Population Genetics and Phylogenetics of DNA Sequence Variation at Multiple Loci within the *Drosophila melanogaster* Species Complex[1]

Jody Hey and Richard M. Kliman
Department of Biological Sciences, Rutgers University

Two regions of the genome, a 1-kbp portion of the *zeste* locus and a 1.1-kbp portion of the *yolk protein 2* locus, were sequenced in six individuals from each of four species: *Drosophila melanogaster, D. simulans, D. mauritiana,* and *D. sechellia.* The species and strains were the same as those of a previous study of a 1.9-kbp region of the *period* locus. No evidence was found for recent balancing or directional selection or for the accumulation of selected differences between species. *Yolk protein 2* has a high level of amino acid replacement variation and a low level of synonymous variation, while *zeste* has the opposite pattern. This contrast is consistent with information on gene function and patterns of codon bias. Polymorphism levels are consistent with a ranking of effective population sizes, from low to high, in the following order: *D. sechellia, D. melanogaster, D. mauritiana,* and *D. simulans.* The apparent species relationships are very similar to those suggested by the *period* locus study. In particular, *D. simulans* appears to be a large population that is still segregating variation that arose before the separation of *D. mauritiana* and *D. sechellia.* It is estimated that the separation of ancestral *D. melanogaster* from the other species occurred 2.5–3.4 Mya. The separations of *D. sechellia* and *D. mauritiana* from ancestral *D. simulans* appear to have occurred 0.58–0.86 Mya, with *D. mauritiana* having diverged from ancestral *D. simulans* 0.1 Myr more recently than *D. sechellia.*

Introduction

This paper describes DNA sequence variation at multiple loci within and among closely related species. These data are used to study the magnitude and form of DNA sequence divergence associated with the formation of new species. Our initial study described DNA sequence variation in a 1.9-kbp region of the *period* locus (*per*) in six individuals from each of the four species of the *Drosophila melanogaster* species complex (Kliman and Hey 1993). The *per* locus proved highly polymorphic within *D. simulans* and *D. mauritiana,* less variable in *D. melanogaster,* and nearly unvaried in *D. sechellia.* Genealogical interpretation of the variable sites suggested that *D. simulans* is a species of large effective population size, still segregating many polymorphisms that arose before *D. mauritiana* and *D. sechellia* split from the ancestral population.

Evolutionary inferences from a single locus, such as those from *per,* are often uncertain because different forces can create similar patterns of variation. Specifically, levels of intraspecific polymorphism—even apparently neutral variation in silent sites

1. Key words: polymorphism, speciation, *simulans, mauritiana, sechellia, zeste, yolk protein, period* locus.

Address for correspondence and reprints: Jody Hey, Rutgers University, Department of Biological Sciences, Nelson Laboratories, P.O. Box 1059, Piscataway, New Jersey 08855.

Mol. Biol. Evol. 10(4):804–822. 1993.
© 1993 by The University of Chicago. All rights reserved.
0737-4038/93/1004-0006$02.00

or introns—are determined by the rate of genetic drift *and* the pattern of natural selection acting on linked non-neutral variation. A clear example of this uncertainty is seen with *per,* where an absence of variation in *D. sechellia* could be due to small effective population size *or* to the recent fixation of a linked favored mutation (Kliman and Hey 1993).

The purpose of extending the research to multiple loci is to discriminate between forces that are expected to affect all loci similarly and forces that act on smaller portions of the genome. The first category consists of forces—such as genetic drift, population subdivision, and speciation—that act on populations. In contrast, natural selection acting on functional variation at individual loci is not expected to affect variation at effectively unlinked loci.

Here, we build on the *per* data set, adding six *zeste* locus sequences and *yolk protein 2* (*yp2*) locus sequences from each of the same four species. The same isofemale lines used by Kliman and Hey (1993) have been used here; for most lines, all three sequences have come from the same chromosome. All three loci are on the X chromosome, and all three loci were selected without a priori expectations of recent patterns of directional or balancing selection.

The *zeste* locus was originally discovered via its involvement with transfection at the *white* locus (Gans 1953). *zeste* is a DNA-binding protein and appears to be involved in the regulation of expression of many loci (Pirrotta et al. 1988). It is located in salivary-gland chromosome band 3A3 (Judd et al. 1972; Mariani et al. 1985), cytologically near *per,* which is located at salivary-gland chromosome bands 3B1–2 (Young and Judd 1978; Smith and Konopka 1981). If the DNA sequence length per salivary-gland chromosome band of this region is typical of the genome, then the eight bands between 3A3 and 3B1 (Bridges 1938) should correspond to ~160 kbp (Spierer et al. 1983). A rough guess of the recombination distance between *per* and *zeste* is 0.6 map units, based on an average of 0.078 map units per band on the X chromosome (Lefevre 1971).

The *yp2* locus is one of three yolk-protein loci in *D. melanogaster* and is located, along with *yolk protein 1,* in salivary-gland chromosome sections 9A–9B (Barnett et al. 1980; Riddell et al. 1981). The three yolk proteins are synthesized in the fat bodies and ovarian follicle cells of females (Brennan et al. 1982) and are thought to function primarily as nutrients during embryogenesis. Deletions at any of the loci result in decreases in fertility and fecundity, but the presence of some viable offspring suggests that the proteins are functionally interchangeable (Bownes et al. 1991).

Material and Methods

Sources of Flies

All strains are the same as those used by Kliman and Hey (1993). 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: *Drosophila melanogaster*—ME-NJ1 and ME-NJ2, Terhunes Farm, N.J., October 1987, M. Kreitman; ME-K1 and ME-K2, Impala, Kenya, August 1989, K. Ardley (via M. Kreitman); ME-LI1 and ME-LI2, Davis Peach Farm, Mt. Sinai, N.Y., 1989, W. Eanes; *D. simulans*—SI-LI1 and SI-LI2, Davis Peach Farm, Mt. Sinai, N.Y., 1989, W. Eanes; SI-CA1 and SI-CA2, Soda Lake, Calif., fall 1989, S. Bryant (via D. Begun); SI-K1 and SI-K2, Impala, Kenya, August 1989, K. Ardley (via M. Kreitman); *D. mauritiana*—MA-1, MA-2, MA-3, MA-4, MA-5, and MA-6, Mauritius (main island) 1981, O. Kitagana (via J. Coyne); *D. sechellia*—SE-C1 and SE-C2, Cousin Island, Seychelles, January 1985,

J. David (via J. Coyne); SE-P1, SE-P2, SE-P3, and SE-P4, Praslin Island, Seychelles, July 1987, Y. Fuyama (via K. Kimura).

DNA Preparation

For all of the *yp2* sequences and most of the *zeste* sequences, the genomic DNA preparations were the same as those used by Kliman and Hey (1993). The exceptions for *zeste* are SE-C2, SE-P3, SE-P4, MA-5, and MA-6. New DNA preparations were made from these lines, with single male flies, by following protocol 48 of Ashburner (1989, pp. 108–109).

Polymerase Chain Reaction (PCR) and Sequencing

For *yp2,* a 1,411-bp region was PCR amplified by using oligonucleotide primers corresponding to bases 40–59 and 1431–1450 of the published Canton-S sequence (Hung and Wensink 1983) (fig. 1). For *zeste,* an ~1,200-base region was PCR amplified by using oligonucleotide primers corresponding to bases 1000–1019 and 2174–2193 of the published *D. melanogaster* sequence (Pirrotta et al. 1987) (fig. 1). PCR and DNA sequencing methods were identical to those of Kliman and Hey (1993), and both strands where sequenced for each locus and line.

