Genet. Res., Camb. (1997), **69**, *pp.* 89–99. *With 4 figures.* Printed in the United Kingdom © 1997 Cambridge University Press

Linkage disequilibrium, mutational analysis and natural selection in the repetitive region of the clock gene, *period*, in *Drosophila melanogaster*

EZIO ROSATO[1], ALEXANDRE A. PEIXOTO[1]†, RODOLFO COSTA[2] AND CHARALAMBOS P. KYRIACOU[1]*

[1] *Department of Genetics, University of Leicester, Leicester LE1 7RH, UK*
[2] *Dipartimento di Biologia, Università degli Studi di Lecce, 73100 Lecce, Italy*

(*Received 6 September 1996 and in revised form 17 January 1997*)

Summary

We have used the method of disequilibrium pattern analysis to examine associations between the threonine-glycine (Thr-Gly) encoding repeat region of the clock gene *period* (*per*) of *Drosophila melanogaster*, and polymorphic sites both upstream and downstream of the repeat, in a number of European fly populations. The results are consistent with the view that selection may be operating on various haplotypes which share the Thr-Gly length alleles encoding 17, 20 and 23 dipeptide pairs, and that the repeat itself may be the focus for selection. These conclusions lend support to a number of other population and behavioural investigations which have provided evidence that selection is acting on the Thr-Gly region. The linkage analysis was also used to infer an approximate mutation rate (μ) for the repeat, of $10^{-5} < \mu < 4 \times 10^{-5}$ per gamete per generation. Direct measurements of the mutation rate using the polymerase chain reaction in a pedigree analysis of tens of thousands of individuals do not contradict this value. Consequently, the Thr-Gly repeat does not have a mutation rate that is as high as some of the non-coding minisatellites, but it is several orders of magnitude higher than the nucleotide substitution rate. The implications of this elevated mutation rate for linkage disequilibria and selection are discussed.

1. Introduction

The *period* (*per*) gene in *Drosophila melanogaster* encodes a critical component of the fly's biological timing (Kyriacou, 1994; Hall, 1995). The central portion of the gene contains a threonine-glycine (Thr-Gly) encoding repeat plus some serine-glycine (Ser-Gly) pairs, and is highly polymorphic in length, not only in *D. melanogaster* (Costa *et al.*, 1991, 1992), but also in other drosophilids, where the alternating Thr-Gly pattern can be interrupted by similar repetitive sequences (Colot *et al.*, 1988; Costa *et al.*, 1991; Peixoto *et al.*, 1993; Rosato *et al.*, 1994; Nielsen *et al.*, 1994). In Lepidoptera, the repetitive region of *per* encodes only a single Thr-Gly pair plus some Ser-Gly dipeptides (Reppert *et al.*, 1994), whereas several pairs of Thr-Gly's and Ser-Gly's are found in the clock gene *frequency* (*frq*) of *Neurospora crassa* (McClung *et al.*, 1989), which appears to be an analogue of the dipteran *per* gene (Aronson *et al.*, 1994). Furthermore, these unusual repetitive sequences have been identified in molecules which are implicated in biological cycles in mammals (Matsui *et al.*, 1993).

Interspecific examination of the repeat and its associated sequences has led to the suggestion that the repeat may coevolve with its flanking environment, implying that the Thr-Gly region is under some form of selection (Peixoto *et al.*, 1993; Nielsen *et al.*, 1994). In addition, detailed analysis of the polymorphisms that constitute the three major Thr-Gly haplotypes found in natural populations of *D. simulans*, have revealed significant departures from neutral expectations (Rosato *et al.*, 1994). In natural populations of *D. melanogaster* the two major length variants, (Thr-Gly)$_{17}$ and (Thr-Gly)$_{20}$, which encode 17 and 20 pairs of alternating dipeptides respectively, are spatially differentiated, producing a latitudinal cline in the Northern Hemisphere (Costa *et al.*, 1992). This pattern could be due to selection, demographic phenomena (drift, partial admixture) or a combination of both. Therefore, an examination of neutrality in this region in *D. melanogaster* would be welcomed.

* Corresponding author. e-mail: cpk@leicester.ac.uk.

† Present address: Department of Biology, Brandeis University, Waltham, MA 02254, USA.

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

Kliman & Hey (1993) have performed a number of tests of neutrality on a region of *per* upstream of the Thr-Gly repeat in the species of the *melanogaster* complex, and found no significant departures from neutrality. Unfortunately the Thr-Gly region cannot be studied by this direct approach because the mutational mechanisms that will be acting on the repeat (Dover, 1987; Jeffreys *et al*., 1985, 1990; Harding *et al*., 1992) will violate the assumptions of the infinite sites model (Watterson, 1975) on which the neutrality tests are based. However, similar tests were applied to the Thr-Gly region of *D. simulans*, but this was only possible because the three major Thr-Gly length alleles were in perfect linkage disequilibrium with specific flanking haplotypes. Thus the HKA and Tajima tests could be focused on the immediately adjacent sequences (Rosato *et al*., 1994). In *D. melanogaster* this is not possible. There are 13 isolength alleles, including seven Thr-Gly length classes (Costa *et al*., 1991, 1992; Peixoto *et al*., 1992), and the same Thr-Gly length variant can be associated with more than one flanking haplotype. Furthermore each haplotype can be similarly associated with more than one length variant (Rosato *et al*., 1996). Consequently the results of any neutrality tests on the flanking material cannot be extrapolated to the repeat itself.

In the case of the Thr-Gly encoding repeat, it might be expected that an elevated mutation rate might not only generate new length variants but also convert existing variants into others at relatively high frequencies. This could lead to an absence of linkage disequilibrium between Thr-Gly alleles and nearby flanking markers, although significant disequilibrium might still be observed between these flanking sites themselves. Reduced linkage disequilibrium of the repeat with the upstream region studied by Kliman & Hey (1993) might suggest that any conclusions about neutrality drawn from studying the sequences 5′ to the repeat may not apply to the downstream Thr-Gly region. Consequently a study of the mutation rate of this region and associated linkage disequilibria is crucial for understanding whether any selection is acting on the repeat. Because of the difficulties of applying the classic tests of neutrality to the repeat region mentioned above, we have applied another method – disequilibrium pattern analysis (Thomson & Klitz, 1987; Klitz & Thomson, 1987) – a technique designed to detect present and past selective events in tightly linked loci. The method is not a statistical test, but predicts patterns of linkage disequilibrium when selection favours one haplotype in a population originally in linkage equilibrium, or when a new mutation occurs at one of the two loci. The technique has been used to detect selection for some haplotypes at the *HLA* locus in Danish populations (Klitz & Thomson, 1987), and to confirm a selective scenario proposed for chromosomal inversions in two *Drosophila* species which showed strong seasonal evidence for selection (Peixoto & Klaczko, 1991). We also applied a three-locus procedure, originally used with *HLA* data from French populations, which has been reported to detect selection at a single site within several linked loci (Robinson *et al*., 1991*b*).

