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

**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.

© 1993 by The University of Chicago. All rights reserved.
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

\[
\text{zeste}
\]

5' \begin{array} {c} \hline \end{array} 3'

\[
\text{yp2}
\]

5' \begin{array} {c} \hline \end{array} 3'

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.
**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.
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*. *yp2* 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

---

**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.
Table 1
Levels of Variation at Different Loci

<table>
<thead>
<tr>
<th>LOCUS</th>
<th>INTRON LENGTH</th>
<th>EXON LENGTH</th>
<th>SILENT SITES*</th>
<th>CHANGESb</th>
<th>Intron</th>
<th>Synonymous</th>
<th>Replacement</th>
<th>S</th>
<th>R</th>
<th>CAI</th>
<th>C^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>zeste</td>
<td>182</td>
<td>805</td>
<td>167</td>
<td>24</td>
<td>41</td>
<td>2</td>
<td>0.132</td>
<td>0.246</td>
<td>0.003</td>
<td>0.467</td>
<td>42.1</td>
</tr>
<tr>
<td>yp2</td>
<td>63</td>
<td>1051</td>
<td>234</td>
<td>9</td>
<td>24</td>
<td>11</td>
<td>0.143</td>
<td>0.103</td>
<td>0.014</td>
<td>0.697</td>
<td>33.7</td>
</tr>
<tr>
<td>per</td>
<td>192</td>
<td>1679</td>
<td>386</td>
<td>40</td>
<td>115</td>
<td>12</td>
<td>0.208</td>
<td>0.298</td>
<td>0.009</td>
<td>0.490</td>
<td>36.6</td>
</tr>
</tbody>
</table>

* Calculated by considering, for each base position of the SI-CA1 sequence, the fraction of possible base changes (\( \frac{1}{4}, \frac{1}{2}, \frac{3}{4} \)) 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).

S No. of intron variable sites divided by total intron length.

R No. of synonymous variable sites divided by no. of silent sites.

C AI 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.

CAI Effective no. of codons, calculated from SI-CA1 according to the method of Wright (1990). Note that higher values reflect lower codon bias.

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 \( \theta \), which is equal to four times the product of the effective population size (\( N \)) and the neutral mutation rate (\( \mu \)). 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 \( \theta \) 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.
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 ($\bar{\theta}$) with Estimated Differences

<table>
<thead>
<tr>
<th>Locus and Species</th>
<th>Exon</th>
<th>Introns</th>
<th>Total</th>
<th>$\bar{\theta}$</th>
<th>$S_{aa}^a$</th>
<th>$S_{ab}^b$</th>
<th>$\bar{\theta}$</th>
<th>$S_{aa}^a$</th>
<th>$S_{ab}^b$</th>
<th>$\hat{\theta}$</th>
<th>PCR Base Pair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zeste:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simulans</td>
<td>0.9</td>
<td>1.8</td>
<td>2.0</td>
<td>2.9</td>
<td>1.2</td>
<td>1.7</td>
<td>7.8</td>
<td>2.8</td>
<td>4.0</td>
<td>0.0078</td>
<td></td>
</tr>
<tr>
<td>sechellia</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>mauritiana</td>
<td>2.9</td>
<td>1.1</td>
<td>1.7</td>
<td>1.7</td>
<td>0.7</td>
<td>1.1</td>
<td>4.5</td>
<td>4.5</td>
<td>2.5</td>
<td>0.0045</td>
<td></td>
</tr>
<tr>
<td>melanogaster</td>
<td>2.5</td>
<td>1.0</td>
<td>1.5</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>2.5</td>
<td>2.5</td>
<td>1.5</td>
<td>0.0025</td>
<td></td>
</tr>
<tr>
<td>YP2:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simulans</td>
<td>0.9</td>
<td>0.4</td>
<td>0.7</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>1.2</td>
<td>0.6</td>
<td>0.8</td>
<td>0.0011</td>
<td></td>
</tr>
<tr>
<td>sechellia</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>0.0003</td>
<td></td>
</tr>
<tr>
<td>mauritiana</td>
<td>1.0</td>
<td>0.5</td>
<td>0.7</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
<td>1.3</td>
<td>0.6</td>
<td>0.9</td>
<td>0.0012</td>
<td></td>
</tr>
<tr>
<td>melanogaster</td>
<td>4.7</td>
<td>1.7</td>
<td>2.5</td>
<td>1.1</td>
<td>0.5</td>
<td>0.8</td>
<td>5.7</td>
<td>2.1</td>
<td>3.0</td>
<td>0.0052</td>
<td></td>
</tr>
<tr>
<td>Per:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>simulans</td>
<td>18.0</td>
<td>6.2</td>
<td>8.8</td>
<td>3.5</td>
<td>1.4</td>
<td>2.0</td>
<td>21.5</td>
<td>7.4</td>
<td>10.5</td>
<td>0.0115</td>
<td></td>
</tr>
<tr>
<td>sechellia</td>
<td>1.2</td>
<td>0.6</td>
<td>0.8</td>
<td>0.5</td>
<td>0.3</td>
<td>0.5</td>
<td>1.7</td>
<td>0.7</td>
<td>1.1</td>
<td>0.0009</td>
<td></td>
</tr>
<tr>
<td>mauritiana</td>
<td>18.3</td>
<td>6.3</td>
<td>9.0</td>
<td>3.7</td>
<td>1.4</td>
<td>2.1</td>
<td>22.1</td>
<td>7.5</td>
<td>10.7</td>
<td>0.0118</td>
<td></td>
</tr>
<tr>
<td>melanogaster</td>
<td>8.4</td>
<td>3.0</td>
<td>4.3</td>
<td>3.3</td>
<td>1.3</td>
<td>1.9</td>
<td>11.7</td>
<td>4.1</td>
<td>5.9</td>
<td>0.0062</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Under the assumptions of no recombination, a Wright-Fisher demographic model (Ewens 1979, p. 16), and an infinite-sites model (Kimura 1969), these values may be taken as estimates of $\bar{\theta} = 3N_\mu$ (see text). The error estimates are made under the same assumptions (Tajima 1983).

