Molecular Polymorphism in the *period* Gene of *Drosophila simulans*

Ezio Rosato,*,† Alexandre A. Peixoto,† Guido Barbujani,‡ Rodolfo Costa* and
Charalambos P. Kyriacou†

*Dipartimento di Biologia, Università di Padova, 35121 Padova, Italy, †Department of Genetics, University of Leicester,
Leicester LE1 7RH, United Kingdom, and ‡Dipartimento di Scienze Statistiche, Università di Bologna, Bologna, Italy

ABSTRACT

The threonine-glycine (Thr-Gly) repeat region of the *period* (*per*) gene of eight natural populations of *Drosophila simulans* from Europe and North Africa was analyzed by polymerase chain reaction, DNA sequencing and heteroduplex formation. Five different length alleles encoding 21, 23, 25 and two different kinds of 24 Thr-Gly pairs in the uninterrupted repeat were found. In the 3' region flanking the repeat 6 nucleotide substitutions (3 synonymous, 3 replacement) were observed in three different combinations that we called haplotypes I, II and III. The complete linkage disequilibrium observed between the haplotypes and these length variants allowed us to infer from the repeat length, the DNA sequence at the 3' polymorphic sites. The haplotypes were homogeneously distributed across Europe and North Africa. The data show statistically significant departures from neutral expectations according to the Tajima test. The results suggest that balancing selection might have played a role in determining the observed levels and patterns of genetic diversity at the *per* gene in *D. simulans*.

THE sex-linked *period* gene in Drosophila represents the best studied "behavioral" gene in any eukaryotic organism. Mutations in this gene disrupt the fly's normal 24-hr circadian behavior (KONOPKA and BENZER 1971), and have corresponding effects on both the male's 60-sec ultradian lovesong rhythm (KYRIACOU and HALL 1980, 1989), and the infradian 10-day developmental cycle (KYRIACOU *et al.* 1990). The *period* gene has been cloned and sequenced in *Drosophila melanogaster* (BARGIELLO and YOUNG 1984; REDDY *et al.* 1984; JACKSON *et al.* 1986; CITRI *et al.* 1987), *Drosophila pseudoobscura* and *Drosophila virilis* (COLOT *et al.* 1988), *Drosophila yakuba* (THACKERAY and KYRIACOU 1990), and *Musca domestica* (A. PICCIN, personal communication). Comparison of these sequences reveals a patchwork of conserved areas interspersed with highly variable regions. One of the variable regions is in the middle of exon 5 and encodes a series of about 20 threonine-glycine (Thr-Gly) repeats in *D. melanogaster* (CITRI *et al.* 1987). The region is polymorphic in length in both laboratory and natural populations (YU *et al.* 1987; COSTA *et al.* 1991), but the major length variants in Europe appear to be the (Thr-Gly)$_{17}$ and the (Thr-Gly)$_{20}$ alleles (COSTA *et al.* 1992). These two variants show a highly significant geographical differentiation in Europe, producing a robust latitudinal cline with the (Thr-Gly)$_{20}$ and (Thr-Gly)$_{17}$ alleles predominant in Northern and Southern Europe, respectively (COSTA *et al.* 1992).

The Thr-Gly repeat is found in all species of the *melanogaster* subgroup (PEIXOTO *et al.* 1992), although length of the motif differs between species. In other species, however, the Thr-Gly encoding repeat is very small, or a different repeat is present which also shows length variation, as in *D. pseudoobscura* (COLOT *et al.* 1988; COSTA *et al.* 1991; PEIXOTO *et al.* 1993; J. NIELSEN, A. A.

PEIXOTO, A. PICCIN, R. COSTA, C. P. KYRIACOU and D. CHALMERS, manuscript submitted for publication). Remarkably the Neurospora clock gene, *frequency* (*frq*) also has a small Thr-Gly repeat and some flanking similarities to *per* (MCCLUNG *et al.* 1989).

WHEELER *et al.* (1991) reported length variation in the Thr-Gly repeat region of *D. simulans*, together with a surprising amount of amino acid polymorphism immediately 3' to the Thr-Gly repeat. The interspecific amino acid differences in the 3'-flanking region appeared to be responsible for the species-specific differences between *D. melanogaster* and *D. simulans* in the male lovesong cycle (WHEELER *et al.* 1991). Furthermore, in *D. melanogaster*, the removal of the uninterrupted Thr-Gly repeat resulted in a marked change in the period of the lovesong cycle (YU *et al.* 1987) as well as in a loss of thermostability of the circadian locomotor activity phenotype (EWER *et al.* 1990). Consequently this region of the *per* gene appears to be important both in determining features of species-specific mating behavior, and in affecting critical functional characteristics of the circadian locomotor activity cycle, with obvious evolutionary implications (ASHBURNER 1987; PETERSEN *et al.* 1988; COYNE 1992).

Given the interest in this part of the *per* gene, especially between the sibling species *D. simulans* and *D. melanogaster*, we have examined in detail the Thr-Gly and flanking regions in several natural populations of *D. simulans* from Europe and North Africa. We observe that there are two major haplotypes in each population which are homogeneously distributed across Europe and North Africa. The geographical pattern appears quite different from the clinal distribution observed in the length polymorphism of the Thr-Gly repeat in *D. melanogaster*. Moreover, our results indicate that the

```
PVHEGSGGSGSSGNFTTASNIHMSSVTNTSIAGTGG - Thr-Gly repeat - NGTNSGTGTGTTSSSRGGSTAIPPVTLTESLLNKHNDEMEKFMLKKHRESRG
------->           ------>                              <-----                    <------  <-------
    5'-external primers                                 3'-internal primer        3'-external primers
```

FIGURE 1.—Amino acid sequence corresponding to the PCR-amplified fragments of the *per* gene in *D. simulans*. The approximate relative positions of the adopted primers are indicated. The amino acid sequence refers to the *sim-Aus* variant (WHEELER *et al.* 1991; PEIXOTO *et al.* 1992). The italicized *TGTG* motif immediately downstream of the 3' internal primer is not present in haplotypes II and III (see text).

pattern of nucleotide polymorphism in the region flanking the repeat does not correspond to the predictions of a neutral model of populations at equilibrium. We suggest that natural selection, perhaps in association with other factors, may have been important in determining the pattern of genetic variation at the *per* locus.

MATERIALS AND METHODS

Natural populations of *D. simulans*: *D. simulans* samples were collected from seven different localities within Europe plus one from Egypt. Samples from Andros (Greece) and Agios Nikolaos (Greece) were collected from the wild in September 1990. The samples from Peniscola (Spain) and Merano (Italy) were collected from the wild in October 1991. The flies were immediately frozen and DNA analysis performed on these individuals. In the case of the Patras (Greece) population, a random sample was taken from the F_1 progeny of more that 200 individuals collected from the wild in October 1991. Two individuals were also collected from Lyon (France) in October 1991. A number of isofemale lines were established immediately from a population collected in Antibes (France) in October 1991, and also from a population collected in Tanta, near Alexandria (Egypt) in October 1992. One male from each isofemale line from these two populations was crossed to females from a *D. simulans* attached-*X* chromosome line (XX*yw*/Y Stock No. 0251.17, Bowling Green Stock Center), and one male from the progeny of each attached-*X* line was studied. Finally, 16 females from a population taken from Gran Ferade in Bordeaux (France) in 1992, were used to set up isofemale lines, one male of each line being analyzed. Fifteen further males from the original wild population were immediately frozen and used for further analysis. In all, eight natural populations were studied, plus two individuals from Lyon.

Polymerase chain reaction (PCR) amplification, gel electrophoresis and DNA sequencing: Single fly DNA was used as a template in the PCR reactions and was prepared using the method of GLOOR and ENGELS (1990). PCR amplification was carried out according to JEFFREYS *et al.* (1988b) for 30 cycles (95° for 1 min, 65° for 1 min, and 70° for 1 min) either in a Perkin-Elmer Cetus or a MJ Research thermocycler. Double-stranded direct DNA sequencing was carried out using the Sequenase version 2.0 kit from U.S. Biochemical Corp. AmpliTaq polymerases from Perkin-Elmer Cetus or Replitherm polymerase from Epicentre Technologies were used.

The primers used for the PCR were the following: 5' primers were 5'-CCCGTCCACGAGGGCAGCGGGGGC-3' (5006–5029) and 5'-ATACACATGAGCAGTGTGAC-3' (5066–5085); 3' primers were 5'-CCGCGCGACTCCCGGTGCTTCTTC-3' (5365–5388) and 5'-TTCTCCATCTCGTCGTTGTG-3' (5336–5355).

A 3' internal primer, 5'-GTGCCGGAATTCGTTCCATT-3' (5234–5253) was also used to check for length variation outside the uninterrupted Thr-Gly encoding repeat (see Figure 1). The primer positions (in parentheses) refer to the *D. melanogaster* sequence published by CITRI *et al.* (1987).

PCR amplified DNAs were electrophoresed through a 3.5% low melting point NuSieve (GTG) agarose gel. TBE buffer

(0.045 M Tris borate, 0.001 M EDTA, pH 8.3) was used, and PCR-amplified DNAs from *D. melanogaster* stocks carrying 17, 20 or 23 uninterrupted pairs of Thr-Gly encoding repeats were used as markers (COSTA *et al.* 1991).