We have also attempted to measure the mutation rate of the Thr-Gly region in *D. melanogaster* both directly and indirectly. The direct approach used a polymerase chain reaction (PCR)-based method designed to detect any changes in Thr-Gly length in a pedigree analysis which involved tens of thousands of offspring from several hundred families. The indirect approach gave an estimate of the mutation rate as a by-product of our investigation of linkage disequilibria. The results of our studies appear to support a selective explanation for the patterns of linkage disequilibria observed

2. Materials and methods

(i) *Direct estimation of the Thr-Gly region length mutation rate*

(*a*) D. melanogaster *lines*

In this study we analysed two lines which are homozygous for the two most common European alleles: $(\text{Thr-Gly})_{17a}$ and $(\text{Thr-Gly})_{20a}$ respectively (Rosato *et al*., 1996). They were derived from two isofemale lines set up in 1991 from flies captured in the wild (Cognac, France), and maintained at 25 °C on a sucrose-yeast-agar medium. Males and females from the same line were randomly mated in single crosses and then analysed by PCR (see below) to identify the length alleles carried by each couple. Those crosses where both the male and the female were carrying the same length allele were further characterized by direct PCR sequencing (see below) and thus, at the end of 1992, several homozygous lines for both the 17 and 20 length alleles were established. Of these lines, one homozygous for the $(\text{Thr-Gly})_{17a}$ and one homozygous for the $(\text{Thr-Gly})_{20a}$ allele, survive to date. From each of these lines one male and one female were randomly chosen and crossed. The two Thr-Gly alleles were then sequenced again from these parents to confirm that Thr-Gly region mutations had not occurred since 1992. The F_1 females were collected as virgin flies and mated with their siblings in single crosses. The same was done for the F_2. In total 328 F_2 crosses for the $(\text{Thr-Gly})_{17a}$ allele and 248 F_2 crosses for the $(\text{Thr-Gly})_{20a}$ allele were established. Progeny from both the F_1 and F_2 crosses were labelled to make possible the identification of each lineage. F_1 and F_2 flies were stored at −20 °C in case a mutant was found in the F_3 which might have been present in the F_1 and F_2. The F_3 flies were collected in 1·5 μl microtubes in groups of 1–10 individuals of the same sex from the same cross and immediately frozen at −20 °C. In total, from the F_3, 7126 males and 7829 females were collected for the $(\text{Thr-Gly})_{17a}$ allele;

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

5400 females and 5443 males were collected for the (Thr-Gly)20a allele.

(*b*) *DNA extraction and PCR amplification*

Fly DNA from groups of 1–10 individuals was prepared following the method of Gloor & Engels (1990) appropriately scaled up for the number of individuals extracted. Males (carrying a single copy of the X-linked *per*$^+$ gene) and females (carrying two copies of *per*$^+$) were treated separately. Preliminary experiments were carried out, by mixing different proportions of DNA from the two length alleles, in order to identify the relative ratios at which a rare 17a or 20a 'mutant' amplification band could be discriminated against a background of the alternative allele. The Thr-Gly region was amplified with the 5′ primer 5′-ATACACATGAGCAGTGTGAC-3′ (5066–5085, nucleotide positions as in Citri *et al*., 1987) and the 3′ primer 5′-TCCATCTCGTCGTTGTGCTT-3′ (5333–5352) with 30 cycles of 95 °C 1 min, 65 °C 1 min and 72 °C 1 min followed by 10 min at 72 °C. The 5′ primer was end-labelled with ^{33}P (Amersham) using T4 polynucleotide kinases (Gibco) as in Sambrook *et al*., (1989). PCR was carried out in 7 μl (67 mM Tris-HCl pH 8·8, 16·6 mM $(NH_4)_2SO_4$, 6·7 mM $MgCl_2$, 6·7 μM EDTA pH 8, 170 μg/ml BSA, 10 mM 2-mercaptoethanol, 1·5 mM dATP, 1·5 mM dGTP, 1·5 mM dCTP, 1·5 mM dTTP, 0·2 μM primer 5′, 0·2 μM primer 3′, 1 U Taq polymerase) volume reactions in a Biometra Trio thermocycler. *Taq* polymerase from Advanced Biotechnologies was used. Three microlitres of Stop Solution (95% formamide, 0·05% bromophenol blue, 0·05% xylene cyanole) was added to the PCR reactions which were then boiled for 2 min and 5 μl was loaded on a 4%, 40 cm denaturing polyacrylamide gel. The gel was run for 2·5 h at 50 °C.

The experiments were performed several times with both male and female DNA using many different ratios of the two alleles. In all cases the results were consistent. For males the identification of both $(Thr\text{-}Gly)_{17a}$ and $(Thr\text{-}Gly)_{20a}$ amplification bands was always unambiguous up to a ratio of 1/200, independently of which of the two was the predominant allele. For females, because a hypothetical mutant allele would be carried by a heterozygous female, we measured the ability to identify both amplification bands by mixing, in various proportions, the DNA extracted from a single heterozygous female with the DNA extracted from females homozygous for either the $(Thr\text{-}Gly)_{17a}$ or the $(Thr\text{-}Gly)_{20a}$ allele. In all cases a single 'mutant' length allele could be detected in ratios up to 1/100 heterozygous/homozygous female DNA. We further confirmed these results (obtained by mixing already extracted DNA) by repeating the experiments in the same conditions that would be used for measuring the mutation rate. DNA was extracted from one male carrying the $(Thr\text{-}Gly)_{17a}$ allele together with that from 9 males carrying the $(Thr\text{-}Gly)_{20a}$ allele, in the same 1·5 μl tube and in a total volume of 500 μl. One microlitre of this extract was then pooled with several 1 μl volumes from extracts of 10 $(Thr\text{-}Gly)_{20a}$ males. Again we obtained an unambiguous detection of the $(Thr\text{-}Gly)_{17a}$ allele up to a ratio of 1/200, and the same result was obtained in the reciprocal test. For the females, the DNA from a single heterozygous female was extracted together with that from 9 females homozygous for either the $(Thr\text{-}Gly)_{17a}$ or the $(Thr\text{-}Gly)_{20a}$ allele. One microlitre of this extract was then pooled with several 1 μl volumes from extracts of 10 females homozygous for either the $(Thr\text{-}Gly)_{20a}$ or the $(Thr\text{-}Gly)_{17a}$ allele respectively. As before, we obtained unambiguous detection of the 'mutant' allele up to a ratio of 1/100. To be conservative we decided to use the technique at 50% of its power. Therefore we analysed the Thr-Gly length mutation rate in the F_3 by pooling DNA extracted from 100 males and from 50 females. Any putative mutation was validated by repeating the PCR on another sample of DNA from the same batch of flies. The separation of the 17 and 20 alleles on the acrylamide gels provided sufficient resolution to identify alleles whose lengths differed from the parentals by less than $(Thr\text{-}Gly)_3$ and consequently the method was suitable for detecting any *de novo* length mutation.

(*c*) *DNA sequencing*

Single fly DNA was extracted as outlined above. The Thr-Gly region was amplified by PCR (30 cycles of 95 °C 1 min, 65 °C 1 min and 70 °C 1 min) in a 50 μl volume reaction (same conditions described above) in either a Perking Elmer Cetus or a Biometra Trio thermocycler. The 5′ primer 5′-AACTATAACGAGAACCTGCT-3′ (4874–4893) and the 3′ primer 5′-CCGCGCGACTCGCGGTGCTTCTTC-3′ (5365–5388) were used. The amplification product was then purified by gel electrophoresis (1·2% NuSieve low melting point agarose gel in TBE buffer: 0·045 M Tris-borate, 0·001 M EDTA, pH 8·3), phenol-chloroform extraction and ethanol precipitation as in Sambrook *et al*., (1989). DNA was resuspended in 15 μl of water and 6 μl was used for each sequencing reaction. Double-stranded direct DNA sequencing was carried out using the chain terminating inhibitors method (Sanger *et al*., 1977). The 5′ primer 5′-ATACACATGAGCAGTGTGAC-3′ (5066–5085) was used in combination with the 3′ primer 5′-TCCATCTCGTCGTTGTGCTT-3′ (5333–5352) to sequence both DNA strands

