J Mol Evol (1991) 32:238–246

Journal of Molecular Evolution

Length Polymorphism in the Threonine-Glycine-Encoding Repeat Region of the *period* Gene in *Drosophila*

Rodolfo Costa,[1,2] Alexandre A. Peixoto,[1] Justin R. Thackeray,[1*]
Raymond Dalgleish,[1] and Charalambos P. Kyriacou[1]

[1] Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK
[2] Dipartimento di Biologia, Universita' di Padova, Via Trieste 75, 35121 Padova, Italy

Summary. Single-fly polymerase chain reaction amplification and direct DNA sequencing revealed high levels of length polymorphism in the threonine-glycine encoding repeat region of the *period* (*per*) gene in natural populations of *Drosophila melanogaster.* DNA comparison of two alleles of identical lengths gave a high number of synonymous substitutions suggesting an ancient time of separation. However detailed examination of the sequences of different Thr-Gly length variants indicated that this divergence could be understood in terms of four deletion/insertion events. In *Drosophila pseudoobscura* a length polymorphism is observed in a five-amino acid degenerate repeat, which corresponds to *melanogaster*'s Thr-Gly domain. In spite of the differences between *D. melanogaster* and *D. pseudoobscura* in the amino acid sequence of the repeats, the predicted secondary structures suggest evolutionary and mechanistic constraints on the *per* protein of these two species.

Key words: *Drosophila* — *per* gene — Repeated sequence — Threonine-glycine — Length polymorphism — Minisatellite

Introduction

The sex-linked *period* (*per*) gene controls biological rhythmicity in *Drosophila.* In *D. melanogaster,* mutations of this gene shorten, lengthen, or obliterate the fly's 24-h circadian cycles and have parallel effects on the 60 s lovesong rhythm of the male (for reviews see Hall and Kyriacou 1990; Kyriacou 1990). The *per* gene has been cloned and sequenced in several *Drosophila* species, revealing large blocks of nonconserved coding DNA intercalated by regions of conservation (Colot et al. 1988; Thackeray and Kyriacou 1990). The most striking feature of the primary structure of the conceptual protein in *D. melanogaster* is a run of alternating threonine-glycine pairs (Jackson et al. 1986; Citri et al. 1987). The region surrounding and including this repeat has been implicated in the determination of the species-specific differences in the *Drosophila* lovesong cycle (Yu et al. 1987; Wheeler et al. 1991). Moreover variability in the number of the encoded Thr-Gly pairs (17, 20, and 23) has been reported among different laboratory strains (Yu et al. 1987). Length polymorphism in coding DNA has also been observed in several other internally repetitive genes (e.g., Muskavitch and Hogness 1982; Goodbourn et al. 1983; Swallow et al. 1987; Lyons et al. 1988; Teumer and Green 1989). Part of such variability may be generated by slippage-like events (Dover 1989), which can be a major factor in some DNA sequence evolution (Tautz et al. 1986; Levinson and Gutman 1987; Treier et al. 1989). The repeat region within the *per* gene thus provides a good opportunity to study any slippage events occurring within the coding minisatellite.

Using single-fly polymerase chain reaction (PCR) amplification and direct DNA sequencing of the amplified products we analyzed the variability in the Thr-Gly-encoding region of the *per* gene in natural

Offprint requests to: C.P. Kyriacou

* *Present address:* Department of Genetics, University of Wisconsin–Madison, 445 Henry Mall, Madison, WI 53706, USA

populations of *D. melanogaster*. We have observed high levels of length polymorphism and have sequenced new length variants. We also describe length variants found in three strains of *D. pseudoobscura* in a region of the *per* gene that appears to correspond to the Thr-Gly domain in *D. melanogaster* (Colot et al. 1988). These results suggest that length variation in this particular domain of the *per* gene may be a common finding within *Drosophila* species and raise the question of whether such polymorphisms might have selective value.

Materials and Methods

Natural Populations of D. melanogaster. Samples from Zakynthos (Greece), Pietrastornina (south Italy), and Merano (north Italy) were collected from the wild in October 1989 using grape baits. Very large samples from each population were obtained and these were immediately subdivided into 60 different mass cultures. Single flies from different cultures were analyzed within five generations from collection. The Bordeaux (France) sample was taken from the progeny of 30 inseminated females collected from the wild in October 1989 and kindly provided by Dr. P. Capy (Gif-sur-Yvette, France).

Strains of D. pseudoobscura. The *D. pseudoobscura* strains were obtained from the National *Drosophila* Species Resource Center (Bowling Green State University, OH, USA). The strain PA (stock number 0121.88) is from Pachuca, Mexico; the strain TU (stock number 0121.0) is from Tucson, Arizona, USA; the BO strain (stock number 0121.35) is from Bogota, Colombia. The population from Bogota is isolated from the main distribution area of the species and is considered a different subspecies (Ayala and Dobzhansky 1974; Orr 1989).

PCR Amplification, Gel Electrophoresis, and Direct DNA Sequencing. Single fly genomic DNA extraction was performed by slightly modifying the method of Jowett (1986). About 100 ng of DNA were used as a template in the PCR amplification reactions, which were carried out in a final volume of 20 μl, according to Jeffreys et al. (1988b). AmpliTaq polymerase from Perkin-Elmer-Cetus was used and reactions were cycled for 1 min at 95°C, 1 min at 65°C, and 1 min at 70°C for 30 cycles in a DNA thermal cycler (either a Techne Programmable Dri-Block PHC-1 or a Perkin-Elmer-Cetus DNA thermal cycler).

In *D. melanogaster* two 24-mer oligonucleotides with the following sequences were used as primers: 5′-CCCGTCCACGAGGGCAGCGGGGGC-3′ and 5′-CCGCGCGACTCCCGGTGCTTCTTC-3′. The first primer corresponds to the coding sequence numbered 5005–5028 in the sequence published by Citri et al. (1987) and the second to nucleotide positions 5364–5387.

