Evolution of the trans-splicing Drosophila locus *mod(mdg4)* in several species of Diptera and Lepidoptera

Key words: molecular evolution, alternative splicing, modifier of mdg4, BTB domain, FLYWCH domain, composite gene

Veiko Krauss[a,*] and Rainer Dorn[b]

[a]Department of Genetics, University of Leipzig, Leipzig, Germany and [b]Institute of Genetics, Biologicum; Martin Luther University Halle, Halle/S., Germany

Abbreviations: Aga, *Anopheles gambiae*; Bmo, *Bombyx mori*; cDNA, complementary DNA; Cel, *Caenorhabditis elegans*; Dme, *Drosophila melanogaster*; Dps, *Drosophila pseudoobscura*; Dvi, *Drosophila virilis*; EST, expressed sequence tag; Hsa, *Homo sapiens*; *lola, longitudinals lacking;* ML, maximum likelyhood; *mod(mdg4), modifier of mdg4);* NJ, neighbour joining; ORF, open reading frame; RT-PCR, reverse transcriptase-coupled polymerase chain reaction;

* Corresponding author: Veiko Krauss, Department of Genetics, University of Leipzig, Johannisallee 21-23, 04103 Leipzig, Germany; Phone: (+049)341-9736751; Fax: (+049)341-9736897; E-mail address: krauss@rz.uni-leipzig.de

Abstract

The *modifier of mdg4* (*mod[mdg4]*) locus of *Drosophila melanogaster* encodes chromatin proteins which are involved in position effect variegation, establishment of chromatin boundaries, nerve pathfinding, meiotic chromosome pairing and apoptosis. It was recently shown that mRNA trans-splicing is involved in the generation of at least 26 different *mod(mdg4)* transcripts. Here, we show that a similar complex *mod(mdg4)* locus exists in *Drosophila pseudoobscura*, *Drosophila virilis*, *Anopheles gambiae* and *Bombyx mori*. As in *Drosophila melanogaster*, most isoforms of these species contain a strongly conserved BTB/POZ domain (hereafter referred to as BTB domain) within the common N-terminal part and a Cys_2His_2 motif containing FLYWCH domain within the isoform-specific C-terminal parts. By sequence comparison, we identified six novel isoforms in *Drosophila melanogaster* and show that altogether 31 isoforms are perfectly conserved by sequence and position in the *mod(mdg4)* locus of the Drosophila species analyzed. We found significant differences in evolutionary speed of synonymous/nonsynonymous divergence between the various isoform specific exons. These results were extended by tree reconstruction analysis based on the evolved FLYWCH domains of predicted Mod(mdg4) proteins in Drosophila and Anopheles. Comparative analysis of *mod(mdg4)* gene structure in species of dipterans implicates that several internal inversions occurred making the mRNA trans-splicing mechanism indispensable for *mod(mdg4)* expression. Finally, we propose a model for the evolution of trans-splicing implementing effective regulation of many alternative gene products in a composite gene structure.

1. Introduction

Alternative mRNA splicing plays a key role in expanding the proteome of multicellular organisms (Graveley 2001; Maniatis and Tasic 2002). Recently, Brett et al. (2002) estimated that about 40% of human and mouse ESTs in the databases are products of alternative splicing. According to the same authors, the occurrence of differential splicing is not significant lower in other metazoans including *Drosophila melanogaster*. Recent findings uncovered a number of genes in Drosophila and human with an extraordinary diversity of alternative splicing (for review see Maniatis and Tasic 2002). However, our knowledge on the functional significance and the underlying regulatory mechanisms of differential splicing is very limited.

mod(mdg4) in *Drosophila melanogaster* (also known as *E[var]3-93D*) encodes more than two dozens of alternatively spliced isoforms. All of them contain a common N-terminal region of 402 amino acid residues but variable C-terminal ends (Büchner et al. 2000; Dorn et al. 2001). Interestingly, *mod(mdg4)* mutations have been independently isolated for their effects on position effect variegation, the properties of insulator sequences, correct pathfinding of growing nerve cells, meiotic pairing of chromosomes and apoptosis (for review see Dorn and Krauss 2003). Except of two mutant alleles disrupting the isoform-specific exon of *mod(mdg4)-67.2*, all mutations affect the common 5' region of the gene. Whereas the two *mod(mdg4)-67.2*-specific alleles are homozygous viable, nearly all other mutations are homozygous lethal. There is additional but limited evidence for a functional differentiation between the Mod(mdg4) isoforms (for review see Dorn and Krauss 2003). However, systematic mutational dissection will be a laborious challenge because of structural complexity and functional diversity of *mod(mdg4)*.

The interest on *mod(mdg4)* was boosted by the finding of mRNA trans-splicing for the first time in arthropods (Labrador et al. 2001; Dorn et al. 2001; Pirrotta 2002). The unusual feature that alternatively spliced exons are encoded by both DNA strands makes trans-splicing indispensable for expression of at least seven of the *mod(mdg4)* mRNA variants. Therefore, *mod(mdg4)* represents a new type of gene structure, where independent transcripts encoding different evolutionary conserved protein domains are fused by trans-

splicing. Recently, Labrador and Corces (2003) compared the genomic structure of *mod(mdg4)* from *Drosophila melanogaster* and *Anopheles gambiae* taking into consideration a partial *Drosophila pseudoobscura mod(mdg4)* sequence. They demonstrate a significant conservation of several Mod(mdg4) isoforms. In contrast to an identical gene structure in *D. pseudoobscura*, they found an extensive exon reshuffling in *A. gambiae*.

We confirmed these results and extended this analysis by including *mod(mdg4)* sequences from *Drosophila virilis* and *Bombyx mori*. Additionally, we performed a detailed computational analysis of the two conserved Mod(mdg4) domains BTB and FLYWCH. A comprehensive comparison of all Mod(mdg4) isoforms points to significant differences in isoform multiplication and evolutionary diversification between species.

2. Materials and Methods

2.1. Library screen and sequencing

D. virilis genomic clones were isolated by two successive screenings of a genomic library (Lanio et al. 1994) according to standard protocols. In the first screen the cDNA clones *mod(mdg4)-67.2* and *mod(mdg4)-58.8* were used as probes. Washes were done twice with 2x sodium citrate (SSC), 0.1% sodium dodecyl sulphate (SDS) for 10 min. In the second screen overlapping recombinant clones have been isolated using a probe deduced from the recombinant clone LDv2-1 obtained in the first screen. Sequencing was performed on a ABI377 (ABI) using BigDye Sequencing Chemistry (ABI). The genomic region covered by the two recombinant λ clones Dv2-1 and Dv3-2 was sequenced on one strand by subcloning and primer walking. Exonic sequences were independently sequenced from both ends as cDNA clones obtained by RT-PCR.

2.2. Sequence sampling and annotation of *mod(mdg4)* gene sequences

Based on 26 known *mod(mdg4)* cDNA sequences (Büchner et al. 2000; Dorn et al. 2001)

we used BLASTn (Altschul et al. 1997) to retrieve the corresponding *Drosophila melanogaster* (Dme) *mod(mdg4)* genomic sequence (AE003734, version 2, 02/14/2003). We searched for additional ORFs between and downstream of known exons using BLAST and identified six novel putative exons coding for a FLYWCH motif-containing polypeptide. We deviate forward primers from common exon 4 of Dme *mod(mdg4)* and backward primers immediately downstream of the ORF ends for RT-PCR to show that all these ORFs represent specific exons of so far undetected *mod(mdg4)* splice variants. The used forward primers were mod-common (5'-GCAACAGTCCCAGAACTACAG-3') and mod-common2 (5´-CGAAATCTGAACCTGACT-3`), and the used backward primers were mod-hpeb1 (5'-AAGATTTAGATTAGAGGAAAGTATC-3'), mod-h62.3-2 (5'-GTAACGAGTAATCAGTCGG-3'), mod-h52.0-2 (5'-TAAACGGAAACAGCATG-3'), mod-h54.2-2 (5'-TTATTTATGTTAAGGTAGAGC-3'), mod-h57.4-2 (5'-GCTTATGGACTAGGTGC-3') and mod-CG15501 (5'-GGGTACATCGAGTTGAC-3'). The RT-PCR fragments were sequenced and, named according to the length of the predicted Mod(mdg4) isoform in kDa, submitted to Genbank/EMBL/DDBJ databases (accession nos. AJ586731, AJ586732, AJ586733, AJ586734, AJ586735, AJ586736).

mod(mdg4)-orthologous DNA sequences from other species were sampled from databases using BLAST. *Drosophila pseudoobscura* (Dps) single trace sequences were screened at the TraceSite of NCBI (http://www.ncbi.nih.gov/blast/tracemb.html), manually assembled and annotated (accession no. BN000431). For *Anopheles gambiae* (Aga), we identified the orthologous gene inside a 2,1Mb contig of the Anopheles genome project (Holt et al. 2002) by BLAST search. In *Bombyx mori* (Bmo), we identified 15 5´primed ESTs coding for a Mod(mdg4)-like BTB domain and independently four 3´primed ESTs coding for different FLYWCH motifs using tBLASTn. By sequence assembling we constructed contigs of four different Bmo *mod(mdg4)* full-length splice variants from these EST sequences (accession no. BN000406).

