DNA Sequence Variation at the *period* Locus Within and Among Species of the *Drosophila melanogaster* Complex

Richard M. Kliman and Jody Hey

Department of Biological Sciences, Nelson Laboratories, Rutgers University, Piscataway, New Jersey 08855-1059

Manuscript received May 11, 1992
Accepted for publication October 19, 1992

ABSTRACT

A 1.9-kilobase region of the *period* locus was sequenced in six individuals of *Drosophila melanogaster* and from six individuals of each of three sibling species: *Drosophila simulans, Drosophila sechellia* and *Drosophila mauritiana*. Extensive genealogical analysis of 174 polymorphic sites reveals a complex history. It appears that *D. simulans*, as a large population still segregating very old lineages, gave rise to the island species *D. mauritiana* and *D. sechellia*. Rather than considering these speciation events as having produced "sister" taxa, it seems more appropriate to consider *D. simulans* a parent species to *D. sechellia* and *D. mauritiana*. The order, in time, of these two phylogenetic events remains unclear. *D. mauritiana* supports a large number of polymorphisms, many of which are shared with *D. simulans*, and so appears to have begun and persisted as a large population. In contrast, *D. sechellia* has very little variation and seems to have experienced a severe population bottleneck. Alternatively, the low variation in *D. sechellia* could be due to recent directional selection and genetic hitchhiking at or near the *per* locus.

EXPERIMENTAL approaches to the study of evolution are traditionally divided into those that address microevolutionary topics (*i.e.*, forces that determine patterns of genetic variation within species) and those that deal with macroevolutionary issues (*i.e.*, speciation processes and other determinants of phylogenetic patterns). Within recently formed species, however, DNA sequence variation may reflect not only processes acting since divergence, but also processes that acted prior to and during speciation. For example, if new species arise via founder events of very few individuals (*e.g.*, MAYR 1954), rapid genetic drift is expected to lead to low levels of neutral or nearly neutral genetic variation. Alternatively, if speciation events are not preferentially associated with times of very small population size, then recently diverged species may share many polymorphisms.

The species of the *Drosophila melanogaster* complex are closely related and afford an excellent opportunity for genetic studies of recent speciation events. Two of the species, *D. melanogaster* and *Drosophila simulans*, are cosmopolitan, while the other two, *Drosophila mauritiana* and *Drosophila sechellia*, are endemic to single island groups (LACHAISE *et al.* 1988). Phylogenetic studies on a wide array of morphological and genetic characters have revealed only that *D. melanogaster* is a sister taxon to the other species. *D. simulans, D. sechellia* and *D. mauritiana* are very similar to one another and, despite considerable effort, a bifurcating phylogeny for these species has not been unambiguously determined (BODMER and ASHBURNER 1984; COHN, THOMPSON and MOORE 1984; COYNE and KREITMAN 1986; LACHAISE *et al.* 1988; CACCONE, AMATO and POWELL 1988). Together these four species represent three speciation events, two of which appear to have occurred very recently and at similar times.

We sampled a 1.9-kilobase pair sequence of the X-linked *period* locus (*per*) from six individuals of each species in the *D. melanogaster* complex. Like the earlier studies, our genealogical analyses support the separation of *D. melanogaster* from the other species. However, the history of the *per* locus within *D. simulans, D. sechellia* and *D. mauritiana* looks complex and does not support a bifurcating phylogeny. It appears that *D. sechellia* and *D. mauritiana* have arisen independently from a large ancestral *D. simulans* population, and that present day *D. simulans* segregates polymorphisms that persist since before the formation of the island species. In addition, the island species appear to have different histories. *D. sechellia* has diverged from *D. simulans* by the accumulation of many fixed differences, while *D. mauritiana* has diverged by the accumulation of many new polymorphisms. Further, *D. mauritiana* appears as variable as *D. simulans*, indicating that divergence of the former from the latter was not accompanied by a population bottleneck.

The sequence data presented in this article have been submitted to the EMBL/GenBank Data Libraries under the accession numbers L07811–L07832.

MATERIALS AND METHODS

Sources of flies: All strains used here began as isofemale lines. For each species and strain, the site and date of capture, the trapper, and the supplier (in parentheses) if different than the trapper are as follows: *D. melanogaster*: ME-NJ1, ME-NJ2, Terhunes Farm, New Jersey, 10/87, M. KREITMAN; ME-K1, ME-K2, Impala, Kenya 8/89, K. ARDLEY (via M. KREITMAN); ME-LI1, ME-LI2, Davis Peach Farm, Mt. Sinai, New York, 1989, W. EANES; *D. simulans*: SI-LI1, SI-LI2, Davis Peach Farm, Mt. Sinai, New York, 1989, W. EANES; SI-CA1, SI-CA2, Soda Lake, California, fall 1989, S. BRYANT (via D. BEGUN); SI-K1, SI-K2, Impala, Kenya, 8/89, K. ARDLEY (via M. KREITMAN); *D. mauritiana*: MA-1, MA-2, MA-3, MA-4, MA-5, MA-6, Mauritius (main island) 1981, O. KITAGANA (via J. COYNE); *D. sechellia*: SE-C1, SE-C2, Cousin I., Seychelles, 1/85, J. DAVID (via J. COYNE); SE-P1, SE-P2, SE-P3, SE-P4, Praslim I., Seychelles, 7/87, Y. FUYAMA (via K. KIMURA).

DNA preparation: Single *X* chromosomes were extracted from *D. melanogaster* and *D. simulans* by crossing single males with multiple attached-*X* females of the same species. Similarly, the *X* chromosomes of single males in strains SE-P1, SE-P2 and SE-C1 of *D. sechellia* and strains MA-1, MA-2, MA-3 and MA-4 of *D. mauritiana* were extracted by crossing each male with several *D. simulans* attached-*X* females for one generation. Since all males derived from these crosses are isogenic with respect to the *X* chromosome, DNA was extracted from multiple male offspring, following the protocol of MCGINNIS, SHERMOEN and BECKENDORF (1983). We extracted DNA from single males of the remaining strains of *D. sechellia* and *D. mauritiana* using the protocol of ASHBURNER (1989).

Polymerase chain reaction (PCR) and sequencing: An approximately 2.1 kilobase pair region of the *per* locus was PCR-amplified using oligonucleotide primers corresponding to bases 2822–2841 and 4856–4875 of the published Oregon R sequence (CITRI *et al.* 1987) (see Figure 1a). PCR was performed in 50 μl reactions using one unit of *AmpliTaq* DNA polymerase (Perkin-Elmer Cetus) with 1–1.5 mM magnesium ion concentration. The resulting fragment was isolated from a 1% agarose TBE (89 mM Tris base, 89 mM boric acid, 2 mM EDTA) [Gel Using The GeneClean II Kit (Bio101), and Eluted With] 50 μL TE (10 mM Tris HCl, pH 8.0, 1 mM EDTA). PCR was repeated using the original PCR product as a template and with one primer phosphorylated by T4 polynucleotide kinase (BRL) (HIGUCHI and OCHMAN 1989). The products of four identical PCR reactions were combined, chloroform extracted, and subsequently purified and concentrated using Ultrafree-MC filters (Millipore). The product was then made single-stranded by treatment with 4 units of λ-exonuclease (BRL), purified again using the Ultrafree-MC filter, and concentrated to 65 μl total volume. Direct sequencing of single-stranded DNA via the chain-termination method (SANGER, NICKLEN and COULSON 1977) was performed on both strands using the modified T7 DNA polymerase (U.S. Biochemical Corp.). Six internal sequencing primers ranging from 16–18 bases in length were chosen for each strand from regions of complete sequence identity among the 24 lines studied. Sequencing reactions were generally split between two 6% polyacrylamide/8 M urea gels (60 cm × 30 cm × 0.4 mm), and run for either 5.5 hr with a NaOAc gradient (SHEEN and SEED 1988) or for 14 hr without a gradient.

RESULTS

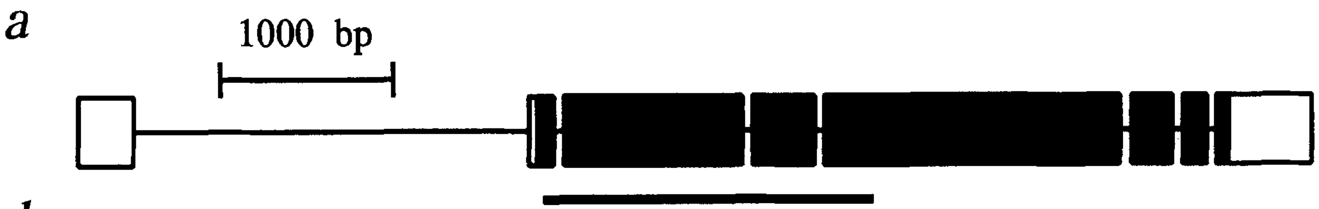
DNA sequence variation summary: A schematic of the *per* locus and a single sequence are presented in Figure 1. The length of the sequenced region varies because of insertion/deletion differences, but includes an average of 190 bases evenly spread among three introns and an average of 1679 bases spread among 4 exons. In *D. sechellia*, sequences SE-P1, SE-P2 and SE-P3 were identical. A total of 174 polymorphic sites were observed across all species, identified as 115 synonymous sites (*i.e.*, not resulting in amino acid substitution), 12 replacement sites (*i.e.*, resulting in an amino acid change), 40 single base differences in introns, two length polymorphisms within exons, and five length polymorphisms within introns. The locations of polymorphisms within species and gene regions are shown in Figure 2, while a breakdown of polymorphism by gene region is shown in Figure 3. Amino acid replacement polymorphisms occurred only in exon 3, the largest exon we sequenced. Also, a region of fixed interspecies sequence length differences occurs in exon 3 (sites 465–470 in Figure 2), with four consecutive glutamate residues in *D. melanogaster* (sites 459–470 in Figure 1b), three in *D. mauritiana* and *D. simulans*, and two in *D. sechellia*.