(ii) *Linkage analysis*

(*a*) *Natural populations of* D. melanogaster

In this study we examined 234 individuals representing random samples of nine natural populations of *D.*

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

Table 1. *Three-locus haplotypes in the nine natural population studied*

	CO	K	PLAT	CON	TJ	STU	NW	LE	MO	Total
CAC-14-CA	1			1				3	1	6
CAT-17-CT	1									1
CAT-17-CA	2	7	1	6	4	10	1	6	2	39
CAC-17-CA	7	11	4	11	4	1	3	16	14	71
AGT-17-CA	1									1
AGC-17-CA	1									1
AGC-20-CA				1					1	2
AGC-20-TA	11		2	6	1	4	2	4	5	35
CAC-20-TA	2						3			5
CAC-20-CA						5		2	3	10
CAT-20-CA			1	3	13			5	5	27
CGT-20-CT			8							8
AGC-21-TA	1								1	2
CAT-23-CT	4		2	2	4	1	2	5	1	21
AGT-23-CA	4		1							5
Total	35	18	19	30	26	21	11	41	33	234

CO, Cognac (France, Oct.–Nov. 1991); K, Kafr-El-Sheik (Egypt, Oct. 1992); CON, Conselve (Italy, Oct. 1992); PLAT, Platanistassa (Cyprus, Sept. 1992); TJ, Tirgu-Jiu (Romania, Aug. 1993); STU, Studina (Romania, Aug. 1993); NW, North Wootten (England, Oct. 1994); LE, Lecce (Italy, Oct. 1994); MO, Modena (Italy, Sept. 1994).

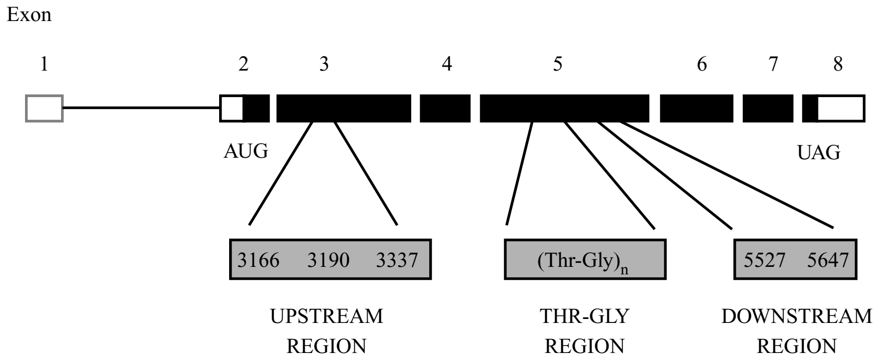


Fig. 1. Position, in the intron/exon map of the *period* gene of *D. melanogaster*, of the three loci analysed in this study. The numbers refer to the Oregon-R sequence (Citri *et al.*, 1987).

melanogaster. Table 1 gives the numbers of individuals analysed for each population and the date of collection. All the individuals under study were males either immediately frozen after capture or derived from isofemale lines established immediately after collection of the sample. Calculating linkage disequilibrium implies that all the individuals come from the same population. However, it was clear from our results (Table 1) that the most frequent associations between loci were generally found in all populations, and therefore for estimating the pairwise disequlibrium values it was reasonable to assume that the individuals were derived from a single population, as in other studies of this type (e.g. Berry & Kreitman, 1993).

(*b*) *Molecular analysis*

Single fly (males) DNA extraction and DNA amplification by PCR were carried out as above. In this study we investigated three sites within *per*, which are referred to as the Thr-Gly region, the upstream (5′) region and the downstream (3′) region (Fig. 1).

The Thr-Gly region was amplified with the 5′ primer 5′-ATACACATGAGCAGTGTGAC-3′ (5066–5085) and the 3′ primer 5′-TCCATCTCG-TCGTTGTGCTT-3′ (5333–5352). The length alleles were identified by electrophoresis through a 3·5% low melting point NuSieve (GTG) agarose gel in TBE buffer. We define the Thr-Gly alleles only in terms of the length of the region and not the interspersion pattern of repetitive units, i.e. the nucleotide sequence.

The upstream (5′) locus is defined by three silent nucleotide polymorphisms at positions 3166 (A/C), 3190 (G/A) and 3337 (C/T). These sites were chosen by examining the published sequences upstream of the Thr-Gly repeat (Citri *et al.*, 1987; Jackson *et al.*, 1986; Kliman & Hey 1993), particularly in the regions that are more variable between *Drosophila* species (Colot *et al.*, 1988). The G/A (3190) and C/T (3337)

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

substitutions determine a restriction polymorphism, respectively for *Bsto*I and *Sin*I. A 656 bp fragment was amplified using the 5′ primer 5′-CGACATGATCATCAAGCGCA-3′ (3079–3098) and the 3′ primer 5′-GGTGCACAAAGTCGATGAAG-3′ (3715–3734) and the composition of those sites was examined by restriction fragment length polymorphism (RFLP) analysis. Similarly a 5′ primer 5′-GGGGCGGGGCAGGCGC**G**GAC-3′ (3146–3165), carrying a mismatch at position 3162 (highlighted in bold) was used in combination with the 3′ primer 5′-GCGGCGAGGGAATCGGAAAG-3′ (3403–3422) to amplify a 277 bp region and to create a *Sin*I site around the nucleotide polymorphism at position 3166.

Using the same criteria as for the 5′ sites, we found two silent nucleotide polymorphisms at positions 5527 (C/T) and 5647 (A/T), downstream (3′) of the repeat which corresponded to restriction polymorphisms for *Bsto*I and *Alu*I, respectively. The RFLP analysis was carried out on a 815 bp fragment amplified by using the 5′ primer 5′-AAGCACAACGACGAGATGGA-3′ (5333–5352) and the 3′ primer 5′-GGCATCGGCTGGTCCATCAT-3′ (6128–6147).

PCR amplifications using all the combination of primers above were carried out for 30 cycles (95 °C for 1 min, 65 °C for 1 min and 70 °C for 1 min).

Digestions with the restriction endonucleases *Alu*I (site 5647), *Bsto*I (sites 3190 and 5527) and *Sin*I (sites 3166 and 3337) (Promega) were set up directly on amplified DNA. Digested DNAs were electrophoresed through a 3·5% low melting point NuSieve (GTG) agarose gel in TBE buffer.

3. Results

(i) *Direct estimation of the Thr-Gly region length mutation rate*

We examined 7126 males and 7829 females for the (Thr-Gly)$_{17a}$ allele and 5400 females and 5443 males for the (Thr-Gly)$_{20a}$ allele. In total we analysed 39027 X chromosomes, taking into account both sexes and both alleles. All putative length mutations, on repeating the PCR amplification on further DNA samples from the 'mutants', turned out to be due to PCR artefacts, and therefore no *de novo* length mutations were found. The upper 95% confidence limit for the mutation rate can be calculated assuming the mutations follow a Poisson distribution where λ is the expected number of mutations (Sokal & Rohlf, 1981) and $e^{-\lambda}$ is the probability of observing exactly zero mutations. Solving for λ such that the probability of zero mutations is 0·05 gives $\lambda = 2·996$, from which the upper 95% confidence limit is $2·996/39027 = 7·7 \times 10^{-5}$.

