Insect Molecular Biology (2001) **10**(6), 573–578

Structure and evolution of the mitochondrial DNA complete control region in the *Drosophila subobscura* subgroup

A. Brehm[1]*, D. J. Harris[2], M. Hernández[3], V. M. Cabrera[3], J. M. Larruga[3], F. M. Pinto[3] and A. M. González[3]

[1]*Centro de Ciências Biologicas e Geologicas, University of Madeira, Penteada, Portugal;* [2]*Centro de Estudos de Ciência Animal (CECA), ICETA-U.P., Campus Agrario de Vairão, Vila do Conde, Portugal;* [3]*Department of Genetics, University of La Laguna, Tenerife, Spain*

Abstract

The complete A + T-rich region of mitochondrial DNA (mtDNA) has been cloned and sequenced in the species of the *Drosophila subobscura* subgroup *D. subobscura*, *D. madeirensis* and *D. guanche*. Comparative analysis of these sequences with others already published has identified new sequence motifs that are conserved in *Drosophila* and other insects. A putative bi-directional promoter and a stop signal are proposed to be involved in the primary mtDNA strand replication of *Drosophila*. This region strongly resolves relationships of the species included in a phylogenetic analysis, both for closely related species and also at deeper phylogenetic levels when only the left and central domains are taken into account.

Keywords: *Drosophila*, control region, promoter, phylogeny.

Introduction

The Control Region (CR) is the only large non-coding region in the mitochondrial DNA (mtDNA). This region has been the object of numerous functional studies in vertebrates, which have identified the transcription initiation sites for each strand (Chang & Clayton, 1984), and the main origin of replication (Clayton, 1982). In invertebrates it has been less studied, although in the crustacean *Artemia franciscana* a main and several potential transcription initiation sites have been located in its non-coding region (Carrodeguas & Vallejo, 1997). In *Drosophila*, in which the CR is also known as the A + T-rich region, the only regulatory feature unequivocally localized is the unidirectional origin of replication that was mapped by electron microscopy in several species (Goddard & Wolstenholme, 1980). In addition, a secondary stem and loop structure has been proposed as the initiation site of the second strand synthesis (Clary & Wolstenholme, 1987) by analogy with mammals (Martens & Clayton, 1979). Comparative studies in vertebrates and invertebrates have subdivided the CR into three domains with different levels of variability. A left domain exists adjacent to the $tRNA^{Ile}$ gene in *Drosophila*, the central domain is the most conserved in all taxa, and the third domain, usually highly variable, is 5′ upstream of the small rRNA gene in *Drosophila*. In mammals, Doda *et al.* (1981) have described termination-associated sequences of the H-strand replication in the left domain. No function has yet been assigned to the left domain in *Drosophila*. The central subregion shows extended nucleotide sequence similarity through several taxonomic levels in vertebrates and in *Drosophila*. Finally, in spite of the high variability of the right domain, it is there where the most conserved sequence blocks in vertebrates are found, which are implicated in several aspects of the mtDNA replication. Although a general model of the CR functions has been proposed for vertebrates (Shadel & Clayton, 1997), it seems that it is not extensible to invertebrates. In *Drosophila*, the complete CR has only been sequenced in *D. virilis* (Clary & Wolstenholme, 1987), in some species that have much shorter A + T-rich regions (Clary & Wolstenholme, 1985; Monnerot *et al.*, 1990), and in *D. melanogaster* which has a 4.6 kb long A + T-rich region (Lewis *et al.*, 1994).

Here, we describe the complete CR sequence of the related species *D. subobscura*, *D. madeirensis* and *D. guanche* that compose the *subobscura* subgroup of the *obscura* group. We compare the phylogenetic estimate of relationships derived from the CR with that previously suggested based on other mitochondrial and nuclear DNA sequences. We also determine structural patterns in the CR, both within the *subobscura* group and with other *Drosophila*.