Results

DNA Sequence Variation Summary

Figure 1 shows a schematic of the *zeste* and *yp2* loci. In the ~1 kbp sequenced from the *zeste* region, a total of 74 varied sites were found (fig. 2). These include 41 synonymous sites, 2 amino acid–replacement sites, 24 single-base sites in introns, 1 intron length variant, and 6 exon length variants. For *yp2,* 47 variable sites were found in ~1.1 kbp of sequence (fig. 3). These include 24 synonymous sites, 11 replacement

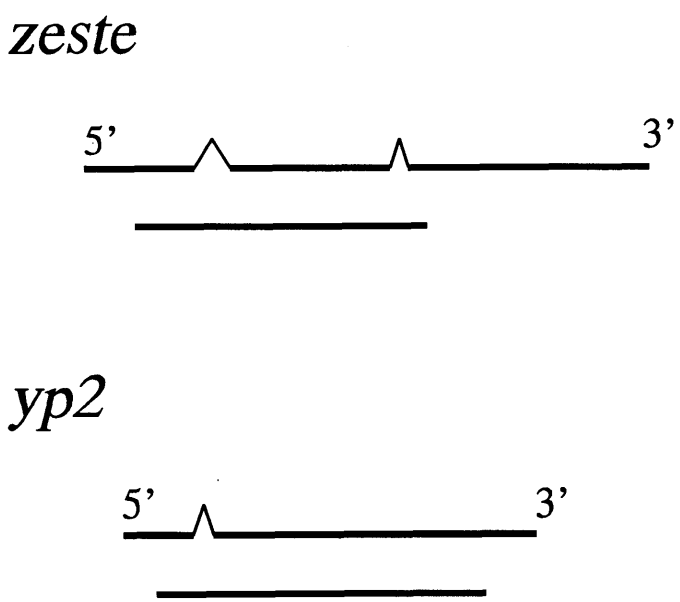

FIG. 1.—*Zeste* and *yp2* loci. The top line for each locus represents the exons and introns between the start and stop codons (Hung and Wensink 1983; Manusukhani et al. 1988). *zeste* is 1,908 bases, and *yp2* is 1,397 bases. The lower line indicates the region sequenced.

Base position	1111 144670566 6802052658	1122222222 7901123444 1263848367	2222222223 5677789990 9834651294	3334444444 7890013333 9516954567	4444 3366 8902	444 667 361
comment	sssssssss	ss		ssssss	sr	ssr
SI-CA1	CGCACGACCC	GTgg*caaat	acagacatgt	CGCTCACAGC	AACA(N)	TGA(E)
SI-CA2	-------T--	----t-----	--g-------	--TC------	----(-)	---(-)
SI-LI1	----------	----------	----------	----------	----(-)	---(-)
SI-LI2	-------T--	----t-----	--g-------	--TC------	----(-)	---(-)
SI-K1	------G--T	----------	----------	---C------	----(-)	---(-)
SI-K2	-A-----T--	----t-----	------ccac	---C--****	**--(-)	---(-)
MA-1	--N-------	----t-----	----------	T--C------	--T-(-)	---(-)
MA-2	----------	----t-----	-t--------	T--C------	----(-)	---(-)
MA-3	----------	----t-----	-t--------	T--C------	----(-)	---(-)
MA-4	T---------	----t-----	-g--------	T--C------	--T-(-)	---(-)
MA-5	----------	----t-----	----------	T--C------	--T-(-)	---(-)
MA-6	T---------	----t-----	-t--------	T--C------	----(-)	---(-)
SE-C1	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
SE-C2	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
SE-P1	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
SE-P2	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
SE-P3	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
SE-P4	-------T--	A--at-g---	------ccac	---C------	----(-)	---(-)
ME-NJ1	--TGT---T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)
ME-NJ2	--TGT---T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)
ME-LI1	--TGT---T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)
ME-LI2	--TGT---T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)
ME-K1	--TGT---T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)
ME-K2	--TGTA--T-	-Aa-tt-gcc	t--attcc--	-A-CTG----	---G(S)	CTT(V)

Base position	4555555566 7124788803 8794103972	6666666666 3333333444 3456789012	6666777777 4469000111 3941789012	7788888888 3801445566 0727890934	888899999 677901112 548103492	99 99 36
comment	sssssssss		sss	ssss s	s	ss
SI-CA1	CCCGCCGGG*	**********	*ACAGCAGTG	CATGCAGCTC	ACgcgagag	CC
SI-CA2	---A----A-	----------	----------	----------	--a------	--
SI-LI1	----------	----------	----------	----------	---------	-N
SI-LI2	---A----A-	----------	----------	----------	--a------	--
SI-K1	----------	----------	-G--------	----------	-------t-	--
SI-K2	---A----A-	----------	----------	--CA------	-------t-	--
MA-1	----------	----------	-G--------	----------	---tc----	N-
MA-2	G---------	----------	-G--------	--N-------	---------	-T
MA-3	G---------	----------	-G--------	----------	---N-----	-T
MA-4	----------	----------	-G--------	----------	---------	--
MA-5	--T--T----	----------	-G--------	----------	----c----	--
MA-6	----------	----------	-G--------	----------	---------	--
SE-C1	----------	----------	-G--------	----------	-T-------	--
SE-C2	----------	----------	-G--------	----------	-T-------	--
SE-P1	----------	----------	-G--------	----------	-T-------	--
SE-P2	----------	----------	-G--------	----------	-T-------	--
SE-P3	----------	----------	-G--------	----------	-T-------	--
SE-P4	----------	----------	-G--------	----------	-T-------	--
ME-NJ1	-AA-T-A--G	CCCAAGCCCA	AG--******	-G--***-**	*----gt-c	T-
ME-NJ2	-AA-T-A--G	CCCAAGCCCA	AG--******	-G--***-**	*----gt-c	T-
ME-LI1	-AA-T-A--G	CCCAAGCCCA	AG--******	-G--***-**	*----gt-c	T-
ME-LI2	-AA-T-A--G	CCCAAGCCCA	AG--******	-G--***T**	*----gt-c	T-
ME-K1	-AA-T----G	CCCAAGCCCA	AGA-******	AG--***-**	*----gt-c	T-
ME-K2	-AA-T--A-G	CCCAA-----	-G-G******	-G--***-**	*----gt-c	T-

