neutral and adaptive patterns in sequence data, but positive selection may be more frequent than thought previously (Kreitman and Akashi 1995). Most molecular evolutionists now agree that most new mutations in proteins are deleterious; there is still disagreement about what proportions of nondeleterious mutant alleles are neutral, nearly neutral, or advantageous (Kreitman 1996; Ohta 1996; Li 1997). There is also some debate about the relative role of drift and positive selection under weak selection because both nearly neutral and episodic selection models are able to produce the identical patterns of polymorphism with certain parameter assumptions (Gillespie 1994).

Rapidly evolving proteins are particularly interesting for this discussion. Three scenarios may explain why proteins evolve rapidly. The first may be a lack of strong functional or structural constraints. In this case, a large number of amino acid residues can be mutated without impairing the function of the protein and it evolves in a neutral fashion. The second may be positive selection for sequence divergence. Some classes of proteins appear to be affected predominantly by positive selection. Such proteins are involved in pathogen-host interaction and the immune system (Hughes *et al.* 1990; Fitch *et al.* 1991; Smith *et al.* 1995; Hughes 1997), sex determination (Whitfield *et al.* 1993; Sutton and Wilkinson 1997), and reproduction (Lee *et al.* 1995; Metz and Palumbi 1996; Tsaur and Wu 1997). A final explanation may be a mixture of the first two explanations: neutral evolution of some residues and positive selection at others.

A major limitation in understanding the factors governing protein evolution is a lack of knowledge about the distribution of evolutionary rates among the vast majority of genes in a genome. Most proteins whose evolution has been studied so far are functionally and

Corresponding author: Karl Schmid, Section of Genetics and Development, 403 Biotechnology Bldg., Cornell University, Ithaca, NY 14853-2703. E-mail: kjs21@cornell.edu

[1] *Present address:* Institut für Genetik, Universität zu Köln, Weyertal 121, 50931 Köln, Germany.

structurally well characterized and evolutionarily conserved. They constitute a nonrandom sample of all genes in a genome and may give a biased picture of the relative roles of mutation, selection, and drift. This is contrasted by the output from genome sequencing projects, where thousands of novel proteins are being identified whose structure, function, and molecular evolution remain largely unknown. As long as there are no complete genome sequences from closely related species available, it is necessary to use a random sample of genes for evaluating the range of evolutionary rates and the factors affecting sequence evolution in a genome.

Previously, we performed such a genome-wide survey and examined the sequence conservation of ~100 different, randomly isolated nonidentical clones from an embryonic cDNA library of *Drosophila melanogaster* to estimate the range and distribution of evolutionary divergence in the Drosophila genome by genomic filter hybridization (Schmid and Tautz 1997). In this screen, about one-third of these clones was classified as fast evolving, because they did not hybridize against genomic DNA from *Drosophila virilis* (40 million year evolutionary distance). More detailed sequence comparisons of 10 fast evolving cDNA clones between *D. melanogaster* and the closely related species *D. yakuba* (12 millon year evolutionary distance) revealed that the numbers of amino acid replacement substitutions are among the highest of currently known Drosophila genes.

Here we describe a survey of nucleotide polymorphism in populations of *D. melanogaster* and *D. simulans* at three fast evolving loci that were isolated in our previous screen. The goal of this study is to test whether the amino acid sequences of the proteins are also variable within species and to use the polymorphism data for tests of neutral evolution. The work described here extends the initial population survey of Schmid and Tautz (1997) because two additional loci and larger numbers of lines were analyzed. Results are compared to other genes that were surveyed in populations of both species to identify differences between fast evolving and conserved genes. Furthermore, we compare levels of polymorphism and divergence among loci and between lineages to differentiate between locus-specific and lineage-specific effects.

MATERIALS AND METHODS

Surveyed genes: Three genes that were classified as fast evolving in our screen were chosen for this analysis. They constitute novel, putative protein coding genes and are characterized by large numbers of nonsynonymous substitutions in comparisons between *D. melanogaster* and *D. yakuba* (Schmid and Tautz 1997). Note that their names are derived from their location in microtiter plates and do not reflect their cytological location in the genome of *D. melanogaster*. Although the genetic and biochemical functions of these genes are not known, there is strong evidence that all three of them are functional genes and not pseudogenes: (1) the ratio of nonsynonymous to synonymous substitutions (K_a/K_s) ratio between *D. melanogaster* and *D. yakuba* is <1, indicating purifying selection; (2) all insertion/deletion mutations between the two species are in frame; (3) the open reading frame (ORF) and expression patterns (K. J. Schmid and D. Tautz, unpublished data) are conserved between species. Figure 1 shows a schematic structure of the genes and the regions that were surveyed in this study.

Clone *anon1A3* encodes a protein of 489 amino acids and is characterized by a highly negative net charge. The gene has no similarity to other sequences in database searches and there are no close homologs in the Drosophila genome, as evaluated by Southern blotting. The gene is expressed in different tissues during embryogenesis: until gastrulation, the transcript is homogeneously distributed in the embryo and then becomes restricted to the developing mesoderm and central nervous system.

The protein encoded by clone *anon1E9* has a length of 588 amino acids and contains six C_2H_2 zinc-finger motifs. Four zinc-finger motifs are arrayed as a tandem in the center of the protein and the other two at the C terminus (Figure 1). Database searches reveal no close similarity to other zinc-finger proteins, and only those residues necessary for maintaining the structure of the fold are identical between *anon1E9* and the best matches. This gene is only maternally expressed during embryogenesis, and the transcript is homogeneously distributed in the early embryo. The transcript can be detected until the cellular blastoderm stage.

Clone *anon1G5* is the fastest evolving among the three genes. The putative protein has a length of 337 amino acids, does not exhibit sequence similarity to other genes, and is a single copy gene. The central region is very divergent between *D. melanogaster* and *D. yakuba* and also contains several insertions and deletions. This gene is expressed throughout embryogenesis and shows no developmental regulation at the transcriptional level.

Lines: Isofemale lines from the following locations were used. The survey of *anon1A3* in *D. melanogaster* includes four lines from Australia, five from North America, five from Asia (Iraq, Japan, and China), nine from Europe (Cyprus, France, Italy, Spain, and the former Soviet Union), and three from East Africa (Kenya and Zimbabwe). The *D. simulans* sample of *anon1A3* includes two lines from the United States, three from Mexico, one from Uruguay, and six from Zimbabwe. Gene *anon1E9* was surveyed in three lines of *D. melanogaster* from Australia, four from North America, one from Asia (Iraq), four from Europe, and three from East Africa. The *D. simulans* sample of *anon1E9* consists of three lines from North America, two from Mexico, one from Uruguay, and two from Zimbabwe. The *D. melanogaster* sample of gene *anon1G5* comprises three lines from Australia, five from North America, one from South America (Peru), two from Asia (Iraq and Japan), three from Europe, and two from East Africa. In the *D. simulans* sample are three lines from North America, four from Mexico, one from South America, and six from Zimbabwe.

The lines were collected by various researchers and given to us by M. Kidwell (*D. melanogaster*) and M. Turelli (*D. simulans*) or maintained at the University of Padua. The number of lines vary between genes, mainly because polymerase chain reaction (PCR) did not work well in all lines or high quality sequences could not be obtained. If only those lines are used for analysis for which we have sequences from all three genes, essentially the same results are observed; we therefore include all sequences from the different lines in the following analysis.