Received 20 April 2001; accepted after revision 16 July 2001. *Correspondence: Fax: +35 12917 05399; e-mail: brehm@uma.pt

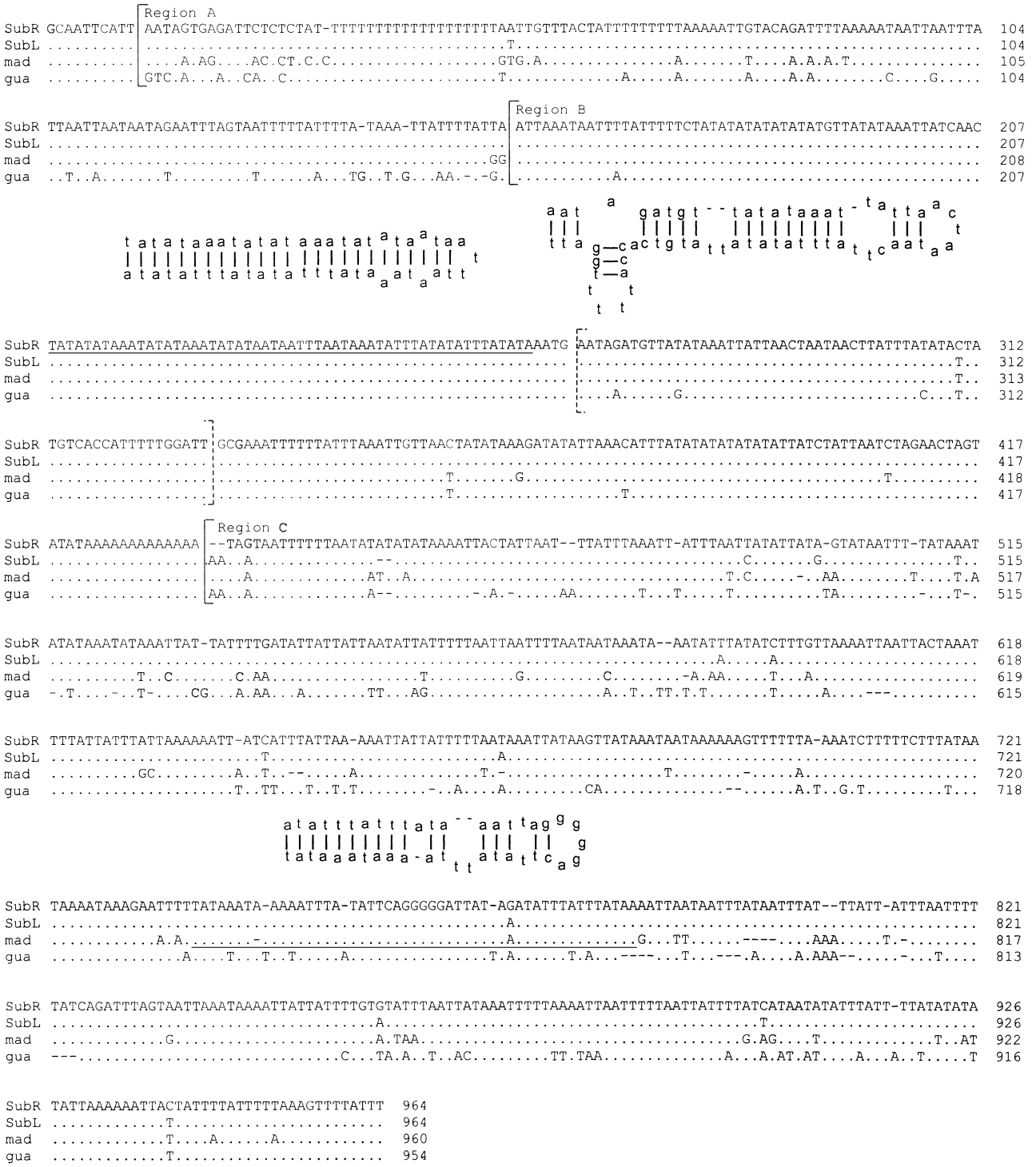

Figure 1. Alignment of the nucleotide sequences analysed. A *dot* indicates a nucleotide that is the same as that in *D. subobscura* from Raíces. A *dash* indicates a nucleotide that is absent. A *letter* indicates a substitution. *Brackets* indicate the limits of the three putative regions. *Dashed brackets* indicate the limits of the sequence that corresponds to the stem-loop structure. The palindromic and putative stop sequences are underlined. The most stable secondary structures are indicated above the stem-loop, the palindromic and the putative stop sequences.

Results

Structure of the A + T-rich region

Length of the CR in the *subobscura* subgroup ranges from 930 nt in *D. guanche* to 940 nt in *D. subobscura*. The overall A + T content for the entire CR in the *D. obscura* group is 93%, which is typical of this region in *Drosophila*. Alignment of these sequences gives a total length of 963 nt (Fig. 1). A screening of nucleotide variation along the sequence using a 20 nt window with the MEGA program (Kumar *et al*., 1993) allows the identification of three regions of different variability. The first, A (bases 11–157, Fig. 1) is immediately adjacent to the $tRNA^{Ile}$ gene and has the highest nucleotide diversity (0.230 ± 0.058); the central, B, bases 158–437, is the most conserved (0.021 ± 0.007); and the third, C, bases 439–973, upstream from the 12S rRNA gene, is more variable (0.150 ± 0.041) than B but less than A. The rate of nucleotide substitution for the regulatory region as a total in the *subobscura* subgroup (0.116 ± 0.03) is four times faster than that of mitochondrial coding regions (0.029 ± 0.006) for the same species (Barrio *et al*., 1994) but the conserved B domain (0.021 ± 0.007) is at the same level.