$^a S_{aa} = $ 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_{ab} = $ 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 (Swoford 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. mauri-
ZESTE

YP2

FIG. 5.—Neighbor-joining trees
Drosophila DNA Sequence Variation

**ZESTE**

![ZESTE tree diagram]

**YP2**

![YP2 tree diagram]

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>
<thead>
<tr>
<th>Table 3</th>
<th>Divergence between Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Species 1–Species 2</td>
<td>zest</td>
</tr>
<tr>
<td>Gross pairwise divergence</td>
<td>9.4</td>
</tr>
<tr>
<td>simulans-mauritiana</td>
<td>simulans-sechellia</td>
</tr>
<tr>
<td>mauritiana-sechellia</td>
<td>melanogaster-simulans</td>
</tr>
<tr>
<td>13.0</td>
<td>37.5</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>12.2</td>
<td>31.7</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>37.5</td>
<td>31.7</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>36.2</td>
<td>31.7</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>36.2</td>
<td>31.7</td>
</tr>
<tr>
<td>Net divergence</td>
<td>3.3</td>
</tr>
<tr>
<td>simulans-mauritiana</td>
<td>simulans-sechellia</td>
</tr>
<tr>
<td>mauritiana-sechellia</td>
<td>melanogaster-simulans</td>
</tr>
<tr>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>3.3</td>
<td>3.2</td>
</tr>
<tr>
<td>Fixed differences</td>
<td>1</td>
</tr>
<tr>
<td>simulans-mauritiana</td>
<td>simulans-sechellia</td>
</tr>
<tr>
<td>mauritiana-sechellia</td>
<td>melanogaster-simulans</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>10</td>
<td>6</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>29</td>
<td>23</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>32</td>
<td>25</td>
</tr>
<tr>
<td>melanogaster-simulans</td>
<td>melanogaster-mauritiana</td>
</tr>
<tr>
<td>36</td>
<td>27</td>
</tr>
</tbody>
</table>

* 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).

* 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).

* 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 $\mu$. By including multiple loci, among which $N$ is constant for a given species, and multiple species, among which $\mu$ is constant for a given locus, estimates of $\theta$ 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)**

<table>
<thead>
<tr>
<th>Species 1–Species 2</th>
<th>per</th>
<th>zeste</th>
<th>yp2</th>
<th>$\hat{T}$</th>
<th>$\hat{f}$</th>
<th>$\chi^2$</th>
<th>$P$</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>melanogaster-simulans</em></td>
<td>12.03</td>
<td>5.08</td>
<td>3.47</td>
<td>5.07</td>
<td>1.60</td>
<td>4.39</td>
<td>0.355</td>
</tr>
<tr>
<td><em>melanogaster-mauritiana</em></td>
<td>12.40</td>
<td>4.32</td>
<td>3.86</td>
<td>5.67</td>
<td>1.30</td>
<td>4.00</td>
<td>0.405</td>
</tr>
<tr>
<td><em>melanogaster-sechellia</em></td>
<td>11.34</td>
<td>4.70</td>
<td>4.54</td>
<td>6.42</td>
<td>0.11</td>
<td>2.34</td>
<td>0.674</td>
</tr>
<tr>
<td><em>simulans-mauritiana</em></td>
<td>24.25</td>
<td>6.54</td>
<td>2.05</td>
<td>0.52</td>
<td>0.81</td>
<td>0.95</td>
<td>0.918</td>
</tr>
<tr>
<td><em>simulans-sechellia</em></td>
<td>23.18</td>
<td>7.54</td>
<td>2.13</td>
<td>1.03</td>
<td>0.07</td>
<td>2.82</td>
<td>0.588</td>
</tr>
<tr>
<td><em>mauritiana-sechellia</em></td>
<td>19.51</td>
<td>4.51</td>
<td>2.70</td>
<td>1.87</td>
<td>0.08</td>
<td>2.15</td>
<td>0.709</td>
</tr>
</tbody>
</table>

* Estimate of $3N\mu$ for species 1.

* Estimate of the time since the common ancestor of the species, in units of $1/2N$ generations, where $N$ is the effective population size of species 1.

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

* Goodness-of-fit statistic.

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

Population Structure

The two African \textit{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 \textit{D. melanogaster yp2} (figs. 5 and 6). At \textit{per} and \textit{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).

\textit{Drosophila simulans} appears different from \textit{D. melanogaster}, lacking any evidence of population structure. The positions of the two African \textit{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 \textit{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 \textit{zeste} and \textit{yp2} have very different functions and have evolved under very different patterns of constraint. The \textit{zeste} product is a regulatory protein that appears highly conserved for amino acid sequence, while \textit{yp2} is primarily an embryo food source, evolving with less constraint on primary structure. Furthermore, \textit{zeste} seems relatively permissive of silent substitutions and has a low level of codon bias, while \textit{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).

\textit{Zeste} and \textit{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 \textit{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 \textit{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, \( \frac{N_{D. melanogaster}}{N_{D. simulans}} \). 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 km²), 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 $3/2N$ 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 $16 \times 10^{-9}$/silent site/year for low-bias genes and as half that for high-bias genes. By taking the average of these values ($12 \times 10^{-9}$) 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


BEGIN, 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.


Drosophila DNA Sequence Variation 821

SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. “Silent” sites in Drosophila
822 Hey and Kliman


MARTIN KREITMAN, reviewing editor

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

Accepted February 16, 1993