Patterns of intraspecific variation: The amount of variation within a species, when considered under a neutral model, is a function of the population size N and the mutation rate μ. A commonly used parameter in models of DNA sequence variation is $\Theta = 4N\mu$ (for a sex-linked locus like *per*, $3N\mu$ is actually more appropriate). If we assume a constant population size and that all mutations are neutral and occur under an infinite sites model (KIMURA 1969), then there are two straightforward ways to estimate Θ. For a species with S polymorphic sites and a sample size of n, Θ can be estimated with the relation

$$\hat{\Theta} = S/\sum_{i=1}^{n-1} (1/i). \tag{1}$$

(WATTERSON 1975). Alternatively, Θ can be estimated by calculating the average number of nucleotide differences observed among all possible pairs of sequences (NEI and LI 1979).

We have observed differences in polymorphism levels among species. Sequences from *D. simulans* and *D. mauritiana* revealed nearly twice as many polymorphic sites as found in *D. melanogaster* and over 10 times as many as seen in *D. sechellia* (Table 1). Figure 4 shows estimates of Θ along with their 95% confidence limits, based on the number of polymorphic sites (see KREITMAN and HUDSON 1991). Although there is some overlap in the confidence intervals of *D. sechellia* and *D. melanogaster*, it seems clear that *D. sechellia* is less variable than either *D. mauritiana* or *D. simulans*. This result is in agreement with estimates of allozyme variation in these species (CARIOU *et al.* 1990). The same pattern is seen when Θ is estimated from the average



b

```
   1  CAAGGTGTCCGACTCGGCCTACTCGAACAGCTGCAGCAAC AGCCAGTCGCAGCGCAGgtgagtccatttgatgtgctcaa tcatgcg...agtagtctaaacgat.......ccatcccc
 121  tttccagTGGCAGCTCCAAATCCCGCCTGAGCGGCAGCCA CTCCTCCGGCAGCAGTGGCTATGGGGGCAAGCCCTCGACG CAGGCCAGCAGCAGCGACATGATCATCAAGCGCAACAAGG
 241  ATAAGTCGCGCAAGAAGAAGAAGAACAAGGGCGCCGGCCA GGGGGCGGGGCAGGCGCAGACCCTGATCTCGGCGTCGACG AGCCTAGAGGGCAGGGACGAGGAGAAGCCGCGGCCCAGCG
 361  GCACTGGGTGCGTAGAACAGCAGATCTGCCGAGAGCTGCA GGATCAACAGCACGGCGAGGATCACAGCGAGCCGCAGGCC ATAGAGCAGCTGCAGCAGGAGGAGGAGGAGGATCAGTCCG
 481  GCTCCGAGTCGGAGGCGGACCGAGTCGAGGGCGTGGCCAA GTCGGAGGCGGCACAGAGCTTTCCGATTCCCTCGCCGCTA TCCGTCACCATTGTCCCGCCCTCGATGGGCGGCTGCGGCG
 601  GAGTGGGCCACGCAGCCGGTCTGGACAGCGGGCTGGCCAA GTTCGACAAGACCTGGGAGGCAGGTCCGGGCAAGCTGGAG TCCATGACCGGGGTGGGCGCAGCGGCGGCGGGCACAGGGC
 721  AGCGCGGGGAGCGGGTGAAGGAGGACAGCTTCTGCTGCGT CATCTCCATGCACGACGGCATCGTCCTGTACACGACGCCC AGCATCACCGATGTCCTGGGCTACCCGCGCGACATGTGGC
 841  TGGGCAGGTCCTTCATCGATTTTGTGCACCTTAAGGACCG CGCCACCTTCGCCAGTCAGATCACCACGGGCATACCCATT GCGGAATCCAGGGGCAGCGTGCCCAAGGACGCCAAGAGCA
 961  CCTTCTGCGTGATGCTGCGTCGCTACCGGGGACTCAAGTC CGGCGGATTCGGCGTCATCGGCAGGCCCGTCAGCTACGAA CCCTTCCGCCTGGGGCTCACCTTCAGGGAGGCTCCGGAGG
1081  AGGCGCGACCGGACAACTACATGGTCTCCAATGGCACCAA CATGCTGCTCGTCATCTGCGCCACTCCGATCAAGAGCAGC TACAAGGgtgagtatctgtgcccgttttggtttcatccag
1201  ccatgtaaccatcgatctctgcccccccccagTTCCCGACG AGATTCTCTCACAGAAGAGCCCCAAGTTCGCCATACGCCA CACGGCCACCGGGATCATATCGCACGTGGACAGCGCGGCG
1321  GTCAGTGCACTGGGCTATCTGCCGCAGGACCTGATTGGCC GCAGCATCATGGACTTCTACCACCACGAGGACCTCTCAGT CATGAAGGAGACCTACGAGACGGTAATGAAGAAGGGTCAG
1441  ACGGCGGGCGCCTCCTTCTGCAGCAAGCCATACCGCTTCC TCATCCAGAACGGTTGCTACGTCCTTCTGGAGACCGAATG GACCAGCTTCGTCAATCCATGGTCCCGCAAGCTGGAATTT
1561  GTCGTCGGACACCATCGCGTCTTTCAGGgtgagttcgcgt gccgacccagtgatccaatgatcctctaaacaggtcattt caccttccagGACCCAAGCAGTGCAACGTCTTCGAGGCGG
1681  CGCCCACATGTAAGCTGAAGATATCGGAGGAGGCACAGAG CCGGAACACGCGAATCAAGGAGGATATCGTGAAGCGCCTG GCGGAGACGGTGTCCCGTCCGTCGGACACGGTCAAGCAGG
1801  AGGTTTCCCGCCGCTGCCAGGCGCTGGCCAGCTTCATGGA AACGCTCATGGACGAGGTGTCCCGGGCGGATCTCAAGC
      ---------+---------+---------+---------+ ---------+---------+---------+---------+ ---------+---------+---------+---------+
```

FIGURE 1.—(a) Map of *per* locus (Oregon R, type A transcript, based on CITRI *et al.* 1987). Introns are indicated by lines, exons by boxes (open boxes, noncoding sequence; closed boxes, coding sequence). The bold line indicates the region sequenced in the present study. (b) Nucleotide sequence of *D. melanogaster*, strain *ME-NJ1* (see Figure 2 legend for details). The sequence shown corresponds to bases 2875–4742 in the published *Oregon R* sequence (CITRI *et al.* 1987). Nucleotide 1 is in the third coding position. Upper case, exon; lower case, intron; gaps relative to other sequences in this study (except *D. yakuba*) are denoted by an asterisk.

number of differences between pairs of sequences (Table 2).

There is considerable shared polymorphism between *D. simulans* and *D. mauritiana* (11 exon sites), two shared polymorphic sites between *D. melanogaster* and *D. mauritiana*, and single shared polymorphic sites in pairs of *D. simulans-D. sechellia* and *D. melanogaster-D. simulans* (plus one shared length polymorphism in intron 2 in the latter pair). *D. melanogaster*, *D. simulans* and *D. mauritiana* displayed 25, 33 and 34 unique polymorphisms (*i.e.*, at least one, but not all, bases at a polymorphic site were unique to that species). Additionally, *D. melanogaster* possessed a unique length polymorphism in Intron 4. Only three unique polymorphisms were seen in *D. sechellia* (see Figure 2 for details).

Interspecific variation: The total and net average pairwise distances were calculated for each pair of species and are shown in Table 3. This table also shows fixed differences, the number of base positions at which all sequences of one species differ from all sequences of the second species (HEY 1991).

We also assessed the number of "unique" fixed differences, which is the number of positions at which a species is different from all others. A total of 35 fixed base differences and four fixed length differences were unique to *D. melanogaster*. *D. sechellia* possessed 14 unique fixed base differences and one fixed length difference (a deletion at sites 465–467 in Figure 2). *D. simulans* and *D. mauritiana* each had a single unique fixed base difference (sites 89 and 497, respectively). At base position 560, *D. melanogaster* and *D. mauritiana* shared a fixed base difference relative to the other two species.

Genealogical analysis: To include length variation in the genealogical analysis, all length polymorphisms were treated as an absence of sequence information in those lines lacking length variants and, instead, were coded as a single binary character added to the ends of the sequence. Thus, all length variants were equally weighted (and weighted the same as base polymorphism), and the base pair variation within regions that were polymorphic for length was included. All genealogical analyses used the published *D. yakuba* sequence (THACKERAY and KYRIACOU 1990) as an outgroup.