DNA preparation, PCR, and sequencing: DNA was prepared from single flies by phenol-chloroform extraction and ethanol precipitation (Sambrook *et al.* 1989). The loci were amplified

with PCR by using the following primers and cycling conditions in 50-μl reactions. Reaction conditions were as suggested by the manufacturer of the AmpliTaq DNA polymerase (Perkin-Elmer, Foster City, CA). Cycling conditions were: 2 min 95°, then 35 cycles of 1 min 94°, 1 min 48°, 2 min 72°, and final extension of 10 min 72°. The following primers were used for amplification: *anon1A3*-1, 5′-GGAGGAGGCGAG GAAGATGT-3′; *anon1A3*-2, 5′-GTTGGCAACATCAGACCA ACT-3′; *anon1E9*-PR3, 5′-AATATATGCTAGCGCACCATG-3′ *anon1E9*-PR2, 5′-ATTTCAACGTTTGCATTTGG-3′; *anon1G5*-PR3, 5′-AAGTATCTAGCCGACGAGGAC-3′; *anon1G5*-PR4, TACCCAGCT CTCATTCATCTC. The PCR products were gel purified with the Jetsorb kit (Genomed, Germany) and directly used for sequencing. Sequencing was carried out on an ABI 377 sequencer with DyeTerminator and AmpliTaq FS chemistry (Perkin-Elmer). Internal primers were used to sequence every base from both directions. Sequences were edited and aligned with ABI Factura, AutoAssembler, and Sequence Navigator programs. GenBank accession numbers are AA433202–AA433290 and AF161723–AF161796. Aligned sequences and figures of variable sites are available at http://www.mbg.cornell.edu/aquadro/sequences.html.

Chromosomal *in situ* hybridization: Chromosomes were prepared from Oregon-R lines from *D. melanogaster* and Soda Lake populations from *D. simulans* according to the protocol of Lim (1993). cDNA inserts (1 μg; cloned into pBluescript) were biotinylated with the BioNick nick translation kit (Gibco BRL, Gaithersburg, MD). Signal detection was achieved with Vectastain (Vector Laboratories, Burlingame, CA) and Detek Hrp (ENZO, Farmingdale, NY) kits. Photographs were taken on a Zeiss microscope with a Pixera digital camera and processed with the GNU image manipulator 1.0 program.

Analysis: The analysis of polymorphism and divergence was carried out using the program *DnaSP 3.0* (Rozas and Rozas 1999). Numbers of substitutions per site were computed with the program *Kestim* (Comeron 1995). θ, an estimate of the mutation parameter $4N_e\mu$ (Watterson 1975), and π, the average number of pairwise differences (Nei 1987), were estimated as measures of nucleotide diversity. Several tests for neutral evolution were applied. Tajima's *D* statistic compares the two different estimates of nucleotide diversity, θ and π, which should be identical under a neutral model (*D* is expected to be zero) (Tajima 1989). *D* is then tested for a significant difference from zero. A related test is Fu and Li's *D* (Fu and Li 1993), which counts the number of singletons in a population sample and tests whether this number is significantly different from the expected number under a neutral model. The HKA test (Hudson *et al.* 1987) tests whether observed levels of polymorphism and sequence divergence are consistent with a neutral equilibrium model. Regional differences in the ratio of polymorphic sites to fixed differences of the sequence data were tested with the program DNA Slider that employs various statistical procedures (see McDonald 1998). The McDonald-Kreitman test was used to compare ratios of silent and replacement substitutions within and between species (McDonald and Kreitman 1991).

Lineage-specific fixed differences and polymorphisms were assigned to either *D. melanogaster* or *D. simulans* lines by comparison to the *D. yakuba* outgroup sequence. The following GenBank accessions of *D. yakuba* homologs were used: AF005844 (*anon1A3*), AF005848 (*anon1E9*), and AF005852 (*anon1G5*). Essentially the same parsimony criteria as described by Akashi (1997) were applied to infer the ancestral state. The relative-rate test of Tajima (1993) was calculated to test whether the number of fixed substitutions differs between the two lineages. The relative-rate test of Muse and Gaut (1994) was calculated with single, randomly chosen alleles from the *D. melanogaster* and *D. simulans* samples and the homologous *D. yakuba* sequence as outgroup.

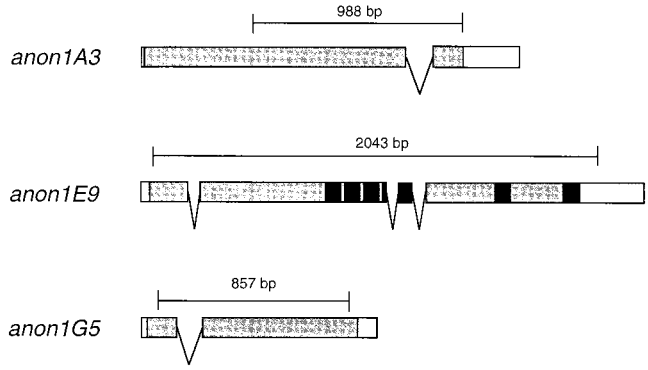

Figure 1.—Sequenced regions of the three loci surveyed in this study. A schematic representation of the cDNA clones and additional introns that were discovered after sequencing of genomic PCR fragments are depicted (the clones are oriented from 5′ to 3′). Gray boxes show the coding regions and white boxes show noncoding regions of the cDNA. The black boxes in *anon1E9* show the zinc-finger domains. Sequenced regions are outlined by the bars above each gene (lengths are given for the aligned *D. melanogaster* and *D. simulans* sequences).

The spatial distribution of substitutions along the coding sequence was tested with the test of Tang and Lewontin (1999), which is based on the empirical cumulative distribution function (ECDF) statistics. This test compares the difference between the observed cumulative distribution of distances between substitutions and a theoretical, homogeneous distribution. Critical values of the test statistic for significance tests are obtained by Monte Carlo simulations of the null model (see Tang and Lewontin 1999 for details). We applied the test to analyze the clustering of silent and replacement polymorphisms and fixed substitutions to identify differences between silent and replacement substitutions and between lineages.

We compared the frequency distributions of silent and replacement polymorphisms in the population samples to detect effects of weak selection (Akashi 1997, 1999; Akashi and Schaeffer 1997). First, we determined whether silent substitutions change a codon from a preferred to an unpreferred one, or *vice versa*. Codons were classified into preferred and unpreferred codons according to Akashi (1995) under the assumption that the same codons are preferred in *D. melanogaster* and *D. simulans* (Akashi and Schaeffer 1997). Second, the frequency distributions of preferred, unpreferred, and replacement substitutions were determined essentially as described by Akashi (1997) and compared by Mann-Whitney U tests. We used two different variants of the tests: the fdMWU test (Akashi and Schaeffer 1997), where only polymorphisms are included in calculating the frequency distribution of the different mutational classes, and the fddMWU test (Akashi 1997), which also includes fixed differences.

RESULTS

A schematic representation of the sequenced regions is shown in Figure 1. Sequence alignments showing polymorphic sites and fixed differences can be found in an appendix provided at our web site (see methods).

Locus *anon1A3*: This locus was sequenced from 26 lines of *D. melanogaster* and 12 lines of *D. simulans*; 930

bp were obtained from the ORF (63% of 1467 bp). The only intron within the surveyed region has a length of 58 bp and is located close to the 3′ end of the ORF. Sixteen polymorphisms were detected in *D. melanogaster* ($\pi = 0.0023$), of which 5 are synonymous and 11 nonsynonymous; 18 polymorphisms (5 synonymous, 11 nonsynonymous, and 2 noncoding) occur in the *D. simulans* sample ($\pi = 0.0045$; Table 1). In *D. melanogaster*, a deletion polymorphism affecting a single amino acid (Val) was found in the Iraq line. There are also two independent, fixed indel mutations; a comparison with the sequence of *D. yakuba* shows that they are caused by an insertion of Glu and Thr, respectively, in *D. melanogaster*. In *D. melanogaster* and *D. simulans*, the gene is located in 71A, on the left arm of chromosome 3.