In *D. pseudoobscura* two 34-mer oligonucleotides with the following sequences were used as primers: 5′-<u>TCACCGGTGAATTC</u>AACTATAACGAGAACCTGCT-3′ and 5′-<u>TCACCGGTGAATTC</u>TTCTCCATCTCGTCGTTGTG-3′. These primers have a 14-nucleotide 5′ extension (underlined) containing an efficiently cleaved EcoRI site that can be used to clone the amplified fragments if necessary (Jeffreys et al. 1990). The first primer corresponds to amino acid positions 579–586 and the second to positions 878–884 in the sequence published by Colot et al. (1988).

Drosophila melanogaster PCR products were electrophoresed through a 2.5–3.5% low melting point NuSieve (GTG) agarose gel. TBE buffer (0.045 M Tris-borate, 0.001 M EDTA) was used, and PCR-amplified DNAs from stocks carrying either 17, 20, or 23 pairs of Thr-Gly's (Yu et al. 1987), hereafter referred to as (Thr-Gly)$_{17}$, (Thr-Gly)$_{20}$, and (Thr-Gly)$_{23}$, respectively, were used as markers. In the case of *D. pseudoobscura* a 1.5% agarose gel was used and the PCR-amplified DNA from the plasmid containing the AY strain cloned gene (Colot et al. 1988) was used as a marker.

Double-stranded direct DNA sequencing of PCR products was carried out according to Bachmann et al. (1990) with the dideoxy chain-termination method (Sanger et al. 1977) using the Sequenase version 2.0 kit from United States Biochemical. The PCR-reamplified DNA to be sequenced was purified in a low melting point agarose gel (NuSieve) and recovered by phenol and double phenol–chloroform extraction followed by ethanol precipitation. The same primers used in PCR amplification were used in the direct DNA sequencing. In the case of *D. pseudoobscura* two additional internal primers (5′-ACATGAGTAGTGCGACCAAC-3′ and 5′-GCTGACAACTATGCAGT-3′) were also used.

Analysis of the DNA Sequences. The sequences were analyzed using the University of Wisconsin Genetics Computer Group (UWGCG) software (Version 4.0; Devereux et al. 1984). DNA sequence alignment was performed using the UWGCG program BESTFIT and final alignment carried out by eye. Codon usage was analyzed using the UWGCG program CODONFREQUENCY.

Protein Secondary Structure Prediction. The JOINT PREDICTION program of the Secondary Structure Prediction Suite Version 3.0 kindly provided by Dr. E. Eliopoulos (University of Leeds) was used to predict the secondary structure of the conceptual *per* proteins of *D. melanogaster* and *D. pseudoobscura*. The program provides a consensus secondary structure obtained from the following eight different methods: Kabat and Wu (1973); Nagano (1973); Burgess et al. (1974); Chou and Fasman (1974); Lim (1974); Dufton and Hider (1977); McLachlan (1977); and Garnier et al. (1978). The graphical representation of the Chou and Fasman (1974) predicted secondary structure was obtained using the UWGCG programs CHOUFAS and PLOTCHOU.

Results

Length Polymorphism in D. melanogaster

Using PCR amplification and agarose gel electrophoresis we analyzed the Thr-Gly-encoding repeat region of 110 single flies from four natural populations of *D. melanogaster* (Zakynthos in Greece, Pietrastornina and Merano in Italy, and Bordeaux in France). Table 1 shows the frequencies of each Thr-Gly length variant found in each population. The Zakynthos population seems to be monomorphic for the (Thr-Gly)$_{17}$ variant. The other three populations were highly polymorphic with the (Thr-Gly)$_{17}$, (Thr-Gly)$_{20}$, and (Thr-Gly)$_{23}$ variants being the most frequent. Rare variants were also observed with mobilities corresponding to 14 and 18 Thr-Gly pairs. Figure 1 illustrates a separation of PCR-amplified variants as obtained by agarose gel electrophoresis (see Materials and Methods). The (Thr-Gly)$_{14}$ (lane *d*), (Thr-Gly)$_{17}$ (lanes *c* and *e*), (Thr-Gly)$_{20}$ (lane *h*), and (Thr-Gly)$_{23}$ (lane *i*) variants are clearly identified. Differences due to only one Thr-

Table 1. Thr–Gly length variant frequencies in four natural populations of *D. melanogaster*

Population	Alleles sampled	Thr-Gly length variants				
		14	17	18	20	23
Zakynthos (Greece)	22	–	1.00	–	–	–
Merano (north Italy)	30	0.03	0.40	0.03	0.33	0.20
Pietrastornina (south Italy)	46[a]	0.02	0.56	–	0.33	0.09
Bordeaux (France)	23	–	0.52	–	0.35	0.13

[a] Eleven females are also included in this sample, i.e., 22 alleles

Gly pair (six nucleotides) are also detectable as shown in the case of the $(\text{Thr-Gly})_{18}$ variant (lane *g*). Lane *f* shows the pattern obtained from a $(\text{Thr-Gly})_{14/17}$ heterozygous female. The two closely spaced bands that migrated more slowly are due to heteroduplex formation.

DNA Sequencing Analysis of Thr-Gly Variants

By direct DNA sequencing of these PCR products we analyzed a number of Thr-Gly variants from each population, including the three rare variants we found [two of these represented the $(\text{Thr-Gly})_{14}$ and one the $(\text{Thr-Gly})_{18}$ variant]. In Fig. 2 the DNA sequences encoding the Thr-Gly repeats are shown.

The length of the Thr-Gly repeat in the rare variants estimated on the basis of their relative electrophoretic mobility (see Fig. 1) was confirmed. The DNA sequence analysis of three $(\text{Thr-Gly})_{17}$ variants (two from Zakynthos, ZA1 and ZA2, and one from Pietrastornina, PI1) and two $(\text{Thr-Gly})_{20}$ (one from Pietrastornina, PI2, and one from Merano, ME1) from wild populations matched those found previously in laboratory stocks (Yu et al. 1987) except for one synonymous substitution (C → A) in the PI1 variant. The sequences of three $(\text{Thr-Gly})_{23}$ variants from different populations (one from Pietrastornina, PI3; one from Merano, ME2; and one from Bordeaux, BX1) were also obtained. These $(\text{Thr-Gly})_{23}$ variants share exactly the same sequence except for one synonymous substitution (C → T) in the last Thr-Gly-encoded pair in the BX1 variant. However, all three had diverged significantly from a $(\text{Thr-Gly})_{23}$ variant previously described in a Canton S laboratory stock, originally collected in Ohio, USA (Lindsley and Grell 1972; Jackson et al. 1986; Yu et al. 1987), labeled 'CS' in Figs. 2 and 3.

