Comparative Analysis of the nonA Region in Drosophila Identifies a Highly Diverged 5′ Gene That May Constrain nonA Promoter Evolution

Susanna Campesan,* David Chalmers,*1 Federica Sandrelli,1 Aram Megighian,2 Alexandre A. Peixoto,*§ Rodolfo Costa1 and Charalambos P. Kyriacou*

*Department of Genetics, University of Leicester, Leicester LE1 7RH, England, 1Dipartimento di Biologia, Università di Padova, 35131 Padova, Italy, 2Dipartimento di Anatomia e Fisiologia Umana, Università di Padova, 35131, Padova, Italy and 3Department de Bioquimica e Biologia Molecular, Fundacao Oswaldo Cruz, Rio de Janeiro, CEP 21045-900, Brazil

Manuscript received May 1, 2000
Accepted for publication November 7, 2000

ABSTRACT

A genomic fragment from Drosophila virilis that contained all the no-on-transientA (nonA) coding information, plus several kilobases of upstream material, was identified. Comparisons of nonA sequences and the gene nonA-like in D. melanogaster, a processed duplication of nonA, suggest that it arose before the split between D. melanogaster and D. virilis. In both species, another gene that lies <350 bp upstream from the nonA transcription starts, and that probably corresponds to the lethal gene l(1)i19, was identified. This gene encodes a protein that shows similarities to GPI1, which is required for the biosynthesis of glycosylphosphatidylinositol (GPI), a component for anchoring eukaryotic proteins to membranes, and so we have named it dGpi1. The molecular evolution of nonA and dGpi1 sequences show remarkable differences, with the latter revealing a level of amino acid divergence that is as high as that of transformer and with extremely low levels of codon bias. Nevertheless, in D. melanogaster hosts, the D. virilis fragment rescues the lethality associated with a mutation of l(1)i19, as well as the viability and visual defects produced by deletion of nonA+. The presence of dGpi1 sequences so close to nonA appears to have constrained the evolution of the nonA promoter.

MUTATIONS in the sex-linked, no-on-transientA (nonA) gene of Drosophila melanogaster produce behavioral defects in vision and in the male courtship song (Hotta and Benzer 1970; Pak et al. 1970; Kulkarni et al. 1988). The gene lies within cytological position 14Cl-2, a region that also contains two lethal complementation groups situated immediately distally to nonA, l(1)i19e and l(1)9.21 (Jones and Rubin 1990). One of these, l(1)i19e, appears to overlap with the 5′ half of nonA, but the sequences corresponding to this lethal gene have yet to be identified (Stanewsky et al. 1995). The l(1)9.21 region encodes the pre-mRNA splicing factor U2AF (Kanaar et al. 1993) and more distally, the region 20–50 kb from nonA has revealed a number of cDNAs and open reading frames, the most studied being cyclophilin-1 (Rutherford 1995). Conceptual translation of nonA predicts a protein of 700 amino acids (Besser et al. 1990; Jones and Rubin 1990). The central segment contains two tandemly repeated 80-amino-acid motifs, common to a family of proteins known for their ability to bind RNA. The RNA recognition motif, RRM, also known as RNP or RBD (Stom and Dreyfuss 1997), is moderately conserved from yeast to humans (Bandziulis et al. 1989).

The original nonA mutants have defects in their visual system but their courtship song is unaffected, whereas the first nonA song mutant, dissonance (later renamed nonAdissonance), has song pulses that appear reasonably normal at the beginning of a song burst, but become polycyclic as the burst progresses (Hotta and Benzer 1970; Pak et al. 1970; Kulkarni et al. 1988). Like other visual mutants, nonAdissonance is also defective in its abnormal electroretinogram (ERG) and optomotor response, suggesting both peripheral and central visual system lesions (Kulkarni et al. 1988; Rendahl et al. 1992, 1996; Stanewsky et al. 1996). Amino acid substitutions within or very close to the second RRM of NONA produce visual but not song abnormalities, whereas the nonAdissonance mutation creates an amino acid substitution in a downstream region notable for the high proportion of charged residues (Rendahl et al. 1996).

The pulse structure of the songs of nonAdissonance mutants resembles, at least superficially, that of D. virilis (Hoikkala and Lumme 1984, 1987). Specifically, pulses of D. virilis are more polycyclic compared to wild-type D. melanogaster, but they also show the additional nonAdissonance feature of increasing the number of cycles per pulse as the song burst progresses (S. Campesan, Y. Dubrova, J. C. Hall and C. P. Kyriacou, unpublished results). These observations stimulated us to attempt to identify

Corresponding author: C. P. Kyriacou, Department of Genetics, University of Leicester, Leicester LE1 7RH, United Kingdom. E-mail: cpk@leicester.ac.uk

1Present address: ETS/FC, 25020 Besancon, France.

