Heredity (2002) 88, 39–45
© 2002 Nature Publishing Group All rights reserved 0018-067X/02 $25.00
www.nature.com/hdy

Nucleotide variation at the *no-on-transient A* gene in *Drosophila littoralis*

S Huttunen[1], S Campesan[2] and A Hoikkala[1]
[1]*Department of Biology, University of Oulu, PO Box 3000, FIN-90014, Oulu, Finland;* [2]*Department of Genetics, University of Leicester, Leicester LE1 7RH, UK*

The *no-on-transient A* (*nonA*) gene encodes a putative RNA-binding protein, and mutations in this gene are known to affect vision, male courtship song and viability in *Drosophila melanogaster*. Here we have sequenced the coding region of the *nonA* gene of *Drosophila littoralis* and compared it with those of *Drosophila virilis* and *D. melanogaster*. All portions of *nonA* appeared to be conserved between *D. littoralis* and *D. virilis*, while the 5′ region of the gene of these two species showed high divergence from that of a more distantly-related species, *D. melanogaster*. The same was true for the glycine repeat regions. No significant deviation from neutrality was observed in the analysis of intraspecific nucleotide variation in 5′ or 3′ region of the *nonA* gene in *D. littoralis* population. Also, comparison of *D. littoralis* sequences with homologous sequence of *D. virilis* suggests that the gene is evolving neutrally in *D. virilis* group. Divergence of the 5′ regions between *D. virilis* group species and *D. melanogaster* could be a result of positive selection, but this finding is obscured by the long divergence time of the species groups.
Heredity (2002) **88,** 39–45. DOI: 10.1038/sj/hdy/6800006

Keywords: *nonA*; *D. littoralis*; *D. virilis*; nucleotide variation

Introduction

The *no-on-transient A* (*nonA*) gene is one of the genes whose mutant alleles have been found to affect behaviour in *Drosophila* flies (see review by Yamamoto *et al*, 1997). The gene was cloned and sequenced in *Drosophila melanogaster* in 1990 simultaneously by two groups: Besser *et al* (1990) and Jones and Rubin (1990). The extensive sequence homology between the central domain of NONA and of the human proteins PSF (PTB-associated splicing factor; Patton *et al*, 1993), HeLa protein $p54^{nrb}$ (nuclear RNA-binding protein; Dong *et al*, 1993) and the product of *nonO* (an octamer-binding protein; Yang *et al*, 1993) strongly suggests conservation of function over a period of divergence between human and fruit flies. Consequently, Dong *et al* (1993) named the phylogenetically conserved central domain of the above-mentioned genes a DBHS domain (*Drosophila* behaviour and human splicing).

nonA-mutations were first isolated in screens for defects in phototactic behaviour in *D. melanogaster* (Hotta and Benzer, 1970), the most severe effects of mutations being optomotor blindness and the lack of transient spikes in the electroretinogram. Kulkarni *et al* (1988) also found some mutants to have disturbances in male courtship song, increasing the amplitude and polycyclicity of the sound pulses as a train of pulses proceeds. Rendahl *et al* (1996) found this mutation, called *dissonance*, to be an allele of *nonA* ($nonA^{diss}$) and showed it to be caused by an amino acid replacement (from Arg to Cys) at site 548 in *nonA* of *D. melanogaster*. Rendahl *et al* (1996) also demonstrated that novel point mutations introduced into the first RNA recognition motif of *nonA* cause disturbances in visual behaviour, male courtship song and the viability of the flies.

The *nonA* gene of *D. virilis* has been sequenced by Campesan *et al* (2001). They showed that the *nonA* genes of *D. virilis* and *D. melanogaster* are diverged at the 5′ region, but that the genes are highly conserved at the central and 3′ region. Species divergence in 5′ region could reflect lowered functional constraint, or it might be a consequence of positive selection affecting this gene region. To find out whether this region shows divergence also between more closely-related species, we have sequenced the coding region of the *nonA* gene of another *D. virilis* group species, *D. littoralis*, and compared it with that of *D. virilis* and *D. melanogaster*. We have also tested the neutrality of sequence evolution in 5′ and 3′ regions of the *nonA* gene by studying intraspecific nucleotide variation on these regions in *D. littoralis* and by comparing *D. littoralis* sequences with the *D. virilis* and *D. melanogaster* sequences. Additional interest in this work is due to the recent finding that *nonA* gene is located at the proximal end of the X chromosome, where the gene or genes causing major differences between the songs of *D. virilis* and *D. littoralis* are known to be located (Päällysaho *et al*, 2001). Thus it is a good candidate gene for causing species differences in male courtship song in the *D. virilis* group.

