Comparative sequence analysis of fruit fly PRF candidates predicted by the Lin et al [26] study

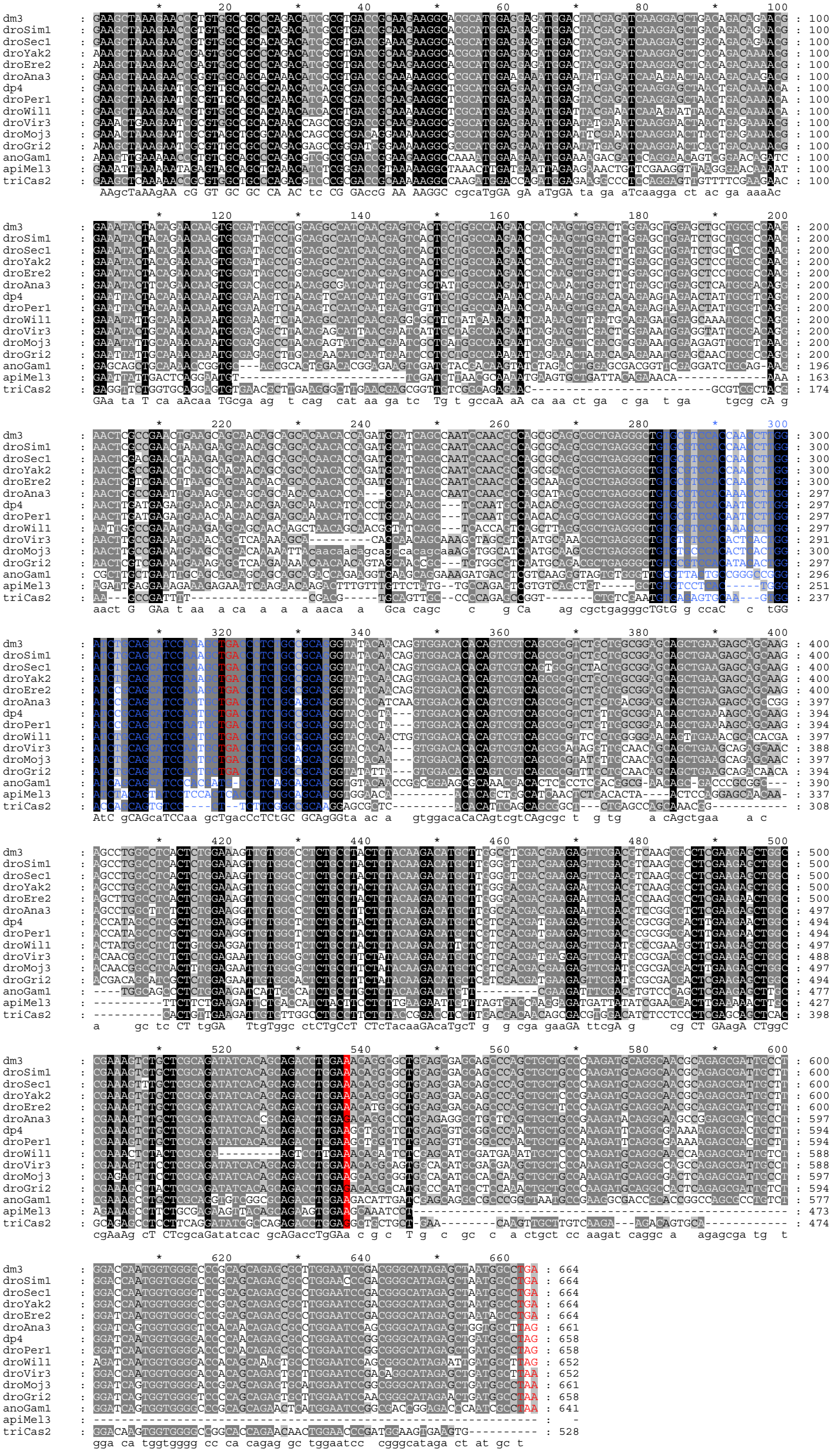
Xbp1. The predicted shift between reading frames has been reported at chr2R:17,032,241 (the coordinates from the 2006 *D. melanogaster* genome assembly are used throughout, the locations for the 2004 assembly were taken from the supplementary material in Lin et al [26] supplementary material and then converted by the LiftOver tool at FlyBase [32]). The alignment of sequences corresponding to this location from 15 insect species is shown in Supplementary Figure 1 in Text S1. The corresponding Refseq mRNA NM_166427 contains two long overlapping ORFs, the CDS is annotated with the start at the first ORF and the end corresponding to the end of the second ORF. The annotated protein sequence corresponds to the product that would be obtained if -1 ribosomal frameshifting would take place at the end of the first ORF. However, no frameshift site is reported in the annotation of NM_166427 mRNA even though the length of the annotated CDS is not a multiple of three and conceptual translation of the corresponding CDS sequence terminates at the internal stop codon. Genbank contains three mRNA entries FJ630003, BT021380 and AY059442; the first two report CDS corresponding to the first ORF, while the third reports CDS corresponding to the second ORF. The evolutionary signature switch between the two ORFs reported by Lin et al [26], occurs in the overlapping region upstream of the stop codon for the first ORF (see Supplementary Figure 1 in Text S1).