Table 2. *Frequencies, linkage disequilibrium* (*D*) *and normalized disequilibrium* (*D′*) *values for the combinations* (5′)–(*Thr-Gly*)$_n$, (*Thr-Gly*)$_n$–(3′) *and* (5′)–(3′)

Haplotype	*n*	Freq	*D*	*D′*
AGC-17	1	0·0043	−0·0783	−0·9482
AGC-20	37	0·1581	0·0946	0·8806
AGC-21	2	0·0085	0·0071	1
AGT-17	1	0·0043	−0·0081	−0·6549
AGT-23	5	0·0214	0·0185	0·8125
CAC-14	6	0·0256	0·0156	1
CAC-17	71	0·3034	0·1136	0·5586
CAC-20	15	0·0641	−0·0821	−0·5615
CAT-17	40	0·1709	−0·0107	−0·0587
CAT-20	27	0·1154	−0·0244	−0·1748
CAT-23	21	0·0897	0·0480	0·6918
CGT-20	8	0·0342	0·0215	1
14-CA	6	0·0256	0·0080	1
17-CA	111	0·4744	0·1421	0·9433
20-CA	39	0·1667	−0·0891	−0·4548
23-CA	5	0·0214	−0·0551	−0·7205
17-CT	2	0·0085	−0·0554	−0·8664
20-CT	8	0·0342	−0·0151	−0·3059
23-CT	21	0·0897	0·0750	0·7783
20-TA	40	0·1709	0·1042	0·9242
21-TA	2	0·0085	0·0070	1
AGC- -TA	37	0·1581	0·1274	0·9086
AGC- -CA	3	0·0128	−0·1048	−0·8910
AGT- -CA	6	0·0256	0·0080	1
CAC- -CA	86	0·3675	0·0970	0·7909
CAC- -CT	1	0·0043	−0·0478	−0·9179
CAC- -TA	5	0·0214	−0·0492	−0·6972
CAT- -CA	66	0·2821	0·0233	0·1986
CAT- -CT	22	0·0940	0·0442	0·5347
CGT- -CT	8	0·0342	0·0297	1

(ii) *Linkage analysis*

(*c*) *Disequilibrium pattern analysis*

Table 2 shows the linkage disequilibrium, *D*, and the normalized disequilibrium values, *D′* (Lewontin, 1964), of the different combinations of upstream (5′) and (Thr-Gly) alleles, of (Thr-Gly) and downstream (3′) alleles, and of (5′) and (3′) alleles. Also, it shows that 12 of 25 (5′)–(Thr-Gly), 9 of 15 (Thr-Gly)–(3′) and 9 of 15 (5′)–(3′) possible haplotypes have been found.

The disequilibrium space is defined by the haplotype disequilibrium *D* parameters on one axis, and by the expected haplotype frequencies in the absence of disequilibrium ($p_i q_j$) on the other axis (Figs. 2–4). Each graph is obtained by considering one allele at one locus in combination with each of the alleles at the other locus. Fig. 2 shows the six graphs obtained for haplotypes from the (5′)–(Thr-Gly) combination of alleles. Haplotypes AGC-20, CAC-17 and CAT-23 show the characteristic patterns for selection (see Section 4). However, these patterns are asymmetric, in that for haplotypes AGC-20 and CAC-17 the pattern is observed only in the (5′) versus all (Thr-Gly) alleles

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

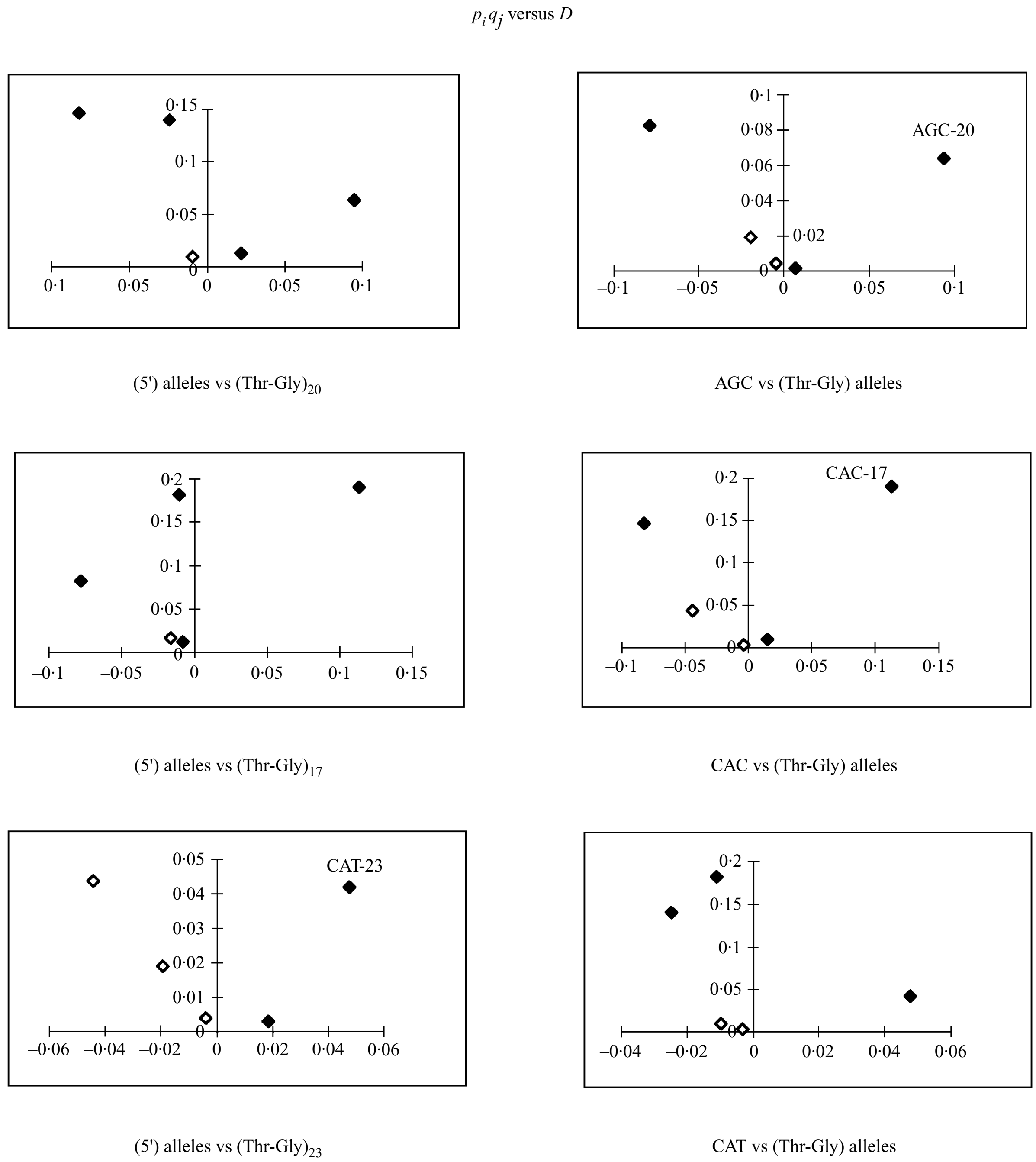
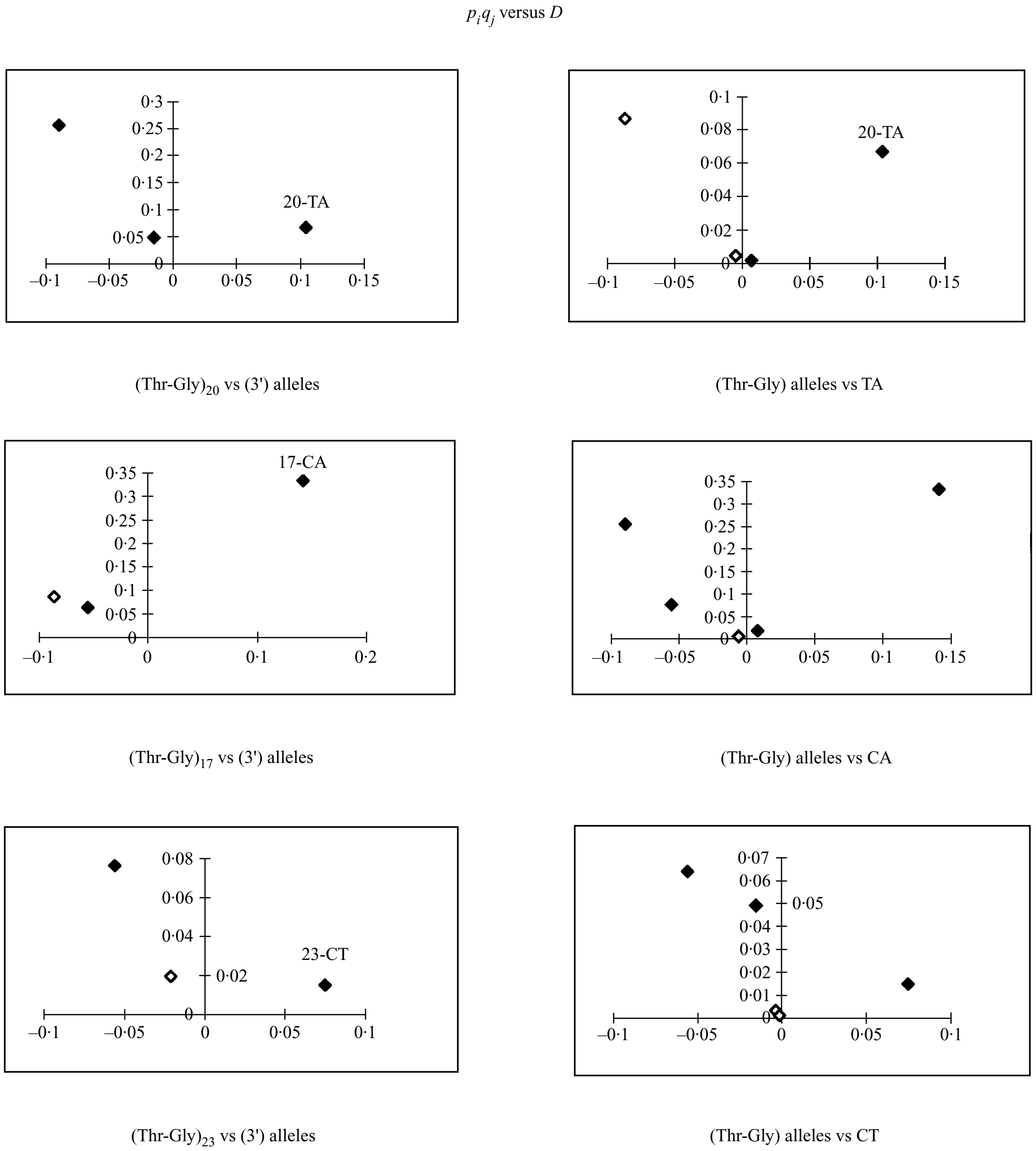


