Polymorphism and Divergence at a Drosophila Pseudogene Locus

Jonathan K. Pritchard[1] and Stephen W. Schaeffer

Department of Biology and Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received December 27, 1996
Accepted for publication June 1, 1997

ABSTRACT

The larval cuticle protein (*Lcp*) cluster in *Drosophila melanogaster* contains four functional genes and a closely related pseudogene. A 630-bp fragment including the larval cuticle pseudogene locus (*Lcpψ*) was nucleotide sequenced in 10 strains of *D. melanogaster* and a 458-bp *Lcpψ* fragment from *D. simulans* was also sequenced. We used these data to test the hypotheses that the rates of synonymous and nonsynonymous substitution are equal, that the absolute levels of variation are higher than in functional genes, and that intraspecific polymorphism is correlated with interspecific divergence. As predicted, synonymous and nonsynonymous substitution rates were equivalent, and overall nucleotide divergence between *D. melanogaster* and *D. simulans* (Jukes-Cantor distance = 0.149 ± 0.150) was extremely high. However, within-species DNA sequence comparisons at *Lcpψ* revealed lower levels of polymorphism ($\hat{\Theta}$ = 0.001 ± 0.001) than at many functional loci in *D. melanogaster*. Using the HUDSON, KREITMAN, and AGUADÉ (HKA) test, we show that the level of polymorphism in *Lcpψ* within *D. melanogaster* is lower than expected given the amount of divergence between *D. melanogaster* and *D. simulans* when the pseudogene data are compared to the *Adh* 5′ flanking region. Because the *Lcpψ* lies in a region of relatively infrequent recombination, we suggest that the low level of within-species polymorphism is the result of background selection.

PSEUDOGENES are DNA segments that show high similarity to functional genes, but are inactive as a result of sequence alterations that include the loss of regulatory sequences or the acquisition of premature stop codons (LI 1983). The neutral theory makes two predictions about the evolution of pseudogenes (KIMURA 1983). First, the rate of synonymous and nonsynonymous substitutions should be equal because the lack of functional constraint on the pseudogene removes the distinction between these two kinds of sites. In fact, comparisons of pseudogenes with their functional parent genes do generally show an accelerated rate of nonsynonymous substitution compared to synonymous substitution (LI 1981; MIYATA and HAYASHITA 1981; GOJOBORI *et al.* 1982). Second, within-species polymorphism and between-species divergence should both be higher in pseudogenes than in functional genes because levels of diversity are inversely related to the strength of selective constraints (KIMURA 1983; HUDSON *et al.* 1987). We present here an intraspecific study of nucleotide sequence variation of a *Drosophila melanogaster* pseudogene to test three hypotheses: (1) the rates of synonymous and nonsynonymous substitution within pseudogenes are equal; (2) pseudogene substitution rates exceed or equal silent site substitution rates in functional genes, and (3) intraspecific polymorphism is correlated with interspecific divergence in pseudogenes.

We have chosen to study a cuticle protein pseudogene first described by SNYDER *et al.* (1982). The small multigene family that encodes the larval cuticle proteins (*Lcp*) is composed of four protein-coding sequences (*Lcp1–Lcp4*) and one pseudogene (*Lcpψ*). This gene family is located within a 9-kb region of the right arm of chromosome *2* at cytogenetic map position 44D (Figure 1). The sequence of the entire cuticle gene complex was given by SNYDER *et al.* (1982) and EISSENBERG and ELGIN (1983). *Lcp1, Lcp2, Lcp3,* and *Lcp4* have all been shown to encode cuticle proteins that are expressed in the late third instar. *Lcp1* and *Lcp2* are transcribed in one direction, while *Lcp3* and *Lcp4* are transcribed in the opposite direction. A short intron splits each of the four functional genes near the 5′ end of its transcript. The pseudogene *Lcpψ* is oriented in the same direction as *Lcp1* and *Lcp2*, and appears to have formed from an unequal crossing-over event between *Lcp1* and *Lcp2*. The size and position of the intron is conserved between the pseudogene and the functional genes.

Lcpψ has been characterized as a pseudogene for several reasons (SNYDER *et al.* 1982). While the pseudogene has a conserved translation initiation codon, a 35-bp deletion upstream has eliminated the TATA box. Furthermore, stop codons at positions 23 and 72 would lead to premature termination of translation. [SNYDER *et al.* (1982) claim that translation would be halted much later, at a UGA termination codon just 19 amino

Corresponding author: Stephen W. Schaeffer, Department of Biology, The Pennsylvania State University, 208 Mueller Labs, University Park, PA 16802-5301. E-mail: sws4@psu.edu
[1]*Current address:* Department of Biological Sciences, Stanford University, Stanford, CA 94305.

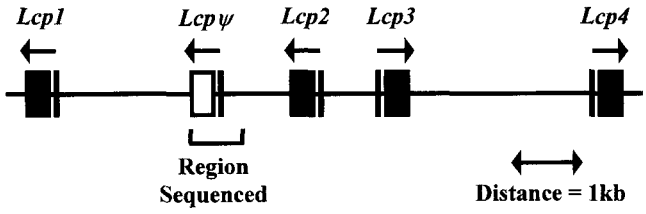

FIGURE 1.—Fine structure of the larval cuticle gene cluster showing the locations of the five gene sequences, *Lcp1, Lcp2, Lcp3, Lcp4,* and *Lcpψ*. The direction of transcription for each of the four genes and the progenitor of the pseudogene is show with an arrow above each gene. The fragment of DNA that includes the pseudogene *Lcpψ*, is shown below the genomic region.

acid residues earlier than in *Lcp1* and *Lcp2*, but reanalysis of the original *Lcpψ* sequence shows that stop codon to be out-of-frame.] In addition, the splicing acceptor sequence for the intron is mutated. They suggest that the substitution of charged amino acids in the signal peptide of *Lcpψ* would prevent secretion because signal peptides are normally comprised of hydrophobic amino acids. Finally, SNYDER *et al.* were unable to detect transcripts of *Lcpψ* at any stage of development. Thus, *Lcpψ* presents an excellent opportunity to study how pseudogenes evolve within and between species of Drosophila.

MATERIALS AND METHODS

Fly strains and genomic DNA preparation: Ten strains of *D. melanogaster* and one strain of *D. simulans* were kindly provided by A. G. CLARK. The *D. melanogaster* and *D. simulans* strains were collected from Winters, California in June 1987. The *D. melanogaster* strains were made homozygous for the second chromosome with the balancer strain *Bl L²/SM5* using methods described by ASHBURNER (1989). These lines are a subset of those used in the study of CLARK (1989), which provides further details about preparation of the fly lines. Genomic DNA was isolated from each isochromosomal line with the procedure of BINGHAM *et al.* (1981).