To reveal a plausible molecular mechanism responsible for the fusion of the two ORFs, we examined multiple alignments of the corresponding region and identified the potential exon shown in Supplementary Figure 1 in Text S1. The splice sites of this exon are highly conserved indicating their likely functionality (though we were unable to identify ESTs supporting the corresponding splice variant). However, if this exon is spliced, it will lead to the formation of an alternative mRNA isoform where two long ORFs would be fused into one. We believe that alternative splicing is the more likely mechanism used for fusion of two ORFs than PRF, since a nucleotide alignment of the overlapping region lacks conserved patterns that resemble potential frameshift-prone patterns.

CG32736. The predicted shift between translational reading frames has been reported at chrX:6,905,921. Supplementary Figure 2 in Text 1S shows nucleotide alignments of the corresponding region. The transition between reading frames, from 0 to -1, coincides with a highly conserved AUG codon corresponding to an ORF in the -1 frame. In most sequences, there are stop codons upstream of this AUG, hence the most plausible explanation for this situation is that *cg32736* encodes a bicistronic mRNA where initiation at the second CDS may actually be reinitiation and depend on termination of the first CDS. Although the reported location is slightly upstream of the conserved AUG, most likely it is due to the lack of conservation of protein sequence at the 3'-end of the first CDS, where purifying selection is weak or lacking. The likely possibility of reinitiation at the downstream CDS in combination with lack of other conserved features that may indicate the existence of a site for programmed ribosomal frameshifting, forced us to conclude that this location is also not a real PRF candidate.

CG14047. The frame switch predicted based on evolutionary signatures has been reported at chrX:2,264,455. CG14047 encodes a protein product from a single ORF and its synthesis does not require ribosomal frameshifting. However, as detected in the Lin et al study [26], evolutionary signatures corresponding to the sequence upstream of the predicted switch indicate that selection acts on the -1 frame. At its 5' end *D. melanogaster* CG14047 contains a long ORF interrupted by a single stop codon. See ORF plot in Supplementary Figure 3 in Text 1S. The location of the frameshift switch is slightly upstream of the stop codon in the alternative frame. We re-examined the Lin et al analysis using the MLOGD program [33] that predicts the likelihood of a particular frame encoding a protein sequence based on purifying selection and detected by the analysis of multiple alignments. The MLOGD plot built using alignments of transcripts (accession numbers Dwil\GK16212-PA, Dmoj\GI15916-PA, Dgri\GH12813-PA, Dvir\GJ16823-PA, Dpse\GA22666-PA, Dper\GL26908-PA, Dana\GF21955-PA, Dyak\GE16950-PA, Dere\GG12618-PA, Dsim\GD16369-PA) obtained from FlyBase [32] is shown in Figure 4 in Text S1. Our