Locus *anon1E9*: At this locus, little nucleotide polymorphism is observed in 15 lines from *D. melanogaster* ($\pi = 0.0007$), but a much higher level is observed among the 8 lines of *D. simulans* ($\pi = 0.0158$). In *D. melanogaster*, 3 of the segregating variants are synonymous, and 4 are nonsynonymous; in *D. simulans*, the numbers are 31 for synonymous and 33 for nonsynonymous variants. In both species, *anon1E9* harbors a small variable trinucleotide microsatellite with 5–9 repeat units of the GAG codon (coding for glutamate). Two alleles with 6 and 7 repeats were observed in *D. melanogaster* and four alleles with 5, 6, 7, and 9 repeat units in *D. simulans*. A second 6-bp deletion polymorphism (deleting Cys and Asn) is found in one strain of *D. simulans*. There are two fixed deletions, a 3-bp deletion in *D. melanogaster* (loss of a Ser) and a 6-bp deletion in *D. simulans* (loss of Ala and Val). In both species, nucleotide polymorphism at noncoding positions is not significantly different from silent positions in the coding region (Table 1). The physical location in *D. melanogaster* is 85B/C, on the right arm of chromosome 3. This region is inverted in *D. simulans* (see below).

Locus *anon1G5*: This locus was sequenced in 16 lines from *D. melanogaster* ($\pi = 0.0042$) and 14 lines from *D. simulans* ($\pi = 0.0125$). There are 6 silent and 4 replacement polymorphic sites among the 16 lines of *D. melanogaster*; there are 17 silent and 20 replacement polymorphic sites in the 14 lines of *D. simulans* (Table 1). Nucleotide diversity is lower in the intron (Table 1), but the difference from silent polymorphism is not significant in either *D. melanogaster* or *D. simulans*. Total polymorphism is threefold higher in *D. simulans* than in *D. melanogaster*. Three indel mutations are fixed between the two species. One deletion (2 bp) is found in the intron; the other two occur in the coding sequences of *D. melanogaster* (insertion of three residues: Ser-Phe-Arg) and *D. simulans* (deletion of two residues: Ser-Val). In *D. simulans*, an indel polymorphism affecting two residues (Ala-Arg) segregates with a frequency of ~50%. The gene maps to 95D/E on the right arm of chromosome 3 in *D. melanogaster* and *D. simulans*.

Nucleotide polymorphism in *D. melanogaster* and *D. simulans*

TABLE 1

Nucleotide diversity in *D. melanogaster* and *D. simulans*

Gene	Alleles	Length (bp)	Number of sites			Polymorphic sites				π			θ:	d^a
			Rep	Syn	Non	Total	Rep	Syn	Non	Total	Rep	Syn	Non	Total	
anon1A3															
D. melanogaster	26	988	720	210	58	16	11	5	0	0.0023	0.0018	0.0044	0.0000	0.0043	55
D. simulans	12	982	716	208	58	18	11	5	2	0.0045	0.0039	0.0062	0.0058	0.0061	
anon1E9															
D. melanogaster	15	2031	1362	384	285	10	4	3	3	0.0007	0.0004	0.0010	0.0002	0.0015	112
D. simulans	8	2016	1354	377	285	80	33	31	16	0.0158	0.0095	0.0309	0.0036	0.0153	
anon1G5															
D. melanogaster	16	829	603	165	61	10	4	6	1	0.0042	0.0016	0.0129	0.0078	0.0040	56
D. simulans	14	818	594	165	59	41	20	17	4	0.0125	0.0104	0.0202	0.0121	0.0165	
Observed range in *D. melanogaster*[b]: $\pi_{Tot} = 0.0000–0.0098$															
$\pi_{Syn} = 0.0000–0.0335$															
Observed range in *D. simulans*[b]: $\pi_{Tot} = 0.0000–0.0224$															
$\pi_{Syn} = 0.0000–0.0797$															

Rep, replacement; Syn, synonymous; Non, noncoding sites.

[a] Number of total differences between two randomly chosen alleles, one from *D. melanogaster* and one from *D. simulans* (interspecific difference).

[b] Data are from Moriyama and Powell (1996).

***simulans*:** The data in Table 1 show that nucleotide diversity differs among genes and also between *D. melanogaster* and *D. simulans*. Still, the polymorphism estimates are well within the range observed for other genes from both species (see Table 1). Note, however, that the level of nucleotide polymorphism between the species varies among the three loci: at *anon1A3* total nucleotide polymorphism (π) is about two times higher in *D. simulans* than in *D. melanogaster*, at *anon1G5* three times higher, and at *anon1E9* 23 times higher (Table 1). In the coding regions within each species, nucleotide diversity at silent sites is on average only threefold higher than at replacement sites. Total nucleotide diversity in *D. simulans* is about five times higher than in *D. melanogaster*; this difference has been noted before (*e.g.*, Aquadro 1992; Moriyama and Powell 1996). In Drosophila, nucleotide polymorphism is correlated with recombination rate (Begun and Aquadro 1992; Aquadro *et al.* 1994). In regions of low recombination, hitchhiking combined with selective sweeps (Maynard Smith and Haigh 1974; Kaplan *et al.* 1989; Stephan *et al.* 1992) or background selection (Charlesworth *et al.* 1993) is hypothesized to remove nucleotide variation at linked loci. The chromosomal location of all three genes was determined by *in situ* hybridization; a measure of recombination rate in *D. melanogaster* (adjusted coefficient of exchange, ACE) was obtained from Kindahl (1994; *anon1A3*, 1.569; *anon1E9*, 0.727; *anon1G5*, 1.739). The observed levels of nucleotide polymorphism at the three loci show a positive correlation with recombination rate in *D. melanogaster*.

Particularly strong evidence for the effect of recombination rate on the level of intraspecific nucleotide polymorphism is observed at locus *anon1E9*. At this locus, nucleotide diversity (π) is 23 times higher in *D. simulans* than in *D. melanogaster*, which is much more than the average difference between both species (Moriyama and Powell 1996). This difference is consistent with variation in the recombination rate between the two species at this locus (Sturtevant 1929; Ohnishi and Voelker 1979). In *D. melanogaster*, *anon1E9* maps to 85B/C in the centromeric region of chromosome 3 (Figure 2A). Two reports described a large inversion of this region between *D. melanogaster* and *D. simulans*. The studies disagree about the exact breakpoints: 84B3 to 92C in Ohnishi and Voelker (1979), and 84F1 to 93F6-7 in Lemeunier *et al.* (1986). Figure 2B shows that this inversion translocated the *anon1E9* locus away from the centromer into a region of a higher recombination rate. This might explain the much higher nucleotide polymorphism at this locus in *D. simulans*.

Tests of neutral evolution: Results of tests of neutral evolution are summarized in Tables 2 and 3. The observed levels of sequence variation at loci *anon1A3* and *anon1G5* in *D. melanogaster* and *D. simulans* and at locus *anon1E9* in *D. simulans* do not reject a neutral model of molecular evolution in the Tajima (1989), Fu and Li (1993), and HKA (Hudson *et al.* 1987) tests. The only significant deviation from neutrality is observed at locus *anon1E9* in *D. melanogaster*. Variation at this locus shows a significant difference from neutrality in the Tajima ($D = -2.156$, $P < 0.01$), Fu and Li ($D = -2.504$, $P < 0.05$), and HKA tests. In the latter test, a comparison with the 5′ *Adh* region of Kreitman and Hudson (1991) that is often used as a supposedly neutral control region rejects neutral evolution due to a lack of polymorphic sites (Table 2). We also applied the tests of McDonald (1998) to detect deviation from neutrality in subregions of the three genes. Across a wide range of recombination rates used in these tests, we have not uncovered a significant deviation from neutrality in any of the three loci in either species (analyses not shown).