The annotation of genomic sequences of *Drosophila pseudoobscura* was done manually in MacVector 7.1.1 (Accelrys) using 32 EST sequences and sequence comparison with *Drosophila melanogaster*. The annotation of the *D. virilis* genomic sequence was done by

cDNA alignment (Dorn et al., submitted) and supplemented by sequence comparison with *Drosophila melanogaster* (accession no. AJ586737). For the initial annotation of the Anopheles *mod(mdg4)* gene EST sequences from this species were used. The Pustell Protein/DNA matrix of MacVector was used for comparison of mosquito genomic sequences with *D. melanogaster* protein sequences to annotate additional specific exons with a significant similarity to Mod(mdg4) isoforms. EST-predicted gene structures which were not supported by splice site consensus sequences were excluded from the analysis (accession no. BN000407).

2.3. Multiple Alignment, secondary structure prediction and tree reconstruction

A multiple alignment of the amino acid sequences of all FLYWCH domains was constructed using MacVector 7.11 (Accelrys). In this alignment ten non-Mod(mdg4) proteins identified through PSI-BLAST (Altschul et al. 1997) using the FLYWCH motif of the Dme Mod(mdg4) variant 55.6 as initial query (profile-inclusion threshold E=0.05) were included. PROFsec secondary structure prediction analysis (Rost 2000) in default mode were independently applied for 22 Dme Mod(mdg4) FLYWCH sequences (all previously described in Büchner et al. 2000 or Dorn et al. 2001), including the Bmo Mod(mdg4) FLYWCH sequence heS00531 and 17 FLYWCH domains from non-Mod(mdg4)-proteins. These analyses were done in a context including more than 35 residues N-terminal and C-terminal (or including the C–terminus) to prevent boundary effects.

The programs MrBayes, version 3.0 (Huelsenbeck and Ronquist 2001), PAUP, version 4.0b10 (Swofford 2002) and TREE-PUZZLE, version 5.0 (Schmidt et al. 2000), were used for phylogenetic analyses. Tree constructions were performed through the bayesian inference method by MrBayes using the JTT substitution model, 500000 replicates (every tenth was saved) and a burnin of 30000 resulting in 20000 trees. We computed the 50% majority consensus tree from this analysis using PAUP4.0b10. Additionally, we used TREE-PUZZLE 5.0 for a maximum likelyhood analysis (quartet puzzling) with 10,000 puzzling steps, the WAG substitution model and assuming rate heterogeneity with eight gamma rate

categories.

2.4. Analysis of nonsynonymous/synonymous sites

Pairwise alignments of coding nucleotide sequences of all specific exons of Mod(mdg4) isoforms of *Drosophila melanogaster, D. pseudoobscura* and *D. virilis* were constructed using MacVector 7.11 (Accelrys). The homology of the aligned nucleotide positions was controlled by simultaneous alignments of the corresponding amino acid sequences. The 5´and 3´ends of the alignments were cut off using the most N–terminal or C-terminal identical amino acid position in between all three species sequences as anchor site. The cut includes these anchor codons itself to avoid bias. Subsequently, the alignments were manually processed to remove all gaps and to ensure homologous triplet positions according to the corresponding amino acid alignment. Pairwise K_A/K_S tests were carried out using the method of Yang and Nielsen (2000), which accounts for differences in nucleotide and codon frequencies as well as transition:transversion rate bias, implemented in PAML 3.12 (Yang 2000). Amino-acid and coding sequence alignments are available on request.

3. Results

3.1. The *mod(mdg4)* locus of *Drosophila melanogaster* is conserved in *D. pseudoobscura*, *D. virilis*, *Anopheles gambiae* and *Bombyx mori*

To study the evolution of the *mod(mdg4)* locus, we first defined the Mod(mdg4) protein in difference to other proteins by three characteristic structural properties. Accordingly, Mod(mdg4) proteins consist of (i) a strongly conserved BTB domain at the N-terminus, (ii) a middle part represented in all isoforms and (iii) a novel Cys_2His_2 motif in the isoform-specific C-terminal part of most isoforms called the FLYWCH domain (Büchner et al. 2000; Dorn and Krauss 2003).

We screened a genomic *Drosophila virilis* (Dvi) library using *D. melanogaster* (Dme) *mod(mdg4)* cDNA fragments. Dme *mod(mdg4)* cDNA sequences were used for *in silico* screening of databases to find similar sequences (see section 2.2) which might represent orthologous genes. After sequencing (*D. virilis*), assembling and annotation, we obtained (i) a genomic sequence of including the complete *Drosophila pseudoobscura* (Dps) *mod(mdg4)* locus from Drosophila Genome Project (http://www.hgsc.bcm.tmc.edu/projects/drosophila/), covering 32 EST sequences representing 9 alternative splice variants, (ii) a partial genomic gene sequence of Dvi *mod(mdg4)* sequenced by ourself, containing the common exons and the 19 most proximal alternative splice forms which were partially covered by twelve C-terminal different cDNAs (described in Dorn et al., submitted), (iii) a genomic sequence including the complete *mod(mdg4)* locus from the *Anopheles gambiae* (Aga) genome project (Holt et al. 2002), annotated with the help of 71 EST sequences representing 20 alternative splice variants, and (iv) 15 EST sequences from *Bombyx mori* (Bmo) representing four alternative splice variants. All these gene sequences encode at their 5´end a strongly conserved BTB domain which shows a much lower BLASTp Expect value to the Dme Mod(mdg4) domain compared to any non-Mod(mdg4) BTB domains. Pairwise sequence comparison with the Dme *mod(mdg4)* protein middle part revealed obvious similarity to the corresponding part of *D. pseudoobscura, D. virilis* and, locally restricted, *A. gambiae* proteins (Fig. 1). Additionally, the majority of the identified alternative exons in all species code for isoform-specific FLYWCH motifs. Therefore, all identified sequences represent *mod(mdg4)*-orthologous genes of the analyzed species.

3.2. The *mod(mdg4)* locus of Drosophila species codes for 31 conserved isoforms

We searched the Dme *mod(mdg4)* genomic regions using tBLASTn to identify undetected specific exons encoding a FLYWCH motif and found six novel putative isoforms. RT-PCR (see section 2.2) was applied to prove both expression and supposed splice sites of these isoforms. The novel variants were named according to their predicted molecular weight as

Mod(mdg4)–53.5, –55.8, –52.4, –55.2, –59.3 and –54.5 (Fig. 2). In *Drosophila pseudoobscura* and *D. virilis*, we found a nearly exact positional correspondence of common and specific *mod(mdg4)* exons to *Drosophila melanogaster*. Therefore, all isoforms were named according to the Dme Mod(mdg4) variants. The prefix h for "homolog" was added to emphasize that all orthologs have deviating molecular weights in other species.

The specific parts of 31 Drosophila Mod(mdg4) isoforms are significantly conserved in Drosophila, except Mod(mdg4)–46.3 (Büchner et al. 2000). Considering that this isoform was identified by a single cDNA clone, we suppose that this clone is the result of dysfunctional exon scipping involving the fifth exon of isoform 54.2 (compare Fig.2). Therefore, we omitted the originally described Mod(mdg4) isoform 46.3 from further analysis. If we additionally take in account that Dme Mod(mdg4) isofomes 59.0 and 1.8 (Labrador and Corces 2003) represent one and the same isoform, the total number of 31 isoforms is in agreement between both studies.

3.3. The BTB domain is highly conserved between Mod(mdg4) orthologues

The most N-terminal region of all Mod(mdg4) proteins consists of a BTB domain which is present in a large number of eukaryotic proteins. This domain was shown to mediate protein-protein interactions resulting in highly stable dimers (Bardwell and Treisman 1994; Ahmad et al. 1998; Li et al. 1999). We aligned the Mod(mdg4) BTB domains with other BTB domains and modeled the secondary structure for comparison with crystallization data of the BTB domain of PLZF (Ahmad et al. 1998; Li et al. 1999) to reveal implications for its function (Fig. 3).

Diagnostic residues were defined as identical amino acid positions present in all five known Mod(mdg4) sequences, but absent in the corresponding positions of the PFAM consensus and in Mod(mdg4)-like BTB domains of other proteins (Fig. 3). Interestingly, the conserved Mod(mdg4) motif HSALxD within the α-helix 4 in alignment positions 80-85, containing three diagnostic residues (HxAxxD), does not seem to have a counterpart in other BTB domains. According to the model of Ahmad et al. (1998), only α–helix 1 and β–sheets 1 and

5 are involved in dimerization. Thus, the HSALxD motif, situated in the external oriented helix 4, might be involved in interactions with other proteins. Chip, an enhancer facilitator protein, represents one putative interactor. This protein interacts with a Dme Mod(mdg4) protein truncated at residue 308 (Gause et al. 2001), which might be mediated by the BTB domain and/or by the region between amino acid positions 120 and 308. However, this second region is not conserved in Anopheles and Bombyx (Fig. 1), which argues for the BTB domain as main interaction site.