analysis is consistent with the Lin et al prediction that purifying selection is acting on the -1 frame upstream of the predicted switch. However as evident from the plot, the long ORF present in *D. melanogaster*, is not conserved. It is interrupted by multiple stop codons in homologous sequences from other flies. Therefore, if frameshifting takes place during CG14047 translation, the site of frameshifting should be located in a relatively short region. Figure 5 in Text 1S shows multiple sequence alignments in the vicinity of the frame switch predicted by Lin et al. While conservation of protein sequence encoded by the -1 frame is apparent, we were unable to reveal any conserved nucleotide pattern in this region that may account for the possibility of ribosomal frameshifting. There might be several alternative plausible explanations why the sequence reported indicates evolutionary selection in the alternative frame. The region is relatively short and the evolutionary signatures may exist in it for stochastic reasons (though estimation of the likelihood of such a situation is not trivial). Perhaps, the corresponding region evolved under positive or disruptive selection or encodes additional regulatory signals in mRNA that results in a pattern of codon changes mimicking purifying selection corresponding to -1 frame. Nonetheless we cannot entirely exclude the possibility of ribosomal frameshifting in CG14047 without experimental investigation.

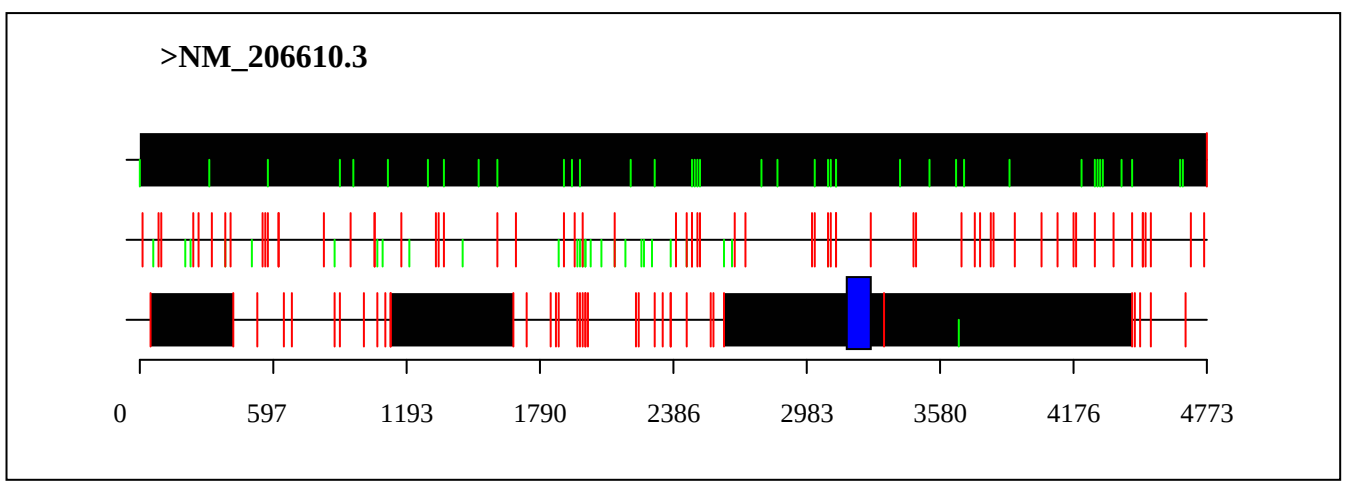


Supplementary Figure 1. **Multiple alignment of fruit fly, mosquito and bee genomic sequences corresponding to a fragment of the Xbp1 CDS.** Blue font indicates the sequence of the presumably alternatively spliced exon. Nucleotides at which a switch in evolutionary signatures has been reported by Lin et al is highlighted with red. The overlapping region of the two ORFs is between the stop codons shown in red. If the exon is retained, translation would terminate at the stop codons at the 3' end of the sequences shown in the alignment. If the exon is spliced, translation of the downstream sequence should occur in a different frame and terminate further downstream.