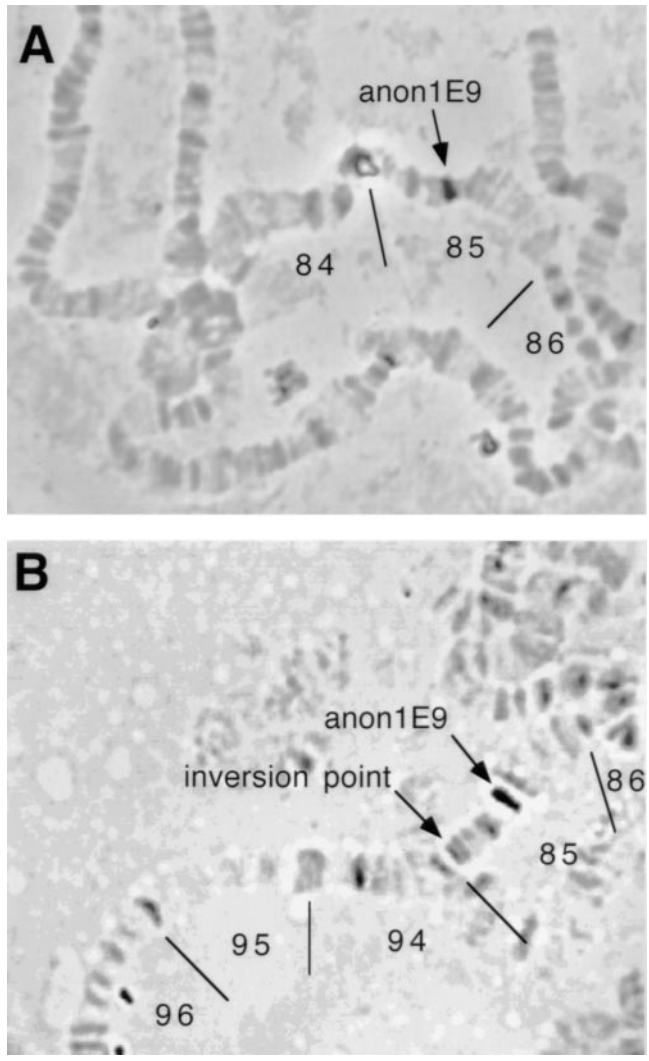

Figure 2.—Chromosomal *in situ* hybridization of gene *anon1E9* in *D. melanogaster* and *D. simulans*. (A) Location of *anon1E9* on the third chromosome of the *D. melanogaster* Oregon R strain. Major polytene band divisions are marked according to the maps in Sorsa (1988). (B) Location of *anon1E9* on the third chromosome of a *D. simulans* strain captured at Soda Lake, California. One of the two inversions breakpoints is marked by an arrow.

TABLE 2

Tests of neutral evolution using estimates of total nucleotide diversity

Gene	Tajima's *D*	Fu and Li's *D* (with outgroup)	HKA test[a] (χ^2)
anon1A3			
D. melanogaster	−1.648	−1.557	1.165
D. simulans	−1.135	−1.513	2.643
anon1E9			
D. melanogaster	−2.156**	−2.504*	6.372*
D. simulans	0.107	0.150	2.540
anon1G5			
D. melanogaster	0.220	0.007	2.165
D. simulans	−1.061	−1.541	0.232

* $P < 0.05$, ** $P < 0.01$.

[a] The HKA test was carried out with the total number of sites. *Adh* 5′ data of Kreitman and Hudson (1991) are used as reference sequence in *D. melanogaster* and *Gld* (Hamblin and Aquadro 1996) in *D. simulans*. Mutual comparisons of the three loci were not significant.

Neutral theory predicts that the ratio of silent to replacement substitutions should be identical for polymorphisms within species and for fixed differences between species. This prediction is tested in the McDonald-Kreitman (MK) test (McDonald and Kreitman 1991). Table 3 shows that the MK test does not reject the null hypothesis of neutral evolution in any of the three loci. The test at locus *anon1E9* is close to significance ($G = 2.98$, $P = 0.08$), because the ratio of replacement to silent substitutions is higher for fixed differences than for polymorphisms. The MK test can be modified with respect to length of regions analyzed. Such tests were carried out with subregions of loci *anon1E9* and *anon1G5*, because in these genes replacement substitutions cluster in certain regions (see below). The coding sequence of gene *anon1E9* was partitioned in four subregions: the N-terminal domain, the first zinc-finger cluster, the linker between the two zinc-finger clusters, and the second zinc-finger cluster (Figure 1). None of the subregion MK tests were significant. The same result was obtained with *anon1G5*, which is characterized by two conserved N- and C-terminal regions and a very rapidly evolving central domain (analyses not shown).

Lineage effects: We used *D. yakuba* as outgroup to assign fixed substitutions to either the *D. melanogaster* or *D. simulans* lineages. The number of these substitutions was then compared between lineages using the relative-rate test described by Tajima (1993). Under the null hypothesis of neutral evolution, there should be no significant differences in the number of substitutions between *D. melanogaster* and *D. simulans* lineages. Table 4 shows that significant rate differences were observed only for locus *anon1A3*. There are more than three times more replacement substitutions in the *D. melanogaster* than in the *D. simulans* lineage (18:6, $\chi^2 = 6.0$, $P < 0.05$). Identical results were obtained with the relative-rate test of Muse and Gaut (1994) using a randomly selected allele from the *D. melanogaster* and *D. simulans* samples and the *D. yakuba* sequence as outgroup (Table 4).

The test by Tang and Lewontin (1999) detected differences in the spatial distribution of substitutions along the coding region in the *D. melanogaster* and *D. simulans* lineages (Table 5). At loci *anon1E9* and *anon1G5*, both replacement polymorphisms and fixed differences are significantly clustered in the *D. simulans*, but not in the *D. melanogaster* lineage. At *anon1E9*, the replacement substitutions are clustered in the linker regions between the zinc-finger domains and at *anon1G5* in the central domain of the protein. No difference between the two lineages was seen at *anon1A3*. The test shows a homogeneous distribution of silent polymorphisms and silent fixed differences in five out of six comparisons. The only significant clustering of synonymous substitutions is seen at locus *anon1G5* in *D. melanogaster*. There, silent polymorphisms are absent in the region that shows a large number of replacement polymorphisms.

TABLE 3

McDonald-Kreitman test of neutral evolution

	Fixed differences (between species)		Polymorphic differences (within species)		
Gene	R	S	R	S	*G* value
anon1A3	26	19	22	10	0.95
anon1E9	44	22	37	34	2.98
anon1G5	22	10	24	22	2.17

R, replacement; S, silent substitutions. None of the *G* values is significant ($P > 0.05$).

TABLE 4

Relative-rate tests

A. Relative-rate test after Tajima (1993)[a]

	Replacement			Silent		
Gene	*MEL*	*SIM*	χ^2	*MEL*	*SIM*	χ^2
anon1A3	18	6	6.00*	10	7	0.53
anon1E9	25	13	3.79	11	9	0.20
anon1G5	5	10	1.67	2	4	0.67

B. Relative-rate test of Muse and Gaut (1994)[b]

	χ^2		
Gene	LRS	LRN	LRB
anon1A3	0.23	5.65*	5.88
anon1E9	0.02	1.75	1.77
anon1G5	3.54	0.15	3.71

* $P < 0.05$.

[a] Fixed silent and replacement substitutions were partitioned between *D. melanogaster* and *D. simulans* lineages using *D. yakuba* as outgroup. Sites with gaps or multiple substitutions were excluded. The significance of the χ^2 statistic was calculated from a χ^2 distribution with 1 d.f.

[b] One random allele was chosen from the *D. melanogaster* and *D. simulans* populations for this analysis. *D. yakuba* was used as an outgroup. The test statistic is the likelihood ratio of different models. Its distribution is not significantly different from a theoretical χ^2 distribution. LRS compares the synonymous rates in the two lineages (1 d.f.), LRN the nonsynonymous rates (1 d.f.), and LRB both rates simultaneously (2 d.f.).