Correspondence: S Huttunen, Department of Biology, University of Oulu, PO Box 3000, FIN-90014, Oulu, Finland.
E-mail: susanna.huttunen@oulu.fi
Received 5 February 2001; accepted 6 August 2001

Materials and methods

DNA extraction and PCR

Genomic DNA from *D. littoralis* strain 1007 (47°N 8°E, Zürich, 1970) males was extracted using the standard protocol of the Puregene DNA Isolation Kit (Gentra Systems). An approximately 4.3 kb region from exon 2 to exon 5 of the *nonA* locus was PCR-amplified using oligonucleotide primers CACAACTTCAAGCGCAGGCC CAA (forward) and CTAAAAACGACGTCGTCCCCATG (reverse) designed from *D. virilis nonA* sequence (GenBank accession number AJ298998). PCR was performed in 50 μl reactions using two units of DynazymeEXT DNA polymerase (Finnzymes), 0.5 μM of primers (Pharmacia), 360 μM of each dNTP and 2 mM of $MgCl_2$ (Finnzymes). The PCR profile was the following: first denaturation at 94°C for 2 min, then denaturation at 94°C for 30 s, annealing and extension at 68°C for 3 min and 30 s repeated for 30 times, and a final synthesis at 68°C for 10 min. The sequence for the 5′ region of the gene was first analysed through cDNA synthesis and then PCR amplified also from genomic DNA (see below; N-terminal region in exon 1). The PCR products were cloned using the TA Cloning Kit (Invitrogen) according to instructions of the supplier.

Genomic DNA was also extracted from a single male of *D. virilis* laboratory strain 1431 (53°N 1°W, England, 1982) and from six wild-caught *D. littoralis* males from three sites in Finland (Sa: Savonlinna, 62°N, 29°E, Ku: Kuopio, 63°N 27°E and Ou: Oulu, 65°N 25°E). The DNA was PCR-amplified using oligonucleotide primers GTTTCTGTACGGAGCTGGA CGGTTG (forward) and GCCGCCACGATTGCGGTTG (reverse) for a 405 bp sequence in exon 1 (N-terminal), primers ACAACA GATGCACC AAAAGCG (forward) and ACTCCTCATC GGTAATGTCATTGG (reverse) for a 514 bp sequence in exon 2 (N-terminal/Central domain), and primers CGCGAATCTGATAATGAGCG (forward) and CTGTC GCTGGTTATTTGCAC (reverse) for a 495 bp sequence in exons 3 and 4 and intron 3 (C-terminal; see Figure 1). For PCR amplification the reaction volume was 50 μl, consisting of ~100 ng of genomic DNA, 1 μM of each primer (Pharmacia), 200 μM of each dNTP (Finnzymes), 1.5 mM of $MgCl_2$ (Finnzymes), 1× standard reaction buffer (Finnzymes) and 2 U of DyNAzyme DNA polymerase (Finnzymes). The amplification profile was 3 min at 94°C, followed by 35 cycles of 30 s at 94°C, 30 s at 58–64°C, 30 s at 72°C, and finally one cycle of 10 min at 72°C. The annealing temperatures for the studied regions were 64°C, 58°C and 60°C, respectively. The resulting fragment was purified from 1% agarose gel using the method of Glenn and Glenn (1994).

RNA isolation and reverse transcriptase PCR (RT-PCR)

Isolation of the total RNA from adult *D. littoralis* 1007 males was performed according to manufacturer's instructions with TriPure®Isolation Reagent Kit (Boehringer Mannheim). To determine the exact exon-intron boundaries without having to sequence the long intron sequences, various lengths of cDNAs of *nonA* were made using Enhanced Avian RT-PCR Kit (Sigma). The 5′RACE System for Rapid Amplification of cDNA Ends, Version 2.0 (GibcoBRL, Life Technologies) was used to isolate the remaining 5′ region of the *D. littoralis nonA* cDNA with a forward Abridged Anchor primer and reverse primers designed from the *D. virilis* sequence. The amplification was carried out using the following program: 94°C 2 min; 35 cycles of 94°C 30 s, 50°C 30 s, 68°C 1 min, and a final extension at 68°C for 10 min.

Sequencing

A different sequencing strategy was used to obtain the first *D. littoralis* sequence (cloning) from that used for the population survey (direct sequencing of PCR products). All cloned sequences were confirmed from at least 10 independent clones to eliminate errors caused by misincorporation of nucleotides by DNA polymerase. Discrepancies arising were resolved also by direct sequencing of additional PCR products amplified from genomic DNA.

The purified PCR products from the two 5′ and one 3′ regions of the *nonA* gene from one *D. virilis* male (strain 1431) and wild-caught *D. littoralis* males were directly sequenced. Sequencing of both DNA strands was performed with an Applied Biosystems model 377 DNA sequencing system, using the ABI PRISM® BigDye™ Terminator Cycle Sequencing Ready Reaction Kit (Perkin Elmer). Approximately 50 ng of the PCR product was added to 4 μl of dye terminator reaction mix and 6.4 pg of sequencing primer (either forward or reverse used in initial PCR) with 2 μl of ×5 TE buffer in a final reaction volume of 20 μl.

Computer analysis

DNA sequences were edited using Chromas version 1.43 (http://trishul.sci.gu.edu.au/~conor/chromas.html) and the sequences were aligned using ClustalX (Thompson *et al*, 1997). The sequence analysis was made with Dnasis (Hitachi Software Engineering Co), Seqweb program