FIG. 2.—Variable sites at *zeste*. The first rows indicate the base position of variable sites within the sequenced region. The first and last bases sequenced correspond to positions 1548 and 2534, respectively, of Mansukhani et al. (1988). In the "comment" row, s = synonymous substitution in an exon; r = amino acid replacement substitution; i = nucleotide substitution within an intron sequence length variant; and absence of a letter denotes nucleotide substitution within an intron. The sequence of *SI-CA1* (*Drosophila simulans*) is used as the reference. Nucleotides identical to the reference in the remaining 23 lines are indicated by a dash. N = an unresolved base. Uppercase letters denote exon sites; and lowercase letters denote intron sites. At amino acid replacement sites, the nucleotide is followed in parentheses by the one-letter code for the resulting amino acid (N = asn; S = ser; E = glu; and V = val). Length variation is indicated by an asterisk (*) in sequences shortened relative to others.

```
Base                                                                                         11
position              11   1111111111111112   2233334   4   4   45555566   6   6677   7889   01
            35   6   702   4555666667778885   6711356   6   6   74458813   4   6906   7579   30
            17   6   061   5489045684570377   7005911   2   7   73652628   0   6957   1397   24
comment     sr   r   ssr        ii        r   ssssssr   r   r   sssssssr   r   sssr   sssr   ss
SI-CA1      GA(K)G(S)TCG(K)gg***agtc**taacA(K)CTCTATC(A)G(A)A(N)CTGCCTGG(R)C(L)TGCC(T)CCCG(A)GC
SI-CA2      --(-)-(-)---(-)----------------(-)-------(-)-(-)-(-)--------(-)-(-)----(-)----(-)--
SI-LI1      --(-)-(-)---(-)--tca-----------(-)-----C-(-)-(-)-(-)--------(-)-(-)G---(-)----(-)--
SI-LI2      --(-)-(-)---(-)--tca--g--------(-)-------(-)-(-)-(-)--------(-)-(-)G---(-)----(-)--
SI-K1       --(-)-(-)---(-)----------------(-)-------(-)-(-)-(-)--------(-)-(-)----(-)----(-)--
SI-K2       --(-)-(-)---(-)----------------(-)-------(-)-(-)-(-)--------(-)-(-)----(-)----(-)--
MA-1        -G(R)-(-)G--(-)--tct-----------(-)-------(-)-(-)-(-)-------A(H)-(-)G---(-)----(-)-T
MA-2        -G(R)-(-)G--(-)--tca-----------(-)-------(-)-(-)-(-)T-------(-)-(-)G---(-)----(-)-T
MA-3        -G(R)-(-)G--(-)--tct-----------(-)-------(-)-(-)-(-)--------(-)-(-)G---(-)----(-)-T
MA-4        -G(R)-(-)G--(-)--tct-----------(-)-------(-)-(-)-(-)--------(-)-(-)G---(-)----(-)-T
MA-5        -G(R)-(-)G--(-)--tct-----------(-)-------(-)-(-)-(-)--------(-)-(-)G---(-)----(-)--
MA-6        -G(R)-(-)G--(-)--tct-----------(-)-------(-)-(-)-(-)--------(-)-(-)G---(-)----(-)-T
SE-C1       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G---(-)----(-)--
SE-C2       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G---(-)----(-)--
SE-P1       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G---(-)----(-)--
SE-P2       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G---(-)----(-)--
SE-P3       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G--T(I)----(-)--
SE-P4       --(-)-(-)---(-)--taa---------g-(-)-------(-)-(-)-(-)-C------(-)A(M)G---(-)----(-)--
ME-NJ1      T-(-)C(T)-T-(-)actcata-tcgg-ctG(R)TCTGG-A(E)-(E)C(T)--ATTC--(-)-(-)GAT-(-)T-TA(T)--
ME-NJ2      T-(-)C(T)-T-(-)actcata-tcgg-ctG(R)TCTGG-A(E)-(E)C(T)--ATTC--(-)-(-)GAT-(-)T-TA(T)--
ME-LI1      T-(-)C(T)-T-(-)actcata-tcgg-ctG(R)TCTGG-A(E)-(E)C(T)--ATTC--(-)-(-)GAT-(-)--TA(T)T-
ME-LI2      T-(-)C(T)-T-(-)actcata-tcgg-ctG(R)TCTGG-A(E)-(E)C(T)--ATTC--(-)-(-)GAT-(-)T-TA(T)--
ME-K1       T-(-)C(T)-TC(N)actcata-tcga*c-G(R)TCTGG-A(E)C(D)C(T)--AT-CA-(-)-(-)G---(-)-TTA(T)--
ME-K2       T-(-)C(T)-TC(N)actcata-tcga*c-G(R)TCTGG-A(E)C(D)C(T)--AT-CA-(-)-(-)G---(-)-TTA(T)--
```

FIG. 3.—Variable sites at *yp2.* The first and last bases correspond to positions 162 and 1275, respectively, in fig. 2 of Hung and Wensink (1983). Symbols are the same as in fig. 2, with the following one-letter amino acid codes: A = ala; E = glu; H = his; I = ile; K = lys; M = met; N = asn; R = arg; S = ser; and T = thr.

sites, 9 single-base intron sites, and 3 intron length polymorphisms. At both loci, approximately half of all variable sites appear as fixed differences between *Drosophila melanogaster* and the other species.

The *zeste* sequences are noteworthy for the relatively large amount of insertion/deletion variation found within coding regions. The alignment shown in figure 2 minimizes the number of single-base substitutions. The largest length difference, a 12-base stretch beginning at position 632, is within a long stretch of Gln-Ala repeats.

Evolutionary Constraint

The level of variation at intron, silent, and replacement sites for each locus is presented in table 1. Calculation of the fraction of all possible coding-region substitutions that would not affect the amino acid sequence provided measures of both the effective number of amino acid replacement sites and the effective number of silent sites. The actual number of variable sites observed at a locus, over all 24 sequences, can then be expressed as a fraction of the available sites. In general, levels of synonymous and intron change are roughly similar and much higher than for replacement changes. However, when we compare the levels of these different classes of variation among loci, we find differences. In particular, *yp2* has approximately one-third the silent variation and ~50% more replacement variation than is found in *per. yp 2* also has over four times the replacement variation seen in *zeste.* Since the fraction of exon lengths that are effectively silent is nearly the same for all three loci, we can test whether the relative proportions of synonymous and replacement variation are different among these three genes. A test of all three loci clearly rejects the null hypothesis of equality ($G = 13.526$ with 2 degrees of freedom, $P = 0.0012$), as do two of the possible two-locus tests (*per* vs. *yp2*—$G = 9.339$ with 1 degree of freedom, $P = 0.0022$; and

Table 1
Levels of Variation at Different Loci

Locus	Intron Length	Exon Length	Silent Sites[a]	Changes[b]			I^c	S^d	R^e	CAI[f]	C_e[g]
				Intron	Synonymous	Replacement					
zeste ..	182	805	167	24	41	2	0.132	0.246	0.003	0.467	42.1
yp2 ...	63	1051	234	9	24	11	0.143	0.103	0.014	0.697	33.7
per[h] ...	192	1679	386	40	115	12	0.208	0.298	0.009	0.490	36.6

[a] Calculated by considering, for each base position of the SI-CAI sequence, the fraction of possible base changes ($^1/_3$, $^2/_3$, $^3/_3$) that would not affect the amino acid sequence. These values were then summed across all exon base positions, and the total was rounded to the nearest integer.

[b] Total no. of variable sites observed across all 24 sequences (figs. 2 and 3; Kliman and Hey 1993, fig. 2).

[c] No. of intron variable sites divided by total intron length.

[d] No. of synonymous variable sites divided by no. of silent sites.

[e] No. of replacement variable sites divided by no. of replacement sites (i.e., exon length minus silent sites).

[f] Codon adaptation index (Sharp and Li 1987), calculated with the codon-usage table for "high bias" genes identified by Shields et al (1988, table 2). For those codons studied by Shields et al. that had zero counts, we followed the suggestion of Bulmer (1988) and used a relative-usage level of 0.01.

[g] Effective no. of codons, calculated from SI-CAI according to the method of Wright (1990). Note that higher values reflect lower codon bias.

[h] Data are from Kliman and Hey (1993).

zeste vs. *yp2*—G = 10.536 with 1 degree of freedom, P = 0.0011). The third two-locus test, with *zeste* and *per,* is not significant (G = 1.094 with 1 degree of freedom, P = 0.296). It appears, therefore, that *yp2* is at odds with the other loci; that is, the sequenced region of the *yp2* locus is more permissive of amino acid sequence variation and is less permissive of silent site variation than are the other loci.

The reduced level of silent variation at *yp2* is also consistent with its high level of codon bias (table 1). If natural selection is limiting codon usage, then it is expected (Shields et al. 1988) that more-biased loci will have a lower rate of substitution for silent sites. During the first 48 h after eclosion, female *Drosophila* synthesize a large amount of the yolk proteins, which ultimately constitute approximately one-third of total hemolymph protein (Gavin and Williamson 1976). Because an increase in codon bias has been associated with highly expressed genes (Bennetzen and Hall 1982; Sharp et al. 1986; Shields et al. 1988), the high codon bias in *yp2* is not surprising. The relatively high level of replacement variation at *yp2* suggests a reduced level of functional constraint, consistent with a role as a nutrient source for the developing embryo.