The comparison of the frequency spectra of replacement, unpreferred, and preferred silent substitutions in different lines provides further evidence for the nature and direction of weak selection within populations. Since the three different types of mutation are interspersed along the sequence, identical frequency distributions of polymorphisms in each class are expected under a neutral model. This prediction forms the basis of tests for neutrality developed by Akashi, which are powerful for detecting weak selection if the assumptions of the test are met (Akashi 1997, 1999; Akashi and Schaeffer 1997). Preferred and unpreferred polymorphisms do not to appear to have different fitness effects at all three loci, and there is little evidence for the strong major codon usage observed in many other Drosophila genes (Akashi 1995). The numbers of preferred and unpreferred silent substitutions are relatively similar to each other in *D. melanogaster* and *D. simulans* (Table 6). In most other Drosophila genes studied so far, the number of unpreferred silent substitutions exceeds preferred substitutions in the *D. melanogaster* line. This is supported by comparisons of frequency distributions of preferred and unpreferred silent substitutions in the fdMWU and fddMWU tests. Frequency distributions are somewhat biased toward low frequencies and are not significantly different from each other at all three loci in both species. Similarly, no significant differences between frequency distributions of replacement and preferred or unpreferred silent polymorphisms are observed, although frequencies of replacement polymorphisms tend to be lower (results not shown).

TABLE 5

Test for heterogeneity in the location of lineage-specific substitutions along the coding sequence (Tang and Lewontin 1999)

		D. melanogaster		*D. simulans*	
	Type[a]	Events[b]	*T*	Events	*T*
anon1A3					
Polymorphic	S	5	0.110	5	0.164
	R	11	0.225	11	0.326
Fixed	S	10	0.168	7	0.138
	R	18	0.133	6	0.398
anon1E9					
Polymorphic	S	3	0.312	31	0.110
	R	4	0.284	33	0.263*
Fixed	S	11	0.139	9	0.364
	R	25	0.321	13	0.348*
anon1G5					
Polymorphic	S	5	0.521*	17	0.213
	R	4	0.281	21	0.296*
Fixed	S	2	0.120	4	0.219
	R	5	0.321	10	0.434*

Lineage-specific fixed differences were determined by aligning the *D. melanogaster* and *D. simulans* sequences with *D. yakuba* (see main text).

* $P < 0.05$.

[a] R, replacement substitution; S, silent substitution.

[b] Number of substitutions used in the simulation.

TABLE 6

Changes in codon preference at fixed silent substitutions

Gene		Unpreferred	Preferred
anon1A3	*D. melanogaster*	4	3
	D. simulans	2	3
anon1E9	*D. melanogaster*	5	3
	D. simulans	6	2
anon1G5	*D. melanogaster*	2	0
	D. simulans	1	1
		mp/su[a]	mu/sp[b]
anon1A3		5	5
anon1E9		11	9
anon1G5		5	7

Fixed changes at silent sites were classified as preferred to unpreferred and unpreferred to preferred according to Akashi (1995).

[a] Silent substitutions encoding preferred codons in *D. melanogaster* (mp) and unpreferred codons in *D. simulans* (su). Two randomly chosen alleles from each species were compared (see Akashi 1995).

[b] mu, unpreferred in *D. melanogaster*; sp, preferred in *D. simulans*.

DISCUSSION

The present survey in *D. melanogaster* and *D. simulans* demonstrates that the proteins encoded by loci *anon1A3*, *anon1E9*, and *anon1G5* exhibit a large degree of amino acid sequence variation not only between (Schmid and Tautz 1997) but also within species. The common characteristic of all three loci is that, in their coding regions, more replacement than silent substitutions are segregating within populations and are fixed between closely related species. At most loci that were studied in Drosophila, the opposite pattern was observed, namely an excess of silent over replacement substitutions within populations and between species. For example, in a survey of nucleotide polymorphism in Drosophila (22 loci from *D. melanogaster*, 12 loci from *D. simulans*; Moriyama and Powell 1996), and in more recent studies of *Gld* (Hamblin and Aquadro 1997), *white* (Kirby and Stephan 1995, 1996), *Tpi* (Hasson *et al.* 1998), and *hunchback* (Tautz and Nigro 1998), more silent than replacement polymorphisms in the coding region are segregating in populations of *D. melanogaster* and *D. simulans*. In the study of Moriyama and Powell (1996), 26.4% of all polymorphisms in *D. melanogaster* and 11.6% in *D. simulans* were replacement polymorphisms. Only at loci encoding the sperm-gland accessory protein *Acp26Aa* (Aguadé *et al.* 1992; Tsaur and Wu 1997; Aguadé 1998; Tsaur *et al.* 1998) and the viral resistance protein *ref(2)p* (Wayne *et al.* 1996) were more replacement than silent polymorphisms observed, and they evolve under positive selection. Therefore, it is interesting that all three loci surveyed in this study show a high proportion of replacement polymorphisms in both species. Three different hypotheses could explain this: a high mutation rate, a lack of constraints (high rate of neutral evolution), or positive selection. These factors will be discussed in turn.

No evidence for a higher mutation rate: It has been suggested that mutation rates may be variable in the genome of Drosophila. Interspecific DNA-DNA hybridization revealed a substantial fraction of single-copy DNA in the Drosophila genome that evolves rapidly (Werman *et al.* 1990). Sequencing of a boundary of fast and slowly evolving genomic regions led to the notion that the differences are not due to selection but to different mutation rates (Martin and Meyerowitz 1986). However, a high mutation rate is not supported as a plausible explanation for the rapid sequence divergence at the loci surveyed in this study. A high mutation rate should also affect silent sites of a locus and, consequently, a high silent substitution rate (in the absence of codon usage bias, which is the case at all three loci) would be expected. Compared to the silent divergence between *D. melanogaster* and *D. simulans* in the genes surveyed by Moriyama and Powell (1996), no larger numbers of silent substitutions per site are observed in interspecific comparisons of the three loci in this study (Table 7). Additionally, in our earlier screen (Schmid and Tautz 1997), 18 pairs of homologous sequences (including the three loci of this study) were compared between *D. melanogaster* and *D. yakuba*. Among all genes, the numbers of synonymous substitutions per site varied only 4-fold, while the numbers for replacement substitutions varied 30-fold. Since the number of silent substitutions per site is similar among all genes and is not correlated with the number of nonsynonymous substitutions, it is unlikely that the rapid evolution of these genes is driven by a high locus-specific mutation rate.

No evidence for strong positive selection: The other two hypotheses, namely neutral evolution and positive selection, were analyzed with various tests for neutral evolution. Kreitman and Akashi (1995) reviewed evidence that patterns of polymorphism and divergence seen at many loci under study in Drosophila are not in accord with the hypothesis that the variation seen is strictly neutral or unaffected by linked sites. Positive selection, purifying selection, and differences in recombination must be taken into account to explain the data. In fact, in the survey of Moriyama and Powell (1996), about half of the loci from *D. melanogaster* and *D. simulans* failed one of the tests for neutrality. Other studies also uncovered certain deviations from neutrality in a number of loci (*Gld*, Hamblin and Aquadro 1997; *concertina*, Wayne and Kreitman 1996; *hunchback*, Tautz and Nigro 1998; *white*, Kirby and Stephan 1995; *ref(2)p*, Wayne *et al.* 1996). At the three loci surveyed in this study, despite the high level of amino acid polymorphism and divergence, neutrality was not rejected by the tests, with the exception of locus *anon1E9* in *D. melanogaster*. Clearly, the rapid evolution of their amino

TABLE 7

Number of nonsynonymous and synonymous substitutions per site between *D. melanogaster*, *D. simulans*, and *D. yakuba*