Heteroduplex characterization of Thr-Gly length variation: First, each *D. simulans* Thr-Gly length variant was analyzed by agarose gel-electrophoresis after two independent PCR amplifications, one using a 3' external, and the other using the 3' internal primer (Figure 1).

Second, the length variants defined above were further characterized by co-amplification of each DNA, with a previously sequenced DNA from an isolength Thr-Gly allele. Co-amplification was made using a 1:1 ratio of the two DNAs. If a co-amplification produced a heteroduplex, this indicated a difference in the DNA sequence of the variant *vs.* the isolength standard. When this occurred, the new isolength variant was sequenced, and subsequently used as a standard for further co-amplifications against all the other isolength alleles. This procedure allowed us to validate whether our isolength sequences have the same or a different pattern in the Thr-Gly encoding ACN GGN codons.

RESULTS

PCR and DNA sequence analysis: We analyzed the Thr-Gly encoding repeat region of 212 single flies from the eight natural populations of *D. simulans*, plus 2 individuals from Lyon. The first approach, using the 5' and the 3' external primers (Figure 1) resulted in a grouping of all the length variants into two electrophoretic classes, the "(Thr-Gly)$_{22}$-like" and "(Thr-Gly)$_{23}$-like" (Figure 2A). The only exceptions were two identical shorter variants from the Antibes population (see below).

Further PCR amplification using the 3' internal primer revealed a more complex pattern, due to a deletion occurring downstream of the 3' internal primer and, consequently, upstream of the 3' external primers (see Figure 2B). The sequencing of these variants revealed that the polymorphism was due to the deletion of 12 nucleotides encoding a "Thr-Gly-Thr-Gly" motif, 6 amino acids downstream of the main uninterrupted Thr-Gly repeat (Figure 1). Length variation due to this deletion has been previously described in this region of the *per* gene of *D. simulans* (WHEELER *et al.* 1991; PEIXOTO *et al.* 1992). Our results show that this deletion is common in *D. simulans* (see Figures 3 and 4), whereas in *D. melanogaster* it is rarely observed (COSTA *et al.* 1992).

The combined use of the 3' internal and external primers led to a new classification of the length variation. All the variants initially classified as "(Thr-Gly)$_{22}$-like" are in fact, uninterrupted (Thr-Gly)$_{24}$ alleles with
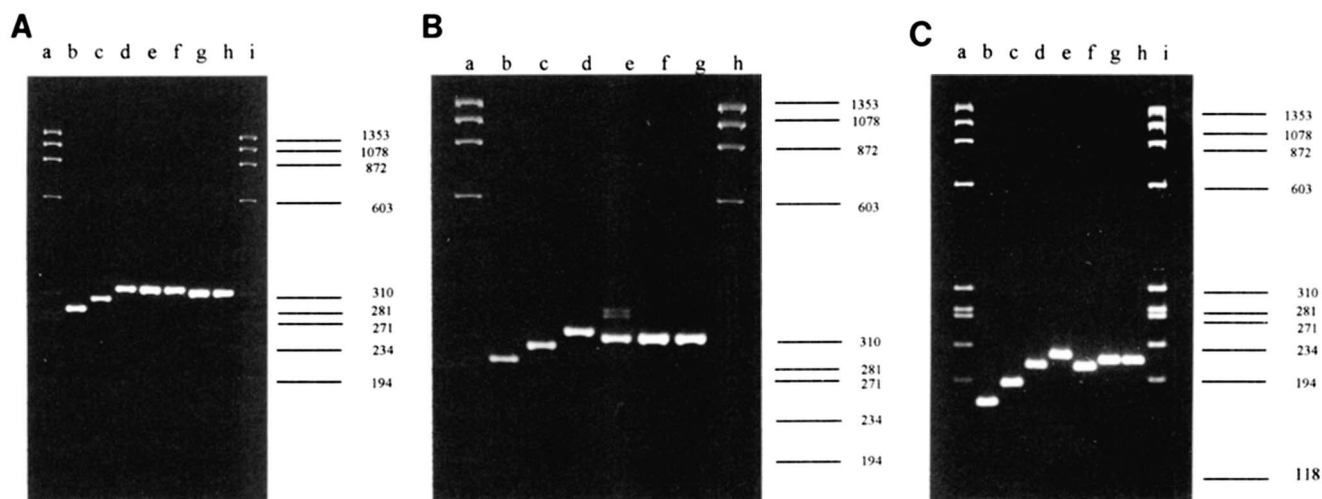
FIGURE 2.—PCR-amplified Thr-Gly length variants separated on a 3.5% low melting point gel (NuSieve). The *D. melanogaster* (Thr-Gly)$_{23}$, (Thr-Gly)$_{20}$, (Thr-Gly)$_{17}$ variants are shown for comparison. (**A**) Amplification with the external primers 5'-ATACACATGAGCAGTGTGAC-3' and 5'-TTCTCCATCTCGTCGTTGTG-3' showing the "22-like" and "23-like" *D. simulans* variants. Lanes a and i, ϕX174 RF/*Hae*III; lane b, *D. melanogaster* (Thr-Gly)$_{17}$; lane c, *D. melanogaster* (Thr-Gly)$_{20}$; lane d, *D. melanogaster* (Thr-Gly)$_{23}$; lanes e and f, *D. simulans* "23-like" variants (AT 11 and AT 20, respectively); lanes g and h, *D. simulans* "22-like" variants (AT 3 and AT 8, respectively). (**B**) Amplification with the 5' external primer 5'-ATACACATGAGCAGTGTGAC-3' and the 3' internal primer 5'-GTGCCGGAATTCGTTCCATT-3'. Lanes a and i, ϕX174 RF/*Hae*III; lane b, *D. melanogaster* (Thr-Gly)$_{17}$; lane c, *D. melanogaster* (Thr-Gly)$_{20}$; lane d, *D. melanogaster* (Thr-Gly)$_{23}$; lane e, *D. simulans* (Thr-Gly)$_{25}$ (AT 20); lane f, *D. simulans* (Thr-Gly)$_{23}$ (AT 11); lane g, *D. simulans* (Thr-Gly)$_{24b}$ (AT 8); lane h, *D. simulans* (Thr-Gly)$_{24a}$ (AT 3). (**C**) Example of heteroduplex formation with PCR amplified fragments using the external primers 5'-ATACACATGAGCAGTGTGAC-3' and 5'-TTCTCCATCTCGTCGTTGTG-3' with (Thr-Gly)$_{24}$. Lanes a and h, ϕX174 RF/*Hae*III; lane b, *D. melanogaster* (Thr-Gly)$_{17}$; lane c, *D. melanogaster* (Thr-Gly)$_{20}$; lane d, *D. melanogaster* (Thr-Gly)$_{23}$; lane e, *D. simulans* (Thr-Gly)$_{24a}$ plus (Thr-Gly)$_{24b}$ (AT 3 + AT 8); lane f, *D. simulans* (Thr-Gly)$_{24b}$ plus (Thr-Gly)$_{24b}$ (AT 8 + AT 15); lane g, *D. simulans* (Thr-Gly)$_{24a}$ plus (Thr-Gly)$_{24a}$ (AT 3 + AT 6).

a deletion of the downstream "Thr-Gly-Thr-Gly" motif (see Figures 3 and 4). DNA sequencing of a number of (Thr-Gly)$_{24}$ alleles (see below) revealed that two different DNA sequences are encoded by this repeat and that these differed in the pattern of the Thr-Gly encoding ACN GGN codons (see Figure 5). These two isolength variants could be distinguished by using the heteroduplex technique as illustrated in Figure 2C (see also MATERIALS AND METHODS). The variants initially classified as "(Thr-Gly)$_{23}$-like" are in fact either (Thr-Gly)$_{23}$ alleles maintaining the "Thr-Gly-Thr-Gly" motif downstream or (Thr-Gly)$_{25}$ alleles with a deletion of this motif. The sequencing of the two shorter variants observed in the Antibes population revealed two identical (Thr-Gly)$_{21}$ alleles with a deletion of the "Thr-Gly-Thr-Gly" downstream motif.

Three to six individuals were chosen from each population and sequenced in the region which includes the uninterrupted repeat, within the fragment amplified between the 5' and 3' external primers. These individuals where not chosen at random with respect to the length of the repeat. Rather, they represented all the length variants found within each population but they were randomly chosen within each group of length variant. Exceptions were the samples from Peniscola (Spain) and Patras (Greece), 30 flies each, in which all the individuals were sequenced as well as the two individuals from Lyon which were also sequenced. Overall, 92 sequences were described in this way.

Figure 3 shows the nucleotide sequences of the 5' and 3' regions flanking the uninterrupted repeat for the 92 individuals. No nucleotide substitutions were observed in the 5' region (84 bp). On the contrary, a total of six nucleotide substitutions (three synonymous, three replacement), were observed in the 3' region (69 bp; Figure 3). These substitutions were distributed in only three different combinations which are associated, in complete linkage disequilibrium, to specific Thr-Gly length variants. These combinations defined three main haplotypes (I, II and III). Haplotype I and II are associated, respectively, with the (Thr-Gly)$_{23}$ and (Thr-Gly)$_{25}$ length variants. Haplotype III is associated with two different (Thr-Gly)$_{24}$ alleles, which differ in the interspersion pattern of the ACN-GGN codons, and also with the rare (Thr-Gly)$_{21}$ length variant which was detected only in the Antibes sample. These haplotype III subtypes are called, respectively, IIIa, IIIb and IIIc.