PCR and nucleotide sequencing: A 630-nucleotide fragment was amplified from each of the 10 strains of *D. melanogaster* with PCR (SAIKI *et al.* 1988). The following oligonucleotide primers were used to amplify *Lcpψ* from *D. melanogaster* genomic DNA. The 5′ primer (*Lcpψ*-B) begins at nucleotide 2462 (5′ CAC ACG AGC CGA TTA CCC TAG TAT 3′) and the 3′ primer (*Lcpψ*-2) begins at nucleotide 1833 (5′ AAG AGT CTT GGT GAT GGT GCG GAT 3′), where the nucleotide positions of the 5′ oligonucleotide base are those designated in SNYDER *et al.* (1982). The combination of *Lcpψ*-2 and *Lcpψ*-B failed to amplify the pseudogene sequence from *D. simulans*. A new 5′ primer (*Lcpψ*-C), which begins at nucleotide 2398 (SNYDER *et al.* 1982), was designed to amplify *Lcpψ* from *D. simulans* (5′ TCA GTC AAC GTT CGT TCT CG 3′). The oligonucleotide pair *Lcpψ*-2 and *Lcpψ*-C amplified a 458-bp fragment. Double-stranded PCR fragments were converted to single-stranded templates for nucleotide sequencing with λ exonuclease (HIGUCHI and OCHMAN 1989). The sequences of both strands of DNA were determined with four oligonucleotide primers using the chain-termination method (SANGER *et al.* 1977). The overlapping sequences for each strain were assembled with the SEQMAN and SEQMANED programs (DNASTAR, Madison, Wisconsin).

The 11 *Lcpψ* sequences of *D. melanogaster* and *D. simulans* determined in this study have been deposited in the GenBank/EMBL data libraries under the accession numbers U17196–U17205 and U68476. The Canton S *Lcpψ* sequence previously determined by SNYDER *et al.* (1982) was used as a reference sequence (GenBank/EMBL data libraries accession numbers: for *Lcp1,* J01080; for *Lcpψ, Lcp2, Lcp3,* and *Lcp4,* J01081) and for comparisons of *Lcpψ* to the paralogous genes *Lcp1* and *Lcp2*.

Nucleotide sequence alignments and statistical analyses: The 48 and 44 nucleotides within the two PCR primers used to amplify the *D. melanogaster* or *D. simulans* genes, respectively, were removed before sequence alignment. The 13 nucleotides downstream from the termination signal at nucleotides 567–569 were excluded from analysis because sequence determinations were ambiguous. The number of bases determined for each *D. melanogaster* sequence varied from 562 to 569 nucleotides depending on the presence or absence of two deletion polymorphisms. The pseudogene itself (including the intron region) was 384 or 386 nucleotides long. We used the previously published alignment of *Lcp1, Lcp2,* and *Lcpψ* of SNYDER *et al.* (1982) to align the 11 pseudogene sequences examined in this study (Figure 2). We moved the position of a four-codon deletion in *Lcp2* that began at nucleotide 281 to its current location that begins at nucleotide 291. This change was suggested by the AALIGN program in DNASTAR, which maximized identity and conservative amino acid substitutions between amino acid sequences *Lcp1* and *Lcp2*. The aligned sequences were assembled with the Eyeball Sequence Editor (ESEE, version: 1.09) (CABOT and BECKENBACH 1989). In comparing the *Lcpψ* sequences, any nucleotide site with two nucleotides present was defined as a segregating or polymorphic site. Nucleotide sites found in insertions or deletions assumed in aligned sequences were excluded from analyses of nucleotide diversity. The *Molecular Evolutionary Genetic Analysis* software package (*MEGA,* version: 1.02) (KUMAR *et al.* 1994) was used to estimate genetic distances between *D. melanogaster* and *D. simulans* and between paralogous loci, while the *DNA Sequence Polymorphism* program (*DnaSP,* version 2.0) (ROZAS and ROZAS 1997) was used in estimating genetic diversity within *D. melanogaster Lcpψ*.

RESULTS

Intra- and interspecific nucleotide polymorphism in *Lcpψ*: The reference sequence obtained by SNYDER *et al.* (1982) from the Canton-S strain of *D. melanogaster* was designated as the $Lcp\psi^{A}$ haplotype. We observed two additional haplotypes or multisite genotypes in our sample of 10 *D. melanogaster* cuticle pseudogene sequences from the Winters population. Seven strains, C01, C08, C10, C12, C20, C23, and C32, had the $Lcp\psi^{B}$ haplotype while the remaining three strains, C09, C11, and C25, had the $Lcp\psi^{C}$ haplotype. The alignment of the three *Lcpψ* haplotypes with the two functional genes *Lcp1* and *Lcp2* is shown in Figure 2.

The derived state for each nucleotide substitution or sequence length variant within the *D. melanogaster Lcpψ* sample may be inferred from a comparison with the parental genes *Lcp1* and *Lcp2,* using a parsimony criterion. A 9-bp deletion polymorphism was observed in the 5′ flanking region at nucleotide 97 in the $Lcp\psi^{C}$ haplotype. Associated with this deletion is an A to C transversion at the nucleotide position immediately 5′

to the lost bases. Given the scarcity of polymorphic nucleotides in this sample, it seems likely that this transversion was generated in the deletion event and will not be considered in further analyses. The only other variable nucleotide position in the Winters population is a C to A transversion at nucleotide 337 that results in an amino acid replacement in *Lcpψ*^{B}.

We observed 11 sequence length variants within the *Lcpψ* coding region in the sampled *D. melanogaster* and *D. simulans* alleles (Table 1). The 2-bp deletion after nucleotide 274 of *Lcpψ*^{A} and *Lcpψ*^{B} is polymorphic within *D. melanogaster.* The sequence length variants at positions 289 and 388 are fixed in *D. melanogaster.* Two deletions and three insertions are unique to *D. simulans,* while one insertion and two deletions are shared between *D. melanogaster* and *D. simulans.*

The extensive sequence length variation results in the premature termination of the *Lcpψ* messages in *D. melanogaster* and *D. simulans. Lcpψ*^{A} and *Lcpψ*^{B} genes would produce a truncated peptide because of termination at codons 23 and 72. Translation of *Lcpψ*^{C} would terminate at codons 29, 90, and 100, while *Lcpψ* in *D. simulans* has five stop codons, at positions 31, 39, 65, 81, and 82 of the coding region. The observation that *Lcpψ* translation would halt at different stop codons in different *D. melanogaster* alleles, and at yet another codon in *D. simulans,* provides further evidence against any function for the *Lcpψ* gene. The sequence length variation in *Lcpψ* in *D. melanogaster* and *D. simulans* also supports the lack of functional constraint on *Lcpψ.* These sequence length variants will not be considered in further analyses of nucleotide diversity. Thus, the number of homologous nucleotides in the *Lcpψ* region excluding insertions and deletions is 533 for the set of 10 *D. melanogaster* genes and 356 when the *D. simulans* sequence is included with the *D. melanogaster* alleles.