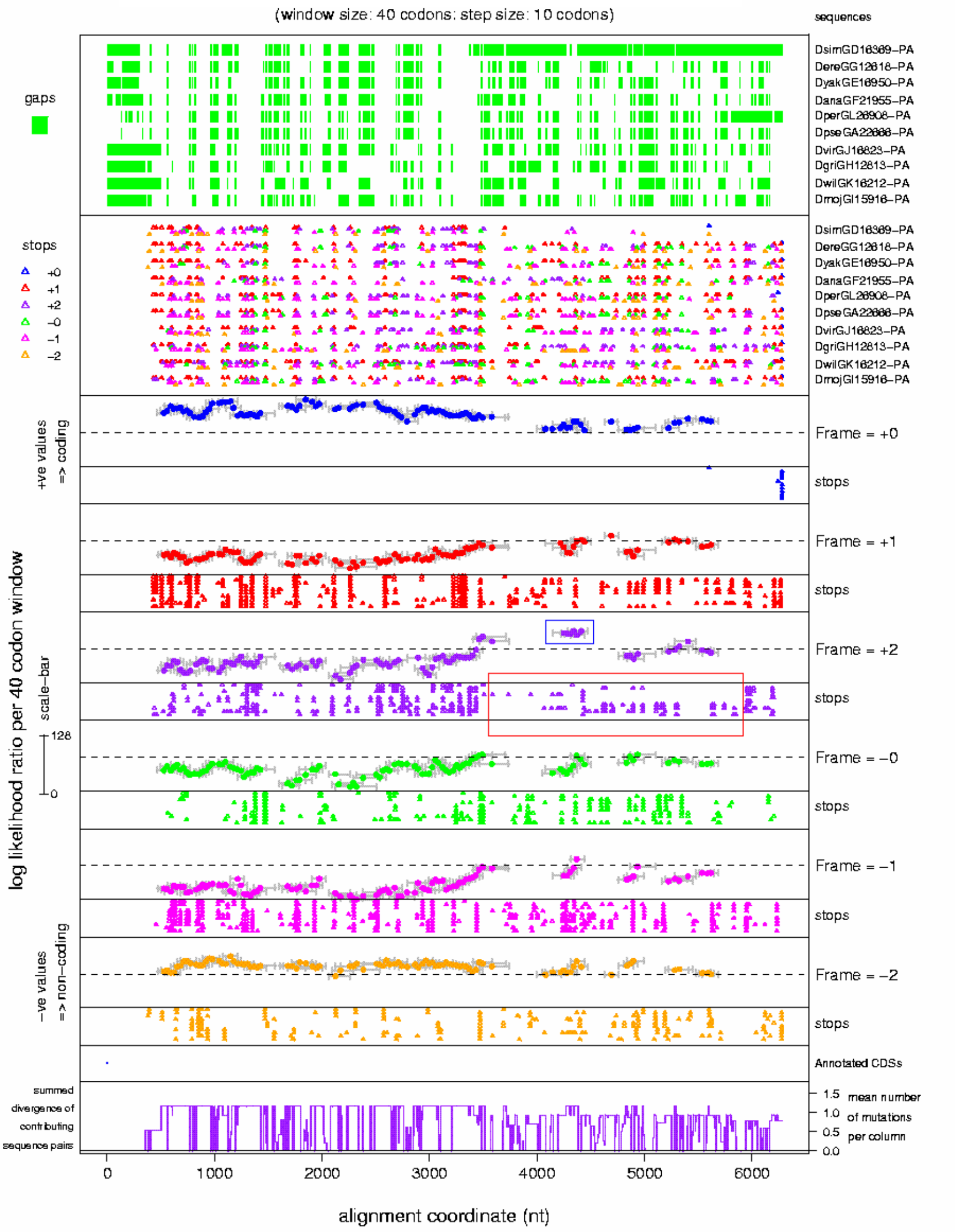
```
B D   D. melanogaster  ACCGCACTTTTAAGCGCCGCCAGGCGAAGAACTACGTGGAAGAGCAGCAGCATCT------GCAG-----
B D       D. simulans  ACCGCACTTTTAAGCGCCGCCAGGCGAAGAACTACGTGGAAGAGCAGCAGCATCT------GCAG-----
B D      D. sechellia  ACCGCACTTTTAAGCGCCGCCAGGCGAAGAACTACGTGGAAGAGCAGCAGCATCT------GCAG-----
B D         D. yakuba  ACCGCACTTTTAAGCGCCGCCAGGCGAAGAATTACGTGGAAGAGCAGCAGCATCT------GCAG-----
            D. erecta  ACCGCACTTTTAAGCGCCGCCAGGCGAAGAACTACGTGGAAGAGCAGCAGCATCT------GCAG-----
          D. ananassae ATCGCACCTTTAAGCGCCGTCAGGCCAAGAACTACGTTGAAGAGCAGCAGCACTT------GCA------
     D. pseudoobscura  ATAGAACATTTAAGCGTCGCCAGGCCAAGAACTACGTGGAGGAGCAGCAGCACCA------TCAGCAGCA
B D     D. persimilis  ATAGAACTTTTAAGCGTCGCCAGGCCAAGAACTACGTGGAGGAGCAGCAGCACCA------TCAGCAGCA
        D. willistoni  ATCGCACATTCAAGCGCCGTCAGGCTAAAAACTACGTAGAGG------AACATCC------ACAG-----
           D. virilis  ATCGCACATTCAAGCGGCGCCAGGCCAAGAACTACATCGAGGAGCAGCAGCAACACCTGGCGCAGCCGCA
        D. mojavensis  ATCGCACGTTCAAGCGCAGGCAGGCGAAAAACTACGTGGAGGATCAACTGCAACA------GCAGCAGCA
         D. grimshawi  ATCGCACATTTAAACGCAGGCAGGCCAAGAACTACGTGGAGGAGCAGCAGCATCATCA---TCAGCAGCA
         T. castaneum  ATAAAGTGTACAAACGCAGACAAGCGAAAATTTTGGTAGAAGAATCCTCACAATA------GCAA-----

       D. melanogaster  -------GCGCGAGCCGCGAATAACACC------------AACTAAgcaa----------aATGCCCGCC
           D. simulans  -------GCGAGAGCCGCGAAAAACACT------------GACTAAgcaa----------aATGCCCGCC
         D. sechellia   -------GCGCGAGCCGCGAAAAACACT------------GTCTAAgcaa----------aATGCCCGCC
            D. yakuba   -------GCGCAAGCCACGCACAGCACT------------CAGTGAacaa----------aATGCCCGCC
            D. erecta   -------GCGCTAGCCGCGAGAAACACT------------GACTGAacaa----------aATGCCCGCC
         D. ananassae   --------CGAAAGACAGGACCAGAGCTCGAAGTAGAGAGAAATTAgcaa---------tcATGCCCGCC
     D. pseudoobscura   -------GCAACAGCAAC-AGCAGCAGT------------AGTCGTgtaactgtgggaatcATGCCCGCT
        D. persimilis   -------GCAACAGCAGC-ACCAGCAGT------------AGTCGTataactgtgggaatcATGCCCGCT
        D. willistoni   ------------AGTACC--ACAGAATA------------AAATCAtcat----------cATGCCCGCT
           D. virilis   ACCCACAGCCACAGCCACAG---CCACA------------GCCTAGg-------------cATGCCCGCG
        D. mojavensis   ----TCAACTGAAGGAAT--------CG------------ATTTAAt-------------cATGCCCGCC
         D. grimshawi   ----GTTGCCAGAGCAACAG--AACACA------------GTTTAAg-------------cATGCCCGCC
         T. castaneum   ------------------------------------------------------------aATGCCGGCA
```