Figure 4 shows the alignment of the putative Per protein in the Thr-Gly region encoded by these 92 sequences. The amino acids are numbered according to their relative position in the alignment. No amino acid substitutions between the haplotypes were observed upstream of the Thr-Gly repeat, but the C terminal region shows a Ser $\rightarrow$ Thr (S $\rightarrow$ T) replacement at position 91 in haplotype III compared to haplotypes I and II. Two further substitutions, Thr $\rightarrow$ Ala (T $\rightarrow$ A) at position 98 and Ile $\rightarrow$ Val (I $\rightarrow$ V) at position 100, were observed in

5'-Flanking region   (Thr-Gly) repeat   3'-Flanking region   Haplotype

Reference 5'-Flanking region: AGTGGCTCCTCGGGCAACTTCACCACCGCCAGTAACATACACATAGAGCAGTGTGCGAAATACGAGCATTGCGGGCACTGGGTGGC

Reference 3'-Flanking region: AATGGAACGAATTCCGGCACCCGGAAACCGGCCACCACCAGTTCATCCAGAGAGGAGCAAGCACCGGCCATACCG

Sample	(Thr-Gly) repeat	Haplotype
PE 13	23	I
PE 15	23	I
PE 19	23	I
PE 34	23	I
PE 40	23	I
PE 47	23	I
PE 60	23	I
PE 71	23	I
PE 73	23	I
PE 76	23	I
PE 80	23	I
PA 1	23	I
PA 2	23	I
PA 13	23	I
PA 15	23	I
PA 16	23	I
PA 17	23	I
PA 20	23	I
PA 21	23	I
PA 24	23	I
PA 27	23	I
PA 31	23	I
PA 32	23	I
PA 33	23	I
PA 58	23	I
ME 4	23	I
LY 1	23	I
AG B1	23	I
AN D6	23	I
AT 11	23	I
TA D2	23	I
GF 55	23	I
PE 3	25	II
PE 12	25	II
PE 28	25	II
PE 39	25	III
PE 51	25	III
PE 61	25	II
PA 6	25	II
PA 9	25	III
PA 22	25	III
PA 26	25	III
PA 29	25	II
ME 7	25	II
AG H3	25	III
AG M1	25	III
AN E5	25	III
AN F2	25	III
AT 20	25	III
TA B10	25	II
GF 75	25	II

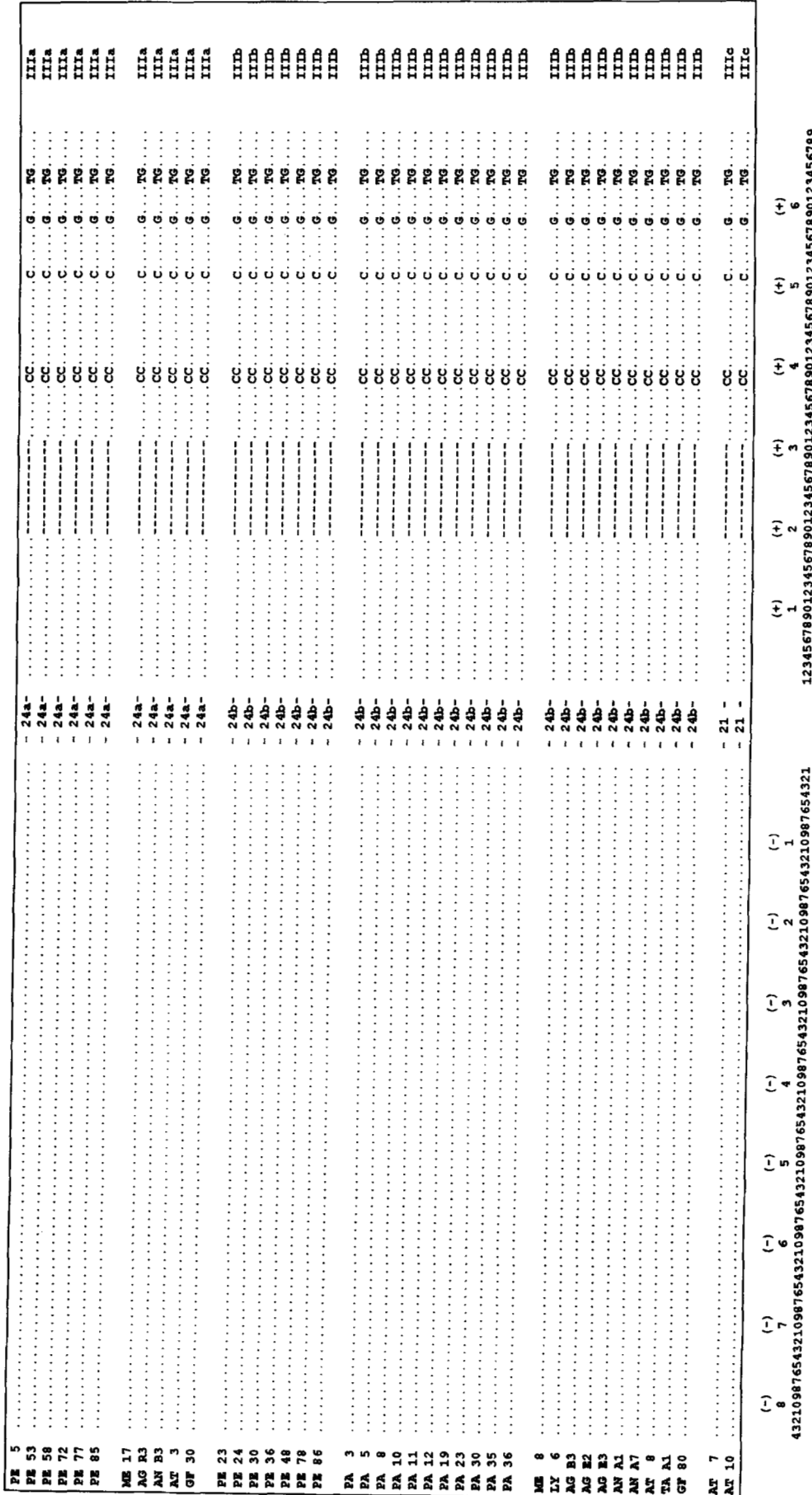FIGURE 3.—DNA sequences from 92 individuals of *D. simulans* in the regions flanking the uninterrupted Thr-Gly repeat. The dots indicate conserved positions, the dashes indicate deletions. PE = Peniscola (Spain), PA = Patras (Greece), ME = Merano (Italy), LY = Lyon (France), AG = Agios Nikolaos (Greece), An = Andros (Greece), AT = Antibes (France), TA = Tanta (Egypt), GF = Grand Ferade (France).