The program PAUP version 2.4 (SWOFFORD 1985) was used to construct a maximum parsimony tree. Preliminary parsimony analysis, as well as neighbor-joining analysis (see below), strongly suggested that the *D. melanogaster* sequences collectively form a sister group to the sequences of the other species. By removing all but one *D. melanogaster* sequence, it was possible to run PAUP using the branch and bound option (HENDY and PENNY 1982), which guarantees the most parsimonious trees. A single most parsimonious tree was found. Upon inclusion of all six *D. melanogaster* sequences, we found three trees of 232 steps (consistency index 0.647). These trees differed only slightly in the branching pattern among the *D. melanogaster* sequences. One of these trees is given in Figure 5. To better convey the length of tip branches,

```
Base                   11111111111111    111112    22    3    3333    33333333344444    4    444444445    55    5556666666
position 6777778889999900000001111244    447894    67    0    0022    23457788900114    4    666667793    66    7781122366
         8035793890134613467890128102    690540    93    0    2504    63164736047061    2    567890372    08    5811703224
comment          ii       i       isr    sssssr    sr    r    sssr    sssssssssssssr    r(1)   s   ssr    sr    sssssssssr
ME-NJ1   tgggtaa***atataga*******ctAC(S) CGCGCG(D) GG(A) A(T) CGGC(L) AAGCAAGCCTAGCA(T) T(I) GAGGAGTGC(A) AC(T) CGCCCTGGAG(G)
ME-NJ2   c----c----------------------(-) ------(-) --(-) -(-) ----(-) --------------(-) -(-) ---------(-) --(-) ------TC--(-)
ME-K1    c----c-------------------*G-(-) -C----(-) --(-) -(-) ----(-) --------------(-) -(-) ---------(-) --(-) -------C-T(V)
ME-K2    c----c----------------------(-) ------(-) --(-) -(-) A---(-) G-------------(-) -(-) ------C--(-) --(-) -------C--(-)
ME-LI1   -----c----------------------(-) ------(-) --(-) -(-) ----(-) --------------(-) -(-) ------C--(-) --(-) -------C--(-)
ME-LI2   -----c----------------------(-) ------(-) --(-) -(-) ----(-) --------------(-) -(-) ------C--(-) --(-) -------C--(-)
SI-CA1   ca---cggggcgta--tcctactctcG-(-) ------(-) -A(T) -(-) -C--(-) G---TG--ACG-G-(-) C(T) --A***C--(-) G-(-) G---T--C--(-)
SI-CA2   ca---cggggcgta--tcctactc-*G-(-) ---C--(-) -A(T) -(-) -C--(-) G----G--ACGA--(-) C(T) ---***C--(-) G-(-) G------C--(-)
SI-K1    c----cggggcgta-attctactc-*G-(-) ------(-) --(-) G(A) -C--(-) G--G-G-AACG---(-) C(T) ---***C--(-) G-(-) G-TTTC-C--(-)
SI-K2    c----cgaggcgta--tcctactctcG-(-) ---CG-(-) --(-) -(-) ----(-) GC---G--ACG-G-(-) C(T) --A***C--(-) G-(-) G---T--C--(-)
SI-LI1   ca---cggggcgta--tcctactc-*G-(-) ---C--(-) --(-) -(-) ----(-) G----G--ACG---(-) C(T) ---***C--(-) GG(S) G------C--(-)
SI-LI2   ca---cggggcgta--tcctactc-*G-(-) ---C-C(H) --(-) -(-) ----(-) G----G--ACG---(-) C(T) ---***C--(-) GG(S) G------C--(-)
SE-C1    caca-cggagcgta--tcctactc-*G-(-) A-----(-) A-(-) -(-) --C-(-) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
SE-C2    caca-cggagcgta--tcctactc-*G-(-) A--C--(-) A-(-) -(-) --C-(-) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
SE-P1    caca-cggagcgta--tcctactc-*G-(-) A-----(-) A-(-) -(-) --C-(-) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
SE-P2    caca-cggagcgta--tcctactc-*G-(-) A-----(-) A-(-) -(-) --C-(-) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
SE-P3    caca-cggagcgta--tcctactc-*G-(-) A-----(-) A-(-) -(-) --C-(-) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
SE-P4    caca-cggagcgta--tcctactc-*G-(-) A-----(-) A-(-) -(-) --CA(M) G----G--ACG-AG(A) C(T) ******C-T(V) G-(-) G------CC-(-)
MA-1     ca---cggagcgtag-tcctactc-*G-(-) ---C--(-) --(-) -(-) -C--(-) G-A--G-AACG---(-) C(T) ---***CA-(-) G-(-) -------C--(-)
MA-2     ca---cggagcgta--tcctactc-*G-(-) ---C--(-) --(-) -(-) -C--(-) G----G--ACG---(-) C(T) ---***CA-(-) G-(-) -------C--(-)
MA-3     ca---cggagcgta--tcctactc-*G-(-) ---C-A(N) --(-) -(-) -C--(-) G----G--ACG---(-) C(T) ---***CA-(-) G-(-) -------C--(-)
MA-4     ca---cggagcgta--tcctactc-*G-(-) -CAC--(-) --(-) -(-) -C--(-) G----GA-ACG---(-) C(T) ---***CA-(-) G-(-) -------C--(-)
MA-5     ca---cggagcgta--tcctactc-*GN    N--C--(-) --(-) -(-) -C--(-) G----G-AACG---(-) C(T) ---***CA-(-) G-(-) -CA----C--(-)
MA-6     ca--acggagcgtag-tcctactc-*GG(C) ---C--(-) --(-) -(-) -C--(-) G-A--G-AACG---(-) C(T) ---***CA-(-) G-(-) ----T--C--(-)
```

```
Base                                     1    11111111111111111111111111111111111111111111111111111111111111111111111111111
position 666677777777    88888889999999990    00000001111111111111222222222233333333444444445555555555566666666677777777888
         789901122346    14567890025667790    12578992357778888889000112225923345569234567890122334455900334489900133479247
         792846928464    22702132569281556    65235175074890234791359474561698940188573516843847095827903273481736536526311
comment  sssssssssssr    ssssssssssssssssr    ssssssssss                  ssssssssssssssssssssssssssss       ssssssssssssss
ME-NJ1   GCGCGAGGGGCC(I) TGCTTCCCCACCCGCCG(G) CGGTGGCCCCagtgccgtgtagcgccccACATCGCCCAATGCCCGTCACCCACCCAgtcaaccATGAGAACTGGGAT
ME-NJ2   -----T------(-) ---C-T-----------(-) ---------------------c-------------------------------------------------------
ME-K1    -----T------(-) ---C-T-----------(-) -----A---------------c---**----------------------------G-----------------A---
ME-K2    -T-----A-A--(-) ---C-T---G-------(-) -----------cc-----a--c---------------------------------G----gat--------------
ME-LI1   -T-----A-A--(-) ---C-T-----------(-) ---------------------c------------------A------------------------------C-----
ME-LI2   -T-----A-A--(-) ---C-T-----------(-) ---------------------c------------------A------------------------------------
SI-CA1   ----A-A-----(-) ---CCT---G-G-----(-) ---C--------c--g-*-a-c---***C-GCTA---CGC-ATT-CGG--TGT--Gag--g--G-CTT-G-CT-AGC
SI-CA2   ----A-A-----(-) ---CC----G-G-----(-) ---C--------c--g-*-a-c---***C-GCT-T-TCGC---TACGG-----GTGag-tg--G-C-A-G-CT-AGC
SI-K1    ----A-A-----(-) -T-CCT---G-G-----(-) T-CC--------c--ta*-a-c---***C-GCT-T-TC-C-A-T-C-----GT--Gag--g--G-C-T-G-CT-AGC
SI-K2    ----A-A-----(-) --TCCT---G-G-AT-C(A) ---C--T--T--c--t-*-a-c---**-C-GCT----CGC-----CGG---GT--Gag-tg--G-C-T-G-CT-AGC
SI-LI1   ----A-A----G(M) ---CC----G-G-----(-) ---C--------c--g-*-a-c---***C-GCT-T-TCGC-----CGG---GT--Gag-tg--G-C-T-G-CT-AGC
SI-LI2   ----A-A-----(-) ---CC----G-G-----(-) ---C--------c--g-*-a-c---***C-GCT-T-TCGC-----CGG---GT--Gag-tg--G-C-T-G-CT-AGC
SE-C1    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----ca-t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-T-G-CT-AGC
SE-C2    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----c--t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-TTG-CT-AGC
SE-P1    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----c--t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-TTG-CT-AGC
SE-P2    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----c--t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-TTG-CT-AGC
SE-P3    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----c--t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-TTG-CT-AGC
SE-P4    ----A-A--A--(-) ---CCT---GT------(-) --CCA--T----ca-t-*-atct--***C-GCT-T-TCGC---T-C-G---GT--Gaga-g--G-C-T-G-CT-AGC
MA-1     --ATA-A-A---(-) C--CCT-T-G--TA---(-) --CC----T---c--t-*aa-c--a***C-GCT----CGC---TACGG-T-GT-TGag--g--G-CTT-G-CT-AGC
MA-2     A---A-A-A---(-) C--CCT--TG--TA-T-(-) TA-C-----Tg-c-tt-*aa-c-t-***C-GCT--T-CGC---T-CGG-TTGT-TGag--g--GCCTT-G-CT-AGC
MA-3     A---A-A-A---(-) C--CCT--TG--TA-T-(-) TA-C-----Tg-c-tt-*aa-c-t-***C-GCT--T-CGC---T-CGG-TTGT-TGag--g--GCCTT-G-CT-AGC
MA-4     A-ATA-A-A-T-(-) ---CCTG-TG-G-----(-) TACC--------c--t-*aa-c---***C-GCT----CGC---TACGG-T-GG-TGag--g--GCC-T-GACT-AGC
MA-5     A---A-A-A-T-(-) C--CCT-T-G-G-----(-) TA-C-----T--c--t-*-a-ct-a***CAGCT----CGC---T-CTG-T-GT-TGag--g--G-CTT-G-CT-AGC
MA-6     A---A-A-----(-) ---CCT-T-G--TA---(-) ---C-----T--c--t-*aa-c-t-***C-GCT----CGC---T-C-GG---G-TGag--g--GCC-T-GACT-AGC
```