Gene	K_a	95% CI	K_s	95% CI	K_a/K_s
anon1A3					
D. melanogaster vs. D. simulans	0.0421	0.0299–0.0658	0.0989	0.0494–0.1238	0.43
D. melanogaster vs. D. yakuba	0.0985	0.0802–0.1389	0.2261	0.1438–0.2727	0.44*
D. simulans vs. D. yakuba	0.0804	0.0642–0.1120	0.2137	0.1251–0.2592	0.38*
anon1E9					
D. melanogaster vs. D. simulans	0.0431	0.0361–0.0614	0.0879	0.0531–0.1064	0.49
D. melanogaster vs. D. yakuba	0.0899	0.0835–0.0123	0.2834	0.1922–0.2929	0.31**
D. simulans vs. D. yakuba	0.0810	0.0747–0.0113	0.2987	0.1998–0.3059	0.27**
anon1G5					
D. melanogaster vs. D. simulans	0.0546	0.0390–0.0849	0.0706	0.0305–0.1075	0.77
D. melanogaster D. yakuba	0.1531	0.1309–0.0220	0.2739	0.1615–0.3139	0.56
D. simulans vs. D. yakuba	0.1755	0.1500–0.2442	0.2691	0.1589–0.3172	0.65

Numbers and confidence intervals (CI) were determined with the program *Kestim* (Comeron 1995). K_a/K_s ratios were tested for significant difference from 1 based on the confidence intervals.
* $P < 0.05$, ** $P < 0.01$.

acid sequences is not driven by strong selection for sequence divergence, which, for example, was implicated in the rapid evolution of the accessory gland protein, *Acp26Aa* (Tsaur and Wu 1997; Aguadé 1998; Tsaur *et al.* 1998). All nucleotide polymorphisms at locus *anon1E9* in *Drosophila melanogaster* are singletons and cause negative Tajima's *D* and Fu and Li's *D* values, which suggest that the excess of rare polymorphisms is due to a recent selective sweep at this locus. However, *anon1E9* may not have been the target of this selective sweep. First of all, the MK test at this locus is not significant, so there is no evidence for selection in the protein. Further, this gene resides in a region of very low recombination, and the lack of polymorphic sites may result from hitchhiking with a recent selective sweep at another linked locus (Maynard Smith and Haigh 1974; Berry *et al.* 1991). As recent theoretical studies on selection incorporating the effects of recombination suggest, background selection may also be strong enough to decrease the level of polymorphism in centromeric regions as seen at locus *anon1E9* (Hudson and Kaplan 1995; Nordborg *et al.* 1996). But Tajima's *D* is highly (and significantly) negative, which is not predicted by background selection (Charlesworth *et al.* 1995). The most compelling evidence against selection-driven divergence at locus *anon1E9* comes from the fact that the region harboring this gene is inverted in *D. simulans* relative to *D. melanogaster*. Because of this chromosomal inversion, *anon1E9* is located in the middle of chromosomal arm 3R in *D. simulans* where recombination rates are higher than in the centromeric region. The observed level of polymorphism in *D. simulans* is 10-fold higher, and in this species, the tests for neutrality do not give any evidence for the hypothesis that the rapid evolution at *anon1E9* results from continuous positive selection.

Nearly neutral polymorphisms: The fixation rate of completely neutral mutations is determined only by the mutation rate (Kimura 1983), while the fixation of nearly neutral mutations is also dependent on the effective population size. In small populations, nearly neutral mutations behave effectively neutral if $N_e s < 1$, and their fate is determined mainly by random drift (Ohta 1973, 1992). Different average heterozygosities of *D. melanogaster* and *D. simulans* genes suggest that the effective total population size of *D. melanogaster* is three to six times smaller than that of *D. simulans* (Aquadro *et al.* 1988; Aquadro 1992; Moriyama and Powell 1996). Under a neutral model, slightly deleterious mutations are expected to be more efficiently removed from *D. simulans* than *D. melanogaster* populations, and slightly advantageous mutations should be more frequently fixed in *D. simulans*. Both the relative-rate test and the test by Tang and Lewontin (1999) detect lineage-specific differences at the three loci, supporting the hypothesis that a substantial number of segregating replacement polymorphisms are not neutral but slightly deleterious. The relative-rate test reveals a significantly larger number of replacement substitutions at locus *anon1A3* in the *D. melanogaster* lineage. The Tang and Lewontin test shows that nonsynonymous substitutions are clustered at *anon1E9* and *anon1G5* in *D. simulans*, but not in *D. melanogaster* (Table 5). A similar pattern was also found in the *G6pd* gene, where a larger number of replacement substitutions could be observed in the *D. simulans* lineage (Eanes *et al.* 1996). The MK test was highly significant in this case due to an excess of fixed replacement substitutions, indicating the occurrence of positive selection in the *D. simulans* lineage. At *anon1G5*, the number of replacement substitutions is also larger in the *D. simulans* than in the *D. melanogaster* lineage, but the difference is not significant in the relative-rate

test, and the MK test gives no evidence for an excess of replacement substitutions. Replacement substitutions are also clustered at *anon1E9* in the *D. simulans* sample, but the number of replacement substitutions in the *D. simulans* lineage is smaller than in the *D. melanogaster* lineage. The lineage effects at *anon1A3* and *anon1E9* loci are probably due to the smaller effective population size in *D. melanogaster*. A certain proportion of the substitutions appears to be slightly deleterious with selection coefficients too small to be "seen" by selection ($N_e s < 1$), but large enough to be removed from *D. simulans* populations, particularly if they occur in constrained regions of the protein. This conclusion is supported by a comparison of the frequency distributions of replacement and silent (preferred and unpreferred) substitutions. In comparison to silent polymorphisms, the distribution of replacement polymorphisms tends to be skewed toward low frequencies, suggesting that most of them are slightly deleterious.

Nucleotide polymorphism and interspecific divergence: Sequences that evolve under a neutral model are expected to show a correlation between interspecific divergence and intraspecific polymorphism (Kimura 1983). This prediction was not met in several studies of polymorphism and divergence in Drosophila, where polymorphism was lower (particularly in regions of low recombination) than expected from the interspecific divergence (Begun and Aquadro 1991, 1992; Berry *et al.* 1991; Langley *et al.* 1993). For example, a survey of the *cubitus interruptus*^{D} locus on the fourth chromosome did not uncover a single polymorphism in *D. melanogaster* and only one in *D. simulans* (Berry *et al.* 1991). Yet, the level of sequence divergence between both species is ~5%. This lack of correlation was explained by genetic hitchhiking with selective sweeps or background selection that removed most or all polymorphism within regions linked to the affected one.

The results of this survey are consistent with the findings of the earlier studies. Levels of nucleotide polymorphism among the three loci are different and correlate with the recombination rate. Under a neutral model, divergence between species should correspond to the observed level of nucleotide polymorphism. This is not observed; rather, the synonymous (K_s) and nonsynonymous divergences (K_a) are very similar among the three loci between *D. melanogaster* and *D. simulans* (Table 7). This is particularly evident at locus *anon1E9*, where *D. melanogaster* exhibits much less polymorphism (*e.g.*, silent sites: $\pi = 0.0001$) than *D. simulans* ($\pi = 0.0032$), yet the numbers of substitutions per site of *D. melanogaster* and *D. simulans* are similar when compared to *D. yakuba* (*D. melanogaster vs. D. yakuba*, $K_s = 0.2834$; *D. simulans vs. D. yakuba*, $K_s = 0.2987$).

Limitations of neutrality tests: Although tests for neutral evolution suggest that most sequence evolution in these genes is neutral or nearly neutral, our results need to be interpreted with caution. The main goal of this study was to determine whether the large variation of amino acids we observed between species also exists within populations of Drosophila. This is achieved most easily by comparing individuals sampled from across the whole geographic distribution of a species. Therefore, we sequenced alleles from worldwide collections of *D. melanogaster* and *D. simulans* lines and only small numbers of alleles from the same local populations. Such a sample, however, does not allow an analysis of the geographic population structure of species or an identification of different patterns of selection in local populations. For example, population-specific sweeps for certain loci were detected in a study of microsatellite variation in separate populations across the world (Schlötterer *et al.* 1997). Also, more detailed analyses of populations of *D. melanogaster* and *D. simulans* have revealed that both species indeed exhibit a considerable amount of population structure (Begun and Aquadro 1993; Hamblin and Veuille 1999). Nucleotide polymorphism of surveyed loci can vary significantly between different populations and affect tests of neutrality if they assume a mutation-drift equilibrium. For example, at the *Gld* locus in *D. melanogaster* (Hamblin and Aquadro 1997), the ratio of replacement to silent substitutions is significantly elevated (in a MK test) in the Chinese population sample, but not in two samples from Africa or a third sample from North America. In our sample, singletons may not necessarily be rare alleles (although they are treated like that in Tajima's test, therefore rendering *D* negative), but could segregate at high frequency in their local populations. A more comprehensive survey might reveal significant population differentiation at the three genes.