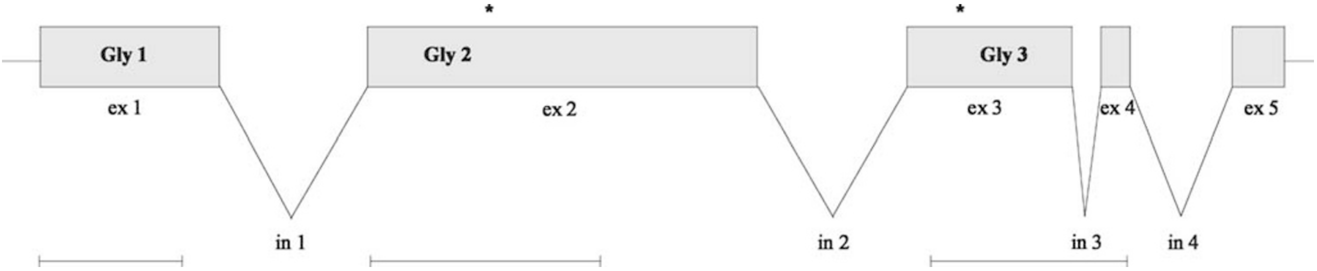

Figure 1 The schematic illustration of the *nonA* gene in *D. littoralis*. Introns are indicated by lines (not in scale) and exons by boxes. Gly1, Gly2 and Gly3 are glycine repeat regions. The black lines below the gene structure show the three regions sequenced from 18 *D. littoralis* males and one *D. virilis* male. The boundaries of amino terminal (N-term), central domain and carboxy terminal region (C-term) are marked by stars.

```
                      N-terminal                                   C-terminal
                   Gly1              Gly2                      Gly3
                                                              1111111111111111111111111111
      1111111122222222223334444444555555666677777778  888888990000000001111111111111111122
    256602333799000112234457135666722567001824577891  157799191233344781166667778888889925
    611219128425367365910842183258458510062191914283  683947597301547363935677890123560132
     s    r    s r s  r s ss   s s    ss      ss       s      ss   ss             i      s
    r rrrs ssss r s ss s s  ssr r ssss  ssssss  srss  * ssssss  rrs  rsssiiiiiiiii iiiiii
ou1 CAGCACCATCCTGGAGTTGTATTGGTTCGTTTCATCTAGGCCGGACGA  CAGACGCCCTAAACCTTATGCTTACGTATTAAAACC
ou2 ................................................  .G............T.............A.......
ou3 .......................C...T.C..................  .G......T.....T.............A.......
ou4 ....................C..C...T.C............A.....  ........TC.........................G
ou5 ..................A.............................  ....................................
ou6 .G.............T................................  ....................................
ku1 .............T.........C...T.C.............A....  ..............T.............A.......
ku2 .......................C...T.C..................  .G..................................
ku3 ...........A...........C...T.C..................  ....................................
ku4 ................................................  ............................A.......
ku5 ..................A....C...T.C..................  .G..........................A.......
ku6 ......T................C........................  .G..........................A.......
sa1 ..................A...AC........................  .............TT.............A.......
sa2 .......................C........................  ....................................
sa3 ..................A....C........................  ....................................
sa4 .......................C...T.C....AT............  ........T...........................
sa5 .G................A.............................  ....................................
sa6 .......................C........................  ..............T.............A.......
vir A.CGTA.CATT.T.TTCC.A-C.CACGTT.CCTT..ACTATTA.CTCG  T.AGTATT..GGT..GGGCCGACCAAAG.GTGTCT.
```

Figure 2 Variability in the analysed *nonA* gene regions in *D. littoralis* and *D. virilis*. Positions of the variable nucleotide sites are presented relative to the sequenced region. Dots represent the same nucleotide as in the first sequence. Abbreviations: r = replacement substitution, s = synonymous substitution, i = intron site, * = triplet incomplete, ou = Oulu, ku = Kuopio, sa = Savonlinna and vir = *D. virilis* (strain 1431). Polymorphic sites among *D. littoralis* sequences are shown on the fourth lane and fixed differences between *D. littoralis* and *D. virilis* on the fifth lane.

version 10.0 of the GCG packet (http://seqweb.csc.fi/gcg-bin/seqweb.cgi) and CodonW, version 1.4.2 (http://bioweb.pasteur.fr/seqanal/interfaces/codonw.html). The estimates for the nucleotide variation were calculated using DnaSP version 3.5 (http://www.bio.ub.es/~julio/DnaSP.html), ProSeq (http://helios.bto.ed.ac.uk/evolgen/filatov/proseq.html) and DNA Slider (http://udel.edu/~mcdonald).

The coding region of the *nonA* gene from *D. littoralis* is deposited in GenBank (accession number AJ296178). The previously published *nonA* sequences from *D. virilis* (GenBank accession number AJ298998) and *D. melanogaster* (GenBank accession number M33496) were also used in this study. The 19 sequences listed in Figure 2 have been submitted to GenBank and have accession numbers AJ304304–369 for *D. littoralis* and AJ30470–72 for *D. virilis*.

Results

The coding region of the *D. littoralis nonA* gene

The translated region of the *nonA* gene in *D. littoralis* is 2097 nucleotides long producing an open reading frame with homology to that of the *D. virilis* and *D. melanogaster nonA* genes. The structure of *nonA* gene (including the promotor region) has been described in *D. melanogaster* eg, by Besser *et al* (1990) and in *D. virilis* by Campesan *et al* (2001). The exon-intron organisation was the same in all three species (Figure 1), and the intron lengths were also almost identical between *D. littoralis* and *D. virilis*. The first and the second introns were about 2 kbp, the third intron 79 bp and the fourth intron was 780 bp long.

