Conservation of gene architecture and domains amidst sequence divergence in the *hsrω* lncRNA gene across the *Drosophila* genus

[1]Ranjan Kumar Sahu[1], Eshita Mutt[2] and Subhash Chandra Lakhotia[1*]

[1] Cytogenetics Laboratory, Department of Zoology, Banaras Hindu University, Varanasi, INDIA

[2] Current address: Laboratory of Biomolecular Research, Paul Scherrer Institute, 5232 Villigen, Switzerland

[*] Corresponding author

Author emails:

Ranjan Kumar Sahu: xray.sahu@gmail.com (ORCID ID: 0000-0002-7893-1361)

Eshita Mutt: eshi.mutt@gmail.com

S. C. Lakhotia: lakhotia@bhu.ac.in (ORCID ID: 0000-0003-1842-8411)

Running title: *hsrω* lncRNA gene conservation in *Drosophila*

Abstract

The non-coding 93D or *hsrω* locus in *Drosophila melanogaster* is important for development and cell stress response, produces multiple long non-coding RNAs using two known transcription start sites and four transcription termination sites and includes two exons, one *omega* intron, one short translatable open reading frame ORFω, long stretch of tandem repeats unique to this locus and an overlapping miRNA gene near its 3' end. Early cytogenetic and cloning studies suggested its functional conservation in the genus but with little sequence conservation of the ORFω and other regions including tandem repeats while its exon-intron junctions showed ultra-conservation. In this bioinformatic study, we used the landmarks known in *Drosophila melanogaster hsrω* to map the *hsrω* orthologue in publicly available annotated and unannotated 34 *Drosophila* species genomes. This gene is present in all the examined *Drosophila* species with conserved syntenic neighbours. Sixteen and sixty nucleotide long sequences spanning, respectively, the 5' and 3' splice junctions of the *omega* intron are ultra-conserved; other regions show varyingly conserved domains. However, the locus-specific tandem repeats are most variable, although a nonamer sequence within them is conserved. Significantly, the ORFω encoding a short alpha helical peptide, with limited sequence conservation, is present in all species indicating its important role. The *mir-4951*, overlapping with the 3' end of *D. melanogaster hsrω*, was identified in 33 *Drosophila* species. Architecture of *hsrω* lncRNA gene appears more conserved than its base sequence, suggesting that the higher order structures of these lncRNAs are critical for their diverse functions.

Keywords: *lncRNA; ORFω; omega* intron; *mir-4951; splice junction; tandem repeats*

Introduction

Major parts of genomes in diverse eukaryotes are comprised of the non-coding (nc) DNA sequences, a large part of which is transcribed as a variety of short and long non-coding RNAs (sncRNA and lncRNA). Although in the past such genomic sequences were often ignored as 'selfish' or 'junk' DNA, recent years have witnessed a remarkable interest in the ncRNA genes because of their well-documented pervasive involvement in cellular regulatory networks in diverse ways and at different levels, ranging from regulation of local chromatin organization, transcription, post-transcriptional processing, transport and stability of RNAs to modulation of activities of different regulatory proteins in the cell (Anastasiadou, et al., 2018; Fatima, et al., 2015; Ferreira and Esteller, 2018; Hirose and Nakagawa, 2016; Kopp and Mendell, 2018; Lakhotia, 1996; Lakhotia, 2016; Lakhotia, 2012; Lakhotia, 2015; Lakhotia, 2017; Lakhotia, 2017; Lakhotia, 2018; Matsumoto and Nakayama, 2018; Noh, et al., 2018; Quinn, et al., 2016; Ransohoff, et al., 2018; Stoiber, et al., 2016). The ncDNA sequences exhibit very wide range of evolutionary conservation ranging from 'ultra-conservation' across phylogenetic groups to remarkable variability even between related species.

The *93D* or *hsrω* in *Drosophila melanogaster* (*D. mel*), one of the first identified lncRNA genes, is known to be essential for normal development as well as for surviving cell stress (Lakhotia, 1987; Lakhotia and Mukherjee, 1982; Lakhotia, 2003; Lakhotia, 2011; Lakhotia, 2012; Lakhotia, 2017; Mohler and Pardue, 1984). Interest in the stress inducible and developmentally active *93D* locus of *D. mel*, later renamed as *hsrω* gene (Bendena, et al., 1989), was generated following a serendipitous observation that the 93D puff was singularly activated after a brief in vitro exposure of larval salivary glands to 1mM benzamide, which otherwise inhibited the on-going transcription at other chromosomal sites (Lakhotia and Mukherjee, 1970). It was found to be one of the most active heat shock induced genes with unusual turnover of its transcripts (Lakhotia and Mukherjee, 1980; Lengyel, et al., 1980; Mukherjee and Lakhotia, 1979; Peters, et al., 1980), which were apparently not translated (Lakhotia and Mukherjee, 1982).

Cloning and sequencing of this locus from three distantly related species, viz., *D. mel*, *D. hydei* (*D. hyd*) and *D. pseudoobscura* (*D. pse*) (Peters, et al., 1984) confirmed its non-coding nature and revealed a comparable organization of the gene although with high sequence variability. Only certain specific regions of this gene displayed high sequence conservation between the three examined species (Garbe, et al., 1986; Garbe, et al., 1989; Garbe and Pardue, 1986). Subsequent studies established a critical importance of this gene in normal development as well as in cell stress response (Lakhotia, 1987; Lakhotia, 2003; Lakhotia, 2011). This gene is transcribed in nearly every cell type in the life of fly (Bendena, et al., 1991; Lakhotia, 2001; Mutsuddi and Lakhotia, 1995). Of its multiple transcripts, the repeat containing nuclear transcripts are essential for organization of the nucleoplasmic omega speckles, which coordinate the dynamic availability of diverse hnRNPs and some other RNA binding proteins in cells under normal and stress conditions (Lakhotia, et al., 1999; Lakhotia, et al., 2012; Mallik and Lakhotia, 2009; Mallik and Lakhotia, 2011; Onorati, et al., 2011; Piccolo, 2018; Prasanth, et al., 2000; Singh and Lakhotia, 2015).

The *hsrω* gene in *D. mel* is located at the 93D4-5 cytogenetic region on right arm of chromosome 3 (genomic coordinates 3R:21,296,127...21,317,836, http://flybase.org/reports/FBgn0001234). Following the advent of genome sequencing, the *hsrω* gene has been annotated in 12 *Drosophila* species at the FlyBase

(http://flybase.org/cgi-bin/gbrowse2/) while unannotated genome sequences of many more species are available at the NCBI database (https://www.ncbi.nlm.nih.gov/). Based on the early studies (Lakhotia, 2011; Peters, et al., 1982), it appeared that this gene had common architectural features with ~10 to 15kb long transcription unit, whose proximal part (~2.6kb) comprised of unique sequence while the distal region (>5kb) contained tandem repeats of a short sequence, unique to this locus (Garbe, et al., 1986; Peters, et al., 1982; Peters, et al., 1984). Intriguingly, the unique part did not show any significant sequence homology between *D. mel*, *D. hyd* and *D. pse* except for remarkably conserved stretches of 16 and 60 nucleotide at the 5′ and 3′ splice junctions, respectively, of the single intron, which we now term as the *omega* intron. The other identified conserved region was a nonamer (ATAGGTAGG) present at nearly equal intervals in the repeat part of the large nuclear transcript in *D. mel* and *D. hyd* (Garbe, et al., 1986). These early studies also indicated this gene to produce two independent primary nucleus limited transcripts, with the >10kb longer (*hsrω-n*) transcript corresponding to the entire transcription unit and the shorter 1.9kb (*hsrω-pre-c*) transcript being derived from the proximal unique region till the first poly-A signal located about 700 bp upstream of the beginning of the stretch of tandem repeats. Splicing of the single intron (~700 bases long) in the 1.9kb nuclear transcript produced the 1.2kb (*hsrω-c*) cytoplasmic transcript, which carried a potentially translatable open reading frame (ORFω) that could encode 23–27 amino acids in the three examined species (Fini, et al., 1989; Garbe, et al., 1986). Intriguingly, the amino acid sequence potentially encoded by the ORFω in the three species did not appear to show high conservation (Garbe, et al., 1986). Absence of strong conservation of its base sequence in the three examined species of *Drosophila* (Garbe, et al., 1986; Garbe, et al., 1989; Garbe and Pardue, 1986; Peters, et al., 1984; Ryseck, et al., 1987) appeared rather incongruous with its known (Lakhotia, 1987; Lakhotia, 2016; Lakhotia, 2011) critical functions in cell regulation.

Currently the NCBI and FlyBase databases annotate the *hsrω* gene of *D. mel* to produce 7 transcripts (Fig. 1). Among these, the *hsrω*-RA (*hsrω-c*), *hsrω*-RB (*hsrω-n1*), *hsrω*-RC (*hsrω-pre-c*) and *hsrω*-RG (*hsrω-n2*) transcripts have been known earlier (Lakhotia, 2011), but the *hsrω*-RD, *hsrω*-RH, and *hsrω*-RF transcripts are new additions on the basis of RNA-sequence data. These seven transcript variants originate because of different combinations of 2 transcription start sites (TSS), 4 transcription termination sites (TTS) and variable splicing of the single intron (see Fig. 1). The TSS for *hsrω*-RD is named as TSS1 while that for all the other four primary transcripts, viz., *hsrω*-RB (*hsrω-n1*), *hsrω*-RC (*hsrω-pre-c*), *hsrω*-RF and *hsrω*-RH, located 494 bp downstream of TSS1, is named TSS2 (http://flybase.org/reports/FBgn0001234). The FlyBase report for this gene states that no orthologs of this gene are known neither in other species of *Drosophila* nor in any other organism.

Comparative genomic analysis of multiple genomes has remarkably improved the resolution to analyze phylogenetic relationships among different species. Since information about genome sequence for a large number of *Drosophila* species is now available in public databases and since the above noted earlier studies indicated a variable conservation of the *hsrω* lncRNA gene, we undertook the present bioinformatic analysis of base sequence and architecture of the *hsrω* gene in 34 species of *Drosophila*, viz., *D. albomicans* (*D. alb*), *D. americana* (*D. ame*), *D. ananassae* (*D. ana*), *D. arizoanae* (*D. ari*), *D. biarmipes* (*D. bia*), *D. bipectinata* (*D. bip*), *D. busckii* (*D. bus*), *D. elegans* (*D. ele*), *D. erecta* (*D. ere*), *D. eugracilis* (*D. eug*), *D. ficusphila* (*D. fic*), *D. grimshawi* (*D. gri*), *D. hydei* (*D. hyd*), *D. kikkawai* (*D. kik*), *D. melanogaster* (*D. mel*), *D. miranda* (*D. mir*), *D.*

mojavensis (*D. moj*), *D. montana* (*D. mon*), *D. nasuta* (*D. nas*), *D. navojoa* (*D. nav*), *D. novomexicana* (*D. nov*), *D. obscura* (*D. obs*), *D. persimilis* (*D. per*), *D. pseudoobscura* (*D. pse*), *D. rhopaloa* (*D. rho*), *D. sechellia* (*D. sec*), *D. serrata* (*D. ser*), *D. simulans* (*D. sim*), *D. subobscura* (*D. sub*), *D. suzuki* (*D. suz*), *D. takahashii* (*D. tak*), *D. virilis* (*D. vir*), *D. willistoni* (*D. wil*), and *D. yakuba* (*D. yak*). For convenience, the 34 species are referred to in the following using only the first three letters (as in parentheses above) of their species names. We show that some features of the architecture of this long gene are remarkably conserved in all the 34 *Drosophila* species examined although the base sequence of the gene shows a peppered conservation with only certain small regions being significantly similar between the 34 species.

Materials and Methods

Retrieval of genomic sequences carrying *hsrω* orthologue in different species

Since sequences flanking the 5' (16 nt) and 3' (60 nt) splice junctions were shown in earlier studies (Garbe, et al., 1986) to be nearly completely identical between *D. mel*, *D. hyd* and *D. pse* species, we used these as query sequences to search for homologous stretches in the 34 *Drosophila* species' genome sequences available, as on 17 Nov 2018, at the NCBI and FlyBase genome databases (12 species being common in NCBI and FB2017_05 October (Dmel Release 6.18) using the online BLAST NCBI (http://www.ncbi.nlm.nih.gov/BLAST/) and BLAST FlyBase (http://flybase.org/blast/) programs. Only the genome sequences that showed presence of both the 5' (16 nt) and 3' (60 nt) splice junction sequences in neighbourhood were selected for further analysis.

The *hsrω* gene in *D. mel* carries a miRNA gene, *mir-4951*, located ~971 bp upstream of the annotated TTS4 (Fig. 1) at the end of its largest transcript *RF* (http://flybase.org/reports/FBgn0001234). Our initial survey revealed that this miRNA sequence is present downstream of the *hsrω* orthologue in all the other 11 species whose genomes are annotated at the FlyBase. Therefore, we queried all the 34 species' genomes with the *mir-4951* sequence of *D. mel* to retrieve its flanking sequences.

The above two sets of retrieved sequences were aligned with Refseq Genome Database **(**Refseq_Genomes) and Whole-genome shotgun contigs (WGS) databases. In most cases, we could retrieve at least 36kb of contiguous genomic sequences (see Results) that covered 2kb upstream and 34kb downstream of the identified 5' splice junction, and used the same for further analyses.

Sequence annotation and mapping

An offline tool SnapGene® (GSL Biotech, snapgene.com) was used for sequence annotation and mapping of different regions of the *hsrω* genes of the 34 *Drosophila* species.

Splice site prediction

The region starting with GT (12th and 13th nt) in the 16 nt 5' junction and ending with AG (39th and 40th nt) in the 60 nt 3' junction (Garbe, et al., 1986) in the splice junction homologous sequences retrieved for each species was annotated as intron. This was further confirmed by BDGP Splice Site Prediction by Neural Network tool (http://www.fruitfly.org/seq_tools/splice.html) and Net Gene 2 predictor program

(http://www.cbs.dtu.dk/services/NetGene2/) using the retrieved sequences from 34 *Drosophila* species genomes as input. Sequences were searched against *Drosophila* as well as *Human* splice site database for the forward strand with a minimum score of 0.4.

Promoter and TSS prediction

Since the *hsrω* gene of *D. mel* has two transcription start sites (TSS) in the proximal region of the gene (Fig. 1), we used the BDGP Neural Network Promoter Prediction program (http://www.fruitfly.org/seq_tools/promoter.html) to search for potential promoter and TSS in 2kb region upstream of the identified 5' splice junction in the retrieved genomic sequences. A minimum score of 0.8 was set as default for search against eukaryotic promoter database with search restricted to the forward strand of the putative *hsrω* gene in each species.

PolyA and 3' end cleavage site prediction

The polyadenylation and transcription termination sites (TTS) in the retrieved sequences were searched using the online tool PolyAh, which recognizes 3'-end cleavage and polyadenylation region (Salamov and Solovyev, 1997) (http://www.softberry.com/berry.phtml?topic=polyah&group=help&subgroup=promoter).

Search for ORFs, including the ORFω orthologue

ORF finder (https://www.ncbi.nlm.nih.gov/orffinder/) and SnapGene® tools were used to search for potential orthologue of the *D. mel* ORFω in the retrieved putative *hsrω* genomic sequences of other species using default parameters that the potential ORF must start with ATG, must be coding for at least 20 amino acids, should be in forward orientation and lie in between TSS2 and 5' splice junction.

Protein secondary structure prediction

In order to study the structural similarities of the putative peptide produced by the ORFω, the predicted amino acid sequences were analysed for their secondary structures using PSIPRED (Jones, 1999) (http://bioinf.cs.ucl.ac.uk/psipred/) tool.

Estimation of repetitive sequences

The length and frequency of tandem repeats in the retrieved genomic sequences were analysed using the Tandem Repeat Finder tool (Benson, 1999) (http://tandem.bu.edu/trf/trf.html) with a minimum period size set to 7 nt and maximum to 500 nt.

Evolutionary conservation analysis

Nucleotide and amino acid sequences from different regions of *hsrω* gene in different species were aligned and percent identity matrix (PIM) was calculated with help of the multiple sequence alignment (MSA) tool T-Coffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/), which provided conservation colour codes as BAD AVG GOOD;

these were used for all alignment tables and figures. Evolutionary relationships were inferred using the Maximum Likelihood method with Mega X software (Kumar, et al., 2018).

Results

1. Early cytogenetic identification of *hsrω* orthologs in different species of *Drosophila*

The unique inducibility of one of the major heat shock puffs in *D. mel* with benzamide (Lakhotia and Mukherjee, 1970; Mukherjee and Lakhotia, 1982), presence of large RNP particles at this puff (Dangli, et al., 1983) and accumulation of several hnRNPs at this site during heat shock (Saumweber, et al., 1980) have been utilized in early cytogenetic studies to identify an orthologue of *93D* or *hsrω* gene in several species of *Drosophila*. In *D. hyd* and related species, the same locus (2-48C) was also known to be selectively induced by brief in vitro exposure of larval salivary glands to vit-B6 or pyridoxal phosphate (Leenders, et al., 1973), although such induction was not seen in *D. mel* (Lakhotia and Singh, 1982). Table 1 summarizes the earlier cytogenetic evidence for existence of *hsrω* orthologue in many *Drosophila* species. These studies showed that a gene sharing the unique inducible properties of *D. mel hsrω* was associated, in all the examined species, with ancestral chromosome element E (Muller, 1940). These cytogenetic studies clearly suggested that a functional orthologue of *hsrω* gene is present in diverse *Drosophila* species although results of the few early DNA:DNA or RNA:RNA in situ hybridization studies (Table 1) and direct sequencing of the cloned regions of this gene in a few species (Drosopoulou, et al., 1996; Garbe, et al., 1986; Garbe, et al., 1989; Peters, et al., 1984) revealed a rather limited conservation of DNA sequence at this locus.

2. Genomes of all *Drosophila* species carry *hsrω* orthologue and show a conserved order of flanking genes

As noted in Material and Methods section, we searched the genomes of 34 *Drosophila* species for presence of the conserved 16 nt 5' and 60 nt 3' splice junction associated sequences, and for the *mir-4951* sequence. Significantly, the two splice junction associated sequences were found in neighbourhood only once in each of the 34 species genomes. In 26 species, the retrieved sequences also carried the *mir-4951* orthologue downstream of the 3' splice junction sequence. However, in eight species (*D. ame, D. eug, D. moj, D. nas, D. obs, D. rho, D. sec*, and *D. suz*), whose genomes are not yet fully assembled, the retrieved scaffolds carrying the two splice junctions did not include the *mir-4951* sequence because the region downstream of the 3' splice junction or upstream of the *mir-4951* in the identified scaffolds was small (Fig. 2). In each of these cases, the scaffold that carried the *mir-4951* orthologue was blasted with the scaffold carrying the 5' and 3' splice junctions to find any overlap. In the case of *D. ame* and *D. suz*, the splice junctions, and the *mir-4951* sequence carrying scaffolds showed overlapping regions (52 bp in *D. ame* and 146 bp in *D. suz*), using which we assembled contigs for the two species. In case of the other six species (*D. eug, D. moj, D. nas, D. obs, D. rho*, and *D. sec*) the two scaffolds did not show any overlap with each other (Fig. 2), or common overlap with any other scaffold. Consequently, the complete *hsrω* region could not be assembled for these six species. A contiguous stretch of ~36kb genomic sequence (2kb upstream and 34kb downstream of the identified 5' splice junction) was used for further analyses in the 28 species while for the other six species, the two separate scaffolds, carrying the splice junctions and *mir-4951* sequence, respectively, were used.

Using the annotated *Drosophila* species genome data at the GBrowse of FlyBase, we confirmed that the genomic coordinates of the retrieved sequences matched, in species where information was available, with the earlier identified chromosomal locations of *hsrω* orthologue (Tables 1 and 2). This strongly indicated that the retrieved sequences indeed included whole or part of the *hsrω* orthologue.

We aligned (Fig. 3) the retrieved sequences/scaffolds for 12 species (*D. mel, D. pse, D. ere, D. gri. D. moj, D. per, D. sec, D. sim, D. vir, D. wil,* and *D. yak*) with the annotated genome data available at FlyBase Dmel Release 6.18 (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/). The retrieved sequences could be associated with annotated genes in 8 of these species, which we rename as the *hsrω* gene (Fig. 3). In the other four species (*D. wil, D. per, D. sec,* and *D. ere*), the retrieved scaffolds were not yet associated with any gene name. In view of the sequence homology, we putatively identified these scaffolds to carry the *hsrω* orthologs in these species (Fig. 3F, G, J, and K).

To further confirm identity of the retrieved genomic sequences with *hsrω*, we examined if genes known to flank the *hsrω* locus in *D. mel* also flanked the genomic sequences retrieved for the other species. The *hsrω* locus in *D. mel* is flanked on the upstream by *SIFaR* and *ETHR* genes and on downstream by *CG16791, mod, tin* and *bap* genes (http://flybase.org/reports/FBgn0001234). Sequences of these genes were blasted with the genomes of 34 species and their association with the retrieved putative *hsrω* orthologue examined. In the six species, for which a contiguous stretch encompassing the conserved splice junctions and the *mir-4951* could not be assembled, the up-stream and down-stream genes were searched in reference to the *omega* splice junction and *mir-4951* carrying scaffolds, respectively. Data in Table 2 show the annotated orthologue and the order of some of the up- and down-stream genes that were found to flank the putative *hsrω* sequences in the examined species genomes. Order of the genes upstream of the *hsrω* is identical with that in *D. mel* in 29 species and that of downstream genes in 31 species. In the case of *D. ana,* the order of the downstream *mod* and *tin* genes is reversed (Tables 2 and 3). The flanking genes could not be identified in the scaffolds retrieved for *D. ame*, *D. bus*, and *D. nas*.

GBrowse (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/) lists two additional ncRNA genes downstream of TSS2 in *D. sim* and one in *D. moj* (Tables 2, Fig 3-C, 3-H) in the same direction as *hsrω*. Based on sequence homology revealed by our pair-wise alignment (data not presented), these additional ncRNA genes appear to correspond to parts of the *hsrω* gene. Two additional ncRNA genes were found to overlap the *hsrω* locus in opposite direction in *D. pse* (Fig 3D). One more ncRNA gene was found in *D. moj* to overlap with the *SIFaR* locus (Table 2). Sequence analysis suggested this to be a gene different from *hsrω*. In *D. gri* genome, an orthologue of *mod* is not annotated. However, the *GD15087* gene located at the corresponding location and annotated as ncRNA gene in *D. gri* actually has a long ORF (including the intronic region) with 94% sequence similarity with the *mod* gene of *D. mel* (data not shown).Therefore, the *GD15087* gene of *D. gri* appears to be an orthologue of the *mod* gene of *D. mel*.