N-terminal region   (Thr-Gly) repeat   C-terminal region   Haplotype

```
PE 13  SGSSGNTTTASNIHMSSVTNTSIAGTGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTG----NGTNSGTGTGTTSSSRGGSTAIP   I
PE 15  ................................................................................----.....TGTG..S.....T.I.   I
PE 19  ................................................................................----.....TGTG..S.....T.I.   I
PE 34  ................................................................................----.....TGTG..S.....T.I.   I
PE 40  ................................................................................----.....TGTG..S.....T.I.   I
PE 47  ................................................................................----.....TGTG..S.....T.I.   I
PE 60  ................................................................................----.....TGTG..S.....T.I.   I
PE 71  ................................................................................----.....TGTG..S.....T.I.   I
PE 73  ................................................................................----.....TGTG..S.....T.I.   I
PE 76  ................................................................................----.....TGTG..S.....T.I.   I
PE 80  ................................................................................----.....TGTG..S.....T.I.   I

PA  1  ................................................................................----.....TGTG..S.....T.I.   I
PA  2  ................................................................................----.....TGTG..S.....T.I.   I
PA 13  ................................................................................----.....TGTG..S.....T.I.   I
PA 15  ................................................................................----.....TGTG..S.....T.I.   I
PA 16  ................................................................................----.....TGTG..S.....T.I.   I
PA 17  ................................................................................----.....TGTG..S.....T.I.   I
PA 20  ................................................................................----.....TGTG..S.....T.I.   I
PA 21  ................................................................................----.....TGTG..S.....T.I.   I
PA 24  ................................................................................----.....TGTG..S.....T.I.   I
PA 27  ................................................................................----.....TGTG..S.....T.I.   I
PA 31  ................................................................................----.....TGTG..S.....T.I.   I
PA 32  ................................................................................----.....TGTG..S.....T.I.   I
PA 33  ................................................................................----.....TGTG..S.....T.I.   I
PA 58  ................................................................................----.....TGTG..S.....T.I.   I

ME  4  ................................................................................----.....TGTG..S.....T.I.   I
LY  1  ................................................................................----.....TGTG..S.....T.I.   I
AG B1  ................................................................................----.....TGTG..S.....T.I.   I
AN D6  ................................................................................----.....TGTG..S.....T.I.   I
AT 11  ................................................................................----.....TGTG..S.....T.I.   I
TA D2  ................................................................................----.....TGTG..S.....T.I.   I
GF 55  ................................................................................----.....TGTG..S.....T.I.   I

PE  3  ............................................................................TGTG.....----..S.....A.V.   II
PE 12  ............................................................................TGTG.....----..S.....A.V.   II
PE 28  ............................................................................TGTG.....----..S.....A.V.   II
PE 39  ............................................................................TGTG.....----..S.....A.V.   II
PE 51  ............................................................................TGTG.....----..S.....A.V.   II
PE 61  ............................................................................TGTG.....----..S.....A.V.   II

PA  6  ............................................................................TGTG.....----..S.....A.V.   II
PA  9  ............................................................................TGTG.....----..S.....A.V.   II
PA 22  ............................................................................TGTG.....----..S.....A.V.   II
PA 26  ............................................................................TGTG.....----..S.....A.V.   II
PA 29  ............................................................................TGTG.....----..S.....A.V.   II

ME  7  ............................................................................TGTG.....----..S.....A.V.   II
AG H3  ............................................................................TGTG.....----..S.....A.V.   II
AG M1  ............................................................................TGTG.....----..S.....A.V.   II
AN E5  ............................................................................TGTG.....----..S.....A.V.   II
AN F2  ............................................................................TGTG.....----..S.....A.V.   II
AT 20  ............................................................................TGTG.....----..S.....A.V.   II
TA B10 ............................................................................TGTG.....----..S.....A.V.   II
GF 75  ............................................................................TGTG.....----..S.....A.V.   II

PE  5  ..........................................................................TG--.....----..T.....A.V.   IIIa
PE 53  ..........................................................................TG--.....----..T.....A.V.   IIIa
PE 58  ..........................................................................TG--.....----..T.....A.V.   IIIa
PE 72  ..........................................................................TG--.....----..T.....A.V.   IIIa
PE 77  ..........................................................................TG--.....----..T.....A.V.   IIIa
PE 85  ..........................................................................TG--.....----..T.....A.V.   IIIa

ME 17  ..........................................................................TG--.....----..T.....A.V.   IIIa
AG R3  ..........................................................................TG--.....----..T.....A.V.   IIIa
AN B3  ..........................................................................TG--.....----..T.....A.V.   IIIa
AT  3  ..........................................................................TG--.....----..T.....A.V.   IIIa
GF 30  ..........................................................................TG--.....----..T.....A.V.   IIIa

PE 23  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 24  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 30  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 36  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 48  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 78  ..........................................................................TG--.....----..T.....A.V.   IIIb
PE 86  ..........................................................................TG--.....----..T.....A.V.   IIIb

PA  3  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA  5  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA  8  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 10  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 11  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 12  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 19  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 23  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 30  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 35  ..........................................................................TG--.....----..T.....A.V.   IIIb
PA 36  ..........................................................................TG--.....----..T.....A.V.   IIIb

ME  8  ..........................................................................TG--.....----..T.....A.V.   IIIb
LY  6  ..........................................................................TG--.....----..T.....A.V.   IIIb
AG B3  ..........................................................................TG--.....----..T.....A.V.   IIIb
AG E2  ..........................................................................TG--.....----..T.....A.V.   IIIb
AG E3  ..........................................................................TG--.....----..T.....A.V.   IIIb
AN A1  ..........................................................................TG--.....----..T.....A.V.   IIIb
AN A7  ..........................................................................TG--.....----..T.....A.V.   IIIb
AT  8  ..........................................................................TG--.....----..T.....A.V.   IIIb
TA A1  ..........................................................................TG--.....----..T.....A.V.   IIIb
GF 80  ..........................................................................TG--.....----..T.....A.V.   IIIb

AT  7  ....................................................------.....TG--.....----..T.....A.V.   IIIc
AT 10  ....................................................------.....TG--.....----..T.....A.V.   IIIc
```

```
        0         0         0         0         0         0         0         0         0         1
        1         2         3         4         5         6         7         8         9         0
123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901
```

FIGURE 4.—The amino acid sequences of the Thr-Gly region for 92 individuals of *D. simulans*. Dots represent invariant amino acid positions, asterisks represent conserved amino acid substitutions, dashes represent deletions. The numbers at the bottom are used to label each amino acid position. PE = Peniscola (Spain), PA = Patras (Greece), ME = Merano (Italy), LY = Lyon (France), AG = Agios Nikolaos (Greece), An = Andros (Greece), AT = Antibes (France), TA = Tanta (Egypt), GF = Grand Ferade (France).

(Thr-Gly) 23

(Thr-Gly) 25

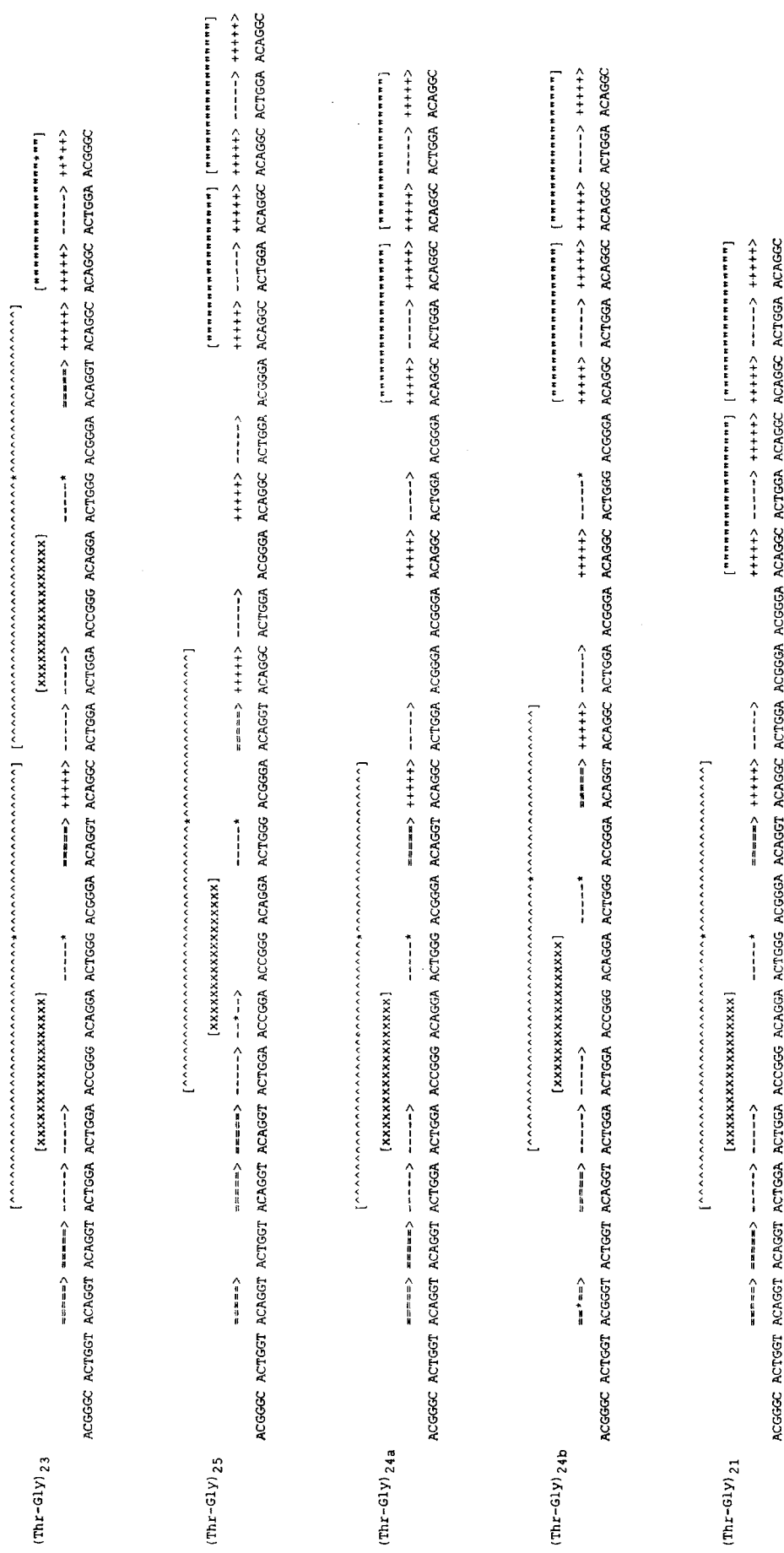(Thr-Gly) 24a

(Thr-Gly) 24b

(Thr-Gly) 21

FIGURE 5.—DNA sequences of the uninterrupted Thr-Gly repeats. Five different alleles corresponding to 21, 23, 24 (a and b) and 25 Thr-Gly pairs were characterized. No point mutations were observed within the sequences for each length variant in the 92 *D. simulans* individuals sequenced. Different symbols are used to show the distribution pattern of different Thr-Gly encoding repeats. Asterisks indicate possible point mutations affecting a codon (see also PEIXOTO *et al.* 1992).

