Supplementary Information

Expansion of the mutually exclusive spliced exome in *Drosophila*

Klas Hatje and Martin Kollmar*

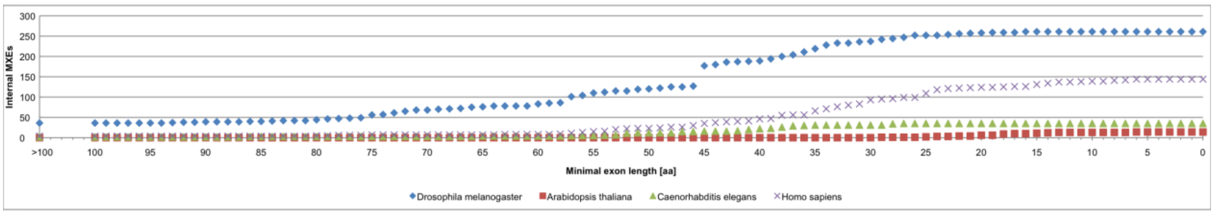
Max-Planck-Institute for Biophysical Chemistry

Department of NMR-based Structural Biology, Group Systems Biology of Motor Proteins,

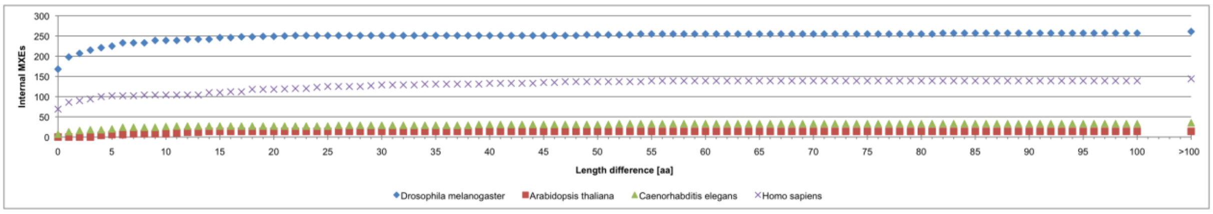
Am Fassberg 11, 37077 Göttingen, Germany

Correspondence and requests for materials should be addressed to M. K.
(email: mako@nmr.mpibpc.mpg.de)

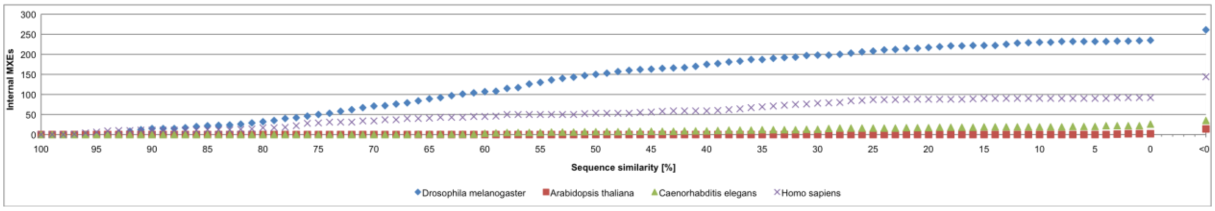
1 Supplementary Figures



Supplementary Figure S1. Number of annotated internal mutually exclusive spliced exons (MXEs) as function of the respective length of the MXE. The two noticeable jumps in the scatter plot of the *Dm* MXEs are due to the MXEs in the large clusters of the DSCAM gene. The shorter the exons are the more probable it becomes that their sequences are featureless and that false positive candidates will be predicted. Therefore, we introduced a parameter “minimal exon length”. Based on the analysis of all annotated MXEs we set this parameter to 15 residues.

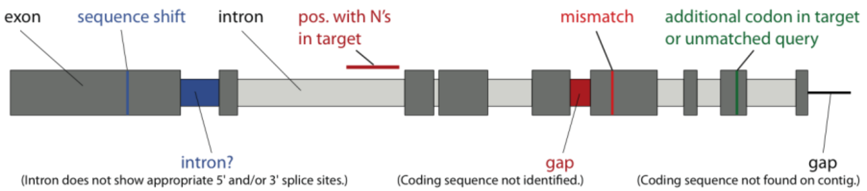


Supplementary Figure S2. Number of annotated internal mutually exclusive spliced exons (MXEs) as function of the minimal length difference to another MXE of the same cluster. To determine a suitable cut-off for the length difference in the search we analysed all internal clusters of annotated MXEs in the *Drosophila melanogaster* genome (*Dm*, Flybase release 5.36). To exclude that the determined characteristics are *Drosophila* specific we also analysed the annotated mutually exclusive exomes of *Homo sapiens* (*Hs*, NCBI release 37.3), *Caenorhabditis elegans* (*Ce*, WormBase release WS230), and *Arabidopsis thaliana* (*At*, TAIR release 167). These species have been chosen because of their widespread taxonomic distribution and their advanced and detailed annotations. For all species analysed the curves look very similar. 64%, 20%, 48% and 0% of the annotated MXEs of *Dm*, *Hs*, *Ce*, and *At*, respectively, have no length difference (86%, 71%, 57% and 43% have length difference of less than five residues). Therefore, a cut-off for the length difference of 20 residues should be appropriate to reconstruct almost all annotated cases and to not include too many mispredictions (95%, 82%, 77% and 100% have length difference of less than 20 residues).

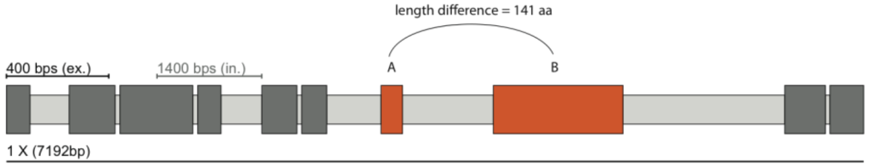


Supplementary Figure S3. Number of annotated internal mutually exclusive spliced exons (MXEs) as function of the sequence similarity to another MXE of the same cluster. In the case of similarity, two slightly different similarity scores can be calculated for a pair of MXEs dependent of which has been used as reference. Here, we included the respective higher scores. In this project, we were supposing that the MXEs of a cluster code for identical secondary structural elements of the protein like in the *Dm* muscle myosin heavy chain. If this condition holds true the MXEs should show a certain degree of sequence similarity. Analysis of the MXEs of *Dm* shows that 94.9% of the MXEs, which show any sequence similarity, have a sequence similarity of more than 15%. In *Hs* and *Ce*, 98% and 86% of the MXEs, which show any sequence similarity, have higher sequence similarities than 15%. Therefore, we decided to use 15% sequence similarity as cut-off for further predictions. However, a few cases of annotated MXEs do not show any sequence similarity and can not be reconstructed with our method (see difference of the two rightmost numbers).

Legend

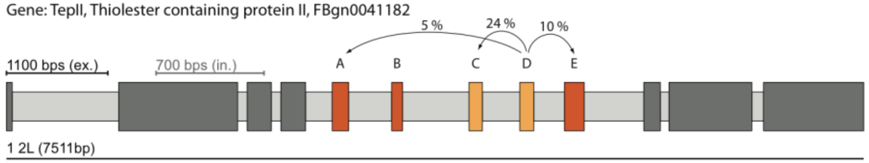

exon
sequence shift
intron
pos. with N's in target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

Gene: PhKgamma, Phosphorylase kinase γ, FBgn0011754



For clarity introns have been scaled down by a factor of 3.42

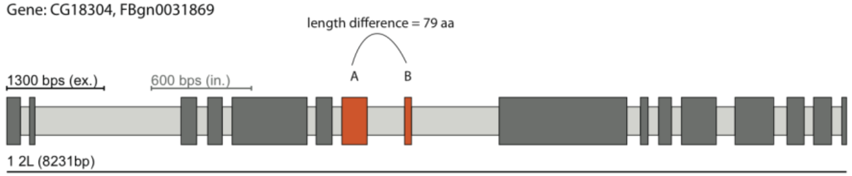
Gene: TepII, Thiolester containing protein II, FBgn0041182



For clarity introns have been scaled up by a factor of 1.67

```
               10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA  ---------------------EFPDYVEDDPEIYAFENNLDALPPMPAIANFPPDTGNTVQP-VEIRKNFADVWIWQSIGRS-
exonB  -------------------AKAIPESLDYQV-----EDSIS-YDEVDAISITSSTKIEL------VRTNFAEVWMWTTSDNGS
exonC  ---------------------EFPIAF-------SLAAPQAAIAGMPGTSSIASHPNQA-PQ---IRKEFPENWIFYNAEN--
exonD  ---------------------ERRIFIRP-----GIGFPRPLFNRVTVAGSLPPNVIPE-PQ---VRKEFPENWIFNIFEN--
exonE  GPLVMSYVFE-GSRHPWITRPRYRVGIRG-----DSGDRISFLSQSLNDRNLKEILLKQTPQRTTIRKEFPETWFFEN-----
```

Gene: CG18304, FBgn0031869

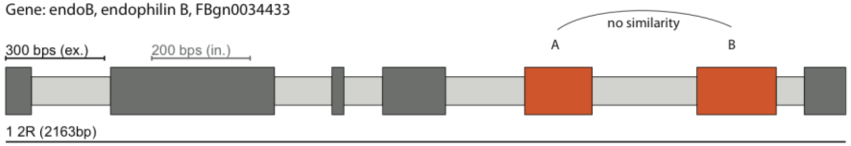


For clarity introns have been scaled up by a factor of 2.23

```
               10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA  LQRLRDRERKRVRFSCGTQTEVPLEVVAFPRGTQTVATVQSDMSTSVENLVTSNVAVTQTDFEVPDRNVSIERETMSSPF
exonB  ---AEHLRKKVTRFEDENESLMMQLKKMATRSR

               90       100       110
       ....|....|....|....|....|....|..
exonA  AGLFPPSSSSRVGQSGSLLFPSAISHVLLSGA
exonB
```

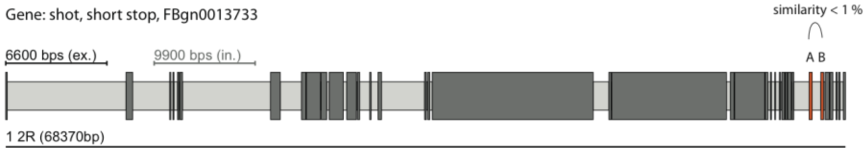
Gene: endoB, endophilin B, FBgn0034433



For clarity introns have been scaled up by a factor of 1.49

```
               10        20        30        40        50        60        70        80
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA  MQHPKPRLRINSEDVDCGPPSSVSMHSCDSDSLGGIALDSDPDPDLDKSLTNLLEDFHIEFDTTAVST
exonB  LGGPTPYIPLDVNEASASKSNISSGAAARGPGNNHSANMAATGHKPNQPMHVSTDQMQRARVLCSYDAKDHTELNLSANE
```

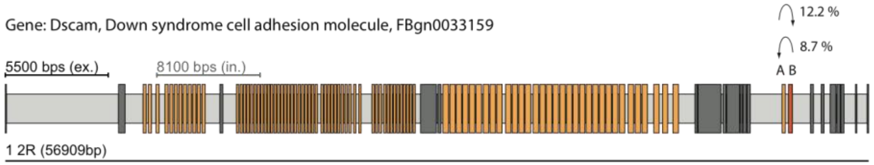
Gene: shot, short stop, FBgn0013733



For clarity introns have been scaled down by a factor of 1.50

```
               10        20        30        40        50        60
       ....|....|....|....|....|....|....|....|....|....|....|....|....|.
exonA  AKGRTNIELREQFILADGVSQSMAAFTPRRSTPNAAATASSSPHAHNGGSSNLPPYMSGQGPIIK
exonB  AD-EHLAELMPIFEKIRAQDQVPCAFPIH-----MGAGGTVFVRCNTSRSVPLSPHVLHCHPTTHW
```

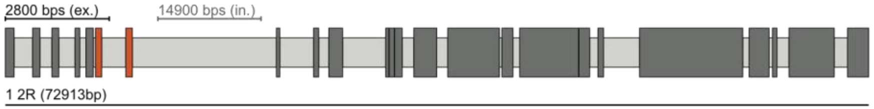
Gene: Dscam, Down syndrome cell adhesion molecule, FBgn0033159



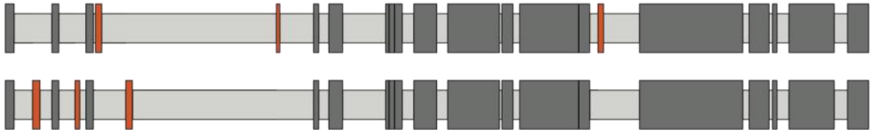
For clarity introns have been scaled down by a factor of 1.47

```
             1600      1610      1620      1630      1640      1650      1660
             |....|....|....|....|....|....|....|....|....|....|....|....|...
exon17A      GTIAPSRDLPELSAEDTIRIILS------NLNLVVPVVAALLVIIIAIIVICILRSKGN--HHK
exon17B      GTIAPLDDGSGHGNVHTRIRLPAWMPEWLDLNFMVPLIATVVVVAVGICVVCVALSRRRADDMR
```

Gene: Nipped-A, FBgn0053554

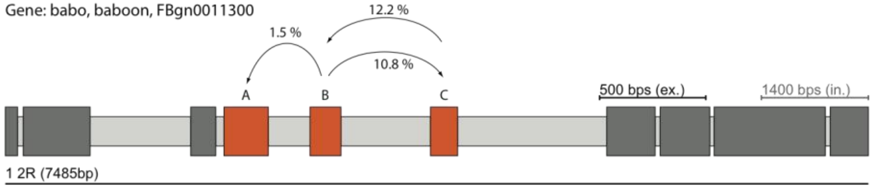


For clarity introns have been scaled down by a factor of 5.36



cDNA evidence for these two different transcripts.

Gene: babo, baboon, FBgn0011300

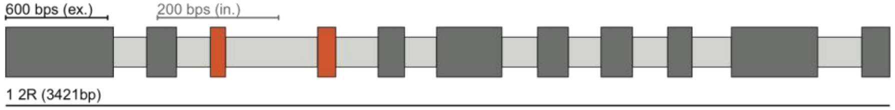


For clarity introns have been scaled down by a factor of 2.78

```
               10        20        30        40        50        60        70
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA     CLHKSQIFPPGRSIWCNDGLHGGPTARPVGRNGAHACCKDRDFCN-RFLWPKTKDQRSDRVEEGRQISVQ
exonB     CMVVKYNMQRSK-------PFECLTSNERFDTYRIDCCKS-DFCNKNEIMKRIFET
exonC     CITDQLP------------PEDPTSCKLNSEAGSSQCCAE-DFCNTRENYSGVLP
```

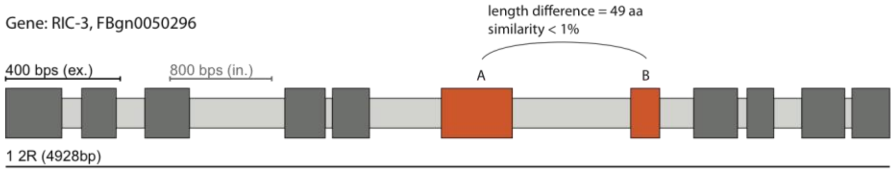
Gene: CG33012, FBgn0053012

According to RNASeq these exons are constitutive exons and are misannotated as MXE in *Dm* r5.36.



For clarity introns have been scaled up by a factor of 3.56

Gene: RIC-3, FBgn0050296

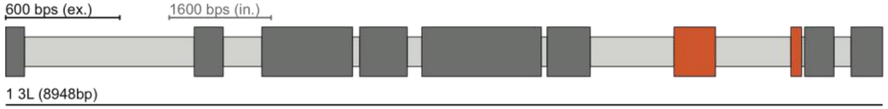


For clarity introns have been scaled down by a factor of 2.25

```
               10        20        30        40        50        60        70        80
          ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA     GAATATAAAKKPAAKDTEKELYNASVSATEVASSLSASLKSHQQLKEAEQLMEIEKLRQKLESTERAMAQLVAEMFDTTAVST
exonB     IVTAIQGLIDAADEQLNGQDKQRATSDTETDSNK
```

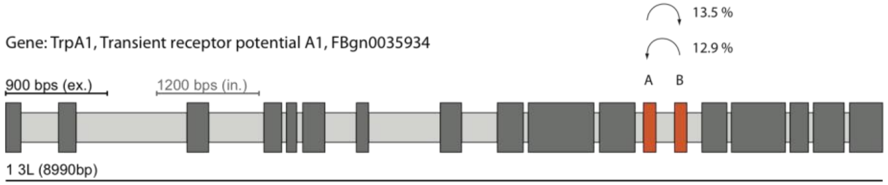
Gene: CG14168, FBgn0036044

Both exons are not part of *Dm* r5.48 anymore. There is no RNASeq evidence for any of them.