Fig. 2. Disequilibrium patterns for the (5′) versus (Thr-Gly) and (Thr-Gly) versus (5′) comparisons. The linkage disequilibrium measure D (abscissa) and the expected haplotype frequencies in the absence of disequilibrium p_iq_j (ordinate) define the disequilibrium space. Open symbols represent haplotypes with $D' = -1$. Haplotypes AGC-20, CAC-17 and CAT-23, which show evidence of selection, have been highlighted only in the graph combinations showing patterns indicative of selection (see text).

comparison, whereas for haplotype CAT-23 the pattern is observed only in the (Thr-Gly) versus all (5′) alleles graph. The AGC-20 and CAT-23 patterns reveal the strongest cases for selection, with the negative D' estimates clustering closely around a single value, but some variation in the D' values exists, instead, for the CAC-17 graph (see Section 4). Fig. 3 shows the six graphs obtained with haplotypes from the (Thr-Gly)–(3′) combination of alleles. Haplotype 20-TA shows the strongest pattern of selection and is the only one showing patterning in both graph combinations. Haplotypes 17-CA and 23-CT also reveal a pattern of selection but with some variation in the negative D' values, and only in the (Thr-Gly) allele versus all (3′) alleles graphs. Finally, Fig. 4 shows the patterns for the (5′)–(3′) combinations of alleles. It is possible to recognize a selection pattern for AGC- -TA and CAT- -CT haplotypes but only in one graph combination which reflects the (5′) versus (3′) and (3′) versus (5′) alleles respectively.