Phylogenetic analysis

For phylogenetic comparisons we included from GenBank sequences from *D. obscura*, *D. ambigua*, *D. yakuba*, *D. teissieri*, *D. melanogaster*, *D. simulans*, *D. mauritania* and *D. virilis*. It was not possible to align region C, which varies in length from approximately 550 bp in the *suboscura* group to 3.5 kb in *D. melanogaster*. Furthermore this region is not reported for some of the taxa. However 434 bp (regions A and B) were included in the analysis for the twelve taxa. Using Modeltest (Posada & Crandall, 1998) under the Akaike information criteria we concluded that the TrN model (equal transversion ratios, transition ratios A/G 4.1, C/T 2.3 relative to transversions), with a gamma distributed rate heterogeneity model (four rate categories, $G = 0.85$; Yang, 1994), was the most appropriate model of evolution for this data. A ten replicate heuristic search with this model produced a single tree with a maximum likelihood score of –1757 (Fig. 2).

Discussion

Putative function of conserved elements in the A + T-rich region

Previously the only conserved element known in region A of *Drosophila* was a thymidylate stretch with a mean of 21 ± 2 nts near the $tRNA^{Ile}$ gene. This run of Ts is also present in the *subobscura* subgroup species with similar number and position (Fig. 1). Due to its localization, this stretch might be expected to form part of a promoter for transcription in the direction from the A + T-rich region to