Genetics 157: 751–764 (February 2001)
the *D. virilis* nonA orthologue and to analyze in detail the courtship song phenotypes of *D. melanogaster* transgenes carrying the *virilis* transgene. Similar interspecific transformation experiments performed with the period (*per*) gene have revealed that *per* carries species-specific behavioral information for both circadian locomotor activity patterns and for ultradian love song cycles (Petersen et al. 1988; Wheeler et al. 1991). Although severe, the nonA<sup>19e</sup> mutation is not amorphic (Stanewsky et al. 1993), so it is preferable for the proposed study that the *D. virilis* nonA transgene be placed on a nonA null background in *D. melanogaster*. This can be provided by the use of the translocation *T(1;4)9e2-10*, in which both nonA and the distal, partially overlapping lethal locus *l(1)19e* are deleted (Jones and Rubin 1990). The most efficient strategy would be to identify a fragment from the *D. virilis* nonA region that encodes both nonA and the adjacent lethal locus and to transform this fragment into the double gene deletion background provided by *T(1;4)9e2-10*.

Our comparative analysis of *D. virilis* and *D. melanogaster* nonA thus also sought to identify the genomic and cDNA sequences corresponding to *l(1)19e*. Consequently, we have isolated a *D. virilis* genomic fragment that encodes nonA and have identified within its promoter a gene we call *dGpi1*, which almost certainly corresponds to *l(1)19e*. We compare the two species nonA sequences with those of *D. melanogaster* nonA-like, which lies in the bithorax complex (Martin et al. 1995), and make further comparisons between the two species coding sequences of *dGpi1*, which reveal extremely high levels of divergence. We also examine whether the presence of the *dGpi1* gene in the promoter region of nonA constrains the evolution of the nonA regulatory region. This is particularly relevant given that the accompanying article by Sandrelli et al. (2001) demonstrates that the transcription unit of *dGpi1* also acts as both a silencer and enhancer of various behavioral and tissue expression phenotypes of nonA. Finally, we investigate whether the transformed *D. virilis* fragment is able to rescue the lethality associated with *l(1)19e* and the nonA ERG and viability defects.

**MATERIALS AND METHODS**

**Identification of *D. virilis* nonA and *l(1)19e*** A small fragment of the *D. virilis* nonA gene was amplified by PCR using 5′ primer 5′-CCGGAGATGTTCAAGCCATA-3′ (4163–4182) and 3′ primer 5′-GCGCTCTGAGTGGACCACAA-3′ (4422–4405), based on the second exon sequence of the *D. melanogaster* nonA gene (nucleotide positions from sequence of Jones and Rubin 1990). The amplified 261-bp fragment was sequenced to confirm its homology with *D. melanogaster* nonA and used to screen an EMBL3 phage *D. virilis* genomic library constructed by Ron Blackman and kindly donated by John Belote. DNA from one putative positive clone was digested with *SalI*, and two fragments of 6.5 and 6.0 kb, respectively, which hybridized to a *D. melanogaster* nonA probe, were subcloned into pUC18. Manual sequencing was performed on various subclones of the two fragments and coding regions were confirmed several times on both strands. Intron-exon boundaries were studied by comparing sequences from RT-PCR products with genomic sequences. In addition 5′ rapid amplification of cDNA ends (RACE) was performed to reveal the transcription start site of both *D. melanogaster* and *D. virilis* nonA. RT-PCR and 5′ and 3′ RACE were also performed on putative *D. melanogaster* and *D. virilis* transcripts from the lethal gene *l(1)19e*, which was believed to be embedded within the 5′ and N-terminal regions of nonA (Jones and Rubin 1990; Stanewsky et al. 1993). The two positive clones were then ligated into the pW8 transformation vector to reconstitute *D. virilis* nonA, including ~3 kb of upstream and 1 kb of downstream sequence, using a number of cloning steps. The integrity of nonA was confirmed by sequencing.

**Analysis of sequence variation in *D. melanogaster* and *D. simulans* nonA promoter fragments** Single *D. melanogaster* males were obtained from five isofemale lines established in 1994 from a natural population sampled in Lecce (Italy), and *D. simulans* males were obtained from three isofemale lines established from a natural population from Zimbabwe. Single fly genomic DNA was prepared as previously described (Gloor and Engels 1990). An 863-bp fragment for *D. melanogaster* and an 880-bp fragment for *D. simulans*, located immediately upstream of the coding region of nonA, were amplified by using the forward primer 5′-GGGGATCCGGCGGCGC TGGTCTCC3′ (positions 974–991 in the sequence of Jones and Rubin 1990) and the reverse primer 5′-TATGATCCCG TACAATCTGTTGGACA-3′ (positions 1849–1868). The amplified fragments were sequenced automatically.

**Computer analyses** All sequence analyses were performed using the programs available from the Genetics Computer Group (GCG) package for molecular biology. A statistical analysis of cryptic simplicity in the coding sequence DNA was performed using the SIMPLE34 program, which generates a Relative Simplicity Factor (RelSF) for each sequence (Hancock and Armstrong 1994). The PSITE program was used to search for functional motifs in the NONA proteins (Solovyev and Kolchanov 1994). SIGNAL SCAN (Prestridge 1991) and TF SEARCH (Akiyama 1995; Heinemeyer et al. 1998) were used to search for functional motifs in the 5′ regulatory region of nonA.