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press



(Thr-Gly)$_{20}$ vs (3') alleles

(Thr-Gly) alleles vs TA

(Thr-Gly)$_{17}$ vs (3') alleles

(Thr-Gly) alleles vs CA

(Thr-Gly)$_{23}$ vs (3') alleles

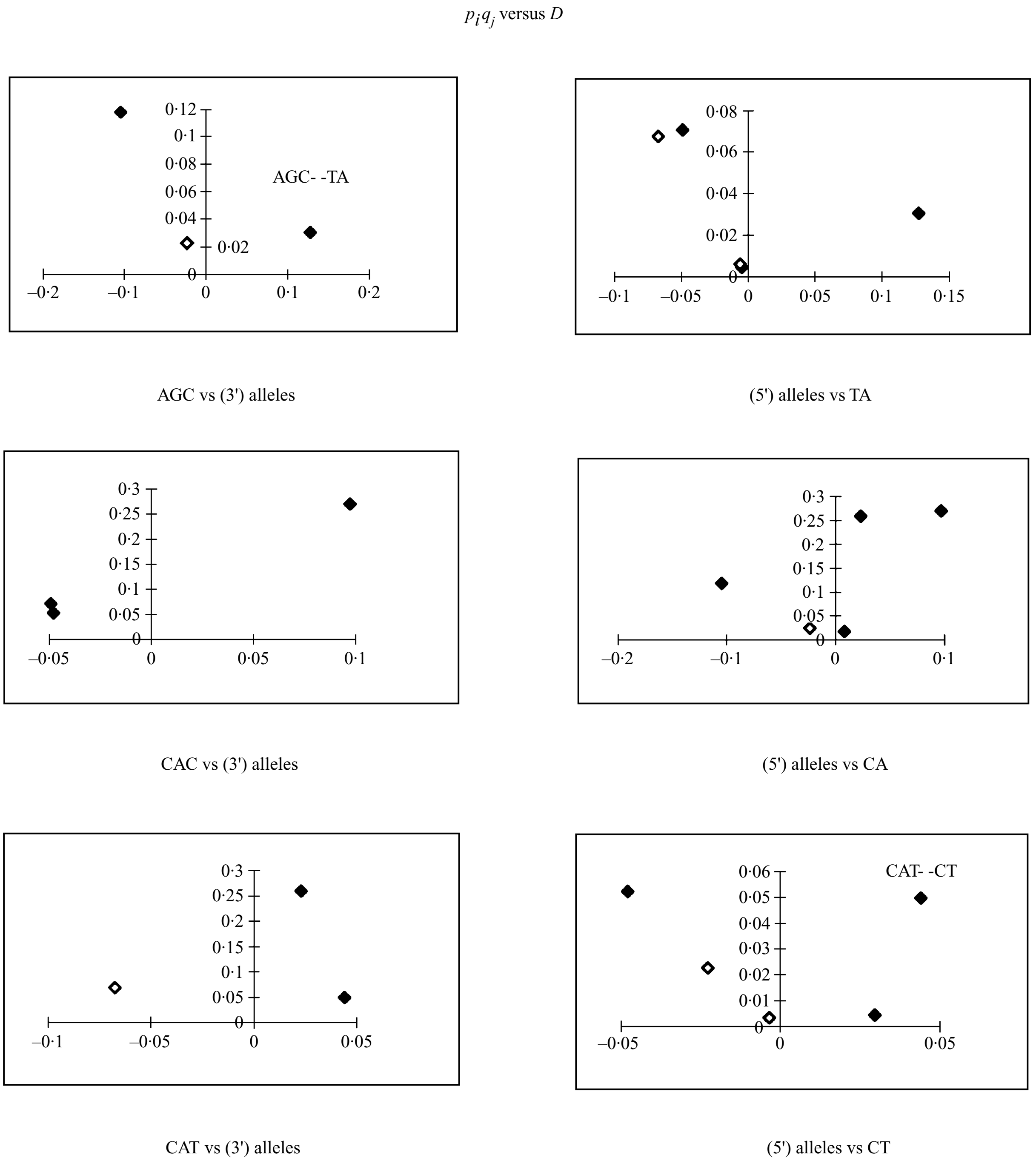
(Thr-Gly) alleles vs CT

Fig. 3. Disequilibrium patterns for the (Thr-Gly) versus (3′) and (3′) versus (Thr-Gly) comparisons. D and p_iq_j represent the abscissa and the ordinate respectively. Open symbols represent haplotypes with $D' = -1$. Haplotypes 20-TA, 17-CA and 23-CT, which show evidence of selection, have been highlighted only in the graph combinations showing patterns indicative of selection (see text).

(*b*) *Three-locus analysis*

The disequilibrium pattern analysis method suggests that selection was or is acting on loci of the AGC-20-TA, CAC-17-CA and CAT-23-CT haplotypes. In an attempt to distinguish which of the three loci might be the focus of selection, we applied the method of Robinson *et al.* (1991*b*) based on the comparison between the pairwise disequilibria existing in a two (less constrained) locus system and a three (more constrained) locus system. Table 3 shows the allele frequencies (p) for each locus, the two-locus normalized pairwise disequilibrium value D', the three-locus normalized pairwise disequilibrium value D'' and their absolute difference $\delta = |D'| - |D''|$ for the three aforementioned haplotypes. For AGC-20-TA all the three δ values are negative. For CAC-17-CA and CAT-23-CT, the δ values for the (5′)–(Thr-Gly) and the (5′)–(3′) comparisons are negative whereas for the (Thr-Gly)–(3′) comparisons they are zero. The pairwise disequilibrium patterns for all three haplotypes is in agreement with the hypothesis of selection

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press



AGC vs (3') alleles

(5') alleles vs TA

CAC vs (3') alleles

(5') alleles vs CA

CAT vs (3') alleles

(5') alleles vs CT

Fig. 4. Disequilibrium patterns for the (5′) versus (3′) and (3′) versus (5′) comparisons. D and p_iq_j represent the abscissa and the ordinate respectively. Open symbols represent haplotypes with $D' = -1$. Haplotypes AGC- -TA and CAT- -CT, which show evidence of selection, have been highlighted only in the graph combinations showing patterns indicative of selection (see text).

Table 3. *Three-locus analysis*

A–B–C	p_A	p_B	p_C	D'_{AB}	D''_{AB}	δ_{AB}	D'_{AC}	D''_{AC}	δ_{AC}	D'_{BC}	D''_{BC}	δ_{BC}
AGC-20-TA	0·1709	0·3718	0·1795	0·8806	0·9338	−0·0531	0·9086	0·9371	−0·0286	0·9242	0·9553	−0·0311
CAC-17-CA	0·3932	0·4829	0·6880	0·5586	0·6098	−0·0513	0·7909	0·8256	−0·0346	0·9433	0·9433	0
CAT-23-CT	0·3761	0·1111	0·1325	0·6918	0·7289	−0·0371	0·5347	0·6200	−0·0853	0·7783	0·7783	0

The (5′), (Thr–Gly) and (3′) loci are represented as A, B and C respectively. p, allele frequency at each locus; D', two-locus normalized pairwise disequilibrium value; D'', three-locus normalized pairwise disequilibrium value, $\delta = |D'| - |D''|$.

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

acting on one or more loci of the three-locus system. However, the method fails to distinguish which of the three loci is the one under selection (Robinson *et al.*, 1991*b*).

4. Discussion

Disequilibrium pattern analysis can detect selection by examining the pattern of distribution in disequilibrium space of the haplotypes observed at two tightly linked loci (Klitz & Thomson, 1987; Thomson & Klitz, 1987). Although it is not a statistical test, the method predicts, as a result of selective events, specific disequilibrium patterns which are not mimicked by drift (when the normalized disequilibrium values move away from the $D' = -1$ region), or mutation, and are unlikely to be imitated by population admixture. There are two criteria by which to identify those two locus haplotypes that have undergone selection: (1) the presence of only one or a few haplotypes in the positive disequilibrium space; (2) related haplotypes (those sharing an allele with a selectively favoured haplotype) have a negative D value proportional to the frequency of the unshared allele, and a standardized disequilibrium D', clustering around a single D' value. This method has been derived from simple deterministic models but Peixoto & Klaczko (1991) have shown that it is able to give meaningful results even in more 'complex' situations. The seasonal cyclic variation in the inversion frequencies (which is a classic case of balancing selection) of *D. mediopunctata* resembles very closely the latitudinal cline for the Thr-Gly region in *D. melanogaster* (Costa *et al.*, 1992), and therefore the application of this analysis to our data seems very appropriate. Moreover because disequilibrium pattern analysis has been empirically developed with computer simulations and no assumptions have been made on the mechanisms giving rise to new alleles, we can apply it directly to repetitive sequences without knowing whether their evolution follows a stepwise model (each new mutation is dependent on the previous allelic states) or an infinite alleles model (each new mutation is independent of the previous allelic states). To our knowledge this is the first time that a repetitive region has been directly investigated for selection, independent of any particular evolutionary model chosen.

We have applied the disequilibrium pattern analysis to our data and observed that a number of haplotypes appear to show patterns consistent with selection. Some caution is needed however, in that the proportion of $D' = -1$ values (a measure of neutrality conditions) is quite high ($0{\cdot}45 = 25/55$) and no statistical test has been developed to exclude selective neutrality in the analysis of a particular pattern. We were able to identify linkage disequilibrium patterns typical of selection for haplotypes AGC-20, CAC-17, CAT-23, 20-TA, 17-CA, 23-CT, AGC- -TA and CAT- -CT. Haplotypes sharing the $(\text{Thr-Gly})_{20}$ allele were those showing the patterns with the strongest indication of selection. Note that, because the AGC-20 and 20-TA and CAT-23 and 23-CT show patterns typical of selection, the AGC- -TA and CAT- -CT patterning we observe may be a consequence of hitchhiking. In contrast, although CAC-17 and 17-CA show patterning, the same is not true for the CAC- -CA haplotype.