In Fig. 2, the alignment of the sequences of the different Thr-Gly variants takes into account putative deletion/insertion events. This alignment has been performed with the aim of obtaining a parsimonious solution to the question of the origin of each Thr-Gly length variant. The approach we adopted minimizes the number of synonymous substitutions between two different variants and assumes that only one deletion/insertion event occurred. Yu et al. (1987) have shown that the $(\text{Thr-Gly})_{17}$, $(\text{Thr-Gly})_{20}$, and $(\text{Thr-Gly})_{23}$ variants from laboratory stocks differ from each other by the presence of one, two, or three copies, respectively, of an 18-bp perfect direct repeat (indicated by the arrow ——————> in Fig. 2). Consequently, they can originate by deletion/insertion events in this region. The $(\text{Thr-Gly})_{18}$ variant we found in the Merano population (ME3) differs from the ZA1 and ZA2 $(\text{Thr-Gly})_{17}$ variant by just a 6-bp repeat (ACAGGT, indicated by the arrow ==> in Fig. 2). This repeat appears in three tandem copies in ME3 and only in two in the $(\text{Thr-Gly})_{17}$ variants.

When the $(\text{Thr-Gly})_{23}$ variant that we found in Merano (ME2) and Pietrastornina (PI3) was compared with the Canton S (CS) $(\text{Thr-Gly})_{23}$ variant, we observed that 21 synonymous substitutions were required for the alignment (Fig. 3). The same figure reveals that 22 substitutions were required in the case of the $(\text{Thr-Gly})_{23}$ BX1 variant. We discuss the possible origin of these variants below.

A parsimonious model indicates that the $(\text{Thr-Gly})_{23}$ variant from Merano (ME2) and Pietrastornina (PI3) gave rise to the $(\text{Thr-Gly})_{18}$ variant (ME3) by the occurrence of a single 30-bp deletion (see Fig. 2). One synonymous substitution (C → A) is also observed in ME3 in the last nucleotide immediately upstream of the last encoded Thr-Gly pair (indicated by an asterisk). The $(\text{Thr-Gly})_{14}$ variant from Merano (ME4) and Pietrastornina (PI4) can also be derived from the same $(\text{Thr-Gly})_{23}$ variant by the occurrence of a single 54-bp deletion. It is interesting to note that the hypothesized deletion breakpoints in the $(\text{Thr-Gly})_{23}$ variants [to give the $(\text{Thr-Gly})_{14}$ and $(\text{Thr-Gly})_{18}$ variants] are preceded by perfect 6-bp motifs (indicated, respectively, by the arrows ——> and ++>) that might work as a possible substrate for slippage or unequal exchange events (Levinson and Gutman 1987; Dover 1989; Jarman and Wells 1989). The $(\text{Thr-Gly})_{17}$ alleles ZA1, ZA2, and PI1 can be derived from the $(\text{Thr-Gly})_{18}$ variant (ME3) by a single 6-bp deletion (plus one synonymous substitution, C → A in the case of PI1). The $(\text{Thr-Gly})_{20}$ (ME1 and PI2) variant can originate from the $(\text{Thr-Gly})_{17}$ variant (ZA1 and ZA2) by an insertion involving the 18-bp motif mentioned above. The CS (Thr-Gly)$_{23}$ variant can in turn originate from the (Thr-Gly)$_{20}$ variant by the same 18-bp insertion. Therefore, by invoking two deletions, two insertions, and one point mutation we can derive the CS (Thr-Gly)$_{23}$ variant from the European (Thr-Gly)$_{23}$ ME2 and PI3 allele. This presents a more parsimonious solution to the origin of these two variants compared to a model that requires 21 (or 22) point mutations. Figure 4 shows a tentative phylogeny of the different Thr-Gly length alleles based on these assumptions.

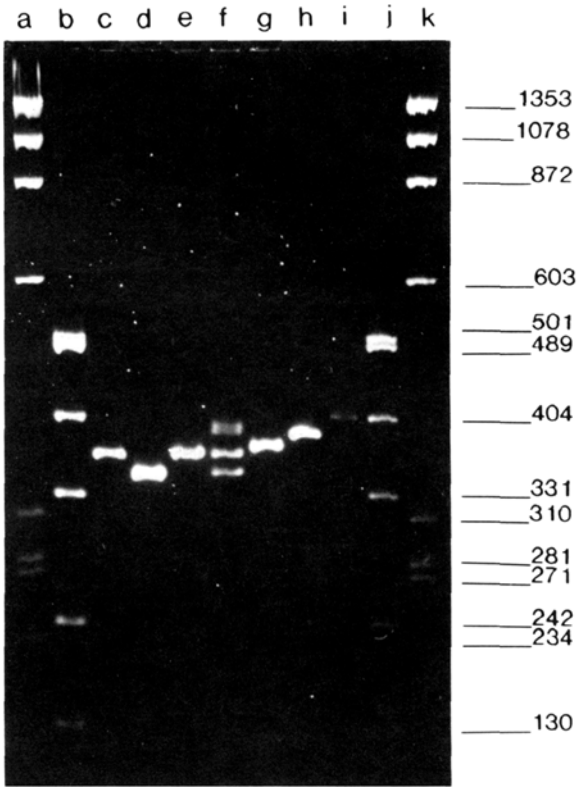

Fig. 1. PCR-amplified Thr-Gly length variants separated on a 3.5% low melting point agarose gel (NuSieve). Lanes *a* and *k*, ϕX174 RF/HaeIII; lanes *b* and *j*, pUC13/HpaII; lanes *c* and *e*, a (Thr-Gly)$_{17}$ variant previously sequenced (Yu et al. 1987); lane *d*, ME4 variant (14) Thr-Gly pairs); lane *f*, PI4/PI1 variants, heterozygous female (14 and 17 Thr-Gly pairs); lane *g*, ME3 variant (18 Thr-Gly pairs); lanes *h* and *i*, respectively, a (Thr-Gly)$_{20}$ and a (Thr-Gly)$_{23}$ variant previously sequenced (Yu et al. 1987). In lane *f* the two slowly migrating, closely spaced bands are due to heteroduplex formation.