Protein comparison

Sequence analysis of *D. littoralis nonA* predicts a protein of 698 amino acids in comparison with 697 amino acids of *D. virilis* NONA and 700 amino acids of *D. melanogaster* NONA. As shown by Campesan *et al* (2001), the product of *D. virilis nonA* gene can be divided into conserved and non-conserved domains: N-terminal region, central domain and C-terminal region. The N- and C-terminals include stretches of repeated amino acids, the C-terminal also being rich in charged amino acids. The conserved central domain contains two tandemly repeated 80 amino acid RNA-binding domains (RRMs) in exon two, both of which contain RNP1 (eight amino acids) and RNP2 (six amino acids) motifs. These consensus sequences contain aromatic and basic residues, implicated in interactions between proteins and single-stranded nucleic acids (Kenan *et al*, 1991).

Excluding gaps, the overall amino acid identity between *D. littoralis* and *D. virilis* was 97.7% (97.8% when conservative substitutions are included, Table 1). The respective percent identities between *D. littoralis* and *D. melanogaster* and between *D. virilis* and *D. melanogaster* were 77.8 and 77.6% (83.7 and 83.7% when conservative substitutions are included). A closer look at identities and similarities at different parts of the gene revealed that all portions of the gene are conserved between *D. virilis* and *D. littoralis*, while the N-termini of the NONA proteins in these two species are highly diverged from that of *D. melanogaster*. The central domain included only six amino acid substitutions between *D. littoralis* and *D. virilis* and 28 substitutions between *D. littoralis* and *D. melanogaster*. RNA-binding sites, which have been found to be highly

Table 1 Comparison of *nonA* amino acid sequences of *D. littoralis* (lit), *D. virilis* (vir) and *D. melanogaster* (mel) overall and separately in amino terminal (N-term), central domain (central) and carboxy terminal (C-term)

	Percent identity[a]				*Percent similarity*[b]			
	Overall	*N-term*	*Central*	*C-term*	*Overall*	*N-term*	*Central*	*C-term*
lit/vir	97.7	95.7	98.3	100.0	97.8	95.7	98.7	100.0
lit/mel	77.8	55.1	90.7	92.4	83.7	66.0	93.7	95.5
vir/mel	77.6	54.7	90.7	92.4	83.7	66.0	93.7	95.5

[a] To calculate the percent identity, amino acids opposite a gap in either species were ignored.
[b] Percent similarity is the sum of the percent identity and the percent of conservative substitutions between the listed species. Conservative substitutions are based upon the biochemical similarity of the amino acids: D/E, K/R/H, N/Q, S/T, I/L/V, F/W/Y and A/G (Smith and Smith, 1990).

conserved between species (Dong *et al*, 1993) accommodated two amino acid substitutions in RNP2 of the second RNA-binding motif between *D. melanogaster* and the two *D. virilis* group species. The site of the *dissonance* song mutation (site 548 in *D. melanogaster*; Rendahl *et al*, 1996) had Arg in all three species.

Changes in amino acid sequence can be the result of either natural selection or neutral evolution. The ratio of nonsynonymous (Ka) and synonymous (Ks) substitutions provides a measure of the strength of selective forces acting on amino acid level. High Ka/Ks ratio (>1) would indicate strong adaptive selection favouring nonsynonymous substitutions, leading to low constraint on protein sequence. As shown in Table 2, Ka/Ks ratios were low in each species comparison, suggesting that *nonA* amino acid sites are not targets of this kind of selection, but are rather under strong constraint, at least in the *D. virilis* group species.

In addition to amino acid substitutions, the species may differ by insertions and deletions. In particular, the repetitive areas of the developmental genes of *Drosophila* are often characterised by these events (Colot *et al*, 1988). As shown in Table 3, all the three species had stretches of repeated amino acid sequences in both terminals of the NONA. Most of the stretches found in exons 1 (Gly 1), 2 (Gly 2 and a shorter repeat) and 3 (Gly 3) consisted of perfect or imperfect glycine repeats. In addition, all species had imperfect QA repeats in exon 2 corresponding to the N-terminal region. Most repeats were shorter in *D. melanogaster* than in the other two species. As is shown in Table 3, the lengths of the repeats also varied a lot within the species *D. littoralis*.

We calculated two measures for codon usage in *nonA* of the three species: the codon bias index (CBI; Bennetzen and Hall, 1982) and the effective number of codons (ENC; Wright, 1990). In a gene with extreme codon bias CBI will equal 1.0, while with random codon usage it will equal 0. ENC (ranges from 20 to 61) is negatively correlated to the level of codon bias. McVean and Vieira (1999) have studied codon bias in 50 orthologous genes in *D. virilis* and *D. melanogaster* and found the average ENC in both species to be very similar (45.52 and 43.36, respectively). In the case of *nonA*, all three *Drosophila* species had low CBI values (0.113 for *D. littoralis*, 0.112 for *D. virilis* and 0.140 for *D. melanogaster*), and higher than average ENC (50.7 for *D. virilis* and 52.5 for *D. littoralis* and *D. melanogaster*) indicating relatively random use of codons. In Gly repeats, GGT and GGC were the most frequently used Gly codons in all three *Drosophila* species, G-ending codons being almost absent in Gly repeat regions (except for one GGG codon in Gly1 in *D. littoralis*). Gly codon usage in repeat regions did not, however, differ significantly from that in the rest of the gene (χ^2-test, data not shown).