The distances between the various genes flanking the putative *hsrω* varied in different species and in some cases, their order and/or orientations also differed, indicating local gene rearrangements (Table 3).

3. Sequences flanking the 5' and 3' splice junctions in the *hsrω* gene are ultra-conserved

The MSA analysis revealed that the 16 nt sequence flanking the 5' splice junction and the 60 nt sequence flanking the 3' splice junction in the proximal region of *hsrω* are very strongly conserved in all the 34 species examined, with conservation scores of 996 and 985**,** respectively (Table 4). Length of the intron flanked by these exon-intron junction sequences varied between 643-785 bp, generally agreeing with the 712 bp length in *D. mel hsrω* gene (Table 4). Compared to the highly conserved 5' and 3' splice junction flanking sequences, the intron sequences in the 34 species are less conserved, with a score of 549. As seen in Fig. 3, at least one of the transcripts in each of the species, for which the *hsrω* transcripts are annotated, carries the *omega* intron at the location identified by our bioinformatic search.

Additional introns are present in the *hsrω* orthologue in other *Drosophila* species. Among the eight *Drosophila* species for which information about transcripts of the *hsrω* gene is annotated at the FlyBase (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/), only the *D. mel hsrω* gene carries a single intron, seven are shown to produce diverse transcripts with variable numbers of introns (Fig. 3). A search for splice-sites in the putative *hsrω* gene region in other species also revealed additional putative intronic sequences. None of these additional introns were found to be flanked by the 16 nt 5' and/or 60 nt 3' splice junctions. In view of the unique features of the single intron in the *D. mel hsrω* gene and its unique conservation across the *Drosophila* species, we designate this as the *omega* intron.

Variations of the conserved 5' or the 3' *omega* splice junction sequences were also individually detectable in a few other genes in *D. mel*: eighty two genes (Supplementary Table S1) showed presence of only the 5' splice junction sequence with 69% to 87% similarity. Interestingly, 41 of these carried the 5' junction sequence in anti-sense orientation while in the other 41 genes it was in sense orientation (Supplementary Table S1). Only 2 genes, different from those sharing the 16 nt *omega* splice junction sequence, showed 33-36% similarity with the *hsrω* 3' splice junction. A survey of known/predicted functions of these genes at the FlyBase indicated that they are involved in diverse functions without any apparent commonality. It is also important note that the 5' and 3' splice junctions of the *omega* intron did not occur together in any other gene in any *Drosophila* species.

4. *mir-4951* located upstream of the last annotated TTS of *D. mel hsrω* locus shows conserved presence and location

The *hsrω* locus in *D. mel* genome is annotated at the FlyBase to include a putative miRNA gene, *mir-4951*, located ~971 bp upstream of the annotated TTS4 (end of the largest transcript *RF* (Fig. 1)). The *mir-4951* homologous sequence is present downstream of the *omega* intron sequence in all the other annotated genomes of *Drosophila* species, viz., *D. ana*, *D. ere*, *D. gri*, *D. moj*, *D. per*, *D. pse*, *D. sim*, *D. vir*, *D. wil*, *D. yak* and *D. sec*, although the *hsrω* locus in *D. sec* is present in two separate scaffolds. In the case of *D. ana*, the *mir-4951* sequence is part of the known *hsrω* transcripts like in *D. mel* (Fig. 3). Unlike the case in *D. mel*, the *mir-4951* gene is so far not annotated as a separate gene in any of the other annotated *Drosophila* species genomes, although as noted above, its sequence is present in all of them.

In 16 other species, whose genome sequences were available at the NCBI data base, we could identify the *omega* intron and *mir-4951* sequence on a single sequenced or assembled contig. In the remaining six species (*D. eug*, *D. mon*, *D. nas*, *D. obs*, *D. rho*, and *D. sec*), the two scaffolds showing regions of homology to the

omega intron and *mir-4951*, respectively did not show any overlap so that a contiguous region including the splice junctions and the *mir-4951*, as in other species, could not be assembled. Since the conserved *omega* intron and the *mir-4951* are present within <34kb in 27 *Drosophila* species, we believe that the separate contigs identified in *D. eug, D. mon, D. nas, D. obs, D. rho,* and *D. sec* include the proximal and distal parts, respectively, of the *hsrω* gene. This is further supported by our finding that the genes upstream and downstream of the *omega* intron and *mir-4951* homologous sequence, respectively, were same in all the species, except in *D. ame, D. bus* and *D. nas* (Tables 2 and 3). Therefore, keeping the *D. mel hsrω* as model, we suggest that the 3' boundary of the *hsrω* locus in other *Drosophila* species is close to the 3' end of *mir-4951* sequence (Figs. 2, 3 and 4). It is expected that further refinements in completion and curation of the genomic sequences in these species would provide more definitive information.

In the case of *D. bus*, sequence for the full *mir-4951* was not recovered in the entire genome sequence, including the 26 Mb long CP012526.1 *D. bus* chromosome 3R scaffold, which harboured the *omega* intron sequence. Surprisingly, a 45 bp region, located ~24Mb upstream of the *omega* intron in the *D. bus* chromosome 3R scaffold CP012526.1 showed similarity with the 3' region of *mir-4951*. It is not clear if this disparity is real or due to some inaccuracies in sequence assembly.

The *D. mel mir-4951* gene is 150 bp long; in other species its length ranges between 131-160 bp. The MSA analysis in 33 species revealed a good conservation with a score of 591, with two regions of the gene showing higher conservation (Table 5). The first region (21 bp), corresponding to the processed 5' *mir-4951* of *D. mel*, showed high conservation score of 991 in 31 species but appeared more divergent in *D. nav* and *D. nov*. The second region (23 bp) corresponding to the processed 3' *mir-4951* had a conservation score of 897 in 32 species (Table 5), being more divergent in *D. mon*. These two regions represent the functional mature miRNAs from the locus.

5. An orthologue of translatable short ORFω is present in all *Drosophila* species

Earlier studies (Garbe, et al., 1986; Garbe, et al., 1989; Ryseck, et al., 1985) showed that the 1.2kb spliced transcript *RA* or *hsrω*-c derived from the proximal region of the *hsrω* gene was cytoplasmic in *D. mel, D. hyd.* and *D. pse*, and that one of the short ORFs, the ORFω, in this transcript was translatable (Fini, et al., 1989). Therefore, we searched for potential ORFs in proximal part of the *hsrω* gene. This analysis suggested presence of several putative ORFs, a total of 11 in the entire gene, including 5 ORFs (≥20aa) within the 1.2kb cytoplasmic transcript in *D mel* (Supplementary Table S2).

MSA of the different ORFs identified by the above search using T-Coffee revealed that only the first predicted ORF, located upstream of the 5' splice junction in each of the 34 species, showed significant degree of conservation with the ORFω of *D. mel* (Table 6). Its location with respect to the TSS2 (see below) and the 5' splice junction was also comparable in different species. The other putative ORFs in the proximal region (till the end of exon 2 in *D. mel*) of the *hsrω* gene in different species showed high variability in their amino acid sequence and locations; many long ORFs (>50aa) were also found in the intron and a few in exon1 and exon2 of the putatively identified *hsrω* gene region; however, their locations and/or sequences did not show any conservation between the species (Supplementary Table S2).

Predicted amino acid sequence of the putative peptide-ω encoded by the ORFω in 34 species (Table 6) showed a conservation score of 627 (Fig. 4). Amino acids in its terminal regions appeared more conserved than those in the middle. The first four N-terminal residues in 27 species are MEKC, MKKC, MQKC, MQMC or MEMC (Table 6). The predicted amino acid sequence of the peptide-ω in 4 species (*ame*, *mon*, *nov* and *vir*) includes a few extra residues prior to the MEKC/MKKC/MQKC/MQMC/MEMC tetrad motif, found as the first four residues in the other 27 species. In 3 other species (*ana*, *bip*, and *bus*), extra residues are seen after the first M of the above tetrad. Interestingly, if these extra residues are ignored, the first four N-terminal residues in these seven species would also be similar to those in the other 27 species (Table 6). Therefore, these additional amino acid residues in the predicted peptide-ω in the seven species need further verification of the genomic sequences.

The C-terminal amino acid sequence is "G/TPT" in 21 species and "RRLK" in 12 species, but the peptide-ω ends with "AVA" in *D. wil* (Table 6).

Secondary structure analysis of the putative peptide-ω in different species revealed that a single alpha helix covers >80% of the peptide in 18 species while a shorter alpha helix, covering <50% of the peptide, is seen in 10 species. Two helices covering >70% of the peptide are predicted in 5 species (Table 6). The peptide-ω in *D. ser* has a short alpha helix and a beta pleated structure while in *D. eug* only two beta pleated structures are predicted (Table 6).

6. Greater conservation of promoter at TSS2 than at TSS1 and possibility of a novel promoter at ORFω

Initial studies (Garbe, et al., 1986; Garbe, et al., 1989; Peters, et al., 1984; Ryseck, et al., 1985) on transcripts of *hsrω* gene identified only one transcription start site (TSS) for the more than one *hsrω* primary transcripts in *D. mel*, *D. hyd* and *D. pse*. However, recent annotations at the FlyBase (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/) suggest two TSS for *D. mel*, only one TSS in five species (*D. ana*, *D. pse*, *D. vir*, *D. yak* and *D. gri*) but more than two in *D. moj* and *D. sim*. Intriguingly, additional ncRNA genes have also been annotated to overlap the *hsrω* locus in *D. sim*, *D. moj* and *D. pse* in sense and/or anti-sense orientations (Fig 3).

We examined the proximal region (~2 kb upstream of the identified 5' *omega* splice junction) of the retrieved *hsrω* homologous sequences for presence of promoter and TSS, in sense orientation, using the BDGP promoter search program.

The TSS1 in *D. mel* corresponds to the site from where the *RD* transcript is initiated (Fig. 1). Promoter search analysis suggested presence of a promoter at the corresponding site in 31 species, other than *D. alb*, *D. tak* and *D. ser*. The predicted TSS1 promoter showed a low degree of sequence conservation between the 31 species (Table 7A). The region between TSS1 and beginning of exon-1 (beginning at the TSS2, see below) in *D. mel* was annotated as UDR (Unique to RD transcript region; Fig. 1); lengths of its putative orthologue in other species ranged from 136-811 bp (Table 2). It may be noted that among the 12 species' *hsrω* gene annotated at FlyBase, only *D. mel* is indicated to have an active TSS1 (Fig. 3 and Table 7C).

Search for homology for the TSS2 of *D. mel* revealed presence of a promoter at ~115-130 bp upstream of the ORFω in 31 species and at ~160 bp upstream in 3 other species (Table 2). Multiple sequence alignment of the

predicted TSS2 promoters of the 33 species revealed a very high homology score with that of *D. mel*, ranging from 0.86-1.00 in different species (Table 7B). The ambiguous sequence available for this region in *D. ana* (81Ns noted in the shotgun sequence scaffold, http://flybase.org/cgi-bin/gbrowse2/dana/scaffold_13340:3,782,971...3,783,051), precluded a definitive identification of the TSS2 promoter in this species. However, the GBrowse transcript annotation data (Fig. 3B) suggests presence of a functional promoter at 74 bp upstream of the ORFω in this species, which partially matched with the putative TSS2 identified in other species (Table 7B).

A 14 bp region (GGTATAAATAGAGC) located 18 bp upstream of the annotated TSS2 in *D. mel* is highly conserved and based on its sequence, this may function as the TATA box for TSS2. The 477 bp region starting from TSS2 and ending at the 5' splice junction is annotated as exon 1 in *D. mel*, and correspondingly, the region between the predicted TSS2 and the 5' splice junction is named as exon 1 in the other 33 species, with its length varying from 414 to 552 bp. Although exon-1 as a whole showed only moderate conservation with a score of 517, a 45 bp region at the 5' end of exon1 showed high conservation; another stretch of 37 bp, located 140 bp downstream of the ORFω in *D. mel* exon-1 also showed good conservation in different species (Fig. 4).

A comparison of the relative locations of the annotated and predicted TSS1 and TSS2 in eight species whose genome annotation data were available at the FlyBase (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/) revealed that in most cases the coordinates of predicted TSS2 matched with those of the annotated transcript sites (Table 7C). As noted above, the TSS1 is annotated only in the case of *D. mel hsrω*, which fully matched with the predicted TSS1 I in this species. The TSS2 is annotated for all the 8 species and interestingly, in five of them the predicted TSS1 nearly matched with the respective annotated TSS. In the case of *D. ana*, this match could not be ascertained because of ambiguity in the available sequence for this region. In the cases of *D. pse* and *D. gri*, the *hsrω* gene region annotations at http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/ suggest the transcription to start downstream of the predicted TSS2 (Fig. 3). Consequently, the annotated and predicted TSS2 in these two species do not match. However, in view of the high similarity between the predicted TSS2 in *D. pse* and *D. gri* with that in other species, we believe that further curation of the *hsrω* transcripts in these species is needed.

We suggest that the region between TSS1 and 5' splice junction of the *omega* intron in *D. mel* be named as exon 1a since its 3' end is same as that of the exon 1. It, however, remains to be seen if a functional TSS1 and a transcript corresponding *D. mel* to *hsrω-RD* is indeed present in other species.

Our promoter search revealed a new potential transcription start site, which has so far not been annotated in *D. mel* or any other species. The core promoter of 50 bp for this potential third TSS (TSS3), predicted by the BDGP promoter search tool, overlaps with the ORFω (Supplementary Table S3) and is present in 23 species with a score of 0.82-0.99. This predicted promoter sequence shows moderate conservation with a score of 432 (Supplementary Table S3). It may be noted that although a TSS has not been annotated at this region in *D. mel* or any other species at the FlyBase, two cDNA sequences (BT122054.1 and M13240.1) in the NCBI database (Garbe and Pardue, 1986) start 93 nt and 26 nt, respectively, upstream of this predicted promoter (Fig. 1).

7. First transcription termination site and exon 2 region are variably conserved

The FlyBase has annotated 4 transcription termination sites (TTS) on the *hsrω* gene of *D. mel*. The first TTS (TTS1) is present 699 bp downstream of the 3' splice site of the *omega* intron and this region is annotated as exon 2 in *D. mel*. The PolyAh output results suggested that a TTS is present between 667 to 799 bp downstream of the 3' splice junction in the *hsrω* gene in all the 34 species. In keeping with the annotation in *D. mel*, we named this TTS as TTS1 and the region between the *omega* intron's 3' splice site to TTS1 as exon 2 (Fig. 4). However, the predicted TTS1 matched with the known TTS at the end of exon 2 only in *D. mel*. In other species for which transcripts are indicated at the FlyBase (Fig. 3), the locations of predicted TTS1 and the annotated TTS1 did not match well. Nevertheless, we considered the sequence between 3' splice site of the *omega* intron and the predicted TTS1 as exon 2 in all the species. Despite the comparable length of exon 2 in different species, its overall conservation score was relatively poor with a score of 543, although some regions along its length showed a higher degree of conservation. These included 75 bp and 30 bp domains, located 199 bp and 379 bp, respectively, downstream of the 3' splice junction in *D. mel hsrω* gene. A 147 bp stretch, 458 bp downstream of the 3' splice site also showed a moderate conservation in different species (Fig. 4).

8. Multiple transcription termination sites may exist in *hsrω* orthologs in other species

We used PolyAh and other tools like ARNold (http://rna.igmors.u-psud.fr/toolbox/arnold/index.php), PolyA Signal Miner (http://dnafsminer.bic.nus.edu.sg/PolyA.html), POLYAR (http://cub.comsats.edu.pk:8080/polyar.htm) PASIGNAL (http://cub.comsats.edu.pk:8080/GEFWebServer.htm) to identify potential TTS. The PolyAh analysis predicted a novel TTS at a distance of 20-27 bp (Table S4) upstream of the TTS1 in all species except *D. wil*. None of the other TTS prediction tools definitively predicted other TTS sites, including those that are annotated for eight species at the FlyBase (Fig. 3). Search for other TTS using the PolyAh tool generated very large numbers of potential termination sites in all the species, with little conservation in terms of their locations. The only other TTS consistently predicted by the PolyAh tool was located downstream of the *mir-4951* in 17 species (*D. ana, D. bip, D. ere, D. gri, D. mel, D. mir, D. moj, D. obs, D. per, D. pse, D. rho, D. sec, D. ser, D. sim, D. sub, D. suz and D. tak*). In the other species, a TTS could not be definitively identified downstream of the identified *mir-4951* sequence.

9. Tandem repeats: the most diverse region of *hsrω* gene

A major part of the *D. mel hsrω* gene contains tandem repeats of 278 bp (generally denoted as 280 bp repeat), unique to this gene and which account for ~50% of the 21.7 Kb long gene. These are located between the TTS2 and TTS3 of the *D. mel hsrω* (Figs. 1 and 4).

Tandem Repeat Finder tool revealed variable sizes and numbers of repeats downstream of the TTS1 in the other 25 species. Intriguingly, tandem repeats were not found in the identified *hsrω* orthologue sequences in 8 species while the absence of genomic sequences for an unknown length of middle part of the identified *hsrω* gene region limited the search for tandem repeats in the case of *D. nas*. In many cases, e.g. *D. per, D. ser, D. rho*, the repeats were not tandemly arranged since some unique sequences intervened between the repeats. The repeat lengths and the number of copies of tandem repeats in 25 species are shown in Table 8. In most species, the larger repeat units were variants of tandem repeats of one or more of smaller repeat units. It is significant that these repeats were found only in the identified *hsrω* gene region in each of the 25 species. Interestingly, the base

sequences of these repeats showed considerable variations between species, resulting in a very low conservation score (Fig. 4). Significantly, however, a nonamer sequence (ATAGGTAGG) that was reported (Garbe, et al., 1986) to be repetitively present in the otherwise diversified tandem repeats in *D. mel* and *D. hyd* was seen to be present with varying frequencies in 29 species, including some of those in which typical tandem repeats were not identifiable (Table 8). *D. wil* did not show the classic ATAGGTAGG nonamer sequence but had 58 copies of a variant nonamer (ATAGGCAGG). Variants were also seen in some species which showed presence of the classic nonamer (Table 8). Intriguingly, neither the classical nor variants of the nonamer were found in the *hsrω* sequence in *D. ana*, *D. ari* and *D. nav*. Presence of repeats or the nonamer motifs could not be ascertained in the *D. nas hsrω* gene, possibly because the retrieved sequence was incomplete.

10. Very low conservation immediately upstream and downstream of the tandem repeat region

The unique sequence regions immediately upstream (UBR, Unique region upstream of the tandem repeats between TTS1 and TTS2 in *D. mel*) and downstream (UAR, Unique region following the stretch of tandem repeat till the TTS3 in *D. mel*) regions flanking the tandem repeat sequences were also examined for conservation in different species. Since the UBR in *D. mel* is 888 bp long, a similar length of sequences downstream of the TTS1 in each of the 34 species was examined by MSA. This revealed that the proximal ~50 bp sequence was highly conserved while the middle region was much less conserved and the ~200 bp at its 3' end showed no conservation at all (Fig. 4). In contrast, a similar MSA analysis for the UAR sequence of *D. mel* in different species showed its conservation only in the 4 species closely related to *D. mel*. The *D. sec* and *D. sim* showed ~70% while *D. yak* and *D. ere* showed ~20% conservation for the UAR region. Little conservation was seen in other species.

MSA of the UFR (Region unique to the *hsrω-RF* transcript, downstream of TTS3 and upstream of the *mir-4951* in *D. mel*) revealed detectable similarity only in 3 species, viz., *D. sim* (~80%), *D. sec* (~70%) and *D. ere* (~30%). In view of such limited conservation across the 34 species, the conservation heat maps for the UAR and UFR could not be prepared.

11. Evolutionary relationship of *hsrω* gene in different species of *Drosophila*

Based on the above analyses, a schematic of the *hsrω* gene, beginning at 2 kb upstream of the 5' splice junction of the *omega* intron and ending with the *mir-4951* gene sequence at the 3' end, in the 34 examined species of *Drosophila* is presented in Fig. 5. We have used the *D. mel hsrω* gene architecture as the base, and accordingly, have tentatively modelled the *hsrω* gene in all species to end with the *mir-4951*. It may, however, be noted that among the 12 annotated *Drosophila* species genomes available at the FlyBase (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/) indicate, the longest *hsrω* transcripts in 6 species terminates much before the *mir-4951* sequence (Fig. 3). Only in the cases of *D. mel* and *D. ana* the longest annotated transcript from this gene covers the *mir-4951* sequence. The transcription start site annotated at GBrowse for the *D. pse hsrω* orthologue differs from that predicted by our bioinformatics analysis. We believe that as more exhaustive genome and transcriptome data for different species become available, a better definition of the *hsrω* gene and its transcripts would be possible. Till that time, we prefer to use the *D. mel hsrω* gene's architecture as the model.

In order to find evolutionary relatedness of the *hsrω* gene in *Drosophila* species, we used sequences from different regions (UDR, Exon-1, Intron, Exon 2, ORFω, UBR, Repeats, and *mir-4951*) to generate phylogenetic trees (Fig. 6). MEGA X software was used to generate the phylogenetic tree following MSA by T-Coffee tool. MEGA X, uses Maximum Likelihood (ML) algorithm Tamura Nei model with nearest neighbour interchange, (Tamura and Nei, 1993) to construct the topology. The 34 species examined in our study belong to 8 sub-genus/species groups, viz., *melanogaster, replete, Hawaiian, immigrans, obscura, virilis, and willistoni, Dorsilopha*, (O'Grady and DeSalle, 2018). The species names in Fig. 6 have been color-coded to indicate these sub-genera/species groups. Sixteen of the 34 species examined in this study belonged to the *melanogaster* group.

For each analysis, branch support was calculated by bootstrap analysis consisting of 500 replicates to generate phylogenetic trees of different regions. Except for members of the *melanogaster* species group, those of the other groups tended to be bunched together in the phylogenetic maps for all the reasons, except the repeat region (Fig. 6G). Members of the more diverse *melanogaster* species group were in most cases, however, clustered in two or even three separate groups although each of these clusters included species that are generally considered to be phylogenetically closer. Significantly, in the case of phylogenetic map of the *omega* intron (including the ultra-conserved splice junction regions), all the 16 *melanogaster* group species were clustered in connected groups, generally matching their phylogenetic relationships (Fig. 6C).