TABLE 1

Frequencies of Thr-Gly length variants in eight natural populations of *D. simulans* estimated by heteroduplex formation

Population	No. of alleles sampled	Thr-Gly length variants (haplotypes)					
		23 (I)	25 (II)	24a (IIIa)	24b (IIIb)	24a+b (IIIa+IIIb)	21 (IIIc)
Merano (Italy)	27	12 (0.444)	5 (0.185)	3 (0.111)	7 (0.259)	10 (0.370)	
Peniscola (Spain)	30	11 (0.367)	6 (0.200)	6 (0.200)	7 (0.233)	13 (0.433)	
Patras (Greece)	30	14 (0.467)	5 (0.167)		11 (0.367)	11 (0.367)	
Agios Nikolaos (Greece)	26	5 (0.192)	9 (0.346)	5 (0.192)	7 (0.269)	12 (0.461)	
Andros (Greece)	23	10 (0.435)	5 (0.217)	2 (0.087)	6 (0.261)	8 (0.384)	
Antibes (France)	18	8 (0.444)	2 (0.111)	3 (0.167)	3 (0.167)	6 (0.333)	2 (0.111)
Tanta (Egypt)	27	14 (0.519)	4 (0.148)		9 (0.333)	9 (0.333)	
Grand Ferade (France)	31	16 (0.516)	3 (0.097)	1 (0.032)	11 (0.355)	12 (0.387)	
Total	212	90 (0.425)	39 (0.184)	20 (0.094)	61 (0.288)	81 (0.382)	2 (0.009)

For each length variant the corresponding haplotypes are indicated in parentheses. The overall frequency of $(Thr-Gly)_{24}$ length variants is also reported because the two haplotypes IIIa and IIIb do not differ at the protein level. For the G test calculation, the two rarest haplotype classes have been pooled in Antibes.

haplotypes II and III (compared with haplotype I). All the amino acid changes between the different haplotypes are conservative changes, for example, both Ser and Thr have –OH groups and both Ile and Val have alkyl side groups.

Figure 5 illustrates the DNA sequences of the uninterrupted Thr-Gly encoding repeat. These sequences differ not only in the number of the ACN-GGN codons encoding the Thr-Gly motif, but also in the interspersion pattern of these codons, as highlighted by arrows. These changes are caused by silent substitutions and slippage-like events, and have been studied in detail within *D. melanogaster* and the *melanogaster* subgroup of species (COSTA *et al.* 1991; PEIXOTO *et al.* 1992). Figure 6 shows two possible alignments of the two $(Thr-Gly)_{24}$ variants. In Figure 6A, 19 point mutations must be invoked in order to derive one sequence from the other. Figure 6B shows how it is possible to relate the two variants to each other more parsimoniously by postulating one deletion, one duplication plus two point mutations.

Analysis of diversity among populations: Table 1 reports the frequencies of the different Thr-Gly alleles in the eight natural populations of *D. simulans* as derived from the heteroduplex analysis. As we observed a complete linkage disequilibrium between the different uninterrupted Thr-Gly sequences and the 3'-polymorphic sites (Figures 3 and 4), the estimate of the frequency of each Thr-Gly length variant can also be considered indirectly as a good estimate of the frequency of the corresponding 3' sequences (see below).

Thr-Gly allele frequencies in the eight populations were compared by G tests. Two criteria were used for

defining the repeat arrays; following EXCOFFIER *et al.* (1992) they will be referred to as the "phenetic" and "evolutionary." Under the phenetic criterion, alleles were considered different if their Thr-Gly repeats differed in length; in this way, haplotypes IIIa and IIIb, both encoding 24 repeats, were pooled. Conversely, under the evolutionary criterion, haplotypes IIIa and IIIb were treated as distinct entities, because their DNA sequence, albeit equal in length, were different.

Allele frequencies appeared homogeneous across localities under the phenetic ($G = 11.62$, 14 d.f., $0.5 < P < 0.9$) as well as under the evolutionary criterion ($G = 31.4$, 21 d.f., $0.05 < P < 0.1$). The Thr-Gly alleles encoding 23 and 24 repeats are the most common, and are represented in almost equal frequencies in all populations.

Tests of neutrality: The HKA (HUDSON *et al.* 1987) and the TAJIMA (1989a) tests of neutrality were initially applied to the two populations (Peniscola and Patras) in which sequences had been obtained for the complete fly samples. The results of the HKA tests are shown in Table 2, while the results of Tajima's tests are shown in Tables 3 and 4. We applied these tests only to the sequences flanking the perfect Thr-Gly repeats, excluding also the downstream deletion. The slippage-like events that have presumably occurred in the repeat region (COSTA *et al.* 1991), contradict the assumptions of the infinite-site model on which the tests are based (WATTERSON 1975; DOVER 1987).

The HKA (HUDSON *et al.* 1987) test compares within-species polymorphism and between-species divergence in two regions of the genome. The underlying hypoth-

The data obtained for the sequences flanking the Thr-Gly repeats of the *D. simulans per* gene was compared to two different loci/gene regions. One of the comparisons is with the data of KLIMAN and HEY (1993) from a fragment of the per gene which lies upstream of the Thr-Gly region and includes exons 2, 3, 4 and part of 5 (see Table 1 of their article). Only the exon sequences were used in this case. The second comparison is with the data of AQUADRO *et al.* (1988) from the rosy locus (see Table 7 of BEGUN and AQUADRO 1991). The published sequence of the *D. melanogaster per* gene (CITRI *et al.* 1987) was used for the calculation of the between-species divergence in the Thr-Gly region. None of the values is significant. The degrees of freedom are equal to one in all cases.

TABLE 3

Results of the Tajima tests of neutrality

Population	Tajima's *D*
Peniscola (Spain)	2.630**
Patras (Greece)	2.903**

Results of the Tajima tests of neutrality for the two populations where the complete sample of flies was sequenced.
** $P < 0.01$.

TABLE 4

Results of the Tajima tests of neutrality

Population	Tajima's *D*, no. of undetected singletons		
	0	1	2
Merano (Italy)	2.809**	2.181*	1.683
Agios Nikolaos (Greece)	1.054	0.352	0.285
Andros (Greece)	2.669**	2.056*	1.571
Antibes (France)	2.629**	2.023*	1.545
Tanta (Egypt)	2.819**	2.191*	1.691
Grand Ferade (France)	2.957**	2.315*	1.803

Results of the Tajima tests of neutrality for the populations where the frequencies of the different haplotypes were predicted from the frequencies of length alleles. The values of the Tajima *D* statistic shown in the last two columns assumes, respectively, one and two undetected singletons in each population (see Table 5).
* $P < 0.05$; ** $P < 0.01$.

esis is that, in the absence of selection, similar levels of genetic diversity should be detected in the comparisons, both within and between species. The sequences flanking the Thr-Gly repeat were compared to the data of KLIMAN and HEY (1993) for another region of *per* lying upstream of the Thr-Gly region and to the data of AQUADRO *et al.* (1988) from the rosy locus (see also BEGUN and AQUADRO 1991). No significant departures from neutral expectations were obtained in either of the two comparisons (see Table 2 for details).

In the Tajima test the number of segregating sites is compared to the average number of nucleotide differences between pairs of sequences (TAJIMA 1989a). Under

FIGURE 6.—Possible alignments of the DNA sequences for the uninterrupted repeat for the two types of (Thr-Gly)$_{24}$ alleles (see text). (A) Nineteen nucleotide substitutions (positions denoted by asterisks) must be invoked to derive one sequence from the other. The numbers at the bottom indicate the number of Thr-Gly repeats. (B) The same sequences related to each other by one deletion, one duplication, represented as gaps in the alignment, and two point mutations, denoted by asterisks (see text).

TABLE 5

Probability that 0, 1 or more undetected polymorphic sites (singletons) were present in the eight samples

Sample	N	n_{seq}	n	P_0	P_1	P_{2+}
Peniscola	30	30	0	1	0	0
Patras	30	30	0	1	0	0
Merano	27	4	23	0.473	0.360	0.167
Agios Nikolaos	23	7	16	0.594	0.314	0.092
Andros	26	6	20	0.522	0.334	0.144
Antibes	18	6	12	0.677	0.269	0.054
Tanta	27	3	24	0.458	0.363	0.179
Grand Ferade	31	4	27	0.416	0.359	0.225

N is the sample size, n_{seq} is the number of individuals whose sequence was determined, n is the number of individuals whose sequence was inferred, P_0 and P_1 are the binomial probabilities to have, respectively, 0 or 1 undetected singleton in the sample, and P_{2+} is the probability to have two or more, calculated as: $P_{2+} = 1 - (0.968^n) - (n \times 0.032 \times 0.968^{n-1})$.

an infinite-site model (WATTERSON 1975), the relationship between the number of segregating (*i.e.*, polymorphic) sites in a sample, and the average number of pairwise differences between sequences, is predictable in the absence of selection. Departures from null expectations, either for an excess or for a deficiency of rare alleles, are considered to indicate different selective regimes. In both populations (Peniscola and Patras), the Tajima's D statistic was positive and highly significant (Table 3) indicating departures from neutral expectations. These were due to an excess of sequence variants (haplotypes I and III) which differed in many substitutions, and which had both intermediate frequencies of around 40%. As a consequence, a deficiency of rare haplotypes is also observed.