**Transformations** Element-mediated transformation was performed using standard methods with the pW8 vector that carries w<sup>+</sup> as a marker (Spradling and Rubin 1982; Klemenz et al. 1987). Embryos microinjected were either w; *Sb e* ΔΔ3-3/TM6 or w<sup>111</sup> when using the latter injections, transposase was provided by coinjection with PUCH<sup>+</sup> ΔΔ3 (a gift from J. M. Dura). A number of independent lines were obtained and the inserts were mapped to at least the chromosomal level. Southern blotting showed that all lines contained single copy insertions. Line 112 was sex linked and mapped close to endogenous nonA, and line 113 integrated on the *Sb e* ΔΔ3-3 chromosome and was crossed off to avoid further transposition via ΔΔ3. Because the 113 insert was homozygous lethal, it was used in a mobilization assay to generate two further hoppers, 168-8 and 67-4, which complemented the lethality of insert 113. Lines 72 and 297 both contained homozygous viable X chromosome inserts. The 297 insert was successfully mobilized to chromosome 3 to give line 297-6. Lines 97, 135, and 191 contained single chromosome 3 insertions, and line 75 carried the transgene on chromosome 2.

**Viability** Females heterozygous for *In(1)FM7* (marked with y w<sup>1</sup>) and the translocation *T(1;4)9e2-10*, marked with y<sup>9</sup> v<sup>1</sup> y<sup>f</sup> f, which carries a deletion uncovering nonA and *l(1)19e* (Stanewsky et al. 1993), were crossed to males carrying an autosomal copy of the *D. virilis* nonA fragment. Ordinarily, translocation males can survive only if the deletion is comple-
RESULTS

Intron-exon structure of D. virilis nonA: The five-exon/four-intron organization found in D. melanogaster nonA is conserved in the D. virilis homolog (Figure 1). The intron-exon boundaries are also conserved as revealed by cDNA and genomic DNA comparisons (data not shown). The approximate lengths of the first two introns (~2.4 and 2.0 kb), as calculated by measuring the length of PCR products obtained by using primers annealing to the exon boundaries, are two and four times, respectively, the sizes of their melanogaster counterparts. The third intron is the same size in both species (~70 bp) but the length of the fourth intron is unknown.

Sequence comparisons: Dot matrix comparisons between the D. virilis and the D. melanogaster nonA coding sequences revealed an area of considerable divergence covering approximately the first half of the gene (data not shown). Plots of each of the two nonA sequences against itself clearly showed numerous large regions of repeated DNA, clustered especially at the beginning and the end of the gene. D. melanogaster nonA appeared much less repetitive than its virilis counterpart, and this was confirmed by computing the RelSF for the two sequences (Hancock and Armstrong 1994), which gave values of 1.552 for D. melanogaster and 1.898 for D. virilis.

Figure 2 shows an alignment of the ~700-amino-acid sequence of D. virilis and D. melanogaster NONA, together with a third D. melanogaster protein encoded by the nonA-like gene (Martin et al. 1995). Overall identity between the two species NONA proteins is just over 75%, while the similarity is ~83% (Table 1). Pairwise comparisons revealed corresponding values of 68 and 72% between D. melanogaster nonA-like and both D. melano-

nagaster and D. virilis nonA. The N-terminal third of the NONA protein up to the RNA-binding domain is the most diverged, with <50% identity (Figure 2). These regions of divergence are constituted in large part by stretches of repeats, particularly tracts of poly(Gly), and a QN and a degenerate GNQGX repeat found in D. melanogaster, which has been replaced by a QA and a very long 29-residue poly(Gly) repeat in D. virilis. The RNA-binding domain (RRM1 + RRM2, residues 295–453 in D. melanogaster nonA) is very well conserved between the two species. The RNP1 octamer and RNP2 hexamer motifs within RRM1 are perfectly conserved, but RNP1 in RRM2 has two changes.

The adjacent charged region (amino acids 454–568; see Figure 2) includes residue 548, in which an aspargine is substituted by cysteine in the nonA-like sequence of D. virilis (Rendahl et al. 1998). This position is conserved in both nonA-like and in D. virilis. Application of the PSITE program revealed that the sequence KRESDE (residues 530–536) spanning the nonA-like site in D. virilis contains putative phosphorylation sites for both cyclic nucleotide-dependent and casein kinase II
protein kinases, whereas the corresponding sequence in D. melanogaster, KREVDNE (residues 547–553), has lost these potential modifications. All other putative post-translational modification sites in the RRM and charged regions are conserved between the two species.

The 5′ regulatory region of nonA: Approximately 2.5 kb of upstream sequence from the D. virilis nonA fragment was initially obtained and compared with the upstream sequence of D. melanogaster (Jones and Rubin 1990). A dot matrix analysis (window, 21; stringency, 14) was performed for the D. virilis and D. melanogaster sequence comparison to graphically highlight the regions of homology (Figure 3). Revealed are seven re-