We also analysed percent identity matrix (PIM) to examine the divergences at different regions of the *hsrω* gene in *Drosophila* species (Fig. 7). The exon-1, *omega* intron, ORFω, exon2 and the *mir-4951* show higher conservation within species subgroups. The exon-1 and exon-2 regions in the *melanogaster* species group appear to share greater similarity with the *obscura* group than with other species. Within the *melanogaster* species group, *D. ana* and *D. bip*, which belong to *ananassae* sub-group, show greater divergence for all regions from other species in this group. As revealed by the phylogenetic tree analysis (Fig. 6), the PIM analysis also showed least conservation of the repeats associated with this gene.

Phylogenetic trees and the PIM show that the sequences for different regions of the *hsrω* gene have diversified rather independently. Further, while within a species group there is greater similarity, the divergence between phylogenetically more distant species is variably greater for different domains of the *hsrω* gene.

Discussion

Early cytogenetic and molecular studies indicated that the non-coding *93D* or the *hsrω* locus of *D. mel* is functionally conserved in diverse species of *Drosophila* with a similar genomic architecture but rapidly diversifying base sequence (Garbe, et al., 1986; Garbe, et al., 1989; Garbe and Pardue, 1986; Lakhotia, 1989; Lakhotia and Singh, 1982; Lakhotia, 2003). Several of the species that were examined in earlier cytogenetic studies for presence of the benzamide and heat shock inducible orthologue of the *93D* locus of *D. mel* (Table 1) are also included in the present bioinformatic search for the *hsrω* orthologue. In all cases, the present bioinformatic analysis identified the same loci that were identified earlier as the *93D* puff orthologue. This confirms that the genomic sequences retrieved by us are indeed orthologue of the *hsrω* gene of *D. mel*. This is further substantiated by the finding that in 31 of the 34 examined species, the genes flanking the retrieved

putative *hsrω* sequences were the same. The remarkable synteny of the *hsrω* locus with its flanking genes is important since the 8 MB region surrounding this set of genes is associated with the *In(3R)Payne*, a common cosmopolitan inversion in *D. mel* whose polymorphism is correlated with geographic clines and, therefore, this region has also been referred to as a '*supergene*' (Anderson, et al., 2003; Durmaz, et al., 2018; Kennington, et al., 2006). It is significant that the genomic region associated with *In(3R)Payne* is also associated with adaptive clinal inversion polymorphism in *D.sub* and *D. buzzatii* populations (Hoffmann, et al., 2004). Since the *hsrω* is an important member of this 'supergene' complex in conferring thermal adaptation in clinal populations of *D. mel* (Collinge, et al., 2008; McColl, et al., 1996; McKechnie, et al., 1998; Rako, et al., 2007), the observed commonality of the flanking genes in other species of *Drosophila* may indicate comparable roles for the *hsrω* orthologue in thermal ad other adaptive features in other species as well. Our inability to identify these genes to flank the putatively identified *hsrω* orthologue in *D. ame, D. bus* and *D. nas* is likely to be due to limitations of the genome sequence as presently available, although at this point we cannot rule out some genomic rearrangements in these species during their diversification. Incomplete curation and/or inaccuracy in the assembly of genome sequences may also account for the identification of more than one *hsrω* orthologue in some species and for our inability to assemble a contiguous stretch representing the *hsrω* gene sequence in a few species. Our analysis also suggests that the *GD15087* in *D. sim* is orthologue of *mod* and the annotation of its orthologue in *D. gri* should be changed from non-coding to a protein coding gene.

Our promoter analysis suggests that the promoter at TSS2 of *D. mel* is the major functional promoter as this shows higher conservation and most of the *hsrω* transcripts in this and the other species, whose transcriptomic information is available at the FlyBase, are initiated from this site. Our bioinformatics analysis predicts existence of the TSS1 promoter in corresponding region in most other species, although the current annotations of *hsrω* transcripts in *Drosophila* species do not show presence of a functional TSS1 except in *D. mel*. It may be noted that the TSS1 in *D. mel hsrω* gene was identified at the FlyBase only recently following in-depth RNA-seq studies. In view of this, we believe that with more detailed transcriptomic studies, existence of TSS1 is likely to be uncovered in other species as well. Since the TSS2 and *mir4951* sequences are present at proximal and distal ends, respectively, of the *hsrω* orthologue in nearly all the *Drosophila* species examined and since TSS1 is also likely to be present at the most proximal end of *hsrω* gene in different species, we tentatively suggest that the *hsrω* orthologue in different species of *Drosophila* start with TSS1 and terminate after *mir4951* as in *D. mel* (Figs. 4 and 5).

It is notable here that the more recent releases of annotated genomes of *Drosophila* species at the GBrowse section of the FlyBase show information for only four species, viz., *D. ana, D. pse, D. sim,* and *D. vir*. This may suggest that the annotated genomes of the 7 other species, shown in Fig. 2 on the basis of an earlier release, are in the process of revision and finalization. We hope that the gaps in our annotation of the *hsrω* orthologue in different species of *Drosophila* would get filled in as more definitive information about genomes of those species become available.

The present bioinformatic analysis failed to definitively predict the transcription termination sites, which must exist in multiple places in view of the known multiple transcripts produced by this gene in different species. This failure may be related to the possibility that most of these transcripts are not polyadenylated and,

consequently, they may not carry the signatures required by the different online tools used for TTS prediction. More exhaustive transcriptomic data are needed to identify the expected multiple TTS sites.

A previous cytological search for the *93D* like locus (Nath and Lakhotia, 1991) suggested absence of a homologous locus that was inducible by benzamide, colchicine, vitamin B6 or heat shocked glands homogenate in other dipterans. The preset search for the ultra-conserved *omega* intron splice junctions also failed to identify a definitive *hsrω* orthologue in these taxa (data not presented). However, presence of tandem repeats and sequence divergence at one of the heat shock induced Balbiani rings in different species of *Chironomus* (Botella, et al., 1991; Martinez-Guitarte, et al., 2008; Martínez□Guitarte and Morcillo, 2014) and the binding of Hsp83 at such a puff site following heat shock (Morcillo, et al., 1993) suggests that a functional equivalent of the *hsrω* locus may exist in other dipteran species as well. Interestingly, one of the heat shock puffs in the mediterranean fruit fly *Ceratitis capitata* has been reported (Semeshin, et al., 1995) to harbor giant electron dense ribonucleoprotein (RNP) granules, similar to the omega speckles seen at the *hsrω* locus and in nucleoplasm (Dangli, et al., 1983; Prasanth, et al., 2000). Together, it seems very likely that a functional equivalent of *Drosophila* orthologue is present in diverse dipterans as well, specific identification of which would need multi-pronged comprehensive cytogenetic, cell biological and genomic approaches. Functional similarity between the cell stress-inducible *Sat III* repetitive transcripts in human cells and the *hsrω* transcripts (Chung, et al., 2018; Jolly and Lakhotia, 2006) suggests that gene loci producing lncRNAs functionally analogous to *hsrω* transcripts indeed exist in diverse taxa as well.

Early cloning and sequence analysis of the *hsrω* genes of *D. mel*, *D. hyd* and *D. pse* (Bendena, et al., 1989; Bendena, et al., 1989; Garbe, et al., 1986; Garbe and Pardue, 1986; Peters, et al., 1984; Ryseck, et al., 1987; Walldorf, et al., 1984) revealed that, unlike the high conservation known for the heat inducible genes, this gene carried divergent sequences but a common organisation in the examined species. In agreement with these early findings, we found a very consistent organization, including length and position, of the proximal ~2Kb region starting from exon 1 in all the species although the nucleotide sequences were conserved only in limited regions. In light of such poor conservation, the ultra-conservation of sequences around the *omega* intron's 5' and 3' splice junctions in all the examined *Drosophila* species is most remarkable. Our preliminary blast analysis (data not presented) suggests that the ultra-conserved exon-intron junctions of the *omega* intron are present in nearly all members of the Drosophilidae family whose genome sequences are available at NCBI.

The preferred GT-AG donor and acceptor sites are reported to display higher conservation in introns in lncRNAs (Ponjavic, et al., 2007). However, the intriguingly unusual ultra-conservation of sequences flanking the *omega* intron splice junctions are unlikely to be related to general splicing functions since these appear unique to the *omega* intron as no other gene in any of the *Drosophila* species carries a combination of the 16 nt 5' and 60 nt 3' splice junctions. Although 82 other genes in the examined *Drosophila* species carried 5'-splice junction sequences that were 69% to 87% similar to the 16 nt long ultra-conserved 5' splice junction of the *omega* intron, none of these genes carried any similarity at the 3'-spice junctions. Since any obvious functional commonality between these 82 genes was not apparent the significance of high similarity with the *omega* intron's 5' junction sequence in these 82 genes is not clear. It is also intriguing that 50% of these showed

similarity in antisense orientation. It remains to be examined if the 16 nt ultra-conserved sequence provides certain structural features to these RNAs in relation to their functions.

With a view to see if the high conservation of *omega* intron's 5' and 3' splice junction associated sequences may be a common feature of lncRNAs, we searched the splice junction sequences in several *Drosophila* lncRNAs *(roX1, roX2, yar, bxd, CR30121, CR30267, CR32835, CR44807, CR44807 and CR44810)* in the 34 different *Drosophila* species. However, none of the splice junctions in these genes in different species showed comparable ultra-conservation of the flanking sequences (Supplementary Table S5), which suggests that the especially high conservation of splice junctions is not common for lncRNAs *per se*. We believe that the uniquely ultra-conserved *omega* splice junction flanks are essential for functionality of these ncRNAs by providing binding sites for some regulatory molecules and/or for generating some specific unique structural features. They may also help in the reported high stability of this intronic RNA (Garbe and Pardue, 1986). These aspects need further studies.

The tandem repeats uniquely associated with the *hsrω* gene in *D. mel* have attracted considerable attention due to their specific association with the omega speckles (Lakhotia, 2011; Prasanth, et al., 2000; Singh and Lakhotia, 2015). Interestingly, the initial cloning and sequencing of the *hsrω* gene from three species in the 1980s (Garbe, et al., 1986; Peters, et al., 1982; Peters, et al., 1984; Ryseck, et al., 1987) indicated the repeats to be the most variable component of this gene. The present bioinformatic analysis also revealed that the *hsrω* gene associated repeats were the most variable component in spite of their being unique to this gene. The lack of any significant sequence similarity between the tandem repeats associated with *hsrω* gene in different species is in agreement with the widely known rapid diversification of repeat sequences during speciation (Chen, et al., 2016; Chodroff, et al., 2010; Corona-Gomez, et al., 2019; Haerty and Ponting, 2013; Kapusta and Feschotte, 2014; Lakhotia, 2017; Lakhotia, 2017; Ponting, 2017; Tavares, et al., 2019; Zhang, et al., 2017).

In view of the significant roles of these tandem repeats in dynamics of omega speckles in *D. mel* (Lakhotia, 2011; Prasanth, et al., 2000; Singh and Lakhotia, 2015), and comparable mobilization of hnRNPs in stressed cells in different species (Table 1), we believe that the nuclear transcripts of this gene perform similar functions in all *Drosophila* species. Apparently, the tandem repeats of the *hsrω* gene discharge their common functions despite the sequence variations. Interestingly, the Human repetitive *Sat III* sequences, functional equivalents of the *hsrω* gene (Chung, et al., 2018; Jolly and Lakhotia, 2006), do not show any similarity with tandem repeats of the *hsrω* gene but still perform comparable functions. As known for many lncRNAs, especially for those associated with membraneless phase-separated nuclear sub-structures, the secondary and higher order structures attained by the specific lncRNA orthologue are critical (Lakhotia, 2017; Lakhotia, 2017; Lin, et al., 2018). Our present findings agree with earlier reports (Garbe, et al., 1986; Garbe, et al., 1989; Peters, et al., 1984; Ryseck, et al., 1987) of high conservation of a nonamer sequence within the poorly conserved tandem repeats at the *hsrω* gene in different species. The heterogeneous RNA binding proteins (hnRNPs) like Hrb87F, Hrb98DE, which are integral parts of the omega speckles (Prasanth, et al., 2000; Singh and Lakhotia, 2012; Singh and Lakhotia, 2015), have binding affinity for the *hsrω* repeat associated nonamer and related sequences (Zu, et al., 1998). It would be interesting to examine if these nonamers indeed provide 'landing sites' for one or more of the hnRNPs during biogenesis of omega speckles which occurs as these transcripts are being produced at the *hsrω* gene

locus (Singh and Lakhotia, 2015). Examination of status of omega speckles in different species and their dynamics under different conditions of cell stress, especially in those that do not show presence of tandem repeats as part of the *hsrω* gene, would be very useful.

The number of the 280 bp tandem repeats in *D. mel hsrω* gene has been found to vary in different strains and to be associated with thermal tolerance (Collinge, et al., 2008; Hogan, et al., 1995; McKechnie, et al., 1998). It is likely that the variations in numbers and sizes of the tandem repeats associated with the *hsrω* orthologue in different species of *Drosophila* may be adaptive in relation to the specific environmental condition under which each species lives in nature. It would be very interesting to comparatively examine these aspects in natural populations of different species.

Earlier analysis of the ORFω in three distantly related species (*D. mel*, *D. hyd* and *D. pse*), suggested this to be poorly conserved (Fini, et al., 1989; Garbe, et al., 1986). Our present analysis using data from variably related species, however, revealed a relatively higher conservation in context of sequence and secondary structure of the ORFω in 34 examined species. This indicates its potential functionality, which is also suggested by its translatability in *D. mel in vitro* (Fini, et al., 1989) as well as *in vivo* (R. K. Sahu and S. C. Lakhotia, unpublished). These need to be explored. Our *in silico* structural analysis suggests that the Omega peptide can form an alpha helical structure, which agrees with recent evidences for the existence of alpha helical short peptides (Baeriswyl, et al., 2019; Bystroff and Garde, 2003; Kelso, et al., 2004; Shin and Hahm, 2004). The first four N-terminal residues, known as capping box sequences, are considered to be important for the stabilization of alpha-helices in proteins and peptides (Aurora and Rosee, 1998; Forood, et al., 1993; Harper and Rose, 1993; Krstenansky, et al., 1989; Marqusee and Baldwin, 1987). It appears that the N-terminal conserved MEKC/MKKC/MQKC/MQMC/MEMC sequences may function as a potential N-Cap for the short omega peptide. The carboxy terminal motifs (Schellman motif and α_L motif) often include Glycine residues at the C-cap sequences (Aurora, et al., 1994). The GPT motif, seen in the Omega peptide in many species, could, therefore, be a potential C-box for the predicted alpha helix. Significance of the RRLK sequence at the Omega peptide C-terminus in other species remains unclear.

The evolutionary relationships of the *hsrω* gene revealed by its sequence comparison in different species generally agree with the previously published phylogenies of different *Drosophila* species (Seetharam and Stuart, 2012; Seetharam and Stuart, 2013; Van Der Linde, et al., 2010). However, as noted above, different domains of this gene seem to have diverged at varying rates even within a single species group/sub-group. Our study, in agreement with earlier findings (Garbe, et al., 1989), shows that while the proximal part, encompassing exon1, the *omega* intron and exon 2 of the *hsrω* gene is better conserved in different species, the distal part is more variable. The higher variability of the distal part is mostly due to the poorly conserved tandem repeats and this seems to be further compounded by the incomplete genomic information for this region in different species. It is expected that as more genomic and transcriptomic data become available, the architecture of the *hsrω* lncRNA gene in different species would be better known. Notwithstanding these limitations at this time, the similar organization, amidst sequence variability, of the two exons, *omega* intron and the ORFω is remarkable and indicates need for greater attention to this region for its functional significance.

It is now generally recognized that compared to the relatively high sequence conservation in protein-coding and a few non-coding genes due to purifying selection, majority of the lncRNA genes show little conservation in their transcribed sequences except for small domains (Chen, et al., 2016; Chodroff, et al., 2010; Corona-Gomez, et al., 2019; Haerty and Ponting, 2013; Kapusta and Feschotte, 2014; Lakhotia, 2017; Lakhotia, 2017; Ponting, 2017; Zhang, et al., 2017). Our analysis of the nucleotide sequence of the relatively large *hsrω* lncRNA gene in the genus *Drosophila* revealed a comparably peppered sequence conservation amidst a similar architecture. Like the nuclear transcripts of the *hsrω* gene which organize the omega speckles (Prasanth, et al., 2000; Singh and Lakhotia, 2015), the large NEAT1 lncRNAs organize nuclear paraspeckles in different mammals, despite poor sequence conservation. It is suggested that the NEAT1 transcripts attain a unique structure because of long-range RNA interactions and thus provide appropriate scaffold for the assembly of paraspeckles (Lin, et al., 2018). It would be interesting to examine structural features of the various *hsrω* transcripts in different species to see if, in spite of the general sequence variability, conserved domains like the 5' and 3' *omega* intron flanking sequences and/or the conserved nonamers help in attaining certain higher order structure/s that provide the required framework for assembly of the omega speckles. Likewise, further studies are necessary to understand the functional significance of the short ORFω, which although present in all species, also shows limited sequence conservation,

The present finding of high sequence diversification amidst a rather conserved architecture of the *hsrω* lncRNA gene, which impinges upon a very wide-ranging regulatory networks (is essential for survival of the organism under normal as well as stress conditions (Lakhotia, 1987; Lakhotia, 2003; Lakhotia, 2011), makes this gene locus a uniquely interesting model for studying evolutionary changes and factors that shape such changes. The availability of increasing numbers of genome sequences of *Drosophila*, other dipterans and insects make future studies very instructive from evolutionary as well as functional perspectives.

Acknowledgements

We thank the Department of Biotechnology, Ministry of Science and Technology, Govt. of India, New Delhi for supporting this work. SCL is currently supported by the Science & Engineering Research Board (SERB), Govt. of India, as SERB Distinguished Fellow. RKS is supported by the Council of Scientific and Industrial Research, New Delhi through research fellowship. EM was supported by Summer Project fellowship of the Science Academies.

Conflict of Interest

Authors declare no conflicting interests

Author contributions

RKS, EM and SCL undertook the bioinformatic analyses and wrote the manuscript.

Funding

This work was supported by a research grant (no. BT/PR6150/COE/34/20/2013) to SCL by the Department of Biotechnology, Ministry of Science and Technology, Govt. of India, New Delhi. SCL is currently supported by the Science & Engineering Research Board (SERB), Govt. of India, as SERB Distinguished Fellow. RKS is supported by the Council of Scientific & Industrial Research, New Delhi.

References

Anastasiadou, E., Jacob, L.S. and Slack, F.J. Non-coding RNA networks in cancer. *Nature Reviews Cancer* 2018;18(1):5.

Anderson, A., *et al.* Thermal tolerance trade-offs associated with the right arm of chromosome 3 and marked by the hsr-omega gene in Drosophila melanogaster. *Heredity* 2003;90(2):195-202.

Ashburner, M. and Bonner, J.J. The induction of gene activity in Drosophila by heat shock. *Cell* 1979;17(2):241-254.

Aurora, R. and Rosee, G.D. Helix capping. *Protein Science* 1998;7(1):21-38.

Aurora, R., Srinivasan, R. and Rose, G.D. Rules for alpha-helix termination by glycine. *Science* 1994;264(5162):1126-1130.

Baeriswyl, S., *et al.* X-ray Crystal Structures of Short Antimicrobial Peptides as Pseudomonas aeruginosa Lectin B Complexes. *ACS chemical biology* 2019;14(4):758-766.

Bendena, W., *et al.* Expression of heat-shock locus hsr-omega in nonstressed cells during development in Drosophila melanogaster. *Developmental biology* 1991;144(1):65-77.

Bendena, W., *et al.* hsrω: A different sort of heat shock locus. *Stress-induced proteins* 1989:3-14.

Bendena, W.G., *et al.* Multiple inducers of the Drosophila heat shock locus 93D (hsr omega): inducer-specific patterns of the three transcripts. *The Journal of cell biology* 1989;108(6):2017-2028.

Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 1999;27(2):573-580.

Botella, L.M., *et al.* Heat-shock induction and cytoplasmic localization of transcripts from telomeric-associated sequences in Chironomus thummi. *Exp Cell Res* 1991;196(2):206-209.

Burma, P.K. and Lakhotia, S. Cytological identity of 93D-like and 87C-like heat shock loci in Drosophila pseudoobscura. *Ind J Exp Biol* 1984;22:577-580.

Bystroff, C. and Garde, S. Helix propensities of short peptides: molecular dynamics versus bioinformatics. *Proteins: Structure, Function, and Bioinformatics* 2003;50(4):552-562.

Chen, J., *et al.* Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biology* 2016;17(1):19.

Chodroff, R.A., *et al.* Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biology* 2010;11(7):R72.

Chowdhuri, D.K. and Lakhotia, S. Different effects of 93D on 87C heat shock puff activity in Drosophila melanogaster and D. simulans. *Chromosoma* 1986;94(4):279-284.

Chung, C.-Y., *et al.* Aberrant activation of non-coding RNA targets of transcriptional elongation complexes contributes to TDP-43 toxicity. *Nature communications* 2018;9(1):4406.

Collinge, J.E., *et al.* Latitudinal and cold-tolerance variation associate with DNA repeat-number variation in the hsr-omega RNA gene of Drosophila melanogaster. *Heredity (Edinb)* 2008;101(3):260-270.

Corona-Gomez, J.A., *et al.* Splicing conservation signals in plant long non-coding RNAs. *bioRxiv* 2019:588954.

Dangli, A., *et al.* Heat-shock puff 93 D from Drosophila melanogaster: accumulation of a RNP-specific antigen associated with giant particles of possible storage function. *Embo J* 1983;2(10):1747-1751.

Derksen, J. Induced RNP production in different cell types of Drosophila. *Cell differentiation* 1975;4(1):1-10.

Drosopoulou, E., Konstantopoulou, I. and Scouras, Z.G. The heat shock genes in theDrosophila montium subgroup: Chromosomal localization and evolutionary implications. *Chromosoma* 1996;105(2):104-110.

Durmaz, E., *et al.* An inversion supergene in Drosophila underpins latitudinal clines in survival traits. *J Evol Biol* 2018;31(9):1354-1364.

Fatima, R., *et al.* Long noncoding RNAs in development and cancer: potential biomarkers and therapeutic targets. *Molecular and cellular therapies* 2015;3(1):5.

Ferreira, H.J. and Esteller, M. Non-coding RNAs, epigenetics, and cancer: tying it all together. *Cancer and Metastasis Reviews* 2018;37(1):55-73.

Fini, M.E., Bendena, W.G. and Pardue, M.L. Unusual behavior of the cytoplasmic transcript of hsr omega: an abundant, stress-inducible RNA that is translated but yields no detectable protein product. *The Journal of cell biology* 1989;108(6):2045-2057.