The intraspecific sequence comparisons can be used to estimate the neutral model parameter Θ (equal to $4N\mu$, where N is the effective population size and μ is the neutral mutation rate) for this region (NEI and TAJIMA 1981; HUDSON 1982). Two commonly used estimators of Θ include one based on the number of segregating sites ($\hat{\Theta}$), and a second, based on the average number of pairwise differences ($\hat{\pi}$). The calculation of these estimators, along with their standard errors over the coalescent and mutation processes, is described by WATTERSON (1975, equations 1.4a and 1.5a) and NEI (1987, equations 10.6 and 10.9), respectively. These derivations assume the Wright-Fisher population and infinite sites mutation models and no recombination within genes.

The estimate of the neutral mutation parameter for the Winters population of *D. melanogaster,* based on the number of segregating sites in the entire region, is $\hat{\Theta} = 0.001 \pm 0.001$, and from the number of pairwise differences is $\hat{\pi} = 0.001 \pm 0.001$. KREITMAN and HUDSON (1991, equation 3) present a method for estimating a 95% confidence interval on Θ that is consistent with the observed number of segregating sites. Equation 3 of KREITMAN and HUDSON (1991) also allows us to determine whether the observed number of segregating sites in *Lcpψ* is consistent with a particular parametric value of Θ. Using this method, we find that Θ is bounded by 0.000 and 0.005 with 95% confidence. If *Lcpψ* shared the value of $\hat{\Theta} = 0.008$ estimated for the 5′ flanking region of *Adh,* then the probability of observing one or fewer segregating sites in *Lcpψ* would be 0.006 (KREITMAN and HUDSON 1991). Hence, we may conclude that our data are inconsistent with a parametric value of $\hat{\Theta} = 0.008$.

We have also split the *Lcpψ* region into flanking, synonymous, nonsynonymous and intron sites, and examined levels of intraspecific variation in each. Summary data are in Table 2. By inspection, the rates of nonsynonymous, synonymous, intron, and 5′ flanking sequence are equivalent within species. Table 2 also shows the Jukes-Cantor substitution distances at synonymous, nonsynonymous, and intron sites when the *Lcpψ*^{B} and *D. simulans* coding sequences are compared. The number of nucleotide substitutions in a sequence is distributed roughly as a Poisson random variable, which we approximate here using the Normal approximation. Hence, we may test these regions for differences in the rates of nucleotide substitution using the z-test statistic (WHITTAM and NEI 1991). The nonsynonymous substitution rate (0.132 ± 0.134) is not significantly different from either the synonymous rates of coding (0.208 ± 0.215) or flanking (0.347 ± 0.374) sequences when the rates are compared with a z test (nonsynonymous *vs.* synonymous rates, $z = 0.29$, $P = 0.76$; and nonsynonymous *vs.* 5′ flanking rates, $z = 0.54$, $P = 0.59$). The overall nucleotide divergence in this region is 0.149 ± 0.150.

We may also use our data to estimate the relative rates of nucleotide substitution and insertion or deletion in neutral DNA sequences. The sequence *Lcpψ*^{B} differs from the *D. simulans Lcpψ* by 48 nucleotide substitutions, and four insertions plus four deletions. Thus we estimate the rate of nucleotide substitutions to be $48/8 = 6.0$ times higher than that of insertion plus deletion. One method of placing confidence intervals on the relative rates of two Poisson processes is to assume that the ratio has a normal distribution, and use the Delta method to estimate the bias and variance of the ratio. That method is inaccurate unless the observed count for each process is at least moderately large. Instead, we have found that the symmetric 95% confidence interval for the ratio is (2.8–14.7), using the following method (COX and HINKLEY 1974, p. 232). To use this method, we suppose that the true relative rate is r nucleotide substitutions for each insertion or deletion. Then condition on observing $48 + 8 = 56$ total events

```
Lcpψ^A ATACTCGTAG TACTAATCCC GTGGATTGC- ---------- CATG---CGA TGCTTTTTAA  46
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
Lcp1   ...G.AC.TA .C.CGT.GTT TG.C...TGC TAAGCTGTCG ....TGA... ..........
Lcp2   GC.AGAC.T- ----GT.TTT TGAC...TGT ATCGCAGAGT ..A.TGGA.. ATT.A.GGGC

Lcpψ^A TGGGTGT-GG GGTATCTTCA TAGTCAACCC AGATTTCAGA GCACCAATCG TATGAAAGAT 105
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... .C--------
Dsim   ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
Lcp1   .......G.. C.C...CG.G A......... .T.AC....C .A......T. A...C.....
Lcp2   CCT.CC.TTT .T.GG.A... .G.G.GTTT. GTGA.AACTT AG.TTTGG.C C.AA..GT.A

Lcpψ^A GTAGCGTATT GACTTGGGTT CAGTTTTGGT GG-------- ---------- ---------- 137
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C -......... .......... .......... .......... .......... ..........
Dsim   ?????????? ?????????? ?????????? ?????????? ?????????? ??????????
Lcp1   ....A..T.. ..TA...... ..C...G... ..CAATCATA TAAAAAGGCT CTGCCCGACC
Lcp2   TA...AAT.C .TT.G.AAAG ..ACCAAAT. ..GAATCATA TAAAAAGACT CTGT.CGACC

Lcpψ^A --------TT ATCAGTCAAC GTTCGTTCTC GAGCAAACAG AAATTGGAAT CTAAAATGTT 189
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   ?????????? ?????????? ?????????? ?CTT...... ....GT...- ----------
Lcp1   ACAATCAG.. .......... .......... ..C..G.... ..G.CA.CC- --..T.....
Lcp2   AAAGTCAG.. .......... .......... ..C..G.... ....CA.CC- --..C.....

Lcpψ^A CATGCTCGTG AGTAAC-GGC AGGATACGAA C--------- -----GTATT CAATGTCTAT 234
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   ......T... .........A. .......... .......... .....C.... ..........
Lcp1   ..A.T.T..A ...GT..C.. .......... .CAACATACT CGATCCCTAA .G.ATG.CTA
Lcp2   ..A.T.T... ...GG.TCA. ....C.TTT. TGAACTCGC. ..CATC..A. TGG.A..AT.

Lcpψ^A GCTCTCGTTC ACGTGGTGAT CTGTGCAGTT TTTGGATTGA --GTGGCTCA TCCCCCA--T 290
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... TA........ ..........
Dsim   ......C... .G........A ......CC.. ......C..G TG...A.C.. .T.....GTG
Lcp1   TT....C... .G..CA.... ...C...... ..G..CC..G CG.....CA. C.....GGTG
Lcp2   T.CTCTA.C. .G...A.... TCTC..C... G.G...G..G CTACC..C.T AG.....GT.
```