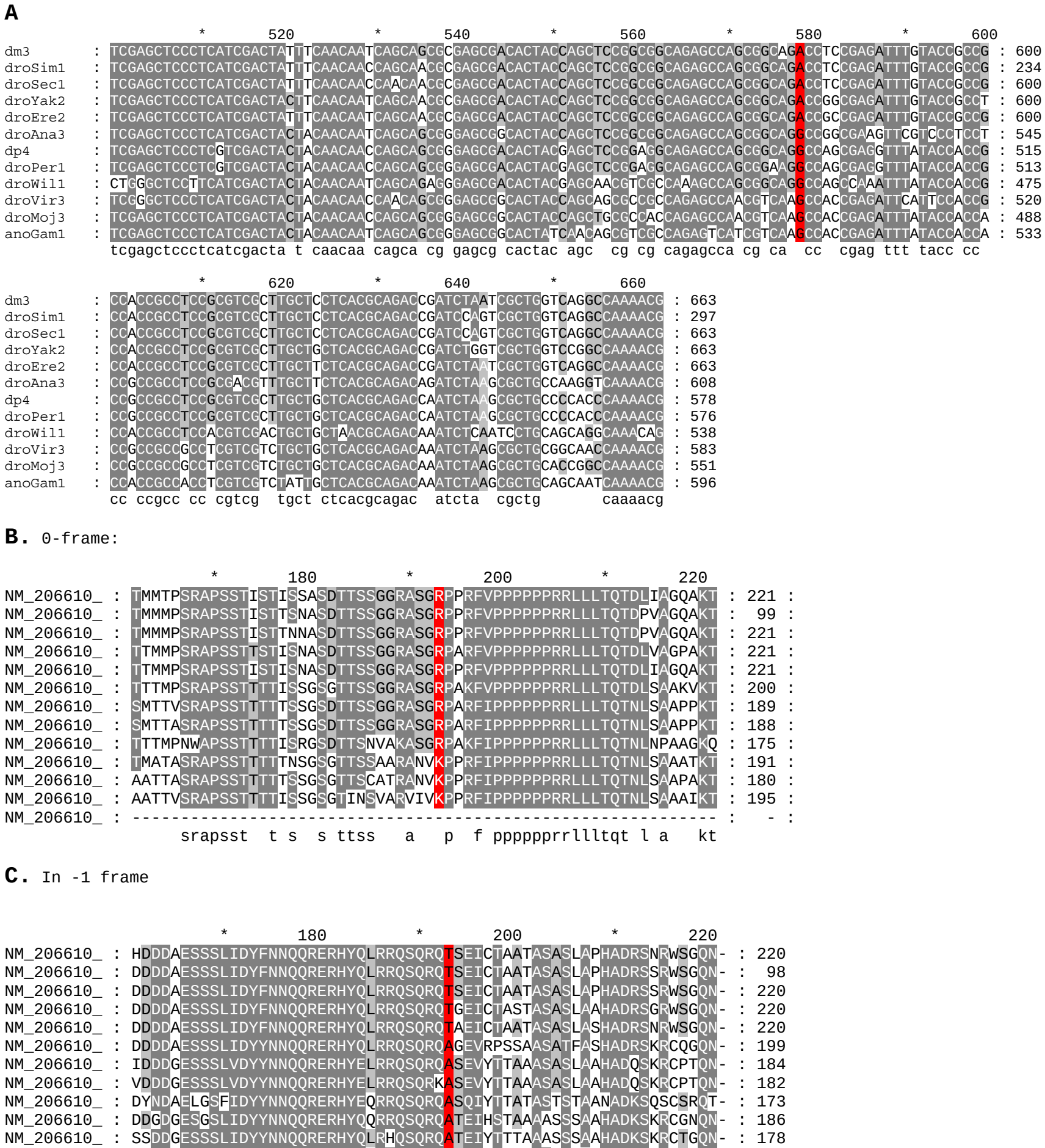
Supplementary Figure 2. **Multiple alignment of the genomic sequences encoding CG32736 in the vicinity of the predicted switch between evolutionary signatures.** The alignment of this region was obtained from the UCSC Genome Browser Multiz Alignment tracks. Red font indicates stop codons of the first ORF, blue font indicates start codons of the second ORF. Red highlighting corresponds to the position of the predicted frame switch.