Forood, B., Feliciano, E.J. and Nambiar, K.P. Stabilization of alpha-helical structures in short peptides via end capping. *Proceedings of the National Academy of Sciences* 1993;90(3):838-842.

Garbe, J., *et al.* A Drosophila heat shock locus with a rapidly diverging sequence but a conserved structure. *Journal of Biological Chemistry* 1986;261(36):16889-16894.

Garbe, J.C., Bendena, W.G. and Pardue, M.L. Sequence evolution of the Drosophila heat shock locus hsr omega. I. The nonrepeated portion of the gene. *Genetics* 1989;122(2):403-415.

Garbe, J.C. and Pardue, M.L. Heat shock locus 93D of Drosophila melanogaster: a spliced RNA most strongly conserved in the intron sequence. *Proceedings of the National Academy of Sciences* 1986;83(6):1812-1816.

Gubenko, I., Subbota, R. and Semeshin, V. Unusual Dvosophila virilis stress puff at 20CD: cytological localization of a heat sensitive locus and some peculiarities of the heat shock response. *Hereditas* 1992;115(3):283-290.
Haerty, W. and Ponting, C.P. Mutations within lncRNAs are effectively selected against in fruitfly but not in human. *Genome Biology* 2013;14(5):R49.
Harper, E.T. and Rose, G.D. Helix stop signals in proteins and peptides: the capping box. *Biochemistry* 1993;32(30):7605-7609.
Hirose, T. and Nakagawa, S. Clues to long noncoding RNA taxonomy. *BBA-Gene Regulatory Mechanisms* 2016;1(1859):1-2.
Hoffmann, A.A., Sgrò, C.M. and Weeks, A.R. Chromosomal inversion polymorphisms and adaptation. *Trends in Ecology & Evolution* 2004;19(9):482-488.
Hogan, N., *et al.* Stability of tandem repeats in the Drosophila melanogaster Hsr-omega nuclear RNA. *Genetics* 1995;139(4):1611-1621.
Jolly, C. and Lakhotia, S.C. Human sat III and Drosophila hsrω transcripts: a common paradigm for regulation of nuclear RNA processing in stressed cells. *Nucleic acids research* 2006;34(19):5508-5514.
Jones, D.T. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of molecular biology* 1999;292(2):195-202.
Kapusta, A. and Feschotte, C. Volatile evolution of long noncoding RNA repertoires: mechanisms and biological implications. *Trends in Genetics* 2014;30(10):439-452.
Kelso, M.J., *et al.* α-turn mimetics: short peptide α-helices composed of cyclic metallopentapeptide modules. *Journal of the American Chemical Society* 2004;126(15):4828-4842.
Kennington, W.J., Partridge, L. and Hoffmann, A.A. Patterns of diversity and linkage disequilibrium within the cosmopolitan inversion *In(3R)Payne* in *Drosophila melanogaster* are indicative of coadaptation. *Genetics* 2006;172(3):1655-1663.
Kopp, F. and Mendell, J.T. Functional classification and experimental dissection of long noncoding RNAs. *Cell* 2018;172(3):393-407.
Krstenansky, J.L., *et al.* Short model peptides having a high α-helical tendency: design and solution properties. *FEBS letters* 1989;242(2):409-413.
Kumar, S., *et al.* MEGA X: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution* 2018;35(6):1547-1549.
Lakhotia, S. The 93D heat shock locus inDrosophila: a review. *Journal of Genetics* 1987;66(2):139-157.
Lakhotia, S. The 93D heat shock locus of Drosophila melanogaster: modulation by genetic and developmental factors. *Genome* 1989;31(2):677-683.
Lakhotia, S. RNA polymerase II dependent genes that do not code for protein. *INDIAN JOURNAL OF BIOCHEMISTRY AND BIOPHYSICS* 1996;33:93-102.
Lakhotia, S. Heat shock response-regulation and functions of coding and non-coding heat shock genes in Drosophila. *PROCEEDINGS-INDIAN NATIONAL SCIENCE ACADEMY PART B* 2001;67(5):247-264.
Lakhotia, S. Non-coding RNAs have key roles in cell regulation. *Proceedings of the Indian National Science Academy* 2016;82(4):1171-1182.
Lakhotia, S. and Mukherjee, A. Activation of a specific puff by benzamide in D. melanogaster. *Dros. Inf. Serv* 1970;45:108.
Lakhotia, S. and Mukherjee, T. Specific activation of puff 93 D of Drosophila melanogaster by benzamide and the effect of benzamide treatment on the heat shock induced puffing activity. *Chromosoma* 1980;81(1):125-136.
Lakhotia, S. and Mukherjee, T. Absence of novel translation products in relation to induced activity of the 93D puff in Drosophila melanogaster. *Chromosoma* 1982;85(3):369-374.
Lakhotia, S., *et al.* The non-coding transcripts of hsr-omega gene in Drosophila: Do they regulate trafficking and availability of nuclear RNA-processing factors. *Curr. Sci* 1999;77:553-563.
Lakhotia, S. and Singh, A.K. Conservation of the 93D puff of Drosophila melanogaster in different species of Drosophila. *Chromosoma* 1982;86(2):265-278.
Lakhotia, S.C. Integration of transcriptional and translational activities in normal and stressed cells by the noncoding hsr-omega transcripts in Drosophila. *Biol. Intl* 2003;45:30-33.
Lakhotia, S.C. Forty years of the 93D puff of Drosophila melanogaster. *Journal of biosciences* 2011;36(3):399-423.
Lakhotia, S.C. Long non coding RNAs coordinate cellular responses to stress. *Wiley Interdisciplinary Reviews: RNA* 2012;3(6):779-796.
Lakhotia, S.C. Divergent actions of long noncoding RNAs on X-chromosome remodelling in mammals and Drosophila achieve the same end result: dosage compensation. *Journal of genetics* 2015;94(4):575-584.
Lakhotia, S.C. From heterochromatin to long noncoding RNAs in Drosophila: expanding the arena of gene function and regulation. In, *Long Non Coding RNA Biology*. Springer; 2017. p. 75-118.

Lakhotia, S.C. Non-coding RNAs demystify constitutive heterochromatin as essential modulator of epigenotype. *The Nucleus* 2017;60(3):299-314.
Lakhotia, S.C. Central dogma, Selfish DNA and noncoding RNAs: a historical perspective *Proc. Indian Natn. Sci. Acad.* 2018;84(2):315-427.
Lakhotia, S.C., *et al.* The large noncoding hsrω-n transcripts are essential for thermotolerance and remobilization of hnRNPs, HP1 and RNA polymerase II during recovery from heat shock in Drosophila. *Chromosoma* 2012;121(1):49-70.
Leenders, H., *et al.* Selective induction of a giant puff in Drosophila hydei by vitamin B6 and derivatives. *Chromosoma* 1973;41(4):447-460.
Lengyel, J.A., *et al.* Transcription and metabolism of RNA from the Drosophila melanogaster heat shock puff site 93D. *Chromosoma* 1980;80(3):237-252.
Lin, Y., *et al.* Structural analyses of NEAT1 lncRNAs suggest long-range RNA interactions that may contribute to paraspeckle architecture. *Nucleic Acids Res* 2018;46(7):3742-3752.
Mallik, M. and Lakhotia, S.C. The developmentally active and stress-inducible noncoding hsrω gene is a novel regulator of apoptosis in Drosophila. *Genetics* 2009;183(3):831-852.
Mallik, M. and Lakhotia, S.C. Pleiotropic consequences of misexpression of the developmentally active and stress-inducible non-coding hsrω gene in Drosophila. *Journal of biosciences* 2011;36(2):265-280.
Marqusee, S. and Baldwin, R.L. Helix stabilization by Glu-... Lys+ salt bridges in short peptides of de novo design. *Proceedings of the National Academy of Sciences* 1987;84(24):8898-8902.
Martinez-Guitarte, J.L., Diez, J.L. and Morcillo, G. Transcription and activation under environmental stress of the complex telomeric repeats of Chironomus thummi. *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology* 2008;16(8):1085-1096.
Martínez□Guitarte, J. and Morcillo, G. Telomeric transcriptome from Chironomus riparius (Diptera), a species with noncanonical telomeres. *Insect Mol . Biol.* 2014;23(3):367-380.
Matsumoto, A. and Nakayama, K.I. Hidden peptides encoded by putative noncoding RNAs. *Cell structure and function* 2018:18005.
McColl, G., Hoffmann, A.A. and McKechnie, S.W. Response of two heat shock genes to selection for knockdown heat resistance in *Drosophila melanogaste*r. *Genetics* 1996;143(4):1615-1627.
McKechnie, S.W., *et al.* Both allelic variation and expression of nuclear and cytoplasmic transcripts of Hsr-omega are closely associated with thermal phenotype in Drosophila. *Proceedings of the National Academy of Sciences* 1998;95(5):2423-2428.
Mohler, J. and Pardue, M.L. Mutational analysis of the region surrounding the 93D heat shock locus of Drosophila melanogaster. *Genetics* 1984;106(2):249-265.
Moltó, M.D., Martínez□Sebastián, M.J. and De Frutos, R. Gene arrangement phylogeny of the E element in Drosophila species of the Obscura group. *Journal of evolutionary biology* 1993;6(5):753-758.
Morcillo, G., *et al.* HSP90 associates with specific heat shock puffs (hsrω) in polytene chromosomes of Drosophila and Chironomus. *Chromosoma* 1993;102(9):648-659.
Mukherjee, T. and Lakhotia, S. 3 H-uridine incorporation in the puff 93D and in chromocentric heterochromatin of heat shocked salivary glands of Drosophila melanogaster. *Chromosoma* 1979;74(1):75-82.
Mukherjee, T. and Lakhotia, S. Heat Shock Puff Activity in Salivary Glands ofDrosophila melanogaster Larvae during Recovery from Anoxia at Two Different Temperatures. *Indian Journal of Experimental Biology* 1982;20:437-439.
Muller, H.J. Bearing of the Drosophila work on systematics. *The new systematics* 1940:185-268.
Mutsuddi, M. and Lakhotia, S. Spatial expression of the hsr□omega (93D) gene in different tissues of drosophila melanogaster and identification of promoter elements controlling its developmental expression. *genesis* 1995;17(4):303-311.
Nath, B. and Lakhotia, S. Search for a Drosophila-93D-like locus in Chironomus and Anopheles. *Cytobios* 1991;65(260):7-13.
Noh, J.H., *et al.* Cytoplasmic functions of long noncoding RNAs. *Wiley Interdisciplinary Reviews: RNA* 2018;9(3):e1471.
O'Grady, P.M. and DeSalle, R. Phylogeny of the genus Drosophila. *Genetics* 2018;209(1):1-25.
Onorati, M.C., *et al.* The ISWI chromatin remodeler organizes the hsrω ncrna–containing omega speckle nuclear compartments. *PLoS Genetics* 2011;7(5):e1002096.
Peters, F., *et al.* Chromosomal arrangement of heat shock locus 2-48B in Drosophila hydei. *Chromosoma* 1982;85(2):237-249.
Peters, F., Lubsen, N. and Sondermeijer, P. Rapid sequence divergence in a heat shock locus of Drosophila. *Chromosoma* 1980;81(2):271-280.
Peters, F.P., *et al.* The unusual structure of heat shock locus 2-48B in Drosophila hydei. *Molecular and General Genetics MGG* 1984;197(3):392-398.

Piccolo, L.L. Drosophila as a Model to Gain Insight into the Role of lncRNAs in Neurological Disorders. In, *Drosophila Models for Human Diseases*. Springer; 2018. p. 119-146.
Ponjavic, J., Ponting, C.P. and Lunter, G. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome research* 2007;17(5):556-565.
Ponting, C.P. Biological function in the twilight zone of sequence conservation. *BMC Biology* 2017;15(1):71.
Prasanth, K., *et al.* Omega speckles-a novel class of nuclear speckles containing hnRNPs associated with noncoding hsr-omega RNA in Drosophila. *Journal of cell science* 2000;113(19):3485-3497.
Quinn, J.J., *et al.* Rapid evolutionary turnover underlies conserved lncRNA–genome interactions. *Genes & development* 2016;30(2):191-207.
Rako, L., *et al.* Candidate genes and thermal phenotypes: identifying ecologically important genetic variation for thermotolerance in the Australian Drosophila melanogaster cline. *Molecular Ecology* 2007;16(14):2948-2957.
Ransohoff, J.D., Wei, Y. and Khavari, P.A. The functions and unique features of long intergenic non-coding RNA. *Nature Reviews Molecular Cell Biology* 2018;19(3):143.
Ryseck, R.-P., *et al.* Heat shock loci 93D of Drosophila melanogaster and 48B of Drosophila hydei exhibit a common structural and transcriptional pattern. *Nucleic acids research* 1987;15(8):3317-3333.
Ryseck, R.-P., Walldorf, U. and Hovemann, B. Two major RNA products are transcribed from heat-shock locus 93D of Drosophila melanogaster. *Chromosoma* 1985;93(1):17-20.
Salamov, A.A. and Solovyev, V. Recognition of 3'-end cleavage and polyadenylation region of human rnRNA precursors. *CABIOS* 1997;13:23-28.
Saumweber, H., *et al.* Monoclonal antibodies against chromosomal proteins of Drosophila melanogaster. *Chromosoma* 1980;80(3):253-275.
Seetharam, A. and Stuart, G.W. Whole genome phylogenies for multiple Drosophila species. *BMC research notes* 2012;5(1):670.
Seetharam, A.S. and Stuart, G.W. Whole genome phylogeny for 21 Drosophila species using predicted 2b-RAD fragments. *PeerJ* 2013;1:e226.
Semeshin, V., *et al.* Electron microscope investigation of polytene chromosomes in the Mediterranean fruit fly Ceratitis capitata. *Genome* 1995;38(4):652-660.
Shin, S.Y. and Hahm, K.-S. A short α-helical antimicrobial peptide with antibacterial selectivity. *Biotechnology letters* 2004;26(9):735-739.
Singh, A.K. and Lakhotia, S.C. The hnRNP A1 homolog Hrp36 is essential for normal development, female fecundity, omega speckle formation and stress tolerance in Drosophila melanogaster. *Journal of biosciences* 2012;37(4):659-678.
Singh, A.K. and Lakhotia, S.C. Dynamics of hnRNPs and omega speckles in normal and heat shocked live cell nuclei of Drosophila melanogaster. *Chromosoma* 2015;124(3):367-383.
Stoiber, M., *et al.* Diverse hormone response networks in 41 independent Drosophila cell lines. *G3: Genes, Genomes, Genetics* 2016;6(3):683-694.
Tamura, K. and Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 1993;10(3):512-526.
Tavares, R.C.A., Pyle, A.M. and Somarowthu, S. Phylogenetic analysis with improved parameters reveals conservation in lncRNA structures. *J. Mol. Biol.* 2019;431(8):1592-1603.
Van Der Linde, K., *et al.* A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genetics research* 2010;92(1):25-38.
Walldorf, U., *et al.* Cloning of heat□shock locus 93D from Drosophila melanogaster. *The EMBO journal* 1984;3(11):2499-2504.
Zhang, B., *et al.* Identification and characterization of a class of MALAT1-like genomic loci. *Cell Reports* 2017;19(8):1723-1738.
Zu, K., SIKES, M.L. and BEYER, A.L. Separable roles in vivo for the two RNA binding domains of a Drosophila A1-hnRNP homolog. *Rna* 1998;4(12):1585-1598.

Table 1. Cytogenetic identification of *hsrω* like loci in different species of *Drosophila*

***Drosophila* Species**	**Polytene chromosome location**	**Inducibility***			**Large RNP particles**	**hnRNP accumulation**	**In situ hybridization with related species' probes**	**References**
		HS	**BM**	**Vit-B6**				
ananassae	2L-2C	Yes	Yes	Yes				(Lakhotia and Singh, 1982)
auraria	2L-42C	Yes	Yes				Yes	(Drosopoulou, et al., 1996)
bicaudata	2L-27	Yes	Yes				Yes	(Drosopoulou, et al., 1996)
bipectinata		Yes	Yes	No				(Lakhotia and Singh, 1982)
eohydei	2-48BC	Yes		Yes			Yes	(Peters, et al., 1980)
guanche	O-85AB	Yes	Yes					(Moltó, et al., 1993)
hydei	2-48BC	Yes	Yes	Yes	Yes	Yes	Yes	(Dangli, et al., 1983; Derksen, 1975 Lakhotia and Singh, 1982; Leenders et al., 1973; Peters, et al., 1980)
jambulina	3L-61	Yes	Yes				Yes	(Drosopoulou, et al., 1996)
kikawai	E-11BC (3R-90)	Yes	Yes	Yes			Yes	(Lakhotia and Singh, 1982) (Drosopoulou, et al., 1996)
madeirensis	O-85AB	Yes	Yes					(Moltó, et al., 1993)
melanogaster	3R-93D	Yes	Yes	No	Yes	Yes		(Ashburner and Bonner, 1979) (Lakhotia and Mukherjee, 1980) (Dangli, et al., 1983; Derksen, 1975)
nasuta	2R-48A	Yes	Yes	Yes				(Lakhotia and Singh, 1982)
neohydei	2-48BC	Yes		Yes			Yes	(Peters, et al., 1980)
pseudoobscura	2-58C	Yes	Yes				Yes	(Burma and Lakhotia, 1984)
seguyi	2L-38	Yes	Yes				Yes	(Drosopoulou, et al., 1996)
serrata	2L-35	Yes	Yes				Yes	(Drosopoulou, et al., 1996)
simulans	3R-93D	Yes	Yes	No		Yes		(Chowdhuri and Lakhotia, 1986)
subobscura	O-85AB	Yes	Yes					(Moltó, et al., 1993)
virilis	2-20CD	Yes	Yes	Yes	Yes	Yes	Yes	(Dangli, et al., 1983; Derksen, 1975 Gubenko, et al., 1992)

* HS = heat shock; BM = benzamide, Vit-B6 = vitamin B6 or pyridoxal phosphate

Table: 2. *hsrω* orthologue, flanking neighbours and lengths of different regions of *hsrω* gene ascertained on bases of data at http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/.

Species	Annotated *hsrω* orthologs	Annotated transcripts	Gene order	Estimated *hsrω* gene length	Estimated lengths of regions of *hsrω* gene				
					TSS1 to TSS2	TSS2 to ORFω	Exon1	Intron	Exon 2
D. alb	NA		*S H C*	21,833	TSS1 NF	157	552	760	678
D. ame	NA		Not found	08,554	273	122	503	755	780
D. ana	*GF27083*	3	*E S H C T B M*	11,882	789	074	414	681	752
D. ari	NA		*S H C*	10,189	269	115	472	733	736
D. bia	NA		*S H C*	17,841	811	118	466	676	704
D. bip	NA		*S H C*	14,262	673	122	465	692	756
D. bus	NA		Not found	UD	476	159	493	643	667
D. ele	NA		*S H C*	20,291	362	127	489	685	700
D. ere	NA		*E S H C M T B*	08,840	321	129	474	695	692
D. eug	NA		*S H C*	UD	448	115	456	702	676
D. fic	NA		*S H C*	17,429	774	121	495	675	698
D. gri	*GH26493*	1	*E S H C M T B*	12,090	413	142	510	731	740
D. hyd	NA		*S H C*	17,291	361	125	480	736	768
D. kik	NA		*S H C*	12,513	610	129	466	729	799
D. mel	*hsrω*	7	*E S H C M T B*	20,245	494	119	477	712	699
D. mir	NA		*S H C*	32,504	328	119	476	741	723
D. moj	*GI26477*	2	*E S# H# C M T B*	22,252	261	117	474	745	749
D. mon	NA		*S H C*	UD	136	120	514	782	761
D. nas	NA		Not found	UD	355	157	552	758	678
D. nav	NA		*S H C*	04,522	256	121	481	785	735
D. nov	NA		*S H C*	19,446	278	122	513	760	779
D. obs	NA		*S H C*	UD	342	115	473	750	719
D. per	NA		*E S H C M T B*	29,551	470	117	476	750	723
D. pse	*GA31955*	2	*E S H C M T B*	21,380	355	119	476	740	723
D. rho	NA		*S H C*	UD	324	116	483	667	711
D. sec	NA		*E S H C M T B*	UD	497	126	481	781	699
D. ser	NA		*S H C*	20,783	TSS1 NF	132	476	719	792
D. sim	*GD28367*	1	*E S H## C M T B*	15,325	497	126	481	712	700
D. sub	NA		*H C*	13,927	570	125	490	747	720
D. suz	NA		*H C*	26,447	706	120	465	683	705
D. tak	NA		*S H C*	11,978	TSS1 NF	121	490	715	710
D. vir	*GJ26899*	14	*E S H C M T B*	14,386	272	119	500	752	759
D. wil	NA		*E S H C M T B*	33,946	341	116	474	711	783
D. yak	*GE28639*	4	*E S H C M T B*	20,850	517	126	477	682	707

Gene abbreviations: E S H C M T B = *ETHR, SIFaR. hsrω, CG16791, mod, tin, bap*, respectively; *M* = annotated in FlyBase as ncRNA in *D. gri*; *S H C* or *H C* = not annotated at FlyBase but identified in the present study; *T B M* = genes in different order; # = ncRNA gene; *S#* = overlapping genes; *M* = Putative orthologue of *mod*; UD = *hsrω* gene length undetermined due to non-overlapping scaffolds; TSS1 NF = TSS1 not found.

Table: 3. Genomic coordinates of *hsrω* and its flanking genes in 12 *Drosophila* species whose genomes are annotated at http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/.