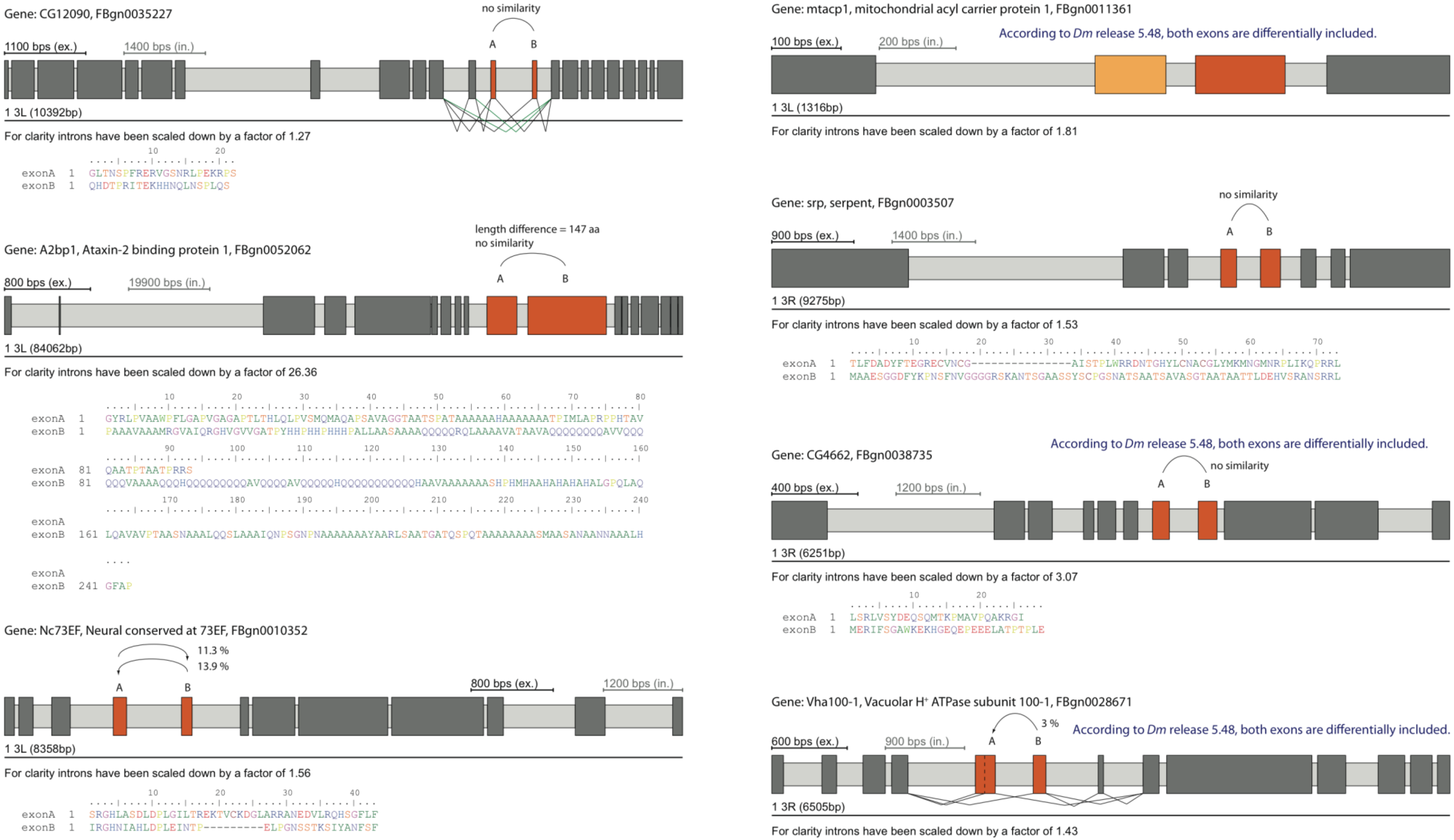
For clarity introns have been scaled down by a factor of 2.99

Gene: TrpA1, Transient receptor potential A1, FBgn0035934



For clarity introns have been scaled down by a factor of 1.32

```
               10        20        30
          ....|....|....|....|....|....|....|..
exonA     IKYSFAFLQCPFMFAKIDEKTGESITTASPIPLPALN
exonB     IKYSFWPYQKTPEQIEA-KRKEFNDPKWRPAPLAVVN
```


Gene: CG12090, FBgn0035227
no similarity
A
B
1100 bps (ex.)
1400 bps (in.)
1 3L (10392bp)
For clarity introns have been scaled down by a factor of 1.27

```
                    10        20
            ....|....|....|....|..
exonA  1   GLTNSPFRERVGSNRLPEKRPS
exonB  1   QHDTPRITEKHHNQLNSPLQS
```

Gene: A2bp1, Ataxin-2 binding protein 1, FBgn0052062
length difference = 147 aa
no similarity
A
B
800 bps (ex.)
19900 bps (in.)
1 3L (84062bp)
For clarity introns have been scaled down by a factor of 26.36

```
                    10        20        30        40        50        60        70        80
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA    1  GYRLPVAAWPFLGAPVGAGAPTLTHLQLPVSMQMAQAPSAVAGGTAATSPATAAAAAAHAAAAAAATPIMLAPRPPHTAV
exonB    1  PAAAVAAAMRGVAIQRGHVGVVGATPYHHPHHPHHHPALLAASAAAAQQQQQRQLAAAAVATAAVAQQQQQQQQAVVQQQ

                    90       100       110       120       130       140       150       160
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA   81  QAATPTAATPRRS
exonB   81  QQQVAAAAQQQHQQQQQQQQQAVQQQQAVQQQQQHQQQQQQQQQQQHAAVAAAAAAASHPHMHAAHAHAHAHALGPQLAQ

                   170       180       190       200       210       220       230       240
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|
exonA
exonB  161  LQAVAVPTAASNAAALQQSLAAAIQNPSGNPNAAAAAAAYAARLSAATGATQSPQTAAAAAAAASMAASANAANNAAALH

            ....
exonA
exonB  241  GFAP
```

Gene: Nc73EF, Neural conserved at 73EF, FBgn0010352
11.3 %
13.9 %
A
B
800 bps (ex.)
1200 bps (in.)
1 3L (8358bp)
For clarity introns have been scaled down by a factor of 1.56

```
                    10        20        30        40
            ....|....|....|....|....|....|....|....|...
exonA  1   SRGHLASDLDPLGILTREKTVCKDGLARRANEDVLRQHSGFLF
exonB  1   IRGHNIAHLDPLEINTP---------ELPGNSSTKSIYANFSF
```

Gene: mtacp1, mitochondrial acyl carrier protein 1, FBgn0011361
According to Dm release 5.48, both exons are differentially included.
100 bps (ex.)
200 bps (in.)
1 3L (1316bp)
For clarity introns have been scaled down by a factor of 1.81
Gene: srp, serpent, FBgn0003507
no similarity
A
B
900 bps (ex.)
1400 bps (in.)
1 3R (9275bp)
For clarity introns have been scaled down by a factor of 1.53

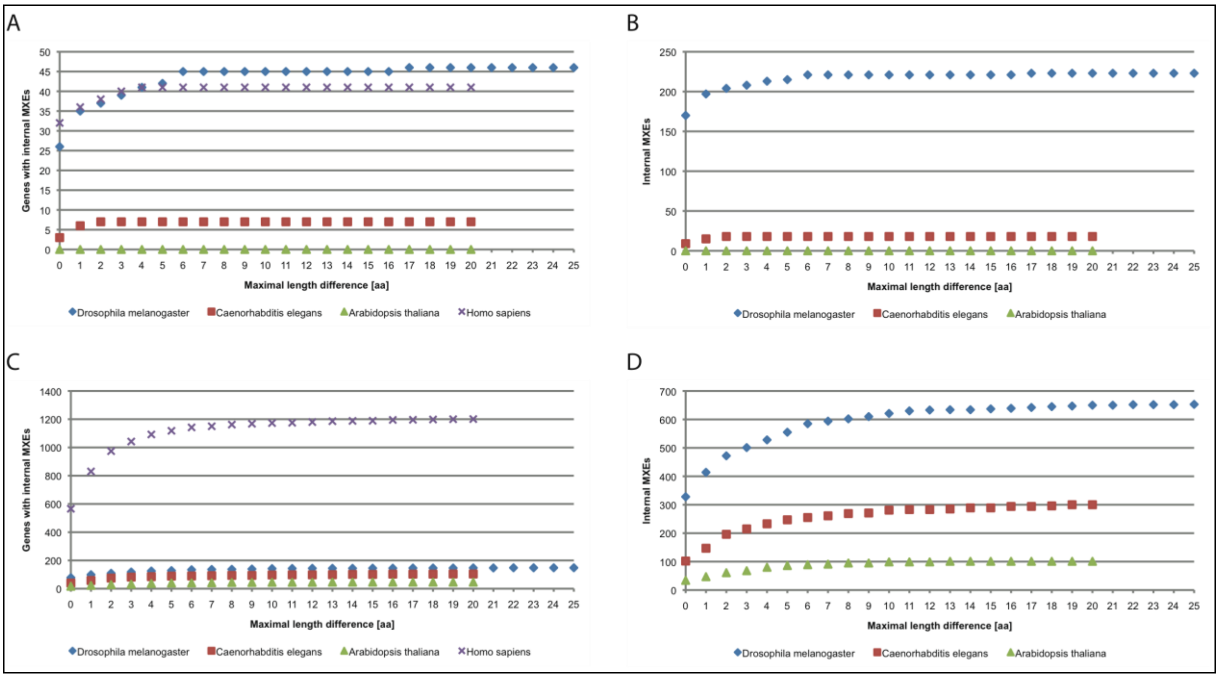
```
                    10        20        30        40        50        60        70
            ....|....|....|....|....|....|....|....|....|....|....|....|....|....|...
exonA  1   TLFDADYFTEGRECVNCG---------------AISTPLWRRDNTGHYLCNACGLYMKMNGMNRPLIKQPRRL
exonB  1   MAAESGGDFYKPNSFNVGGGGRSKANTSGAASSYSCPGSNATSAATSAVASGTAATAATTLDEHVSRANSRRL
```

Gene: CG4662, FBgn0038735
According to Dm release 5.48, both exons are differentially included.
no similarity
A
B
400 bps (ex.)
1200 bps (in.)
1 3R (6251bp)
For clarity introns have been scaled down by a factor of 3.07

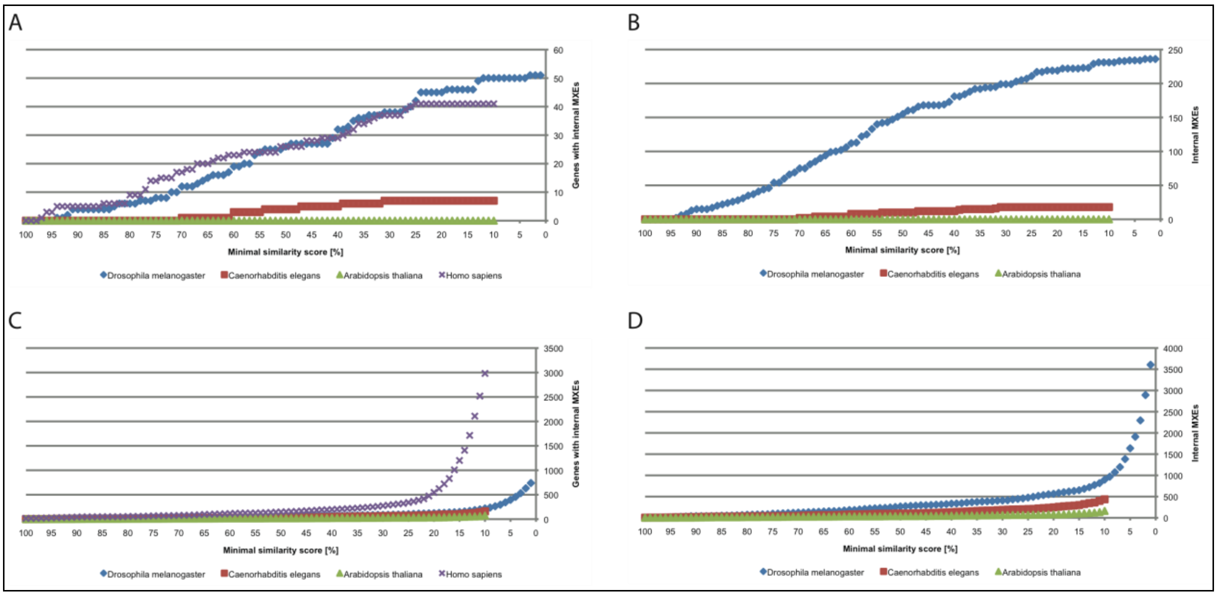
```
                    10        20
            ....|....|....|....|....|....
exonA  1   LSRLVSYDEQSQMTKPMAVPQAKRGI
exonB  1   MERIFSGAWKEKHGEQEPEEELATPTPLE
```

Gene: Vha100-1, Vacuolar H+ ATPase subunit 100-1, FBgn0028671
3 %
According to Dm release 5.48, both exons are differentially included.
A
B
600 bps (ex.)
900 bps (in.)
1 3R (6505bp)
For clarity introns have been scaled down by a factor of 1.43

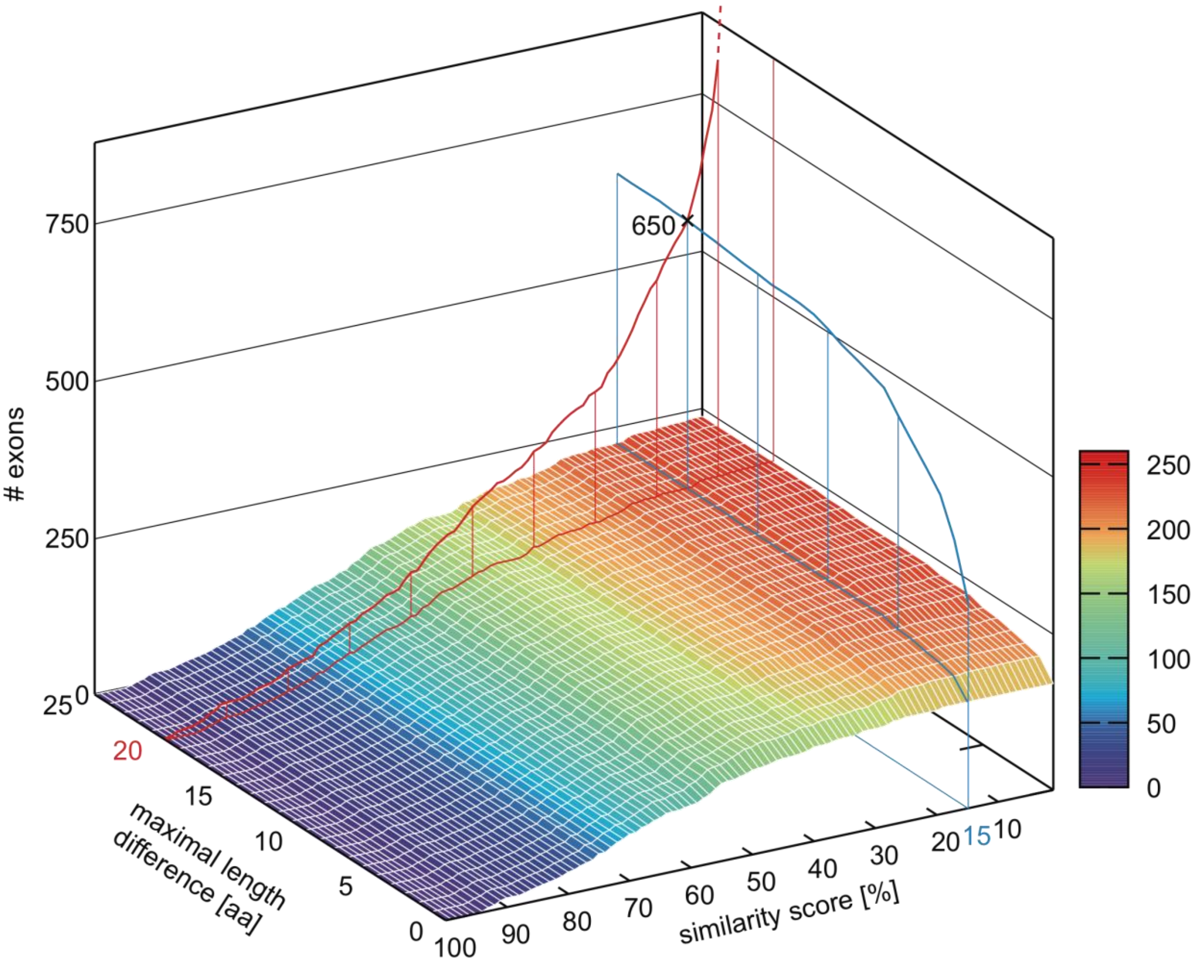
Supplementary Figure S4. Genes containing annotated mutually exclusive spliced exons (MXEs), which could not be reconstructed using the default parameters. These MXEs are shown in dark orange. MXEs found with the default prediction parameters are shown in light orange. Of the annotated MXEs, which we could not reconstruct, four pairs of exons do not show any sequence similarity, three have length differences of more than 50 aa, three are annotated as differentially included in the latest release (Dm r5.48), one pair does not consist of neighboring exons, and two pairs of exons have completely been removed from the latest annotation. Thus, the sensitivity of our method is considerably higher than 83.5% (218 of the annotated internal MXEs reconstructed). All transcripts are represented 5’ to 3’. The color coding is explained in the legend.