3.4. The FLYWCH motif is found in most Mod(mdg4) isoforms and in other proteins of bilaterian animals

The second conserved region of Mod(mdg4) proteins was first detected as a specific Cys_2His_2 pattern (Büchner et al. 2000; Dorn et al. 2001) and we recently named it FLYWCH domain according to strongly conserved residues within a peptide of 60 amino acid residues (Dorn and Krauss 2003). To investigate the occurrence of this module in other proteins, we performed iterative PSI-BLAST (Altschul et al. 1997; see section 2.3) for screening the Genbank nr Protein databank using occasional manual sequence exclusion and five iterations until convergence. In this analysis, we collected all 22 previously described Dme Mod(mdg4) isoforms containing FLYWCH domains and several predicted proteins with FLYWCH motifs from *Anopheles gambiae* encoded by specific *mod(mdg4)* exons. This result indicates the specificity of our screen. Additionally, ten non-Mod(mdg4)-protein sequences containing FLYWCH domains from several species were retrieved. All these proteins do not contain other identifiable domains, but several contain multiple FLYWCH copies. The maximum of five copies was found in the hypothetical human protein KIA 1552 (accession no. AB046772). All FLYWCH-containing proteins are exclusively found in Bilateria which argues for a relatively recent evolutionary origin of this domain.

All identified FLYWCH sequences were aligned using the ClustalW-Algorithm (Supplementary Fig. S1 at http://www.uni-leipzig.de/~genetics/S1.GIF). We used a secondary structure prediction program (Rost 2000) for 40 selected FLYWCH sequences and compiled

a 90% consensus (see section 2.3). All conserved sequence elements with the exception of the C-terminal HNH motif form exclusively β- sheets in the 23 tested Mod(mdg4) sequences (Supplementary Fig. S2 at http://www.uni-leipzig.de/~genetics/S2.GIF). Such a secondary structure is also predicted for most of 17 selected non-Mod(mdg4)-FLYWCH domains. This is remarkable because other Cys-His-rich conserved sequences typically adopt a ββα structure, the classic fold of Cys_2His_2 zinc finger domains (Laity et al. 2001). However, an α- Helix is often predicted for unconserved sequences immediately C-terminal to the FLYWCH domain, whereas the HNH conserved motif is generally involved in the loop between a N-terminal β- sheet and a C-terminal α- helix. Thus, it remains possible that the FLYWCH domain might adopt a fold similar to metal-chelating domains.

We further investigated the evolution of the FLYWCH domain using different tree reconstruction methods (see section 2.3). A tree reconstructed by Bayesian inference gives several hints about the evolutionary history of the domain (Supplementary Fig. S3 at http://www.uni-leipzig.de/~genetics/S3.GIF). Between Anopheles and Drosophila, we detect orthologous relationships between 19 singletons or groups of specific variants. We decided to use these relationships for the nomenclature of Anopheles Mod(mdg4) isoforms (see below). Because of the different number of isoforms, 31 in Drosophila and 41 in Anopheles, such a partial orthologous relationship was expected and is useful to evaluate the structural evolution of this complex locus.

In contrast to our results, Labrador and Corces (2003) identified 35 isoforms in Anopheles. Isoforms which are additionally identified in our study are Aga Mod(mdg)-h55.1a, -h52.2, -v35, -v36, -v39 and -v40. We confirmed 12 of 13 orthologous relationships supported by Labrador and Corces (2003), whereas the orthology between Dme Mod(mdg4)-54.7 and Aga Mod(mdg4)-v41 could not sustained by our analysis. In addition, we identified seven orthologous relationships not described by the former study (Fig. 4).

3.5. Structural relationship of specific splice variants suggest locus-internal rearrangements during Dipteran evolution

The exon-intron-structure of the four common exons of *mod(mdg4)* is strongly conserved between *Anopheles gambiae* and all Drosophila species analyzed (Fig. 1). In contrast, tree reconstruction analysis of specific exons (Supplementary Fig. S3 at **http://www.uni-leipzig.de/~genetics/S3.GIF**) implicate only a partial positional correspondence of the Drosophila and Anopheles genomic structure in the regions of alternatively spliced exons. In Figure 4, the positions of all *mod(mdg4)* specific exons of Drosophila and Anopheles are schematically shown. Because of two inversions of transcriptional orientation found in all Drosophila species analyzed, we suppose that the Anopheles gene structure is very likely more similar to the ancient one. Therefore, the Anopheles-type of arrangement of specific exons was used to infer a hypothetical *mod(mdg4)* complex structure of the last common ancestor of flies and mosquitos. Each of the proposed 19 conserved specific variants exist in one or more copies in both Drosophila and Anopheles (Fig. 4). The resulting relationships are partially supported by introns which interrupt the specific C-terminal ORFs of some isoforms. Altogether six such introns are found which are conserved in all analyzed Drosophila species. There is one exception, the *mod(mdg4)-54.5* intron, which was probably gained in the *D. melanogaster* lineage, because both *D. pseudoobscura* and *Anopheles gambiae* do not contain an intron in the corresponding isoform. The specific introns in *mod(mdg4)-67.2*, 59.1 and 54.2 are not found in Anopheles indicating that they were lost in Anopheles or gained in Drosophila. However, whereas only one specific intron position (in *mod(mdg4)-58.0*) was exactly conserved between Anopheles and Drosophila, two other introns in *mod(mdg4)-59.0* and 58.8 slide only by 2 and 14 nucleotide positions, respectively. The amino acid conservation is strong around the 58.0 intron position and relaxed in case of 59.0 and 58.8 (data not shown). Therefore, intron sliding might be counteracted by functional conservation of the affected amino acid positions.

On the other hand, the positions of those orthologous exons (*mod(mdg4)–58.0*, *–58.8* and *–59.0*) is different in Drosophila and Anopheles. Additionally, the specific exons of *mod(mdg4)–55.1*, *–55.6* and *–67.2* have different locations. The two inversions of transcriptional orientation found in the Drosophila species cannot account for this exon rearrangements. Thus, the question arises if the emerging incongruence between sequence

similarity and relative genomic location of *mod(mdg4)* isoform-specific exons might be a misinterpreted result of convergent evolution. Figure 4 presents two counter-arguments: (i) A block of six positionally conserved orthologous specific exons (51.4, 59.1, 56.3, 57.4, 58.6 and 53.4) is found in Drosophila and Anopheles. (ii) Species-specific multiplications of exons occurred mostly through local tandem duplication as found in Drosophila at 54.6/56.3 and 59.3/57.4 and in Anopheles at v5/v6, h55.7a/h55.7b/h55.7c, v22/v23, h67.2a/h67.2b/h67.2c and v39/v40. Disperse duplications are less frequent: 53.6/52.2, 53.1/59.1/54.2 (Drosophila), and h55.1a/h55.1b (Anopheles, see also Fig. 1*B*). Interestingly, the stronger evidence for a disperse multiplication of exons in Drosophila coincides with the partial inversions of transcriptional orientation in this genus. At least eight newly emerged variants through recent duplications in Anopheles in difference to only five new duplicated variants in Drosophila points to a role of exon multiplication in establishment of the different numbers of variants (41 versus 31) between both genera.

3.6. Mod(mdg4) isoforms show specific evolutionary rates in Drosophila

To measure the relative selective constraint of all 31 Mod(mdg4) isoforms of *D. melanogaster*, *D. pseudoobscura* and *D. virilis*, we determined the ratio of nonsynonymous to synonymous substitutions per site ($\omega = K_A/K_S$) for alignable coding sequences (Fig. 5). We found a high variation of the K_A/K_S ratios between different isoforms and different species which nevertheless corresponds to ratios already described (Bergman et al. 2002) for a sample of genes between *D. melanogaster* and *D. pseudoobscura* ($\omega = K_A/K_S = 0.146/2.313 = 0.0631$) and between *D. melanogaster* and *D. littoralis*, a close relative of *D. virilis* ($\omega = K_A/K_S = 0.170/2.166 = 0.0785$). In the common part of all isoforms, the strong conservation of the BTB domain-coding region of exons 2 and 3 and the significant relaxed conservation of exon 4 were supported by all three pairwise species comparisons (Fig. 5). Most specific exons which were not found in Anopheles show a higher K_A/K_S ratio than the average given by Bergman et al. (2002). This applies to the Drosophila-specific variants 53.5, 55.8, 52.4, 65.0 and 55.3, however, 54.7 and 52.0 are stronger conserved than the

average. Altogether, the degree of conservation found inside the genus Drosophila appears not to be predictive for the extent of conservation between Drosophila and mosquitos.