Figure 2.—CLUSTAL alignment of NONA protein in D. melanogaster and D. virilis (EMBL database accession no. AJ298998) together with D. melanogaster NONA-like. The two adjacent 80-residue RRMs are underlined, and within these the RNP-I octapeptides and RNP-2 hexapeptides are italicized. Inverted triangles show the position of the introns within the translation products of D. melanogaster and D. virilis nonA. The nonA-like gene is intronless.
gions of moderate-to-good conservation, which are interspersed with regions of complete divergence between the two sequences, while the terminal third is very diverged. As the vital gene \(l(1)i19e\) may be encoded within the 5’ region of \(nonA\) and could overlap with the N-terminal half of \(nonA\) coding sequences (Jones and Rubin 1990; Stanewsky et al. 1993), this pattern of divergence and conservation could reflect the intron-exon pattern, respectively, of \(l(1)i19e\). Primers were generated based on putative conserved coding regions of both species genes, and 5’ and 3’ RACE performed. The amplified cDNA fragments were sequenced and Figure 4 shows a CLUSTAL alignment of the two putative \(l(1)i19e\) coding regions. The gene has an 8 exon-7 intron structure (Figure 1) and encodes a protein of 481 amino acids in \(D. melanogaster\) and 473 in \(D. virilis\). The ends of the two transcripts, as detected by 3’ RACE, fall 424 and 336 bp before the initiating ATG codons of \(nonA\) in \(D. melanogaster\) and \(D. virilis\), respectively. Overall identity between the two Drosophila proteins is 59% and similarity is 67% (Table 1). However, the N and C termini are much more diverged, with identities of 43% (residues 1–214) and 33% (residues 420–481), respectively, compared to the central region (residues 215–419), whose identity is 82%.

A BLAST search of the databases using both sequences revealed similarity with the \(gpi1\) genes of mammals, \(Caenorhabditis elegans\), and yeasts. These encode a component necessary for the first step in the biosynthesis of glycosylphosphatidylinositol (GPI), which is used to anchor eukaryotic proteins to membranes. Figure 4 also shows the CLUSTAL alignments of these various GPI proteins. The similarity between fly and human GPI is 34%, whereas identity is 23%. This rises to corresponding figures of 37 and 28% when compared to \(C. elegans\) and falls slightly when compared to the two yeast species. The alignment shows very few conserved residues among all species, and so putative secondary structure was investigated to look for similarities between the Drosophila and other species proteins. Hydropathy analysis (Kyte and Doolittle 1982; Engelman et al. 1986) reveals that in spite of their low overall identity, there is extensive similarity between the \(D. melanogaster\) and \(Saccharomyces cerevisiae\) proteins in both the number and spatial patterning of hydrophobic regions that may represent transmembrane domains (Figure 5). Consequently, this analysis suggests that they may be homologous proteins, and we suggest naming this Drosophila gene \(dGpi1\). Whether this corresponds to \(l(1)i19e\) will be discussed below.

Comparisons among several \(D. virilis\) and \(D. melanogaster\) homologous proteins revealed identities ranging from 50 to 83% (Table 1). \(dGPI1\) has similarity and identity scores very similar to those of \(transformer\), making it one of the most diverged genes known in Drosophila (O’Neil and Belote 1992).

**Molecular evolution of \(nonA\) and \(dGpi1\):** Considerable divergence has been found in the first half of \(nonA\) and in the N- and C-terminal regions of the \(dGPI1\) proteins. This could reflect a lack of functional importance and freedom from selective constraints or could serve adaptive, species-specific characteristics, particularly in the case of a “behavioral” gene such as \(nonA\) (e.g., Wheeler et al. 1991). We therefore used the \(K_s/K_a\) test to examine the ratio of nonsynonymous \((K_s)\) to synonymous \((K_a)\) substitutions (Li and Graur 1991). A ratio greater than

### Table 1: Drosophila protein identity

<table>
<thead>
<tr>
<th>%</th>
<th>dGPI1</th>
<th>NONA</th>
<th>EN</th>
<th>HB</th>
<th>KNI</th>
<th>NOS</th>
<th>OSK</th>
<th>PER</th>
<th>RUNT</th>
<th>SEV</th>
<th>TRA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similarity</td>
<td>67</td>
<td>83</td>
<td>88</td>
<td>89</td>
<td>83</td>
<td>78</td>
<td>74</td>
<td>78</td>
<td>87</td>
<td>78</td>
<td>66</td>
</tr>
<tr>
<td>Identity</td>
<td>59</td>
<td>75</td>
<td>83</td>
<td>81</td>
<td>75</td>
<td>64</td>
<td>59</td>
<td>68</td>
<td>81</td>
<td>65</td>
<td>50</td>
</tr>
</tbody>
</table>

Percentage similarity and identity between \(D. melanogaster\) and \(D. virilis\) homologous proteins was calculated using the method of Smith and Waterman (1981). The proteins compared between the two species include \(dGPI1\), NONA, engrailed (EN), hunchback (HB), knirps (KNI), nanos (NOS), oskar (OSK), period (PER), runt, sevenless (SEV), and transformer (TRA).
Figure 4.—CLUSTAL alignment of GPI proteins in a variety of species. The corresponding intron positions within the two Drosophila GPI1 genes are shown as inverted triangles. EMBL accession numbers for D. melanogaster and D. virilis Gpi1 are AJ298995 and AJ298996, respectively, and for the nonA promoter region for D. virilis, AJ298997.
unity indicates the action of selection in fixing nonsynonymous variation at a rate faster than the neutral mutation rate. For comparison, the test was also performed on a number of other D. virilis and D. melanogaster homologues (Table 2). In all these genes the rate of nonsynonymous substitutions is occurring at a much lower rate than the synonymous level. Even the high $K_s/K_a$ ratio for $tra$ is still far from unity. The striking feature of these results is the astonishingly high $K_a$ value observed for $dGpi1$ (3.377 sem 0.236), almost three times that of nonA and twice that of $tra$. This is not an artefact of poor alignment because when the $dGpi1$ gene is divided into the conserved central region vs. the nonconserved N- and C-terminal regions together, $K_a$ values > 3 are obtained for both portions. The most diverged part of nonA (from nucleotide 1 to 813) gives a $K_a$ value of 2.155, much lower than that of $dGpi1$.