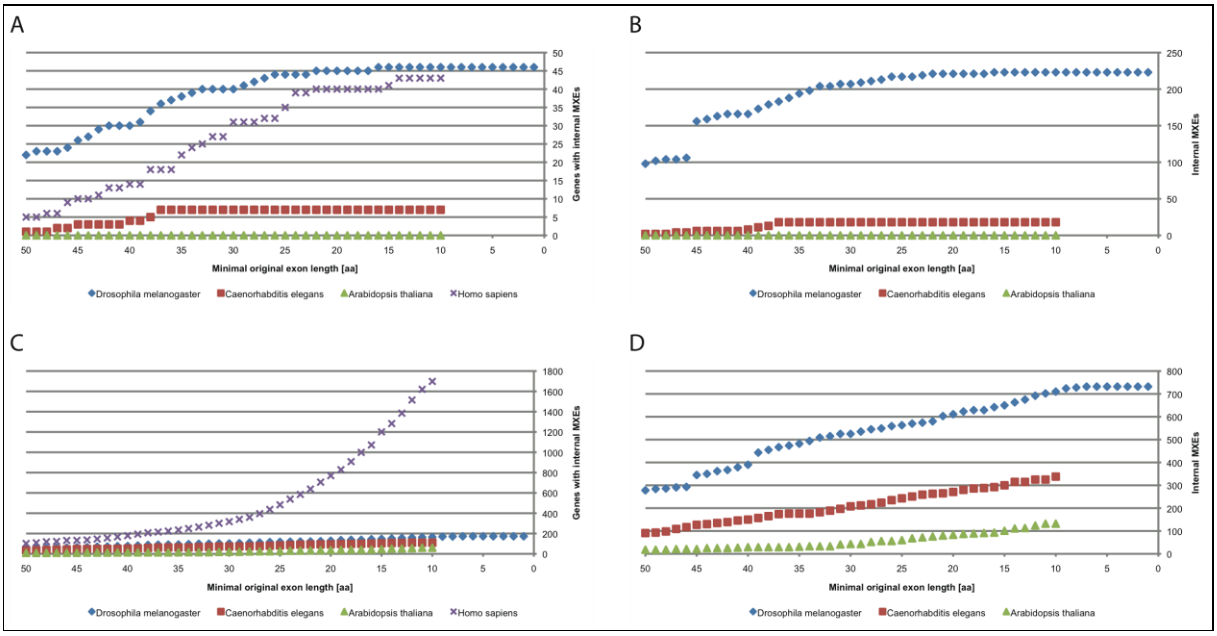
Supplementary Figure S5. Reconstructed and predicted internal mutually exclusive spliced exons (MXEs) at a similarity score cut-off of 15%. Apart from the MXEs that we cannot reconstruct because they are out of the scope of our preconditions (no sequence similarity, huge length difference), we assessed the sensitivity of our method when using a length difference of 20 residues and a similarity score of 15% as standard cut-offs. Given a similarity score of at least 15%, the analysis of the reconstructed MXEs shows that all annotated MXEs have length differences of less than 20 residues (A, B). A similar distribution is found for the length difference of the internal MXEs that we predict newly (C, D). A) Number of genes containing annotated internal MXEs that could be reconstructed at a given length difference cut-off having a similarity score of at least 15%. B) Number of annotated internal MXEs that could be reconstructed at a given length difference cut-off having a similarity score of at least 15%. C) Number of genes containing predicted internal MXEs (including annotated MXEs that could be reconstructed) with a similarity score of at least 15% at a given length difference. D) Number of internal MXE candidates (including annotated MXEs that could be reconstructed) with a similarity score of at least 15% at a given length difference.



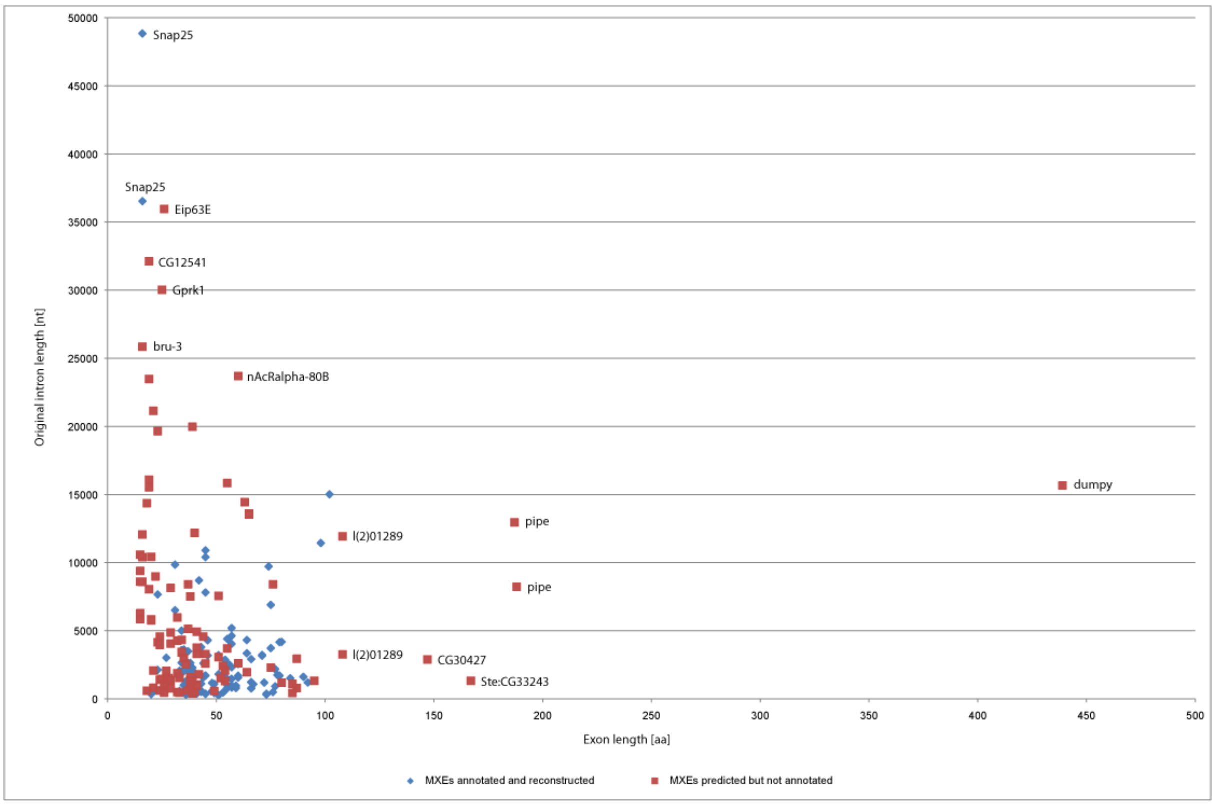
Supplementary Figure S6. Reconstructed and predicted internal mutually exclusive spliced exons (MXEs) at a length difference cut-off of 20 residues. To assess the suitability of the sequence similarity cut-off of 15% within the preconditions of our prediction method, we analysed the distribution of the annotated exons with a length difference of less than 20 residues (A, B). In contrast to the MXEs of *Hs* and *Ce*, the MXEs of *Dm* do not show a pronounced plateau. The number of predicted MXE candidates even shows an exponential increase below a similarity score of 10% (*Dm*) and 15% (*Hs*), respectively (C, D). A) Number of genes containing annotated internal MXEs that could be reconstructed at a given sequence similarity score cut-off and having a length difference of less than 20 aa. B) Number of internal MXEs that could be reconstructed at a given sequence similarity score cut-off and having a length difference of less than 20 aa. C) Number of genes containing internal MXE candidates (including annotated MXEs that could be reconstructed) predicted at a given sequence similarity score cut-off and having a length difference of less than 20 aa. D) Number of internal MXE candidates (including annotated MXEs that could be reconstructed) predicted at a given sequence similarity score cut-off and having a length difference of less than 20 aa.



Supplementary Figure S7. Assessing annotated and predicted mutually exclusive spliced exons (MXEs) in *Drosophila melanogaster*. This figure comprises information from the previous figures (Supplemental Figs. S4 and S5) for *Drosophila melanogaster* and shows the dependence of the number of internal MXEs on the maximal length difference and similarity between search exon and MXE candidate. The figure is similar to Fig. 1A of the main manuscript except that the number of exons is shown here in contrast to the number of genes, reflecting that many genes contain several clusters of MXEs and clusters with more than two MXEs. The colored grid denotes the number of MXEs as annotated in FlyBase r5.36 that were also predicted by WebScipio. The red and blue lines mark the number of predicted MXE candidates at the maximal length difference of 20 amino acids and at the minimal similarity score of 15%, respectively.

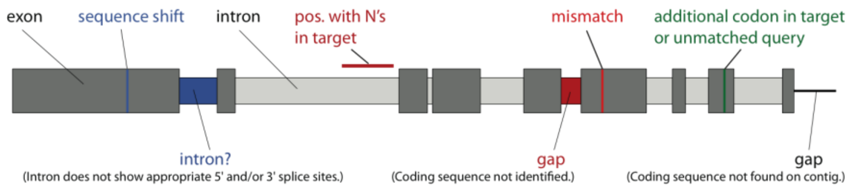


Supplementary Figure S8. Reconstructed and predicted internal mutually exclusive spliced exons (MXEs) in dependence of a minimal original exon length. The sequences of very short exons do not contain enough complexity to exclude the identification of “similar” exons, especially if they are surrounded by long introns. Luckily, short exons within genes are rather rare and are predominantly found at gene borders. In order to avoid the inclusion of many false positives we introduced the parameter “minimal original exon length”. Annotated MXEs, which we can reconstruct with a length difference cut-off of 20 residues and a similarity score cut-off of 15%, are all longer than ten residues (A, B). For the initial search for MXE candidates in *Drosophila* we set this parameter to one residue (A, B). However, only a few candidates were found for exons shorter than 15 residues. Therefore, we set the minimal original exon length parameter to 15 residues for the analysis of the *Drosophila* genome and for the search for MXE candidates in the other model organisms (C, D). The value seems appropriate for *Caenorhabditis* and *Arabidopsis* while the number of MXE candidates is increasing exponentially in dependence of the search exon length in human. This is most probably due to the much longer introns in human compared to the other species analysed. A) Number of genes containing annotated internal MXEs in dependency of the length of the MXEs that could be reconstructed at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa. B) Number of annotated internal MXEs in dependency of the length of the MXEs that could be reconstructed at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa. C) Number of genes containing internal MXE candidates in dependency of the length of the MXEs that were predicted at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa. D) Number of internal MXE candidates in dependency of the length of the MXEs that were predicted at a sequence similarity score cut-off of 15% and a length difference of less than 20 aa.

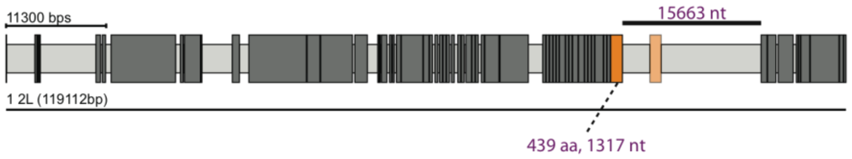


Supplementary Figure S9. Scatter plot of the internal mutually exclusive spliced exon (MXE) candidates. Blue, annotated in r5.36; red, predicted MXEs. This figure is similar to Fig. 1B of the main manuscript. However, some of the genes containing either very long introns or very long exons, for which MXE candidates were predicted, are indicated. If exons are short the complexity of the translations will be low and chances will thus be high to predict false positive candidates, especially if the surrounding introns are long. The introns surrounding annotated MXEs vary from 50 to 50,000 nucleotides. Although most introns range up to 15,000 nucleotides we therefore cannot assume that potential MXE candidates in longer introns are false predictions. MXE candidates, which are also conserved in other arthropods, were found for example in very long introns of the *nAcRalpha-80B* and *bruno-3* genes. In the case of long exons, it is very unlikely that by chance the translation of intronic region shows sequence similarity to neighbouring exons. However, if long exon candidates are found in long introns these could also, instead of being part of a cluster of MXEs, belong for example to mis-annotated tandemly arrayed gene duplicates or belong to the very rare cases of clusters of exons, which share sequence homology and are spliced as cluster. Here, we also found false positive MXE candidates, that are annotated in the latest FlyBase release as belonging to different tandemly arrayed gene duplicates (*CG33243* gene region; FlyBase r5.48), that were derived from isoforms containing different, mutually exclusive clusters of exons (*CG30427* gene; *pipe* gene[49]) and that were part of some isoforms of the gigantic *dumpy* gene that displays a complex pattern of alternative splicing[50].

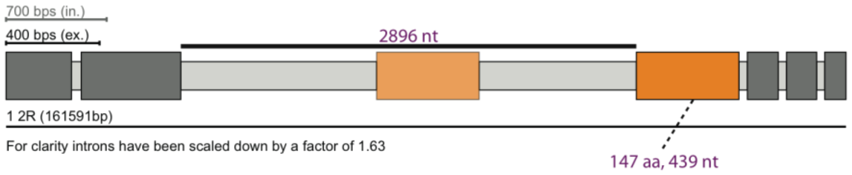
Legend


exon
sequence shift
intron
pos. with N's in target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

Gene: dp, dumpy, FBgn0053196
Polypeptide: FBpp0293673



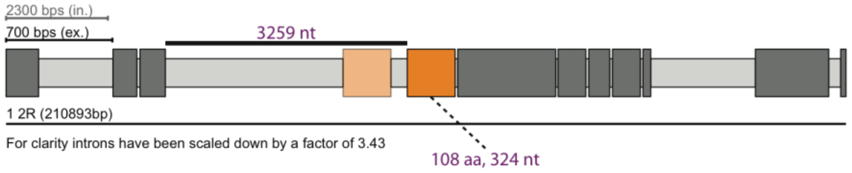
Gene: CG30427, FBgn0043792
Polypeptide: CG30427-PB, FBpp0072320



For clarity introns have been scaled down by a factor of 1.63

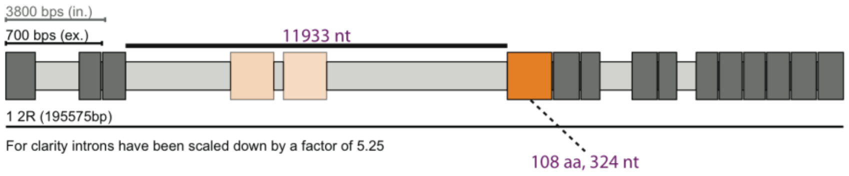
mutually exclusive splicing of cluster of C-terminal exons

Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PA, FBpp0085469



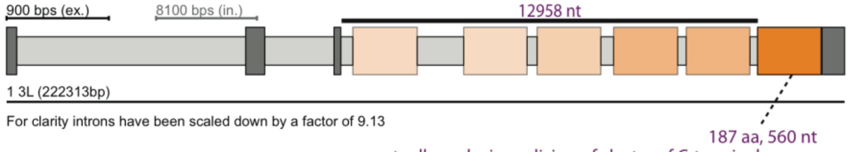
For clarity introns have been scaled down by a factor of 3.43

Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PH, FBpp0290635



For clarity introns have been scaled down by a factor of 5.25

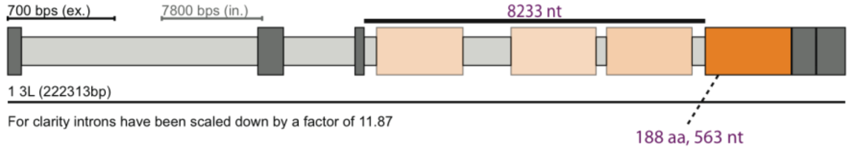
Gene: pip, pipe, FBgn0003089
Polypeptide: pip-PA, FBpp0074777



For clarity introns have been scaled down by a factor of 9.13

mutually exclusive splicing of cluster of C-terminal exons

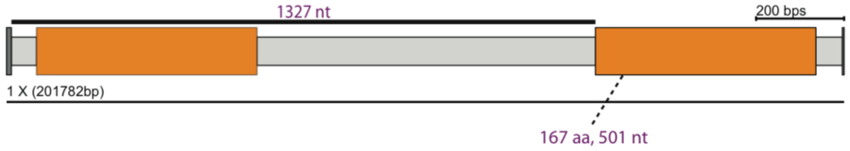
Gene: pip, pipe, FBgn0003089
Polypeptide: pip-PG, FBpp0074775