FIGURE 2.—Polymorphic sites among *D. melanogaster*, *D. simulans*, *D. sechellia* and *D. mauritiana*. First four rows refer to base position relative to Figure 1b. "Comment" row: s, synonymous substitution in an exon; r, amino acid replacement substitution; an absence of a letter, nucleotide substitution within an intron; i, nucleotide substitution within an intron sequence length polymorphism. The sequence of *ME-NJ1* (*D. melanogaster*) is used as the reference. Nucleotides identical to the reference in the remaining 23 lines are indicated by a dash. N, unresolved base. Upper case, exon sites; lower case, intron. At amino acid replacement sites, the nucleotide is followed in parentheses by the one-letter code for the resulting amino acid (A, Ala; C, Cys; D, Asp; G, Gly; H, His; I, Ile; L, Leu; M, Met; N, Asn, S, Ser; T, Thr; V, Val; "–" = same as reference). Length polymorphism is indicated by an asterisk in sequences shortened relative to others.

changes at noninformative sites (*i.e.*, those in which only a single sequence carries the less common base) were included in Figure 5.

We also performed a neighbor-joining analysis (SAITOU and NEI 1987) with 200 bootstrap replications. The programs SEQBOOT, DNADIST, NEIGHBOR and CONSENSE from PHYLIP version 3.4 (FELSENSTEIN 1989) were used to create a majority rule consensus tree. DNADIST was run using the multiple hits correction of KIMURA (1981) with a transition/transversion ratio of 1.6, based on the ratio seen among variable sites in the present study. The majority rule consensus tree is presented in Figure 6a. It is noteworthy that the basal nodes of *D. melanogaster*, *D. mauritiana* and *D. sechellia* occur in all 200 of the gene trees produced by bootstrapping. Similarly, the node

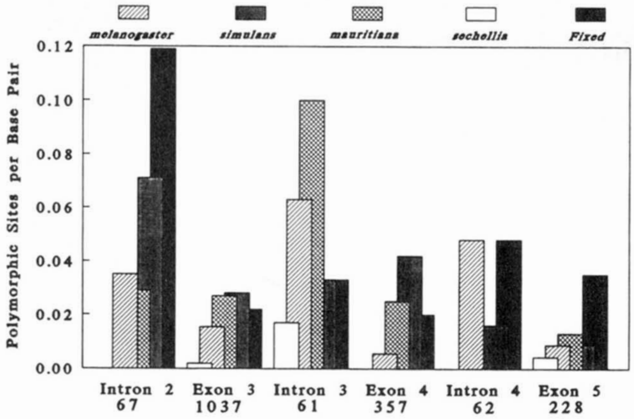


FIGURE 3.—The number of polymorphic sites per base pair classified by gene regions. Polymorphisms are indicated as occurring within particular species or as fixed differences between species. The average length of each region is given below the gene region label. The sequenced portion of exon 2 (57 bases) was invariant and is not shown.

at the base of the *D. simulans*/*D. mauritiana*/*D. sechellia* cluster also occurred 100% of the time. In addition, we performed a standard neighbor-joining analysis on the data set in order to present intra- and interspecies DNA sequence variation in the form of a distance tree (Figure 6b).

The trees presented in Figures 5 and 6 are very similar. In particular, the maximum parsimony tree (Figure 5) and bootstrapped neighbor-joining tree (Figure 6a) differ only among lines of *D. mauritiana* and in the branches leading to ME-NJ2 (*D. melanogaster*) and SI-CA1 (*D. simulans*). Several characteristics common to all trees are noteworthy. (1) As with other phylogenetic studies (LACHAISE *et al.* 1988; CACCONE, AMATO and POWELL 1988), *D. melanogaster* lineages are clearly separated from those of the other species. (2) All of the lineages within *D. sechellia* and *D. mauritiana* form discrete clusters. (3) One of the *D. simulans* lineages, SI-K1, is paired with the cluster of *D. sechellia* lineages. (4) The earliest nodes of the tree, excluding the root, separate lineages of *D. simulans* (except in the case of the standard neighbor-joining tree in Figure 6b; here, five *D. simulans* lineages form a sister group to *D. mauritiana*, *D. sechellia* and *D. simulans* line SI-K1). Additionally, based on the standard neighbor-joining tree and on the number of steps required to produce the maximum parsimony tree, branches within species are relatively very short in *D. sechellia*, intermediate within *D. melanogaster*, and long within *D. simulans* and *D. mauritiana*.

Recombination: If recombination has occurred in the history of a sample of sequences, then that history may be misrepresented by a bifurcating genealogy. Figure 7 shows all sites within each species at which different bases appear in two or more sequences. Within each of the species, except *D. sechellia*, many pairs of sites exhibit all four "gametic" types. While these patterns could be accounted for by multiple mutation events, it is likely that a majority of these conflicts result from recombination. The algorithm in Appendix 2 of HUDSON and KAPLAN (1985) applied to the patterns in Figure 7 reveals evidence for two recombination events in *D. melanogaster*, seven events in *D. simulans*, and nine events in *D. mauritiana*. These numbers are the minimum necessary to explain the observations and can be expected to greatly underestimate the actual number of recombination events that have occurred in the history of a sample (HUDSON and KAPLAN 1985). There is no evidence for recombination in *D. sechellia*. However, this is expected because of the low polymorphism, regardless of the actual amount of recombination at *per* in this species.

TABLE 1
The number of polymorphic sites within species

	Exons			Introns		
Species	Synonymous	Replacement	No. bases	Base	Length	No. bases[a]
D. melanogaster	19	1	1682	9	2	186
D. simulans	40	6	1679	8	2	192
D. sechellia	2	1	1676	1	0	191
D. mauritiana	38	2	1679	8	0	191

[a] Intron lengths are an average because of length polymorphism.

More evidence for recombination comes from the 11 polymorphisms shared by *D. mauritiana* and *D. simulans*. Polymorphisms shared via descent are incompatible with a genealogical interpretation of the single branch that leads to the *D. mauritiana* sequences in Figures 5 and 6. If this basal branch does represent a single ancestral lineage to the *D. mauritiana* sample, then all polymorphism shared with *D. simulans* must result from independent identical mutations. Considering that only 2.6% and 3.0% of all sites are variable within these species, respectively (see Table 1), most of the shared polymorphism must be explained by descent.

The intent of maximum parsimony analysis is to reconstruct the true genealogy of the sampled sequences; in this view, a bifurcating tree reflects only the processes of gene replication and transmission. Since the history of the portion of the genome studied here appears to include recombination within species, it would be improper to interpret Figure 5 as an estimate of the *per* locus genealogy. The true genealogy appears to be a complex intercalated network, especially within *D. simulans* and *D. mauritiana*. Also, because of shared polymorphism, the branch between these species represents divergence, and does *not* represent the transmission of a single lineage. Still, based on fixed differences between species and the bootstrapped neighbor-joining analysis (Figure 6a), the branches leading to the *D. melanogaster* and D. sechellia basal nodes may, in fact, appropriately repre-

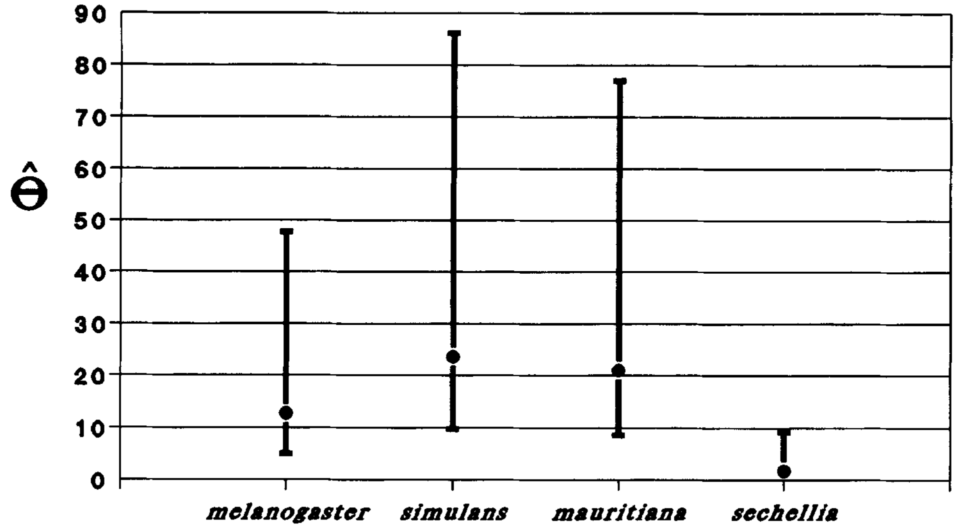


FIGURE 4.—Estimates of Θ calculated using Equation 1. Bars represent the 95% confidence interval for each species (KREITMAN and HUDSON 1991).