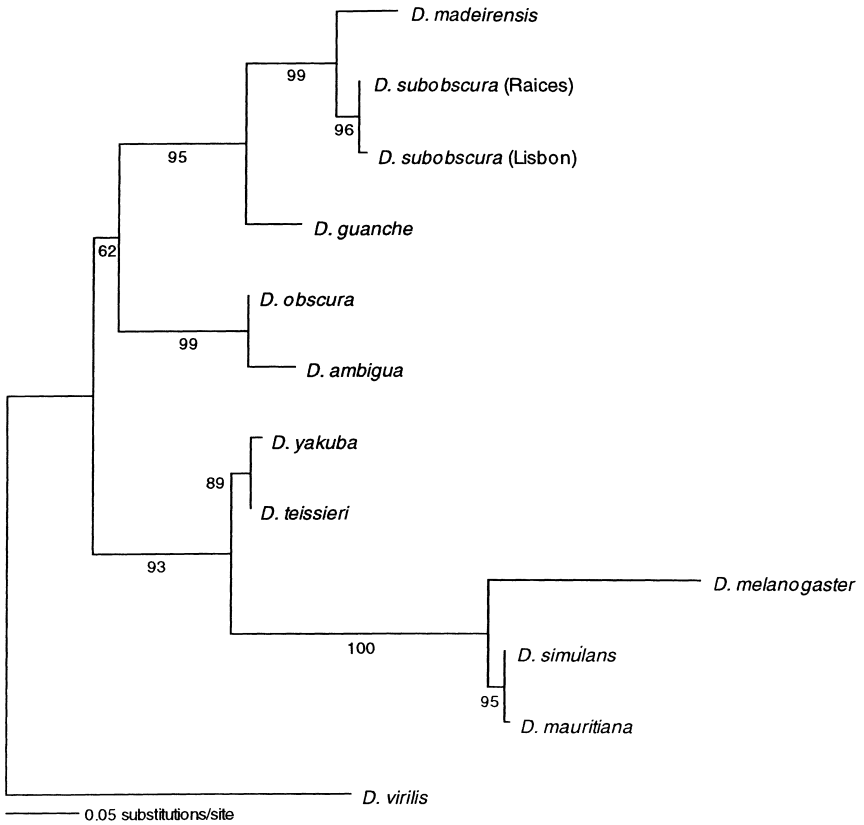

Figure 2. Single tree derived from a maximum likelihood, ten-replicate heuristic search with random sequence additions using the TrN + G model of evolution. Numbers below nodes indicate bootstrap support (1000 replicates).

the $tRNA^{Ile}$ gene (Clary & Wolstenholme, 1987). The central domain B contains the largest and most conserved region among all the *Drosophila* species studied (Fig. 1). Clary & Wolstenholme (1987) identified in the centre of this region a possible stem-and-loop structure, which was suggested as a possible secondary origin of mtDNA replication in *Drosophila*. We detected a highly conserved palindromic structure localized within the B region at the left of this putative stem-loop motif, which is not only present in all the *Drosophila* species studied (underlined in Fig. 1) but also in the same relative position in other insects such as *Anopheles* and *Locusta*. It is one of the most favourable secondary structures detected in the region and its mean free energy is higher (DG = –14) than those described previously. The right border of the B region ends in a very conserved run of As, which in the *subobscura* species has a mean length of 14 ± 1 nt, which is within the range of all the *Drosophila* species examined (15 ± 4 nts). Due to its position adjacent to the innermost conserved type II repeat in *D. melanogaster*, a role in replication has been suggested for this A-stretch (Lewis *et al.*, 1994). The C region has little homology even within the *Drosophila* species studied. Alignment of this region is only possible among closely related species such as those of the *subobscura* subgroup in this paper (Fig. 1) or among those of the *melanogaster* subgroup (Inohira *et al.*, 1997). No conserved sequence motifs have been described for this region. However, we have detected in it a well conserved block that is not only present in the *subobscura* subgroup species and in all the *Drosophila* sequenced for this region, but in other distantly related insects. The motif is a stretch of four or five Gs. In addition, although the primary sequences of the flanking regions are different in distantly related species, a stem-loop-structure with the Gs displayed in the loop is possible in all of them (underlined in Fig. 1). In the fast evolving region C, the presence of the G stretch stands out, not only for its maintenance across unrelated taxa but also for the oddity that such a motif exists in a region that is 95% A + T. From the primary sequence, secondary structure and position 3′ downstream of the primary origin of replication, this G motif might be associated with the stop of the primary strand replication in *Drosophila* and other invertebrates.