One possible way to explain the high synonymous rate for $dGpi1$ would be to invoke low levels of codon bias. With no selection for specific codons, the third position would be relatively free of constraints, providing an avenue for inflation of $K_a$ values. The Relative Synonymous Codon Usage (RSCU) index was calculated for a number of different genes in $D. melanogaster$ and $D. virilis$ (Sharp et al. 1988 and see legend of Table 3). It can be seen that $dGpi1$ has the lowest overall codon bias levels in both species, and this may contribute toward its high $K_a$ levels. In addition, the substitution rates for all seven small introns in $dGpi1$, which range from 51 to 70 bp in length in both $D. melanogaster$ and $D. virilis$ (see Figure 1), and for the large first intron of nonA (1298 bp in $D. melanogaster$, 1444 in $D. virilis$), were also calculated. This was done after first removing the canonical donor and acceptor dinucleotides, the short conserved pyrimidine tracts close to the 3’ end, and the single conserved adenine branch point from each intron. Intron nucleotide identity was 37% for $dGpi1$ and 41% for nonA, revealing no obvious increase in the mutation rate of $dGpi1$ that could explain its extraordinary $K_a$ values.

The proximity of the two genes raises the issue of whether $dGpi1$ sequences act as promoter and enhancer regions for nonA expression. This has been studied in the accompanying article by Sandrelli et al. (2001) using various deleted fragments of the nonA upstream regions. It is clear from these results that enhancers and silencers of nonA expression must overlap with $dGpi1$ sequences. We therefore examined potential transcription factor binding sites in the ~2.3-kb region immediately upstream of the transcription start of nonA in both species (see MATERIALS AND METHODS). Putative binding sites might suggest which trans-acting factors could be involved in nonA regulation. The most significant sites (scores ≥ 90) include those for Broad-Complex (BR-C), situated ~350 and 300 bp upstream of the $melanogaster$ and $virilis$ nonA transcription starts, respectively (see Figures 1 and 6), and those for Deformed and heat-shock factors, which were found within the $dGpi1$ sequences (Figure 6). Reducing the stringency of the match between the binding site consensus and the target sequences (scores ≥ 85) revealed two more BR-C sites in the intergenic region of both species.

The presence of $dGpi1$ could thus constrain the evolution of the nonA regulatory region. To explore this further, we adopted a neutrality test (McDonald and

### Table 2

<table>
<thead>
<tr>
<th></th>
<th>nonA</th>
<th>en</th>
<th>per</th>
<th>tim</th>
<th>tra</th>
<th>dGpi1</th>
<th>nos</th>
</tr>
</thead>
<tbody>
<tr>
<td>$K_s$</td>
<td>1.242</td>
<td>1.035</td>
<td>1.234</td>
<td>1.264</td>
<td>1.515</td>
<td>3.377</td>
<td>2.113</td>
</tr>
<tr>
<td>$K_a$</td>
<td>0.190</td>
<td>0.130</td>
<td>0.243</td>
<td>0.102</td>
<td>0.493</td>
<td>0.327</td>
<td>0.266</td>
</tr>
<tr>
<td>$K_a/K_s$</td>
<td>0.153</td>
<td>0.126</td>
<td>0.197</td>
<td>0.081</td>
<td>0.325</td>
<td>0.097</td>
<td>0.126</td>
</tr>
</tbody>
</table>

Synonymous ($K_s$) and nonsynonymous ($K_a$) substitution rates between $D. melanogaster$ and $D. virilis$ were calculated for nonA, engrailed (en), period (per), timeless (tim), transformer (tra), dGpi1, and nanos (nos) using the method of Li and Grauer (1991).
between These sequences corresponded to the intergenic region on a T(1;4)9e2-10®xation of such changes, even in the face of high K

downstream to the transcription start of nonA sequence up to its stop codon and the sequence dGpi1 sequence from D. melanogaster of D. simulans, three from a natural population collected in Lecce, Italy, of D. melanogaster, plus the reference sequence from D. melanogaster (Jones and Rubin 1990). These sequences corresponded to the intergenic region between dGpi1 and nonA, plus ~400 bp of the 3’ half of dGpi1 (representing exons 6, 7, and 8; 823 bp for D. melanogaster and 880 bp for D. simulans). The TF SEARCH program was used on the reference D. melanogaster sequence as before using a stringency of 0.9, and each nucleotide was classified as to whether it lay within or outside of a putative binding site. A total of 85 differences (substitutions, insertions, and deletions) were found among the sequences studied. Of these, 24 were fixed changes between the two species and 61 were polymorphisms. We divided the region into two: the dGpi1 sequence up to its stop codon and the sequence downstream to the transcription start of nonA (Figure 6). Table 4 shows that the dGpi1 sequences reveal no significant association between the type of change (fixed vs. polymorphic) and whether the sequence represents a putative binding site (P = 1.00). However, in the intergenic region, a significant increase is observed in the number of fixed changes within putative binding sites relative to polymorphisms (P = 0.0398). These results suggest the possibility of adaptive changes in the putative binding sites within the intergenic region, whereas dGpi1 imposes additional constraints on the fixation of such changes, even in the face of high K values.