Length Variants in D. pseudoobscura

Figure 5 shows the sequences of the region corresponding the Thr-Gly domain in *D. melanogaster* in three different strains of *D. pseudoobscura* (PA from Pachuca, Mexico; TU from Tucson, AZ, USA; and BO from Bogota, Colombia) compared with the sequence of the Ayala reference strain (AY) already published (Colot et al. 1988). This sequence analysis was carried out after size differences in PCR-amplified fragments were detected by agarose gel electrophoresis.

Because only four different sequences were available, no attempt was made to reconstruct a possible origin of the different variants. Moreover, a preliminary analysis of the sequences revealed more complexity than in *D. melanogaster*. Therefore the sequences were aligned with the published AY sequence (Colot et al. 1988). The length variants from the three strains show a number of synonymous and nonsynonymous nucleotide substitutions when compared to the AY reference strain. Additionally, our sequencing analysis indicates that the length polymorphism is due to deletion/insertion events involving 15 nucleotides. The three strains (PA, TU, and BO) share one insertion and two synonymous substitutions in comparison with the AY strain. The PA strain sequence also has one additional synonymous substitution. The TU strain shows an additional insertion, three more synonymous substitutions, and three other nonsynonymous substitutions, two of them in the same codon. The BO strain shows one additional insertion, which is different from the insertion in TU but encodes the same five amino acids. One further nonsynonymous substitution is found in BO. It is intriguing that TU presents more differences in relation to AY and PA than does BO in spite of the fact that the latter is a different subspecies (Ayala and Dobzhansky 1974; Orr 1989).

In *D. pseudoobscura* the repeat region encodes a short run of Thr-Gly imperfect repeats (interrupted once by the sequence Thr-Ile-Ile-Ala-Thr-Ser-Gly), followed by a large amino acid sequence that contains a long degenerate repeat of a five-amino acid motif (Colot et al. 1988). A detailed analysis of the DNA sequence in this region shows the presence of 32–34 (32 in AY, 33 in PA, and 34 in TU and BO) degenerate tandem repeats of a 15-nucleotide motif (indicated by arrows in Fig. 4) with the GGCGCCGACAACTCT consensus sequence (see Table 2). Each repeat in the four sequences always has at least eight bases identical to the consensus. A consensus motif of five amino acids (Gly-Ala-Asp-Asn-Ser) can also be drawn from the conceptual protein (Table 2). Finally, it can be deduced from the sequence data that the length polymorphism in the repeat region of *D. pseudoobscura* is due to a variation in the number of the 15-nucleotide motif and not in the number of the encoded Thr-Gly pairs.

Predicted Secondary Structure of the Conceptual per Protein

Using several different methods (see Materials and Methods), a predicted consensus secondary structure of the conceptual protein of *D. melanogaster*