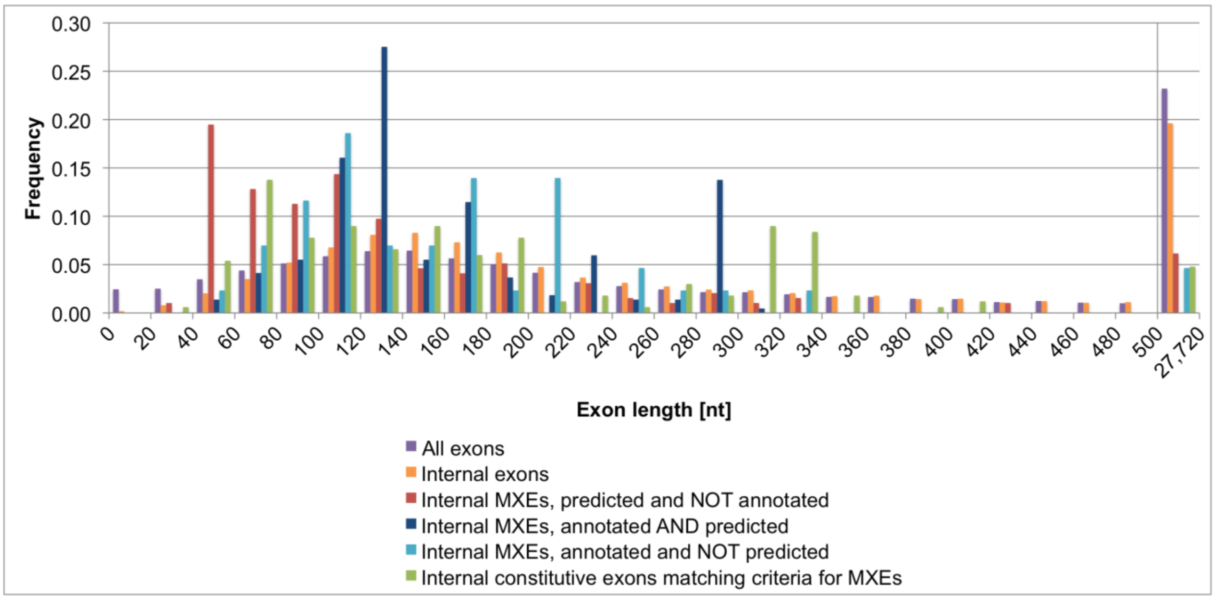
For clarity introns have been scaled down by a factor of 11.87

mutually exclusive splicing of cluster of C-terminal exons

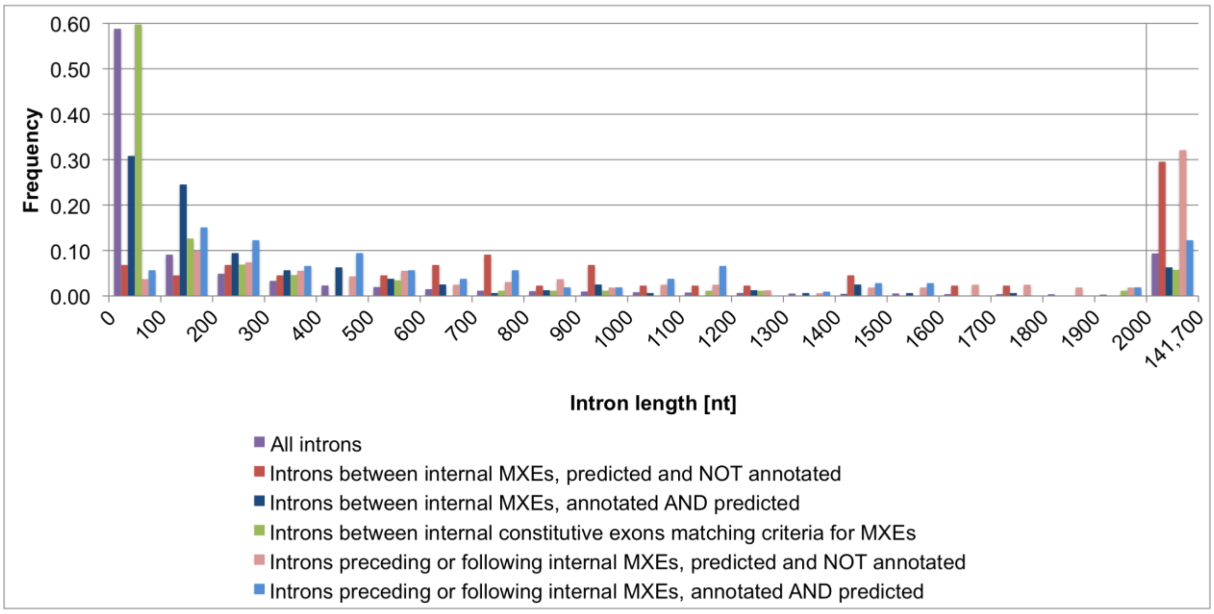
Gene: Ste:CG33243, FBgn0053243
Polypeptide: Ste:CG33243-PB, FBpp0289369



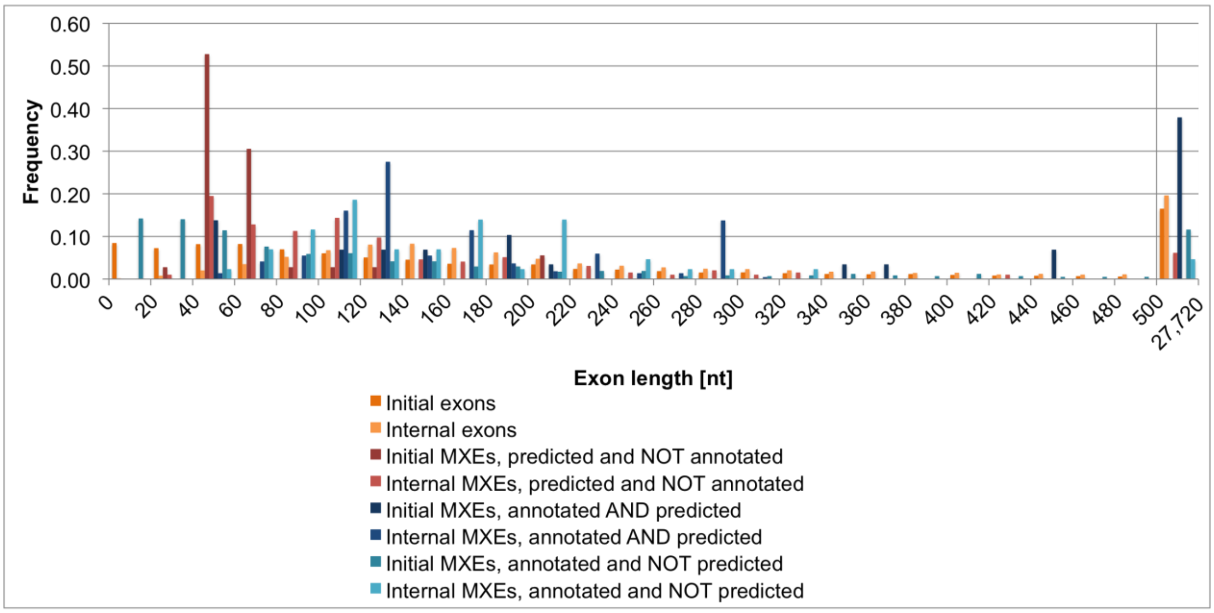
Supplementary Figure S10. Examples of mutually exclusive spliced exon (MXE) candidates found for long introns. All transcripts are represented 5’ to 3’. Colored big bars represent MXEs. The darkest colored bar is the exon that was included in the query sequence, while the lighter colored bars represent identified MXEs. The higher the similarity between the candidate and the query exon the darker the color of the candidate (100% identity would result in the same color). The opacity of the colors of each alternative exon corresponds to the alignment score of the alternative exon to the original one. The color coding is explained in the legend.



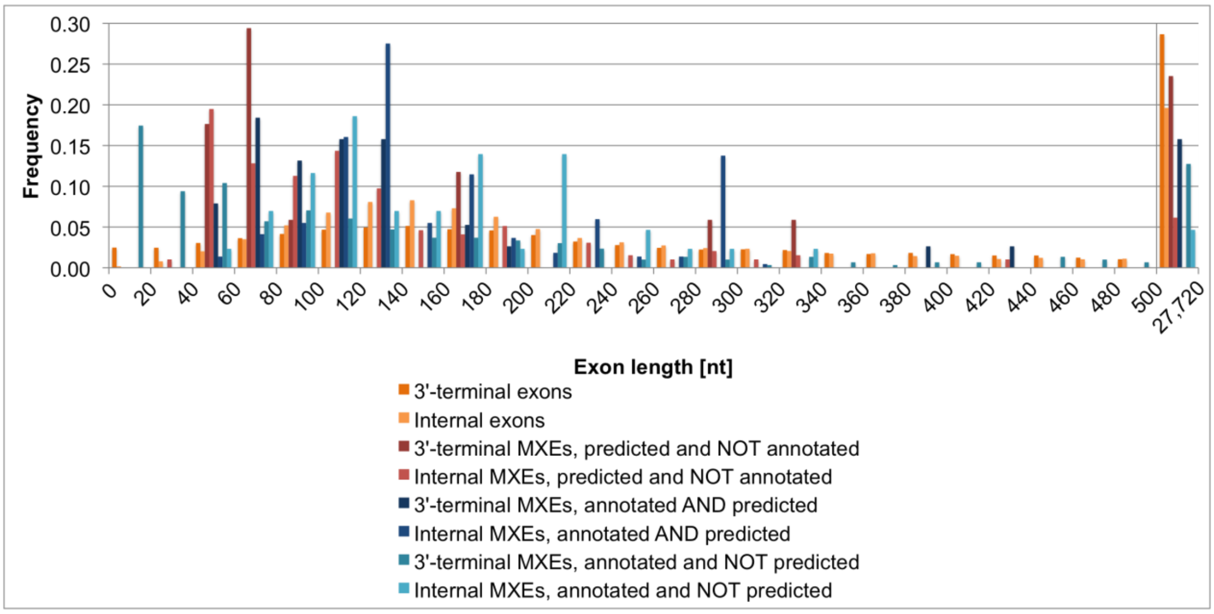
Supplementary Figure S11. Comparison of exon lengths. Various subsets of annotated and predicted mutually exclusive spliced exons (MXEs) are compared to all exons and internal constitutive exons sharing our criteria for MXEs. The exon lengths of the annotated and predicted MXEs show almost the same distribution like all exons of *Drosophila* with a broad peak around 140 residues. Interestingly, there is a second smaller peak for the length of MXEs at 300 amino acids. The comparison of the annotated MXEs to the predicted MXE candidates shows similar distributions meaning that the predictions represent normal MXEs. The internal MXEs that are annotated and that we cannot reconstruct also display a similar distribution but in addition tend to represent larger exons as compared to the other sets. Surprisingly, the constitutive exons sharing our criteria for MXEs show three striking peaks at 80, 320 and 340 residues but show a local minimum at 140 residues. This supports the notion that the predicted MXEs rather represent MXEs than potential constitutively spliced exons.



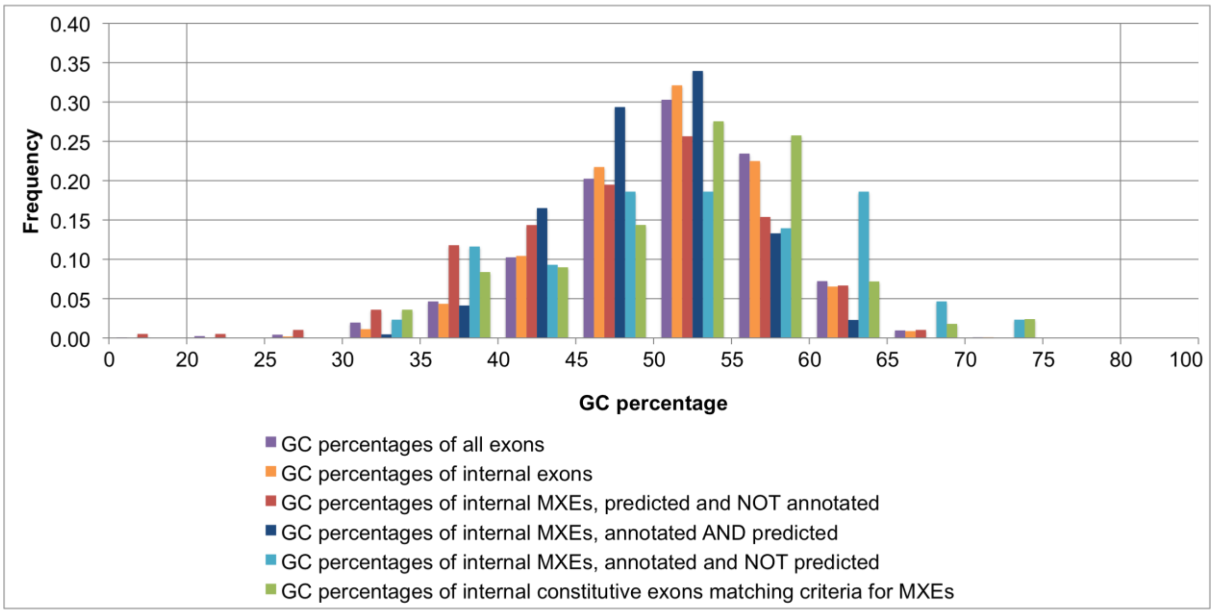
Supplementary Figure S12. Comparison of intron lengths. Introns next to various subsets of annotated and predicted mutually exclusive spliced exons (MXEs) are compared to all introns and introns next to internal constitutive exons sharing our criteria for MXEs. The comparison of the intron lengths shows a broad distribution with a tendency to rather short introns (< 300 bp).



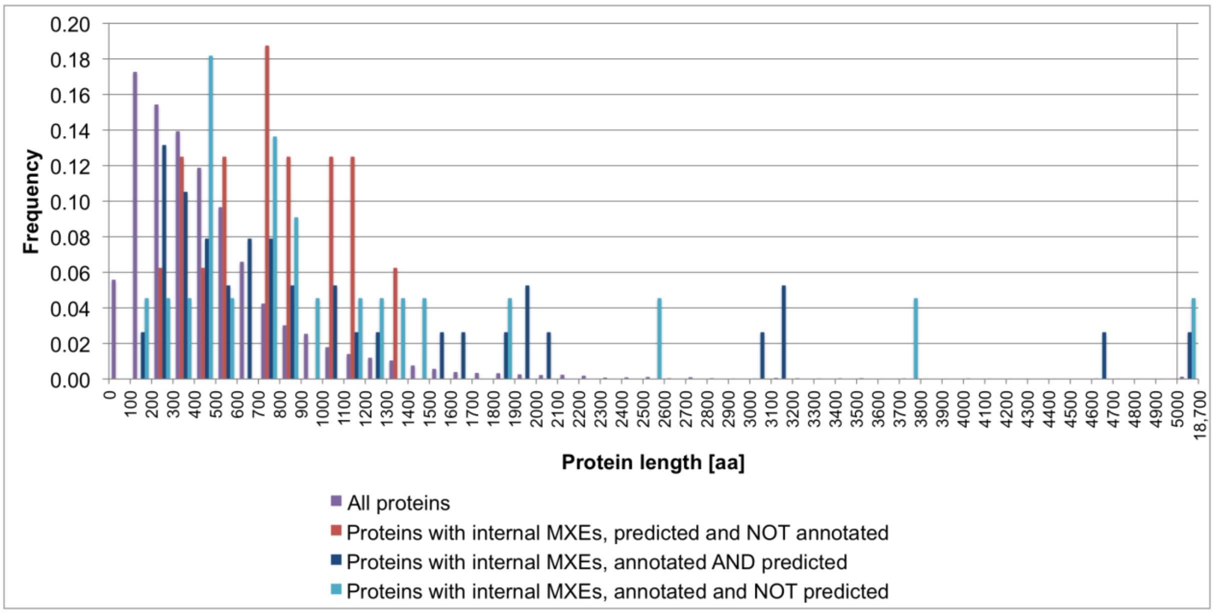
Supplementary Figure S13. Comparison of exon lengths of initial exons of multi-exon genes. Various subsets of annotated and predicted initial exons matching the criteria for mutually exclusive spliced exons (MXEs) are compared to all exons and internal MXEs. Because the algorithm is based on protein coding sequence it could be possible that the initial and terminal exons of the coding region are not the initial and terminal exons of the transcripts. In this case, these exons would be regarded as internal exons. Therefore, we also analysed candidate exons of initial and terminal exons that share the criteria of MXEs. In general, initial and terminal exons of multi-exon genes are considerably shorter than internal exons. Some of these match the criteria of MXEs. Of those, almost all code for at least 40 residues. In these cases it is unlikely that pseudo-duplicates of low-complexity exons were found.



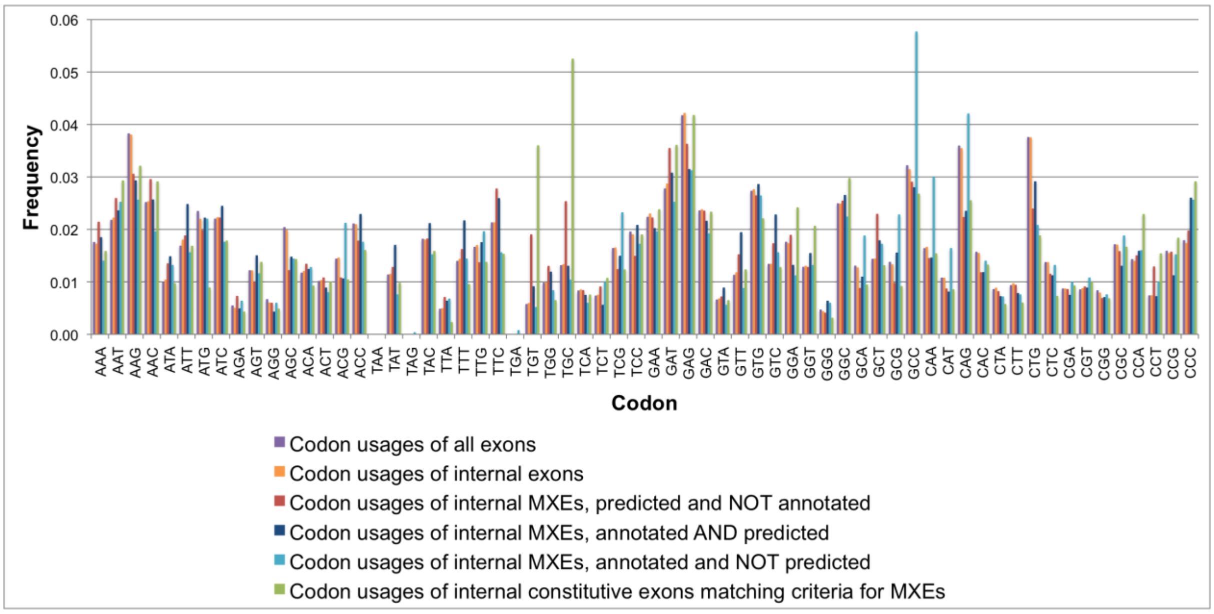
Supplementary Figure S14. Comparison of exon lengths of terminal exons of multi-exon genes. Various subsets of annotated and predicted terminal exons matching the criteria for mutually exclusive spliced exons (MXEs) are compared to all exons and internal MXEs. Because the algorithm is based on protein coding sequence it could be possible that the initial and terminal exons of the coding region are not the initial and terminal exons of the transcripts. In this case, these exons would be regarded as internal exons. Therefore, we also analysed candidate exons of initial and terminal exons that share the criteria of MXEs. In general, initial and terminal exons of multi-exon genes are considerably shorter than internal exons. Some of these match the criteria of MXEs. Of those, almost all code for at least 40 residues. In these cases it is unlikely that pseudo-duplicates of low-complexity exons were found.