Supplementary Figure 3. **Sequence cruncher 1.9 plot of ORFs in CG14047. ORFs (>300nts) are shown by black boxes.** Red dashes indicate stop codons, green – ATG codons. A region at which purifying selection is acting in the alternative -1 frame is shown with a blue box. The switch between frames predicted by Lin et al corresponds to the end of the region marked by the blue box.



Supplementary Figure 4. **MLOGD plot of multiple alignment of sequences encoding CG14047 in different flies.** Blue box is used to indicate the region at which purifying selection is acting on protein sequence in the alternative -1 frame (marked with blue box in the Supplementary Figure 3). Red box indicates a region corresponding to the location of long alternative ORF in *D. melanogaster* (see Supplementary Figure 3).



Supplementary Figure 5. **Multiple alignments of nucleotide and encoded protein sequences in the vicinity of the predicted evolutionary switch for CG14047. A.** Nucleotide alignment. **B.** Alignment of protein sequences corresponding to the zero frame. **C.** Alignment of protein sequences corresponding to the -1 frame. Red indicates stop codons that interrupt the long alternative ORF in *D. melanogaster* (see Supplementary Figure 3). Symbols highlighted in red correspond to the location of the predicted switch between frames. It is clear that the protein sequence upstream of the switch is more conserved if derived from the -1 frame, while the 0-frame encoded sequence is more conserved downstream of the switch.