Most of the isoforms are conserved to a similar degree in all species. In contrast, the isoforms 55.2 and 55.3 show a much stronger constraint between *D. melanogaster* and *D. pseudoobscura* which corresponds to the phylogenetic sister relationship of these species, but might indicate a faster evolution of these isoforms in the *D. virilis* evolutionary lineage. Furthermore, the non-FLYWCH isoform 58.0 shows a stronger conservation between *D. virilis* and *D. pseudoobscura* and between *D. virilis* and *D. melanogaster* than between *D. melanogaster* and *D. pseudoobscura*, which might argue for a disruptive evolution of the Mod(mdg4)–58.0 isoform since the separation of the *D. melanogaster* and *D. pseudoobscura* lineages. In contrast to these predicted evolutionary shifts, the isoforms 52.4, 54.7, 59.0 and 52.2 show the strongest conservation between *D. pseudoobscura* and *D. virilis*, which argues for a relatively fast evolution in the *D. melanogaster* lineage.

Some caveats in the interpretation of these results are in order. Individual codon alignments used are between 49 and 200 codons (mean 91 codons) long. Accordingly, their nucleotide substitution rate may deviate substantially from the norm. Second, for regularly measured synonymous substitution rates per site of $K_S > 1$, potentially large inaccuracies in the estimates of nucleotide divergence are expected to result from multiple substitutions per site. Nevertheless, only seven out of 31 specific Mod(mdg4) isoforms of Drosophila show no significant evolutionary relationship with one of the 41 Anopheles specific variants (Fig. 4). From these seven isoforms, 55.3, 52.4 and 54.7 show also unequal evolutionary rates between Drosophila species (Fig. 5). Thus, whereas the majority of Mod(mdg4) variants reveal a strong structural conservation, species-specific functional changes might drive evolution of some isoforms and decrease the extent of conservation between orthologous isoforms of different species.

4. Discussion

In this study, we identified and compared *mod(mdg4)* loci of different insect taxa. This

complex locus produces more than two dozens of isoforms via trans-splicing in Drosophila (for review see Dorn and Krauss 2003) and was identified using the simultaneous occurrence of the BTB and the FLYWCH domain as Mod(mdg4)-specific criteria. Its conservation in Diptera and Lepidoptera indicates that the differentially spliced *mod(mdg4)* gene already existed before the divergence of these taxa 333 to 352 million years ago (Gaunt and Miles 2002).

The structure of the locus is nearly perfect conserved in the genus Drosophila. In contrast, the Anopheles *mod(mdg4)* locus probably encodes at least 41 splice variants as supported by sequence similarity to Drosophila specific exons and/or by ESTs and single cDNA reads. Phylogenetic analysis of the FLYWCH motifs from Drosophila, Anopheles and Bombyx isoforms suggests that lineage-specific duplications of ancient specific exons played an important role during the evolution of the locus. At the same time, some specific exons might be further evolved resulting in loss of any sequence similarity to other variants. Consecutively, in both Drosophila and Anopheles evolutionary lineages non-FLYWCH isoforms appear. However, also some of these show orthologous relationships (55.1 and h55.1a/b, 58.0 and h58.0). There is no direct evidence for *de novo* establishment of alternatively spliced exons as might be anticipated by recruitment of exons from other genes.

The common part of all isoforms in the analyzed species consists of two regions: N-terminal the BTB domain and C-terminal an only weakly conserved region of about 150 (Bombyx) up to about 320 amino acids (Anopheles). Functional equivalence of the BTB domains of GAGA and Mod(mdg4) proteins was shown in *Drosophila melanogaster* by successful replacement of the BTB domain of GAGA by the BTB domain of Mod(mdg4) (Read et al. 2000), suggesting that homodimerization is a key feature of both domains. Additionally, the BTB domain may be the main interaction site of the Chip enhancer facilitator protein (Gause et al. 2001). In contrast, the middle part of Mod(mdg4) proteins (coded by exon 4 in Dipterans) shows only a moderate conservation between the Drosophila species ($0.1147 < K_A/K_S < 0.1325$) which is lower compared to the majority of specific exons (Fig. 5). This is compatible with the conservation of an acidic amino acid subsequence inside this region between Drosophila and Anopheles (Fig. 1: between alignment positions 335 and

370) and may reflect coevolution with putative interacting proteins or a secondary role as connecting part of the conserved domains.

4.1. Evolution of the alternatively spliced exons

The specific parts of Mod(mdg4) isoforms are assumed to account for distinct functions of different isoforms, which is supported by several data. First, the isoform Dme 67.2 interacts with the Su(Hw) protein via the specific C-terminus (Gause et al. 2001; Ghosh et al. 2001). Second, the isoform Dme 56.3 (Doom) interacts with the inhibitor of apoptosis protein of Baculovirus OpIAP (Harvey et al. 1997). Third, the isoforms 67.2 and 58.0 localize predominantly to different sites on polytene chromosomes of *D. melanogaster* (Büchner et al. 2000). It is tempting to speculate that the multiply diversified FLYWCH domain, occuring in most isoforms, plays an important role in these specific interactions. This includes the possibility of a direct involvement in DNA binding. The N-terminal region of 158 amino acids of the Caenorhabditis PEB-1 protein contains a FLYWCH motif and shows DNA interaction (Thatcher et al. 2001).

Independent of the nature of the FLYWCH-mediated molecular interactions, they are supposed to be conserved among orthologous isoforms since the divergence of Culicimorpha (Anopheles) and Brachycera (Drosophila). Concomitantly, some novel isoforms have been evolved in both evolutionary lineages. From these novel variants, at least 54.7, 65.0 and 52.0 show a remarkable degree of sequence conservation between the Drosophila species (Fig. 5), but have no orthologous counterparts in Anopheles. A comparable situation was recently described for the BTB-transcription regulator lola (Goeke et al. 2003; Ohsako et al. 2003). This locus encodes 20 isoforms generated by alternative splicing, which are involved in axon guidance. All of them are conserved between *D. melanogaster* and *D. pseudoobscura*, but only eight have orthologs in Anopheles. Altogether these findings suggest that different isoforms of Mod(mdg4) fulfil taxa-specific roles in insect chromatin. Accordingly, functional conservation of specific isoforms should be limited to monophyletic groups of different ages. In particular, the isoforms Mod(mdg4)–64.2, –60.1, –62.3, –55.2,

–59.0, –51.4 and –58.6, which are strongly conserved between the Drosophila species (Fig. 5), and unambiguously related to a corresponding ortholog in Anopheles (Supplementary Fig. S3 at http://www.uni-leipzig.de/~genetics/S3.GIF), show slow evolution. These isoforms probably fulfil ancient functions in all Dipterans and might exist also in other insects. On the other side, novel isoforms of Mod(mdg4) obviously emerged through duplication of existing ones (Fig. 4). In case of functionalization of both copies, (i) one copy may acquire a novel, beneficial function and become readapted and preserved by natural selection, whereas the other copy keeps its original function (neofunctionalization), or (ii) both copies may become partially compromised by mutation accumulation to the point at which their total capacity is reduced to the level of the single-copy ancestral variant (subfunctionalization) (Force et al. 1999). Mod(mdg4) isoforms might have evolved along both pathways. For example, the Anopheles/Drosophila ortholog pairs 59.1/h59.1 and 59.3/h59.3 are sister sequences to the Drosophila-only paralogues 53.1 and 57.4, respectively. Corresponding orthologues might be lost in Anopheles, but it is also possible that a duplication in the Drosophila lineage enabled the ancient specific exons 53.1 and 57.4 to adopt novel function(s), which could mask the paralogous duplication. In contrast, subfunctionalization might be represented by the tandem duplications resulting in the specific exons 54.6/56.3 (Drosophila) and h55.7a/h55.7b (Anopheles). In stark contrast to this reshuffling of isoforms during Dipteran evolution, all isoforms remain significantly conserved in the genus Drosophila and no novel isoform emerges. This finding argues against a role of Mod(mdg4) proteins in establishment of species-specific properties, at least in the genus Drosophila.

4.2. A model for evolution of a composite gene

The unusual structure of the conserved Drosophila *mod(mdg4)* locus is represented by different directions of transcription (Dorn et al. 2001; Labrador et al. 2001). Therefore, *mod(mdg4)* requires a minimum of two promoters and a trans-splicing mechanism to express all isoforms. Interestingly, in Anopheles all *mod(mdg4)* exons are transcribed in the

same direction. We do not know if trans-splicing occurs in Anopheles. In contrast to Labrador and Corces (2003), we argue that trans-splicing should have evolved significantly before inversions of some exons can take place as evident in the Drosophila evolutionary lineage. We propose that the trans-splicing mechanism was positively selected by the advantage of a complex regulation of expression of numerous alternatively spliced exons, regardless of transcriptional orientations.