FIGURE 2.—Sequence variation in the pseudogene *Lcpψ*. The sequences of three *Lcpψ* haplotypes from *D. melanogaster* and one from *D. simulans*, along with the functional cuticle genes, *Lcp1* and *Lcp2*, are aligned as in Figure 5 of SNYDER *et al.* (1982) (except for a minor alteration noted in the text). *Lcpψ*^{A} was used as the consensus sequence such that nucleotides that are identical with *Lcpψ*^{A} are indicated with a dot and deletions in a sequence are shown by a dash. The numbers on the right edge of the figure indicate the positions of the aligned nucleotides within the reference sequence *Lcpψ*^{A}.

(either point substitutions or insertions/deletions), the observed number of point substitutions is binomial with mean $56r/(1 + r)$. Given that we observed 48 point substitutions, we can find confidence intervals for r under the binomial distribution.

The Hudson, Kreitman, and Aguadé test of neutrality: We used the statistic suggested by HUDSON, KREITMAN and AGUADÉ (1987) (HKA test) to test for departures from neutral expectations. The neutral theory predicts that intraspecific polymorphism will be proportional to interspecific divergence (KIMURA 1983). The HKA test rejects a neutral model if the ratio of polymorphism to divergence differs significantly among independent loci (HUDSON *et al.* 1987). The *Lcpψ* region was compared to the 5′ flanking region of alcohol dehydrogenase (*Adh*) in *D. melanogaster* (KREITMAN and HUDSON 1991). The within-species polymorphism for *Lcpψ* was estimated from the 10 sequences collected in the Winters population. The number of segregating sites in the 5′ flanking region of *Adh* was estimated from a worldwide sample of 11 alleles. The between-species divergence distances were calculated using comparisons of each locus with its homologue in *D. simulans*. In the case of *Lcpψ*, we used the sequence between nucleotides 160 and 569 in *Lcpψ*^{B} because these were the only bases available in *D. simulans*. In performing this test, we pooled all available sites, including nonsynonymous, synonymous and noncoding because all nucleotides should be neutral sites due to the lack of functional constraint on *Lcpψ*. We used the HKA test with the

```
Lcpψ^A CCCCCATCCG TAGGACGTTT CGAAAGTGAC AACGCCGATG TCCACTACCA ATCATACGAT 350
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... ......C... ..........
Dsim   ....AT...T .........- ---------- ---------- -.....C.G. ..T-G.....
Lcp1   ....AT...C ....C....C G..GGA..T. C......... ...TT.C..G ...CG.T...
Lcp2   ---------- --TCC..C.C ...TGA..TA C....T.... ...TT.C..G ...GG....C

Lcpψ^A GTTTATACCG ATGGATTCGA TTCCAGCAGT CTGCACATAT TCAACGGAAT CGAGCGGGCC 410
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   ...CG.G... .......T.. ......T--- .......... C......... .........T
Lcp1   ...CG.G... .......... .......--- .......CC. C......... .....A....
Lcp2   ...CG.G... .C........ C.......--- .......CC. CA........ .....A....

Lcpψ^A GCCAAAGCAG ATATTGATGG TATTTTTCAG GGCACTATAA AATGGATCAC CCCCGAGGGC 470
Lcpψ^B .......... ....C..... .......... .......... .......... ..........
Lcpψ^C .......... ....C..... .......... .......... .......... ..........
Dsim   .......... ..G.C..... ......C... ..TT....T. .....G.... ..A.......
Lcp1   ....GC.GT. ..GCCC.... C.ACA.C..C ....ACT.CG GC......T. A.........
Lcp2   ....GC.GT. ..GCCC.... C.ACA.C..C ....ACT.CG GC......T. A.........
```

FIGURE 2.—*Continued*

```
Lcpψ^A GATCACGTCG AGGTCACGTA CGACGCTGGT GAGAACGGAT ACCA--GGCT TCGAT--CAG 526
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   .......... .......... ..G....AA. .......... G..TCT...G .....GC...
Lcp1   ..G....... ....T.A... ..T...CAA. .......... .......C.C ...-G..G..
Lcp2   ..G.....T. ....A.A... ..T...GAA. ..A....... .......C.C ...-G..G..

Lcpψ^A CCTTGATTCC GACTCCTGCA CCAATCC--G A-AGGC---- ---------- ---------- 559
Lcpψ^B .......... .......... .......... .......... .......... ..........
Lcpψ^C .......... .......... .......... .......... .......... ..........
Dsim   ..CC...... .......C.. ..T......C GG........ .......... ..........
Lcp1   ...G...C.. C......C.T .......CA. ..G.C.ATCG GCCGCGCCGT CGCCTGGCTA
Lcp2   ...G...C.. C......C.T .......CA. ..G.C.ATCG CCCGCGCCGT TGCCTGGCTG

Lcpψ^A --GTTCCACT GACCGGTACC CGAGCATCCG CACCATCACC AA                    599
Lcpψ^B .......... ..???????? ?????????? ?????????? ??
Lcpψ^C .......... ..???????? ?????????? ?????????? ??
Dsim   .......T.. C..T....T. ...A.????? ?????????? ??
Lcp1   GA..C....C C...A.C... ......C..C .GT......T A.
Lcp2   GA..CT...C C...A.C... ......C..C .GT......T A.
```

modification for differences in sample size suggested by BERRY *et al.* (1991) because the test was done with 10 *Lcpψ* genes and 11 *Adh* genes. The comparison of the *Lcpψ* data with the 5′ flanking region of *Adh* probably reflects the most reasonable test of neutrality because this segment of *Adh* has the least functional constraint on its sequence and probably has not been influenced by positive Darwinian selection in its recent history (KREITMAN and HUDSON 1991). Comparisons of the *Adh* or *Adh-related* coding sequences with the *Lcpψ* region are likely to be inappropriate for the HKA test because these regions are believed to be subject to balancing and directional selection, respectively (KREITMAN and HUDSON 1991).

The observed and expected numbers of segregating and divergent sites used in the HKA test are shown in Table 3. The HKA results suggest that the level of variation observed in *Lcpψ* was significantly less than expected given the ratio of polymorphism and divergence in the 5′ flanking sequence of *Adh* ($\chi^2 = 5.31$, $P = 0.02$). The HKA test did not detect a significant departure from neutrality when either synonymous or nonsynonymous sites were examined separately (data not shown).