Note that because a complete linkage disequilibrium exists between the Thr-Gly length variants and the DNA sequences at these 3' polymorphic sites in the 92 individuals assayed, the DNA sequence for the samples from the other populations could be, in principle, inferred by examining the length of the Thr-Gly repeat. The error associated with this procedure is smaller than 3.2%, which is the upper 95% confidence limit of the ratio of observed new haplotypes (0) to the total number of individuals assayed (92) (SOKAL and ROHLF 1981). Tajima's test statistic was then calculated independently for each of the other 6 populations from the DNA sequences inferred from the length of the Thr-Gly repeat. In all six populations the Tajima's D statistic was also positive, and it reached high levels of significance in five of them (Table 4), Agios Nikolaos (Greece) being the only exception. As before, the departures from neutral equilibrium expectations were in all cases due to an excess of sequence variants with intermediate frequency.

We also repeated the tests allowing for the probability, P_1, of the existence of one undetected polymorphic site (singleton) in each one of these six populations (see Table 5 for details). As before all the values are positive and significant except for Agios Nikolaos (see Table 4). However, if two or more singletons escaped detection the results of the Tajima test would be nonsignificant (Table 4, last column). This probability, that two or

more undetected singletons existed in the same sample, P_{2+}, can be calculated as 1 minus the binomial probability to have zero or one undetected singleton, given a probability of their existence as 0.032 or less. For each sample, P_{2+} fell in the range between 0 and 0.225 (Table 5), and therefore we cannot rule out the probability of one false significant D value. However, the probability of having two false significant D values is the product of two P_{2+} values, which varies again according to the samples considered, but never exceeds 0.040 (Tanta and Grand Ferade). Therefore, even if we accept the highly conservative view that we missed the maximum number of singletons consistent with our sample sizes, it is highly unlikely that we had more than one false significant D value in the Tajima tests.

Thus our conclusions with the Tajima D statistic rest firmly with the results based on the complete data sets from Patras and Peniscola. Our further tests with the six incompletely sequenced populations add further weight to our conclusion but are tempered by the assumptions underlying the analysis.

DISCUSSION

Both *D. melanogaster* and *D. simulans* show length variation in the Thr-Gly encoding repeat of the *per* gene (COSTA *et al.* 1991, 1992; PEIXOTO *et al.* 1992). However, a major difference between the species exists downstream of the repeat where in *D. melanogaster*, we have yet to detect any amino acid variation in the two common European variants, the (Thr-Gly)$_{17}$ and the (Thr-Gly)$_{20}$ alleles (A. PEIXOTO and E. ROSATO, unpublished observations), whereas six nucleotide substitutions, only three of which are silent, have been observed in *D. simulans*. Such an intraspecific level of variability in the 3' flanking region is of the same order of magnitude as the variability between the two species (WHEELER *et al.* 1991; PEIXOTO *et al.* 1992). However, this apparently higher level of variability is not significantly different from that observed in other loci (see below). A further difference between *D. simulans* and *D. melanogaster* lies in the geographical pattern of variation within this region of

the *per* locus. *D. melanogaster* shows a cline that is strongly correlated with latitude (COSTA *et al.* 1992), whereas the G-tests performed in this study reveal that the frequencies of the different haplotypes do not differ significantly among populations of *D. simulans*.

Tables 3 and 4 show that for both populations where the complete fly samples were sequenced, Peniscola and Patras, as well as for 5 out of 6 other populations where the haplotype frequencies were inferred from the Thr-Gly length variation, the Tajima *D* statistic was positive and highly significant. This was due to an excess of haplotypes I and III which differ in many substitutions and are observed in intermediate frequencies, each representing roughly 40% of all haplotypes examined. In addition, the departures from the expected haplotype frequencies are parallel in the seven populations of interest.

In general, deviations from a model of populations at equilibrium suggest that evolution has not occurred through a process of differentiation occurring at a constant rate through time (EXCOFFIER 1990). Accordingly, it seems necessary to identify one or more evolutionary factors that may have contributed to determining such deviations. In principle, these factors include balancing selection, incomplete admixture, and other demographic phenomena such as recent population bottlenecks. Admixture and bottlenecks affect individual populations. It is hard to imagine that they can have occurred independently, and yet have had the same effect on populations distributed over a large region of Europe. On the contrary, our results require an explanation for the concordant departures from equilibrium expectations at the different localities. Furthermore, bottlenecks amplify random allele-frequency fluctuations and should lead local populations to diverge. However we observe only insignificant differences between them. Therefore, although bottlenecks and partial admixture may well have occurred in individual populations and may have played a relevant evolutionary role at the local level, other factors must be invoked to explain the continental pattern described here.

The sibling species *D. melanogaster* and *D. simulans* have a cosmopolitan distribution; average nucleotide heterozygosity in the latter appears to be three to sixfold as high as in the former, which could reflect differences in effective population sizes between the two species (KREITMAN 1991; AQUADRO 1992; KLIMAN and HEY 1993; HEY and KLIMAN 1993). However, large population sizes imply large opportunities for mutation to generate new variants (KIMURA 1983). In *D. melanogaster*, rare length alleles for the Thr-Gly region were found in most of the populations (COSTA *et al.* 1991). In *D. simulans*, on the contrary, only in the Antibes sample were rare length alleles detected, and in seven of our eight samples Tajima tests showed a significant deficiency of rare haplotypes. As a consequence, effects of population sizes cannot, by themselves, explain the findings of this study.

The results of the Tajima tests (Table 3 and 4) show significant departures from a strictly neutral model of molecular evolution. On the other hand, the HKA test (Table 2) failed to show any significant excess of diversity at the sequences flanking the Thr-Gly repeats in *D. simulans*, relative to the expectations of the neutral theory derived from comparison to an independent locus, *rosy* (AQUADRO *et al.* 1988, 1993), and to another region of *per* (KLIMAN and HEY 1993). This is not the first time that different results have been obtained between these two tests in comparisons with these species (KREITMAN and HUDSON 1991; BEGUN and AQUADRO 1991). Both tests compare the data to neutral expectations, but they differ in their meaning of neutrality. The HKA test examines whether the number of polymorphic sites is significantly different from an expected value which is based on the comparison of the amount of polymorphism within *D. simulans* at two loci, and the divergence at these loci between *D. simulans* and its sibling species *D. melanogaster*. The Tajima test asks whether the average number of pairwise nucleotide differences is consistent with the observed number of segregating sites.

The most trivial explanation for the different results between the two tests is that the rejection of the null hypothesis in the Tajima test may be due to a type I error. Perhaps the eight independent populations represent the descendants of just one well mixed population which relatively recently expanded into Europe, as with *D. melanogaster* (DAVID and CAPY 1988).

Alternatively, the HKA test has several assumptions which may have been violated with our populations. For example one of these is that there should be no ancestral polymorphism segregating between the two species (FORD *et al.* 1994). In fact PEIXOTO and co-workers (1992) have examined the evolution of the Thr-Gly repeat and flanking region in the *D. melanogaster* subgroup of species and suggested that the major variants seen in *D. simulans* probably derive from an ancient polymorphism which could have been present in the common ancestor of *D. simulans* and its sibling species. Similar conclusions were reached by KLIMAN and HEY (1993) based on a polymorphic 1.9-kb fragment of *per* which lies upstream of the Thr-Gly repeat. Secondly the assumption that there should be free recombination between loci is also violated in the comparison of our *per* sequences with those from KLIMAN and HEY (1993), thereby making the HKA test more conservative. Finally the upstream sequences of *per* and *rosy*, that we used were taken from different worldwide localities and this may have artificially inflated the variation in these sequences. Again this would have had the effect of making the HKA test more stringent (BEGUN and AQUADRO 1991).

Both the Tajima and the HKA test assume a constant population size at equilibrium and infinite sites model of mutation. This means that the number of nucleotide

```
mel            AGCTCATCCAAAGGCGGAAGCGCCGCCATACCG
sim I          AGTTCATCCAGAGGAGGAAGCACCGCCATACCG
sim II         AGCTCATCCAGAGGCGGAAGCGCCGCTGTACCG
sim III        ACCTCATCCAGAGGCGGAAGCGCCGCTGTACCG
varying sites  12      3 4     5    67
```

FIGURE 7.—The polymorphic sites in the *D. simulans* haplotypes compared with the same region in *D. melanogaster* (CITRI *et al.* 1987).
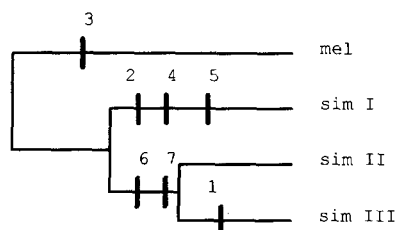

FIGURE 8.—A parsimony tree of the *D. simulans* variants rooted with the *melanogaster* sequence.

differences between haplotypes should be proportional to the time elapsed since their common ancestry. Figure 7 shows the sequences for each haplotype compared to that of *D. melanogaster*, and we have constructed the most parsimonious phylogenetic tree for these data (Figure 8). Note that the number of base changes in each branch of the *D. simulans* lineage are the same suggesting that the two major haplotypes are ancient. Figure 9 shows the repetitive Thr-Gly sequences for the same haplotypes and again, we have generated a parsimonious phylogeny which incorporates the putative deletion-insertion events by which we can derive the different variants. The tree in Figure 10 looks extremely similar to that shown in Figure 8 based on the downstream non-repetitive sequences, supporting the view that the numbers of differences between the different haplotypes are proportional to the time since their common ancestry. Thus the infinite sites assumption underlying our statistical tests appears to be supported. However we must add that a number of different phylogenies are possible when comparing the repetitive sequences and even though they all give similar topologies, the number of insertion-deletion events is not always similar on the branches that descend from each side of the root. It is extremely difficult to choose between the different trees as there is no sensible criterion on which to judge whether the different insertion-deletion events are all equally probable. In other words this approach using the repetitive sequence cannot give an unambiguous answer to the question of whether the infinite sites model of mutation can apply.