Species, chromosome arm and/or scaffold no.	**Genomic coordinates (approximate)**						
D. mel **3R**	*ETHR*[+] 21247 kb	*SiFAR*[-] 21294 kb	*hsrω* [+] 21296 kb	*CG16791* [-] 21350 kb	*mod(mdg4)*[-] 21377 kb	*tin*[+] 21379 kb	*bap*[+] 21389 kb
D. ana **Sc_13340**	GF17700[+] 3746 kb	GF17259[-] 3779 kb	GF27083[+] 3783 kb	GF17258[-] 3829 kb	GF18944[+] 23479 kb	GF18599[+] 17555 kb	GF18600[+] 17564 kb
D. ere **Sc_4770**	GG24498[-] 4587 kb	GG14716[+] 4542 kb	*hsrω* [-] 4538 kb	GG14705[+] 4504 kb	GG14539[+] 4468 kb	GG24587[-] 4466 kb	GG24598[-] 4456 kb
D. gri **Sc_15074**	GH21933[+] 2602 kb	GH18883[-] 2637 kb	GH26493[+] 2641 kb	GH18872[-] 2680 kb	GH18751[-] 2716 kb	GH22021[+] 2717 kb	GH22032[+] 2727 kb
D. moj **Sc_6540**	GI21940[-] 33876 kb	GI10832[+] 33838 kb	GI26477[-] 33836 kb	GI10831[+] 33781 kb	GI10821[+] 33752 kb	GI21950[-] 33750 kb	GI21951[-] 33740 kb
D. per **Sc_19**	GL27327[+] 1304 kb	GL27160[-] 1334 kb	*hsrω* [+] 1339 kb	GL27159[-] 1401 kb	GL27151[-] 1430 kb	GL27335[+] 1433 kb	GL27337[+] 1442 kb
D. pse **Sc_2**	GA30100[+] 5586 kb	GA10579[-] 5617 kb	GA31955[+] 5622 kb	GA14153[-] 5675 kb	GA30094[-] 5703 kb	GA20667[+] 5707 kb	GA20672[+] 5715 kb
D. sec **Sc_38 and 6**	GM15081[-] 48 kb	GM15085[+] 11 kb	*hsrω* [-] 6 kb	GM11029[+] 4341 kb	GM11009[+] 4288 kb	tin[-] 4286 kb	bap[-] 4275 kb
D. sim **Sc_3R**	GD19410[-] 4235 kb	GD19990[+] 4189 kb	GD28367[-] 4185 kb	GD19989[+] 4137 kb	GD15087[+] 4108 kb	tin[-] 4107 kb	bap[-] 4097 kb
D. vir **Sc_12822**	GJ14300[-] 1257 kb	GJ14438[+] 1221 kb	GJ26899[-] 1206 kb	GJ14437[+] 1141 kb	mod(mdg4)[+] 1110 kb	GJ14310[-] 1108 kb	GJ14311[-] 1098 kb
D. wil **Sc_2_1100000004943**	GK14494[+] 16064 kb	GK12768[-] 16097 kb	*hsrω* [+] 16101 kb	GK12767[-] 16170 kb	GK12763[-] 16199 kb	GK14502[+] 16201 kb	GK14503[+] 16213 kb
D. yak **3R**	GE25720[+] 5876 kb	GE24999[-] 5904 kb	GE28639[+] 5923 kb	GE24998[-] 5979 kb	GE24979[-] 6009 kb	GE25731[+] 6011 kb	GE25732[+] 6028 kb

[+] and [-] = Genes in forward or reverse direction, respectively, with reference to the GBrowse chromosomal scale; *hsrω* = mapped *hsrω*, Genes = Linkage order reversed in *D. ana*; G15087 = identified in this study as an orthologue of *mod*.

Table: 4. MSA and conservation scores of 16 nt and 60 nt sequences flanking, respectively, the 5' and 3' exon-intron splice junctions of the *omega* intron in different *Drosophila* species.

Species	5' Splice Junction	3' Splice Junction	Intron Length
D. alb	TAGGATACCA--GGTATA	TA-ACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAAAAGTTGATA	760
D. ame	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	755
D. ana	AAGGAAGCCA--GGTACA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-TGTTGATA	681
D. ari	TAGGAAGACA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	733
D. bia	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	676
D. bip	AAGGAAGCCA--GGTACA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-TGTTGATA	692
D. bus	TAGCTCGCCA--GGTATA	TTAACCAAGCAACATGTATTTCT-TCTTTGAACTTTATAGTTGGGCGTTGAA-AGTTGATA	643
D. ele	TAGGAAGCCA--GGTATA	CTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	685
D. ere	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	695
D. eug	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	702
D. fic	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	675
D. gri	TAGCAAGCCA--GGCTTG	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTGAAA-AGTTGATA	731
D. hyd	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	736
D. kik	GAGGAAGCCA--GGTACA	TTAACC-AGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	729
D. mel	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	712
D. mir	TAGGAAGCCA--GGTATT	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	741
D. moj	TAGGAAGACA--GGTATA	TTAACCAAGCATTATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	745
D. mon	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	782
D. nas	TAGGATACCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AAGTTTAT	758
D. nav	TAGGAAGACA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	785
D. nov	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	760
D. obs	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACATTATAGTTGGGCGTTGAA-AGTTGATA	750
D. per	TAGGAAGCCA--GGTATT	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	750
D. pse	TAGGAAGCCA--GGTATT	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	740
D. rho	TAGGAAGCCA--GGTATA	CTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	667
D. sec	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	781
D. ser	GAGGAAGCCA--GGTACA	TTAACC-AGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	719
D. sim	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	712
D. sub	TAGGAAGCCA--GGTACA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-TGTTGATA	747
D. suz	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	683
D. tak	TAGGAAGCCAGAGGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	715
D. vir	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	752
D. wil	TAGGAATCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	711
D. yak	TAGGAAGCCA--GGTATA	TTAACCAAGCAATATGTATTTCTTTCTCTAAACTTTATAGTTGGGCGTTGAA-AGTTGATA	682
	** ** **	*** **** ********** *** * *** ********** ** *	

Table: 5. MSA and conservation scores of 5'and 3' regions of mature *mir-4951* transcripts overlapping with *hsrω* in different species of *Drosophila.*

Species	Mature 5'*mir-4951*	Mature 3'*mir-4951*	TTS	Length
D. alb	TCGGGTGAGGTTATATT-AGAT	CCAC---GA-AAAATAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TCAAAA	154
D. ame	TCGGGTGAGGTTT---TATGAT	CCAA---A--AAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TAAAAA	160
D. ana	TCGGGTGAGGTTAT-AC-TGAT	CC-----CA-AGACTAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TATAAA	138
D. ari	TCGGGTGAGGTTT---TATGAT	CCCA---CC--AAAAAAACC-TCACCCACCA	AACATATGAACAAAGA-GA-TACAAA	150
D. bia	TCGGGTGAGGTTT---C-TGAT	CCA-------CAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-A-AAAA	155
D. bip	TCGGGTGAGGTCA--AC-TGTT	CC-----CA-AAACTAACC-TCACCCACCA	AATATATGAAAAAAGA-GA-TACAAA	136
D. bus	----------	----------	----------	SNF
D. ele	TCGGGTGAGGTTT---C-TGAT	CCCT----A-TAAAAAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-TATAAA	133
D. ere	TCGGGTGAGGTTT---C-TGAT	CCAT---AC-CAAAAAACC-TCACCCACCA	AATGTATGAACAGAGA-GA-TAAAAA	135
D. eug	TCGGGTGAGGTTT---C-TGAT	CCAT------CAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TACAAA	149
D. fic	TCGGGTGAGGTTT---C-TGAT	CCAC------CTAAAAACC-TCACCCACCA	AATGTATGAACAAAGA-GA---AAAA	142
D. gri	TCGGGTGAGGTTT---T-AGAT	CAAAAAAAAAAAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TA-AAA	131
D. hyd	TCGGGTGAGGTTT---TATGAT	CCAC---A--AAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TAAAAA	139
D. kik	TCGGGTGAGGTTA---C-TGAT	CCAC--CTT-AAAATAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-ACAAAA	156
D. mel	TCGGGTGAGGTTT---C-TGAT	CCAT------CAAAAAACC-TCACCCACCA	AATGTATGACCAGAGA-GA-TAAAAA	150
D. mir	TCGGGTGAGGTTA---T-GGAT	CCAC----C-CAAATAACC-TCACCCACCA	AAT-TATGAACAAAGA-GA-AAAAAA	132
D. moj	TCGGGTGAGGTTT---TATGAT	CCCA---CC-AAAAAAACC-TCACCCACCA	AACATATGAACAAAGA-GA-TACAAA	148
D. mon	GTGGGTGAGGTTT---T-TGTT	TGTT----C-AAGATGGGAAGGAGC-GAGA	ATTGCATCAGAAAACCTCACCCGAAA	157
D. nas	TCGGGTGAGGTTATATT-AGAT	CCAT---GA-AAAATAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TCAAAA	154
D. nav	TATGATGCAGTA----CTCACT	CCCA---CC-AAAAAAACC-TCACCCAACA	AAAATATGAACAAAGA-GA-TAAAAA	140
D. nov	TATGATGCAATTC---TCGCTC	CCAA------AAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-TAAAAA	148
D. obs	TCGGGTGAGGTTA---T-GGAT	CCAC----C-CAAATAACC-TCACCCACAA	AATGTATGAACAAAGA-GA-AAAAAA	144
D. per	TCGGGTGAGGTTA---T-GGAT	CCAC----C-CAAATAACC-TCACCCACCA	AAT-TATGAACAAAGA-GA-AAAAAA	132
D. pse	TCGGGTGAGGTTA---T-GGAT	CCAC----C-CAAATAACC-TCACCCACCA	AAT-TATGAACAAAGA-GA-AAAAAA	132
D. rho	TCGGGTGAGGTTT---C-TGAT	CTTT------AAAAAAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-T--AAA	145
D. sec	TCGGGTGAGGTTT---C-TGAT	CCAT------CAAAAAACC-TCACCCACCA	AATGTATGACCAGAGA-GA-TAAAAA	150
D. ser	TCGGGTGAGGTTA---C-TGAT	CCC----TA-AAAATAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-ACAAAA	159
D. sim	TCGGGTGAGGTTT---C-TGAT	CCAT------CAAAAAACC-TCACCCACCA	AATGTATGACCAGAGA-GA-TAAAAA	150
D. sub	TCGGGTGAGGTTA---T-GGAT	CCAC----C-ACAATAACC-TCACCCACAA	AATGTATGAACAAAGA-GA-GAAAAA	142
D. suz	TCGGGTGAGGTTT---C-TGAT	CCAC----A-CAAAAAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-AAAAAA	150
D. tak	TCGGGTGAGGTTT---C-TGAT	CCAC------CTAAAAACC-TCACCCACCA	AATGTATATTCAAAGA-GA-ATAAAA	136
D. vir	TCGGGTGAGGTTT---TATGAT	CCAA---CCAAAAAAAACC-TCACCCACCA	AATGTATGAACAAAGA-GA-TAAAAA	165
D. wil	TCGGGTGAGGTTA---G-TGAT	CC-----AA-AAAAAAACC-TCACCCACCA	AATATATGAACAAAGA-GA-AACAAA	143
D. yak	TCGGGTGAGGTTT---C-TGAT	CCAA---AT-CAAAAAACC-TCACCCACCA	AAAGTATGAACAGAGA-GA-TAAAAA	153
	* ** *	* * *	* ** * * * * ***	

Conservation colour code = BAD AVG GOOD

SNF- Sequence not found.

Table: 6. MSA and conservation scores of predicted amino acid sequences of the peptide-□.

Species	Predicted amino acid sequence of the putative peptide-□	Length	Predicted secondary structure
D. alb	--------M-----KKCKTTALKA-KLR--SYNMA--RRL-K	23	
D. ame	MYIYCSTAM-----QKCAETVGMA-NLR--SYNMA--RRL-K	31	
D. ana	--------MY----QKCDNRVAEA-EK-RKQIGTVA-EGP-T	26	
D. ari	--------M-----KKCAESVGKA-KLR--SYIMA--RRL-K	23	
D. bia	--------M-----EKCKNRVHRD-QQ-QYEFRKMT-PGP-T	25	
D. bip	--------MY----QKCDNRGAEA-EK-RKQIGTVA-EGP-T	26	
D. bus	--------MISHQQQMCRKFTKMS-RQR--SYNMA--RRL-K	28	
D. ele	--------M-----EKCKNRVAEA-EL-QYEYCKMT-PGP-T	25	
D. ere	--------M-----EKCKNRVPAD-EQQ-NEYCKMQ--GP-T	24	
D. eug	--------M-----EKCKNCVAEG-DGQLYDYCKMT-PGP-T	26	
D. fic	--------M-----EKCENRVAEE-EQ-QEENCKIS-EGP-T	25	
D. gri	--------M-----EKCRIAVGMS-KLR--SYIMA--RRL-K	23	
D. hyd	--------M-----EKCTVFVAKA-RLR--SYNMS--RRL-K	23	
D. kik	--------M-----EKCKNHSKGG-AEEQFCY-QMT-EGP-T	25	
D. mel	--------M-----EKCKNRVPAD-EQQQYEYCKMQGQGP-T	27	
D. mir	--------M-----EKCGKRVVKA-KQR--PYSSMA-TTP-T	24	
D. moj	--------M-----KKCAESVGKA-KLR--SYIMA--RRL-K	23	
D. mon	MHIYCSTAM-----KKCAETFGKS-KLR--SYNMA--RRL-K	31	
D. nas	--------M-----KKCKTTALKA-KLR--SYNMA--RRL-K	23	
D. nav	--------M-----KKCAEGVGRA-KLR--SYIMA--RRL-K	23	
D. nov	MYIYCSTAM-----QKCAETVGKA-NLR--SYNMA--RRL-K	31	
D. obs	--------M-----EKCGKRVVTA-KQK--PYASMA-KTP-T	24	
D. per	--------M-----EKCGKRVVEA-KQR--PYSSMA-TTP-T	24	
D. pse	--------M-----EKCGKRVVKA-KQR--PYSSMA-TTP-T	24	
D. rho	--------M-----EKCKNRVAEA-EQ-QYKYWKLT-PGP-T	25	
D. sec	--------M-----EKCKNRVPGG-EQQ-YEYCKMQGQGP-T	26	
D. ser	--------M-----EKCKIRSEGAVEEQ-SSYQAM--EGP-T	26	
D. sim	--------M-----EKCKNRVPGG-EQQ-YEYCKMQGAGP-T	26	
D. sub	--------M-----EKCGKRVAKA-KQR--PYSSME-KTP-T	24	
D. suz	--------M-----EKCKNRDADE-QQ-QYEFCKMT-PGP-T	25	
D. tak	--------M-----EKCKNRVAPD-EQ-QYEYCKMT-PGP-T	25	
D. vir	MYIYCSTAM-----QMCAETVGKA-NLR--SYNMA--RRL-K	31	
D. wil	--------M-----EKCKKSKQKQ--H-CYQYGRKS-NIAVA	25	
D. yak	--------M-----EMCKNCVPAD-EQQ-CDYCKMQGQGP-T	26	
	* : *		

Conservation colour code = BAD AVG GOOD

Cylinders represent alpha helical structures and yellow arrows represent beta pleated structures in the last column.

Table: 7. Predicted promoters 1 and 2 of *hsrω* in different species of *Drosophila*

A. Promoter 1

Species	Promoter-1 (TSS1)	Score
D. alb	----------	PNF
D. ame	GTTAA-A-T-ATATGTATAA-G-TG-GCATA-CCT-TGAAAAGGGACCAT-TT--A-ATGCT	0.97
D. ana	GTAGT-GAC-AACTCTATGAATC-A-GGG-GAGGCC-AGT-TGTCGCACT--C--T-GGAAA	0.92
D. ari	ACATA-T-GAGTTTAAAAAA---AT-GCATA-TTCATATGTACATAT-AT-TT--T-TGCTA	0.81
D. bia	CAGTG-G-T-CAATAAATAA-T-CT-GCCTC-AATTGGTATCTT-AAAAA-T--CT-TAAGT	0.85
D. bip	GCATT-T-T-ACATAAATAA-GT-A-GCG-ACAACC-CTA-TGAATCAGTGGG--G-GCCAG	0.96
D. bus	CAATT-T-C-AAACATAAAT-GA-A-GCGCGAGAGA-TTAACAGAACATT--C--T-ATGAG	0.92
D. ele	AACAC-T-A-GTTTATATAA-T-GC-GCTTG-CTA-AGAAGAGATAAACT-TA--A-TTTAA	0.95
D. ere	TGAGC-A---GTTTATAAAA-C-AGAAGATA-GAA-TATAGCTGAAACAG-AA--A-ATTCT	0.90
D. eug	TTAAT-G-C-TTATAAAAAA---AC-GAAGT-TTATTAATTATAATTATA-AT--C-TATGC	0.91
D. fic	ATGAATA-T-AAATGTATAA-A-TG-GGTTT-TAA-TATTTCTTCAA-AA-AC--A-AAACA	0.96
D. gri	GTCAG-T-C-CCATATATCT-G-TC-ACGGT-CGCACAAATACT-GTCTG-G--CA-ACACC	0.89
D. hyd	ACTCG-C-A-CAATAAAAAT-C-TG-GCAAC-ATCATTTATACA-TTAAA-A--AC-AACAT	0.81
D. kik	TGAGT-A-A-ACATTTACGA-T-TT-GTCTTTGTC-TTTCTCTC-GAAAA-TG--C-TATCT	0.89
D. mel	TGTGA-A-T-CAATAAAAAA-AAGC-ATGCG-GTC-CCTCTCTC-GAAAA-CT--G-AACAT	0.96
D. mir	ATTCA-G-C-GTGTTAAAAG-A-GG-GAAAA-GAG-AAATTTTATAATCG-TG--T-AAGAC	0.82
D. moj	TCTAT-T-A-CTATATATAA-AATC-A-A-TTCCAAAATATTAGCTTATT--T--T-TCAGC	0.95
D. mon	GGGCA-G-A-GTATATAAAG-CAGT-GAG-GCACGT-AGGCAG-CGAACA-CA--G-CAGTG	0.98
D. nas	TCCCA-C---CGATATAAAA-ATGC-TCTTT-CAA-TGAGTCTAAGTTAA-TT--T-GTTAT	0.85
D. nav	ATTTG-A-G-TTTTAAAAAA---AC-GCAAA-ATCATATTTAAATTTTGC-CA--G-GCATA	0.99
D. nov	GTTAA-A-T-ATATGTATAA-G-TG-GCATA-CCT-TGAAAAGAGACCAT-TT--T-ATGCT	0.97
D. obs	ACAGA---T-ATTTATACAG-TGGC-GCCCC-TTCCACAACATT-GGTAT-T--TT-ATATT	0.94
D. per	CGAAT-T-T-GTATATAAAT-AA-T-ACA-TAAATA-ATACATTCGTGCT-TG--C-TTTTA	0.81
D. pse	CGAAT-T-T-GTATATAAAT-AA-T-ACA-TAAATA-ATACATTCGTGCT-TG--C-TTTTA	0.81
D. rho	CTAAC-T-A-TCGTAAAAAA-A-GT-CCATC-A-AATTTGGATGGGGAAC-TA--C-TGGGA	0.97
D. sec	TGTGA-A-T-CAATAAAAAA-C-AT---GCG-GTC-TCTCTCTC-GAAAA-CTGAACATTAT	0.95
D. ser	----------	PNF
D. sim	TGTGA-A-T-CAATAAAAAA-ACAT---GCG-GTC-CCTCTCTC-GAAAA-CTGAA-CATTA	0.90
D. sub	CGGGA-T-A-AATTGTAAAA-TAGC-GGG-TAGCTGGA-ATTATAGCTTT--T--T-CCCAA	0.87
D. suz	TTCTA-T-T-ATAAATTTAA-AG-T-GCGCAGTTCT-TTGTTTATTCAAC--A--T-TCAAC	0.98
D. tak	----------	PNF
D. vir	GTTAA-A-T-ATATGTATAA-G-TG-TCACA-CCT-TGAAAAGGGACAAT-TC--T-TGCCG	0.91
D. wil	AAGCG-G-G-TTATATAAAA-A-AG-ATATA-GAA-AAGTAGTGCTTAAG-TG--C-GAAAC	0.93
D. yak	CAGCC-C----TATAAATCA-GTATAACCTA-GTGGCCTCTCTC-GAAAA-CT--G-ACGAT	0.87

B. Promoter 2

Species	Promoter-2 (TSS2)	Score
D. alb	ACT---CGGGGTATAAATAGAGGCTCGCGC-TCCGACAGCA-GTCATTC--GAAAAA	1.00
D. ame	CCG---CTGGGTATAAATAGAGCTCGCA-GCAGGCATCTCA-GTCATTC--GAAAAA	1.00
*D. ana**	NNNN---NNNNNNNNNNNNNNNNNNNNN--NNNNNN-NT-GTAGCTGAAAAAAGT	SNA
D. ari	CTC----TGGGTATAAATACAGGCGGACTACTCACATCGCA-GTCATTC--GAAAGT	1.00
D. bia	GGA---TCGGGTATAAATAGGGCCGCCATGGCCGGAGA-AC-GTCACTC--TCCAAG	1.00
D. bip	GCT---TCAGGTATAAATAGAGCTCGTTTTGCAACAGC-AC-GTCACTC--TCGAAA	0.99
D. bus	CAC----AAAGTATAAATAGAGGTGCGCGT-TTGCGATACATTTCATTC--GAATAT	0.95
D. ele	GGA---TCGGGTATAAATAGGCCCGCCCCGAGCTGAGC-AC-GCCACTC--TCAAAG	1.00
D. ere	CGG---ATAAGTATAAATAGAGCCGCCTCTGTCGGGTC-AC-GTCACTC--TCAAAG	1.00
D. eug	GGG---CAGGGTATAAATAGAGCTGCCTCGGCCGGAGA-AC-GTCACTT--TCAAAG	1.00
D. fic	CCA---GCGGGTATAAATAGGGCCGCCTTTTTCGGAGA-AC-GTCACTA--TCAAAT	1.00
D. gri	GGC---CAGCGTATAAATAGAGGCCACAGC-TCGCAATGCA-TTCATTC--GTAAAA	1.00
D. hyd	CCG---CTAGGTATAAATAAAGGTGACTGC-CTGGACTGCA-GTCATTC--GAAAAA	1.00
D. kik	GAC---TCGGGTATAAATAGAGGCGTCGAGGCCAGAGA-AC-GGCACTC--TCAAAA	1.00
D. mel	CGG---TTGGGTATAAATAGAGCCGCCTCAGTCCGGTC-AC-GTCACTC--TCAAAT	1.00
D. mir	AGC---GCTGGTATAAATAGAACTGGTCGAACCAGATA-CT-GCCATTC--GCAAAA	0.92
D. moj	CTC----TGGGTATAAATACAGGCGGACTACTCGCATCGCA-GTCATTC--GAAAAT	1.00
D. mon	CCA---CCGGGTATAAATAGAGCTCGCA-GCAGGCATCGCA-GTCATTC--GAAAAA	1.00
D. nas	ACT---CGGGGTATAAATAGAAGCTCGCGC-TCAGACAGCA-GTCATTC--GAAAAA	1.00
D. nav	CAGCCACTGGGTATAAATAGAGGCGGACTACTCGCTTCGCA-GTTATTC------GA	1.00
D. nov	CCG---CTGGGTATAAATAGAGCTCGCA-GTAGGCATCTCA-GTCATTC--GAAAAA	1.00
D. obs	TTG---GCATGTATAAATAGAACTGGCCGAACCAGACA-TT-GCCACTC--GCAAAA	1.00
D. per	AGC---GCTGGTATAAATAGAACTGGTCGAACCAGATA-CT-GCCATTC--GCAAAA	0.92
D. pse	AGC---GCTGGTATAAATGGAACTGGTCGAACCAGATA-CT-GCCATTC--GCAAAA	0.86
D. rho	AAA---TCGTGTATAAATAGGGCCGCATCAAAATGAGC-TC-GCCACTC--TCGAAG	1.00
D. sec	CGG---TTGGGTATAAATAGAGCCGCCTCAGTCCGGTC-AC-GTCACTC--TCAAAT	1.00
D. ser	GCG---TCGGGTATAAATAGAGGCATCGAGGCCAGAGA-AC-GGCACTC--TCAAAA	0.99
D. sim	CGG---TTGGGTATAAATAGAGCCGCCTCAGTCCGGTC-AC-GTCACTC--TCAAAT	1.00
D. sub	CGG---CAGGGTATAAATAGAACTGGCCGTGCCAGATA-AC-GCCACTC--GCAAAA	0.99
D. suz	GGA---TCGGGTATAAATAGAGGCACCTCGCCCGGAGA-AC-GTCACTC--TCCAAG	1.00
D. tak	CGG---ATGGGTATAAATAGGGCCGCCTCGTCCGGAGA-AC-GTCACTC--TCGCAG	1.00
D. vir	CCG---CTGGGTATAAATACAGCTCGCA-GCAGGCATTTCA-GTCATTC--GGAAAA	1.00

```
D. wil   CGC---TCAGGTATAAATAGGACCGCACATGCCAAGCA-AT-GTCATTC--GAAAAA   0.98
D. yak   CGG---ATTGGTATAAATAGAGCTGCCTCACATGGGGC-AC-GTCACTC--TCAATG   0.99
                                                         *
```