Table 2 The ratio of nonsynonymous (Ka) to synonymous (Ks) substitutions per site in the *nonA* coding region between *D. littoralis* (lit), *D. virilis* (vir) and *D. melanogaster* (mel) overall and separately 5′, central and 3′ region

	Overall	*5′ region*	*Central*	*3′ region*
lit/vir	0.064	0.086	0.052	0.029
lit/mel	0.115	0.196	0.039	0.094
vir/mel	0.120	0.198	0.040	0.076

Table 3 Repeat length variation in the coding region of *nonA* gene in *D. littoralis* (lit) populations from Oulu, Kuopio and Savonlinna and the strain from Zurich, and the presence of the same repeats in *D. virilis* (vir) *and D. melanogaster* (mel)

Species	*Gly 1*	*Gln-Ala*[a]	*Gly 2*	*Gly 3*
lit (Ou)	G_9SG_9 $G_{19–20}$	$(QA)_{18}$	$G_{20–21}$	$G_3N_2G_{12–13}VG_3VG$
lit (Ku)	GVG_{17} GVG_7SG_8 $G_{18–19}$	$(QA)_{18}$	G_{20}	$G_3N_2G_{12–13}VG_3VG$
lit (Sa)	$G_9SG_{8–9}$ $G_{18–19}$	$(QA)_{18}$	$G_{20–22}$	$G_3N_2G_{9–13}VG_3VG$
lit (Zurich)	G_9SG_8	$(QA)_{20}$[c]	G_{20}	$G_3N_2G_{12}VG_3VG$
vir (1431)	G_{11}	$(QA)_{20}$[c]	G_{29}	$G_3NG_{12}VG$
mel	G_8[b]	$(QN)_9$[d]	G_{10}[b]	G_6

[a] Imperfect repeat, where Q can be replaced by H and A can be replaced by L.
[b] Glycine repeat is interrupted by other amino acids.
[c] Imperfect repeat, where Q can be replaced by H and A can be replaced by L or N.
[d] Imperfect repeat, where Q can be replaced by H and N can be replaced by G, R or V.
Amino acid symbols: A = Alanine, G = Glycine, H = Histidine, L = Leusine, N = Asparagine, Q = Glutamine, R = Arginine, V = Valine; S = Serine.

Intraspecific nucleotide variation in *nonA* gene of *D. littoralis*

We analysed altogether ~1.3 kb from the *nonA* locus of six males from each of the three Finnish *D. littoralis* populations (Ou, Ku and Sa). The regions examined included sequences from 5′ end (part of exons 1 and 2) and from 3′ end (part of exons 3 and 4 and intron 3; see Figure 1). The *D. littoralis* samples had, altogether, 22 polymorphic sites including three amino acid variants in exon 1 at nucleotide positions 131 (Ser/Leu), 206 (Gly/Val) and 229 (Gly/Ser). Only one polymorphic site (A/T polymorphism) was present in intron 3 (see Figure 2) .

The imperfect Gln-Ala repeat (in exon 2) was invariable in length among all *D. littoralis* samples, while the lengths of the Gly repeats varied between individuals (Table 3). The length of a perfect stretch of glycines in exon 2 (Gly 2) varied between 20 and 22 amino acids. Variation in Gly repeats 1 and 3 was even larger, these repeats being often interrupted by Ser, Val or Asn. Eight out of 18 (44%) of synonymous mutations were associated with the Gly regions representing about 19% of the analysed sites.

Possible divergence between the three *D. littoralis* samples from Oulu, Kuopio and Savonlinna was studied by calculating F_{st} values for silent sites in 5′ and 3′ parts of the gene. The significance of the F_{st} values was tested by permutation tests (Hudson *et al*, 1992). For the three populations F_{st} for the 5′ end was 0.0576 ($P = 0.1850$ for 1000 permutations) and for 3′ end 0.0000 ($P = 0.4560$). Consequently, the data for the three samples were combined to study nucleotide variation in Finnish *D. littoralis* population.

We calculated two measures of nucleotide diversity for *D. littoralis* data: π is the observed average proportion of nucleotide differences between sequences (Nei, 1987) and θ is based on the number of segregating sites, being sensitive to existence of deleterious alleles at low frequency (Watterson, 1975). Insertion/deletion (indel) variation was not included in the estimates of nucleotide variability. Only the 5′ region had nonsynonymous mutations (three; Table 4). The average variability at synonymous sites in in this region was as small or even smaller than at the 3′ end, and variation in intron sites was surprisingly low. Tajima's D statistics (Tajima, 1989) was –0.99020, $P > 0.10$ (two-tailed test, assuming beta distribution) for the coding region in 5′ region and –0.5045, $P > 0.10$ for the coding region in 3′ region. Negative D values suggest that the population size may recently have undergone drastic changes or that the samples included some deleterious alleles at low frequency. The nonsignificance of the D values shows, however, that the expectation of selective neutrality of mutations was not violated.