DISCUSSION

Several lines of evidence suggest that *Lcpψ* is not a functional gene. A number of these were listed in the Introduction. In addition, we have found that *Lcpψ* has

TABLE 1

Insertion and deletions in the coding region of *Lcpψ*

Event[a]	Δ2 bp	Δ2 bp	Δ22 bp	Δ1 bp	∇3 bp	∇2 bp	∇1 bp	∇2 bp	Δ2 bp	∇1 bp	Δ26 bp
Type[b]	Pm	Fm	Fs	Fs	Fm	Fs	Fms	Fs	Fms	Fs	Fms
Position[c]	274	289	310	344	388	514	522	523	553	555	559
Lcpψ^{A}	+	+	−	−	+	−	+	−	+	−	+
Lcpψ^{B}	+	+	−	−	+	−	+	−	+	−	+
Lcpψ^{C}	−	+	−	−	+	−	+	−	+	−	+
Lcpψ sim	−	−	+	+	−	+	+	+	+	+	+
Lcp1	−	−	−	−	−	−	−	−	−	−	−
Lcp2	−	−	−	−	−	−	−	−	−	−	−

[a] Event indicates whether the sequence length variant was inferred to be an insertion (∇) or deletion (Δ) based on comparisons with the *Lcp1* and *Lcp2* sequences. The size of the insertion or deletion is indicated in bp.

[b] Type indicates whether the indel is polymorphic within *D. melanogaster* (Pm), fixed in *D. melanogaster* (Fm), fixed in *D. simulans*, or fixed prior to the divergence of *D. melanogaster* and *D. simulans* (Fms).

[c] Position indicates the location in *Lcpψ*^{A} where the insertion or deletion occurred. The nucleotide numbers are the same as in Figure 2.

accumulated 11 sequence length variants since gene inactivation (Table 1). We observed six deletions and five insertions in *Lcpψ* in *D. melanogaster* or *D. simulans*, relative to the parental genes *Lcp1* and *Lcp2*. These length variants result in different premature stop codons (see RESULTS). The divergence of *Lcpψ* from its two parental genes *Lcp1* and *Lcp2* also argues that there has been an acceleration in the nonsynonymous substitution rates in *Lcpψ* (Table 4, data analysis not shown).

The patterns of interspecific divergence that we observed in *Lcpψ* are consistent with neutral theory predictions for pseudogenes. First, synonymous and nonsynonymous substitution rates were statistically equivalent when *Lcpψ* was compared between (or within) species. This is consistent with the loss of selective constraints on nonsynonymous sites associated with the gene inactivation event.

Second, interspecific divergence between *D. melanogaster* and *D. simulans* is greater than for any other gene that has been sequenced in the two species (Table 2). The divergence at synonymous sites in *Lcpψ* (Jukes-Cantor distance = 0.208) is greater than for other comparisons of *D. melanogaster* and *D. simulans* genes (BEGUN and AQUADRO 1992), which vary from 0.046 in myosin light-chain gene 1 (LEICHT *et al.* 1993) to 0.167 in accessory protein 26Aa (AGUADÉ *et al.* 1992). Overall Jukes-Cantor divergence (0.149) is also greater in *Lcpψ* than in other genes, for which divergence varies from a low of 0.023 in heat shock protein 70A (LEIGH BROWN and ISH-HOROWICZ 1981; LEIGH BROWN 1983) to 0.072 in metallothionein (LANGE *et al.* 1990).

Thus, the pattern of nucleotide substitution between homologues and paralogues of *Lcpψ* indicates a lack of functional constraint and a a rapid rate of neutral evolution of the pseudogene sequences of both *D. mela-*

TABLE 2

Within-species variability and between-species divergence (Jukes-Cantor distances) for different kinds of sites

Region	Nuc[a]	Within-species variability ($\hat{\pi}$)[b]
Within *D. melanogaster*		
5′ flanking	174.0	0.000 ± 0.000
Introl	50.0	0.000 ± 0.000
Exons 1 and 2, synonymous	70.5	0.000 ± 0.000
Exons 1 and 2, nonsynonymous	238.5	0.002 ± 0.002
Synonymous and noncoding sites	294.5	0.000 ± 0.000
All sites	533.0	0.001 ± 0.001
Between *D. melanogaster* and *D. simulans*		
5′ flanking	18.0	0.347 ± 0.374
Intron	50.0	0.085 ± 0.094
Exons 1 and 2, synonymous	68.8	0.208 ± 0.215
Exons 1 and 2, nonsynonymous	219.2	0.132 ± 0.134
Synonymous and noncoding sites	136.8	0.176 ± 0.180
All sites	356.0	0.149 ± 0.150

[a] Nuc is the number of nucleotides within a region.

[b] The estimates of $\hat{\pi}$ within *D. melanogaster* are based on the strains collected from Winters, California and do not include the SNYDER *et al.* (1982) sequence. The standard errors (±SE) are based on the sum of the sampling and stochastic variances. The *Lcpψ*^{B} sequence was used in the comparison of *D. melanogaster* and *D. simulans* sequences.

TABLE 3

HKA test for the *Lcpψ* gene

Region	S^a	$E(S)$	D^b	$E(D)$
Lcpψ	1	9.49	48	39.51
5′ flanking *Adh*	30	21.51	78	86.49

[a] S is the number of segregating sites within *D. melanogaster*.

[b] D is the number of divergent sites between *Lcpψ*^{B} and the *D. simulans* sequence.

TABLE 4

Jukes-Cantor distances at intron, synonymous, and nonsynonymous sites

Comparison	Intron	Synonymous	Nonsynonymous
Lcpψ^A and *Lcpψ D. simulans*	0.085 ± 0.094	0.227 ± 0.235	0.132 ± 0.235
Lcp1 and *Lcpψ*^{A}	0.572 ± 0.582	0.451 ± 0.458	0.253 ± 0.255
Lcp1 and *Lcpψ D. simulans*	0.490 ± 0.500	0.485 ± 0.492	0.245 ± 0.492
Lcp2 and *Lcpψ*^{A}	0.824 ± 0.834	0.665 ± 0.673	0.325 ± 0.328
Lcp2 and *Lcpψ D. simulans*	0.766 ± 0.776	0.659 ± 0.666	0.314 ± 0.316
Lcp1 and *Lcp2*	0.902 ± 0.910	0.325 ± 0.331	0.054 ± 0.056

The standard errors (±SE) are based on the sum of the sampling and stochastic variances.

nogaster and *D. simulans* following the inactivation of *Lcpψ* in a common ancestor of the two species.