Mitochondrial DNA promoter sequences are generally not well conserved except for their A + T rich composition (Jaehning, 1993). We performed a search for possible promoters using the NNPP program (Reese *et al.*, 1996), concentrating on putative promoters that are localized in similar positions and orientation in different species. A pair of symmetric promoters, one in each strand, appear localized in region B in the conserved block upstream of the stem-loop motif, with the consensus sequence AYRTTATA. This sequence also appears in phylogenetically distinct organisms such as yeast (Christianson & Rabinowitz, 1983) and chickens (L'Abbé *et al.*, 1991).

Phylogenetic value of the A + T-rich region

Within the *suboscura* group *D. madeirensis* is most closely related to *D. subobscura* (99% bootstrap support) and *D. guanche* is the sister group to this pair (95% support). This is concordant with topologies derived from other mitochondrial and nuclear gene regions (O'Grady, 1999). However, despite the relatively short segment used, bootstrap support for relationships is much higher than obtained with other mitochondrial genes even those for which longer sequences were included (O'Grady, 1999). Therefore while the C region of the control region may not be phylogenetically useful the other parts are very effective at recovering the mtDNA tree. Caution should still be exercised when comparing this to the species tree however, because some of these taxa, for example *D. simulans* and *D. mauritana*, have been shown to have introgressed mtDNA (Ballard, 2000). Clearly in this case the gene tree will not necessary correspond to the species tree.

Experimental procedures

Laboratory methods

Two specimens of each species were sequenced for the entire A + T-rich region. For the widespread *D. subobscura* species one strain (subR) was from Tenerife (Canary Islands) and the other (subL) from Lisbon (Portugal). For the endemic species *D. guanche* (gua) and *D. madeirensis* (mad) the specimens were from the Tenerife island and Madeira island, respectively. Total DNA was extracted from individual adult flies by the CTAB method (Towner, 1991) with minor modifications. The entire A + T-rich region was amplified in two overlapping fragments: Fragment I, closest to the $tRNA^{Ile}$ (450–451 bp), using primers R1 (5′-CCTATCAAggTAACCCTTTTTATCAg-3′) and R2 (5′-gTgTATACTAAgTTCTAAATTAATAg-3′ – Monforte *et al.*, 1993). Fragment II, closest to the small rRNA was amplified using primer R10 (5′-TgTAACCATTTTTggATTgCgA-3′) designed from our sequences, and R8 (5′-AACTAATAACAAATTTTTAAgCCA-3′, Clary & Wolstenholme, 1987). PCR amplifications were performed in a mix reaction containing 0.1 mM of 7-deaza dGTP and dCTP, and 0.3 mM of dATP and dTTP. In the cycle profile reduced extension temperatures are typically used to amplify the A + T-rich region (Xin-Zhuan *et al.*, 1996) –20 s at 90 °C, 1 min at 48 °C and 1 min at 68 °C for thirty-five cycles. In order to read accurately the sequences after the stretch of Ts, found at the beginning of fragment I, we proceeded to clone these fragments in the plasmid pBS II KS (+/–). In the presence of dTTPs alone, Taq DNA polymerase was used to add a single T to the blunt ends of the EcoRV digested vector in order to clone efficiently the PCR products (Ausubel *et al.*, 1995). The ligation was performed with the Ready to Go T4 ligase kit (Pharmacia) and the ligate was used to transform the XL1-blue *E. coli* strain (Stratagene). Nucleotide sequences were obtained by the double-stranded dideoxy terminator procedure with the Promega fmol DNA sequencing system, using primers isotopically (γ-^{32}P) ATP end labelled. Because of the high level of A and T in the sequences, modified nucleotide mix was used, containing 30 mM of dATP and dTTP, and 3 mM of dCTP and 7-deaza dGTP. The nucleotide mix d/ddG was 9 mM ddGTP, d/ddA was 900 mM ddATP,

d/ddT was 1500 mM ddTTP and d/ddC was 45 mM ddCTP. The sequencing reactions were performed using the cycle profile: 15 s at 92 °C, 15 s at 52–55 °C and 15 s at 65 °C for forty-five cycles. In addition to the PCR amplification primers, R11 (5′-ATATAgT TgA-TAAT T TATATAACA-3′) and R7 (5′-AATAAATAT T TATAATCCCCCT-gAAT-3′) were used to sequence completely fragments I and II, respectively. Clones were sequenced with the M13 universal primers.