An additional problem is that current tests of neutral evolution are useful for detecting strong positive selection, but do not reject the null hypothesis of neutral evolution if selection coefficients are small. Power analyses have shown that Tajima's *D* and Fu and Li's *D* fail to detect a selective sweep when it occurred in the distant past or very recently and that their power is low with small sample sizes (Simonsen *et al.* 1995). Similar results were obtained in an analysis of the HKA test (M. Ford and C. F. Aquadro, unpublished results). This situation becomes even more complicated because weak and episodic selection models produce patterns of nucleotide polymorphism under realistic parameters that are indistinguishable from neutral evolution in a test like Tajima's *D* (Gillespie 1994). The existence of weak selection and the problems associated with detecting it are now widely acknowledged (Akashi 1996; Kreitman 1996; Ohta 1996; Ohta and Gillespie 1996; Wayne and Simonsen 1998).

Although strong positive selection does not seem to drive the rapid evolution of the three loci, we do not entirely exclude (for reasons discussed above) the possibility that at least a certain proportion of the large number of replacement polymorphisms may be subject to

weak positive or balancing selection. For example, in the complete absence of positive selection, one would expect a higher nonsynonymous rate in the *D. melanogaster* lineage, because of its smaller effective population size; not only completely neutral but also slightly deleterious substitutions should get fixed in this lineage. Indeed, at loci *anon1A3* and *anon1E9*, more replacement substitutions occur in the *D. melanogaster* lineage. In the most rapidly evolving gene *anon1G5*, however, more replacement substitutions occur in the *D. simulans* lineage (Table 4). Although the relative-rate test and the other tests for neutral evolution do not reject neutral evolution, the existence of some positive selection cannot be entirely excluded.

Implications for genome-wide surveys of nucleotide polymorphism: The three loci we surveyed for this study constitute a random sample of protein coding genes from the genome of Drosophila with regard to phenotypic effects. Although their biochemical functions are probably very different, their common characteristic is the fast evolution of their amino acid sequence as shown in our previous screen (Schmid and Tautz 1997) and in this study. Because of the random isolation of these clones, it is possible to estimate the fraction of genes in the Drosophila genome that are expected to show similar rates of evolution. In the original screen, about one-third of ~100 clones was scored as fast evolving by genomic cross-hybridization experiments. Sequence comparisons of 10 clones with their *D. yakuba* homologs lead to the estimate that ~20% of the Drosophila genes are fast evolving and exhibit a large number of replacement polymorphisms. Since the Drosophila genome probably has a similar number of genes as *Caenorhabditis elegans* (~19,000; *C. elegans* Sequencing Consortium 1998), several thousand Drosophila genes can be expected to evolve with few evolutionary constraints.

We propose that a similar proportion of rapidly evolving genes can be expected in the genomes of other eukaryotes. All three genes of this study have no or only low sequence similarity to genes from other species and therefore are "orphans." Since orphans are also common in other eukaryotes whose genome has been partially or completely sequenced (Goffeau *et al.* 1996; Bevan *et al.* 1998; *C. elegans* Sequencing Consortium 1998), it is probable that these fast evolving genes are ubiquitous components of eukaryotic genomes. It will be interesting to explore the long-term evolution of these rapidly evolving genes and their utility for phylogenetic analyses of closely related taxa. It will also be of critical importance to understand the relationship between the rapid sequence evolution and the structure and function of the proteins encoded by these genes. If there is only little conservation on sequence level, it may not be possible to identify homologs in other phyla (if they exist there at all). For example, we were not able to detect significant sequence similarity between *anon1A3* and *anon1G5*, and the genes from the *C. elegans* genome. In these cases, additional studies such as a genetic analysis or a determination of the protein structure will be necessary for identifying the function of these proteins. It will also be important to study whether these genes contribute to the phenotypic differences between species (Tautz and Schmid 1997).

This article is dedicated to the memory of our collaborator Loredana Nigro who sadly died in October 1998. We thank M. Hamblin for advice about *in situ* hybridization and the members of the Aquadro lab for discussion. This work was supported by a postdoctoral fellowship of the Deutsche Forschungsgemeinschaft (DFG) to K.J.S., an European Molecular Biology Organization short-term fellowship to L.N., a National Institutes of Health grant to C.F.A., and various DFG grants to D.T.

LITERATURE CITED

Aguadé, M., 1998 Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150:** 1079–1089.

Aguadé, M., N. Miyashita and C. H. Langley, 1992 Polymorphism and divergence in the *Mst26A* male accessory gland gene region in Drosophila. Genetics **132:** 755–770.

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 acid substitution, and larger proteins in *D. melanogaster*. Genetics **144:** 1297–1307.

Akashi, H., 1997 Codon bias in Drosophila: population genetics of mutation-selection drift. Gene **205:** 269–278.

Akashi, H., 1999 Inferring the fitness effects of DNA mutations from polymorphism and divergence data: statistical power to detect directional selection under stationarity and free recombination. Genetics **151:** 221–238.

Akashi, H., and S. W. Schaeffer, 1997 Natural selection and the frequency distribution of "silent" DNA polymorphism in Drosophila. Genetics **146:** 295–307.

Aquadro, C. F., 1992 Why is the genome variable? Insights from *Drosophila*. Trends Genet. **8:** 355–362.

Aquadro, C. F., K. M. Lado and W. A. Noon, 1988 The *rosy* region of *Drosophila melanogaster* and *Drosophila simulans*. I. Contrasting levels of naturally occurring DNA restriction map variation an divergence. Genetics **119:** 875–888.

Aquadro, C. F., D. J. Begun and E. C. Kindahl, 1994 Selection, recombination and DNA polymorphism in Drosophila, pp. 46–56 in *Non-neutral Evolution: Theories and Molecular Data*, edited by B. Golding. Chapman & Hall, New York.

Begun, D. J., and C. F. Aquadro, 1991 Molecular population genetics of the distal portion of the X chromosome in *Drosophila*: evidence for genetic hitchhiking of the *yellow-achaete* region. Genetics **129:** 1147–1158.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature **356:** 519–520.

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

Berry, A. J., J. W. Ajioka and M. Kreitman, 1991 Lack of polymorphism on the Drosophila fourth chromosome resulting from selection. Genetics **129:** 1111–1117.

Bevan, M., I. Bancroft, E. Bent, K. Love, H. Goodman *et al.*, 1998 Analysis of 1.9 Mb of contiguous sequence from chromosome 4 of *Arabidopsis thaliana*. Nature **391:** 485–488.

C. elegans Sequencing Consortium, 1998 Genome sequence of the nematode *C. elegans*: a platform for investigating biology. Science **282:** 2012–2018.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics **134:** 1289–1303.

Charlesworth, D., B. Charlesworth and M. T. Morgan, 1995 The pattern of neutral molecular variation under the background selection model. Genetics **141:** 1619–1632.

Comeron, J. M., 1995 A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J. Mol. Evol. **41:** 1152–1159.

Eanes, W. F., M. Kirchner, J. Yoon, C. H. Biermann, M. A. McCartney *et al.*, 1996 Historical selection, amino acid polymorphism and lineage-specific divergence at the *G6pd* locus in *Drosophila melanogaster* and *D. simulans*. Genetics **144:** 1027–1041.