While the pattern of *Lcpψ* evolution is consistent with a lack of functional constraints on gene sequence, the estimate of nucleotide diversity in *Lcpψ* of 0.001 is among the lowest values found in *D. melanogaster* (MORIYAMA and POWELL 1996), which vary for sequenced coding regions from 0.000 in *cubitus interruptus Dominant* (BERRY *et al.* 1991; WAYNE and KREITMAN 1996) to 0.010 in the *proximal amylase* gene (INOMATA *et al.* 1995). Furthermore, the HKA test shows that the number of segregating sites in *Lcpψ* is lower than expected given the level of divergence between *D. melanogaster* and *D. simulans* (Table 3). We are now left to explain why *Lcpψ* has low levels of polymorphism despite the evidence that its function has been lost and that it has a correspondingly high rate of substitution between species.

Because 10 of the sequences came from a single population in California, it might be suggested that the low level of variation is due to inbreeding in the California population. If this were the case, however, one might expect that the published (Canton-S) sequence would be quite divergent from the California flies. Instead the published sequence differs from seven of these lines at just one nucleotide site. Furthermore, the study of Clark (1989) examined the variation of metabolic enzyme activities in these lines (and other lines from the site) and found considerable heritable variation.

As a hint to a possible explanation for the lack of nucleotide diversity in *Lcpψ*, we point out that *Lcpψ* lies in a region that is subject to fairly low rates of recombination. The local recombination rate that *Lcpψ* experiences may be expressed by the coefficient of exchange (LINDSLEY and SANDLER 1977). The coefficient of exchange is the frequency of recombination events per cytological band. The coefficient of exchange estimated for the region 44D, the cytological location of the *Lcp* gene family, is 0.0147 ± 0.0005, while the coefficients of exchange determined for the second chromosome range from 0.0032 to 0.0606. By contrast, the *Adh* 5′ flanking region, which we used for comparison in the HKA test, lies in a region of significantly higher recombination, with a coefficient of exchange almost three times higher than at *Lcpψ* (5′ *Adh* COE = 0.0419 ± 0.0027; KINDAHL 1994).

Two evolutionary mechanisms acting in regions of low recombination may be used to explain the paucity of genetic diversity in *Lcpψ* within *D. melanogaster:* background selection or genetic hitchhiking. Background selection reduces nucleotide variation in regions of low recombination as a result of selection against linked mildly deleterious mutations (CHARLESWORTH *et al.* 1993; HUDSON and KAPLAN 1995; CHARLESWORTH 1996). Alternatively, genetic hitchhiking can lower levels of diversity during selective fixation of advantageous mutations (MAYNARD SMITH and HAIGH 1974; KAPLAN *et al.* 1989; STEPHAN *et al.* 1992; WIEHE and STEPHAN 1993). Typical models of these processes study genetic variation at a neutral locus surrounded by loci that can mutate to either advantageous [*e.g.,* WIEHE and STEPHAN (1993)] or deleterious [*e.g.,* HUDSON and KAPLAN (1995)] forms. In both cases, it is generally assumed that the frequency of selected mutations is constant per unit of physical distance: hence, the number of selected loci within some linkage distance of the neutral locus is inversely related to the local recombination rate.

The earliest background selection models were not adequate to explain the reductions in genetic diversity in regions of low recombination (CHARLESWORTH *et al.* 1993; CHARLESWORTH 1994) because they only focused on slightly deleterious mutations resulting from point mutations. CHARLESWORTH (1996) has recently demonstrated that background selection is sufficient to explain the reduction of genetic variation for most *D. melanogaster* genes when conventional mutations and transposable elements are the source of deleterious mutations. Figure 3 in CHARLESWORTH (1996) shows that the level of variation in *Lcpψ* falls close to the theoretical levels of heterozygosity predicted under a background selection model. Background selection could only be ruled out if it were found that the Winters population was the only population with a reduction in diversity at *Lcpψ* (CHARLESWORTH 1996, page 143 last paragraph).

The *Lcpψ* divergence rate at synonymous sites between *D. melanogaster* and *D. simulans* provides an opportunity to estimate the selection parameter, $4N_e s$, from the ratio of selected to neutral fixation probabilities,

$$\frac{u}{u_0} = \frac{s}{(1 - e^{-s})}$$

TABLE 5

Estimates of $4N_es$ for synonymous sites determined from the ratio of fixation probabilities in functional genes to *Lcpψ*

Gene	ds	Ratio	$4Ns$	Citation
ac	0.114	0.548	−1.11	MARTIN-CAMPOS *et al.* (1992)
Acp26Aa	0.167	0.803	−0.43	AGUADÉ *et al.* (1992)
Acp26Ab	0.059	0.284	−2.15	AGUADÉ *et al.* (1992)
Adh	0.052	0.250	−2.34	KRIETMAN (1991)
Adhr	0.133	0.639	−0.84	KREITMAN (1991)
Amy-d	0.098	0.471	−1.36	INOMATA *et al.* (1995)
Amy-p	0.108	0.519	−1.20	INOMATA *et al.* (1995)
ase	0.059	0.284	−2.15	HILTON *et al.* (1994)
cta	0.056	0.269	−2.23	WAYNE and KREITMAN (1996)
dpp	0.036	0.173	−2.87	NEWFELD *et al.* (1997)
Est-6	0.129	0.620	−0.89	KAROTAM *et al.* (1995)
Mlc1	0.046	0.221	−2.52	LEICHT *et al.* (1993)
Pgd	0.148	0.712	−0.65	BEGUN and AQUADRO (1994)
Pgi	0.072	0.346	−1.85	J. H. MCDONALD and M. KREITMAN, personal communication
pn	0.093	0.447	−1.44	SIMMONS *et al.* (1994)
ref(2)P	0.063	0.303	−2.06	WAYNE *et al.* (1996)
Rh3	0.093	0.447	−1.44	AYALA (1993)
Sod	0.096	0.462	−1.39	HUDSON *et al.* (1994)
su(f)	0.102	0.490	−1.29	LANGLEY *et al.* (1993)
tra	0.143	0.688	−0.71	WALTHOUR and SCHAEFFER (1994)
v	0.150	0.721	−0.63	BEGUN and AQUADRO (1995)
w	0.111	0.534	−1.15	KIRBY and STEPHAN (1996)
Zw	0.105	0.505	−1.25	EANES *et al.* (1993)
Average			−1.45	

where u is the fixation probability for a selected locus, u_0 is the fixation probability for a strictly neutral locus, and S is $4N_es$ (KIMURA 1983, equation 3.14). If we assume that all mutations at *Lcpψ* are neutral, and that the total mutation rate is constant across genes, then we can estimate u/u_0 by the ratio of d/d_0 where d and d_0 are the silent site divergence rates at the selected locus and the neutral locus, respectively. We have performed these calculations using the silent site substitution rates of 22 functional genes for which the sequence of the entire coding region is available in both *D. melanogaster* and *D. simulans* (Table 5). The average value of $4N_es$ is −1.45 suggesting that most synonymous changes are slightly deleterious, consistent with the "major codon preference" model (SHARP and LI 1986; LI 1987; BULMER 1988). Under this model, a single synonymous codon tends to be selected for either because it increases translational efficiency or accuracy. Thus, selection will disfavor synonymous mutations that alter major codons of genes with high to moderate codon bias (AKASHI 1994, 1995; AKASHI and SCHAEFFER 1997). These results are consistent with the estimates of $4N_es$ that AKASHI (1995) obtained for unpreferred synonymous changes using ratios of polymorphism to divergence at *per* and *Pgi* in *D. melanogaster* and *D. simulans*.