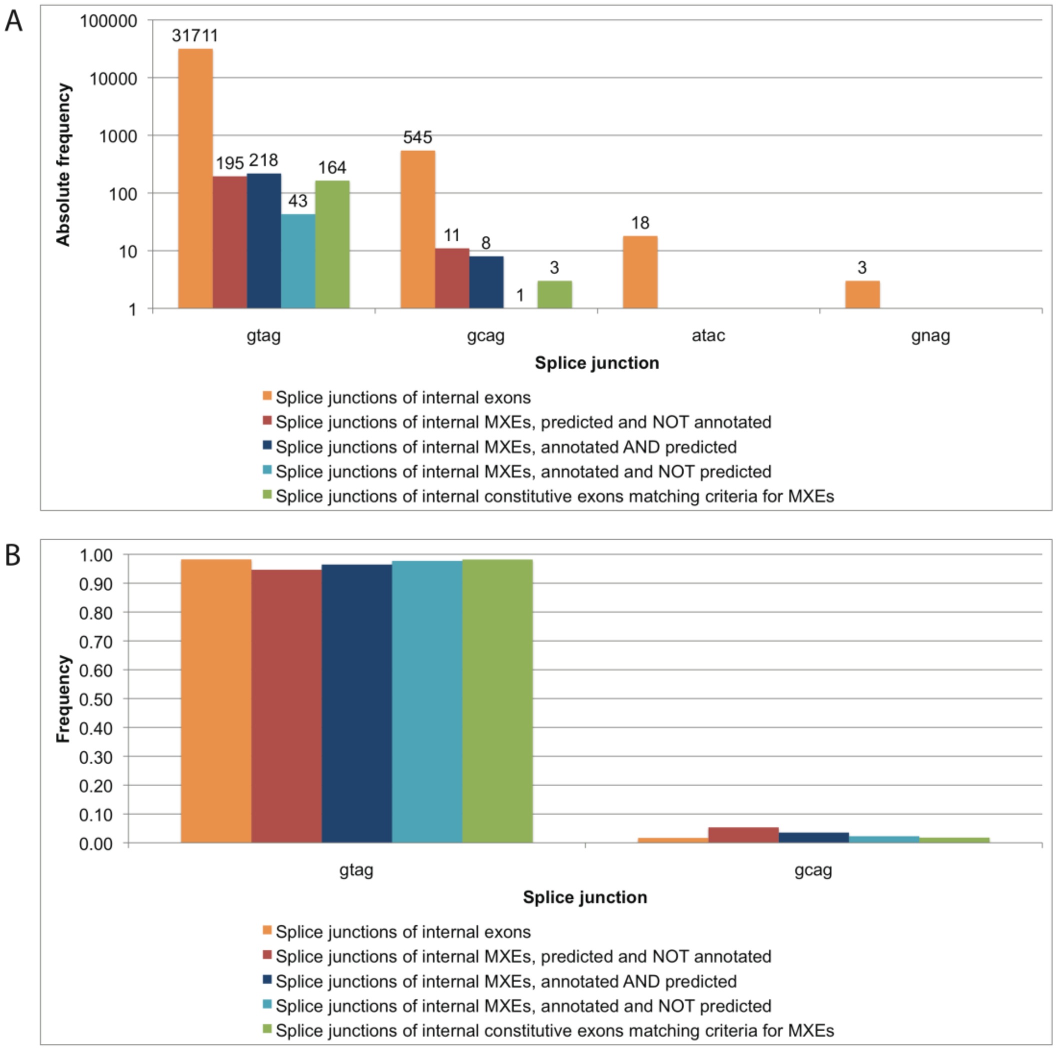
Supplementary Figure S15. Comparison of GC content of exons. The GC content of all exons (reference) is compared to the GC content of annotated and predicted internal mutually exclusive spliced exons (MXEs) and to internal constitutive exons sharing our criteria for MXEs. The GC content of all exons shows a broad distribution around 55%. The MXEs, which we cannot reconstruct, and the constitutive exons sharing our criteria of MXEs have a broader GC content distribution with a remarkably higher percentage of exons with GC contents of 60 to 75%. The distribution of the GC content of the predicted MXEs is similar to the distribution of the annotated MXEs except for a slight increase of exons with GC contents of 40 to 45%.



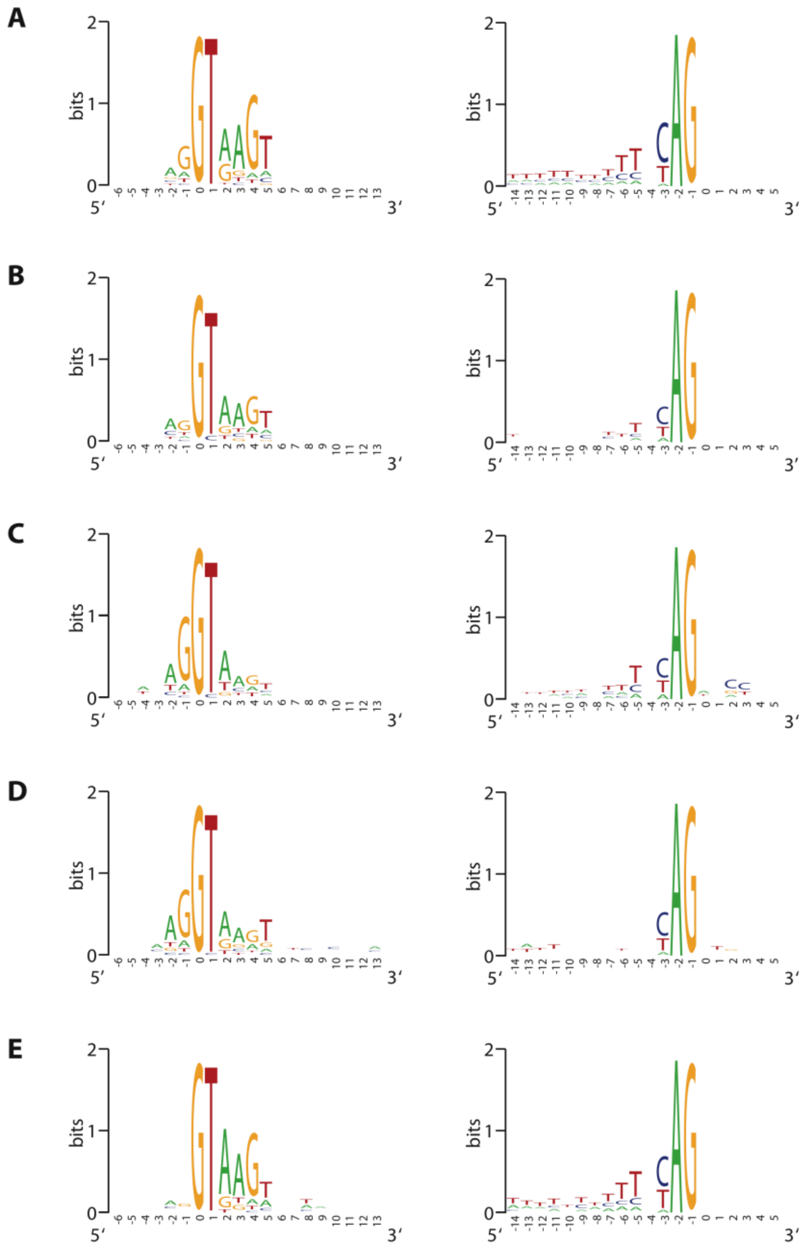
Supplementary Figure S16. Comparison of the lengths of the translations of one isoform per gene. For the reconstruction of the translations of the genes containing mutually exclusive spliced exons (MXEs) only one isoform has been chosen and only one exon of each cluster. For the protein lengths of all proteins, only the isoforms "A" were considered. To assess whether MXEs are predominantly found in proteins of a certain size, we analysed the lengths of the translations. Here, from each alternatively spliced gene (independently of alternative splicing type) only one transcript and the corresponding translation were considered. Proteins built with MXEs are relatively longer than the average proteins. The distribution of the proteins with annotated MXEs and with predicted MXE candidates is very similar.



Supplementary Figure S17. Comparison of the codon usage. Codon usage in all exons is compared to that of genes containing annotated or predicted mutually exclusive spliced exons (MXEs) and to that of internal constitutive exons sharing our criteria for MXEs. The codon usage of the MXEs (annotated and predicted) is very similar to the codon usage of all or all internal exons except for the codons AAG, AGC CAG and CTG that are slightly less represented in MXEs. Strikingly, the percentage of cysteine-coding codons (TGT and TGC) is five times higher in constitutive exons sharing our criteria of MXEs compared to all exons, and the MXEs, that are annotated in FlyBase but that we cannot reconstruct, have a considerably higher content of alanines (GCC codon) and glutamines (CAA and CAG codons).

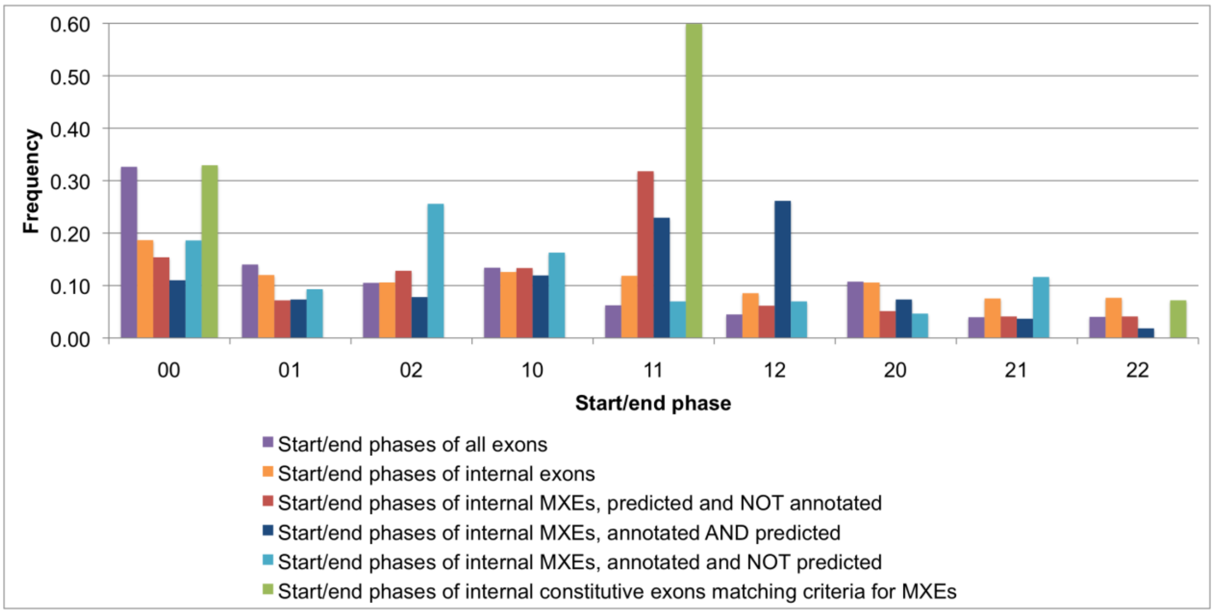


Supplementary Figure S18. Comparison of splice junctions. The splice junctions of all introns are compared to those of the putative introns between a mutually exclusive spliced exon (MXE) and the next constitutive exon before and after a cluster of MXEs. MXEs are separated in annotated or predicted MXEs and compared to internal constitutive exons sharing our criteria for MXEs. Total numbers are given in (A) and percentages in (B). As known, by far most introns have the splice junctions GT---AG followed by the GC---AG slice junctions (A). Only a few of the annotated introns have other splice junctions. The percentage of the GC---AG splice junction in introns surrounding MXEs is slightly higher than that of all introns (B). These numbers are, however, hard to interpret because the total number of MXEs spliced by GC---AG is very low.



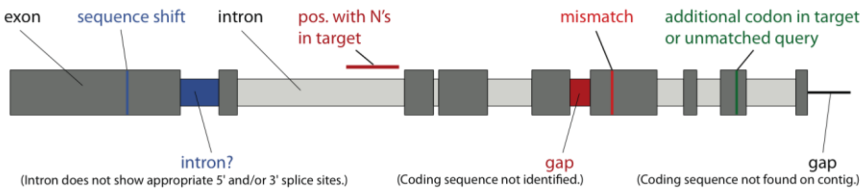
Supplementary Figure S19. Conservation of intron splice junctions. The weblogos were generated from the aligned 14 nucleotides of the intron and six nucleotides of the exon of both the 5’- and 3’-splice sites. The height of the letters represents the degree of conservation. A) All internal introns. B) Predicted internal mutually exclusive spliced exons (MXEs) that were not annotated. C) Annotated and reconstructed internal MXEs. D) Annotated but not reconstructed internal MXEs. E) Internal constitutive exons matching our criteria of MXEs. Splice junctions display sequence conservation beyond the two-base splice site.

Characteristic to all internal exons (pattern strongly dominated by constitutive exons) and the constitutive exons sharing our criteria of MXEs are the considerably stronger conservation of the bases AGT in positions +4, +5 and +6 of the intron. In contrast, the introns following the MXEs (annotated and predicted) have a stronger conserved G in position -1. The 3’ ends of the introns before the MXEs have similar patterns as compared to all introns.

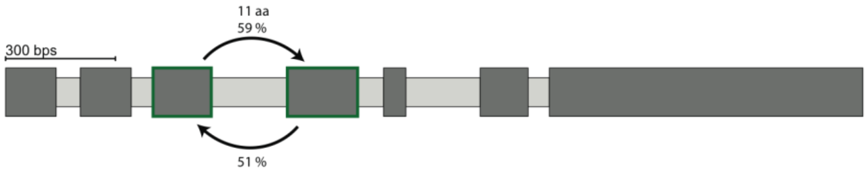


Supplementary Figure S20. Comparison of start/end phases of exons. A strong indication for mutually exclusive splicing is the impossibility to incorporate more than one of the mutually exclusive spliced exons (MXEs) of a cluster into the final transcript because of the incompatibility of the splice site phases. Exons can be classified based on the phase of the flanking intron: symmetric exons are 0-0 (intron interrupts the reading frame between two consecutive codons), 1-1 (intron interrupts the reading frame between the first and second base of a codon) and 2-2, and asymmetric exons are 0-1, 0-2, 1-0, 1-2, etc. Symmetric exons are the only ones that can be spliced in succession without changing the reading frame. Thus, constitutive exons sharing our criteria of MXEs comprise only symmetric exons. Compared to the annotated MXEs, the predicted MXEs show a slightly higher percentage of symmetric exons. Therefore, these potential exon candidates could also be spliced constitutively or they could be incorporated in a differentially included manner.

Legend

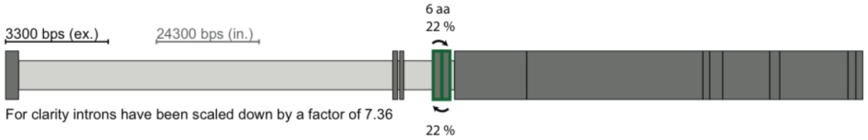

exon
sequence shift
intron
pos. with N's in target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

Gene: C901, FBgn0021742
Polypeptide: C901-PA , FBpp0073256



Constitutively spliced exons.

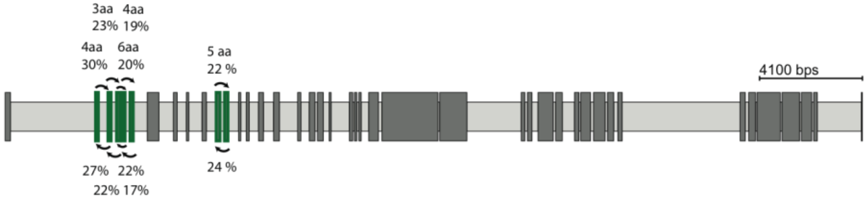
Gene: Megalin, FBgn0261260
Polypeptide: Megalin-PA, FBpp0291363



For clarity introns have been scaled down by a factor of 7.36

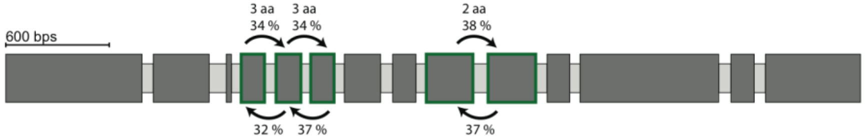
Constitutively spliced exons.

Gene: trol, terribly reduced optic lobes, FBgn0261451
Polypeptide: trol-PD, FBpp0070440



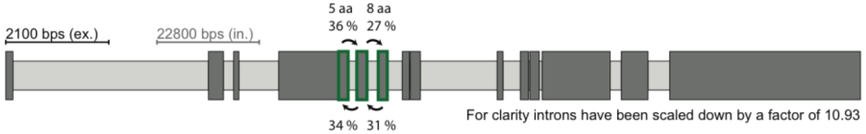
Constitutively spliced exons.

Gene: CG15570, FBgn0029697
Polypeptide: CG15570-PA, FBpp0070613



Constitutively spliced exons.

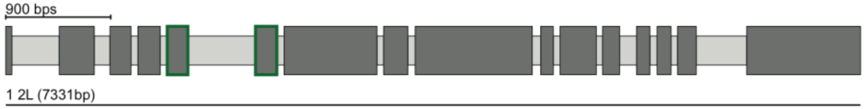
Gene: Ten-a, Tenascin accessory, FBgn0259240
Polypeptide: Ten-a-PD, FBpp0289136



For clarity introns have been scaled down by a factor of 10.93

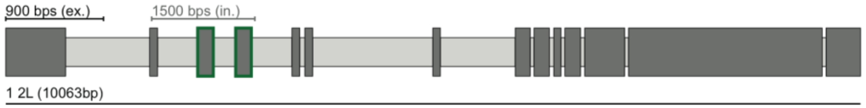
Constitutively spliced exons.