Intraspecific Variation

A commonly used parameter in models of DNA sequence variation in diploid populations is θ, which is equal to four times the product of the effective population size (N) and the neutral mutation rate (μ). For the present case of X chromosome sequences, it is more accurate to think of $\hat{\theta}$ as an estimate of $3N\mu$. If the assumptions are made that mutations are neutral and occur under an infinite-sites model (Kimura 1969) and that the population is at stationarity under a Wright-Fisher model (Ewens 1979), then θ can be estimated either by using the total number of polymorphic sites (fig. 4) or by using the average number of differences between sequences (table 2).

Analyses at all three loci show that variation in *D. sechellia* is at or near zero. In the nearly 4 kbp sequenced in each of the six lines, only five polymorphic sites have been found in *D. sechellia.* Comparisons also show that, for *D. simulans* and *D.*

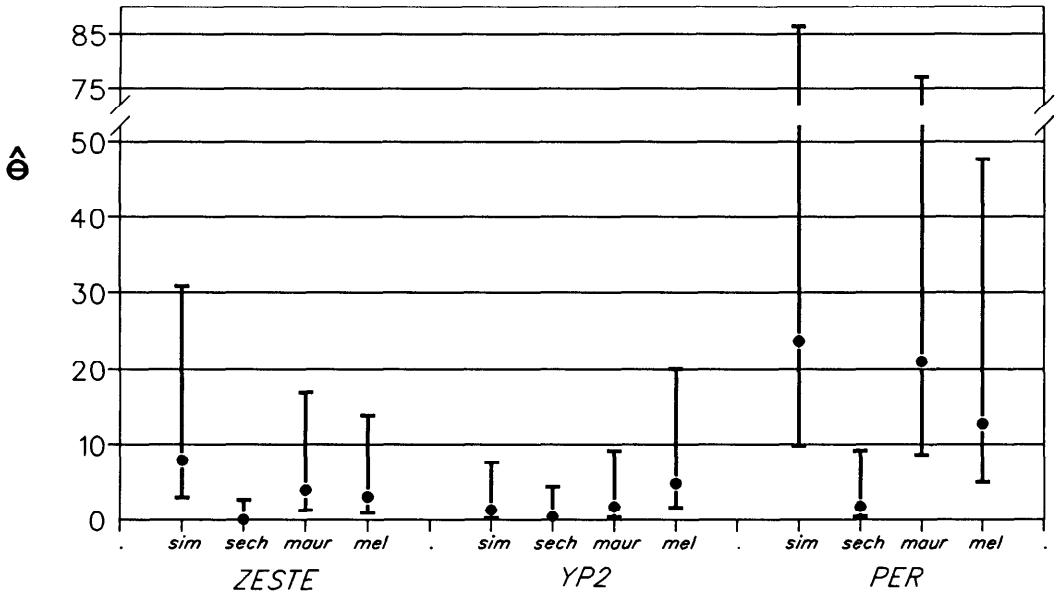

FIG. 4.—Estimates of θ from the number of polymorphic sites, calculated with expression (17) of Hudson (1990). Bars represent the 95% confidence interval for each species, calculated according to the procedure of Kreitman and Hudson (1991).

mauritiana, per is the most variable locus, followed by *zeste* and *yp2,* in that order. This is true after adjusting for sequence length (table 2) and is not explained by different intron lengths or the proportion of effectively silent sites (table 1).

Drosophila melanogaster does not fit the pattern of *D. simulans* and *D. mauritiana.* At *yp2, D. melanogaster* shows nearly as much variation at the base-pair level as at *per* (0.0052 vs 0.0062) and is the most variable species. Almost all of this variation occurs as fixed differences between the two identical African sequences (ME-K1 and ME-K2) and four similar sequences from North America. Base-pair variation at *zeste* in *D. melanogaster* (0.0025; table 2) was similar to that found, by using restriction enzymes, in a 20-kb region around *zeste* (0.004; Aguadé et al. 1989).

Genealogical Inference

As in the *per* genealogical analyses (Kliman and Hey 1993), all length variants were treated as an absence of sequence information in those lines requiring the insertion of gaps; however, each length variant was also coded as a single binary character (indicating presence or absence of the DNA sequence) added to the end of each nucleotide sequence. Thus, all insertion/deletion variants, regardless of length, are equally weighted and are weighted the same as are base variants. This approach permits inclusion of base-pair variation within regions that are also polymorphic for length. Outgroups were not used, though the large divergence between *D. melanogaster* and the other species, for both *zeste* and *yp2,* suggests that the root is along this branch, consistent with all other data (Bodmer and Ashburner 1984; Cohn et al. 1984; Coyne and Kreitman 1986; Caccone et al. 1988; Lachaise et al. 1988; Kliman and Hey 1993).

Distance matrices were created by using the program DNAdist (PHYLIP version 3.4; Felsenstein 1989) with the multiple-hits correction of Kimura (1981). The transition/transversion ratios were 3.0 for *zeste* and 1.5 for *yp2,* based on the observed

Table 2
Average Number of Pairwise Differences within Species ($\hat{\theta}$) with Estimated Differences

	Structure									
	Exon			Intron			Total			
Locus and Species	$\hat{\theta}$	S_{sa}[a]	S_{st}[b]	$\hat{\theta}$	S_{sa}[a]	S_{st}[b]	$\hat{\theta}$	S_{sa}[a]	S_{st}[b]	$\hat{\theta}$ per Base Pair
Zeste:										
simulans	4.9	1.8	2.6	2.9	1.2	1.7	7.8	2.8	4.0	0.0078
sechellia	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0000
mauritiana	2.9	1.1	1.7	1.7	0.7	1.1	4.5	4.5	2.5	0.0045
melanogaster ...	2.5	1.0	1.5	0.0	0.0	0.0	2.5	2.5	1.5	0.0025
YP2:										
simulans	0.9	0.4	0.7	0.3	0.2	0.4	1.2	0.6	0.8	0.0011
sechellia	0.3	0.2	0.4	0.0	0.0	0.0	0.3	0.2	0.4	0.0003
mauritiana	1.0	0.5	0.7	0.3	0.2	0.4	1.3	0.6	0.9	0.0012
melanogaster ...	4.7	1.7	2.5	1.1	0.5	0.8	5.7	2.1	3.0	0.0052
Per:[c]										
simulans	18.0	6.2	8.8	3.5	1.4	2.0	21.5	7.4	10.5	0.0115
sechellia	1.2	0.6	0.8	0.5	0.3	0.5	1.7	0.7	1.1	0.0009
mauritiana	18.3	6.3	9.0	3.7	1.4	2.1	22.1	7.5	10.7	0.0118
melanogaster ...	8.4	3.0	4.3	3.3	1.3	1.9	11.7	4.1	5.9	0.0062

NOTE.—Under the assumptions of no recombination, a Wright-Fisher demographic model; Ewens 1979, p. 16), and an inifinite-sites 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).

[a] S_{sa} = sampling error, which is a measure of the variation expected among samples from the same population and is calculated as the square root of the sampling variance of Tajima [1983, expression (32)].

[b] S_{st} = stochastic error, which is a measure of the variance expected among populations of identical sizes and is calculated as the square root of the stochastic variance of Tajima [1983, expression (31)].

[c] Data are from Kliman and Hey (1993).

number of variable sites (figs. 2 and 3). Neighbor-joining trees (Saitou and Nei 1987) were produced by using the PHYLIP program NEIGHBOR (fig. 5). Neighbor-joining bootstrap trees were produced by using NEIGHBOR in conjunction with the programs SEQBOOT, DNAdist, and CONSENSE. Majority-rule consensus trees based on 200 replicates are shown in figure 6.

Maximum-parsimony analysis for both loci was problematic because of the small number of "informative" sites within species. Both loci yielded similar sets of most parsimonious trees (Swofford 1985). All most parsimonious trees resembled those of figures 5 and 6, in that *D. melanogaster* and *D. mauritiana* formed discrete clades. Trees varied in their topology among the *D. simulans* sequences, as well as in the positioning of the *D. mauritiana* and *D. sechellia* clusters with respect to each other and to the *D. simulans* sequences.