TABLE 2

The average number of pairwise differences within species

Species	Exon			Intron			Total			Total per base pair
		S_{sa}	S_{st}		S_{sa}	S_{st}		S_{sa}	S_{st}	
D. melanogaster	8.4	3.0	4.3	3.3	1.3	1.9	11.7	4.1	5.9	0.0062
D. simulans	18.0	6.2	8.8	3.5	1.4	2.0	21.5	7.4	10.5	0.0115
D. sechellia	1.2	0.6	0.8	0.5	0.3	0.5	1.7	0.7	1.1	0.0009
D. mauritiana	18.3	6.3	9.0	3.7	1.4	2.1	22.0	7.5	10.7	0.0118

By employing the assumptions of no recombination, a Wright-Fisher demographic model (FISHER 1930; WRIGHT 1931; EWENS 1979), and an infinite sites mutation model (KIMURA 1969), these values may be taken as estimates of $\Theta = 3N\mu$ (see text). The error estimates are made under the same assumptions (TAJIMA 1983). The sampling error, s_{sa}, is a measure of the variation expected among samples from the same population. It is calculated as the square root of the sampling variance from expression 32 of TAJIMA (1983). The stochastic error s_{st}, is a measure of the variation expected among populations of identical population sizes. It is calculated as the square root of the stochastic variance from expression 31 of TAJIMA (1983).

TABLE 3

Differences between species

	Species 2			
Species 1	*melanogaster*	*simulans*	*mauritiana*	*sechellia*
melanogaster		48.8 65.2	58.1 74.9	66.3 72.9
simulans	37		13.1 34.7	21.5 33.1
mauritiana	44	3		31.3 43.2
sechellia	60	19	21	

Above the diagonal are values for average pairwise distance between species, with net average distance [see NEI (1987), p. 276] above the gross average pairwise distance. Below the diagonal is the number of fixed differences between species (see text).

sent a portion of the true genealogy. However, the topology of the maximum parsimony tree should be considered more like a distance tree (as in Figure 6b) than a true genealogy.

Tests of natural selection: We compared the ratio of amino acid replacement mutations to synonymous mutations for changes that have become fixed between species relative to those that are polymorphic within species (McDONALD and KREITMAN 1991). [Note: if different polymorphism occurred at the same site in different species, then the site was counted

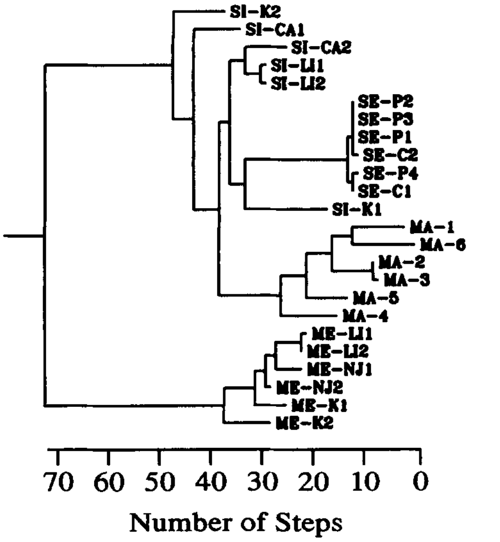


FIGURE 5.—A maximum parsimony tree. Branch lengths represent the number of steps required to construct the tree.

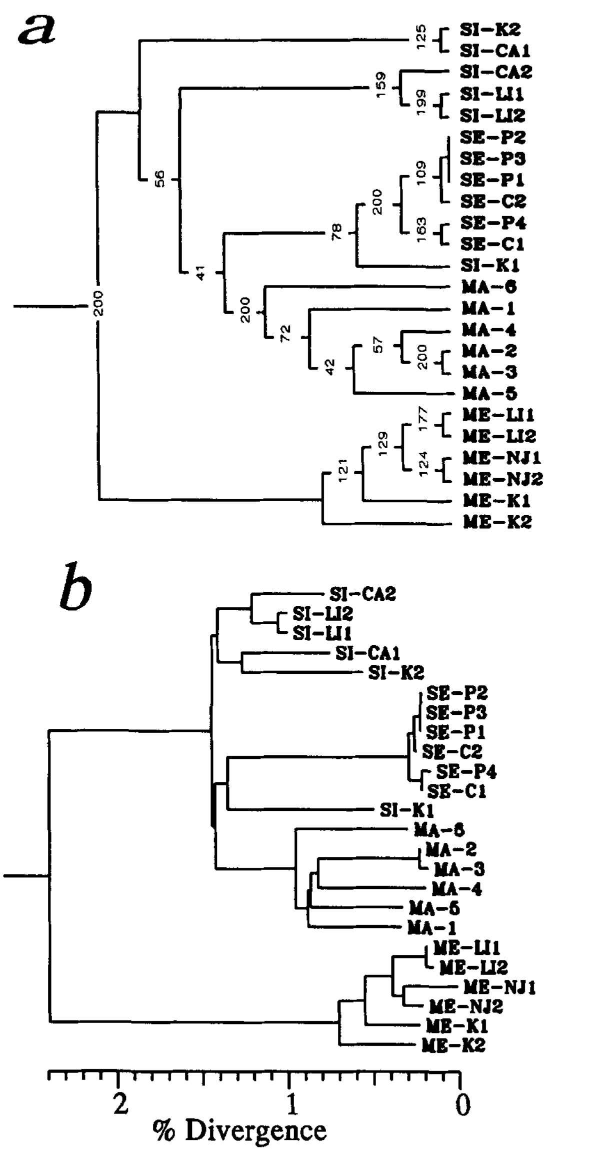
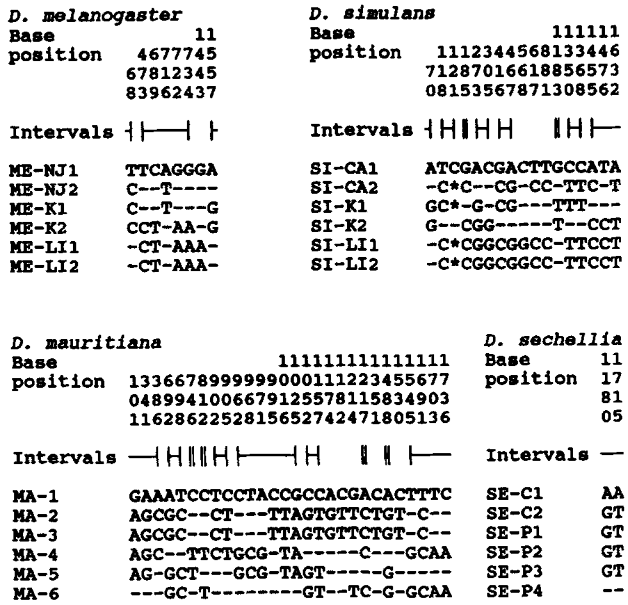


FIGURE 6.—Neighbor-joining trees. (a) A tree representing the result of bootstrapping. Bootstrap values indicate how many times out of 200 runs the lineages to the right were clustered. (b) A standard neighbor-joining tree, in which all sites are weighted equally.



```
D. melanogaster                  D. simulans
Base              11             Base                 111111
position     4677745             position   1112344568133446
            67812345                       7128701661885657 3
            83962437                       08153567871308562

Intervals                        Intervals

ME-NJ1      TTCAGGGA             SI-CA1     ATCGACGACTTGCCATA
ME-NJ2      C--T----             SI-CA2     -C*C--CG-CC-TTC-T
ME-K1       C--T---G             SI-K1      GC*-G-CG---TTT---
ME-K2       CCT-AA-G             SI-K2      G--CGG-----T--CCT
ME-LI1      -CT-AAA-             SI-LI1     -C*CGGCGGCC-TTCCT
ME-LI2      -CT-AAA-             SI-LI2     -C*CGGCGGCC-TTCCT


D. mauritiana                                D. sechellia
Base                     111111111111111     Base        11
position    1336678999999000111223455677     position    17
            0489941006679125578115834903                 81
            1162862225281565274247180513 6               05

Intervals                                    Intervals

MA-1        GAAATCCTCCTACCGCCACGACACTTTC     SE-C1       AA
MA-2        AGCGC--CT---TTAGTGTTCTGT-C--     SE-C2       GT
MA-3        AGCGC--CT---TTAGTGTTCTGT-C--     SE-P1       GT
MA-4        AGC--TTCTGCG-TA-----C---GCAA     SE-P2       GT
MA-5        AG-GCT---GCG-TAGT-----G-----     SE-P3       GT
MA-6        ---GC-T---------GT--TC-G-GCAA    SE-P4       --
```

FIGURE 7.—Evidence for recombination. All polymorphic sites at which different bases are each found in two or more individuals within a species (*i.e.*, genealogically informative sites) are shown. Assuming homoplasy is not due to multiple substitutions, recombination events must have occurred between sites left unconnected by a horizontal line in the row labelled "Intervals" (*e.g.*, between sites 68 and 473 in *D. melanogaster*). When a single site is incompatible with either adjacent site (*e.g.*, site 185 in *D. simulans*), this is indicated by a double horizontal line.

twice; otherwise the site was counted once.] If some amino acid replacement mutations are favored and become fixed between species by the action of natural selection, the ratio of replacement to synonymous variation should be higher for interspecific comparisons (MCDONALD and KREITMAN 1991). From comparisons of the coding region within and among *D. melanogaster* and the three sibling species, we found a total of 84 synonymous polymorphic sites, 10 replacement polymorphic sites, 35 fixed synonymous substitutions, and 3 fixed replacement substitutions. The ratios are nearly equal, consistent with the neutral model ($G = 0.159$, $P > 0.9$).