Table 4 Nucleotide diversity in the *nonA* gene of *D. littoralis*

	N-term, nsyn (577.04)	*C-term, nsyn (281.83)*	*N-term, syn (184.96)*	*C-term, syn (72.17)*	*C-term, intron (79)*
S	3	0	12	6	1
π	0.00112	0.0000	0.01392	0.02047	0.00662
θ	0.00151	0.0000	0.0.886	0.02417	0.00368

The number of nonsynonymous, synonymous and intron sites analysed for N- and C-terminals are shown in brackets. *S* is the number of segregating sites, π is the average number of differences per base pair and θ is Watterson's estimator based on the number of segregating sites (Watterson, 1975). Because of X-linkage, $\theta = 3N_e\mu$, where N_e is the effective population size and μ is the mutation rate per nucleotide site per generation.

Recombination is known to affect the level of DNA polymorphism in *Drosophila* (Begun and Aquadro, 1992). Our *nonA* sequences showed evidence for a minimum of five recombination events in the history of the population samples, which shows that *nonA* gene, located close to proximal end of the X chromosome in *D. littoralis* (Päällysaho *et al*, 2001), experiences recombination. All four possible gametic types were found in 21 out of 231 pairwise comparisons involving 22 polymorphic sites. Significant linkage equilibrium was detected only once (by χ^2 method with sequential Bonferroni correction) in 36 pairwise comparisons involving nine informative polymorphic sites. The distance between the linked alleles was very short (6 bp). In our sample the recombination parameter R, which is based on the variance of the average number of nucleotide differences between pairs of sequences, was 0.0862. For an X-linked locus $R = 3Nr$ (N is the population size and r is the recombination rate per sequence; Hudson, 1987).

The Swiss *D. littoralis* strain 1007, for which we sequenced the whole coding region but no introns, differed from the Finnish *D. littoralis* samples by 1 fixed synonymous substitution in a region sequenced for all strains. The average divergence level, D (Nei, 1987), was 0.0248 for synonymous sites and 0.0011 for nonsynonymous sites.

Divergence between *D. littoralis* and *D. virilis*

The combined data from *D. littoralis* samples were used for comparisons with a single *nonA* sequence from *D. virilis*. Instead of using here the *D. virilis* sequence by Campesan *et al* (2001), which reported only coding region sites, we sequenced the relevant 5′ and 3′ regions of *nonA* from *D. virilis* strain 1431. Sequences of the two *D. virilis* strains differed from each other in the coding region examined by one synonymous base substitution. The *D. virilis* 1431 sequence differed from *D. littoralis* sequences in 62 sites. Fifteen of these sites were located in intron 3. The average divergence level D (Nei, 1987) between the *D. littoralis* and *D. virilis nonA* sequences was 0.1566 for synonymous sites, 0.0137 for nonsynonymous sites and 0.2271 for the introns, ie synonymous sites are diverging 10 times faster than replacement sites, but slower than intron sites. The standard error for the intron divergence level is, however, rather high as it is based on only 68 sites (79 sites in *D. littoralis* and 68 sites in *D. virilis*).

We applied the McDonald-Kreitman test (McDonald and Kreitman, 1991), which compares the number of sites that are polymorphic within the species to those fixed between the species for replacement vs. synonymous sites excluding gaps, to the *D. littoralis–D. virilis* data. Under neutrality, the ratios should be equal. There were a total of 17 synonymous polymorphic sites and three replacement polymorphic sites in *D. littoralis* sample, and 35 fixed synonymous subsititutions and 11 fixed replacement substitutions between *D. littoralis* and *D. virilis* (a region of 1312 bp). The ratios were nearly equal, consistent with the neutral model ($G = 0.696$, $P = 0.4041$).

The ratio of polymorphisms to fixed differences across the *nonA* sequence was studied with the DNA Slider program (McDonald, 1998). Heterogeneity was measured by the number of runs of polymorphic sites and fixed differences, where a run is a series of one or more sites of one kind preceded and followed by sites of the other kind. We used here two test statistics: the number of runs, which is most powerful for detecting patterns containing several peaks of polymorphism, and the Kolmogorov–Smirnov statistic, which is most powerful for detecting patterns in which one end of the gene has high polymorphism and the other end has low polymorphism (McDonald, 1996, 1998). Significance of these statistics was tested by performing repeated Monte Carlo simulations of a coalescent model using parameters estimated from silent site variation in the data and an arbitrary value for recombination (1000 simulations with five recombination parameters: $R = 2$, 4, 6, 8 and 10). The data contained 29 runs showing no deviation from heterogeneity under a neutral model (the *P* values with different recombination parameters varied from 0.5340 to 0.6225). The Kolmogorov–Smirnov statistic was 0.04584. The *P* values for this test with different recombination parameters were nonsignificant (0.3850 and 0.4420), which suggests that there was no single change across a gene from low to high polymorphism, ie the 5′ region did not show higher polymorphism than the 3′ end.

Divergence between *D. littoralis* and *D. melanogaster*

The *D. melanogaster* sequence differed from the *D. littoralis* sequences at 353 sites. The average divergence level between these two species was 0.6266 for synonymous and 0.2686 for nonsynonymous sites. These numbers may, however, not be quite accurate because of possible saturation in the number of fixed synonymous differences during the long diverge time between the species.