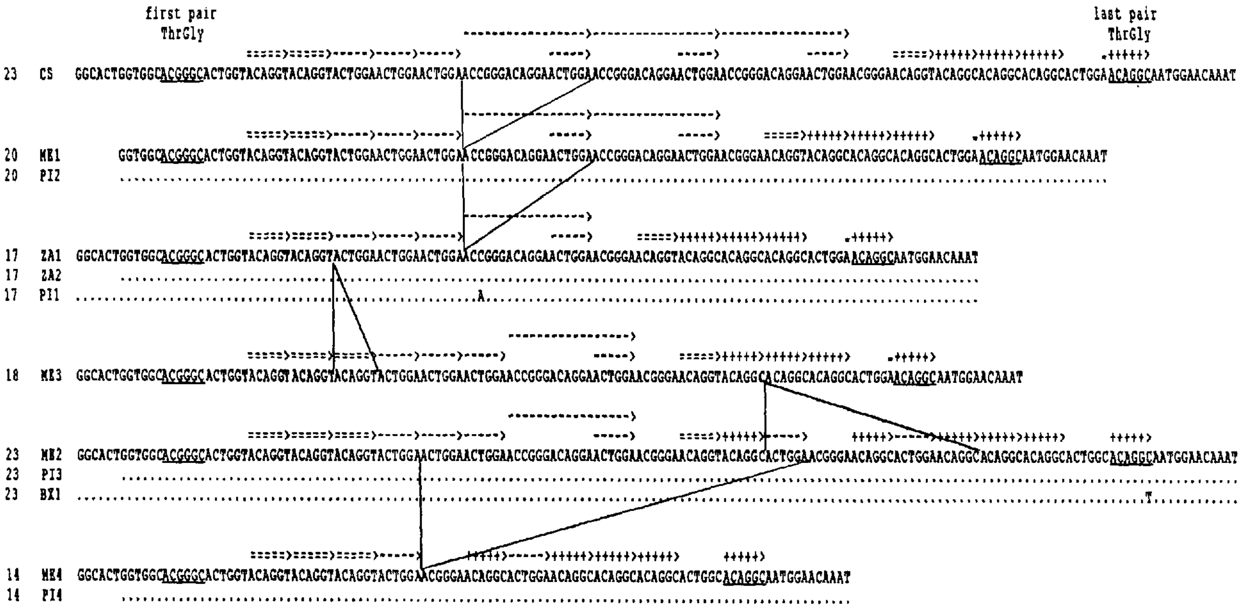

Fig. 2. DNA sequences of Thr-Gly length variants from different natural populations compared with the Canton S (CS) sequence (Jackson et al. 1986; Yu et al. 1987). ME1 and PI2 are $(\text{Thr-Gly})_{20}$ variants from Merano and Pietrastornina, respectively; ZA1, ZA2, and PI1 are $(\text{Thr-Gly})_{17}$ variants from Zakynthos and Pietrastornina; ME3 is the $(\text{Thr-Gly})_{18}$ variant from Merano; ME2, PI3, and BX1 are $(\text{Thr-Gly})_{23}$ variants from Merano, Pietrastornina, and Bordeaux, respectively; ME4 and PI4 are $(\text{Thr-Gly})_{14}$ variants from Merano and Pietrastornina, respectively. The first and last pairs of the encoded Thr-Gly run are underlined. The arrows ——————>, ==>, ++>, and ——>) indicate direct repeats that may be involved in the origin of different alleles (see text for details). The lines connecting different length variants show the putative deletion/insertion events. Only point mutations between variants of the same length are indicated. The asterisks indicate a synonymous nucleotide difference between the ME2, PI3, BX1, ME4, PI4 group of alleles and the ME3, PI1, ZA2, ZA1, PI2, ME1, CS group. The number preceding the acronym used to label each sequenced allele refers to the encoded number of Thr-Gly pairs.

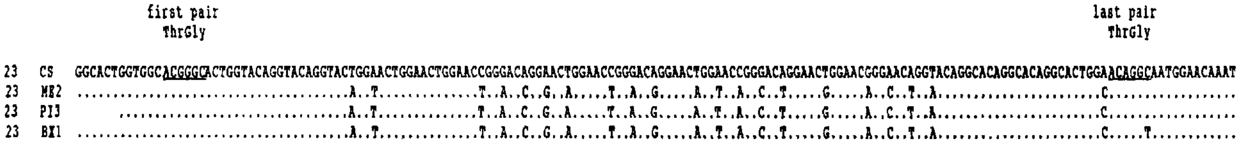

Fig. 3. Alignment between the Canton S (CS) (Jackson et al. 1986; Yu et al. 1987) and the Merano (ME2), Pietrastornina (PI3), and Bordeaux (BX1) $(\text{Thr-Gly})_{23}$ length variants. See also legend in Fig. 2.

was obtained. A continuous series of turns for the Thr-Gly-encoding region is predicted. This region of turns seems to subdivide the conceptual *per* protein into two main globular domains. Similar results have been obtained for the degenerate five-amino acid repeat in the corresponding region of the *per* protein of *D. pseudoobscura.* Figure 6 shows a graphic representation of the obtained *per* protein secondary structure of both species when the Chou and Fasman (1974) method is applied.

Codon Usage within the Repeats

We examined the codon usage for Thr and Gly within the repeat of the *D. melanogaster* CS sequence (Jackson et al. 1986) and compared it with the usage in the remainder of the gene. We obtained a G value of 12.66 (df = 3, $P < 0.01$) for Thr and a G value of 14.94 (df = 3, $P < 0.01$) in the case of Gly. In the *D. pseudoobscura* AY sequence (Colot et al. 1988) a significant difference in codon usage was also observed for Ser ($G = 58.42$, df = 5, $P \ll 0.01$) out of the consensus five amino acids encoded by the degenerate repeat, when compared to the rest of the gene.

Discussion

The repeat encoding region in the *per* genes of *D. melanogaster* and *D. pseudoobscura* confirms some of the expectations of theoretical models concerning the evolution of tandem-repetitive DNA sequences through replication slippage and/or unequal sister-chromatid exchange (Smith 1976; Dover 1987). For example, we observed the existence of higher-order structures such as the 18-bp "cassette" and the clusters of 6-bp direct repeats in *D. melanogaster.* Another likely consequence of replication slippage and/

or unequal sister-chromatid exchange in this coding region is the observed significant bias in the codon usage within the repeats.

The phylogeny of the different *D. melanogaster* Thr-Gly length variants shown in Fig. 4 and derived from the data in Fig. 2 was obtained using an arbitrarily chosen parsimony criterion, which minimized both the number of point mutations and deletion/insertion events between two alleles. It assumes that the alleles originated by slippage and/or unequal sister-chromatid exchange. Indeed, these seem to be the major mechanisms by which new alleles arise in minisatellite sequences (Jeffreys et al. 1990). The resulting model predicts that one of the $(\text{Thr-Gly})_{23}$ alleles found in the European populations (ME2, PI3, and BX1) is representative of the ancestral state compared to the others we examined, because they can all be derived simply from these $(\text{Thr-Gly})_{23}$ alleles. Furthermore our results show that without a careful analysis of several DNA sequences from other natural Thr-Gly alleles, we would have made serious error in interpreting the evolutionary history of the two $(\text{Thr-Gly})_{23}$ variants. Gene conversion and unequal crossing-over between the different length variants may also play a role in the evolution of this region. Perhaps this could explain the apparently complex pattern of polymorphism in *D. pseudoobscura* where in the TU strain a 3-bp substitution occurs as a cluster.

It is expected that the tandem-repetitive coding sequences in the *per* genes of these two species may be under different constraints from noncoding repetitive DNA sequences (Jeffreys et al. 1988a, 1990; Jarman and Wells 1989). An obvious requirement is the maintenance of the reading frame. However it is interesting to observe that, with the exception of the rare $(\text{Thr-Gly})_{18}$ variant, all the other Thr-Gly alleles in *D. melanogaster* differ by multiples of three Thr-Gly pairs. Could this suggest some kind of functional constraint at the level of the protein? Our observation of a convergence producing two kinds of $(\text{Thr-Gly})_{23}$ variants (CS, originally collected from the USA, and those found in European natural populations) further supports this intriguing possibility. In this respect it would be interesting to know whether the CS $(\text{Thr-Gly})_{23}$ variant found in the laboratory reflects a common variant in American natural populations.

Analysis of the predicted secondary structure of the conceptual protein of *D. melanogaster* using eight different methods indicates that the Thr-Gly region may represent a series of turns that act as a spacer dividing the *per* protein into two main globular domains. Similar secondary structure is obtained for the *D. pseudoobscura per* protein in which the long *D. melanogaster* Thr-Gly repeat is replaced by a degenerate five-amino acid repeat. This suggests that in spite of the high divergence in the primary protein sequences between the two species in this region, the secondary structure is quite conserved.