Rescue of l(1)I19e and nonA mutant phenotypes: The extremely high levels of divergence in the dGpi1 coding sequence, which probably corresponds to l(1)i19e, generates considerable doubt as to whether the 12.5-kb D. virilis fragment we have studied will rescue the lethality associated with T(1;4)9e2-10, in which both nonA and l(1)i19e are deleted. We therefore studied the viability of male progeny carrying the virilis transgene on a T(1;4)9e2-10 background by crossing T(1;4)9e2-10/In(1)FM7 females with males carrying a single balanced autosomal transgenic copy of the D. virilis fragment (insert/balancer). The viability results are shown in Table 5 and reveal considerable heterogeneity between the various lines studied (χ² = 31.05, d.f. = 6, P < 0.01), yet it is clear that in all but line 191, at least one-third of the male progeny from all crosses produce T(1;4)9e2-10/Y; insert/+ individuals. Even in line 191 there was no significant departure from the 1:2 ratio of T(1;4)9e2-10/Y to In(1)FM7 males (χ² = 2.43, d.f. = 1), further confirming that the virilis fragment rescues both the mutant l(1)i19e and nonA viabilities to normal levels.

Finally, we examined the ERG, a sensitive measure of nonA function (Reed et al. 1996; Stanewsky et al. 1996), of males carrying the T(1;4)9e2-10/Y; insert/+ genotype from lines 135, 191, and 297.6 and compared them with a Canton-S wild type. Table 6 shows that all flies demonstrated both ON and OFF transients, and ANOVA revealed no significant differences between any of the genotypes in the amplitudes of either response (ON, F = 1.39, d.f. = 1, 61; OFF, F = 0.90, d.f. = 1, 69).

**DISCUSSION**

_D. virilis_ shows an elevation in the amount of repetitive DNA in both the coding and upstream regulatory regions of nonA compared to _D. melanogaster_, mirroring similar observations that were made in comparisons between these two species involving the _hunchback_ gene (Hancock et al. 1999). Comparative analyses of homologous genes reveal that areas of high divergence are often associated with regions of repetitive DNA, both in coding (Treier et al. 1989; Peixoto et al. 1993) and noncoding regions (Tautz et al. 1987; Hancock and Dover 1988). Not surprisingly, therefore, much of the divergence between _D. melanogaster_ and _D. virilis_ in the N-terminal regions of nonA involves repetitive motifs such as QN, GNQGGX, and poly(G).

Poly(G) motifs [another long poly(G) stretch is found in the C terminus of _D. virilis_ NONA] are of particular interest as several known RNA-binding proteins, such as the hnRNP proteins A1 and A2 and the nucleolar prerRNA-binding protein, Nuclein, have auxiliary domains constituted by glycine-rich regions (Bandziulis et al. 1991) that has been used in an attempt to identify adaptive changes in putative regulatory sequences (Jenkins et al. 1995). Five sequences were obtained from a natural population collected in Lecce, Italy, of _D. melanogaster_, three from a natural population of _D. simulans_ from Zimbabwe, plus the reference sequence from _D. melanogaster_ (Jones and Rubin 1990). The number of codons within each gene that has an RSCU index >1.5, <2, and >2 is shown for _D. virilis_ and _D. melanogaster_ (parentheses). RSCU is defined as the observed number of codons divided by the expected number if all codons were used equally (Sharp et al. 1988). Note the extremely low overall codon bias of dGpi1.
unique example of a processed duplicated behavioral (O'Neil et al. 1989). These auxiliary domains may be involved in protein-protein interactions, but have also been shown to be involved in the polynucleotide binding properties of RNA-binding domains (BANDZIULIS et al. 1989). Thus the long N-terminal stretch of 29 Gly residues in D. viridis compared to D. melanogaster nonA may have functional relevance. Interestingly, two-dimensional plots of NONA from the two species (using the Peptidestructure and Plostructure programs from the GCG package) revealed a marked difference in the N-terminal regions (data not shown). In D. viridis NONA, a large uninterrupted domain of turns is predicted from the poly(Gly) tract and is preceded by a long a-helical conformation produced by the QA repeats. The corresponding region in D. melanogaster NONA has no helical conformation and very short, frequently interrupted areas of turns (data not shown).

The areas of high conservation between the two nonA sequences correspond to the RRM1s. Mutational studies have revealed that the first RRM domain (RRM1) in nonA is necessary for all the known functions of NONA (RENDahl et al. 1996; STANESKY et al. 1996). Mutations in this region not only cause severe defects in both visual and song phenotypes, but also invariably reduce the viability of the affected flies. On the contrary, mutations in the RRM2 domain either have little or no effect or produce impairments of the visual system only (RENDahl et al. 1996; STANESKY et al. 1996). In this regard, we note that the RNP1 region of RRM2 has a lower level of conservation than RRM1 (see Figure 2). The NONA protein can therefore be roughly divided into two, the N-terminal diverged fragment and the central and C-terminal conserved regions. However, within the C-terminus lies the charged region in which is located the site of the nonA i19e song mutation (RENDahl et al. 1996). It was therefore of interest that a single substitution between the two species in the region including this mutant site generated additional potential post-translational modifications in D. viridis. If nonA does act as a reservoir for species-specific song information (S. CAMPesan, Y. DUBROVA, J. C. Hall and C. P. KYRIACOU, unpublished results), then perhaps this difference in sequence might be relevant, because it lies in an area of the NONA protein that has some influence on the song phenotype (RENDahl et al. 1996).