Discussion

Comparison of gene sequences between distantly related species has been a commonly used technique to identify biologically functional regions in genes. *D. virilis* and *D. melanogaster*, which have diverged 60 million years ago (Beverley and Wilson, 1984), have been a very popular species pair in such studies. For example, genes like *engrailed*, *hunchback*, *master mind* and *neuralized* (see Zhou and Boulianne, 1994) are highly (>80% amino acid identity) conserved between the two species. The present study showed very high (98%) conservation in *nonA* amino acid sequence between two *D. virilis* group species, *D. littoralis* and *D. virilis* (diverged 20 Mya; Spicer, 1991). The central domain and C-terminus of the NONA were conserved (91–92%) also between the two *D. virilis* group species and *D. melanogaster*, while the N-terminal regions showed divergence (only 55% identity) between the species groups.

Comparison of gene sequences of distantly related species is not a very informative strategy when studying the reasons for gene divergence or conservation, because of the problems with sequence alignment and the possible saturation of sites for synonymous substitutions. One way to overcome these problems is to study sequence variation within the species and compare this variation to fixed differences between closely and distantly related species. We have used this strategy to find out whether the divergence of the 5′ sequence between *D. melanogaster* and *D. virilis* group species is due to lowered functional constraint, or whether it might be a consequence of positive selection affecting this region.

nonA showed a high rate of synonymous substitutions and weak codon bias in both terminals in *D. littoralis*, which suggests that these regions are not affected by selection for speed and accuracy of translation (see Akashi, 1994). The average level of variability at synonymous sites was 1.40% in the 5′ region and 2.09% in the 3′ end, variation in both regions being higher than the average (0.70%) found earlier for X-linked genes in *D. virilis* (Vieira and Charlesworth, 1999). In *D. virilis*, *nonA* is located at the proximal end of the X chromosome, bands 15D/16A, and in *D. littoralis* it is even closer to the centromere (Päällysaho *et al*, 2001). Our results on *nonA* suggest that the finding of Vieira and Charlesworth (1999) of no reduction in variation and no evidence for suppression of recombination at the base of *D. virilis* X chromosome is relevant also for *D. littoralis*. This is in contrast to *D. melanogaster*, where the base of all chromosome arms is a region of low recombination (Ashburner, 1989).

Comparison of *nonA* sequence polymorphism in *D. littoralis* with fixed differences between this species and *D. virilis* supports the neutral mutation hypothesis assuming purifying (negative) selection to be the most prevalent type of selection affecting the gene. Also, the fact that the ratio of polymorphisms to fixed differences was fairly uniform across the gene suggests that the gene is evolving neutrally in the *D. virilis* group species (McDonald, 1998).

Both the 5′ and 3′ regions of *nonA* gene had repeat regions showing high variability in length, but not in position, within and between the species. The presence of repeated sequences coding for amino acid stretches and their poor conservation between species are features shared with many developmental genes of *Drosophila* (see Colot *et al*, 1988). The repeat regions, and more generally the nucleotide sequences that comprise the variable regions of *per* protein, have been found to be hot spots for events that lead to DNA (and therefore protein) changes (Colot *et al*, 1988). The most obvious role of repeats could be at the level of protein secondary structure (Newfeld *et al*, 1993). Among the human proteins resembling NONA by their structure PSF and p54^{nrb} are unusually rich in Gln and Pro in their N-termini (Dong *et al*, 1993). In addition, NONO has a repeat of 10 Gln in the same place as the three *Drosophila* species have their last Gly repeat, preceding the central domain (Yang *et al*, 1993).

Rendahl *et al* (1996) have suggested that *nonA* has a role in mRNA processing in the central nervous system. However, all mutants of this gene that show a deficient song phenotype are also affected in vision (Stanewsky *et al*, 1996), and the possibility that *nonA* mutant phenotypes (like male song) are caused by nonspecific effects such as generalised effects of protein stability cannot be excluded (Rendahl *et al*, 1996). In the present study *D. virilis*, *D. littoralis* and *D. melanogaster* did not have any amino acid differences in RNP1 or at the site of *dissonance* mutation, which shows that species differences in male courtship song cannot be caused by point mutations at these sites. The regions, which varied most profoundly between *D. littoralis* males and between *D. littoralis* and *D. virilis* corresponded to the Gly repeats located in N- and C-terminals in NONA. Whether these repeats could affect species-specificity of the male song characters, as Thr-Gly repeat in *per* gene has been found to do in *D. melanogaster* (Wheeler *et al*, 1991), remains to be studied.

Acknowledgements

This work was supported by the Academy of Finland to SH and AH (projects 36166 and 44960). SC was supported by a grant from EC predoctoral training fellowship. We are grateful for CP Kyriacou and the two anonymous reviewers for constructive critics on the manuscript.

References

Akashi H (1994). Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* **136**: 927–935.

Ashburner M (1989). *Drosophila: A Laboratory Handbook*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, New York.

Begun DJ, Aquadro CF (1992). Levels of naturally occurring DNA polymorphism correlate recombination rates in *D. melanogaster*. *Nature* **356**: 519–520.

Bennetzen JL, Hall BD (1982). Codon selection in yeast. *J Biol Chem* **257**: 3026–3031.

Besser HV, Schnabel P, Weiland C, Fritz E, Stanewsky R, Saumweber H (1990). The puff-specific *Drosophila* protein Bj6, encoded by the gene *no-on-transientA*, shows homology to RNA-binding proteins. *Chromosoma* **100**: 37–47.

Beverley SH, Wilson AC (1984). Molecular evolution in *Drosophila* and higher *Diptera*. II. A time scale for fly evolution. *J Mol Evol* **21**: 1–13.