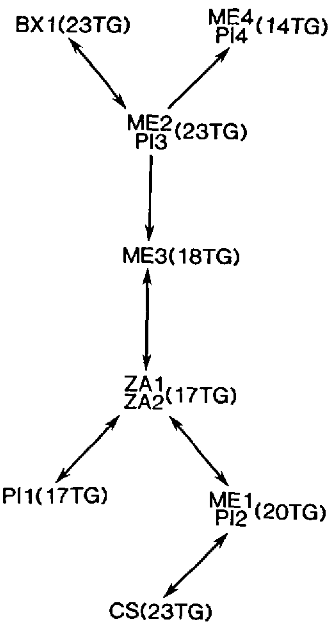

Fig. 4. Hypothetical phylogeny of the different Thr-Gly alleles based on the alignment of the sequences in Fig. 2. Arrows indicate the direction of the hypothesized deletions/insertions and point mutation events needed to derive one allele from another.

Could the length polymorphism in the repeats have selective value? A recent result from Ewer et al. (1990), who examined the circadian behavior of arrhythmic *D. melanogaster* flies that had been transformed with a *per* gene from which the Thr-Gly-encoding motif had been deleted (Yu et al. 1987), may be relevant. They observed that the circadian period in locomotor activity of these transformants became temperature dependent. This result suggests that the Thr-Gly repeat may play a role in the thermostability of the *per* protein. The absence of such temperature sensitivity is a cardinal feature of a true biological clock and is termed temperature compensation (Pittendrigh 1954). Consequently, perhaps different length variants may have altered temperature compensation properties when challenged with different temperature regimes.

Finally, in *D. melanogaster* the island population from Zakynthos appears monomorphic for the $(\text{Thr-Gly})_{17}$ variant. If other length variants exist in this population, they will certainly have low frequencies. We can almost certainly exclude inbreeding and founder effects to explain this lack of variation as two gene–enzyme systems, esterase-6 (*Est-6*) and alcohol dehydrogenase (*Adh*), were studied in this population, and the results revealed high levels of polymorphism (R. Costa, unpublished). The pattern of allele frequencies was typical for Mediterranean natural populations (Anderson and Oakeshott 1984; David et al. 1988). The apparent monomorphism

```
AY  CCAAAAGTGGGGTCCTCGGATGTGAGCAGCACCCGCGAGGATGCCCGCAGCACGCTTAGCCCCCTGAACGGCTTCGAGGGCAGTGGCGCCAGTGGCTCCT
PA  ....................................................................................................
TU  ....................................................................................................
BO  ....................................................................................................

                                                                                     imperfect
                                                                    first           Thr-Gly pairs
AY  CAGGCCACTTGACCAGCGGCAGCAATATACACATGAGTAGTGCGACCAACACAAGCAATGCTGGAACGGGCACTGGTACGGTCACGGGAACCGGCACAAT
PA  ....................................................................................................
TU  ....................................................................................................
BO  ....................................................................................................

                            last            Gly
                       -------------->-------------->-------------->-------------->-------------->------
AY  AATAGCCACCTCCGGGACCGGCACTGTCACCTGCGCCTCCGGCAACATGGACGCCAACACCTCTGCGGCCTTCAACATTGCCGCCAACACCTCTGCCGCC
PA  ....................................................................................................
TU  ....................................................................................................
BO  ........................................A...........................................................
                                            Ser

       Asn
    -------->-------------->-------------->-------------->-------------->-------------->-------------->-
AY  GACAACTTTGGCGCCGATACCTCTGCCGCTGACACCTCTGGCGCCGACACCTCTGCTGCTGACAACTAT---------------GGCCCCGGCAACTTTG
PA  .....................................................................GCAGTCGACAACTAT................
TU  ...C.................................................................GCAGTCGACAACTAT................
BO  .....................................................................GCAGTCGACAACTAT................
       His                                                               AlaValAspAsnTyr

                                                                                   Asp
    ------------->-------------->-------------->-------------->-------------->-------------->-----------
AY  GCGCAGAAAACTCTTGCGCCGATAACTCTGGCGCCGAAAACTCATGCGCAGATAACTCTGGCGTCGATAACTCCCGGCCCGATAACTCTGGGGCCGATAA
PA  ....T...............................................................................................
TU  ....T.........................................................................TTG..................
BO  ....T...............................................................................................
                                                                                   Cys

    --->-------------->-------------->-------------->-------------->-------------->-------------->------
AY  CTCTGCGGCCGATAACTTT---------------GGGCCCGACAATTCCGGGGCCGACAATTCC---------------GGGCCCGACAACACTGGACCC
PA  ...................---------------.................C............---------------.....................
TU  ...................GGCGCCGACAACTCT..A...........................---------------..A..................
BO  ...................---------------..............................GGCGCCGACAATTCC.....................
                       GlyAlaAspAsnSer                              GlyAlaAspAsnSer

    -------->-------------->-------------->-------------->-------------->-------------->-------------->-
AY  GACAATTCCGGCGCCGAAAACTCTCGGGCCGAAAACTCTCGAGCCGATAACTCTAGACCCGACCACCCTAGACCCGACATCTCTGGCGCCAGCAATTCTC
PA  ....................................................................................................
TU  ....................................................................................................
BO  ....................................................................................................

    ------------->-------------->-------------->
AY  GACCCGACAAAACTGGACCCGACAAGTCGGGCGCCGAAAACTCTGCCTCTGGATCGGGATCGGGCACCTCCGGCAACGAAGGTCCCTCCAGTGGTGGGCA
PA  ....................................................................................................
TU  ....................................................................................................
BO  ....................................................................................................

AY  GGACACCAGGACCACCGCTGGGACT
PA  .........A...............
TU  .........A...............
BO  .........A...............
```

Fig. 5. Sequences of length variants in the Thr-Gly-encoding region of different strains of *D. pseudoobscura* compared with the published sequence of the Ayala (AY) reference strain (Colot et al. 1988). PA, Pachuca, Mexico; TU, Tucson, Arizona, USA; BO, Bogota, Colombia. The first and last pairs of the Thr-Gly imperfect repeat are underlined. The arrows (——————>) indicate the degenerate tandem repeats of the 15-nucleotide motif GGCGCCGACAACTCT. Each repeat in the four sequences has at least eight bases identical to the consensus. Synonymous and nonsynonymous substitutions as well as the insertions found in relation to the AY strain are also shown.