It is well established that by 5´prime end-specific differential cis-splicing several alternative exons can be expressed spatially and temporally restricted using distinct promoters. This was shown, for example, for two genes of the human tandem-duplicated paralogous gene cluster of protocadherin (Tasic et al. 2002). These promoters concomitantly express all constitutive exons of one gene copy. This mode of expression is not compatible to 3´prime end-specific differential splicing. However, Tasic et al. 2002 also showed trans-splicing between specific exons and constitutive exons of different gene copies. In this case trans-splicing occurred very infrequent, but it reveals an evolutionary pathway to evolve trans-splicing as an additional mechanism of gene regulation (Fig. 6).

The following steps could be involved in emerging of trans-splicing at *mod(mdg4)*. According to our model (Fig. 6), the ancestral locus consisted of a common 5´region and two or more alternatively spliced, specific 3´exons. Subsequently, tandem duplication of this gene leads to twofold expression of corresponding isoforms and allowed its evolutionary differentiation. The tandem duplicated loci showed initially high level of expression, which increases the probability of pre mRNA molecules to be in tight vicinity. Thus, the productive interaction of a functional 5´splice site of one pre mRNA molecule with a functional 3´splice site of another pre mRNA molecule leads to a trans-spliced mRNA. This trans-splicing becomes important, if the common exons of the downstream gene copy are lost by deletion. If the promotor of the downstream gene copy is preserved, pre mRNAs are produced which contain a functional 3´splice site but no functional 5´splice site. Hence, the downstream transcript can exclusively be used as an acceptor for pre mRNAs of the upstream gene copy. At this stage, the downstream gene copy loses its autonomy because of incapability to produce a functional mRNA. Subsequently, both the promoter and the downstream specific

exons may evolve to acquire new functions. At the same time, mRNA trans-splicing provides an additional regulatory mechanism for differential expression of the corresponding isoforms. Further multiplication of specific exons would be more feasible if an additional copy of the now specialized promoter of specific exons is included in subsequent duplications.

The *mod(mdg4)*-related locus *lola* encodes 20 isoforms which are generated by cis-splicing and trans-splicing (Horiuchi et al. 2003). A functional promoter driving the expression of a trans-splicing specific exon was identified. All exons are encoded by one DNA strand. Thus, the *lola* locus is in line with the suppositions made by our model. As a consequence of establishing trans-splicing, intragenic inversions as found in Drosophila *mod(mdg4)* do not abolish the expression of the locus.

Therefore, *mod(mdg4)* and *lola* might be first examples of a yet undiscovered type of gene structure, which we propose to name composite gene. We define two obligatory criteria for a composite gene: (i) Independent transcripts contribute a constant region represented in all mature mRNAs and variable regions contributing alternatively spliced specific exons. The specific exons can be expressed as one or more transcripts which contain the appropriate trans-splicing acceptor sites. (ii) The pre mRNAs are combined via trans-splicing. This mechanism allows the generation of pre mRNAs from both DNA strands. The current structure of *mod(mdg4)* and *lola*, with many different isoforms in Diptera, can thus be considered as a balancing act between the necessity of increased functional diversity of proteins and a limited gene number. Further detailed analysis of transcriptoms and proteoms will prove if composite genes are merely exceptional or common gene structures.

Acknowledgments

We would like to thank A. Herbst and P. Doebberthin for help in sequencing. The D. virilis genomic library was kindly provided by H. Kress (Berlin). We gratefully acknowledge the sequencing of the yet unpublished genome of *Drosophila pseudoobscura* by the Human Genome Sequencing Center at the Baylor College of Medicine. This work was partly supported by a grant from the Deutsche Forschungsgemeinschaft to R. D.

References

Ahmad, K.F., Engel, C.K., Prive, G.G., 1998. Crystal structure of the BTB domain from PLZF. Proc. Natl. Acad. Sci. USA 95, 12123–12128.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". Nucleic Acids Res. 25, 3389-3402.

Bardwell, V.J., Treisman, R., 1994. The POZ domain: a conserved protein–protein interaction motif. Genes Dev. 8, 1664–1677.

Bergman, C.M., Pfeiffer, B.D., Rincon-Limas, D.E. et al. (17 co-authors), 2002. Assessing the impact of comparative genomic sequence data on the functional annotation of the Drosophila genome. Genome Biol. 3, RESEARCH0086. Epub.

Brett, D., Pospisil, H., Valcarcel, J., Reich, J., Bork, P., 2002. Alternative splicing and genome complexity. Nat. Genet. 30, 29-30.

Büchner, K., Roth, P., Schotta, G., Krauss, V., Saumweber, H., Reuter, G., Dorn, R., 2000. Genetic and molecular complexity of the position effect variegation modifier *mod(mdg4)* in Drosophila. Genetics 155, 141–157.

Dorn, R., Reuter, G., Loewendorf, A., 2001. Transgene analysis proves mRNA trans-splicing at the complex *mod(mdg4)* locus in Drosophila. Proc. Natl. Acad. Sci. USA 98, 9724–9729.

Dorn, R., Krauss, V., 2003. The *modifier of mdg4* locus in Drosophila: functional complexity is resolved by trans splicing. Genetica 117, 165-177.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.L., Postlethwait, J., 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151, 1531-1545.

Gaunt M.W., Miles, M.A., 2002. An insect molecular clock dates the origin of the insects and

accords with palaeontological and biogeographic landmarks. Mol. Biol. Evol. 19, 748-761.

Gause, M., Morcillo, P., Dorsett, D., 2001. Insulation of enhancer–promoter communication by a gypsy transposon insert in the Drosophila *cut* gene: cooperation between suppressor of hairy-wing and modifier of mdg4 proteins. Mol. Cell. Biol. 21, 4807–4817.

Ghosh, D., Gerasimova, T.I., Corces, V.G., 2001. Interactions between the Su(Hw) and Mod(mdg4) proteins required for gypsy insulator function. EMBO J. 20, 2518–2527.

Goeke, S., Greene, E.A., Grant, P.K., Gates, M.A., Crowner, D., Aigaki, T., Giniger, E., 2003. Alternative splicing of *lola* generates 19 transcription factors controlling axon guidance in Drosophila. Nat. Neurosci. 6, 917-924.

Graveley, B.R., 2001. Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17, 100-107.

Harvey, A.J., Bidwai, A.P., Miller, L.K., 1997. Doom, a product of the Drosophila *mod(mdg4)* gene, induces apoptosis and binds to baculovirus inhibitor-of-apoptosis proteins. Mol. Cell. Biol. 17, 2835–2843.

Holt, R.A., Subramanian, G.M., Halpern, A. et al. (123 co-authors). 2002. The genome sequence of the malaria mosquito *Anopheles gambiae*. Science 298, 129-149.

Horiuchi, T., Giniger, E., Aigaki, T., 2003. Alternative trans-splicing of constant and variable exons of a Drosophila axon guidance gene, *lola*. Genes Dev. 17, 2496-2501.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754-755.

Labrador, M., Mongelard, F., Plata-Rengifo, P., Baxter, E.M., Corces, V.G., Gerasimova, T.I., 2001. Protein encoding by both DNA strands. Nature 409, 1000.

Labrador, M., Corces, V.G., 2003. Extensive exon reshuffling over evolutionary time coupled to trans-splicing in Drosophila. Genome Res. 13, 2220-2228.

Laity, J.H., Lee, B.M., Wright, P.E., 2001. Zinc finger proteins: new insights into structural and functional diversity. Curr. Opin. Struct. Biol. 11, 39-46.

Lanio, W., Swida, U., Kress, H., 1994. Molecular cloning of the *Drosophila virilis* larval glue protein gene *Lgp-3* and its comparative analysis with other Drosophila glue protein genes. Biochim. Biophys. Acta 1219, 576-580.

Li, X., Peng, H., Schultz, D.C., Lopez-Guisa, J.M., Rauscher III, F.J., and Marmorstein, R., 1999. Structure-function studies of the BTB/POZ transcriptional repression domain from the pro-myelocyticleukemia zinc finger oncoprotein. Cancer Res. 59, 5275–5282.

Maniatis, T., Tasic, B., 2002. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature 418, 236-243.

Melnick, A., Ahmad, K.F., Arai, S., Polinger, A., Ball, H., Borden, K.L., G. W. Carlile, G.W., Prive, G.G., Licht, J.D., 2000. In-depth mutational analysis of the promyelocytic leukemia zinc finger BTB/POZ domain reveals motifs and residues required for biological and transcriptional functions. Mol. Cell. Biol. 20, 6550-6567.

Melnick, A., Carlile, G.W., Ahmad, K.F., Kiang, C.L., Corcoran, C., Bardwell, V., Prive, G.G., Licht, J.D., 2002. Critical residues within the BTB domain of PLZF and Bcl-6 modulate interaction with corepressors. Mol. Cell. Biol. 22, 1804-1818.

Notredame, C., Higgins, D.G., Heringa, J., 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J. Mol. Biol. 302, 205-217.

Ohsako, T., Horiuchi, T., Matsuo, T., Komaya S., Aigaki, T., 2003. Drosophila *lola* encodes a family of BTB-transcription regulators with highly variable C-terminal domains containing zinc finger motifs. Gene 311, 59-69.

Pirrotta, V., 2002. Trans-splicing in Drosophila. Bioessays 24, 988-991.