Fitch, W. M., J. M. E. Leiter, X. Li and P. Palese, 1991 Positive Darwinian evolution in human influenza A viruses. Proc. Natl. Acad. Sci. USA **88:** 4270–4274.

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

Gillespie, J. H., 1994 Alternatives to the neutral theory, pp. 1–17 in *Non-neutral Evolution*, edited by B. Golding. Chapman & Hall, New York.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon *et al.*, 1996 Life with 6000 genes. Science **274:** 563–567.

Hamblin, M. T., and C. F. Aquadro, 1996 High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model. Mol. Biol. Evol. **13:** 1133–1140.

Hamblin, M. T., and C. F. Aquadro, 1997 Contrasting patterns of non-neutral nucleotide sequence variation at the *Glucose dehydrogenase* locus in different populations of *Drosophila melanogaster*. Genetics **145:** 1053–1062.

Hamblin, M. T., and M. Veuille, 1999 Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics **153:** 305–317.

Hasson, E., I.-N. Wang, L.-W. Zeng, M. Kreitman and W. F. Eanes, 1998 Nucleotide variation in the Triosephosphate isomerase (*Tpi*) locus of *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **15:** 756–769.

Hudson, R. R., and N. L. Kaplan, 1995 Deleterious background selection with recombination. Genetics **141:** 1605–1617.

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

Hughes, A. L., 1997 Rapid evolution of immunoglobulin superfamily C2 domains expressed in immune system cells. Mol. Biol. Evol. **14:** 1–5.

Hughes, A. L., T. Ota and M. Nei, 1990 Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I Major-Histocompatibility-Complex molecules. Mol. Biol. Evol. **7:** 515–524.

Kaplan, N., R. R. Hudson and C. H. Langley, 1989 The "hitchhiking effect" revisited. Genetics **116:** 153–159.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution*. Cambridge University Press, Cambridge.

Kindahl, E. C., 1994 Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*. Ph.D. Thesis, Cornell University.

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of variation at the *white* gene of *Drosophila melanogaster*. Genetics **144:** 636–645.

Kreitman, M., 1996 The neutral theory is dead. Long live the neutral theory. BioEssays **18:** 678–683.

Kreitman, M., and H. Akashi, 1995 Molecular evidence for natural selection. Ann. Rev. Ecol. Syst. **26:** 403–422.

Kreitman, M., and R. 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.

Langley, C. H., J. McDonald, N. Miyashita and M. Aguadé, 1993 Lack of correlation between interspecific divergence and intraspecific polymorphism at the suppressor of forked region in *Drosophila melanogaster* and *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **90:** 1800–1803.

Lee, Y.-H., T. Ota and V. Vacquier, 1995 Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol. Biol. Evol. **12:** 231–238.

Lemeunier, F., J. R. David and L. Tsacas, 1986 The *melanogaster* species group, pp. 147–256 in *The Genetics and Biology of Drosophila*, Vol. 3e, edited by M. Ashburner and E. Novitsiki. Academic Press, London.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Lim, J. K., 1993 *In situ* hybridization with biotinylated DNA. Dros. Inf. Serv. **72:** 73–76.

Martin, C. H., and E. M. Meyerowitz, 1986 Characterization of the boundaries between adjacent rapidly and slowly evolving genomic regions in Drosophila. Proc. Natl. Acad. Sci. USA **83:** 8654–8658.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23:** 23–35.

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

McDonald, J. H., 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.

Metz, E. C., and S. R. Palumbi, 1996 Positive selection and sequence rearrangements generate extensive polymorphism in the gamete recognition protein bindin. Mol. Biol. Evol. **13:** 397–406.

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

Muse, S., and B. S. Gaut, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol. Biol. Evol. **11:** 715–724.

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

Nordborg, M., D. Charlesworth and B. Charlesworth, 1996 The effect of recombination on background selection. Genet. Res. **63:** 159–174.

Ohnishi, S., and R. A. Voelker, 1979 Comparative studies of allozyme loci in *Drosophila melanogaster* and *Drosophila simulans*. II. Gene arrangements on the third chromosome. Jpn. J. Genet. **54:** 203–209.

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

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Ann. Rev. Ecol. Syst. **23:** 263–286.

Ohta, T., 1996 The current significance and standing of neutral and nearly neutral theories. BioEssays **18:** 673–677.

Ohta, T., and J. H. Gillespie, 1996 The development of neutral and nearly neutral theories. Theor. Pop. Biol. **49:** 128–142.

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

Sambrook, J., E. F. Fritsch and T. Maniatis, 1989 *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Schlötterer, C., C. Vogl and D. Tautz, 1997 Polymorphism and locus-specific effects on polymorphism at microsatellite loci in natural *Drosophila melanogaster* populations. Genetics **146:** 309–320.

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

Simonsen, K. L., G. A. Churchill and C. F. Aquadro, 1995 Properties of statistical tests of neutrality for DNA polymorphism data. Genetics **141:** 413–429.

Smith, N. H., J. Maynard Smith and B. G. Spratt, 1995 Sequence evolution if the *porB* gene of *Neisseria gonorrhoeae* and *Neisseria meningitidis*: evidence of positive Darwinian selection. Mol. Biol. Evol. **12:** 363–370.

Sorsa, V., 1988 *Chromosome Maps of Drosophila*, Vol. II. CRC Press, Boca Raton, FL.

Stephan, W., T. H. E. Wiehe and M. W. Lenz, 1992 The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. Theor. Popul. Biol. **41:** 237–254.

Sturtevant, A. H., 1929 The genetics of *Drosophila simulans*. Carnegie Inst. Wash. Publ. No. **339:** 1–62.

Sutton, K. A., and M. F. Wilkinson, 1997 Rapid evolution of a homeodomain: evidence for positive selection. J. Mol. Evol. **45:** 579–588.

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

Tajima, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics **135:** 599–607.

Tang, H., and R. C. Lewontin, 1999 Locating regions of differential variability in DNA and protein sequences. Genetics **153:** 485–495.

Tautz, D., and L. Nigro, 1998 Microevolutionary divergence pattern of the segmentation gene *hunchback* in *Drosophila*. Mol. Biol. Evol. **15:** 1403–1411.

Tautz, D., and K. J. Schmid, 1997 From genes to individuals—developmental genes and the generation of the phenotype. Proc. Roy. Soc. Ser. B **353:** 231–240.

Tsaur, S.-C., and C.-I. Wu, 1997 Positive selection and molecular evolution of a gene of male reproduction, *Acp26Aa* of *Drosophila*. Mol. Biol. Evol. **14:** 544–549.

Tsaur, S.-C., C.-T. Ting and C.-I. Wu, 1998 Positive selection driving the evolution of a gene of male reproduction, *Acp26Aa*, of Drosophila. II. Divergence versus polymorphisms. Mol. Biol. Evol. **15:** 1040–1046.

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

Wayne, M., and M. Kreitman, 1996 Reduced variation at *concertina*, a heterochromatic locus in *Drosophila*. Genet. Res. Cambr. **68:** 101–108.

Wayne, M. L., and K. L. Simonsen, 1998 Statistical tests of neutrality in the age of weak selection. Trends Ecol. Evol. **13:** 236–240.

Wayne, M. L., D. Contamine and M. Kreitman, 1996 Molecular population genetics of *ref(2)p*, a locus which confers viral resistance in *Drosophila*. Mol. Biol. Evol. **13:** 191–199.

Werman, S., E. H. Davidson and R. J. Britten, 1990 Rapid evolution in a fraction of the Drosophila nuclear genome. J. Mol. Evol. **30:** 281–289.

Whitfield, L. S., R. Lovell-Badge and P. N. Goodfellow, 1993 Rapid sequence evolution of the mammalian sex-determining gene *SRY*. Nature **364:** 713–715.

Communicating editor: A. G. Clark