In comparing variable sites at which two different nucleotides appear in a sample of six individuals, the two nucleotides will be observed in either 1:5, 2:4 or 3:3 patterns. From expression (50) in TAJIMA (1989), we expect these three polymorphism patterns, under the neutral model, to occur in 1.2:0.75:0.33 proportions in each of our species. Figure 8 shows the observed and expected frequencies of the three patterns. Although *D. melanogaster* and *D. simulans* had a greater than expected number of 1:5 sites, while *D. mauritiana* had a surplus of 2:4 sites, application of TAJIMA's test (1989) revealed no significant departures of observed frequencies from neutral expectations (results not shown).

DISCUSSION

Levels of polymorphism: From the levels of intraspecific sequence polymorphism, it appears that *D. melanogaster*, *D. simulans* and *D. mauritiana* have maintained large effective population sizes (N_e). *D. sechellia* appears to have had a historically much smaller N_e or a recent episode of directional selection and genetic hitchhiking at or near the *per* locus.

The observed intraspecies variation agrees with studies of restriction fragment length polymorphism in *D. melanogaster* and *D. simulans*, which consistently reveal more variation in the latter (AQUADRO, LADO and NOON 1988; BEGUN and AQUADRO 1991). We also note that the level of *per* sequence variation in *D. melanogaster* (from Table 2, $\hat{\theta}$ per base pair = 0.0062)

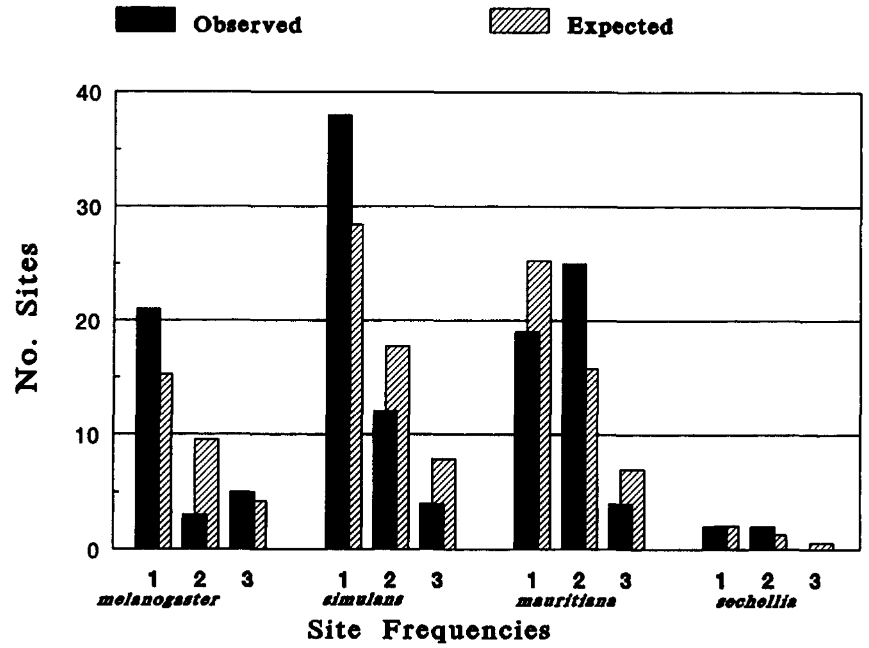


FIGURE 8.—Frequencies of polymorphism patterns. For each species, the number of sites with 1:5, 2:4 and 3:3 polymorphism patterns (labeled as 1, 2 and 3, respectively) are shown, along with expected frequencies.

was within the range of values observed for other loci in this species (AGUADE, MIYASHITA and LANGLEY 1989; AQUADRO *et al.* 1986; EANES *et al.* 1989; EANES, LABATE and AJIOKA 1989; MIYASHITA and LANGLEY 1988; SCHAEFFER, AQUADRO and LANGLEY 1988; AQUADRO, LADO and NOON 1988; BEGUN and AQUADRO 1991).

Genealogical analysis: The divergence of *D. melanogaster* from the common ancestor for all *D. simulans, D. sechellia* and *D. mauritiana* is thought to have occurred about 2.5 million years ago (LACHAISE *et al.* 1988). Our results are consistent with this relatively ancient split. Many unique fixed differences and unique polymorphisms contribute to the separation of *D. melanogaster* from the other species in our gene trees. The large number of fixed differences is also reflected in the deep branch connecting the basal node of *D. melanogaster* with that of *D. simulans* in the neighbor-joining tree (Figure 6b).

Both maximum parsimony and neighbor-joining methods indicated clustering of *per* gene copies within *D. sechellia, D. mauritiana* and *D. melanogaster* (Figures 5 and 6). Since these are acknowledged species, this result may seem trivial. However, the clusters are not required by the biological species concept. If species divergence has occurred relatively recently, and if there has not been a severe population bottleneck, it would be reasonable to find, for example, that some *per* gene copies in *D. sechellia* are more closely related to copies sampled from *D. mauritiana* than to the others from *D. sechellia* (TAJIMA 1983). Put another way, there is no guarantee that the most recent common ancestors to all *per* copies in the respective species existed after the split from the ancestral population.

It is possible that certain features of the tree topologies, such as the tightness of clustering within most species and the absence of intercalation between species, would change with the inclusion of more sequences. In particular, we would like to assess the extent to which the conclusions drawn from a small sample of sequences from one species can be extended to the entire tree for that species. Consider a sample of sequences drawn from a single species and assume the population fits a standard Wright-Fisher model (FISHER 1930; WRIGHT 1931; EWENS 1979) with constant effective population size N (*i.e.*, a discrete generation model with each generation of $2N$ gene copies formed by sampling $2N$ times, with replacement, from the previous generation). We also assume, for the moment, that recombination does not occur. The placement of *per* on the X chromosome does not affect the applicability of this model if the sample size n is much smaller than N. A random sample of n gene copies will have a genealogy that is a sample of the genealogy for all $2N$ gene copies. We would like to assess $P(n)$, the probability that the basal node of the sample genealogy is also the basal node of the population genealogy. Consider that the basal node of the population genealogy divides the population into two sets of lineages, and let the frequency in the population of those gene copies that descend from one side of the basal node be denoted by p. It follows that the frequency of those gene copies that descend from the other side of the basal node have frequency $1 - p$. If $2N \gg n$, then the probability that the sample of n genes is not made up of lineages from only one side of the basal node is, to a good approximation,

$$1 - p^n - (1 - p)^n. \qquad (2)$$

HARDING (1971), in a study of random bifurcating histories, has shown that p can take on any of the values $1/2N$, $2/2N \ldots 2N/2N$ with equal likelihood. If we treat p as a continuous, uniformly distributed, random variable, then

$$P(n) = \int_0^1 1 - p^n - (1-p)^n \, dp = \frac{n-1}{n+1}. \quad (3)$$

Since $P(6) = 0.714$ we find, under the model, a large chance that the basal node for the sample genealogy of a species is also the basal node for the entire species.

A critical assumption of this analysis is that sequences are drawn randomly from the entire population. In this light, it is important to note that *D. melanogaster* and *D. simulans* sequences are drawn from both North America and Africa, with the latter thought to be the site of the oldest populations of these species (LACHAISE *et al.* 1988). Similarly, the *D. sechellia* sequences have come from two different islands in the Seychelles.

We can also gain a rough assessment of the degree to which additional sequences can be expected, on average, to add to the depth of the tree. The relationship between population size and the time depth of genealogies is made explicit in KINGMAN's coalescent (1982a,b). For a Wright-Fisher population, the expected time since the basal node of the entire genealogy is approximately $4N$ generations and the expected time depth of a genealogy based on a sample of n sequences, in generations, is $4N(1 - 1/n)$. Thus for $n = 6$, the expectation of total depth of the genealogy is 5/6 of that for the entire species.

In sum, three points indicate that additional sequences would not appreciably alter the main feature of the trees in Figures 5 and 6. First, from Equation 3, there is a good chance that we have found the deepest nodes within species' gene trees. Second, we expect that even if additional sequences were to yield deeper nodes, they would not greatly increase the depth of the basal nodes of the within-species gene trees. Finally, there are the considerable number of fixed differences that separate the *D. sechellia* sequences from those of *D. simulans*, as well as *D. melanogaster* from the three remaining species.

It is possible that, by chance, polymorphism and divergence at *per* is not representative of the genome and thus, our conclusions are subject to the stochastic variance that occurs among unlinked loci. Under simplifying assumptions, including no recombination, stochastic variance can be estimated for intraspecific variation and accounts for much of the total variance (Table 2). However, *D. simulans* and *D. mauritiana* have had a large amount of recombination and, thus, the 1.9-kilobase region of *per* in these species behaves like multiple partially linked loci. This means that the evolutionary history of this region of *per* is more representative of the *D. simulans* and *D. mauritiana* genomes, though the effect is not easily quantified. Recombination reduces the stochastic variance of intraspecific variation (HUDSON 1983), reducing the chance that (1) additional *per* sequences would sharply alter the trees and (2) that other loci would show very different patterns of variation. It also means that the stochastic errors in Table 2 and the confidence intervals in Figure 4 for *D. simulans* and *D. mauritiana* are overestimates (HUDSON 1983).