With both a neighbor-joining distance tree and a majority-rule consensus tree, we can refer to branch lengths and to the level of confidence in topology, respectively. With a single exception (switching the position of MA-2 and MA-5 in *yp2*), both trees have the same topology for a given locus.

At both loci, *D. melanogaster* and *D. sechellia* sequences clearly cluster within their respective species designations and are distinct from the sequences of *D. mauritiana* and *D. simulans*. Relative to the other species, *D. sechellia* has five unique fixed differences at *yp2* and seven unique fixed differences at *zeste*. In the case of *D. maur-*

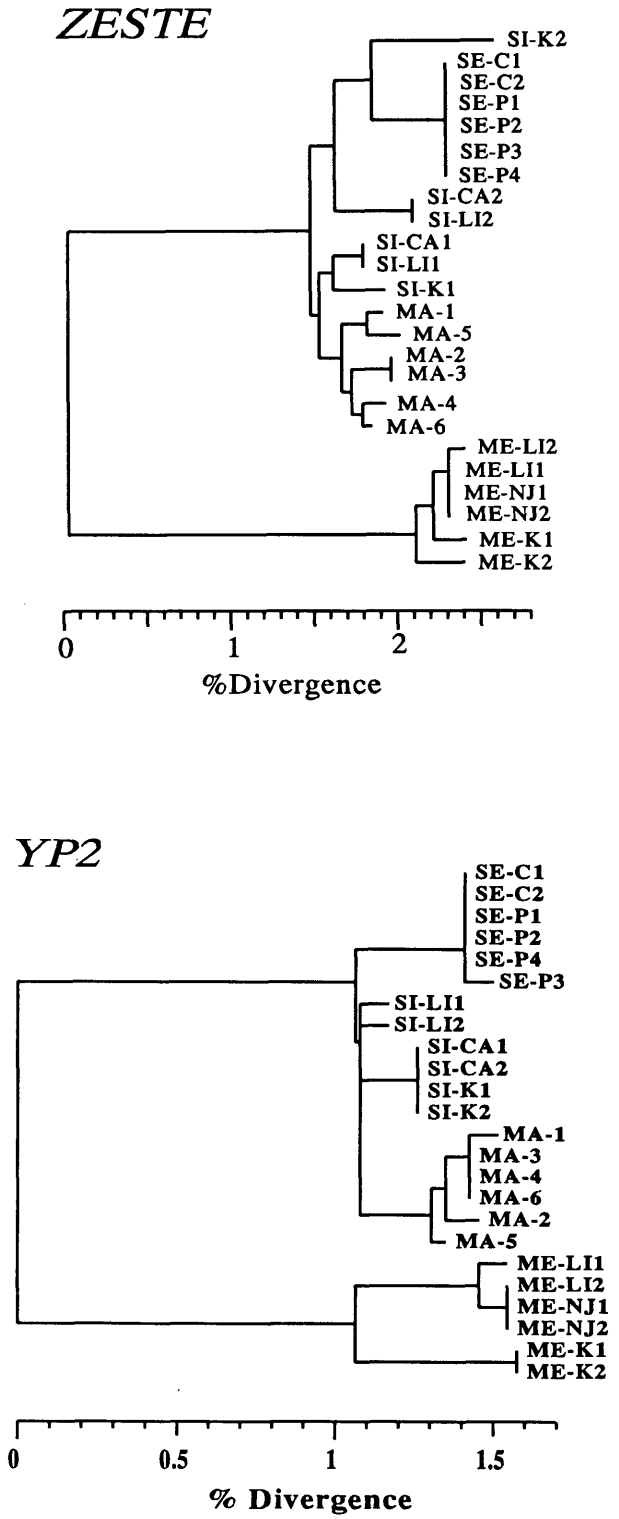

FIG. 5.—Neighbor-joining trees

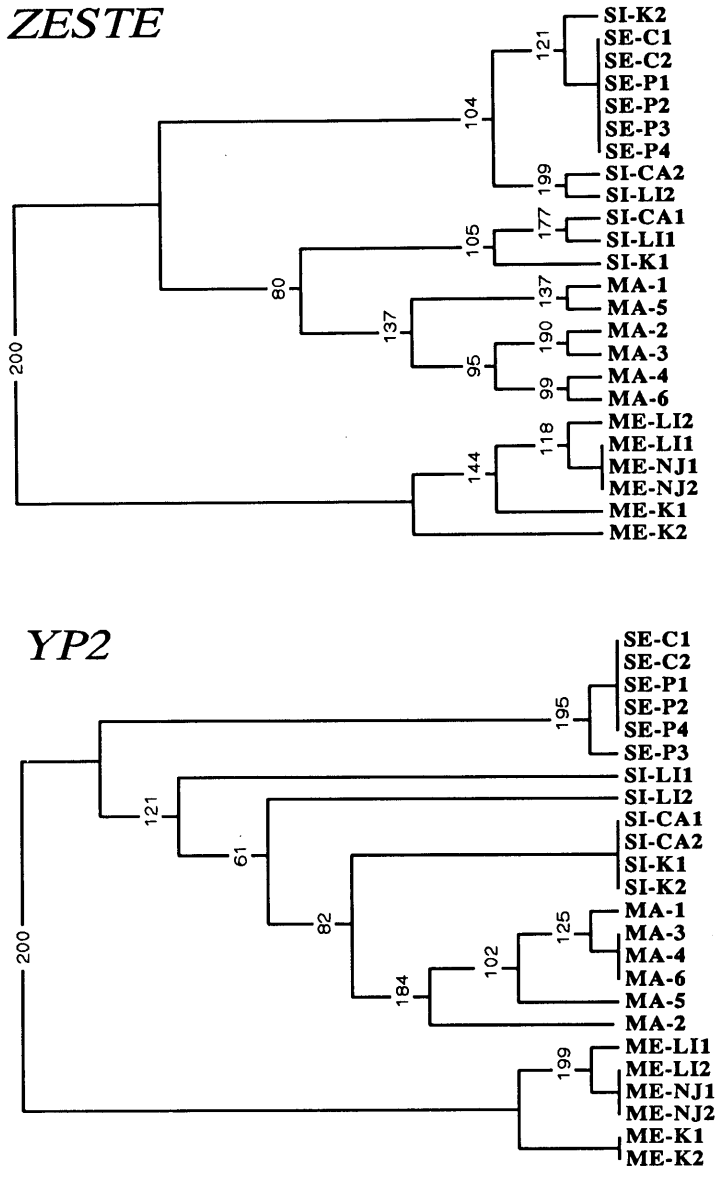

FIG. 6.—Majority-rule consensus trees based on 200 neighbor-joining trees from bootstrapped data sets. Branches that did not occur in 50% of the trees, but that are consistent with those that did, are included.

itiana sequences, a discrete cluster is supported by both data sets, though more strongly at *yp2* (184 replicates) than at *zeste* (137 replicates). Close examination of all 200 *zeste* replicates revealed three classes of trees: 137 replicates had a discrete *D. mauritiana* cluster positioned among *D. simulans* lineages (as shown in fig. 6); in 32 replicates, *D. simulans* and *D. sechellia* sequences together formed a discrete cluster, while *D. mauritiana* sequences connected to the tree via multiple deep branches between the *D. melanogaster* cluster and the *D. simulans-D. sechellia* cluster; and in 31 replicates, neither the *D. simulans–D. sechellia* cluster nor the *D. mauritiana* cluster was separate from the other.