Table 2. Consensus sequence of the 15-nucleotide degenerate repeat of *D. pseudoobscura*

G	G	C	G	C	C	G	A	C	A	A	C	T	C	T
75	75	53	66	97	84	88	84	60	97	78	81	84	78	75
	Gly			Ala			Asp			Asn			Ser	
	50			63			63			69			69	

The numbers below each base represent their percentage frequency of occurrence as calculated for the 32 repeats of the Ayala reference strain (AY) (Colot et al. 1988). A consensus motif of five amino acids (Gly-Ala-Asp-Asn-Ser) can also be drawn from the conceptual protein. The numbers below each amino acid represent their frequency of occurrence in percentage

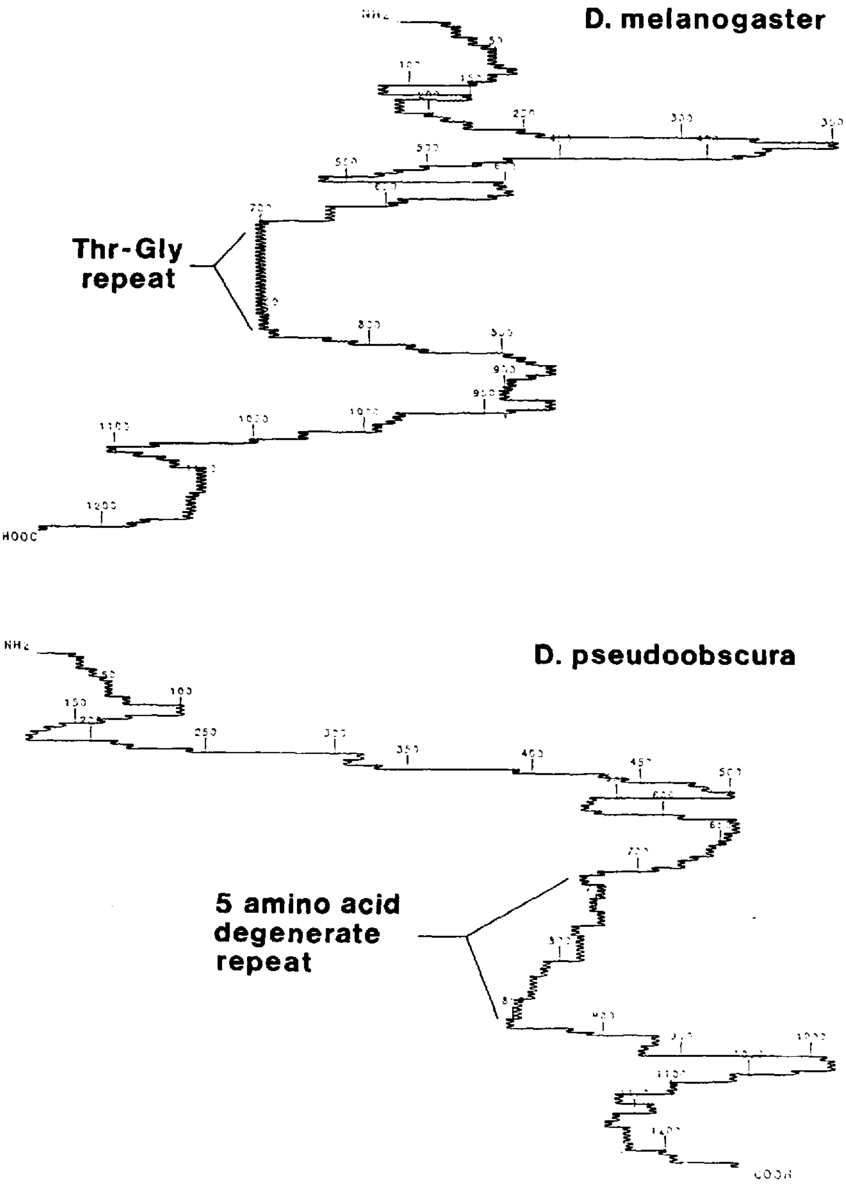

Fig. 6. The predicted secondary structure of the *per* protein in *D. melanogaster* and *D. pseudoobscura* using the Chou and Fasman (1974) method. The position of the Thr-Gly repeat of *D. melanogaster* and the five-amino acid degenerate repeat of *D. pseudoobscura* are indicated. Both these regions are characterized by a series of turns that subdivide the conceptual proteins into two main globular domains. The amino acids of the *per* protein are numbered from the amino terminus.

at the *per* locus could be due to selection or to the accidental fixation of the $(\text{Thr-Gly})_{17}$ variant by genetic turnover mechanisms (Dover 1989).

In summary our results suggest that length variation will be a common finding in the repetitive encoding region of the *per* gene in *Drosophila* species, irrespective of the amino acid moiety that is encoded by the repeat motif. Furthermore, our sequence analyses highlight some of the dangers inherent in drawing premature evolutionary histories from superficial DNA examination of internally repetitive genes.

Acknowledgments. We acknowledge the support of an EMBO short-term fellowship and a NATO senior fellowship to R.C., a Brazilian CNPq scholarship to A.A.P., a SERC studentship to J.R.T., and SERC grant GRF41778 to C.P.K. We are grateful to Dr. John Keyte for synthesizing the primers and Dr. Michael Rosbash for sending us the plasmid containing the *D. pseudoobscura* material. We thank Mark Blight for his help with the secondary structure prediction programs and we appreciate comments on the manuscript from Mark Blight, John Armour, Mike Ritchie, Nicola Royle, Paul Sharp, Gabby Dover, and Alec Jeffreys.