Gene: CG10186, FBgn0032797
Polypeptide: CG10186-PA , FBpp0080818



Constitutively spliced exons.

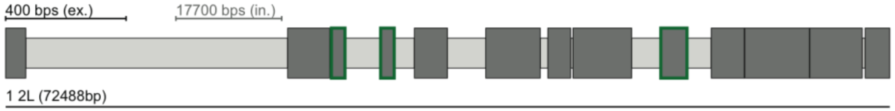
Gene: rk, rickets, FBgn0003255
Polypeptide: rk-PA, FBpp0080183



For clarity introns have been scaled down by a factor of 1.58

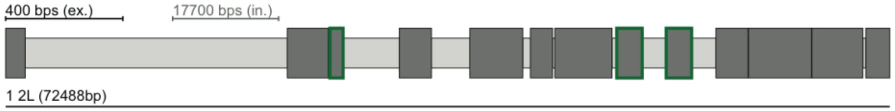
Constitutively spliced exons.

Gene: nAcRalpha-30D, nicotinic Acetylcholine Receptor α 30D, FBgn0032151
Polypeptide: nAcRα-30D-PD , FBpp0079503

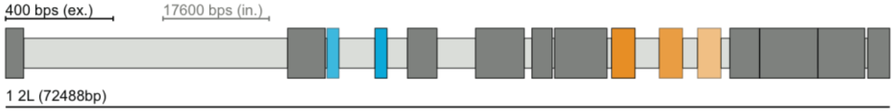


For clarity introns have been scaled down by a factor of 50.34

Polypeptide: nAcRα-30D-PE , FBpp0079502



For clarity introns have been scaled down by a factor of 48.97



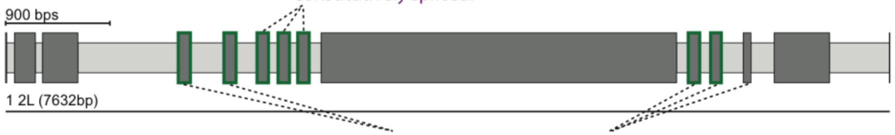
For clarity introns have been scaled down by a factor of 44.59

RNASeq supports differentially included splicing, last exon candidate not reported before.

RNASeq supports mutually exclusive splicing

Gene: CG8086, FBgn0032010
Polypeptide: CG8086-PG , FBpp0297483

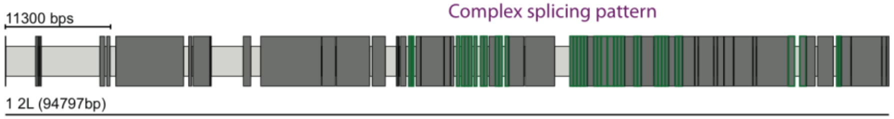
Constitutively spliced.



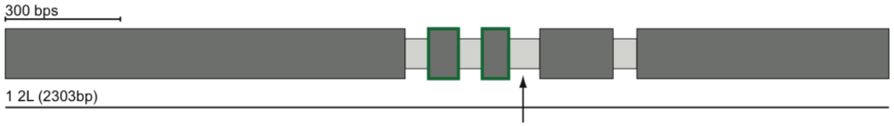
Differentially included splicing supported by many cDNAs.

Gene: dp, dumpy, FBgn0053196
Polypeptide: FBpp0288445

Complex splicing pattern

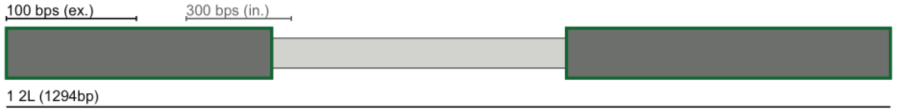


Gene: Rrp1, Recombination repair protein 1, FBgn0004584
Polypeptide: Rrp1-PA, FBpp0077362



Intron retention

Gene: CG10039, FBgn0031581
Polypeptide: CG10039-PA, FBpp0077196

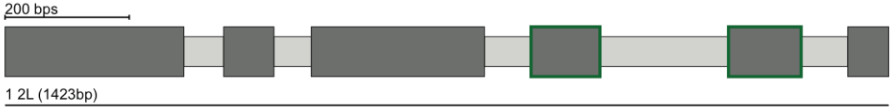


For clarity introns have been scaled down by a factor of 3.71

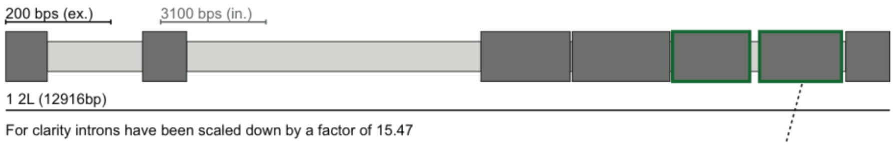
Two separate genes. Tandem gene duplicates. Annotation corrected in FlyBase.

Gene: Tsp26A, Tetraspanin 26A, FBgn0031760
Polypeptide: Tsp26A-PB, FBpp0111937

Differentially included splicing supported by cDNAs and RNASeq.

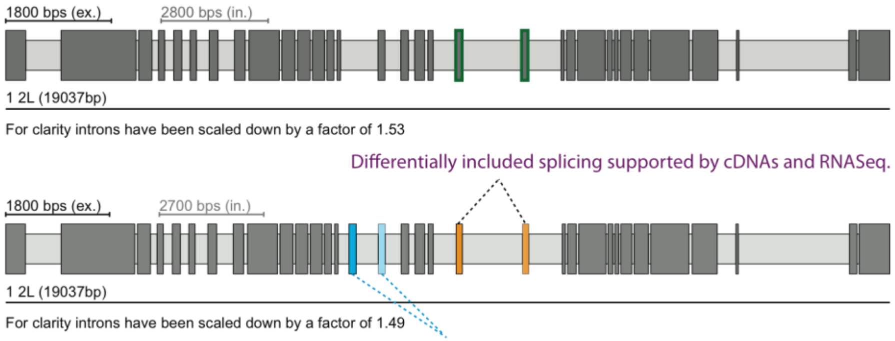


Gene: stai, stathmin, FBgn0051641
Polypeptide: stai-PB, FBpp0078828


200 bps (ex.)
3100 bps (in.)
1 2L (12916bp)
For clarity introns have been scaled down by a factor of 15.47

Differentially included exon.

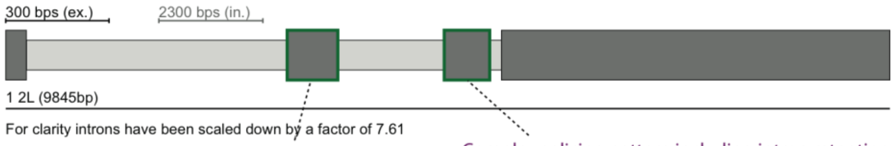
Gene: Ca-alpha1D, Ca^{2+}-channel protein α1 subunit D, FBgn0001991
Polypeptide: Ca-alpha1D-PC, FBpp0089047


1800 bps (ex.)
2800 bps (in.)
1 2L (19037bp)
For clarity introns have been scaled down by a factor of 1.53
1800 bps (ex.)
2700 bps (in.)
1 2L (19037bp)
For clarity introns have been scaled down by a factor of 1.49

Differentially included splicing supported by cDNAs and RNASeq.

Mutually exclusive spliced exons, exonA contains three 5' splice sites.

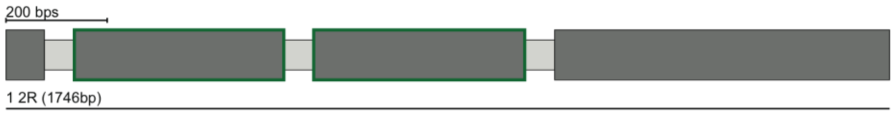
Gene: CG5674, FBgn0032656
Polypeptide: CG5674-PA, FBpp0080574


300 bps (ex.)
2300 bps (in.)
1 2L (9845bp)
For clarity introns have been scaled down by a factor of 7.61

Differentially included exon.

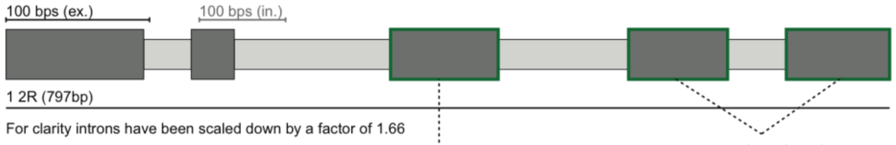
Complex splicing pattern including intron retention and several 5' splice sites.

Gene: CG10494, FBgn0034634
Polypeptide: CG10494-PA, CG10494-PA



Constitutively spliced exons.

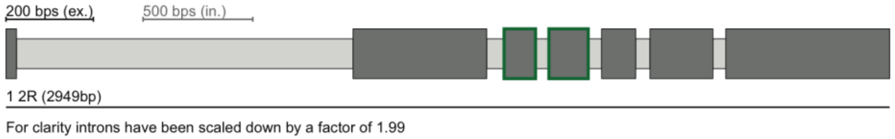
Gene: CG13428, FBgn0034515
Polypeptide: CG13428-PA, FBpp0085579


100 bps (ex.)
100 bps (in.)
1 2R (797bp)
For clarity introns have been scaled down by a factor of 1.66

Differentially included exon.

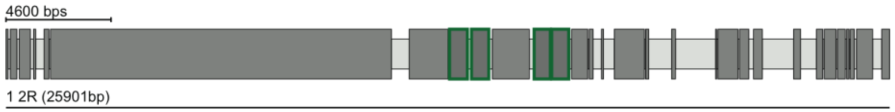
Constitutively spliced exons.

Gene: CG15615, FBgn0034159
Polypeptide: CG15615-PB, FBpp0289779


200 bps (ex.)
500 bps (in.)
1 2R (2949bp)
For clarity introns have been scaled down by a factor of 1.99

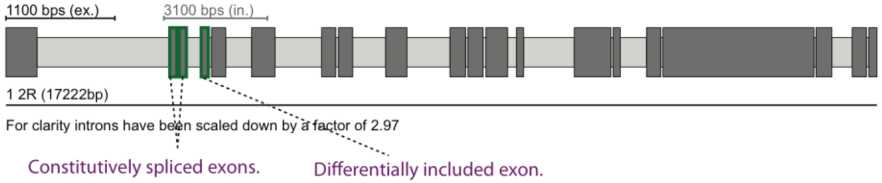
Constitutively spliced exons.
Intron has even been lost in other *Drosophila* species.

Gene: Strn-Mlck, Stretchin-Mlck, FBgn0013988
Polypeptide: Strn-Mlck-PD, FBpp0086409



Constitutively spliced exons.

Gene: sli, slit, FBgn0003425
Polypeptide: sli-PC, FBpp0086438

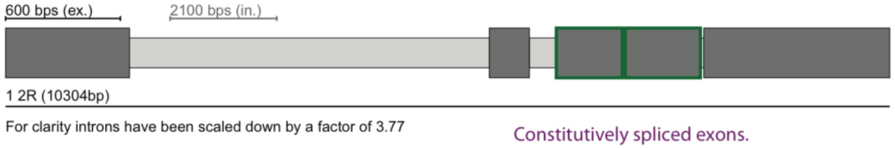


For clarity introns have been scaled down by a factor of 2.97

Constitutively spliced exons.

Differentially included exon.

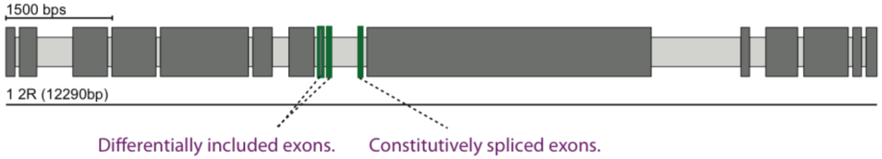
Gene: rgr, regular, FBgn0033310
Polypeptide: rgr-PA, FBpp0087772



For clarity introns have been scaled down by a factor of 3.77

Constitutively spliced exons.

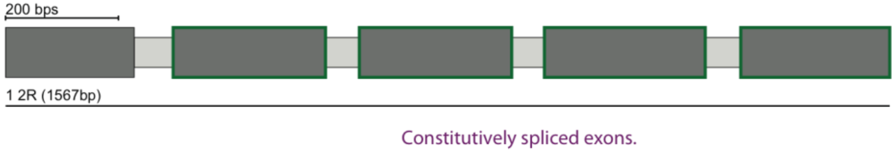
Gene: tou, toutatis, FBgn0033636
Polypeptide: tou-PA, FBpp0087193



Differentially included exons.

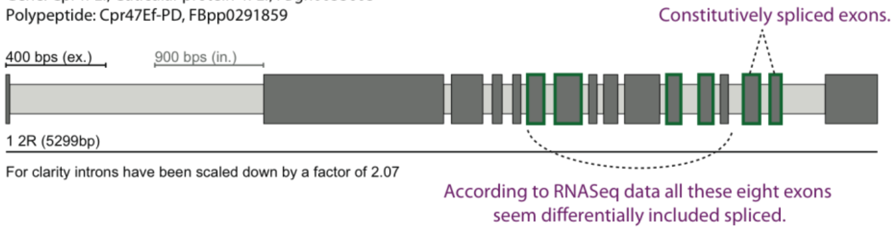
Constitutively spliced exons.

Gene: CG6357, FBgn0033875
Polypeptide: CG6357-PA, FBpp0086764



Constitutively spliced exons.

Gene: Cpr47Ef, Cuticular protein 47Ef, FBgn0033603
Polypeptide: Cpr47Ef-PD, FBpp0291859

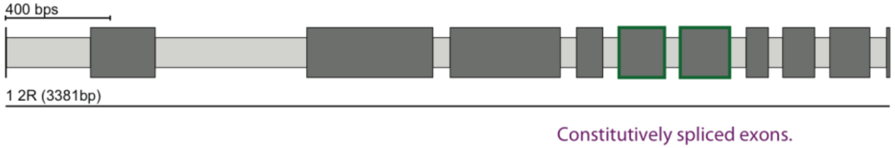


Constitutively spliced exons.

For clarity introns have been scaled down by a factor of 2.07

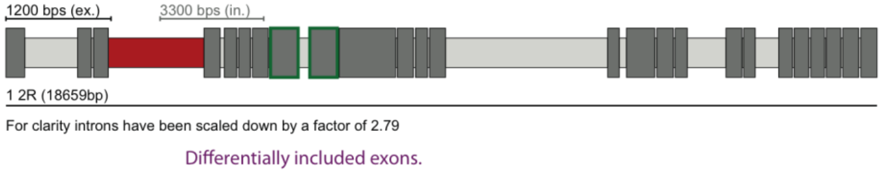
According to RNASeq data all these eight exons seem differentially included spliced.

Gene: Dek, FBgn0026533
Polypeptide: Dek-PA, FBpp0099855



Constitutively spliced exons.

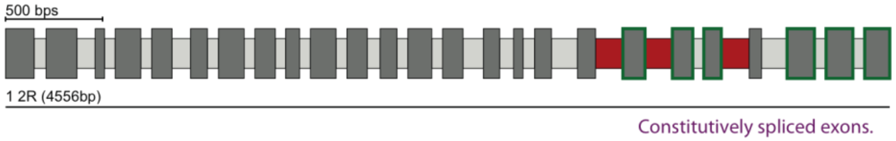
Gene: l(2)01289, lethal (2) 01289, FBgn0010482
Polypeptide: l(2)01289-PB, FBpp0085470



For clarity introns have been scaled down by a factor of 2.79

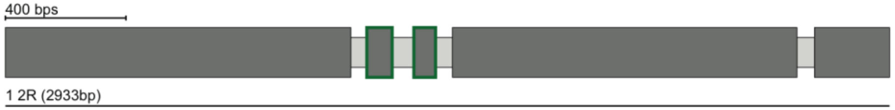
Differentially included exons.

Gene: CG30395, FBgn0050395
Polypeptide: CG30395-PB, FBpp0289463



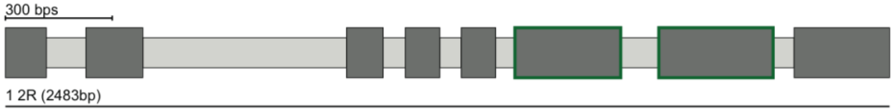
Constitutively spliced exons.

Gene: CG9861, FBgn0034844
Polypeptide: CG9861-PA, FBpp0071911



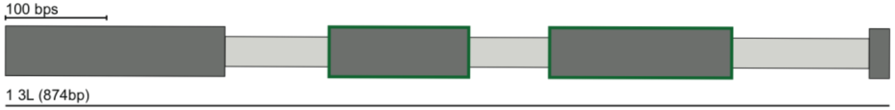
Constitutively spliced exons.

Gene: Mlp60A, Muscle LIM protein at 60A, FBgn0259209
Polypeptide: Mlp60A-PB, FBpp0288975



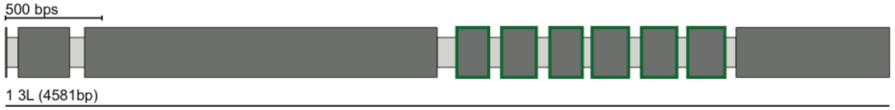
Constitutively spliced exons.

Gene: miple, FBgn0027111
Polypeptide: miple-PA, FBpp0072405



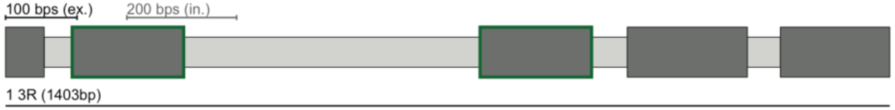
Constitutively spliced exons.