Read, D., Butte, M.J., Dernburg, A.F., Frasch, M., Kornberg, T.B., 2000. Functional studies of

the BTB domain in the Drosophila GAGA and Mod(mdg4) proteins. Nucl. Acid. Res. 28, 3864–3870.

Rost, B. 2000. Better secondary structure prediction through more data, Columbia University. World Wide Web URL: http://cubic.bioc.columbia.edu/predictprotein.

Schmidt, H. A., Strimmer, K., Vingron, M., v. Haeseler, A., 2001. TREE-PUZZLE 5.0. Maximum likelihood analysis for nucleotide, amino acid, and two-state data. http://www.tree-puzzle.de/

Swofford, D.L., 2002. PAUP*. Phylogenetic analysis using Parsimony (*and other methods), Version 4.0b10. Sinauer, Sunderland, Massachusetts.

Tasic, B., Nabholz, C.E., Baldwin, K.K., Kim, Y., Rueckert, E.H., Ribich, S.A., Cramer, P., Wu, Q., Axel, R., Maniatis, T., 2002. Promoter choice determines splice site selection in protocadherin alpha and gamma pre-mRNA splicing. Mol. Cell. 10, 21-33.

Thatcher, J. D., Fernandez, A.P., Beaster-Jones, L., Haun, C., Okkema, P.G., 2001. The *Caenorhabditis elegans peb-1* gene encodes a novel DNA-binding protein involved in morphogenesis of the pharynx, vulva, and hindgut. Dev. Biol. 229, 480–493.

Yang, Z., 2000. Phylogenetic Analysis by Maximum Likelyhood (PAML) 3.0. London, University College London.

Yang, Z., Nielsen, R., 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol. Biol. Evol. 17, 32-43.

Fig. 1. Global Alignment of selected Mod(mdg4) isoforms from different insects. Multiple sequence alignments were computed by T-Coffee (Notredame et al. 2000). (A) Alignment of the isoform Mod(mdg4)–64.2 which represents the most proximal located specific exon in the Drosophila orthologues. In Bombyx, where no orthologous isoform was identified so far, the isoform heS00531 (accession no. **BN000406**) was chosen. Both conserved domains are marked by grey bars. Intron positions found in dipterans are indicated by triangels. (B) Alignment of the conserved, specific parts of Mod(mdg4)–55.1, representing a variant without FLYWCH domain. A threefold Cys_2 motif is marked. A consensus threshold of 51% was implemented.

Fig. 2. The genomic structure of *mod(mdg4)* in *Drosophila melanogaster* (modified from Dorn et al. 2001). The alternative splice site at the 3´boundary of exon 4 is used to generate all mature mRNAs indicated by the molecular weight of the deduced proteins. Exons shown below the rules are encoded by the same DNA strand, whereas those encoded by the antiparallel DNA strand are shown above the rules. Direction of transcription is indicated by arrows. Untranslated regions are shown as empty boxes and translated regions as filled boxes. All ORFs containing FLYWCH motifs are black filled, those without this motif are shown in grey. Note the overlap of several specific exons of different isoforms. The sequenced region of the orthologous *Drosophila virilis* locus is marked on the rule. mRNA variants supported by *D. virilis* cDNA sequences are underlined, and those supported by *D. pseudoobscura* ESTs are marked with a line above.

Fig. 3. Alignment of BTB domains of Mod(mdg4) proteins and selected BTB domains of other proteins and the deduced amino acid signature of Mod(mdg4) BTB domain. Residues are black boxed if the majority at the position is identical. Grey boxed residues mark a majority of similar amino acids. BTB domains of the Mod(mdg4) proteins (*D. pseudoobscura* and *D. virilis* were omitted because of the high degree of sequence identity to *D. melanogaster*) , together with their strict sequence consensus and a minimal consensus of their modelled BTB domain secondary structures (determined using PROF = divergent profile-based neural

network prediction trained and tested with PSI-BLAST; Rost 2000), in comparison with the sequence consensus of all BTB domains (Pfam database, Release 7.4), are shown. The substraction of this consensus sequence from the Mod(mdg4)-BTB consensus reveals several diagnostic residues of Mod(mdg4)-specific BTB domains. Sequence and secondary structure (from crystallization analysis; Ahmad et al. 1998) of the PLZF BTB domain is given together with three BTB domain sequences which are most similar to Mod(mdg4) BTB domains. Above the Drosophila Mod(mdg4) BTB sequences four in vivo or in vitro substituted amino acid positions, which interfere severely with the function of the protein, are shown in black triangels (Read et al. 2000; Kornberg unpublished in Melnick et al. 2000). Similarly, below the PLZF sequence amino acid substitutions in this protein which interfere severely (filled triangles) or not significantly (open triangles) with the function of PLZF are shown (Li et al. 1999; Melnick et al. 2000, 2002).

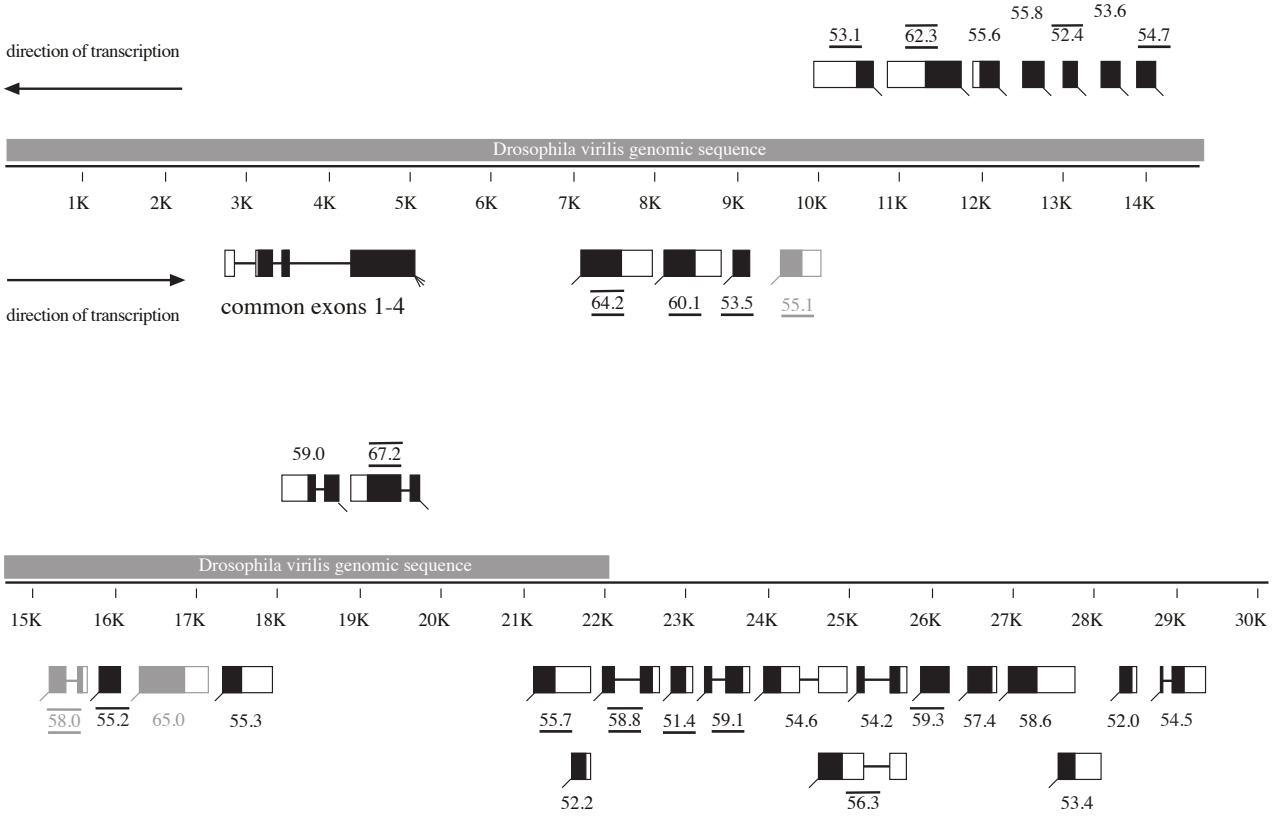
Fig. 4. Comparison of the Drosophila and Anopheles *mod(mdg4)* composite gene. 5´ common exons one to four are omitted. All specific exons are shown schematically, not to scale. The direction of transcription is indicated by a single slanted line representing the 3´splice site of the exons. Specific exons coding for a FLYWCH-containing polypeptide are shown as filled boxes, specific exons coding for a polypeptide without FLYWCH motif are shown as empty boxes. The nomenclature of the Anopheles exons corresponds to the supposed orthologous Drosophila exons or, alternatively, to their relative position in the locus. The *mod(mdg4)* locus of the last common ancestor of flies and mosquitos contained a minimum of 19 specific exons which are symbolized between the Drosophila and the Anopheles variants. All these variants exist in one or more copies in both Drosophila and Anopheles evolutionary lineages, which is illustrated by lines in both directions. The evolutionary relationships of isoforms were reconstructed by computing trees using NJ, ML and Bayesian Inference methods, respectively, and BLASTp searches and pairwise sequence comparison by MacVector 7.1 using the PAM250 matrix. Full lines indicate support by at least two different trees (FLYWCH-containing variants) or a reziproke maximal similarity score between the two putative peptide sequences and a corresponding BLASTp high