Comparison of the nonA genes with nonA-like revealed lower identity scores between the nonA and nonA-like proteins than between the nonA orthologues. The nonA-like gene is found within the bithorax complex of chromosome 3 and is unusual because it encodes a single open reading frame (MARTin et al. 1995). This suggests that the duplication event giving rise to nonA-like was mediated by an RNA intermediate followed by transposition. The absence of stop codons suggests that nonA-like may be functional, although the available sequence is genomic only. Consequently, nonA-like may represent a unique example of a processed duplicated behavioral gene. On the basis of the identity scores, we can assume that the duplication event occurred before the D. melanogaster-D. viridis split.

Comparison of the 5’ region of nonA revealed the presence of dGpi1, which may correspond to l(1)i19e. The protein sequence has a low level of identity with the product of the gpi1 gene of yeast (LEIDICH and ORLEAN 1996), which is used in GPI synthesis to anchor proteins in the endoplasmic reticulum. The biosynthesis of GPI requires sequential additions of sugar molecules to phosphatidylinositol (PI) in a number of steps (ENGlund 1993; McCONville and FERGUSON 1993). The first stage requires the synthesis of N-acetylglucosaminyl phosphatidylinositol (GlcNAc-PI) from UDP-GlcNAc and PI, catalyzed by GPI-GlcNAc transferase (GPI-GnT, DOERING et al. 1989), and involves more than one gene. In yeast, one of these genes is GPI1; it encodes a 609-residue polypeptide with several hydrophobic membrane-spanning domains (LEIDICH and ORLEAN 1996). The amino acid identity between GPI1 and the Drosophila protein is low, but the hydropathy profile is very similar to that of S. cerevisiae. We have therefore taken the liberty of naming this fly gene dGpi1.

The dGpi1 gene almost certainly corresponds to l(1)i19e for a number of reasons. First, it lies in a region of overlap between nonA and l(1)i19e, as predicted (JONES and RUBIN 1990). Second, a nonA fragment with the same 5’ end as that of 270XS16 (beginning at amino acid residue 148 of dGpi1; see Figure 1), and encompassing the whole promoter region up to the nonA start codon, does not rescue the inviability of transformants carrying the l(1)i19e mutation (SANDRELLI et al. 2001, accompanying article). In contrast, the original 270XS16 fragment that carries the 5’ and coding regions of nonA has been reported to rescue this lethality to some degree, suggesting that l(1)i19e sequences may “straddle” the 5’ end of 270XS16 (JONES and RUBIN 1990). The N-terminus of GPI1 proteins is so diverged (see Figure 4) that it could be that, in some transformant lines, flanking regions at the 5’ end of the insert may promote the transcription and translation of either a truncated dGPI1 or a fusion dGPI1 product that may rescue viability. It is certain that, because these rescuing (or nonrescuing) fragments begin in the first exon of dGpi1 (Figure 1), the sequences corresponding to l(1)i19e cannot be encoded downstream of those for dGpi1. The only caveat (on the basis of our results only) is that the D. viridis fragment that rescues T(1;4)9e2-10 still has ~1 kb of unsequenced material upstream of the transcriptional start of dGpi1, so a gene could be encoded immediately 5’ of dGpi1. However, if one accepts the arguments outlined above concerning the sporadic rescue of l(1)i19e with 270XS16 (JONES and RUBIN 1990), then dGpi1 is l(1)i19e.