Theoretically, if patterning is exhibited only in one graph combination [i.e. AGC vs (Thr-Gly) alleles] but not in the reciprocal [i.e. $(\text{Thr-Gly})_{20}$ vs (5′) alleles, see Fig. 2], then the allele showing patterning (AGC), when tested against the other locus alleles, is the more recently derived of the two, and usually the less frequent. In other words, the model on which the disequilibrium pattern analysis method is based assumes that patterning asymmetries may be due to a new allele appearing at one locus and creating a new haplotype with an existing allele at the second locus, with the new haplotype increasing in frequency due to direct selection or hitchhiking. Because the older allele would have had more time to be associated with a greater number of alleles at the other locus, this would explain the greater probability for it to present patterns not typical of selection, thereby generating asymmetries.

According to this criterion, in the (5′)–(Thr-Gly) comparisons, AGC would represent a more recent allele than $(\text{Thr-Gly})_{20}$, CAC more recent than $(\text{Thr-Gly})_{17}$ and $(\text{Thr-Gly})_{23}$ more recent than CAT. In the (Thr-Gly)–(3′) comparisons, $(\text{Thr-Gly})_{23}$ would be more recent than CT and $(\text{Thr-Gly})_{17}$ more recent than CA. Finally in the (3′)–(5′) comparisons, AGC would be more recent than TA, and CAT older than CT. For the Thr-Gly-(3′) comparisons and the (5′)-(3′) comparisons we do not have any external reference to use for speculating about the relative ages of the alleles at the two loci of the selected haplotypes. However, we do have this information for the (5′)–(Thr-Gly) comparison. *D. simulans* shares with *D. melanogaster* the (5′) locus AGC allele but not the (Thr-Gly) locus alleles $(\text{Thr-Gly})_{20}$, $(\text{Thr-Gly})_{17}$ and $(\text{Thr-Gly})_{23}$ (Peixoto *et al.*, 1992; Kliman & Hey, 1993; Rosato *et al.*, 1994). Therefore AGC (found in *simulans*) cannot be younger than $(\text{Thr-Gly})_{20}$ (never found in *simulans*). Why does the disequilibrium pattern analysis method give the 'wrong' assessment of the alleles' respective ages? It is because the (5′) and the (Thr-Gly) loci are non-repetitive and repetitive sequences respectively, experiencing different mutational processes, having a different mutation rate, and therefore evolving at a different pace. For these reasons it is not advisable to use disequilibrium pattern analysis to compare the history of the (5′) and (Thr-Gly) alleles or the (Thr-Gly) and (3′) alleles and therefore we did not explore these implications of the method any further.

To investigate which locus in the [(5′)–(Thr-Gly)–(3′)] system is under selection, we applied the

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

method of Robinson *et al.* (1991*b*). The basic idea is that in a two-locus system the pairwise disequilibrium term, *D*, is less constrained than in a three-locus system. In a two-locus system, A–B, the allele frequencies impose a constraint on the minimum and maximum values of the pairwise disequilibrium term D_{AB}. In a three-locus system, A–B–C, the additional pairwise disequilibrium terms D_{BC} and D_{AC} (if nonzero) impose additional constraints on D_{AB} (Robinson *et al.*, 1991*a*). On this basis a new normalized pairwise disequilibrium measure, D''_{AB} (and similarly D''_{AC} and D''_{BC}) can be defined, which takes into account these additional constraints (Robinson *et al.*, 1991*a*). The absolute difference, $\delta = |D'| - |D''|$, between the two-locus normalized pairwise disequilibrium measure D' and the three-locus normalized pairwise disequilibrium measure D'' shows, under selection, three basic patterns. Pattern I is such that δ for unselected loci (δ_{uu}) is positive whereas δ for unselected and selected loci (δ_{us}) is negative so that $\delta_{\mathrm{uu}} \geqslant 0 \geqslant \delta_{\mathrm{us}}$. Pattern II is such that $\delta_{\mathrm{uu}} > \delta_{\mathrm{us}} \geqslant 0$ and finally pattern III is such that $\delta_{\mathrm{uu}} \leqslant 0$ and $\delta_{\mathrm{us}} \leqslant 0$ (Robinson *et al.*, 1991*b*). Pattern I and pattern II (but only when $\delta_{\mathrm{uu}} \gg \delta_{\mathrm{us}}$) are the only two patterns which permit the identification of the locus subject to selection. Our results gave pattern III, and therefore do not indicate which is the selected locus. However, it is interesting to note that this method and the disequilibrium pattern analysis, which is based on different assumptions, both suggest that selection was/is operating on the AGC-20-TA, CAC-17-CA and CAT-23-CT haplotypes. Robinson *et al.*, (1991*b*) suggest that pattern III occurs more frequently when (in the A–B–C system) the recombination distance between A and B differs from the recombination distance between B and C. The fourfold difference in the recombination distance between (5′)–(Thr-Gly) compared with (Thr-Gly)–(3′) may thus predispose towards pattern III. Moreover, they suggest that the magnitudes of the δ values are, in general, correlated with the strength of selection and that directional selection yields higher δ values than balancing selection. This seems in agreement with a balancing selection scenario being responsible for the $(\text{Thr-Gly})_{17}$ and $(\text{Thr-Gly})_{20}$ latitudinal cline across Europe in natural populations of *D. melanogaster* (Costa *et al.*, 1992).

Can we use the linkage data to give us a rough idea about the Thr-Gly length mutation rate? In the (5′)–(Thr-Gly) comparison, if the length mutation rate (μ) was significantly higher than the amount of recombination between the two loci, we would not see any disequilibrium. Given the significant disequilibrium that we find, it seems reasonable to postulate that μ has the same order of magnitude or lower than the amount of recombination between the two loci. Aguadé *et al.* (1989) have estimated the recombination rate for *per* as 2×10^{-5} recombinants/kb. The distance between the 5′ sites and Thr-Gly loci is 2 kb, and so the amount of recombination between them of 4×10^{-5} must also represent the upper limit of the Thr-Gly length mutation rate. This is fairly close to the upper estimate of $7{\cdot}7 \times 10^{-5}$ that we reached through our attempts at a direct measurement of the Thr-Gly length mutation rate. What about the lower limit? The exchange of alleles between haplotypes can be due to either recombination or Thr-Gly mutation, and generally we are not able to distinguish between the two types of events. However, the situation for the (Thr-Gly)–TA haplotypes is different. Rosato *et al.* (1996) demonstrated that the $(\text{Thr-Gly})_{21}$ allele is derived from $(\text{Thr-Gly})_{20}$ by an intra-allelic mutational event, and not by recombination. TA is never in association with any other allele except $(\text{Thr-Gly})_{20}$ and $(\text{Thr-Gly})_{21}$, and this suggests that the mutation rate must be higher than the recombination rate between the two loci, which are 0·5 kb apart. Applying the *per* recombination rate as above, this gives us a lower limit for the mutation rate of the Thr-Gly repeat of $2 \times 10^{-5} \times 0{\cdot}5$, or 10^{-5}.

To identify a mutation rate of about 10^{-5}, our experiment to measure the length mutation rate should have been at least one order of magnitude larger, which was quite impractical. More sophisticated techniques, such as SP-PCR (Jeffreys *et al.*, 1994), are not appropriate for this range of mutation rates, because of the difficulty of demonstrating that the ‘mutant’ amplification band truly corresponds to a mutated allele, and not to a PCR artefact. Our aim of directly measuring the length mutation rate of the Thr-Gly region thus was partially successful, in that it at least defines the upper limit, and enables us to reject the idea that the $(\text{Thr-Gly})_{17}$ and $(\text{Thr-Gly})_{20}$ alleles are ‘flipping’ from one allelic state to the other at very high frequencies, thereby destroying any linkage disequilibrium between the repeat and flanking sites (Costa *et al.*, 1991). The value measured represents an estimate for the mutation rate based on analysis of homozygous female progeny. It is possible that the mutation rate in heterozygous females could be higher. However, analysis of heterozygotes has the disadvantage that mutation from one allele length to the other cannot be detected, whereas in homozygotes our method can potentially identify all *de novo* length variants.