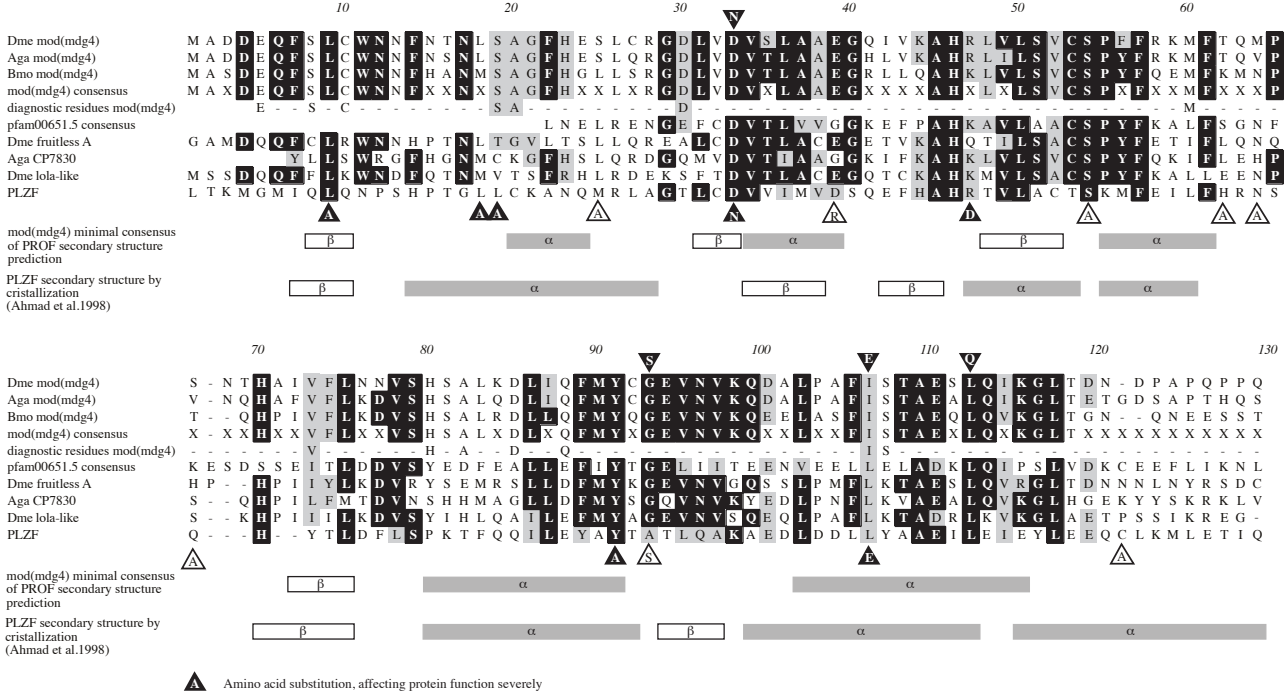
scoring pair with E<10 (variants without FLYWCH motifs). Broken lines indicate a putative relationship supported by one tree analysis and non-reziproke similarity scores or by one tree analysis and the relative location of exons. The relationships implicate recent duplications of isoforms and novel isoforms without clear provenance in both evolutionary lineages.

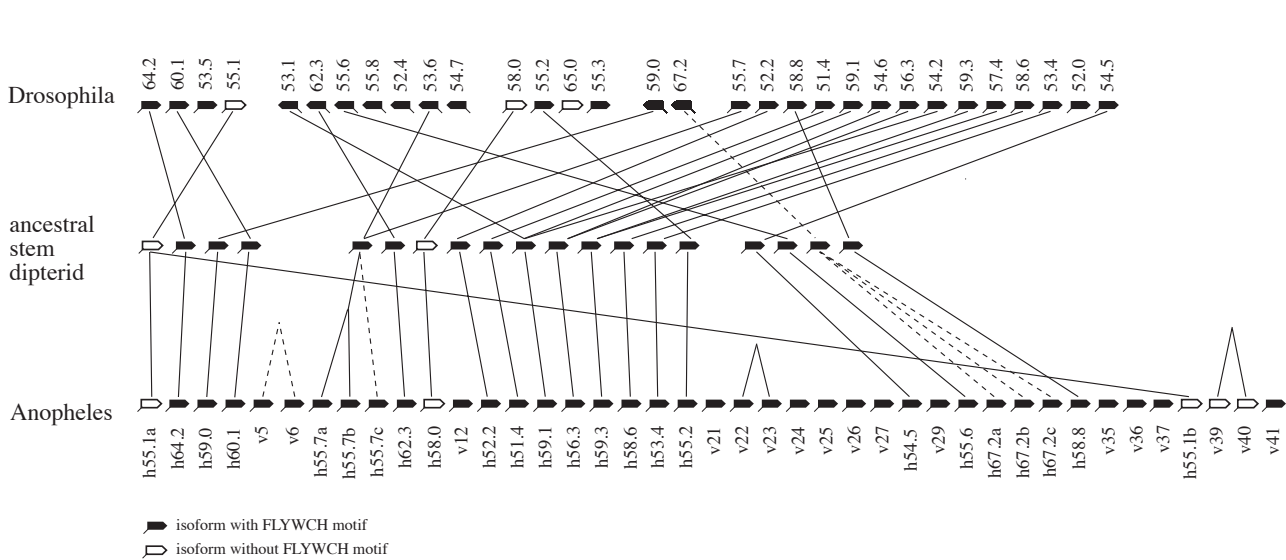
Fig. 5. Nonsynonymous/synonymous substitution ratios ($\omega = K_A/K_S$) of *mod(mdg4)* exons of the analyzed Drosophila species, computed according to Yang and Nielsen (2000). Each point represents the alignable partition of a single pair of orthologous *mod(mdg4)* coding nucleotide sequences from two different Drosophila species. Three values (two for 55.1 and one for 54.2) were omitted from the diagram because of to high synonymous substitution rates which make the analysis inapplicable. Two broken lines representing average ω values from analysis of 41 genes from *D. melanogaster* and *D. pseudoobscura* and 31 genes from *D. melanogaster* and *D. littoralis* (Bergman et al. 2002) are given for comparison.

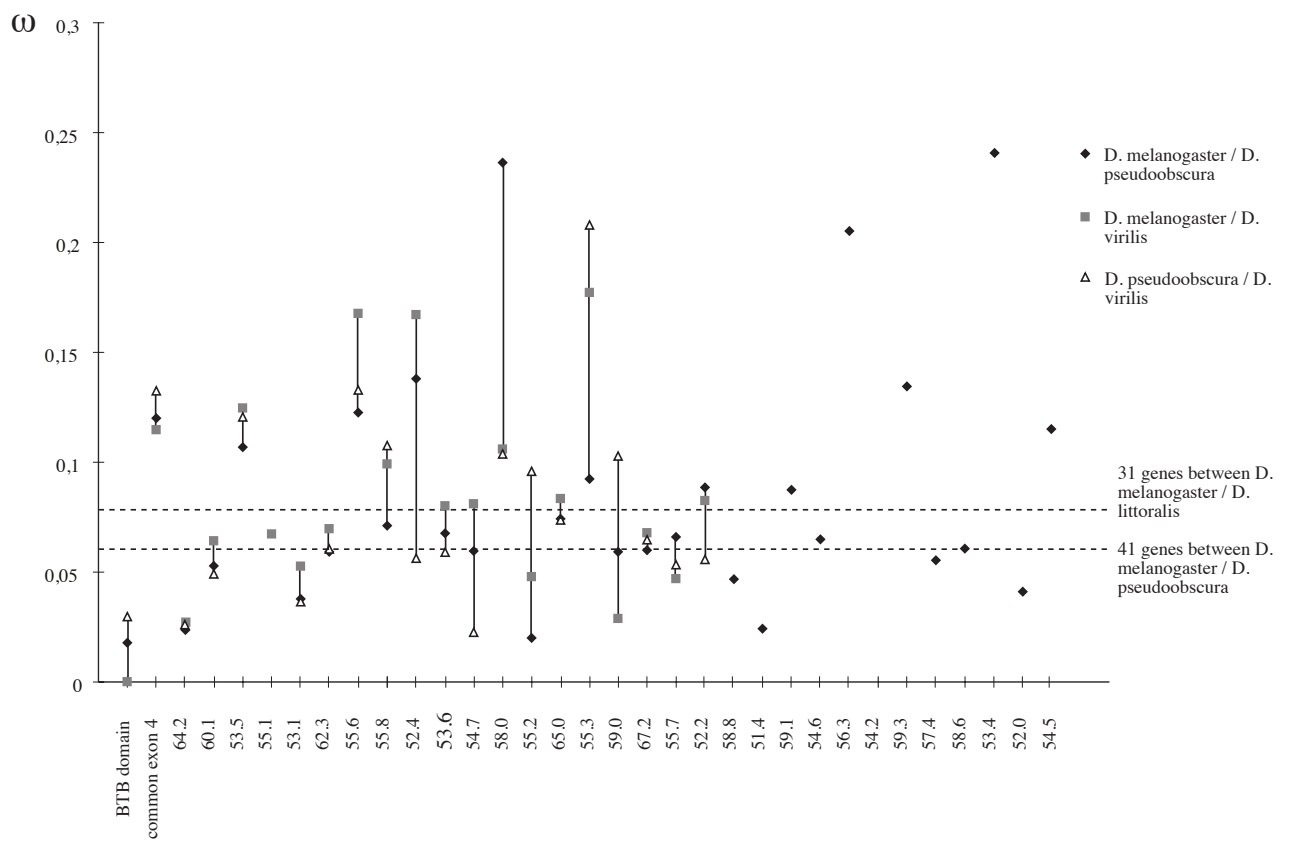
Fig. 6. A model for the generation of a composite gene (an alternative trans-splicing gene) in evolution. For simplicity, gene copies consist of one constitutive and two variable (specific) exons. See text for details.