Divergence patterns of *D. simulans, D. sechellia,* and *D. mauritiana* are consistent for the three genes. Regardless of how measured (table 3), divergence is greatest between *D. sechellia* and *D. mauritiana,* less so between *D. simulans* and *D. sechellia,* and least between *D. simulans* and *D. mauritiana.* This pattern is compatible with the history suggested on the basis of *per* alone (Kliman and Hey 1993); that is, *D. mauritiana* and *D. sechellia* arose independently from an ancestral *D. simulans,* and the two species have diverged while *D. simulans* has changed little since the time of the divergence. Note that, for both the number of fixed differences and net divergence (table 3), 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.* This scenario is also consistent with the trees in figures 5 and 6. *Drosophila mauritiana* and *D. sechellia* form clusters that arise from different points in the tree, and most of the deepest nodes of the tree (excepting the probable root) are among *D. simulans* lineages. Thus any measure of divergence between *D. mauritiana* and *D. sechellia* will include variation present in the ancestral *D. simulans* population.

Accumulating Selected Differences between Species

If some amino acid replacement mutations are favored by natural selection and become fixed within species, we may expect a higher proportion of replacement-site

Table 3
Divergence between Species

Species 1–Species 2	*zeste*	*yp2*	*per*
Gross pairwise divergence:[a]			
simulans-mauritiana	9.4	4.4	34.7
simulans-sechellia	12.2	4.5	33.1
mauritiana-sechellia	13.0	8.2	43.2
melanogaster-simulans	37.5	28.7	64.9
melanogaster-mauritiana	36.2	31.7	74.7
melanogaster-sechellia	37.7	31.2	72.7
Net divergence:[b]			
simulans-mauritiana	3.3	3.2	12.9
simulans-sechellia	8.3	3.7	21.4
mauritiana-sechellia	10.8	7.3	31.3
melanogaster-simulans	32.3	25.2	48.4
melanogaster-mauritiana	34.0	28.1	57.9
melanogaster-sechellia	36.2	28.1	66.1
Fixed differences:[c]			
simulans-mauritiana	1	2	3
simulans-sechellia	4	4	18
mauritiana-sechellia	10	6	21
melanogaster-simulans	29	23	37
melanogaster-mauritiana	32	25	44
melanogaster-sechellia	36	27	60

[a] Mean no. of nucleotide differences between all pairs of sequences within a species (in the case of six sequences for each species, there are 36 sequence comparisons).

[b] Gross pairwise divergence minus the average of the values of average pairwise diversity within each of the two species given in table 2 (Nei 1987, p. 276).

[c] No. of base-pair positions at which all of the sequences from species 1 are different from all of the sequences of species 2.

differences in interspecific contrasts than in intraspecific contrasts. McDonald and Kreitman (1991) devised a straightforward test of whether the relative proportions of replacement and silent substitutions are the same within and between species. In the case of *zeste,* a test cannot be made, because only two replacement polymorphisms (both between species) were found. In the case of *yp2,* there are 11 silent and 4 replacement polymorphisms within species, and there are 13 silent and 7 replacement fixed differences between species. The ratios are similar, and there is no evidence for a departure from neutrality ($G = 0.279$, with 1 degree of freedom; $P = 0.597$). A similar observation was made with *per* (Kliman and Hey 1993), though in that case the proportion of replacement changes is lower (table 1).

Recent Hitchhiking or Balancing Selection

If natural selection discerns functional differences among different copies of a locus within a species, the effect may be to elevate or decrease associated levels of neutral variation. In the case of balancing selection, whereby one or more functional alleles persist in a species for a long period of time, neutral divergence between the functional alleles is expected to accumulate and exceed strictly neutral expectations (Strobeck 1983; Hudson and Kaplan 1988). In the case of recent directional selection, whereby a rare sequence increases in frequency and becomes fixed in the population, linkage will lead to reduced variation on either side of the site of selection (Maynard-Smith and Haigh 1974; Kaplan et al. 1989). It is possible to test whether data from multiple loci and multiple species fit the neutral model, where variation is a function of both the effective population size N and the neutral mutation rate μ. By including multiple loci, among which N is constant for a given species, and multiple species, among which μ is constant for a given locus, estimates of θ are constrained. The results of two-species, three-locus tests (Hudson et al. 1987) are shown in table 4. In no case do the observations appear inconsistent with neutral expectations. Thus the observations of differing levels of polymorphism among loci and among species are consistent with variation in neutral mutation rates and in population sizes, respectively.

In general, the lack of statistical evidence for either balancing selection or recent hitchhiking is expected, given the similar patterns of species differences that are seen

Table 4
HKA Tests for Three Loci and Two Species (Hudson et al. 1987)

	$\hat{\theta}$[a]						
SPECIES 1–SPECIES 2	*per*	*zeste*	*yp2*	$\hat{T}$[b]	$\hat{f}$[c]	X^2[d]	P[e]
melanogaster-simulans	12.03	5.08	3.47	5.07	1.60	4.39	0.355
melanogaster-mauritiana	12.40	4.32	3.86	5.67	1.30	4.00	0.405
melanogaster-sechellia	11.34	4.70	4.54	6.42	0.11	2.34	0.674
simulans-mauritiana	24.25	6.54	2.05	0.52	0.81	0.95	0.918
simulans-sechellia	23.18	7.54	2.13	1.03	0.07	2.82	0.588
mauritiana-sechellia	19.51	4.51	2.70	1.87	0.08	2.15	0.709

[a] Estimate of $3N\mu$ for species 1.

[b] Estimate of the time since the common ancestor of the species, in units of $\frac{3}{2}N$ generations, where N is the effective population size of species 1.

[c] Estimate of the scalar by which estimates of $3N\mu$ for species 1 are multiplied to get those of species 2.

[d] Goodness-of-fit statistic.

[e] Probability of observing an X^2 greater than or equal to the actual value, when a χ^2 distribution with 4 degrees of freedom is assumed.

among loci (table 2 and fig. 4). An exception is *D. melanogaster* variation at *yp2* which, unlike that at *per* and *zeste,* exceeds that of the other species. At *zeste, D. melanogaster* has approximately one-third the variation of *D. simulans,* while, at *yp2, D. melanogaster* has nearly four times the variation seen in *D. simulans.* However, neither locus has a large number of polymorphisms in either species, and an HKA test of the *D. melanogaster–D. simulans* contrast at *yp2* and *zeste* is not significant ($X^2 = 0.2681$; 2 degrees of freedom, $P = 0.262$; Hudson et al. 1987).

Population Structure

The two African *D. melanogaster yp2* sequences differ, at 11 positions, from the North American sequences. This distance is reflected in the depth of a node within the neighbor-joining tree for *D. melanogaster yp2* (figs. 5 and 6). At *per* and *zeste,* the African sequences are also separated from the North American sequences, but the distances are not as great (Kliman and Hey 1993; fig. 5).

Drosophila simulans appears different from *D. melanogaster,* lacking any evidence of population structure. The positions of the two African *D. simulans* sequences, as well as those of the four North American lines, vary greatly among the three loci (figs. 5 and 6; Kliman and Hey 1993).

Discussion

The larger purpose of this work is to assess the evolutionary history of the species in the *Drosophila melanogaster* complex. More specifically, four interrelated topics are addressed: natural selection at these loci, population sizes of the species, speciation processes, and the phylogeny of the species.

Natural Selection

For the present purpose, natural selection can be divided into two classes of effects: (1) the sorting of functional from nonfunctional or deleterious gene copies and (2) the adaptive sorting among functionally different gene copies. In the former, which is expected to occur continuously, the population is steadily purged of detrimental variants. These deleterious variants are thus always rare and have little or no effect on the structure of the genealogy. In the case of adaptive sorting of gene copies (e.g., balancing or directional selection), however, natural selection can have a large effect on the genealogy (see Results).