***D. mauritiana* vs. *D. simulans*:** The two species, *D. mauritiana* and *D. simulans*, share several characteristics. Estimates of Θ, based on average pairwise sequence difference and on the number of polymorphic sites, were similar (Tables 1 and 2, Figure 4). The high estimates of Θ in *D. mauritiana* are consistent with the reported ecology of the species; although it is limited in its range to the islands of Mauritius, it is widespread and broad-niched (DAVID *et al.* 1989). The relatively high degree of polymorphism within *D. mauritiana* also corresponds to prior estimates of allozyme heterozygosity and proportion of polymorphic loci, where *D. mauritiana* actually appears more variable than the other three species in this study (CARIOU *et al.* 1990). *D. simulans* and *D. mauritiana* also share eleven polymorphisms, and there is little fixed sequence difference between them. Additionally, both species possess a large number of unique polymorphisms.

The trees in Figures 5 and 6 indicate that the two species have very different histories. Because of recombination, the *D. simulans* and *D. mauritiana* portions of the trees cannot be taken as genealogies, but they do show the variation in *D. mauritiana* sequences having arisen more recently than *D. simulans* variation. Our interpretation is that *D. simulans* still segregates many ancient polymorphisms; that some of these are shared with *D. mauritiana*; and that *D. mauritiana* has accumulated many new polymorphisms. The presence of 11 shared polymorphisms between *D. simulans* and *D. mauritiana* would seem to rule out a small founder population in the origin of *D. mauritiana*, a speciation model commonly applied to insular species (MAYR 1954). It has been argued that a small founding population experiences a rapid change in natural selection so that genetic variation persists despite a small effective population size (CARSON 1975; TEMPLETON 1980). However, all of the 11 shared polymorphisms are at silent sites (Figure 2) and they are unlikely to have been linked to nearby selected polymorphisms, given the large amount of recombination in the history of these species at *per*. The simplest explanation for the shared polymorphism is that *D. mauritiana* arose and has persisted with a large effective population size.

Because both species have similar estimates of Θ, and since *D. mauritiana* appears younger than *D.*

TABLE 4

Relative rate tests comparing the divergence of *D. simulans* sequences with the divergence of *D. mauritiana* and *D. sechellia* sequences

	Sequence 2					
Sequence 1	SI-CA1	SI-CA2	SI-K1	SI-K2	SI-LI1	SI-LI2
MA-1	0.0035	0.0069*	0.0040	0.0035	0.0086**	0.0086**
	0.0033	0.0031	0.0035	0.0034	0.0031	0.0031
MA-2	0.0047	0.0081*	0.0052	0.0047	0.0098**	0.0098**
	0.0033	0.0034	0.0037	0.0035	0.0033	0.0033
MA-3	0.0053	0.0087*	0.0058	0.0053	0.0104**	0.0104**
	0.0033	0.0034	0.0038	0.0035	0.0033	0.0033
MA-4	0.0046	0.0080*	0.0052	0.0046	0.0097**	0.0097**
	0.0034	0.0031	0.0036	0.0036	0.0032	0.0032
MA-5	0.0035	0.0069*	0.0041	0.0035	0.0086**	0.0086**
	0.0032	0.0032	0.0034	0.0034	0.0031	0.0031
MA-6	0.0017	0.0051	0.0023	0.0018	0.0069*	0.0069*
	0.0033	0.0032	0.0035	0.0033	0.0032	0.0032
SE-C1	0.0030	0.0064	0.0035	0.0030	0.0081**	0.0081**
	0.0033	0.0033	0.0034	0.0035	0.0030	0.0030
SE-C2	0.0035	0.0069*	0.0041	0.0035	0.0086**	0.0086**
	0.0033	0.0032	0.0034	0.0034	0.0029	0.0029
SE-P1	0.0030	0.0063	0.0035	0.0030	0.0081**	0.0081**
	0.0033	0.0033	0.0034	0.0035	0.0030	0.0030
SE-P2	0.0030	0.0063	0.0035	0.0030	0.0081**	0.0081**
	0.0033	0.0033	0.0034	0.0035	0.0030	0.0030
SE-P3	0.0030	0.0063	0.0035	0.0030	0.0081**	0.0081**
	0.0033	0.0033	0.0034	0.0035	0.0030	0.0030
SE-P4	0.0035	0.0069*	0.0041	0.0036	0.0086**	0.0086**
	0.0033	0.0033	0.0034	0.0035	0.0030	0.0030

ME-NJ1 was used as an outgroup in all tests. The upper value is the number of substitutions per base pair between sequence 1 and the outgroup minus the number of substitutions per base pair between sequence 2 and the outgroup. A positive value indicates more divergence for sequence 1 than for sequence 2. The lower value is the standard error of the difference. Standard errors were calculated and tests made using the method of Wu and Li (1985).

* $P < 0.05$; ** $P < 0.01$.

simulans, *D. mauritiana* may be evolving faster at the nucleotide level than *D. simulans*. In order to test whether the *D. simulans* and *D. mauritiana* lineages have diverged at equal rates, we used ME-NJ1 as an outgroup and performed relative rate tests on each of the 36 different pairs of *D. simulans* and *D. mauritiana* sequences. All of the tests attributed more substitutions to the *D. mauritiana* branch than to the *D. simulans* branch (Table 4), and three of the *D. simulans* lines (SI-CA1, SI-LI1 and SI-LI2) had significantly shorter branches relative to some or all of the *D. mauritiana* lines (Table 4). On the whole, these results support the view that *D. simulans* is evolving more slowly than *D. mauritiana*. Unfortunately, because of the complex genealogy and the apparent history of recombination, the interdependence of these contrasts cannot be assessed.

***D. sechellia* vs. *D. mauritiana*:** Based on low intraspecies variation, and a high degree of divergence from *D. simulans*, *D. sechellia* appears to have a historically low effective population size. Specifically, it has many unique fixed sequence differences and only four observed polymorphic sites (one of which is shared with *D. simulans*). Unlike *D. mauritiana*, *D. sechellia* is highly specialized, breeding only on the rotting fruit of *Morinda citrifolia* on the Seychelles (LEMEUNIER and ASHBURNER 1984). The contrasting ecologies of the two island species may explain differences in N_e, and consequently the observed levels of genetic variation (CARIOU *et al.* 1990). Our results are consistent with a scenario suggested by R'KHA, CAPY and DAVID (1991), in which a founding *D. simulans* population gave rise to a specialist adapted to *M. citrifolia*.

Like *D. mauritiana*, *D. sechellia* appears to be evolving more rapidly than *D. simulans*. Relative rate tests were carried out on *D. simulans* and *D. sechellia* in the same way as with *D. mauritiana* and *D. simulans*, with nearly identical results (Table 4).

Natural selection: In theory, the observed differences among species in levels of *per* sequence variation could be caused by differences in the action of natural selection. However, analyses of our data do not support this. Balancing selection does not appear to have been a determinant of the gene trees in Figures 5 and 6, since this should create a deep basal node separating lineages that correspond to different functional alleles (STROBECK 1983; HUDSON and KAPLAN 1988). None of the within species gene trees exhibits a fork noteworthy for its depth in relation to others. Note that the apparently large amounts of recombination in the history of *D. simulans* and *D. mauritiana* severely limit the extent to which balancing selection is expected to increase levels of polymorphism in adjacent sequences (HUDSON and KAPLAN 1988).

We also find no compelling evidence for directional selection acting on this locus. The *G*-test did not detect deviations from neutral expectations in the ratio of replacement to synonymous substitutions within and between species. Thus, unlike at *Adh* (MCDONALD and KREITMAN 1991), there is no evidence that this portion of *per* has undergone repeated adaptive fixation of amino acid replacement mutations, at least since the common ancestor of these species. Likewise, we found no significant departure from neutral expectations in the observed frequencies of patterns of polymorphism within any of the species (Figure 8). These results agree with a restriction-site study by BEGUN and AQUADRO (1991), which also provided no evidence for selection around *per* in *D. melanogaster* or *D. simulans* using the multilocus test of HUDSON, KREITMAN and AGUADE (1987).

It is possible that the lack of variation within *D. sechellia* is the result of recent directional selection. The genealogical effect of directional selection, whereby an advantageous mutation quickly becomes fixed within a population, is to create very short branches with nodes tightly clustered in time (MAYNARD-SMITH and HAIGH 1974; KAPLAN, HUDSON and LANGLEY 1989). Since *per* is associated with species differences in the male courtship song (WHEELER *et al.* 1991), one scenario is that amino acid replacements in *per* were targeted by selection to enhance premating isolation between sympatric populations of *D. simulans* and *D. sechellia*. Within the region sequenced, the only amino acid substitution fixed within *D. sechellia* is an alanine to valine substitution at position 532 (Figure 2). However, since we have not sequenced the entire locus and adjacent regions, we can not rule out the possibility that the lack of sequence variation is the result of hitchhiking accompanying adaptive fixation at a closely linked site.