References

Anderson PR, Oakeshott JG (1984) Parallel geographic patterns of allozyme variation in two sibling *Drosophila* species. Nature 308:729–731

Ayala FJ, Dobzhansky T (1974) A new subspecies of *Drosophila pseudoobscura* (Diptera: Drosophilidae). Pan-Pac Entomol 50: 211–219

Bachmann B, Luke W, Hunsmann G (1990) Improvement of PCR amplified DNA sequencing with aid of detergents. Nucleic Acids Res 18:1309

Burgess AW, Ponnuswamy PK, Scheraga HA (1974) Analysis of conformation of amino acid residues and prediction of backbone topography in proteins. Isr J Chem 12:239–286

Chou PY, Fasman GD (1974) Predictions of protein conformation. Biochemistry 13:222–245

Citri Y, Colot HV, Jacquier AC, Yu Q, Hall JC, Baltimore D, Rosbash M (1987) A family of unusually spliced biologically active transcripts encoded by a *Drosophila* clock gene. Nature 326:42–47

Colot HV, Hall JC, Rosbash M (1988) Interspecific comparison of the *period* gene of *Drosophila* reveals large blocks of non-conserved coding DNA. EMBO J 7:3929–3937

David JR, Alonso-Moraga A, Borai F, Capy P, Mercot H, McEvey SF, Munoz-Serrano A, Tsakas S (1989) Latitudinal variation of *Adh* gene frequencies in *Drosophila melanogaster*: a Mediterranean instability. Heredity 62:11–16

Devereux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res 12:387–396

Dover GA (1987) DNA turnover and the molecular clock. J Mol Evol 26:47–58

Dover GA (1989) Slips, strings and species. Trends Genet 5: 100–102

Dufton MJ, Hider RC (1977) Snake toxin secondary structure predictions. Structure activity relationships. J Mol Biol 115: 177–193

Ewer J, Hamblen-Coyle M, Rosbash M, Hall JC (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

Garnier J, Osguthorpe DJ, Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120: 97–120

Goodbourn SEY, Higgs DR, Clegg JB, Weatherall DJ (1983) Molecular basis of length polymorphism in the human ζ-globin gene complex. Proc Natl Acad Sci USA 80:5022–5026

Hall JC, Kyriacou CP (1990) Genetics of biological rhythms in *Drosophila*. Adv Insect Physiol 22:221–298

Jackson FR, Bargiello TA, Yun S-H, Young MW (1986) Product of the *per* locus of *Drosophila* shares homology with proteoglycans. Nature 320:185–188

Jarman AP, Wells RA (1989) Hypervariable minisatellites: recombinators or innocent bystanders? Trends Genet 5:367–371

Jeffreys AJ, Royle NJ, Wilson V, Wong Z (1988a) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. Nature 332:278–281

Jeffreys AJ, Wilson V, Neumann R, Keyte J (1988b) Amplification of human minisatellites by polymerase chain reaction: towards DNA fingerprinting of single cells. Nucleic Acids Res 16:10953–10971

Jeffreys AJ, Neumann R, Wilson V (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

Jowett T (1986) Preparation of nucleic acids. In: Roberts DB (ed) *Drosophila*. A practical approach. IRL Press, Oxford, pp 275–286

Kabat EA, Wu TT (1973) The influence of nearest-neighbour amino acids on the conformation of the middle amino acid in proteins: comparisons of predicted and experimental determination of β-sheets in concanavalin A. Proc Natl Acad Sci USA 70:1473–1477

Kyriacou CP (1990) The molecular ethology of the *period* gene in *Drosophila*. Behav Genet 20:191–211

Levinson G, Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol 4:203–221

Lim VI (1974) Structural principles of the globular organization of protein chains. A stereochemical theory of globular protein secondary structure. J Mol Biol 88:857–872

Lindsley DL, Grell EH (1972) Genetic variations of *D. melanogaster*. Carnegie Inst Washington Publ 627:1–471

Lyons KM, Stein JH, Smithies O (1988) Length polymorphism in human proline-rich protein genes generated by intragenic unequal crossing-over. Genetics 120:267–278

McLachlan AD (1977) Quantum chemistry and protein folding: the art of the possible. Int J Quantum Chem [Suppl 1] 12: 371–385

Muskavitch MAT, Hogness DS (1982) An expandable gene that encodes a *Drosophila* glue protein is not expressed in variants lacking remote upstream sequences. Cell 29:1041–1051

Nagano K (1973) Logical analysis of the mechanism of protein folding. I. Predictions of helices, loops and β structures from primary structure. J Mol Biol 75:401–420

Orr HA (1989) The genetics of sterility in hybrids between two subspecies of *Drosophila*. Evolution 43:180–189

Pittendrigh CS (1954) On the temperature independence in the clock system controlling emergence time in *Drosophila*. Proc Natl Acad Sci USA 40:1018–1029

Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463–5467

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191:528–535

Swallow DM, Gendler S, Taylor-Papadimitriou J, Bramwell ME (1987) The human tumor-associated epithelial mucins are coded by an expressed hypervariable gene locus PUM. Nature 328:82–84

Tautz D, Trick M, Dover GA (1986) Cryptic simplicity in DNA is a major source of genetic variation. Nature 322:652–656

Teumer J, Green H (1989) Divergent evolution of part of the involucrin gene in the hominoids: unique intragenic duplications in the gorilla and human. Proc Natl Acad Sci USA 86:1283–1286

Thackeray JR, Kyriacou CP (1990) Molecular evolution in the *Drosophila yakuba period* locus. J Mol Evol 31:389–401

Treier M, Pfeifle C, Tautz D (1989) Comparison of the gap segmentation gene *hunchback* between *Drosophila melanogaster* and *Drosophila virilis* reveals novel modes of evolutionary change. EMBO J 8:1517–1525

Wheeler DA, Kyriacou CP, Greenacre M, Yu Q, Rutila J, Rosbash M, Hall JC (1991) Molecular transfer of a species-specific courtship behaviour from *Drosophila simulans* to *Drosophila melanogaster*. Science (in press)

Yu Q, Colot HV, Kyriacou CP, Hall JC, Rosbash M (1987) Behaviour modification by in vitro mutagenesis of a variable region within the *period* gene of *Drosophila*. Nature 326:765–769

Received September 4, 1990/Revised November 19, 1990