Structural and phylogenetic sequence analysis

The new nucleotide sequences reported here are available on GenBank, accession numbers AJ132899–AJ132902. Previously published sequences from other *Drosophila* for the same region were imported from GenBank. Multiple sequence alignments were performed with CLUSTAL X (Thompson *et al*., 1997) and later maximized for sequence similarity by visual inspection. The MFOLD program (Mathews *et al*., 1999) was employed to identify potential secondary structures. Putative promoters were determined using the program NNPP (Reese *et al*., 1996). Sequences were phylogenetically analysed using PAUP* (Swofford, 2001). When estimating phylogenetic relationships among sequences, one assumes a model of evolution regardless of the optimality criteria employed. We used the approach suggested by Huelsenbeck & Crandall (1997) to test fifty-six alternative models of evolution, employing PAUP* and Modeltest (Posada & Crandall, 1998), outlined in detail in Harris & Crandall, 2000). Once a model of evolution was chosen, it was used to estimate a tree using maximum likelihood (Felsenstein, 1981). Confidence in resulting nodes was assessed using the bootstrap technique (Felsenstein, 1985) with 1000 replicates.

Acknowledgements

This work was supported by grant Stride: STRDB/C/BIO/381/92 from Fundação para a Ciência e Tecnologia, Lisbon, to A.B.

References

Ausubel, F.M., Brent, R., Kingston, R.E., Moore, O.D., Seidman, J.G., Smith, J.A. and Struhl, K. (1995) *Current Protocols In Molecular Biology*, Vol. 2, Unit 15.7. John Wiley and Sons Inc, Massachusetts.

Ballard, J.L. (2000) When one is not enough: introgression of mitochondrial DNA in *Drosophila*. *Mol Biol Evol* **17**: 1126–1130.

Barrio, E., Latorre, A. and Moya, A. (1994) Phylogeny of the *Drosophila obscura* species group deduced from mitochondrial DNA sequences. *J Mol Evol* **39**: 478–488.

Carrodeguas, J.A. and Vallejo, C.G. (1997) Mitochondrial transcription initiation in the crustacean *Artemia Franciscana*. *Eur J Biochem* **250**: 514–523.

Chang, D.D. and Clayton, D.A. (1984) Precise identification of individual promoters for transcription of each strand of human mitochondrial DNA. *Cell* **36**: 635–643.

Christianson, T. and Rabinowitz, M. (1983) Identification of multiple transcriptional initiation sites on the yeast mitochondrial genome by *in vitro* capping with guanylytransferase. *J Biol Chem* **258**: 14025–14033.

Clary, D.O. and Wolstenholme, D.R. (1985) The mitochondrial DNA molecule of *Drosophila yakuba*: nucleotide sequence, gene organization and genetic code. *J Mol Evol* **22**: 252–271.

Clary, D.O. and Wolstenholme, D.R. (1987) *Drosophila* mitochondrial DNA: conserved sequences in the A + T-rich region and supporting evidence for a secondary structure model of the small RNA. *J Mol Evol* **25**: 116–125.

Clayton, D.A. (1982) Replication of animal mitochondrial DNA. *Cell* **28**: 693–705.

Doda, J.N., Wright, C.T. and Clayton, D.A. (1981) Elongation of displacement-loopstrands in human and mouse mitochondrial DNA is arrested near specific template sequences. *Proc Natl Acad Sci USA* **78**: 6116–6120.

Felsenstein, J. (1981) Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol* **17**: 368–376.

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783–791.