Another assumption of both tests is a constant population size at equilibrium. If the population size for *D. simulans* has recently decreased, rare variants may have been lost, and the number of polymorphic sites may have been reduced. However the average number of pairwise differences would be affected less than the number of polymorphic sites (TAJIMA 1993b). The implications are that the HKA test would give a nonsignificant result because it would not take into account

```
mel  17   a  b  c    c  d  d  d  e   f   d   g   c  h  h  h  d  h

sim 23   a  b  c    c  d  d  e  f   d₂  g   c   h  d  d  e  f  d₂  g  c  h  h  d  h₁

sim 24a  a  b  c    c  d  d  e  f   d₂  g   c   h  d  g  g  h  d   g  h  d  h  h  d  h

sim 24b  a  b  c₁  b  c  d  d  e   f   d₂  g   c  h  d  g  h  d₃  g  h  d  h  h  d  h

sim 25   a  b  c    b  c  c  d  d₄  e   f   d₂  g  c  h  d  g  h   d  g  h  d  h  h   d  h
```

```
a = ACG GGC      d = ACT GGA      e = ACC GGG      h = ACA GGC
b = ACT GGT      d₂= ACT GGG      f = ACA GGA      h₁= ACG GGC
c = ACA GGT      d₃= ACT GGG      g = ACG GGA
                 d₄= ACC GGA
```

FIGURE 9.—The repetitive Thr-Gly region of the *D. simulans* variants compared to the $(Thr-Gly)_{17}$ allele of *D. melanogaster* (COSTA *et al.* 1991). Each Thr-Gly encoding hexamer is represented by a lower case letter. Subscripts (for example d_2, d_3, etc.) denote Thr-Gly encoding "cassettes" which share common descent.
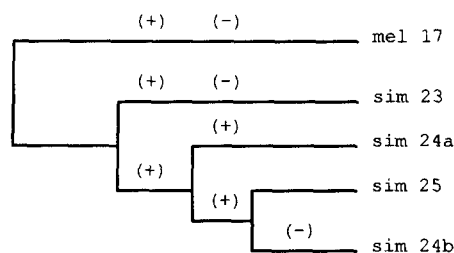
FIGURE 10.—One of the possible parsimony trees based on the presence and absence of deletion and duplication (see text).

average pairwise differences, but a significant D value in the Tajima test would still be obtained. However this explanations seems unlikely because KLIMAN and HEY (1993) did not report significant D values for un upstream region of *per* as well as for sequences within the *zeste* and the *yp2* genes, and a reduction in population size would be expected to affect all loci equally.

Perhaps the different portions of the genome we are considering did not evolve at the same rate. Selective effects may have reduced variation at one site, while leaving other sites unaffected. Figure 8 shows that both the major haplotypes, I and III, look old. However, the trees in Figure 8 show that the two $(Thr-Gly)_{24}$ variants, a and b, which we classify as haplotype III, do not connect to the same depth in the tree. In fact the $(Thr-Gly)_{24b}$ variant appears to be more recent than the $(Thr-Gly)_{24a}$. Interestingly the $(Thr-Gly)_{24b}$ allele is also the most frequent variant and, in the contrast to the $(Thr-Gly)_{24a}$, has been found in all the localities examined. Perhaps a selective substitution is in progress involving the $(Thr-Gly)_{24b}$ variant. In evolutionary processes, when the frequency of an allele does not reflect its age it suggests that the frequency may not be determined by drift and selection might come into play. Note that because the two $(Thr-Gly)_{24}$ variants have the same amino acid sequence in the Thr-Gly region, we might speculate that any selection could be acting on a nearby linked site. Examining the rhythmic behavior of these two variants might

reveal differences between them which could be visible to selection.

If natural selection has acted to change the effective population size of this portion of genome we have studied the positive value for D could reflect the effects of balancing selection. The variants involved would remain in the population for a longer period of time than expected by the action of drift, thereby accumulating excess polymorphism. The extent of the extra polymorphism will depend on the mutation rate, effective population size and recombination rate (HUDSON and KAPLAN 1988). As the number of extra polymorphic sites will depend on time, these will not be expected to be numerous if the balanced polymorphism is a new phenomenon, for example if it was a feature of a relatively recent colonization of Europe. However if the haplotypes are in high frequencies, as in our case, the contribution to the average pairwise differences even of a small number of polymorphic sites would be relatively large giving the positive value for D. In contrast, the HKA test has relatively little power to detect departures from neutrality given a small number of extra polymorphic sites, and the use of the χ^2 distribution to asses the test's significance assumes that measures of variation are approximately normally distributed which is more likely when the number of polymorphic sites is large (H. HILTON, R. KLIMAN and J. HEY, manuscript submitted for publication).

In addition to the results of Tajima's test there is further evidence in favor of selection as a factor determining genetic homogeneity among populations at the *per* locus. The insignificant G test for heterogeneity calculated under the phenetic criterion indicates that variants of equal length occur at the same frequency across localities. However, the processes whereby these variants accumulated may have been different in certain populations, as shown by the fact that the G test, calculated under the evolutionary criterion, is much closer to significance. In other words, what is constant across localities is the frequency of length phenotypes, and perhaps not the particular DNA sequences encoding them; different alleles of equal length have reached the same pooled frequency in each population. Thus the excess of haplotypes I and III may result from some form of balancing selection. In Wright's terms the coexistence of two length variants at higher-than-expected frequencies may represent an adaptive peak, which all populations, except perhaps for Agios Nikolaos, have climbed.

Whatever the interpretation of the population homogeneity observed in *D. simulans*, it is interesting to note that once again, different pattern of polymorphism are observed in *D. melanogaster* and *D. simulans* (ANDERSON and OAKESHOTT 1984; AQUADRO 1992). Why does the same gene show such a radically different pattern of geographical variation in *D. melanogaster*, where a robust latitudinal cline is observed in Thr-Gly length (COSTA

et al. 1992) compared to *D. simulans* with its apparent spatial homogeneity? This is puzzling but seems to conform to DOBZHANSKY's (1962) distinction between what he called "rigid" and "flexible" polymorphisms. An analogous situation exists where "temporal fluctuation" takes the place of "geographical variation" in inversion frequencies of the sibling species *D. pseudoobscura* and *Drosophila persimilis*. Inversion frequencies fluctuate significantly less in *D. persimilis* than in *D. pseudoobscura* both between seasons and between years (COYNE *et al.* 1992).

The evidence obtained so far does not allow us to define the relative roles of the Thr-Gly repeat, and of the flanking regions, in the interplay with any external factors which might conceivably be determining levels and patterns of genetic variation. At this present time one view might be that different selective regimes maintain the cline in *D. melanogaster* and the geographically homogeneous distribution of haplotype frequencies in *D. simulans*. Similar conclusions with regard to *D. simulans* have been made by AQUADRO and co-workers with their American populations (C. AQUADRO, personal communication). In these selection regimes, an important role could have been played by differences in the effective population size between the two species. In a larger population, such as *D. simulans*, genetic drift is less important as an evolutionary factor, but selection against slightly deleterious mutants is more efficient. This hypothesis has been proposed as an explanation for the lower allozyme variation in *D. simulans* than in *D melanogaster* (AQUADRO *et al.* 1988). It could account for the lower genetic variance observed for some adaptive characters (DAVID and BOUCQUET 1975; HOFFMANN and PARSONS 1993) and in our case for the deficiency of rare variants in the Thr-Gly region of *per* in *D. simulans*. Thus, although both species were genetically versatile enough to colonize temperate regions, the greater "potentially adaptive" genetic variation in *D. melanogaster* may have permitted a spatial differentiation in Europe in contrast to *D. simulans*. Perhaps this is why *D. melanogaster* shows clines for many more characters than *D. simulans* (DAVID and BOUCQUET 1975; ANDERSON and OAKESHOTT 1984; HOFFMANN and PARSONS 1993; AQUADRO 1992).

However, it may be that a linkage disequilibrium exists between this region and another site in *D. simulans* causing the different spatial patterns. In *D. melanogaster* the (Thr-Gly)$_{17}$ and (Thr-Gly)$_{20}$ variants differ by one deletion/insertion event involving three Thr-Gly pairs (COSTA *et al.* 1991). Repetitive DNA can mutate at extremely high rates (JEFFREYS *et al.* 1988a, 1990) and it is therefore conceivable that the (Thr-Gly)$_{17}$ allele may be "flipping" to the (Thr-Gly)$_{20}$ variant and vice versa at high frequencies. This would effectively destroy any strong linkage disequilibrium because flanking markers

will continuously be exchanged between the two variants. In *D. simulans*, on the other hand, it is difficult to see how the (Thr-Gly)$_{23}$ (haplotype I) and (Thr-Gly)$_{24}$ (haplotype III) variants can be derived from each other without invoking a number of overlapping deletion/insertion events, as well as point substitutions. Thus the mutation rate of the (Thr-Gly)$_{23}$ ⇔ (Thr-Gly)$_{24}$ will be relatively slow and linkage disequilibrium could become a significant factor to explain the different pattern we observe between the species. Clearly the (Thr-Gly)$_{23}$ and (Thr-Gly)$_{24}$ *D. simulans* variants represent an ancient polymorphism (PEIXOTO *et al.* 1992), and the regions adjacent but further from the Thr-Gly region may retain many ancestral differences. An interesting possibility is that the linkage disequilibrium between the Thr-Gly repeats and the downstream sequences observed in *D. simulans* might result from the type of epistatic selection that has been recently observed for other loci (BERRY and KREITMAN 1992; SCHAEFFER and MILLER 1993; MIYASHITA *et al.* 1993).