In conclusion, we applied two mathematical methods, based on different assumptions, to the Thr-Gly encoding repeat and nearby loci of the *D. melanogaster period* gene, as a novel line of enquiry for revealing any natural selection in the region. The analyses do not contradict the view, based on several other studies (Costa *et al.*, 1992; Peixoto et al., 1993; Castiglione-Morelli *et al.*, 1995), that the Thr-Gly region itself may be the target for selection.

A. A. P. and E. R. were funded by a NERC project grant to C. P. K. We also wish to acknowledge the CEC for a grant to C. P. K. and R. C., and MURST/British Council for an award promoting Anglo-Italian cooperation. We thank H. Dobbs and A. Gallippi for technical assistance.

https://doi.org/10.1017/S001667239700267X Published online by Cambridge University Press

References

Aguadé, M., Miyashita, N. & Langley, C. H. (1989). Reduced variation in the yellow-achaete-scute region in natural populations of *Drosophila melanogaster*. *Genetics* **122**, 607–615.

Aronson, B. D., Johnson, K. A., Loros, J. J. & Dunlap, J. C. (1994). Negative feedback defines a circadian clock: autoregulation of the clock gene *frequency*. *Science* **263**, 1578–1584.

Berry, A. & Kreitman, M. (1993). Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. *Genetics* **134**, 869–893.

Castiglione-Morelli, M. A., Guantieri, V., Villani, V., Kyriacou, C. P., Costa, R. & Tamburro, A. M. (1995). Conformational study of the Thr-Gly repeat in the *Drosophila* clock protein Period. *Proceeding of the Royal Society of London, Series B* **260**, 155–163.

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

Colot, H. V., Hall, J. C. & Rosbah, M. (1988). Interspecific comparison of the *period* gene of *Drosophila* reveals large blocks of non-conserved coding DNA. *EMBO Journal* **7**, 3929–3937.

Costa, R., Peixoto, A. A., Thackeray, J. R., Dalgleish, R. & Kyriacou, C. P. (1991). Length polymorphism in the Threonine-Glycine encoding repeat region of the *period* gene in *Drosophila*. *Journal of Molecular Evolution* **32**, 238–246.

Costa, R., Peixoto, A., Barbujani, G. & Kyriacou, C. P. (1992). A latitudinal cline in a *Drosophila* clock gene. *Proceeding of the Royal Society of London, Series B* **250**, 43–49.

Dover, G. A. (1987). DNA turnover and the molecular clock. *Journal of Molecular Evolution* **26**, 47–58.

Gloor, G. & Engels, W. (1990). Single fly DNA preps for PCR. *Drosophila Information Newsletter* **1** (January).

Hall, J. C. (1995). Tripping along the trail to the molecular mechanisms of biological clocks. *Trends in Neurosciences* **18**, 230–240.

Harding, R. M., Boyce, A. J. & Clegg, J. B. (1992). The evolution of tandemly repeptitive DNA: recombination rules. *Genetics* **132**, 847–859.

Jackson, F. R., Bargiello, T. A., Yun, S-H. & Young, M. W. (1986). Product of *per* of *Drosophila* shares homology with proteoglycans. *Nature*, **320**, 185–188.

Jeffreys, A. J., Wilson, V. & Thein S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**, 67–73.

Jeffreys, A. J., 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.

Jeffreys, A. J., Tamaki, K., MacLeod, A., Monckton, G., Neil, D. L. & Armour, J. A. L. (1994). Complex gene conversion events in germline mutation at human minisatellies. *Nature Genetics* **6**, 136–145.

Kliman, R. M. and Hey, J. (1993). DNA sequence variation at the *period* locus within and among species of the *Drosophila melanogaster* complex. *Genetics* **133**, 375–387.

Klitz, W. & Thomson, G. (1987). Disequilibrium pattern analysis. II. Application to Danish *HLA A* and *B* locus data. *Genetics* **116**, 633–643.

Kyriacou, C. P. (1994). Clock research *per*-ring along: it's about time! *Trends in Genetics* **10**, 69–71.

Lewontin, R. C. (1964). The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67.

Matsui, M., Mitsui, Y. & Ishida, N. (1993). Circadian regulation of *per* repeat messenger RNA in the suprachiasmatic nucleus of rat-brain. *Neuroscience Letters* **163**, 189–192.

McClung, C. R. B., Fox, B. A. & Dunlap, J. C. (1989). The *Neurospora* clock gene *frequency* shares a sequence element with the *Drosophila* clock gene *period. Nature* **339**, 558–562.

Nielsen J., Peixoto, A. A., Piccin, A., Costa, R., Kyriacou, C. P. & Chalmers, D. (1994). Big flies, small repeats: the 'Thr-Gly' region of the *period* gene in Diptera. Molecular Biology and Evolution **11**, 839–853.

Peixoto, A. A. & Klaczko, L. B. (1991). Linkage disequilibrium analysis of chromosomal inversion polymorphisms of *Drosophila*. *Genetics* **129**, 773–777.

Peixoto, A. A., Costa, R., Wheeler, D. A., Hall, J. C. & Kyriacou, C. P. (l992). Evolution of the threonine-glycine repeat region of the *period* gene in the *melanogaster* species subgroup of *Drosophila*. *Journal of Molecular Evolution* **35**, 411–419.

Peixoto, A., Campesan, S., Costa, R. & Kyriacou, C. P. (1993). Molecular evolution of a repetitive region within the *per* gene of *Drosophila*. *Molecular Biology and Evolution* **10**, 127–139.

Reppert, S. M., Tsai, T., Roca, A. L. & Sauman, I. (1994). Cloning of a structural and functional homolog of the circadian clock gene period, from the Giant Silkmoth *Antheraea pernyi*. *Neuron* **13**, 1167–1176.

Robinson, W. P., Asmussen, A. A. & Thomson, G. (1991*a*). Three-locus systems impose additional constraints on pairwise disequilibria. *Genetics* **129**, 925–930.

Robinson, W. P., Cambon-Thomsen, A., Borot, N., Klitz, W. & Thomson, G. (1991*b*). Selection, hitchhiking and disequilibrium analysis at three linked loci with application to *HLA* data. *Genetics* **129**, 931–948.

Rosato, E., Peixoto, A. A., Barbujani, G., Costa, R. & Kyriacou, C. P. (1994). Molecular evolution of the *period* gene in *Drosophila simulans*. *Genetics* **138**, 693–707.

Rosato, E., Gallippi, A., Peixoto, A. A., Kyriacou, C. P. & Costa, R. (1996). Mutational mechanisms, phylogeny, and evolution of a repetitive region within a clock gene of *Drosophila melanogaster*. *Journal of Molecular Evolution* **42**, 392–408.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). *Molecular Cloning: A Laboratory manual*, 2nd edn. Cold Spring Harbor, New York: Cold Spring Harbor Laboratory Press.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain terminating inhibitors. *Proceedings of the National Academy of Sciences of the USA* **74**, 5463–5467.

Sokal, R. R. & Rohlf, F. J. (1981). *Biometry*. San Francisco: W. H. Freeman.

Thomson, G. & Klitz, W. (1987). Disequilibrium pattern analysis. I. Theory. *Genetics* **116**, 623–632.

Watterson, G. A. (1975). On the number of segregating sites in genetic models without recombination. *Theoretical Population Biology* **7**, 256–276.