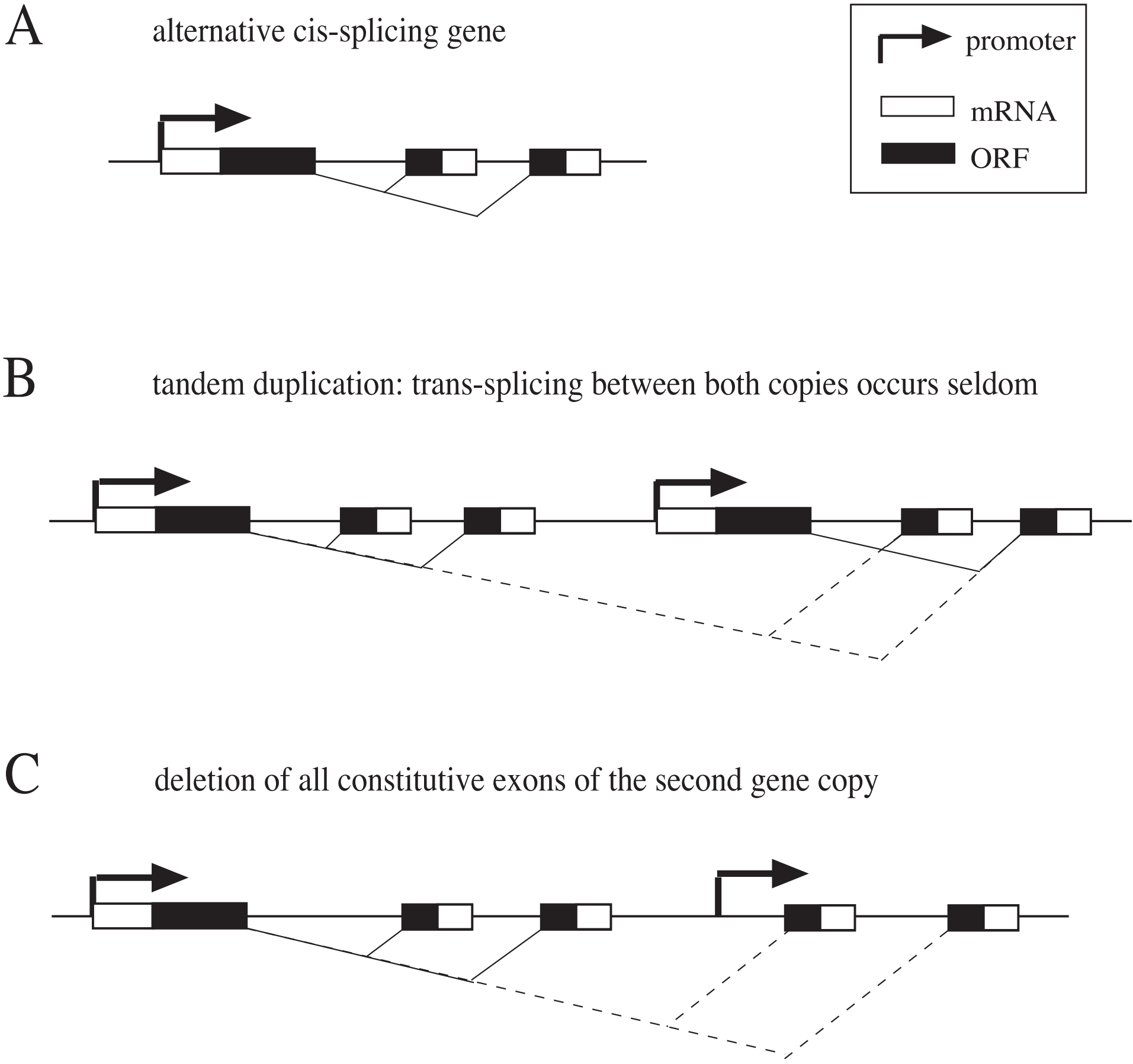

A
Intron 2
Intron 3
Intron 4, Anopheles
Intron 4, Drosophila
Dme 64.2
Dps h64.2
Dvi h64.2
Aga h64.2
Bmo heS00531
BTB/POZ domain
FLYWCH domain
B
Dme 55.1
Dvi 55.1
Dps 55.1
Aga 55.1b
Aga 55.1a
threefold CxxC motif


direction of transcription
53.1
62.3
55.6
55.8
52.4
53.6
54.7
Drosophila virilis genomic sequence
1K
2K
3K
4K
5K
6K
7K
8K
9K
10K
11K
12K
13K
14K
direction of transcription
common exons 1-4
64.2
60.1
53.5
55.1
59.0
67.2
Drosophila virilis genomic sequence
15K
16K
17K
18K
19K
20K
21K
22K
23K
24K
25K
26K
27K
28K
29K
30K
58.0
55.2
65.0
55.3
55.7
58.8
51.4
59.1
54.6
54.2
59.3
57.4
58.6
52.0
54.5
52.2
56.3
53.4


10 20 30 40 50 60
Dme mod(mdg4)
Aga mod(mdg4)
Bmo mod(mdg4)
mod(mdg4) consensus
diagnostic residues mod(mdg4)
pfam00651.5 consensus
Dme fruitless A
Aga CP7830
Dme lola-like
PLZF
mod(mdg4) minimal consensus of PROF secondary structure prediction
PLZF secondary structure by cristallization (Ahmad et al.1998)
70 80 90 100 110 120 130
Dme mod(mdg4)
Aga mod(mdg4)
Bmo mod(mdg4)
mod(mdg4) consensus
diagnostic residues mod(mdg4)
pfam00651.5 consensus
Dme fruitless A
Aga CP7830
Dme lola-like
PLZF
mod(mdg4) minimal consensus of PROF secondary structure prediction
PLZF secondary structure by cristallization (Ahmad et al.1998)
Amino acid substitution, affecting protein function severely
Amino acid substitution, affecting protein function not or weakly


Drosophila
ancestral stem dipterid
Anopheles
64.2
60.1
53.5
55.1
53.1
62.3
55.6
55.8
52.4
53.6
54.7
58.0
55.2
65.0
55.3
59.0
67.2
55.7
52.2
58.8
51.4
59.1
54.6
56.3
54.2
59.3
57.4
58.6
53.4
52.0
54.5
h55.1a
h64.2
h59.0
h60.1
v5
v6
h55.7a
h55.7b
h55.7c
h62.3
h58.0
v12
h52.2
h51.4
h59.1
h56.3
h59.3
h58.6
h53.4
h55.2
v21
v22
v23
v24
v25
v26
v27
h54.5
v29
h55.6
h67.2a
h67.2b
h67.2c
h58.8
v35
v36
v37
h55.1b
v39
v40
v41
isoform with FLYWCH motif
isoform without FLYWCH motif


ω
0,3
0,25
0,2
0,15
0,1
0,05
0
D. melanogaster / D. pseudoobscura
D. melanogaster / D. virilis
D. pseudoobscura / D. virilis
31 genes between D. melanogaster / D. littoralis
41 genes between D. melanogaster / D. pseudoobscura
BTB domain
common exon 4
64.2
60.1
53.5
55.1
53.1
62.3
55.6
55.8
52.4
53.6
54.7
58.0
55.2
65.0
55.3
59.0
67.2
55.7
52.2
58.8
51.4
59.1
54.6
56.3
54.2
59.3
57.4
58.6
53.4
52.0
54.5


A
alternative cis-splicing gene
promoter
mRNA
ORF
B
tandem duplication: trans-splicing between both copies occurs seldom
C
deletion of all constitutive exons of the second gene copy