Campesan S, Chalmers D, Sandrelli F, Megighian A, Peixoto AA, Costa R *et al* (2001). Comparative analysis of the *nonA* region in *Drosophila* identifies a highly diverged 5′ gene that may constrain *nonA* promoter evolution. *Genetics* **157**: 751–764.

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.

Dong B, Horowitz DS, Kobayashi R, Krainer R (1993). Purification and cDNA cloning of HeLa cell p54^{nrb}, a nuclear protein with two RNA recognition motifs and extensive homology to human splicing factor PSF and *Drosophila* NONA/BJ6. *Nucleic Acids Res* **21**: 4085–4092.

Glenn TC, Glenn SJ (1994). Rapid elution of DNA from agarose gels using polyester plug spin inserts (PEPSIs). *TIG* **10**: 344.

Hotta Y, Benzer S (1970). Genetic dissection of the *Drosophila* nervous system by means of mosaics. *Proc Natl Acad Sci USA* **67**: 1156–1163.

Hudson RR (1987). Estimating the recombination parameter of a finite population model without selection. *Renet Res* **50**: 245–250.

Hudson RR, Boos DD, Kaplan NL (1992). A statistical test for detecting population subdivision. *Mol Biol Evol* **9**: 138–151.

Jones KR, Rubin GM (1990). Molecular analysis of *no-on-transientA*, a gene required for normal vision in *Drosophila*. *Neuron* **4**: 711–723.

Kenan DJ, Query CC, Keene JD (1991). RNA recognition: towards identifying determinants of specificity. *Trends Biochem Sci* **16**: 214–220.

Kulkarni SJ, Steinlauf AF, Hall JC (1988). The *dissonance* mutant of courtship song in *Drosophila melanogaster*: isolation, behavior and cytogenetics. *Genetics* **118**: 267–285.

McDonald JH (1996). Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol* **13**: 253–260.

McDonald JH (1998). Improved test for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. *Mol Biol Evol* **15**: 377–384.

McDonald JH, Kreitman M (1991). Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351**: 652–654.

McVean GAT, Vieira J (1999). The evolution of codon preferences in *Drosophila*: A maximum-likelihood approach to parameter estimation and hypothesis testing. *J Mol Evol* **49**: 63–75.

Nei M (1987). *Molecular Evolutionary Genetics*. Columbia University Press: New York.

Newfeld SJ, Schmid AT, Yedvobnick B (1993). Homopolymer length variation in the *Drosophila* gene *mastermind*. *J Mol Evol* **37**: 483–495.

Patton JG, Porro EB, Galceran J, Tempst P, Nadal-Ginard B (1993). Cloning and characterization pf PSF, a novel pre-mRNA splicing factor. *Genes Dev* **7**: 393–406.

Päällysaho S, Huttunen S, Hoikkala A (2001). Identification of X chromosomal restriction fragment length polymorphism markers and their use in a gene localisation study in *Drosophila virilis* and *D. littoralis*. *Genome* **44**: 1–7.

Rendahl KG, Gaukhshteyn, N, Wheeler DA, Fry TA, Hall JC (1996). Defects in courtship and vision caused by amino acid substitutions in a putative RNA-binding protein encoded by the *no-on-transient A* (*nonA*) gene of *Drosophila*. *J Neurosci* **16**: 1511–1522.

Smith RF, Smith TF (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc Natl Acad Sci USA* **87**: 118–122.

Spicer GS (1991). Molecular evolution and phylogeny of the *Drosophila virilis* species group as inferred by two-dimensional electrophoresis. *J Mol Evol* **33**: 379–394.

Stanewsky R, Fry TA, Reim I, Saumweber H, Hall JC (1996). Bioassaying putative RNA-binding motifs in a protein encoded by a gene that influences courtship and visually mediated behavior in *Drosophila*: *in vitro* mutagenesis of *nonA*. *Genetics* **143**: 259–275.

Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG (1997). The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **24**: 4876–4882.

Wheeler DA, Kyriacou CP, Greenacre ML, Yu Q, Rutila J, Rosbash M, Hall JC (1991). Molecular transfer of a species-specific courtship behaviour from *Drosophila simulans* to *Drosophila melanogaster*. *Science* **25**: 1082–1085.

Vieira J, Charlesworth B (1999). X chromosome DNA variation in *Drosophila virilis*. *Proc R Soc Lond B* **266**: 1905–1912.

Watterson GA (1975). On the number of segregating sites in genetical models without recombination. *Theor Pop Biol* **7**, 256–275.

Wright F (1990). The 'effective number of codons' used in a gene. *Gene* **87**: 23–29.

Yamamoto D, Jallon J-M, Komatsu A (1997). Genetic dissection of sexual behavior in *Drosophila melanogaster*. *Annu Rev Entomol* **42**: 551–585.

Yang Y-S, Hanke JH, Carayannopoulos L, Craft CM, Capra JD, Tucker PW (1993). NonO, a Non-POU-domain-containing, octamer-binding protein, is the mammalian homolog of *Drosophila nonA^{diss}*. *Mol Cell Biol* **13**: 5593–5602.

Zhou L, Boulianne GL (1994). Comparison of the *neuralized* genes of *Drosophila virilis* and *D. melanogaster*. *Genome* **37**: 840–847.