Gene: CG6947, FBgn0036233
Polypeptide: CG6947-PA, FBpp0075777



Constitutively spliced exons.

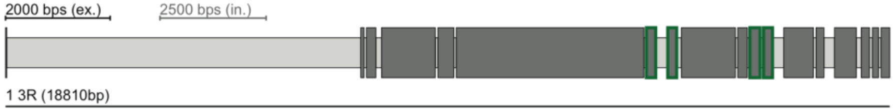
Gene: CG33483, FBgn0053483
Polypeptide: CG33483-PB, FBpp0292484



For clarity introns have been scaled down by a factor of 1.30

Constitutively spliced exons.

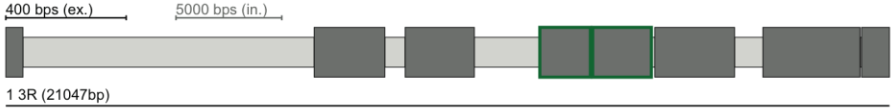
Gene: Ppn, Papilin, FBgn0003137
Polypeptide: Ppn-PE, FBpp0291051



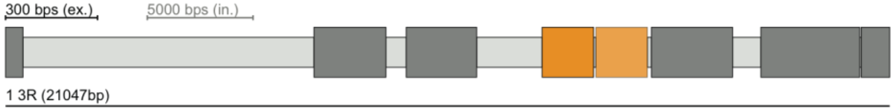
For clarity introns have been scaled down by a factor of 1.23

Differentially included exons.

Gene: betaTub97EF, β-Tubulin at 97EF, FBgn0003890
Polypeptide: βTub97EF-PA , FBpp0084630



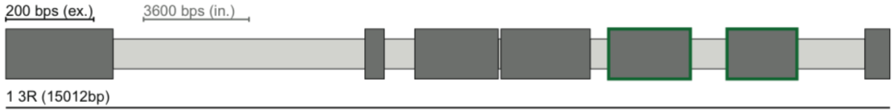
For clarity introns have been scaled down by a factor of 14.22



For clarity introns have been scaled down by a factor of 14.49

Mutually exclusive exons.
Misannotated as constitutive exons.

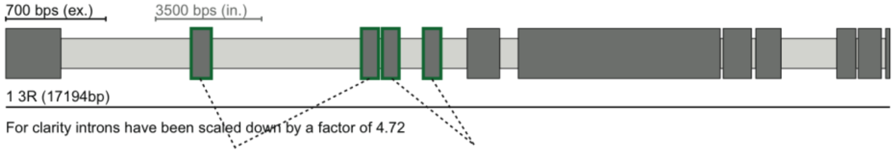
Gene: tau, FBgn0051057
Polypeptide: tau-PA, FBpp0084567



For clarity introns have been scaled down by a factor of 15.01

Differentially included exons.

Gene: LpR1, Lipophorin receptor 1, FBgn0066101
Polypeptide: LpR1-PK, FBpp0290685

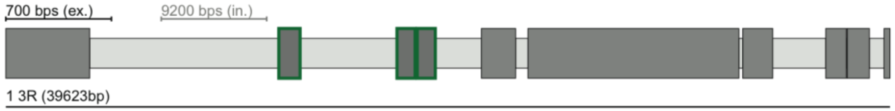


For clarity introns have been scaled down by a factor of 4.72

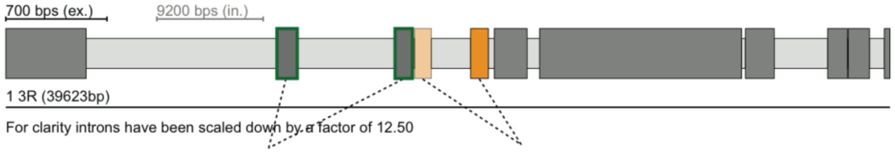
Constitutively spliced exons.

Differentially included exons.

Gene: LpR2, Lipophorin receptor 2, FBgn0051092
Polypeptide: LpR2-PA, FBpp0084301



For clarity introns have been scaled down by a factor of 13.16

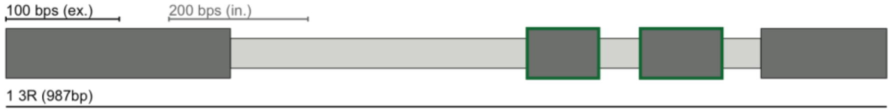


For clarity introns have been scaled down by a factor of 12.50

Constitutively spliced exons.

Mutually exclusive exons

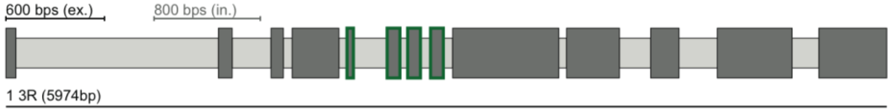
Gene: CG31406, FBgn0051406
Polypeptide: CG31406-PA, FBpp0081713



For clarity introns have been scaled down by a factor of 1.63

Constitutively spliced exons.

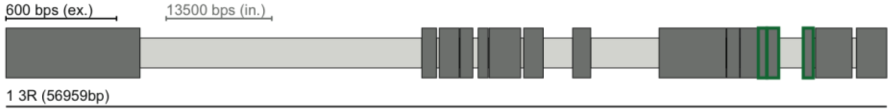
Gene: CG9297, FBgn0038181
Polypeptide: CG9297-PA, FBpp0082295



For clarity introns have been scaled down by a factor of 1.24

Constitutively spliced exons.

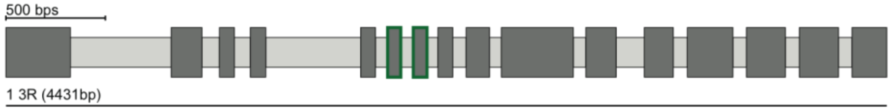
Gene: CG42342, FBgn0259244
Polypeptide: CG42342-PD, FBpp0289172



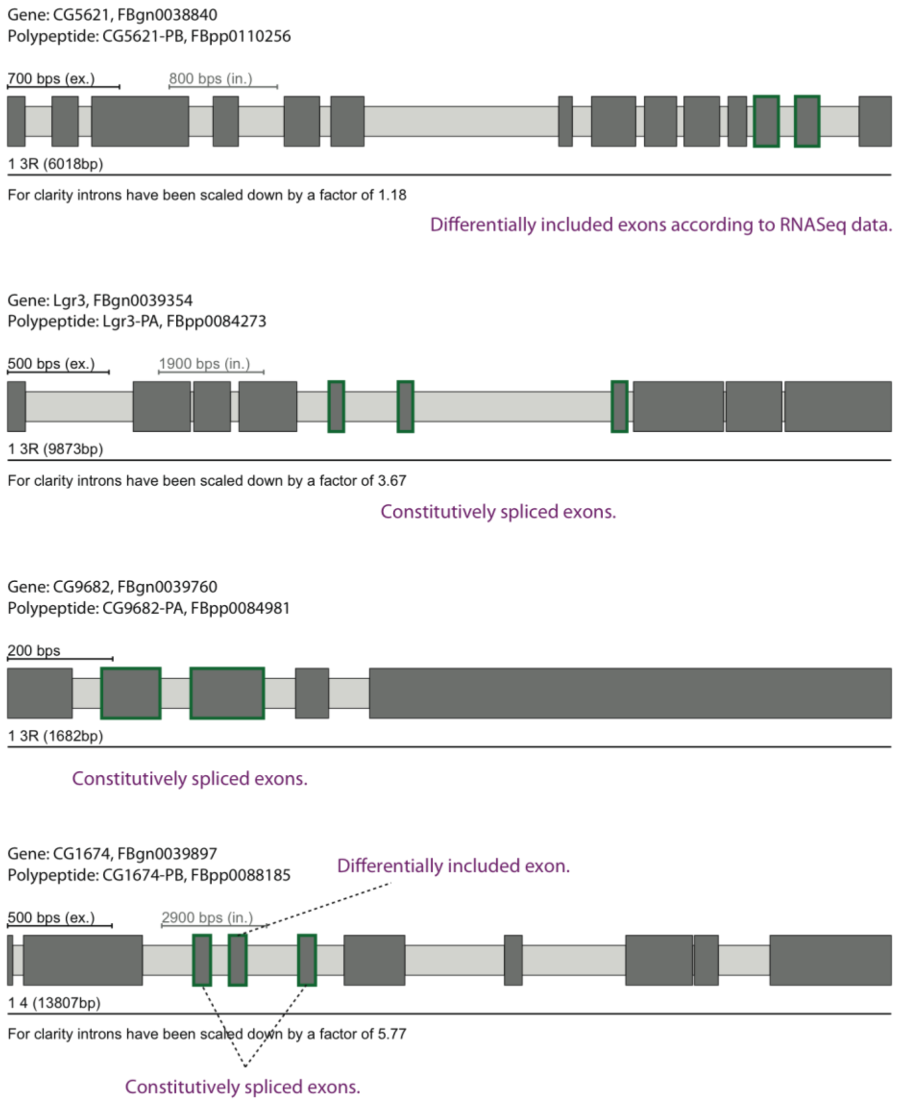
For clarity introns have been scaled down by a factor of 23.57

Constitutively spliced exons.

Gene: Fsh, Fsh-Tsh-like receptor, FBgn0016650
Polypeptide: Fsh-PA, FBpp0082933



Constitutively spliced exons.


Gene: CG5621, FBgn0038840
Polypeptide: CG5621-PB, FBpp0110256
700 bps (ex.)
800 bps (in.)
1 3R (6018bp)
For clarity introns have been scaled down by a factor of 1.18

Differentially included exons according to RNASeq data.

Gene: Lgr3, FBgn0039354
Polypeptide: Lgr3-PA, FBpp0084273
500 bps (ex.)
1900 bps (in.)
1 3R (9873bp)
For clarity introns have been scaled down by a factor of 3.67

Constitutively spliced exons.

Gene: CG9682, FBgn0039760
Polypeptide: CG9682-PA, FBpp0084981
200 bps
1 3R (1682bp)

Constitutively spliced exons.

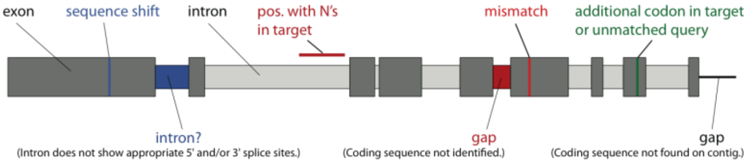
Gene: CG1674, FBgn0039897
Polypeptide: CG1674-PB, FBpp0088185
500 bps (ex.)
2900 bps (in.)
1 4 (13807bp)
For clarity introns have been scaled down by a factor of 5.77

Differentially included exon.

Constitutively spliced exons.

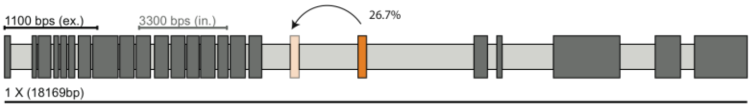
Supplementary Figure S21. List of genes containing constitutive exons matching the prediction parameters for mutually exclusive spliced exons (MXEs). Several of these exons are even annotated as MXEs in the latest Flybase release on RNA-Seq evidence, including a cluster of MXEs in the βTub97EF gene, the Lipophorin receptor 1 gene, and the nicotinic Acetylcholine Receptor α 30 D gene. Another gene is now split into two tandemly arrayed duplicates (CG10039 is now CG43773 and CG43774). The putative constitutive exons in 15 other genes are now annotated as differentially included or as other types of alternative splice forms. All transcripts are represented 5’ to 3’. The color coding is explained in the legend. Colored big bars represent MXEs. The darkest colored bar is the exon that was included in the query sequence, while the lighter colored bars represent identified MXEs. The higher the similarity between the candidate and the query exon the darker the color of the candidate (100% identity would result in the same color). The opacity of the colors of each alternative exon corresponds to the alignment score of the alternative exon to the original one. The green strokes mark constitutive exons that match our criteria for MXEs.

Legend


exon
sequence shift
intron
pos. with N's in target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

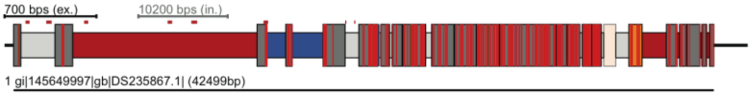
Gene: mmd, mind-meld, FBgn0259110
Polypeptide: mmd-PD, FBpp0288810

Exon A is annotated in r5.48.
Supported by RNA-Seq data.
Conserved in Anopheles gambiae, Pediculus humanus corporis, dana, dere, dgri, dmoj, dper, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exon B has an alternative splice site at the 3'-end.



For clarity introns have been scaled down by a factor of 3.17

Cross-species search in Pediculus humanus corporis

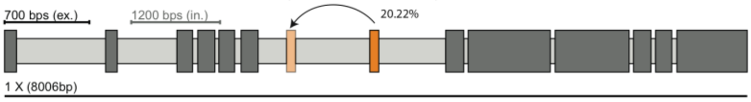


For clarity introns have been scaled down by a factor of 15.00

```
                   10        20        30
           ....|....|....|....|....|....|....|..
exonA      GTSDQNRISTLTMVIILTVIVKCVFISFATLAVCYR-
exonB      --ENYHGSNTVFLVGVLMSVVGFVFITFTLMALCYR-
AngExonA   -DHDQGKVSTLAMVIMLVVIVKCVFLCFALMAVCYR-
AngExonB   --ANYHGSNTVFLVGVLMSVVGGVFITFALMALCYRS
PdcExonA   -GGNNNNLSTLAMVFILVGVVKGVFICFTLMAVCYR-
PdcExonB   --ENYHSTNTGFLVGVLMSVVGGVFILFALMALCYR-
```

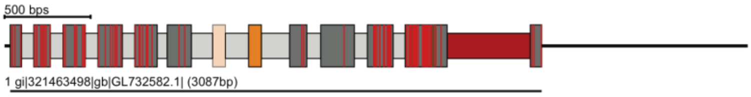
Gene: g, garnet, FBgn0001087
Polypeptide: g-PB, FBpp0073673

Exon A is annotaed in r5.48.
Supported by EST and RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.
RNA-Seq: Exon A has an alternative splice site at the 5'-end.



For clarity introns have been scaled down by a factor of 1.64

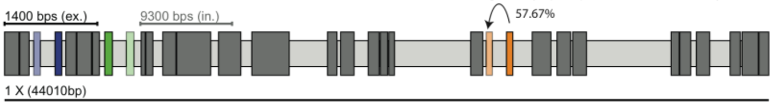
Cross-species search in Daphnia pulex



```
                   10        20
           ....|....|....|....|....|
exonA      FASLTTIEPALGRKLTQPLIEIISS
exonB      FGALTPLEPRLGKKLIEPLTNLIHS
DapExonA   LNAMTQCDSRLSKCISQPLIAIIKS
DapExonB   FGALTPLEPRLGKKLIEPLTNLIHS
```

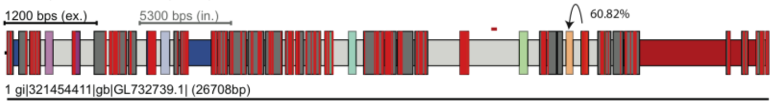
Gene: cac, cacophony, FBgn0263111
Polypeptide: cac-PA, FBpp0298319

Exon A is annotaed in r5.48.
Supported by RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex, Pediculus humanus corporis dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.