Goddard, J.M. and Wolstenholme, D.R. (1980) Origin and direction of replication in mitochondrial DNA molecules from the genus *Drosophila*. *Nucl Acids Res* **25**: 741–757.

Harris, D.J. and Crandall, K.A. (2000) Intragenomic variation within ITS1 and ITS2 of freshwater crayfishes (Decapoda: Cambaridae): implications for phylogenetic and microsatellite studies. *Mol Biol Evol* **17**: 284–291.

Hulsenbeck, J.P. and Crandall, K.A. (1997) Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann Rev Ecol Syst* **28**: 437–466.

Inohira, K., Hara, T. and Matsuura, E.T. (1997) Nucleotide sequence divergence in the A + T-rich region of mitochondrial DNA in *Drosophila simulans* and *Drosophila mauritiana*. *Mol Biol Evol* **14**: 814–822.

Jaehning, J. (1993) Mitochondrial transcription: is a pattern emerging? *Mol Microbiol* **8**: 1–4.

Kumar, S., Tamura, K. and Nei, M. (1993) *MEGA: Molecular Evolutionary Genetics Analysis*, Version 1.01. The Pennsylvania State University, USA.

L'Abbé, D.L., Duhaime, J.-F., Lang, B.F. and Morais, R. (1991) The transcription of DNA in chicken mitochondria initiates from one major bidirectional promoter. *J Biol Chem* **266**: 10844–10850.

Lewis, D.L., Farr, C.L., Farquhar, A.L. and Kaguni, L.S. (1994) Sequence, organization, and evolution of the A + T region of *Drosophila melanogaster* mitochondrial DNA. *Mol Biol Evol* **11**: 523–538.

Martens, P.A. and Clayton, D.A. (1979) Mechanism of mitochondrial DNA replication in mouse L-cells: localization and sequence of the light-strand origin of replication. *J Mol Biol* **135**: 327–351.

Mathews, D.H., Sabina, J., Zuker, M. and Turner, D.H. (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J Mol Biol* **288**: 911–940.

Monforte, A., Barrio, E. and Latorre, A. (1993) Characterization of the length polymorphism in the A + T rich region of the *Drosophila obscura* group species. *J Mol Evol* **36**: 214–223.

Monnerot, M., Solignac, M. and Wolstenholme, D.R. (1990) Discrepancy in divergence of the mitochondrial and nuclear genomes of *Drosophila teissieri* and *Drosophila yakuba*. *J Mol Evol* **30**: 500–508.

O'Grady, P.M. (1999) Reevaluation of phylogeny in the *Drosophila obscura* species group based on combined analysis of nucleotide sequences. *Mol Phylogenet Evol* **12**: 124–139.

Posada, D. and Crandall, K.A. (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.

Reese, M.G., Harris, N.L. and Eeckman, F.H. (1996) Large scale sequencing specific neural networks for promoter and splice

site recognition. In *Biocomputing: Proceedings of the 1996 Pacific Symposium* (Hunter, L., Klein, T.E., ed.),. World Scientific Publishing Co., Singapore.

Shadel, G.S. and Clayton, D.A. (1997) Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* **66**: 409–435.

Swofford, D.L. (2001) *PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods) 4.0.B3a*. Sinauer Associates, Sunderland, Massachusetts, USA.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmongin, F. and Higgins, D.G. (1997) The ClustalX windows interface: flexible strategies for multiple sequence alignement aided by quality analysis tools. *Nucl Acids Res* **24**: 4876–4882.

Towner, P. (1991) Purification of DNA. In *Essential Molecular Biology. A Practical Approach*, Vol. I (Brown, T.A. ed.) IRL Press, Oxford.

Xin-Zhuan, S., Yimin, W., Sifri, C.D. and Wellens, T.E. (1996) Reduced extension temperatures required for PCR amplification of extremely A + T-rich DNA. *Nucl Acids Res* **24**: 1574–1575.

Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**: 306–314.