To choose between the competing hypotheses which may explain these observations, there is a need for the different Thr-Gly haplotypes in each species to be investigated with respect to their behavioral phenotypes. This approach has generated considerable insight in understanding the clinal pattern observed in *D. melanogaster* (L. SAWYER, A. A. PEIXOTO, R. COSTA, and C. P. KYRIACOU, unpublished observations) Furthermore, the construction of *D. simulans* hybrid genes where the Thr-Gly repeat length and the downstream 3' haplotypes are "mixed and matched," followed by transformation experiments may reveal subtle changes in rhythmic behavior. Such changes in circadian and ultradian cycles may have implications for fitness, so a thorough examination of these behavioral phenotypes may contribute to our understanding of the mechanisms which maintain these polymorphisms.

LITERATURE CITED

ANDERSON, P. R., and J. G. OAKESHOTT, 1984 Parallel geographical patterns of allozyme variation in two sibling *Drosophila* species. Nature 308: 729–731.

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

AQUADRO, C. F., 1993 Molecular population genetics of *Drosophila*, pp. 222–266 in *Molecular Approaches to Fundamental and Applied Entomology*, edited by J. OAKESHOTT and M. J. WHITTEN. Springer Verlag, Berlin.

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., 1987 Love-song and circadian rhythm. Nature 326: 741.

BARGIELLO, T. A., and M. W. YOUNG, 1984 Molecular genetics of a biological clock in *Drosophila*. Proc. Natl. Acad. Sci. USA 81: 2142–2146.

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

BERRY, A., and M. KREITMAN, 1993 Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics 134: 869–893.

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

COLOT, H. V., J. C. HALL and M. ROSBASH, 1988 Interspecific comparison of the *period* gene of *Drosophila* reveals large blocks of non-conserved coding DNA. EMBO J. 7: 3929–3937.

COSTA, R., A. A. PEIXOTO, J. R. THACKERAY, R. DALGLEISH and C. P. KYRIACOU, 1991 Length polymorphism in the threonine-glycine encoding repeat region of the *period* gene in *Drosophila*. J. Mol. Evol. 32: 238–246.

COSTA, R., A. A. PEIXOTO, G. BARBUJANI and C. P. KYRIACOU, 1992 A latitudinal cline in a *Drosophila* clock gene. Proc. R. Soc. Lond. Ser. B 250: 43–49.

COYNE, J. A., 1992 Genetics and speciation. Nature 355: 511–515.

COYNE, J. A., J. R. POWELL and C. E. TAYLOR, 1992 Temporal stability of third-chromosome inversion frequencies in *D. persimilis* and *D. pseudoobscura*. Evolution 46: 1558–1563.

DAVID, J. R., and BOCQUET C., 1975 Similarities and differences in latitudinal adaptation of two *Drosophila* sibling species. Nature 257: 588–590.

DAVID J. R., and P. CAPY, 1988 Genetic variation of D. melanogaster natural populations. Trends Genet. 4: 106–111.

DOBZHANSKY, T., 1962 Rigid *versus* flexible chromosomal polymorphism in *Drosophila*. Am. Nat. 96: 321–328.

DOVER, G. A., 1987 DNA turnover and the molecular clock. J. Mol. Evol. 26: 47–58.

EWER J., M. HAMBLEN-COYLE, M., ROSBASH and J. C. HALL, 1990 Requirement for *period* gene expression in the adult and not during development for locomotor activity rhythms of imaginal *Drosophila melanogaster*. J. Neurogenet. 7: 31–73.

EXCOFFIER, L., 1990 Evolution of human mitochondrial DNA: Evidence for departure from a pure neutral model of populations at equilibrium. J. Mol. Evol. 30: 125–139.

EXCOFFIER, L., P. E. SMOUSE and J. M. QUATTRO, 1992 Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mithocondrial DNA restriction data. Genetics 13: 479–491.

FORD M. J., C. K. YOON and C. AQUADRO, 1994 Molecular evolution of the *period* gene in *Drosophila athabasca*. Mol. Biol. Evol. 11: 169–182.

GLOOR G., and W. ENGELS, 1990 Single-fly DNA preps for PCR. Drosophila Inf. Newsl. 1: January.

HEY, J., and R. M. KLIMAN, 1993 Population genetics and phylogenetics of DNA sequence variation at multiple loci within the *Drosophila melanogaster* species complex. Mol. Biol. Evol. 10: 804–822.

HOFFMANN, A. A., and PARSONS P. A., 1993 Direct and correlated responses to selection for desiccation resistance: a comparison of *D. melanogaster* and *D. simulans*. J. Evol. Biol. 6: 643–657.

HUDSON, R. R., and N. L. KAPLAN, 1988 The coalescent process in models with selection and recombination. Genetics 116: 153–159.

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

JACKSON, F. R., T. A. BARGIELLO, S. H. YUN and M. W. YOUNG, 1986 Product of the per locus shares homology with proteoglycans. Nature 320: 185–188.

JEFFREYS A. J., N. J. ROYLE, V. WILSON and Z. WONG, 1988a Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332: 278–281.

JEFFREYS A. J., V. WILSON, R. NEUMANN and J. KEYTE, 1988b Amplification of human minisatellites by polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res. **16:** 10953–10971.

JEFFREYS A. J., R. NEUMANN, and V. WILSON, 1990 Repeat unit sequence variation in minisatellites: A novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. Cell **60:** 473–485.

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

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.

KONOPKA, R. J., and S. BENZER, 1971 Clock mutants of *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **68:** 2112–2116.

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

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

KYRIACOU, C. P., and J. C. HALL, 1980 Circadian rhythm mutations in *Drosophila melanogaster* affect short-term fluctuations in the male's courtship song. Proc. Natl. Acad. Sci. USA **77:** 6929–6933.

KYRIACOU, C. P., and J. C. HALL, 1989 Spectral analysis of *Drosophila* courtship song rhythms. Anim. Behav. **37:** 850–859.

KYRIACOU, C. P., M. OLDROYD, J. WOOD, M. SHARP and M. HILL, 1990 Clock mutations alter developmental timing in *Drosophila*. Heredity **64:** 395–401.

McCLUNG, C. R., B. A. FOX and J. C. DUNLAP, 1989 *frequency*, a clock gene in *Neurospora* shares a sequence element with the Drosophila clock gene *period*. Nature **339:** 558–562.

MIYASHITA, N. T., M. AGUADE and C. H. LANGLEY, 1993 Linkage disequilibrium in the *white* locus region of *Drosophila melanogaster*. Genet. Res. **62:** 101–109.

PEIXOTO, A. A., R. COSTA, D. A. WHEELER, J. C. HALL and C. P. KYRIACOU, 1992 Evolution of the threonine-glycine repeat region of the *period* gene in the *melanogaster* species subgroup of *Drosophila*. J. Mol. Evol. **35:** 711–719.

PEIXOTO, A. A., S. CAMPESAN, R. COSTA and C. P. KYRIACOU, 1993 Molecular evolution of a repetitive region within the *per* gene of *Drosophila*. Mol. Biol. Evol. **10:** 127–139.

PETERSEN, G., J. C. HALL and M. ROSBASH, 1988 The *period* gene of *Drosophila* carries species-specific behavioural instructions. EMBO J. **7:** 3939–3947.

REDDY, P., W. A. ZEHRING, D. A. WHEELER, V. PIRROTTA, C. HADFIELD *et al.*, 1984 Molecular analysis of the *period* locus in *Drosophila melanogaster* and identification of a transcript involved in biological rhythms. Cell **38:** 701–710.

SCHAEFFER, S. W., and E. L. MILLER, 1993 Estimates of linkage disequilibrium and the recombination parameter determined from segregating nucleotides sites in the alcohol dehydrogenase region of *Drosophila pseudoobscura*. Genetics **135:** 541–552.

SOKAL R. R., and F. J. ROHLF, 1981 *Biometry*. W. H. Freeman, San Francisco.

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

TAJIMA, F., 1989b The effect of change in population size on DNA polymorphism. Genetics **123:** 597–601.

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

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

WHEELER, D. A., C. P. KYRIACOU, M. GREENACRE, Q. YU, J. RUTILA *et al.*, 1991 Molecular transfer of a species-specific courtship behaviour from *Drosophila simulans* to *Drosophila melanogaster*. Science **251:** 1082–1085.

YU, Q., H. V. COLOT, C. P. KYRIACOU, J. C. HALL and M. ROSBASH, 1987 Behaviour modification by *in vitro* mutagenesis of a variable region within the *period* gene of *Drosophila*. Nature **326:** 765–769.