It is clear that *zeste* and *yp2* have very different functions and have evolved under very different patterns of constraint. The *zeste* product is a regulatory protein that appears highly conserved for amino acid sequence, while *yp2* is primarily an embryo food source, evolving with less constraint on primary structure. Furthermore, *zeste* seems relatively permissive of silent substitutions and has a low level of codon bias, while *yp2* has fewer silent substitutions and a very high level of codon bias (table 1). These differences are likely caused by differing levels and patterns of expression required of these loci (Shields et al. 1988).

Zeste and *yp2* were included in this study without prior expectations of the recent action of balancing or directional selection. In this respect, this study differs from studies of *alcohol dehydrogenase* (Kreitman 1983; Hudson et al. 1987), which, prior to studies of DNA sequence variation (Oakeshott et al. 1982), were thought to be under balancing selection, and from studies of *cubitus interruptus Dominant* (Berry et al. 1991), for which it was thought that recent hitchhiking was likely. Unlike these other studies, we find no evidence for recent balancing or directional selection. This

could be due to a lack of statistical power resulting from either the small number of sequences per species or the relatively short regions sequenced. This absence of power is evident in that the large contrast created by the finding of considerable variation in *D. melanogaster* at *yp2* is not a significant departure from the neutral model. However, this is an exception, and patterns of variation among species are fairly consistent across loci, despite differences in levels and types of variation among loci. The number of sequences per species is also not low when considered from the viewpoint of sampling from a bifurcating genealogy. We can ask how many sequences must be included so that the sample genealogy is an accurate reflection of the genealogy for the entire population. It turns out that only a small number of sequences are needed to have a good chance of obtaining sample genealogies that include the earliest nodes of the genealogy for the entire population (Harding 1971; Felsenstein 1992; Kliman and Hey 1993), and this is especially true if the sequencing stocks have geographically diverse origins, as is the case with the *D. melanogaster* and *D. simulans* samples.

The absence of evidence for adaptive sorting among gene copies at *zeste* and *yp2* simplifies the tasks of addressing issues of population size and speciation. In general, these population-level forces are expected to affect all parts of the genome similarly. The apparently similar genealogies among different loci parsimoniously support the view that these patterns have been largely determined by population-level processes.

Population Sizes

From data on *zeste, yp2,* and *per,* we find *D. simulans* and *D. mauritiana* similarly variable and about twice as variable as *D. melanogaster.* The single exception is that the *D. melanogaster* sample has much more variation at *yp2. Drosophila sechellia* consistently has by far the least variation. If we assume that genetic drift and mutation are the primary agents determining these levels and that locus-specific mutation rates are constant across species, then inferences about the relative population sizes among species should mirror the relative levels of sequence polymorphism. Table 4 lists, for each species pair, the estimated ratio of population sizes, $\hat{f}$. When *D. melanogaster* is assigned an arbitrary value of 1.0, simultaneous solution of the values in table 4 yields scalars of proportionality for the other species: 1.600 for *D. simulans,* 1.296 for *D. mauritiana,* and 0.111 for *D. sechellia.*

The differences between *D. melanogaster* and *D. simulans* are consistent with observations from studies on restriction-fragment-length polymorphisms (Aquadro et al. 1988; Begun and Aquadro 1991). Less expected are the large amount of variation in *D. mauritiana* and the large difference in variation between *D. mauritiana* and the other island endemic, *D. sechellia.* It is reported that *D. mauritiana* may be more numerous and widespread than *D. sechellia* (Lachaise et al. 1988; also see Speciation below). Still unexplained is why *D. mauritiana* apparently has an effective population size comparable to that of *D. simulans* and larger than that of *D. melanogaster.*

Speciation

To briefly summarize part of a recent review of the biogeography of this species complex (Lachaise et al. 1988), we note that the two cosmopolitan species, *D. melanogaster* and *D. simulans,* spread out of western and eastern Africa, respectively; *D. mauritiana* is largely limited to the island of Mauritius (1,865 km^2), where it is common, though a few individuals have been collected on Rodriguez Island 500 km to the east; *D. sechellia* is limited to a few small islands of the Seychelles, where it is apparently restricted to a single host, *Morinda citrifolia. Drosophila simulans* has not

been found on Mauritius, but it has been found on one of the islands occupied by *D. sechellia* (Cariou et al. 1990; R'Kha et al. 1991).

In general, the DNA sequence data support the species designations. Except for *D. simulans,* genealogies estimated on the basis of all three loci (figs. 5 and 6; Kliman and Hey 1993) show that sequences cluster by species. However, this is a poor test of species designations, as an intermixed genealogical pattern could also have been consistent with recently formed species (Tajima 1983; Coyne and Kreitman 1986). The sequence data do imply an absence of gene flow between the species.

In the case of *D. simulans,* sequences do not form discrete clusters in the estimated genealogies, and several of the nodes appear earlier in the trees than do those that distinguish *D. mauritiana* and *D. sechellia.* Both the structure of these trees and the levels of divergence among these species (table 3) support (*a*) a model in which *D. mauritiana* and *D. sechellia* arose independently from ancestral *D. simulans* and (*b*) the inference that modern *D. simulans* has changed relatively little since that time. In light of the apparent clustering within the other species, the pattern of early nodes among *D. simulans* lineages might suggest that this sample includes sequences from multiple "cryptic" species. However, the *D. simulans* data from all three loci came from the same six chromosomes and together reveal considerable recombination. Among the *D. simulans per* lineages, there appear to have been a minimum of seven recombination events (Kliman and Hey 1993). Also, the topologies of the estimated genealogies for the six *D. simulans* chromosomes are very different for the different loci (Kliman and Hey 1993; figs. 5 and 6), indicating recombination between the loci.

Given that *D. mauritiana* and *D. sechellia* are endemic to different oceanic islands, a plausible model for their formation is that they diverged from ancestral *D. simulans* after the isolation of a small number of "founder" individuals (Mayr 1954; Carson 1975; Templeton 1980). *Drosophila mauritiana* currently exhibits a level of sequence variation similar to that of *D. simulans,* indicating a relatively large effective population size. It is possible that the population size of *D. mauritiana* expanded greatly after an initial founder event, but many of the *D. mauritiana* polymorphisms (11 in the case of *per* and 1 in the case of *yp2*) are shared with *D. simulans.* If it is assumed that these polymorphisms are identical by descent, then they must have persisted through the initial isolation of *D. mauritiana.* In this light, a very small effective population size of *D. mauritiana* at any time in its history seems unlikely. The repeated observation of very little variation in *D. sechellia* indicates a small effective population size for this species. However, we can not distinguish between a recent population bottleneck and the case where the population size has been small since the species formation.

Phylogeny

The *zeste* and *yp2* data support a relatively ancient split of *D. melanogaster* from the other species (Bodmer and Ashburner 1984; Cohn et al. 1984; Coyne and Kreitman 1986; Caccone et al. 1988; Lachaise et al. 1988). The data, like those from *per* (Kliman and Hey 1993), also do not support *D. mauritiana* and *D. sechellia* as being the most closely related species pair. However, despite having estimated genealogies and six sequences for each species, we are uncertain about the relative timing of the origin of *D. mauritiana* and *D. sechellia* from ancestral *D. simulans.*

By considering the estimated genealogies together with the results of the Hudson-Kreitman-Aguade (HKA) tests, we find some evidence that *D. sechellia* separated before *D. mauritiana.* The HKA test provides an estimate of the time of divergence

between two species. From table 4 the times of the splits involving *D. mauritiana* and *D. sechellia* are 0.52 and 1.03, respectively. These times are in units of 3/2 *N* generations (3/2 rather than 2 because the loci are sex linked), where *N*, in this case, is the effective population size of *D. simulans*.