* NNNNNNN represent unidentified nucleotide sequences in this region of *D. ana* genome while the TGAAAAAA sequence matches with the annotated 1st transcription site (Fig. 3) in this species.
Conservation colour code = BAD AVG GOOD

C. The annotated (http://fb2017_05.flybase.org/cgi-bin/gbrowse/dmel/) and predicted Transcription Start Sites (TSS) in proximal region of *hsrω* gene show good concordance in species with annotated genomes while the first Transcription Termination Site (TTS) seems to vary

Species	**Distances (in bp) between**					
	5' splice junction & TSS1		**5' splice junction & TSS2**		**3' splice junction & TTS1**	
	Annotated	**Predicted**	**Annotated**	**Predicted**	**Annotated**	**Predicted**
D. mel	971	971	477	475	699	699
D. ana	NA	1203	414	SNA	550	752
D. moj	NA	735	474	473	786	749
D. sim	NA	978	481	479	2262	700
D. pse	NA	831	16	474	5256	723
D. gri	NA	923	408	508	497	740
D. vir	NA	772	497	498	7468	759
D. yak	NA	994	496	475	13546	707

NA = not annotated at FlyBase; SNA = full sequence not available due to ambiguity; Shaded numbers reflect high mismatch between annotated and predicted distances

Table 8. Repeat units in the *hsrω genes* in different species of *Drosophila.*

Species	Lengths of repeat units and their tandem repetition frequencies	ATAGGTAGG (and variant) Nonamer Frequency	Estimated Total Length
D. alb	No repeats found	20+12 (TTAGGTAGG)	00,000
D. ame	No repeats found	00	00,000
D. ana	No repeats found	04 (+14 TTAGGTAGG)	00,000
D. ari	No repeats found	00	00,000
D. bia	82x2, 152x1.8, 180x2, 228x16, 254x7, 338x2, 399x2, 483x2.5	36	14,912
D. bip	179x3.5 (located downstream of *mir-4951*)	03 (+16 TTAGGTAGG)	00,636
D. bus	61x2, 135x5.4	06	00,722
D. ele	55x2, 164x2, 180x2, 232x3	01 (+27 TTAGGTAGG)	01,010
D. ere	114x2, 230x5.5	12	00,481
*D. eug**	96x39, 61x14, 78x4, 87x2, 157x16, 175x5, 195x2.5, 253x6, 294x3.3	03 (+14 TGTGCTAGG)	04,542
D. fic	77x2	11	00,145
D. gri	60x5, 110x2, 155x25.7, 464x8.6	01 (+27 CTAGGTAGG)	04,205
D. hyd	57x2, 115x92, 170x4, 190x2, 228x44, 258x2, 363x2.3	45 (+39 TTAGGTAG)	10,826
D. kik	62x18.6, 176x17, 245x2.2, 309x4.2, 350x8.4, 429x2.2	31	04,144
D. mel	87x3,278x37	68	10,535
D. mir	343x2, 356x3.7	27	02,167
D. moj	60x187, 124x56.6, 156x1.8, 180x49.5, 246x38.6, 418x1.8, 483x9.8	00 (+76 TTAGGTAGG)	11,935
*D. mon**	87x3, 132x2.2, 162x4, 219x2.3, 294x2.2	00 (+5 CCTATTAGG)	01,386
*D. nas**	Sequence not available	00	--------
D. nav	No repeats found	00	00,000
D. nov	85x57.7, 99x2, 140x23, 215x27, 303x2, 365x8, 455x2.4	00 (+85 TTAGGTAGG)	11,078
*D. obs**	No repeats found	05	00,000
D. per	156x2, 360x9	43	03,604
D. pse	No repeats found	22	00,000
*D. rho**	135x2.6, 142x2, 260x11.6	02 (+28 TTAGGTAGG)	03,027
*D. sec**	76x9, 260x2, 280x22, 354x5	37	06,579
D. ser	50x4.1, 67x33, 160x25, 213x14, 232x2, 280x15, 368x2, 468x3, 493x10.3	53	09,156
D. sim	278X12	29	03,331
D. sub	No repeats found	01	00,000
D. suz	128x2.1, 311x2.5, 336x4.2, 354x5	41	04,306
D. tak	205x4, 247x7, 490x2	18	03,396
D. vir	36x236, 42x2.7, 83x27.5, 114x2, 124x6, 137x41, 220x14.4, 273x2.7, 356x6.2, 437x2	25	05,883
D. wil	63x2, 135x2, 197x125, 332x2	00 (+58 ATAGGCAGG)	24,363
D. yak	89x56, 128x2.2, 180x14.5, 220x4, 288x36,376x2, 463x9.5	31	09,349

* Complete *hsrω* gene sequences are not available for these species

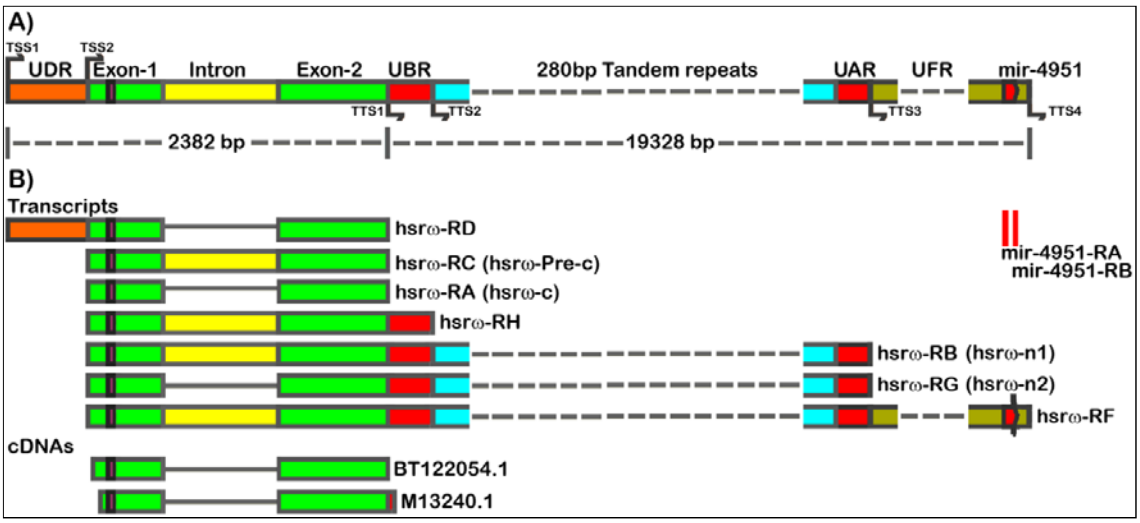


Fig: 1. Genomic coordinates and predicted transcripts of the *hsrω* gene in *D. mel* (http://flybase.org/reports/FBgn0001234). A Organization of the *hsrω* gene as available at FlyBase: two upward arrows in indicate the two transcription start sites (TSS1 and TSS2) while the four downward arrows indicate the four annotated transcription termination sites (TTS1-TTS4). **B** Transcripts produced by the *hsrω* gene: unspliced intron in transcripts is shown as yellow box while the spliced out intron is indicated by a black line joining the exon 1 and exon 2 regions (green boxes). UDR = region unique to RD transcript, UBR = unique region before tandem repeats between TTS1 and TTS2, UAR = unique region after tandem repeats till the TTS3; UFR = unique region specific to the RF transcript, *mir-4951* = miRNA gene overlapping with the *hsrω* gene at its 3' end, broken lines indicate the 280 bp tandem repeats (redrawn after http://flybase.org/reports/FBgn0001234).

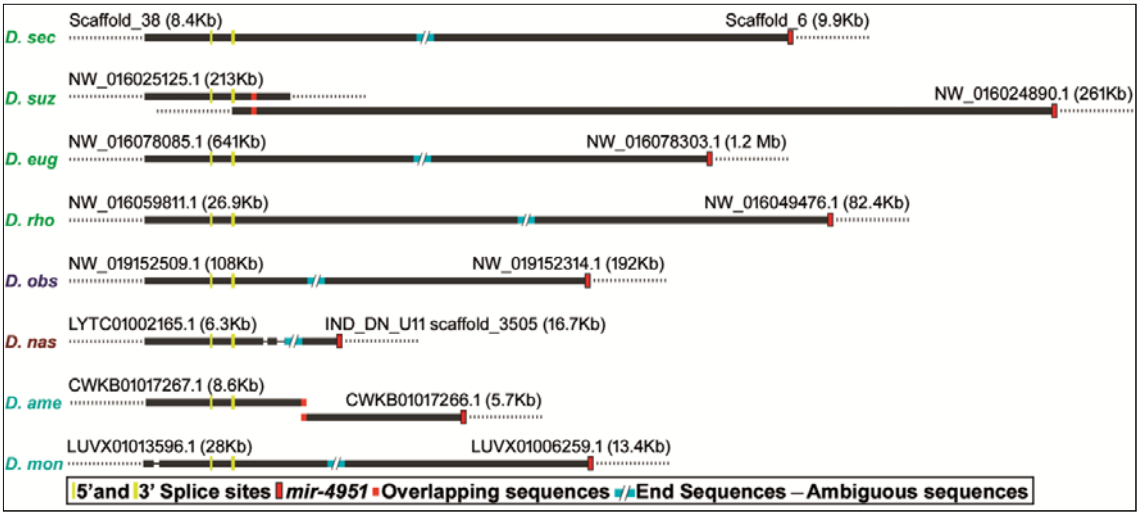


Fig: 2. Strategy for assembly of *hsrω* carrying contigs in species with the splice junctions and *mir-4951* retrieved in separate scaffolds. Schematics of the scaffolds carrying splice junctions or the *mir-4951* sequences in different species are shown as solid thick (putative *hsrω* regions) and dotted lines (flanking regions). The yellow and red vertical bars represent the splice junctions and the *mir-4951* sequences, respectively; orange vertical bars indicate regions of sequence overlap between the two scaffolds in *D. ame* and *D. suz*, while slanted turquoise bars represent non-overlapping but presumably facing ends of the two scaffolds in six species.

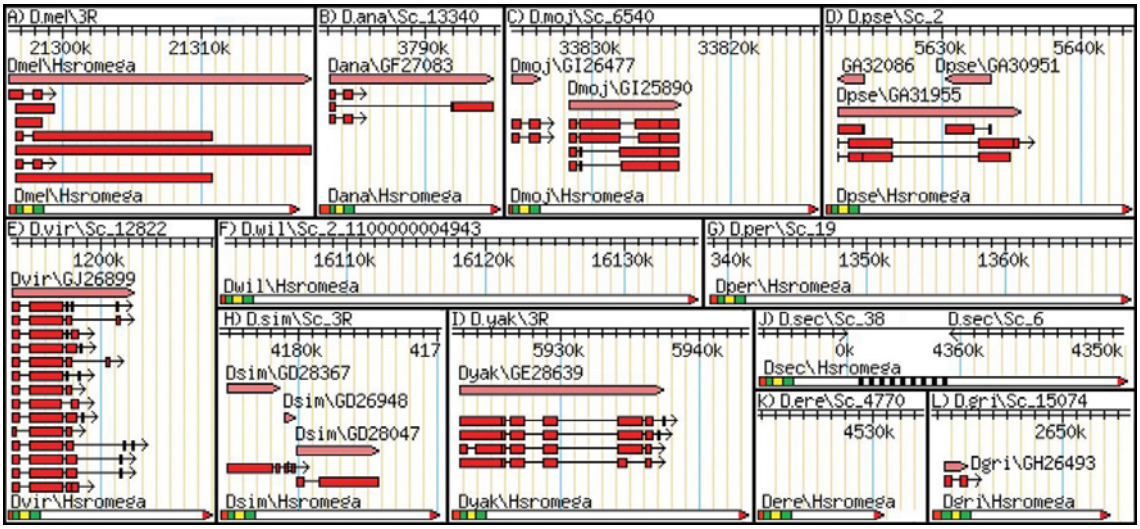


Fig: 3. A-L GBrowse genomic co-ordinates of *hsrω* locus in the 12 annotated *Drosophila* genomes available at the FlyBase (Dmel Release 6.18). A-L Species and the respective contig names are noted above the chromosomal scale for each species. Light red bars denote the *hsrω* gene with the direction of transcription indicated by the pointed end; the deep red bars represent the annotated transcripts; no transcripts are annotated for the *hsrω* gene in *D. wil* (**F**)**,** *D. per* (**G**), *D. sec* (**J**) and *D. ere* (**K**). A manually drawn map at the lower end of each panel denotes the putative *hsrω* gene, starting with TSS1 and ending at the *mir-4951*; the light red box at proximal end corresponds to region between TSS1 and TSS2 (as in *D. mel*), green boxes represent the two exons, the yellow box the *omega* intron while the red arrow head represents the *mir-4951* sequence; alternating black and white bars in middle of the *D. sec hsrω* gene (**J**) indicate the missing part between the two identified scaffolds.

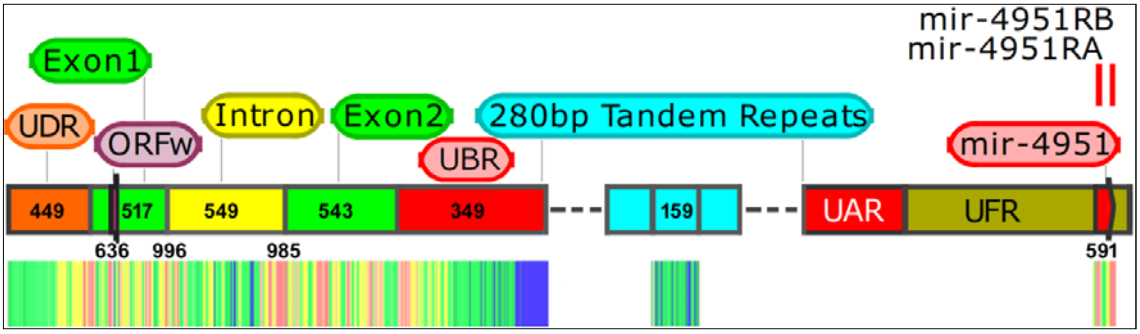


Fig: 4. Common architecture but variable sequence conservation across the *hsrω* gene in different *Drosophila* species. Overall architecture of the *hsrω* gene in *D. mel* (upper part) and heat map of conservation of different regions taking all the 34 species together (lower part). Numeric values represent conservation score of the corresponding regions. Conservation colour code = BAD AVG GOOD

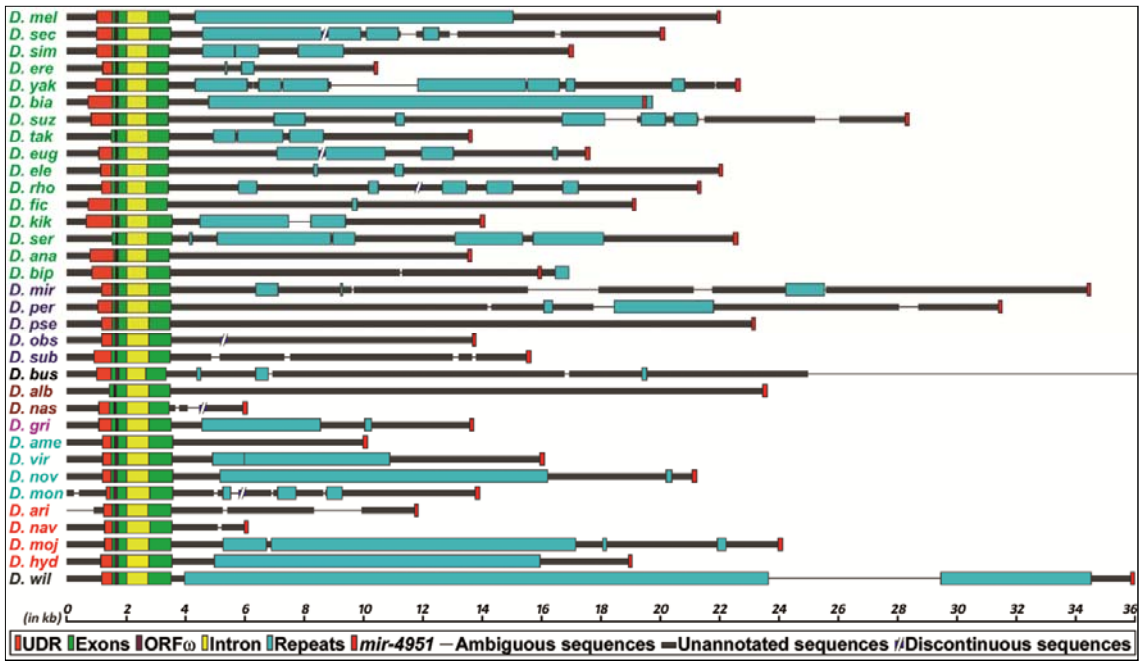


Fig: 5. Predicted architecture of *hsrω* gene in the genus *Drosophila.* A full length SnapGene annotated map of *hsrω* in all species beginning from 2kb upstream of 5' splice junction and ending with *mir-4951*. Different domains in this gene includes two exons (green), the ORFω (purple), the *omega* intron and the UDR (orange) in the proximal part and the *mir-4951* at the 3' end and the most diversified tandem repeat region (sky-blue) in the middle part. The discontinuity in six species indicates that the two scaffolds, presumably representing proximal and distal regions of the *hsrω*, did not show overlap to enable assembly of a contiguous region; the thin light grey lines indicate ambiguous sequence.

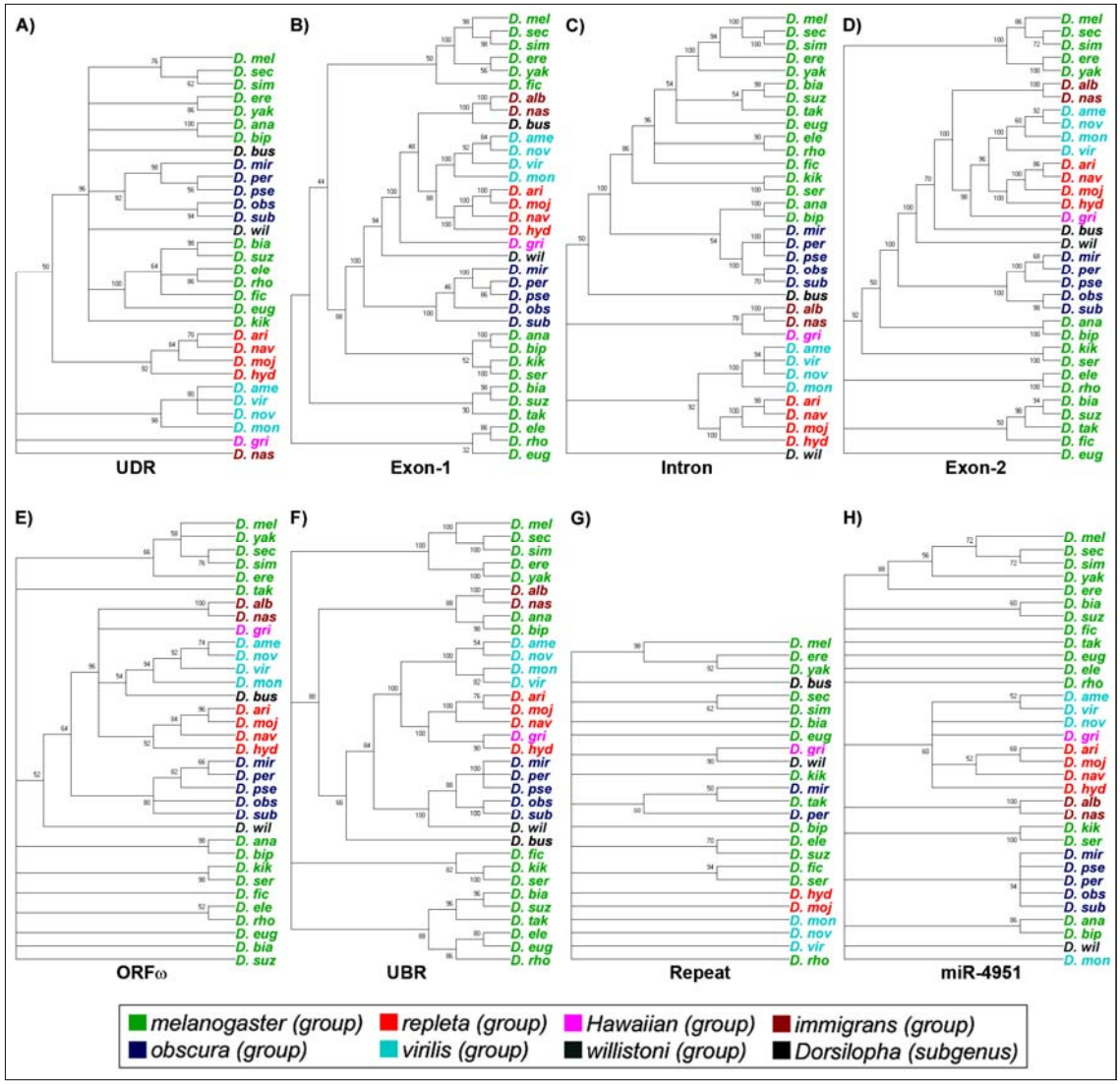


Fig: 6. Molecular phylogenetic relationships for different regions of the *hsrω* gene by Maximum Likelihood method. Phylogenetic trees for different regions (UDR, Exon-1, Intron, Exon-2, ORFω, UBR, Repeat and *mir-4951*) of *hsrω* gene from 34 species are shown in **A-H**. Numeric value at the base of each branch point represents the homology index between two branches.