As pointed out earlier, the frequencies of polymorphism patterns did not differ significantly from neutral expectations (see Figure 8). However, the two species, *D. simulans* and *D. mauritiana*, appear qualitatively different from *each other* in their polymorphism patterns. *D. mauritiana* had 19 "1:5" sites and 29 "2:4" or "3:3" sites, while these values were 38 and 16, respectively, in *D. simulans*. The ratios are significantly different ($G = 9.91$, $P < 0.005$). This test is not appropriate if polymorphic sites occur on a common genealogy and are not independent of each other (TAJIMA 1989). However, because there appears to have been considerable recombination within each of these species (Figure 7), different sites are much more independent than if linkage where complete. This comparison should also be interpreted conservatively because the test was applied *a posteriori* and to the most extreme contrast. If, in fact, the polymorphism distributions of *D. simulans* and *D. mauritiana* are different, then one possible cause is different levels of purifying selection in the two species. In effect, there may be a lower neutral mutation rate in *D. simulans* than in *D. mauritiana*, perhaps due to larger N_e in the former. In a small population, genetic drift may allow slightly deleterious mutants to persist longer, and attain higher frequencies, than in a large population, perhaps explaining the relative surplus of "2:4" and "3:3" sites in *D. mauritiana*. This could also account for the apparently higher rate of divergence seen on the branches leading to *D. mauritiana* and *D. sechellia* sequences relative to some *D. simulans* sequences (Table 4).

Phylogeny of the *D. melanogaster* species complex: One possible outcome of genealogical analysis using multiple sequences sampled from closely related species is that the phylogenetic relationships of species becomes evident. In the present study, *D. sechellia* and *D. mauritiana* appear to be independently derived from an ancient population that also gave rise to modern *D. simulans*. The relationships suggested by Figures 5 and 6 do not support a phylogeny that makes *D. sechellia* and *D. mauritiana* most closely related (see CACCONE, AMATO and POWELL 1988). Such a species tree is also unlikely on biogeographic grounds, since both species are endemic to different small and isolated island groups (LACHAISE *et al.* 1988). However, the estimated *per* gene trees also do not support a particular order to the branching of the island species from ancestral *D. simulans*. First, the order of the events leading to the formation of *D. mauritiana* and *D. sechellia* sequence clusters is not discernible. This is not counter to the concept of a bifurcating species tree, but it prevents us from settling on a *single* tree. Second, the gene trees do not support an interpretation, commonly applied to bifurcating phylogenies, of divergence of descendant species following speciation. On the basis of *per* sequences (in particular, by the presence of apparently ancient polymorphisms and the lack of accumulated unique fixed differences), *D. simulans* is little changed since divergence of the island species. In contrast, *D. sechellia* is separated from the other species by many fixed differences, while *D. mauritiana* possesses numerous unique, recent polymorphisms. This pattern of variation is evident in Table 3, where for both the number of fixed differences and net pairwise divergence, the divergence between *D. mauritiana* and *D. sechellia* is approximated by the sum of the divergence values between each of these species and *D. simulans*. Thus, rather than considering *D. simulans*, *D. mauritiana* and *D. sechellia* as "sister" species, it might be more useful to consider *D. simulans* the "parent" species and the island species as "daughters."

We would like to thank two anonymous reviewers for their helpful comments. This research was supported by National Science Foundation grant BSR 8918164 to J.H.

LITERATURE CITED

AGUADE, M., N. MIYASHITA and C. H. LANGLEY, 1989 Restriction-map variation at the *Zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 123–130.

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 and divergence. Genetics **119:** 875–888.

AQUADRO, C. F., S. F. DESSE, M. F. BLAND, C. F. LANGLEY and C. C. LAURIE-AHLBERG, 1986 Molecular population genetics of the alcohol dehydrogenase gene region of Drosophila melanogaster. Genetics **114:** 1165–1190.

ASHBURNER, M., 1989 Protocol 48: preparation of DNA from single flies, pp. 108–109 in *Drosophila: A Laboratory Manual*. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

BEGUN D., 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.

BODMER, M., and M. ASHBURNER, 1984 Conservation and change in the DNA sequences coding for alcohol dehydrogenase in sibling species of Drosophila. Nature **309:** 425–430.

CACCONE, A., G. D. AMATO and J. R. POWELL, 1988 Rates and patterns of scnDNA and mtDNA divergence within the *Drosophila melanogaster* subgroup. Genetics **118:** 671–683.

CARIOU, M.-L., M. SOLIGNAC, M. MONNEROT and J. R. DAVID, 1990 Low allozyme and mtDNA variability in the island endemic species *Drosophila sechellia* (*D. melanogaster* complex). *Experientia* **46:** 101–104.

CARSON, H. L., 1975 The genetics of speciation at the diploid level. Am. Nat. **109:** 83–92.

CITRI, Y., H. V. COLOT, A. C. JACQUIER. Q. YU, J. C. HALL, D. BALTIMORE and M. ROSBASH, 1987 A family of unusually spliced biologically active transcripts encoded by a *Drosophila* clock gene. *Nature* **326:** 42–47.

COHN, V. H., M. A. THOMPSON and G. P. MOORE, 1984 Nucleotide sequence comparison of the Adh gene in three Drosophilids. J. Mol. Evol. **20:** 31–37.

COYNE, J. A., and M. KREITMAN, 1986 Evolutionary genetics of two sibling species, *Drosophila simulans* and *D. sechellia*. Evolution **40:** 673–691.

DAVID, J. R., S. F. MCEVEY, M. SOLIGNAC and L. TSACAS, 1989 *Drosophila* communities on Mauritius and the ecological niche of *Drosophila mauritiana* (*Diptera, Drosophilidae*). Rev. Zool. Afr. **103:** 107–116.

EANES, W. F., J. LABATE and J. W. AJIOKA, 1989 Restriction-map variation with the *yellow-achaete-scute* region in five populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 492–502.

EANES, W. F., J. W. AJIOKA, J. HEY and C. WESLEY, 1989 Restriction-map variation associated with G6PD polymorphism in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. **6:** 384–397.

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

FELSENSTEIN, J., 1989 PHYLIP. Phylogeny Inference Package, version 3.2. Cladistics **5:** 164–166.

FISHER, R. A., 1930 *The Genetical Theory of Natural Selection*. Oxford University Press, London.

HARDING, E. F., 1971 The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Prob. **3:** 44–77.

HENDY, M. D., and D. PENNY, 1982 Branch and bound algorithms to determine minimal evolutionary trees. Math. Biosci. **59:** 277–290.

HEY, J., 1991 The structure of genealogies and the distribution of fixed differences between DNA sequence samples from natural populations. Genetics **128:** 831–840.

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.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

HUDSON, R. R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics **120:** 831–840.

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

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

KIMURA, M., 1969 The number of heterozygous nucleotide sites maintained in a finite population due to a steady flux of mutations. Genetics **61:** 893–903.

KIMURA, M., 1981 Estimation of evolutionary distances between homologous nucleotide sequences. Proc. Natl. Acad. Sci. USA **78:** 454–458.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. J. Appl. Prob. **19A:** 27–43.

KINGMAN, J. F. C., 1982b The coalescent. Stochast. Process. Appl. **13:** 235–248.

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

LACHAISE, D., M.-L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS and M. ASHBURNER, 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evol. Biol. **22:** 159–225.

LEMEUNIER, F., and M. ASHBURNER, 1984 Relationships within the *melanogaster* species subgroup of the genus *Drosophila* (Sophophora). Chromosoma **89:** 343–351.

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

MAYR, E., 1954 Change of genetic environment and evolution, pp. 157–180 in *Evolution as a Process*, edited by J. HUXLEY, C. HARDY and E. B. FORD. Allen & Unwin, London.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature **351:** 652–654.

MCGINNIS, W., A. W. SHERMOEN and S. K. BECKENDORF, 1983 A transposable element inserted just 5′ to a Drosophila glue gene alters gene expression and chromatin structure. Cell **34:** 75–84.

MIYASHITA, N., and C. H. LANGLEY, 1988 Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics **120:** 199–212.

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

NEI, M., and W.-H. LI, 1979 Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc. Natl. Acad. Sci. USA **76:** 5269–5273.

R'KHA, S., P. CAPY and J. R. DAVID, 1991 Host-plant specialization in the *Drosophila melanogaster* species complex: a physiological, behavioral, and genetical analysis. Proc. Natl. Acad. Sci. USA **88:** 1835–1839.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SANGER, F., S. NICKLEN and A. R. COULSON, 1977 DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. USA **74:** 5463–5466.

SCHAEFFER, S. W., C. F. AQUADRO and C. H. LANGLEY, 1988 Restriction-map variation in the *Notch* region of *Drosophila melanogaster.* Mol. Biol. Evol. **5:** 30–40.

SHEEN, J., and B. SEED, 1988 Electrolyte gradient gels for DNA sequencing. BioTechniques **6:** 942–944.

STROBECK, C., 1983 Expected linkage disequilibrium for a neutral locus linked to a chromosomal arrangement. Genetics **103:** 545–555.

SWOFFORD, D. L., 1985 PAUP version 2.4. Illinois Natural History Survey, Champaign, Ill.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite population. Genetics **105:** 437–460.

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

TEMPLETON, A. R., 1980 The theory of speciation via the founder principle. Genetics **94:** 1011–1038.

THACKERAY, J. R., and C. P. KYRIACOU, 1990 Molecular evolution in the *Drosophila yakuba period* locus. J. Mol. Evol. **31:** 389–401.

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

WHEELER, D. A., C. P. KYRIACOU, M. L. GREENACRE, Q. YU, J. E. RUTILA, M. ROSBASH and J. C. HALL, 1991 Molecular transfer of a species-specific behavior from *Drosophila simulans* to *Drosophila melanogaster.* Science **251:** 1082–1085.

WRIGHT, S., 1931 Evolution in Mendelian populations. Genetics **16:** 97–159.

WU, C.-I., and W.-H. LI, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Communicating editor: A. G. CLARK