The divergence of the dGpi1 gene between D. melanogaster and D. viridis is almost as high as that of transformer (O’NEIL and BELOTE 1992). The central region, which
Figure 6.—Alignment of sequences for five *D. melanogaster* (863 bp) and three *D. simulans* (880 bp) natural haplotypes 5' of the *nonA* transcriptional start. The reference sequence is that obtained by Jones and Rubin (1990) and is reported from nucleotide 992 to 1855 (GenBank accession no. M33496). This corresponds to nucleotides –917 to –54 in relation to the initiating methionine codon of *nonA* (see Figure 1). A dot represents bases that are identical to the Jones and Rubin (1990) sequence; a dash represents single base deletion; *dGpi1* introns are shown in italics. Putative binding sites (or their complementary sequences) are boxed. The long turquoise box represents sequences downstream of the *dGpi1p* stop codon, which is asterisked. The *Kpn4GAL4* construct from Sandrelli et al. (2001, accompanying article) includes nucleotides 992–1866 (see text). Nucleotides 992–1236 carry regulatory sequences for *nonA* expression that include silencers as well as enhancers, whereas nucleotides 1236–1686 carry enhancers only (Sandrelli et al. 2001, accompanying article). The transcription termination of *dGpi1* in relation to the initiating codon of *nonA* (–424 bp) is shown as an inverted arrowhead. The red arrow indicates the *nonA* transcription start found by 5' RACE (–89 bp) and is 11 bp upstream of the one proposed by Stanewsky et al. (1993). Binding sites are represented with A, C, G, or T plus the IUBS code (K, G or T; M, A or C; N, any base; R, A or G; W, A or T; Y, C or T). Eighty-five differences are present among the analyzed sequences, either within *D. melanogaster* or *D. simulans*. Of these, 78 are single nucleotide substitutions and 7 involve insertions or deletions. In *D. simulans*, five insertions relative to the *D. melanogaster* sequence are present: a single nucleotide in position 1528 of Jones and Rubin’s sequences, two nucleotides in position 1533, 33 nucleotides in position 1533, 21 nucleotides in position 1558, and 13 nucleotides in position 1681. In addition, in *D. simulans* there are two deletions involving two nucleotides in position 1442–3 and eight nucleotides in position 1651–9. The EMBL database accession nos. are AJ296020 for *D. melanogaster* and AJ296021 for *D. simulans*. 
The transformant line number is shown in the left-hand column. The genotypes of F₁ males from the cross T(1;4)9e2-10/In(1)FM7 females × +/Y; ∨/+ males (insert indicated with ∨/+ denotes second or third chromosome balancer) carrying an automosomal copy of the viridis transgene are shown in columns A and B. (A) Number of viable males carrying T(1;4)9e2-10 and the viridis insert. (B) Number of males with FM7 and the insert or males with In(1)FM7 only. (C) Percentage of A males.
ecdysone signal to downstream genes during metamorphosis (DiBello et al. 1991). The amino-terminal BTB motif is embedded within the BR-C core and is alternatively spliced to give four isoforms, Z1 to Z4, each carrying different pairs of zinc fingers and each having slightly different, but overlapping, functions (Bayer et al. 1997; Sandrelli et al. 1997). One of these functions is to regulate the proper attachment of the thoracic musculature, and the Z1 and Z4 isoforms are able to rescue the muscle defects in BR-C mutants (Sandrelli et al. 1997). Note that in Figure 6 it is the Z4 BR-C binding site that is conserved (also in D. simulans). The relationship between thoracic musculature and the courtship song provides a possible rationale for the presence of the BR-C site on a “song gene” such as nonA (Ewing 1977).

We applied a modified version of the McDonald-Kreitman test to inspect variation within and outside these putative binding sites in a number of D. melanogaster and D. simulans sequences (Jenkins et al. 1995). These sequences correspond almost exactly to those that are used to drive expression of GAL4 in the Kpm4GAL4 construct used by Sandrelli et al. (2001, accompanying article; Figure 6). The distal fragment (from −917 to −673 bp upstream of the initiating nonA methionine codon or nucleotides 992 to 1236 in Figure 6) contains dGpi1 coding sequences and carries silencers for nonA expression in the thoracic muscles and enhancers for the visual optomotor response (Sandrelli et al. 2001, accompanying article). The proximal fragment from −673 to −223 (nucleotides 1236-1686 in Figure 6) encodes the C-terminal sequences of dGpi1, plus the intergenic spacer between it and nonA, and carries enhancers for all tissue-specific nonA expression and the optomotor response (Sandrelli et al. 2001, accompanying article).

In fulfilling these roles in nonA regulation, the transcription unit of dGpi1 might be under different selective constraints compared to the intergenic spacer. Indeed, a significant excess of fixed changes relative to polymorphisms in the binding regions of the intergenic, as opposed to dGpi1, sequences was observed, so the presence of dGpi1 is placing constraints on the fixation of adaptive changes in the putative binding sites. The high levels of polymorphism in the binding regions within dGpi1 are consistent with the low codon bias in this gene and would serve as a barrier against adaptive fixation. We realize that this analysis is speculative and will rely on future work to show that these binding sites are biologically relevant. Nevertheless, if the algorithms we used were identifying completely nonfunctional sites, it is difficult to understand why a significant result would be obtained at all with this neutrality test, let alone in the intergenic region only.

Finally, in spite of extensive divergence, the dGpi1 sequence within the D. virilis fragment is nevertheless able to rescue the lethality associated with T(1;4)9c2-10. The nonA sequences contained in the same fragment also rescue the nonA-associated ERG defect caused by the translocation and can be used to study whether nonA carries species-specific song information in these transformants (S. Campesan, Y. Dubrova, J. C. Hall and C. P. Kyriacou, unpublished results). In conclusion, the comparative analysis of nonA has clarified the molecular genetics of this genomic region and revealed some interesting and unusual evolutionary dynamics. These appear to reflect the unique regulatory relationships between dGpi1 and nonA that are identified in the accompanying article (Sandrelli et al. 2001). S.C. thanks the European Community for a predoctoral fellowship. This work was supported by a Human Frontiers Science Programme and Biotechnology and Biological Sciences Research Council (BBRSC) grant to C.P.K., Ministero per Università e la Ricerca Scientifica e Tecnologica (MURST) grants to R.C. and A.M., a MURST studentship for a “dottorato di ricerca” to F.S., a CNPq fellowship to A.A.P., and a Wellcome Trust International Research Development Award to A.A.P. and C.P.K.

LITERATURE CITED