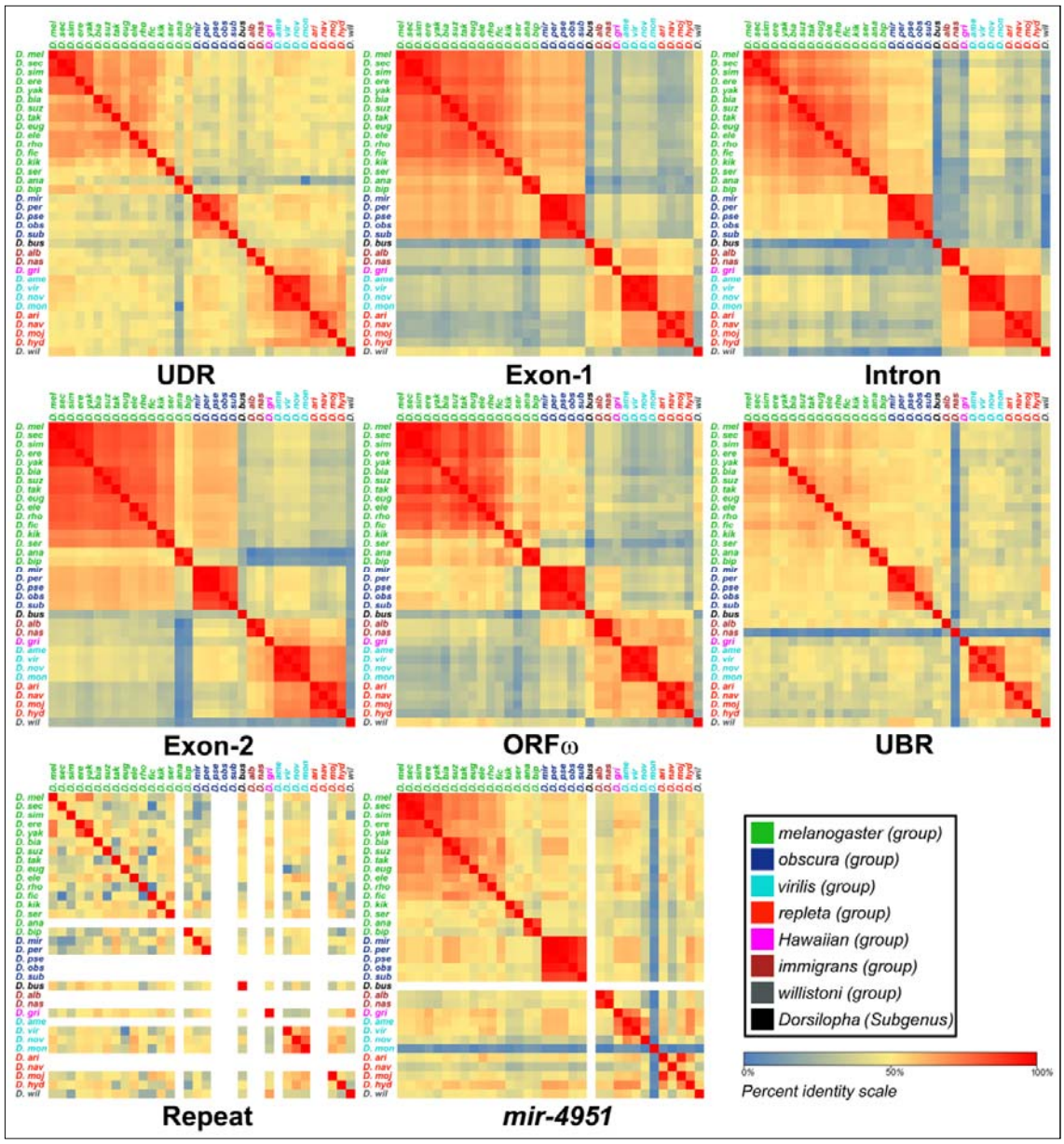


Fig. 7. Percent Identity Matrix analysis confirms variable but independent divergences in sequences in different domains of the *hsrω* gene in the 34 Drosophila species. Colour codes for heat-maps and for species groups are given below the matrices. White Rows and columns in PIM for Repeats and *mir-4951* reflect absence of the corresponding sequences in some species.

Supplementary Table S1. List of *Drosophila* genes sharing splice junctions similar to the 5' or 3' *omega* splice junction sequences

A. *Drosophila melanogaster* genes sharing homology with 16 nt *omega* intron 5' splice junction

Anti-sense orientation			Sense orientation		
Genes	**Coverage**	**Identity**	**Genes**	**Coverage**	**Identity**
1. abd-A	11	11	*1. AOX2*	11	11
2. AGO3	11	11	*2. Apoltp*	11	11
3. Appl	11	11	*3. aur A*	11	11
4. bru 3	13	13	*4. CanA1*	11	11
5. Cac	11	11	*5. CG14190*	11	11
6. CG11353	12	12	*6. CG1492*	11	11
7. CG1416	11	11	*7. CG17230*	12	12
8. CG1545	11	11	*8. CG5986*	11	11
9. CG5756	11	11	*9. CG8176*	12	12
10. CG8405	11	11	*10. Cpr50Cb*	11	11
11. CG9492	15	14	*11. Emp*	11	11
12. Cht6	11	11	*12. GNBP3*	11	11
13. CR44289	12	12	*13. Hang*	11	11
14. Cstf 64	11	11	*14. heph*	12	12
15. Dpr12	11	11	*15. Hr38*	11	11
16. Drl	11	11	*16. Hr4*	11	11
17. Ear	11	11	*17. hyd*	11	11
18. Eip78c	11	11	*18. Ide*	11	11
19. ERR	11	11	*19. Klar*	11	11
20. Fur1	11	11	*20. Lac*	11	11
21. Gmd	12	12	*21. Lmpt*	11	11
22. Gprs	11	11	*22. mspo*	11	11
23. gstE13	11	11	*23. Mthl10*	11	11
24. hppy	11	11	*24. Mus205*	12	12
25. INO80	11	11	*25. Myo28B1*	11	11
26. kay	11	11	*26. Nek2*	11	11
27. Loco	11	11	*27. nompB*	11	11
28. Mctp	11	11	*28. Ocm*	11	11
29. mtd	11	11	*29. Or67b*	11	11
30. nemy	11	11	*30. pnt*	11	11
31. Not1	13	13	*31. Prm*	11	11
32. opa	11	11	*32. PUS1*	11	11
33. Or30a	11	11	*33. Rab27*	11	11
34. Ovo	11	11	*34. ref1*	11	11
35. pasi1	11	11	*35. Rip11*	12	12
36. Pxd	11	11	*36. Tai*	11	11
37. pzl	11	11	*37. Tmc*	11	11
38. Sky	11	11	*38. trv*	11	11
39. SLO2	11	11	*39. Usp7*	11	11
40. T3dh	11	11	*40. Vav*	11	11
41. Tgi	11	11	*41. Wdr62*	11	11

B. *Drosophila melanogaster* genes sharing homology with 60 nt *omega* intron 3' splice junction

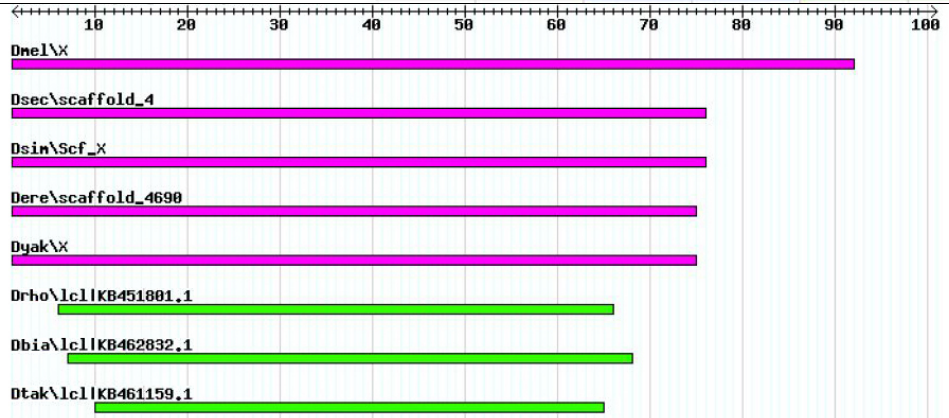
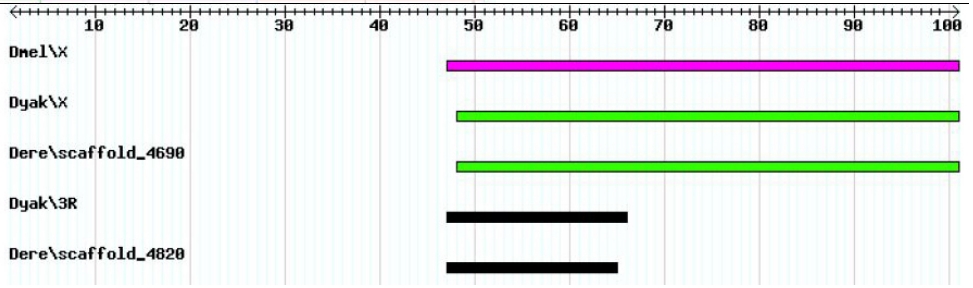
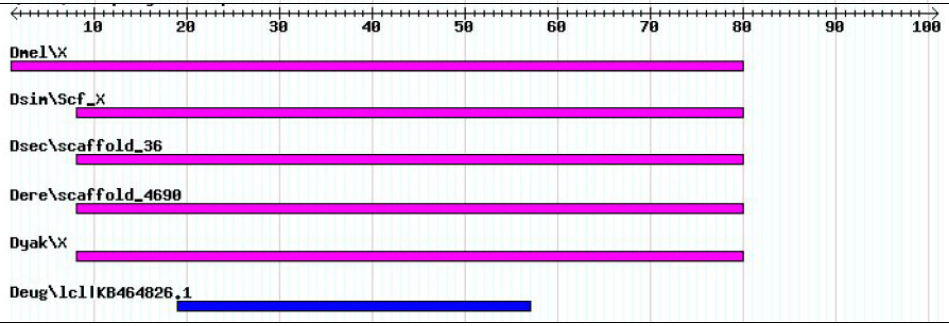
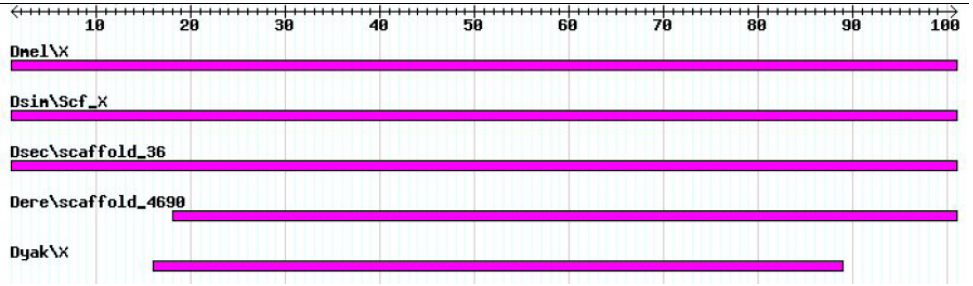
Anti-sense orientation			Sense orientation		
Genes	**Coverage**	**Identity**	**Genes**	**Coverage**	**Identity**
1. ALK	24	22	*1. fasciclin 3*	20	19

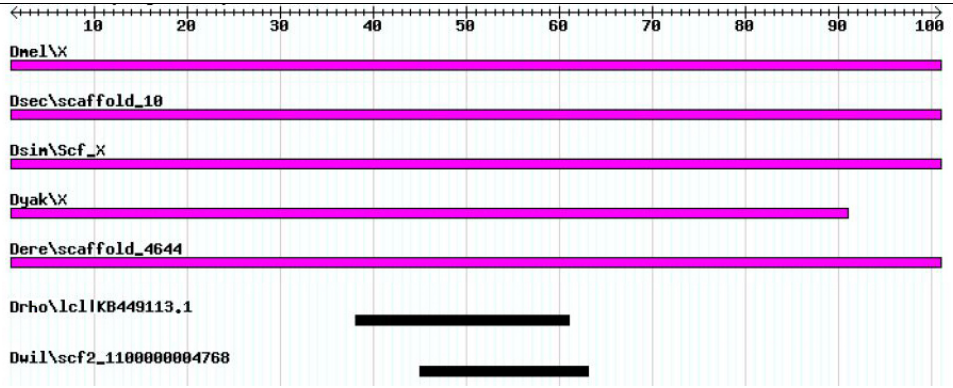
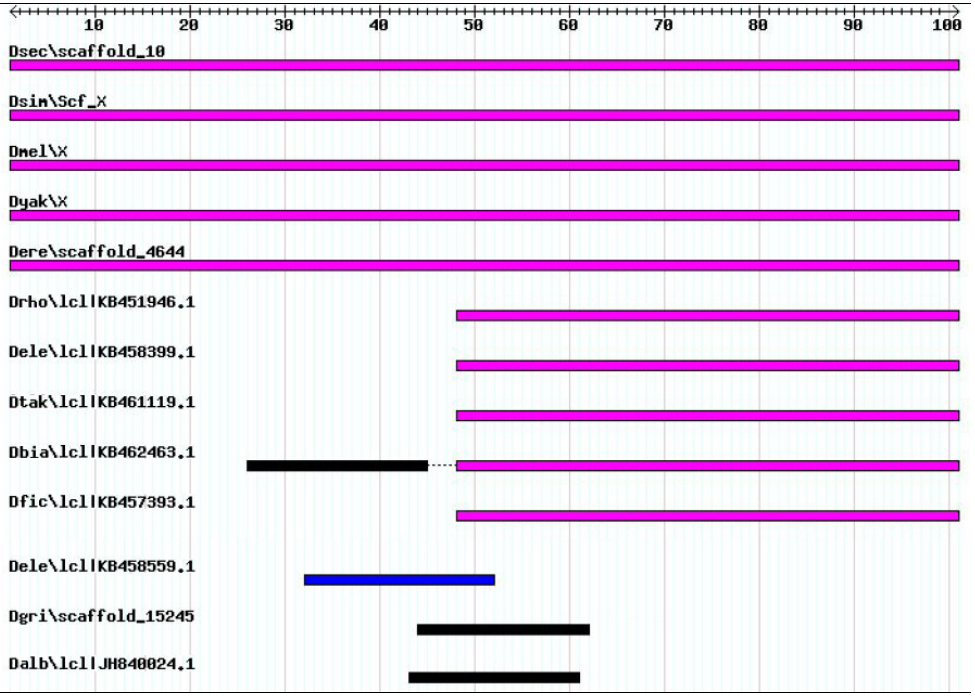
Supplementary Table S3. Putative TSS3, predicted to overlap with the ORF-□, in *Drosophila* species

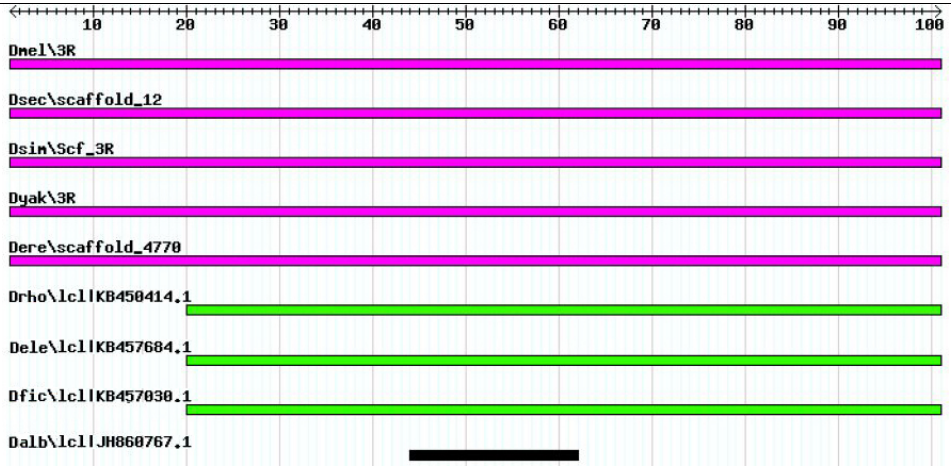
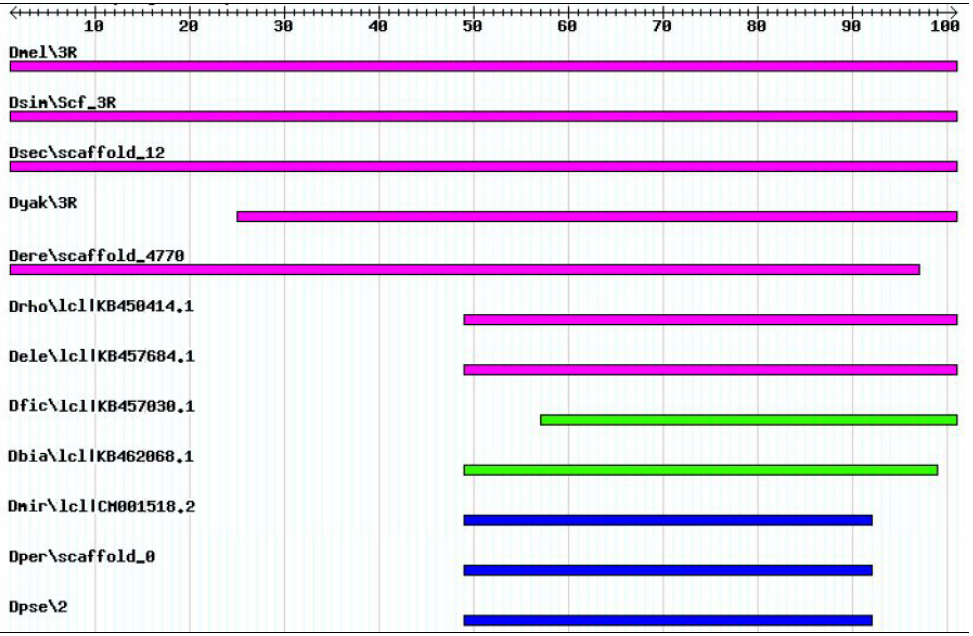
Species	TSS3 sequence	Score
D. alb	TGAA-AA----AGTGTAAAACAACTGCCT--TAAAGGCAAAGCTGAGA----AGTT--ACAAC	0.85
D. ame	ATACC-A----TGTATATATATTGTTCAA--CGGCA---ATGCAAAAGTGTGCAG---AAACA	0.96
D. ana	----------	NF
D. ari	----------	NF
D. bia	GAA---A----AGTGTAAAAATCGTGTCC--ACCGAGACCAGCAGCAG----TACGAGTTTCG	0.97
D. bip	----------	NF
D. bus	----------	NF
D. ele	ATGGAAA----AGTGTAAAAATCGCGTCG--CAGAAGCAGAGCTGCAG----TAC---GAGTA	0.99
D. ere	ATGGAAA----AGTGTAAAAATCGTGTCC--CAGCAGACGAGCAGCAG----AAC---GAGTA	0.97
D. eug	ATGGAAA----AGTGTAAAAATTGTGTCG--CAGAAGGAGACGGACAG----CTG---TACGA	0.97
D. fic	----------	NF
D. gri	----------	NF
D. hyd	ATGGAGA----AGTGTACAGTATTTGTTG--CGAAGGCGAGATTGCGG----AGC---TATAA	0.93
D. kik	----------	NF
D. mel	ATGGAAA----AGTGTAAAAATCGTGTCC--CAGCAGACGAGCAGCAG----CAG---TACGA	0.99
D. mir	CTGGC-T----AGCATATATAATCATCAA--CGAAAAAAATGGAAAAGTGT---G---GAAAA	0.82
D. moj	ATAGT-TGCTATATATAAAGATTACTCAA--CTGCA---ATGAAAAAGTGT-------GCAGA	0.94
D. mon	ATACC-A----TGCATATATATTGTTCAA--CGGCG---ATGAAAAAGTGTGCAG---AAACA	0.88
D. nas	TGAA-AA----AGTGTAAAACAACTGCCT--TAAAGGCAAAGCTGAGA----AGTT--ACAAC	0.85
D. nav	----------	NF
D. nov	ATACC-A----TGTATATATATTGTTCAA--CGGCA---ATGCAAAAGTGTGCAG---AAACA	0.96
D. obs	----------	NF
D. per	CTGGC-T----AGCATATATAATCATCAA--CGAAAAAAATGGAAAAGTGT---G---GAAAA	0.82
D. pse	CTGGC-T----AGCATATATAATCATCAA--CGAAAAAAATGGAAAAGTGT---G---GAAAA	0.82
D. rho	ATGGAAA----AGTGTAAAAATCGCGTCG--CTGAAGCCGAGCAGCAG----TAC---AAGTA	0.98
D. sec	ATGGAAA----AGTGTAAAAATCGTGTCC--CAGGAGGCGAGCAGCAG----TAC---GAGTA	0.98
D. ser	----------	NF
D. sim	ATGGAAA----AGTGTAAAAATCGTGTCC--CAGGAGGCGAGCAGCAG----TAC---GAGTA	0.98
D. sub	----------	NF
D. suz	ATGGAAA----AGTGTAAAAATCGTGACG--CAGACGAGCAGCAGCAG----TAC---GAGTT	0.98
D. tak	ATGGAAA----AGTGTAAAAATCGTGTCG--CACCAGACGAGCAGCAG----TAC---GAGTA	0.98
D. vir	ATACC-A----TGTATATATATTGTTCAA--CGGCA---ATGCAAATGTGTGCAG---AAACA	0.96
D. wil	TTTGTGCA---TTTATATATATAGAATGAGTCATACGTTATGGAAAAGT----------GTAA	0.98
D. yak	ATGGAAA----TGTGTAAAAATTGTGTCC--CAGCAGACGAGCAGCAG----TGC---GATTA	0.98
	** *	