Estimated speciation dates can be obtained by using dated fossils and by assuming evolutionary-rate constancy. The data of the Sophophoran radiation of the genus *Drosophila* may be ~30–35 Mya, on the basis of a small number of Oligocene-Miocene fossils (Throckmorton 1975). Sharp and Li (1989) conservatively assumed that it occurred 40 Mya and estimated the substitution rate for silent sites in *Drosophila* at 16×10^{-9}/silent site/year for low-bias genes and as half that for high-bias genes. By taking the average of these values (12×10^{-9}) and applying it to the divergence observed at silent sites of *per, yp2,* and *zeste,* we can estimate the times of divergence of *D. mauritiana* and *D. sechellia* from *D. simulans*. For the divergence between *D. mauritiana* and *D. simulans,* net divergence (Nei 1987, p. 276) for all silent sites is as follows: *per,* 9.73; *yp2,* 2.06; and *zeste,* 2.71. The same calculations for *D. sechellia* and *D. simulans* yield the following values: *per,* 11.38; *yp2,* 1.40; and *zeste,* 3.40. The weighted average per silent site (i.e., summed and divided by the total number of silent sites listed in table 1) is 0.0184 for the *D. mauritiana* divergence and 0.0205 for the *D. sechellia* divergence. With the rate from Sharp and Li (1989), these values lead to estimates of 0.77 Myr since the *D. mauritiana* divergence and 0.86 Myr since the *D. sechellia* divergence. These values are based on a value of 40 Myr since the Sophophoran radiation and would be reduced by three-fourths if a date of 30 Myr is considered to be more accurate. This analysis can also be applied to the divergence between *D. melanogaster* and *D. simulans*. In this case the net divergence values are as follows: *per,* 35.2; *yp2,* 12.9; and *zeste,* 16.8. The weighted average number of substitutions per silent site is 0.0825; and the estimated time since divergence began is 3.4 Myr for the 40-Mya date and is 2.55 Myr if scaled to a 30-Mya date.

Sequence Availability

GenBank accession numbers for the *zeste* sequences are L13043–L13066; and those for the *yp2* sequences are L14417–L14428.

Acknowledgements

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

LITERATURE CITED

AGUADÉ, 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.

ASHBURNER, M. 1989. *Drosophila:* a laboratory manual. Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

BARNETT, T., C. PACHL, J. P. GERGEN, and P. C. WENSINK. 1980. The isolation and characterization of *Drosophila* yolk protein genes. Cell **21**:729–738.

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.

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.

BENNETZEN, J. L., and B. D. HALL. 1982. Codon selection in yeast. J. Biol. Chem. **247**:3026–3031.

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.

BOWNES, M., K. LINERUTH, and D. MAUCHLINE. 1991. Egg production and fertility in *Drosophila* depend upon the number of yolk-protein gene copies. Mol. Gen. Genet. **228**:324–327.

BRENNAN, M. D., A. J. WEINER, T. GORALSKI, and A. P. MAHOWALD. 1982. The follicle cells are a major site of vitellogenin synthesis in *Drosophila melanogaster.* Dev. Biol. **89**:225–236.

BRIDGES, C. B. 1938. A revised map of the salivary gland X chromosome of *Drosophila melanogaster.* J. Hered. **29**:11–13.

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

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.

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.

EWENS, W. J. 1979. Mathematical population genetics. Springer, New York.

FELSENSTEIN, J. 1989. PHYLIP: phylogeny inference package, version 3.2. Cladistics **5**:164–166.

———. 1992. Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. Genet. Res. **59**:139–147.

GANS, M. 1953. Étude génetique et physiologique du mutant *z* de *Drosophila melanogaster.* Bull. Biol. Fr. Belge Suppl. **38**:1–90.

GAVIN, A. G., and J. H. WILLIAMSON. 1976. Synthesis and deposition of yolk protein in adult *Drosophila melanogaster.* J. Insect Physiol. **22**:1457–1464.

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

HUDSON, R. R. 1990. Gene genealogies and the coalescent process. Pp. 1–44 *in* P. H. HARVEY and L. PARTRIDGE, eds. Oxford surveys in evolutionary biology. Vol. 7. Oxford University Press, New York.

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. AGUADÉ. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

HUNG, M.-C., and P. C. WENSINK. 1983. Sequence and structure conservation in yolk proteins and their genes. J. Mol. Biol. **164**:481–492.

JUDD, B. H., M. W. SHEN, and T. C. KAUFMAN. 1972. The anatomy and function of a segment of the X chromosome of *Drosophila melanogaster.* Genetics **71**:139–156.

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.

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

KLIMAN, R. M., and J. HEY. 1993. DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. Genetics **133**:375–387.

KREITMAN, M. 1983. Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304**:412–417.

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.

LEFEVRE, G., JR. 1971. Salivary chromosome bands and the frequency of crossing over in *Drosophila melanogaster*. Genetics **67**:497–513.

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

MANSUKHANI, A., P. H. GUNARATNE, P. W. SHERWOOD, B. J. SNEATH, and M. L. GOLDBERG. 1988. Nucleotide sequence and structural analysis of the *zeste* locus of *Drosophila melanogaster*. Mol. Gen. Genet. **211**:121–128.

MARIANI, C., V. PIRROTTA, and E. MANET. 1985. Isolation and characterization of the *zeste* locus of *Drosophila*. EMBO J. **4**:2045–2052.

MAYNARD-SMITH, J., and J. HAIGH. 1974. The hitchhiking effect of a favorable gene. Genet. Res. **23**:23–25.

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

NEI, M. 1987. Molecular evolutionary genetics. Columbia University Press, N.Y.

OAKESHOTT, J. G., J. B. GIBSON, P. R. ANDERSON, W. R. KNIBB, D. G. ANDERSON, and G. K. CHAMBERS. 1982. Alcohol dehydrogenase and glycerol-3-phosphate dehydrogenase clines in *Drosophila melanogaster* on three continents. Evolution **269**:491–493.

PIRROTTA, V., S. BICKEL, and C. MARIANI. 1988. Developmental expression of the *Drosophila zeste* gene and localization of *zeste* protein on polytene chromosomes. Genes Dev. **2**:1839–1850.

PIRROTTA, V., E. MANET, E. HARDON, S. F. BICKEL, and M. BENSON. 1987. Structure and sequence of the *Drosophila zeste* gene. EMBO J. **6**:791–799.

RIDDELL, D. C., M. J. HIGGINS, B. J. MCMILLAN, and B. N. WHITE. 1981. Structural analysis of the three vitellogenin genes in *Drosophila melanogaster*. Nucleic Acids Res. **9**:1323–1338.

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.

SHARP, P. M., and W.-H. LI. 1987. The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. **15**:1281–1295.

———. 1989. On the rate of DNA sequence evolution in *Drosophila*. J. Mol. Evol. **28**:398–402.

SHARP, P. M., T. M. F. TUOHY, and K. R. MOSURSKI. 1986. Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res. **14**:5125–5143.

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. "Silent" sites in *Drosophila*

genes are not neutral: evidence of selection among synonymous codons. Mol. Biol. Evol. **5:** 704–716.

SMITH, R. F., and R. J. KONOPKA. 1981. Circadian clock phenotypes of chromosome aberrations with a breakpoint at the *per* locus. Mol. Gen. Genet. **183:**243–251.

SPIERER, P., A. SPIERER, W. BENDER, and D. HOGNESS. 1983. Molecular mapping of genetic and chromomeric units in *Drosophila melanogaster.* J. Mol. Biol. **168:**35–50.

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.

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

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

THROCKMORTON, L. H. 1975. The phylogeny, ecology and geography of *Drosophila.* Pp. 421–469 *in* R. C. KING, ed. Handbook of genetics. Vol. **3.** Plenum, New York.

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

YOUNG, M. W., and B. H. JUDD. 1978. Nonessential sequences, genes, and the polytene chromosome bands of *Drosophila melanogaster.* Genetics **88:**723–742.

MARTIN KREITMAN, reviewing editor

Received October 18, 1992; revision received February 16, 1993

Accepted February 16, 1993