Our data also provide some insight into the generation of length variation in neutral DNA. As noted above (Table 1) there are six deletions and five insertions in *Lcpψ* relative to the parental genes *Lcp1* and *Lcp2*. The deletions and insertions had average lengths of 9.2 and 1.8 bases, respectively. The average loss of DNA from *Lcpψ* is consistent with the hypothesis that gene inactivation generally leads to a reduction in gene length (GRAUR *et al.* 1989; GU and LI 1995; PETROV *et al.* 1996; SAITOU and UEDA 1994), however, most of the net loss in sequence length resulted from just two large deletion events, of 22 and 26 bp, respectively.

PETROV *et al.* (1996) studied length variation in inserts of the retrotransposon *Helena* in Drosophila. They found a striking pattern of evolution toward shorter sequence length (23 deletions and no insertions). They suggest that such a process may be generally active in Drosophila pseudogenes. However, we observed a very different pattern of length variation, in which insertions were relatively common, albeit short in length (1–3 bp). We can use two methods to classify our length variation as insertions or deletions: (1) assuming that all length variation arose in the *Lcpψ* lineage, as above, or (2) using a parsimony criterion. Method 1 produces an estimate of six deletions and five insertions; method 2 is somewhat stricter and leads to an estimate of four insertions, four deletions, and three unclassified length variants. We have compared our results to those of PETROV *et al.* (1996), using a 2 × 2 contingency table of insertions and deletions, in *Lcpψ* and *Helena*. Fisher's exact test shows that the rows and columns of this table are not independent ($P < 0.01$, using either method of classification of the length variation at *Lcpψ*). This result implies that insertions are significantly more fre-

quent at *Lcpψ* than at *Helena*. The reason for this difference is not clear, however it argues that *Helena* is not a good model for the evolution of all Drosophila pseudogenes.

In summary, we have examined intra- and interspecific diversity in nucleotide sequences in the *Lcpψ* pseudogene of *D. melanogaster*. The interspecific fixation rates at synonymous sites of *Lcpψ* were used to show that on average, silent mutations in functional genes are slightly deleterious. The intraspecific nucleotide diversity at *Lcpψ* appears to be at odds with an abundance of evidence indicating a lack of selective constraint at this locus. Under a strict neutral model, it is expected that within-species variation should be proportional to between-species divergence. Using the HKA test however, we have shown a significant departure from this prediction. Such an uncoupling of polymorphism and substitution rate seems most likely to be the result of linked selection against slightly deleterious alleles.

We thank C. F. AQUADRO, D. J. BEGUN, C. T. BERGSTROM, J. BRAVERMAN, B. CHARLESWORTH, D. CHARLESWORTH, A. G. CLARK, M. W. FELDMAN, M. M. TANAKA and W. STEPHAN for helpful comments and discussions used in the preparation of this manuscript. We also thank J. H. MCDONALD and M. KREITMAN for use of their unpublished *Pgi* data and Dr. R. K. SELANDER for the use of the Applied Biosystems 373A DNA Sequencer, and S. A. PLOCK and S. L. O'BRYAN for technical assistance with the *D. simulans* sequence. This work was supported in part by several grants: National Institutes of Health (NIH) grant GM-42472 to S. W. S., Howard Hughes Predoctoral Fellowship to J. K. P., and NIH grant GM-28016 to M. W. FELDMAN.

LITERATURE CITED

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., 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., and S. W. SCHAEFFER, 1997 Natural selection and the frequency distributions of "silent" polymorphism in Drosophila. Genetics **146:** 295–307.

ASHBURNER, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

AYALA, F. J., 1993 Molecular evolution of the *Rh3* gene in *Drosophila.* Genetica **92:** 23–32.

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. J., and C. F. AQUADRO, 1994 Evolutionary inferences from DNA variation at the 6-phosphogluconate dehydrogenase locus in natural populations of Drosophila: selection and geographic differentiation. Genetics **136:** 155–171.

BEGUN, D. J., and C. F. AQUADRO, 1995 Molecular variation at the *vermillion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans.* Genetics **140:** 1019–1032.

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.

BINGHAM, P. M., R. LEVIS and G. M. RUBIN, 1981 Cloning of DNA sequences from the *white* locus of D. melanogaster by a novel and general method. Cell **25:** 693–704.

BULMER, M., 1988 Are codon usage patterns in unicellular organisms determined by selection-mutation balance. J. Evol. Biol. **1:** 15–26.

CABOT, E. L., and A. T. BECKENBACH, 1989 Simultaneous editing of multiple nucleic acid and protein sequences with ESEE. Comput. Appl. Biosci. **5:** 233–234.

CHARLESWORTH, B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. Camb. **63:** 213–227.

CHARLESWORTH, B., 1996 Background selection and patterns of genetic diversity in *Drosophila melanogaster.* Genet. Res. Camb. **68:** 131–149.

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

CLARK, A. G., 1989 Causes and consequences of variation in energy storage in *Drosophila melanogaster.* Genetics **123:** 131–144.

COX, D. R., and D. V. HINKLEY, 1974 *Theoretical Statistics.* Chapman and Hall, New York.

EANES, W. F., M. KIRCHNER and J. YOON, 1993 Evidence for adaptive evolution of the G6pd gene in the *Drosophila melanogaster* and *Drosophila simulans* lineages. Proc. Natl. Acad. Sci. USA **90:** 7475–7479.

EISSENBERG, J. C., and S. C. R. ELGIN, 1983 Analysis of DNA structural patterns and sequence organization at the larval cuticle locus in *Drosophila melanogaster.* Mol. Cell. Biol. **3:** 1724–1729.