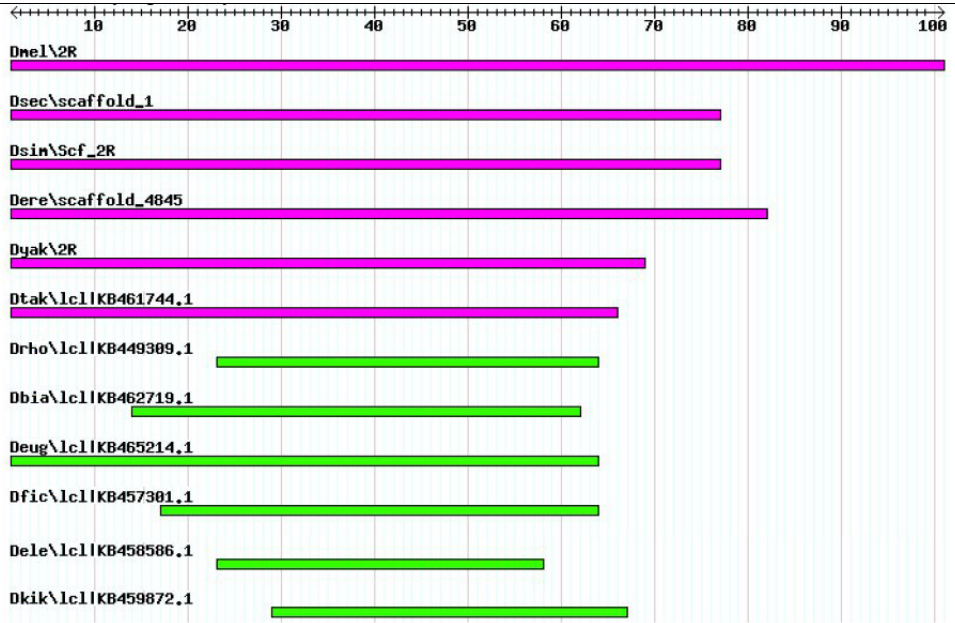
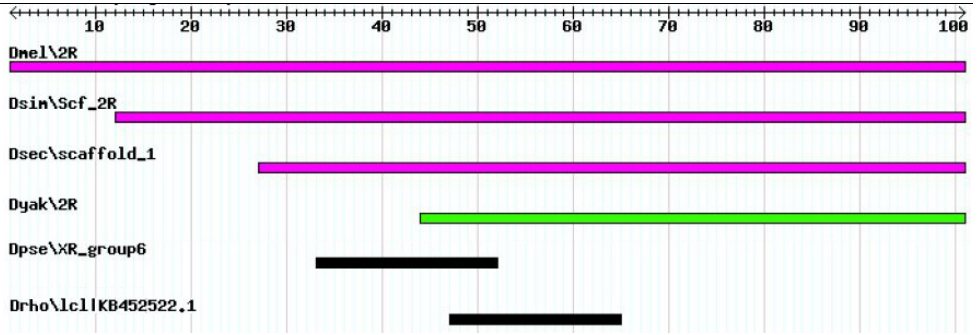
NF = sequence not found

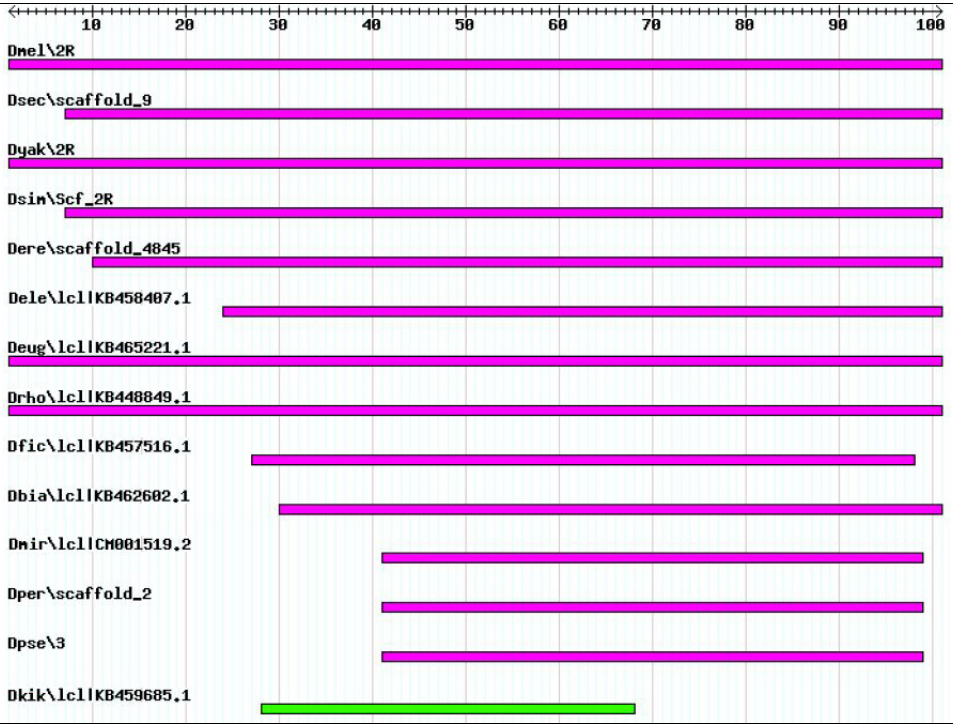
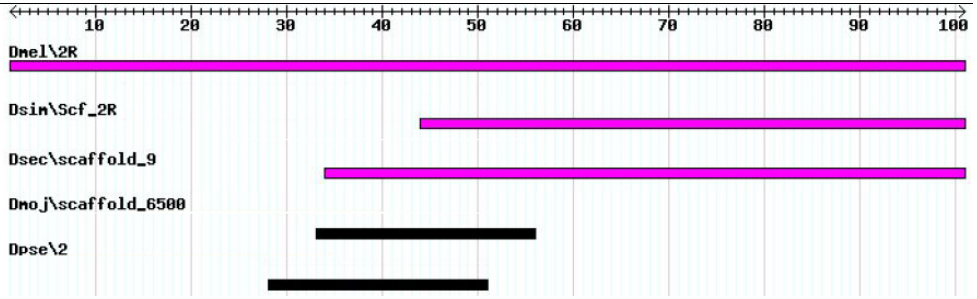
Supplementary Table S5. Relative conservation of base sequences flanking 5' and 3' splice junctions of 1st intron in different lncRNA genes in different *Drosophila* species

Gene	5' Splice junction	3' Splice junction
roX1		
roX2		

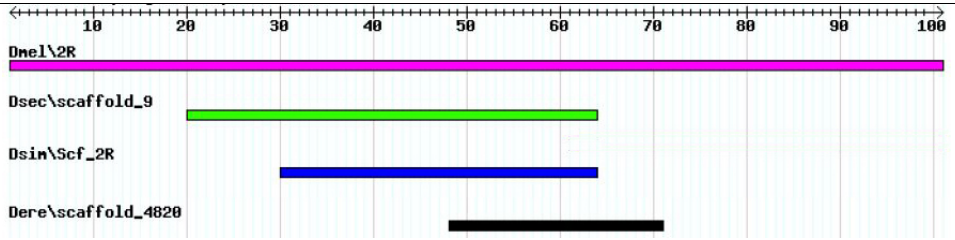
yar	 10 20 30 40 50 60 70 80 90 100 Dmel\X Dsec\scaffold_10 Dsim\Scf_X Dyak\X Dere\scaffold_4644 Drho\lcl\KB449113.1 Dwil\scf2_1100000004768	 10 20 30 40 50 60 70 80 90 100 Dsec\scaffold_10 Dsim\Scf_X Dmel\X Dyak\X Dere\scaffold_4644 Drho\lcl\KB451946.1 Dele\lcl\KB458399.1 Dtak\lcl\KB461119.1 Dbia\lcl\KB462463.1 Dfic\lcl\KB457393.1 Dele\lcl\KB458559.1 Dgri\scaffold_15245 Dalb\lcl\JH840024.1

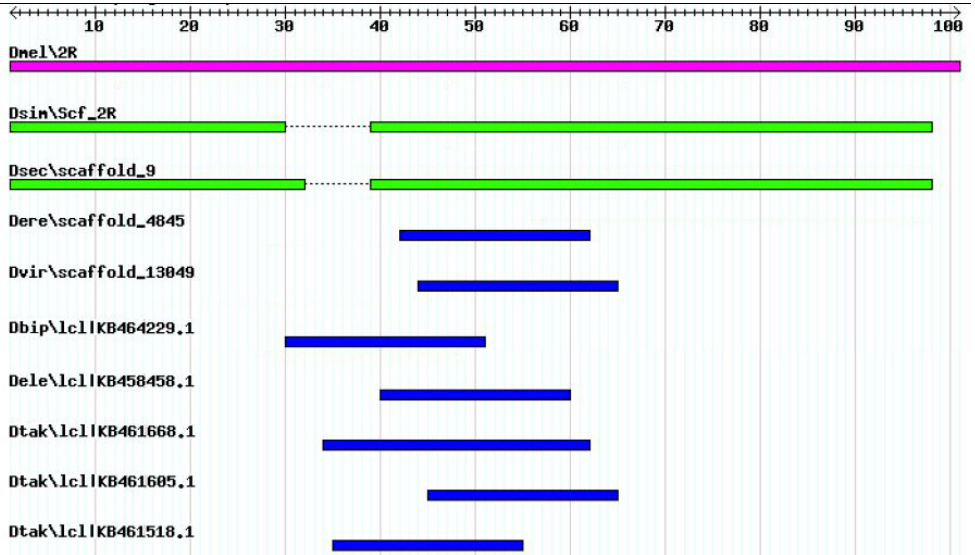
bxd	 10 20 30 40 50 60 70 80 90 100 Dmel\3R Dsec\scaffold_12 Dsim\Scf_3R Dyak\3R Dere\scaffold_4770 Drho\lcl\KB450414.1 Dele\lcl\KB457684.1 Dfic\lcl\KB457030.1 Dalb\lcl\JH860767.1	 10 20 30 40 50 60 70 80 90 100 Dmel\3R Dsim\Scf_3R Dsec\scaffold_12 Dyak\3R Dere\scaffold_4770 Drho\lcl\KB450414.1 Dele\lcl\KB457684.1 Dfic\lcl\KB457030.1 Dbia\lcl\KB462068.1 Dmir\lcl\CM001518.2 Dper\scaffold_0 Dpse\2

CR30121	 10 20 30 40 50 60 70 80 90 100 Dmel\2R Dsec\scaffold_1 Dsim\Scf_2R Dere\scaffold_4845 Dyak\2R Dtak\lcl\KB461744.1 Drho\lcl\KB449309.1 Dbia\lcl\KB462719.1 Deug\lcl\KB465214.1 Dfic\lcl\KB457301.1 Dele\lcl\KB458586.1 Dkik\lcl\KB459872.1	 10 20 30 40 50 60 70 80 90 100 Dmel\2R Dsim\Scf_2R Dsec\scaffold_1 Dyak\2R Dpse\XR_group6 Drho\lcl\KB452522.1

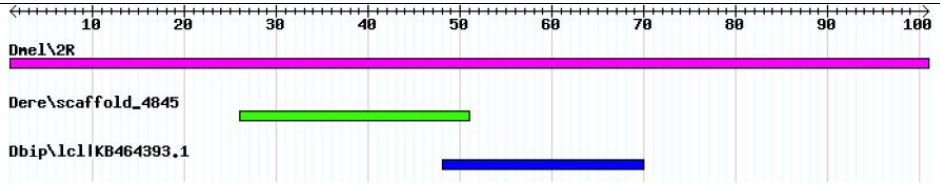
CR30267	 10 20 30 40 50 60 70 80 90 100 Dmel\2R Dsec\scaffold_9 Dyak\2R Dsim\Scf_2R Dere\scaffold_4845 Dele\lcl\KB458407.1 Deug\lcl\KB465221.1 Drho\lcl\KB448849.1 Dfic\lcl\KB457516.1 Dbia\lcl\KB462602.1 Dmir\lcl\CM001519.2 Dper\scaffold_2 Dpse\3 Dkik\lcl\KB459685.1	 10 20 30 40 50 60 70 80 90 100 Dmel\2R Dsim\Scf_2R Dsec\scaffold_9 Dmoj\scaffold_6500 Dpse\2

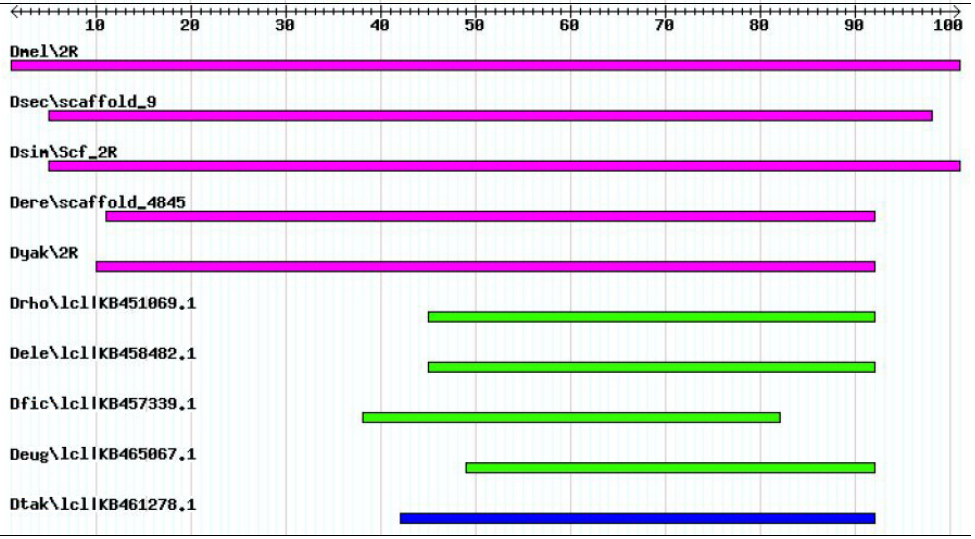
CR32835

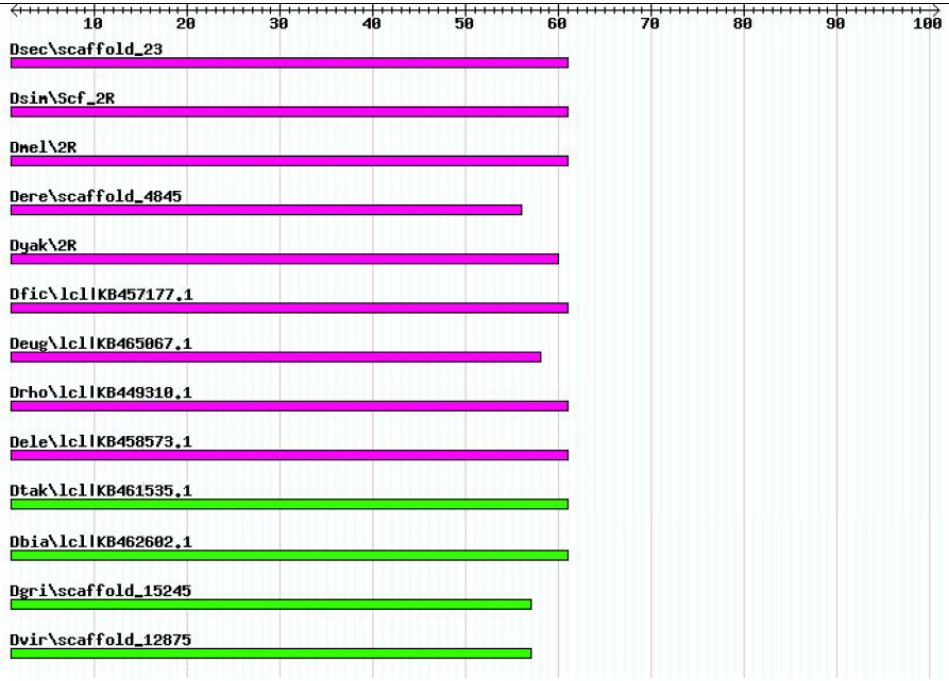
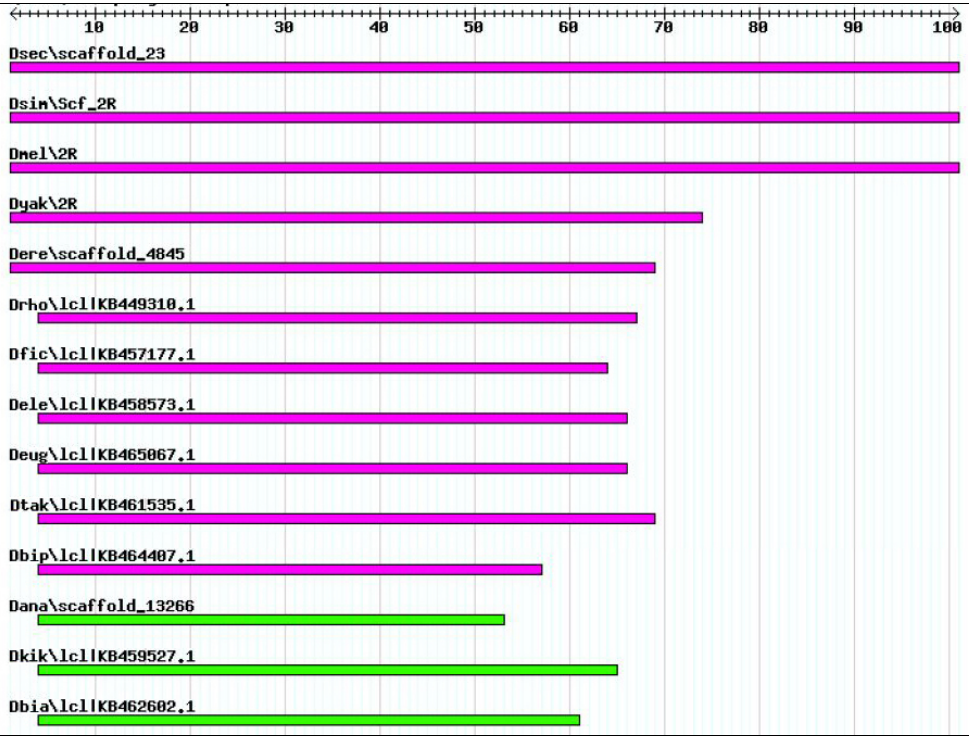


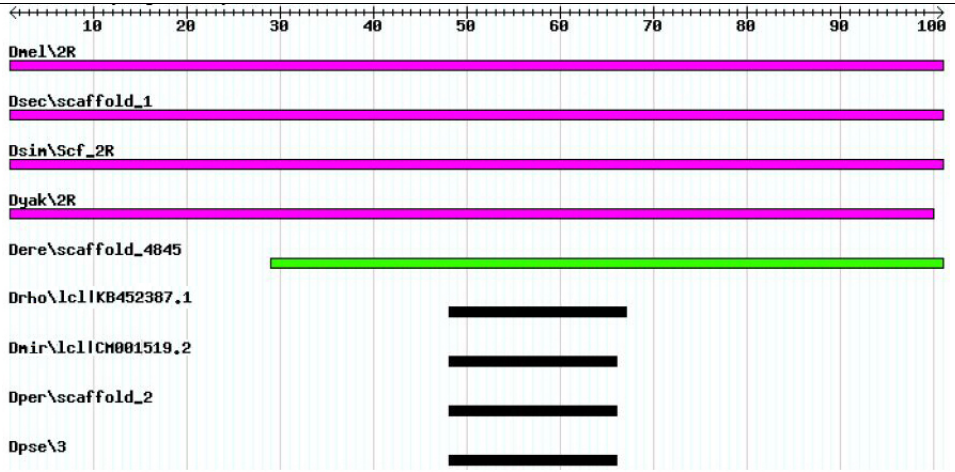
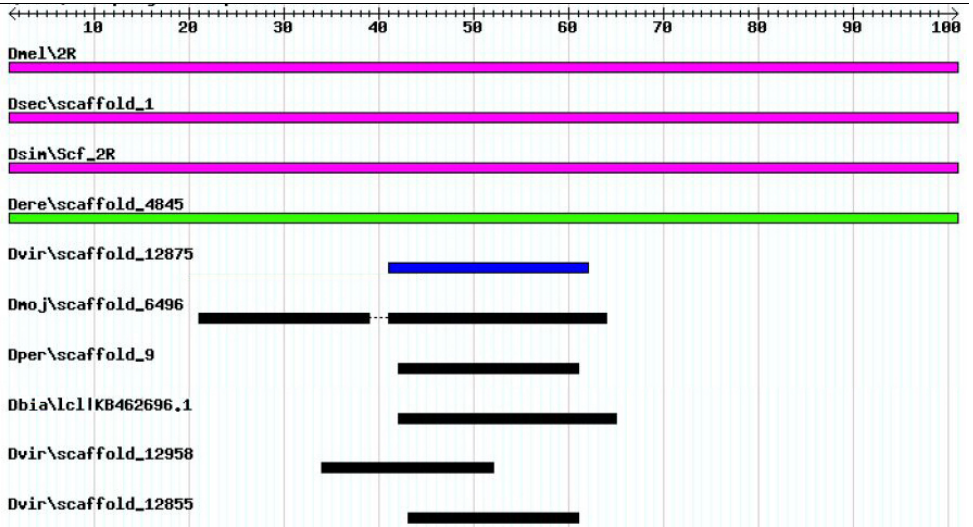


CR44807





CR44810	 10 20 30 40 50 60 70 80 90 100 Dsec\scaffold_23 Dsim\Scf_2R Dmel\2R Dere\scaffold_4845 Dyak\2R Dfic\lcl\KB457177.1 Deug\lcl\KB465067.1 Drho\lcl\KB449310.1 Dele\lcl\KB458573.1 Dtak\lcl\KB461535.1 Dbia\lcl\KB462602.1 Dgri\scaffold_15245 Dvir\scaffold_12875	 10 20 30 40 50 60 70 80 90 100 Dsec\scaffold_23 Dsim\Scf_2R Dmel\2R Dyak\2R Dere\scaffold_4845 Drho\lcl\KB449310.1 Dfic\lcl\KB457177.1 Dele\lcl\KB458573.1 Deug\lcl\KB465067.1 Dtak\lcl\KB461535.1 Dbip\lcl\KB464407.1 Dana\scaffold_13266 Dkik\lcl\KB459527.1 Dbia\lcl\KB462602.1

asRNA CR45146		

100 nt sequence (50 nt upstream and 50 nt downstream of the actual splice junction) of the given lncRNA gene (column 1) in each species was used for BLAST with the corresponding region in *D. mel* gene. The extent of similarity is indicated by horizontal colour bars, reflecting the Conservation score code given at top. Relative location, in relation to the nt scale given at top of each panel, of the horizontal bars indicates the region of similarity. *Drosophila* species which did not show any similarity in this region are not shown.

For Supplementary Tables S2 and S4, see the attached xlsx files