GOJOBORI, T., W-H. LI and D. GRAUR, 1982 Patterns of nucleotide substitution in pseudogenes and functional genes. J. Mol. Evol. **18:** 360–369.

GRAUR, D., Y. SHUALI and W-H. LI, 1989 Deletions in processed pseudogenes accumulate faster in rodents than in humans. J. Mol. Evol. **28:** 279–285.

GU, X., and W-H. LI, 1995 The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. J. Mol. Evol. **40:** 464–473.

HIGUCHI, R. G., and H. OCHMAN, 1989 Production of single-stranded DNA templates by exonuclease digestion following the polymerase chain reaction. Nucleic Acids Res. **17:** 5865.

HILTON, H., R. M. KLIMAN and J. HEY, 1994 Using hitchhiking genes to study adaptation and divergence during speciation within the *Drosophila melanogaster* species complex. Evolution **48:** 1900–1913.

HUDSON, R. R., 1982 Estimating genetic variability with restriction endonucleases. Genetics **100:** 711–719.

HUDSON, R. R., K. BAILEY, D. SKARECKY, J. KWIATOWSKI and F. J. AYALA, 1994 Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster.* Genetics **136:** 1329–1340.

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

INOMATA, N., H. SHIBATA, E. OKUYAMA and T. YAMAZAKI, 1995 Evolutionary relationships and sequence variation of α-amylase variants encoded by duplicated genes in the *Amy* locus of *Drosophila melanogaster.* Genetics **141:** 237–244.

KAPLAN, N. L., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

KAROTAM, J., T. M. BOYCE and J. G. OAKESHOTT, 1995 Nucleotide variation at the hypervariable esterase 6 isozyme locus of *Drosophila simulans.* Mol. Biol. Evol. **12:** 113–122.

KIMURA, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, New York.

KINDAHL, E. C., 1994 Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster.* Ph.D. dissertation, Cornell University, Ithaca, NY.

KIRBY, D. A., and W. STEPHAN, 1996 Multi-locus selection and the structure of variation at the white gene of Drosophila melanogaster. Genetics **144:** 635–45.

KREITMAN, M., 1991 Detecting selection at the level of DNA, pp. 204–221 in *Evolution at the Molecular Level,* edited by R. K. SELANDER, A. G. CLARK and T. S. WHITTAM. Sinauer, Sunderland, MA.

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.

KUMAR, S., K. TAMURA and M. NEI, 1994 MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput. Appl. Biosci. **10:** 189–191.

LANGE, B. W., C. H. LANGLEY and W. STEPHAN, 1990 Molecular evolution of Drosophila metallothionein genes. Genetics **126:** 921–932.

LANGLEY, C. H., J. MACDONALD, 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.

LEICHT, B. G., E. M. S. LYCKEGAARD, C. M. BENEDICT and A. G. CLARK, 1993 Conservation of alternative splicing and genomic organization of the myosin alkali light-chain (*Mlc1*) gene among *Drosophila* species. Mol. Biol. Evol. **10:** 769–790.

LEIGH BROWN, A. J., 1983 Variation at the 87A heat shock locus in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **80:** 5350–5354.

LEIGH BROWN, A. J., and D. ISH-HOROWICZ, 1981 Evolution of the 87A and 87C heat-shock loci in *Drosophila*. Nature **290:** 677–682.

LI, W-H., 1981 Pseudogenes as a paradigm of neutral evolution. Nature **292:** 237–239.

LI, W-H., 1983 Evolution of duplicate genes and pseudogenes, pp. 14–37 in *Evolution of Genes and Proteins,* edited by M. NEI and R. K. KOEHN. Sinauer Associates, Sunderland, MA.

LI, W-H., 1987 Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. J. Mol. Evol. **24:** 337–345.

LINDSLEY, D. L., and L. SANDLER, 1977 The genetic analysis of meiosis in female *Drosophila melanogaster*. Philos. Trans. R. Soc. Lond. Biol. **277B:** 295–312.

MARTIN-CAMPOS, J. M., J. M. COMERON, N. MIYASHITA and M. AGUADÉ, 1992 Intraspecific and interspecific variation at the *y-ac-sc* region of *Drosophila simulans* and *Drosophila melanogaster*. Genetics **130:** 805–816.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. Camb. **23:** 23–35.

MIYATA, T., and H. HAYASHITA, 1981 Extraordinary high evolutionary rate of pseudogenes: evidence for the presence of selective pressure against changes between synonymous codons. Proc. Natl. Acad. Sci. USA **78:** 5739–5743.

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

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

NEI, M., and F. TAJIMA, 1981 DNA polymorphism detectable by restriction endonucleases. Genetics **97:** 145–163.

NEWFELD, S. J., R. W. PADGET, S. D. FINDLEY, B. G. RICHTER, M. SANICOLA, 1997 Molecular evolution at the decapentaplegic locus in Drosophila [In Process Citation]. Genetics **145:** 297–309.

PETROV, D. A., E. R. LOZOVSKAYA and D. L. HARTL, 1996 High intrinsic rate of DNA loss in *Drosophila*. Nature **384:** 346–349.

ROZAS, J., and R. ROZAS, 1997 DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Appl. Biosci. **13:** 307–311.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI, 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science **239:** 487–491.

SAITOU, N., and S. UEDA, 1994 Evolutionary rates of insertion and deletion in noncoding nucleotide sequences of primates. Mol. Biol. Evol. **11:** 504–512.

SANGER, F., S. NICKEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5467.

SHARP, P. M., and W.-H. LI, 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. **24:** 28–38.

SIMMONS, G. M., W. KWOK, P. MATULONIS and T. VENKATESH, 1994 Polymorphism and divergence at the *prune* locus in *Drosophila melanogaster*. Mol. Biol. Evol. **11:** 666–671.

SNYDER, M., M. HUNKAPILLER, D. YUEN, D. SILVERT, J. FRISTROM, 1982 Cuticle protein genes of Drosophila: structure, organization and evolution of four clustered genes. Cell **29:** 1027–1040.

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. Theoret. Popul. Biol. **41:** 237–254.

WALTHOUR, C. S., and S. W. SCHAEFFER, 1994 Molecular population genetics of sex determination genes: the *transformer* gene of *Drosophila melanogaster*. Genetics **136:** 1367–1372.

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

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.

WAYNE, M. L., and M. KREITMAN, 1996 Reduced variation at *concertina*, a heterochromatic locus in *Drosophila*. Genet. Res. Camb. **68:** 101–108.

WHITTAM, T. S., and M. NEI, 1991 Neutral mutation hypothesis test. Nature **354:** 115–116.

WIEHE, T. H. E., and W. STEPHAN, 1993 Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol. Biol. Evol. **10:** 842–855.

Communicating editor: D. CHARLESWORTH