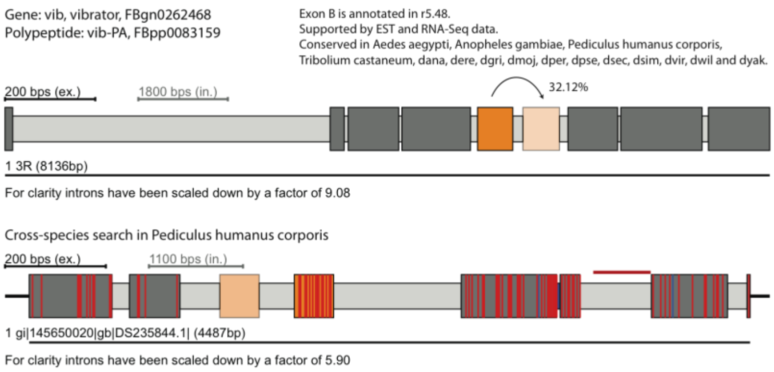
For clarity introns have been scaled down by a factor of 6.73

Cross-species search in Daphnia pulex

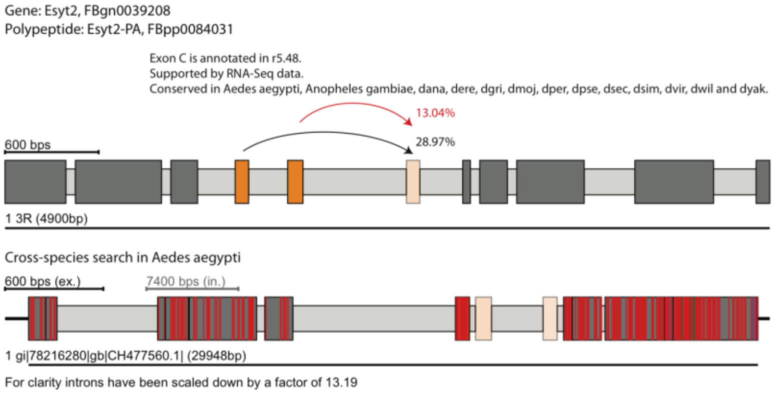


For clarity introns have been scaled down by a factor of 4.46

```
                   10        20        30
           ....|....|....|....|....|....|..
exonA      VFGNIRYDP-DTQLNRHNNFQSFSGGIMLLFR
exonB      VFGNIKLGTVENSITRHNNFQSFIQGVMLLFR
DapExonA   VFGNLHLDP-DSSVNRHNNFQSFIGGLLLLFR
DapExonB   VFGNILLEPGTTHIHRHNNFRSFIQGLMLLFR
```


Gene: vib, vibrator, FBgn0262468
Polypeptide: vib-PA, FBpp0083159
Exon B is annotated in r5.48.
Supported by EST and RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
200 bps (ex.)
1800 bps (in.)
32.12%
1 3R (8136bp)
For clarity introns have been scaled down by a factor of 9.08
Cross-species search in Pediculus humanus corporis
200 bps (ex.)
1100 bps (in.)
1 gi|145650020|gb|DS235844.1| (4487bp)
For clarity introns have been scaled down by a factor of 5.90

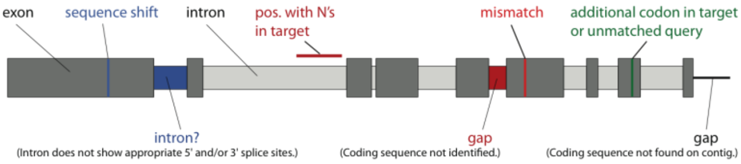
```
                  10        20
            ....|....|....|....|....|..
exonA       NPKFMKDAFKIIIDTLHV-GDAGDSEN
exonB       NPGYMDKNFKIDIYSQHIENDLGTVDN
PdcExonA    NPLYMKEKFHLTIESHHL-IDDGQNEN
PdcExonB    NPGYMKENFLIMIESFHI-NDSGYQEN
```


Gene: Esyt2, FBgn0039208
Polypeptide: Esyt2-PA, FBpp0084031
Exon C is annotated in r5.48.
Supported by RNA-Seq data.
Conserved in Aedes aegypti, Anopheles gambiae, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
13.04%
28.97%
600 bps
1 3R (4900bp)
Cross-species search in Aedes aegypti
600 bps (ex.)
7400 bps (in.)
1 gi|78216280|gb|CH477560.1| (29948bp)
For clarity introns have been scaled down by a factor of 13.19

```
                  10        20        30
            ....|....|....|....|....|....|...
exonA       ATVFIEMGQFVEIQLKDSDDS----KKDENLGR
exonB       ACIFTTIGHYIGFSLWDYDQTMPGVQSDDVLGR
exonC       AVVEVSQHAILVLRLFDWDRT----SDDESLGR
AeaExonA    AFIHAESGQQLQIVLNDKDAG----GDDELLGR
AeaExonB    AEVNATLGQEIELNLWDWDPGFPGVQNDDYLGR
AeaExonC    ACVDVSHQTLIGIKLFDWDRT----GDHDPLGR
```

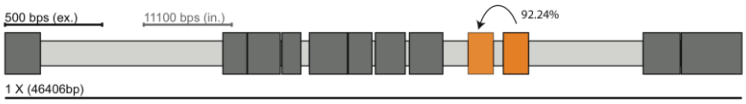
Supplementary Figure S22. Genes containing newly predicted mutually exclusive spliced exons (MXEs) which were not annotated in Flybase release 5.36, but are annotated in Flybase release 5.48. All transcripts are represented 5’ to 3’. The color coding is explained in the legend. Colored big bars represent MXEs. The darkest colored bar is the exon that was included in the query sequence, while the lighter colored bars represent identified MXEs. The higher the similarity between the candidate and the query exon the darker the color of the candidate (100% identity would result in the same color). The opacity of the colors of each alternative exon corresponds to the alignment score of the alternative exon to the original one.

Legend


exon
sequence shift
intron
pos. with N's in target
mismatch
additional codon in target or unmatched query
intron?
(Intron does not show appropriate 5' and/or 3' splice sites.)
gap
(Coding sequence not identified.)
gap
(Coding sequence not found on contig.)

Gene: Sh, Shaker, FBgn0003380
Polypeptide: Sh-PB, FBpp0088600

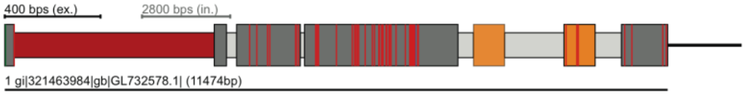
RNA-Seq supports 3'-end of exon A.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Daphnia pulex, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dpse, dsec, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 24.50

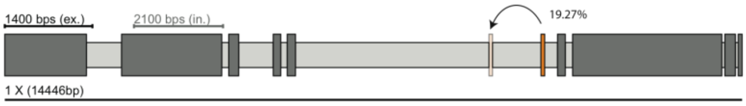
```
                   10        20        30        40
           ....|....|....|....|....|....|....|....|....
exonA      GVVLFSSAVYFAEAGSDSSFFKSIPDGFWWAVVTMTTVGYGDMR
exonB      GVVLFSSAVYFAEAGSENSFFKSIPDAFWWAVVTMTTVGYGDMT
DapExonA   GVILFSSAVYFAEAGSEYSYFKSIPDAFWWAVVTMTTVGYGDMR
DapExonB   GVILFSSAVYFAEAGSEVSHFKSIPDAFWWAVVTMTTVGYGDMT
```

Cross-species search in Daphnia pulex



For clarity introns have been scaled down by a factor of 7.60

Gene: fs(1)h, female sterile (1) homeotic, FBgn0004656
Polypeptide: fs(1)h-PB, FBpp0071074

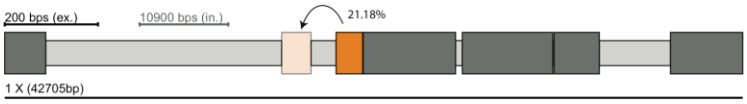


For clarity introns have been scaled down by a factor of 1.48

```
                10        20
         ....|....|....|....|
exonA    GRGNKKKGRKKSGR-RELRN
exonB    GDGDERPPRKKKSRDSNGSN
```

Gene: CG12541, FBgn0029930
Polypeptide: CG12541-PD, FBpp0289984

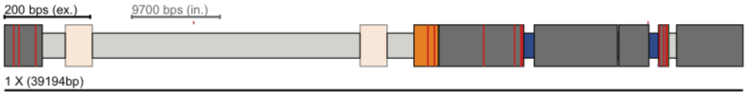
Conserved in dere and yak.



For clarity introns have been scaled down by a factor of 57.31

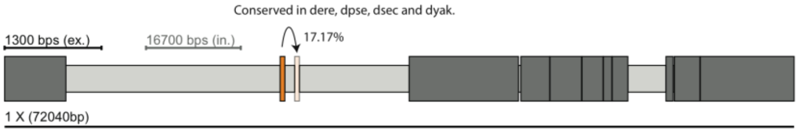
```
                     10        20
              ....|....|....|....|.
exonA         VRFMRSLMIAERASTKASLKY
exonB         V--VRLEVFAEEVTTAASLSE
dyakExonA     VIVAHTFAFEICVVTLAMCSS
dyakExonB     VRFMGSQVFAVRLSAKASLKY
dyakExonC     V--VRLEVFAEELTTAAALSE
```

Cross-species search in dyak



For clarity introns have been scaled down by a factor of 47.22

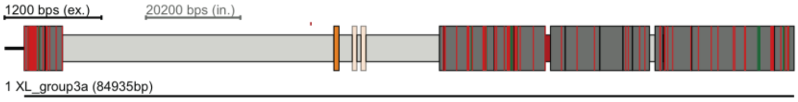
Gene (r5.36): CG42248
Polypeptide(r5.36): CG42248-PD, FBpp0288785
Gene (r5.48): CG43867, FBgn0264449
Polypeptide (r5.48): CG43867-PD, FBpp0304858



For clarity introns have been scaled down by a factor of 13.14

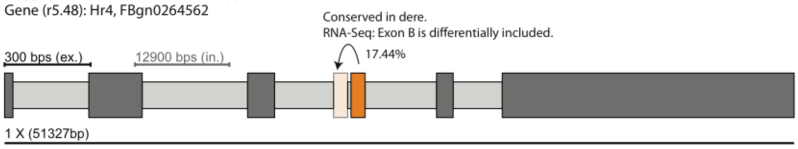
```
                    10        20
            ....|....|....|....|.
exonA       LTELEQRVIEAEERAEEAEDK
exonB       ASTWQLAVLESVENAGKSARK
dpseExonA   LTELEQRVIEAEERAEEAEDK
dpseExonB   LRGIERN--TARERESDVEER
dpseExonC   ATAREQRSCAACERESAARTC
```

Cross-species search in dpse



For clarity introns have been scaled down by a factor of 17.28

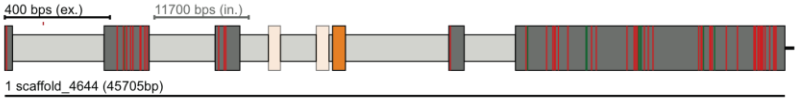
Gene (r5.36): CG3600
Polypeptide(r5.36): CG3600-PC, FBpp0288868
Gene (r5.48): Hr4, FBgn0264562



For clarity introns have been scaled down by a factor of 39.21

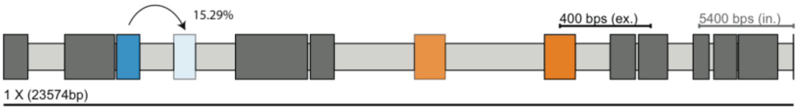
```
                     10
            ....|....|....|.
exonA       RARSAVGQRPVGGRFI
exonB       TCQAEEGQSSAGSHYT
dereExonA   RC-HELGERSSTSTWN
dereExonB   SERSAVGQRPVGGRFI
dereExonC   TCQAEEGQSSAGSHYT
```

Cross-species search in dpse



For clarity introns have been scaled down by a factor of 32.62

Gene: SK, small conductance calcium-activated potassium channel, FBgn0029761
Polypeptide: SK-PH, FBpp0289694

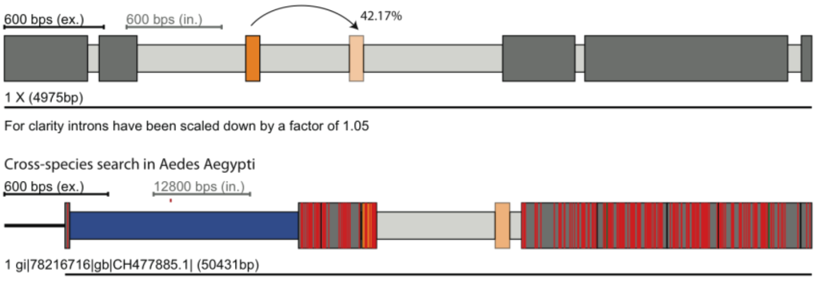


For clarity introns have been scaled down by a factor of 13.00

```
                    10        20        30
            ....|....|....|....|....|....|....
exonA       ASFYSTALKTLISVSTVILLGLIVAYHALEVQVR
exonB       KSHNSYSLHTICSLSLSII--IITPNQCLPPQIN
```

Gene: mys, myospheroid, FBgn0004657
Polypeptide: mys-PA, FBpp0071061

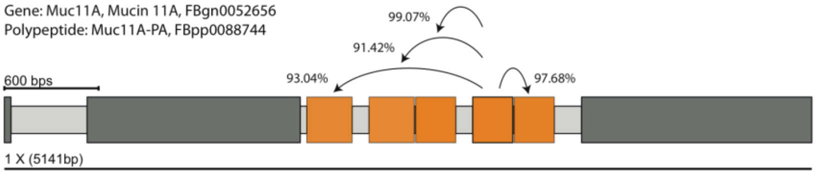
RNA-Seq data supports exon B.
Conserved in Aedes aegypti, Anopheles gambiae, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exons are differentially included.


600 bps (ex.)
600 bps (in.)
42.17%
1 X (4975bp)
For clarity introns have been scaled down by a factor of 1.05
Cross-species search in Aedes Aegypti
600 bps (ex.)
12800 bps (in.)
1 gi|78216716|gb|CH477885.1| (50431bp)

For clarity introns have been scaled down by a factor of 22.94

```
                    10        20
           ....|....|....|....|....|...
exonA      LEHPCENCKAPYGYQNHMPLNNNTESFS
exonB      LVEPCANCTATYGFHHQMVLDKNITQFT
AeaExonA   LEHPCDGCEAPYGYKNHMSLSVDTSRFS
AeaExonB   LREPCPQCAAPYGYHNLMPLSVDTHRFT
```

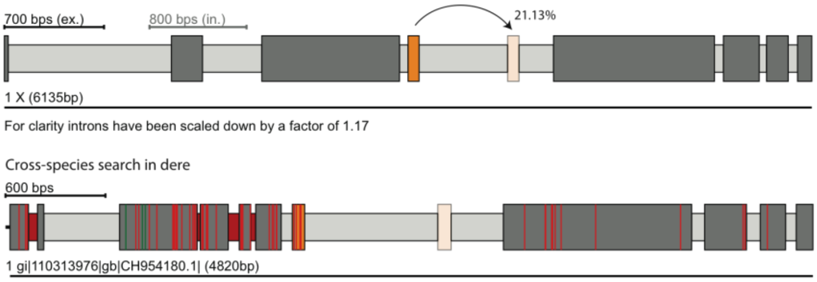
Gene: Muc11A, Mucin 11A, FBgn0052656
Polypeptide: Muc11A-PA, FBpp0088744



```
                10        20        30        40        50        60        70        80        90
       ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|..
exonA  ADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPAEESSAAPGSPADVTTAAPGAPADGSSAAP
exonB  AEGSSAAPGAPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonC  AEGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonD  ADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAP
exonE  AEGSSAAPGAPADVTTAAPGAPADGSSAAPGAPADGSSAAPGSPADVTTAAPGAPADGSSAAPGAPADVTTAAPGAPADGSSAAP
```

Gene: Top1, Topoisomerase 1, FBgn0004924
Polypeptide: Top1-PA, FBpp0073822

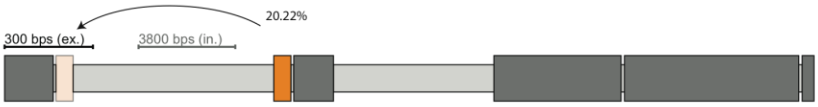
RNA-Seq data supports 3'-end of exon B.
Conserved in dere, dsec, dsim and dyak.
RNA-Seq: Exon A has an alternative splice site at 5'-end.


700 bps (ex.)
800 bps (in.)
21.13%
1 X (6135bp)
For clarity introns have been scaled down by a factor of 1.17
Cross-species search in dere
600 bps
1 gi|110313976|gb|CH954180.1| (4820bp)

```
                      10        20
             ....|....|....|....|....|..
exonA        EP-EPAVSPGKRQKAKAKVEEEEVWRW
exonB        RMLAVATVAGKRRRVRRKSVQEEQIRW
dereExonA    EP-E-VVSPTKRQKAKVKEEEEEVWRW
dereExonB    RMVAKVTNDGKKRRVRRKSVQEEQVRW
```

Gene: mol, moladietz, FBgn0086711
Polypeptide: mol-PA , FBpp0080238

Conserved in Aedes aegypti, dsec, dsim and dwil.

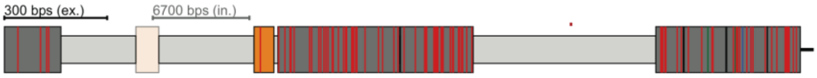


1 2L (16153bp)

For clarity introns have been scaled down by a factor of 11.57

```
                 10        20
             ....|....|....|....|..
exonA        KARWFKLINLYYFKNATLL
exonB        KFTTFSTVTLSLFVGLVIL
AeaExonA     QFNWFSTRSVVRFSSLVHLKIQ
AeaExonB     KFTTFTTVTLSLFVGLVIL
```

Cross-species search in Aedes aegypti

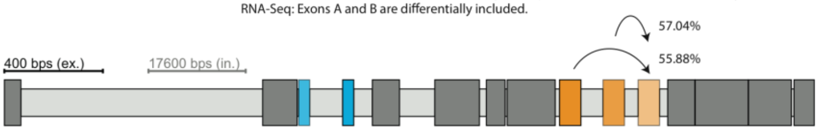


1 gi|78216149|gb|CH477448.1| (26216bp)

For clarity introns have been scaled down by a factor of 23.76

Gene: nAcRalpha-30D, nicotinic acetylcholine receptor alpha 30D, FBgn0032151

Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsim, dvir, dwil and dyak.
RNA-Seq: Exons A and B are differentially included.

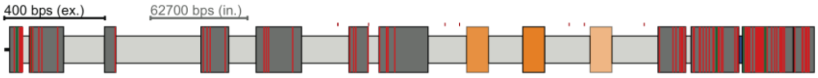


1 2L (72488bp)

For clarity introns have been scaled down by a factor of 44.59

```
                10        20
            ....|....|....|....|....|....
exonA       GVTILLSLTVFLNLVAESMPTTSDAVPLI
exonB       GVTILLSLTVFLNLVAETLPQVSDAIPLL
exonC       GVTILLSQTVFSLLVGNVITKTSEAVPLL
AmExonA     GVTILLSLTVFLNLVAESMPTTSDAVPLI
AmExonB     GVTILLSLTVFLNLVAETLPQVSDAIPLL
AmExonC     GVTILLSQTVFSLLVNHVLTRTSEAVPLI
```

Cross-species search in Apis mellifera

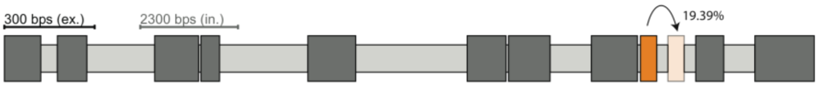


1 gi|318087635|gb|CM000055.5| (246068bp)

For clarity introns have been scaled down by a factor of 162.56

Gene: CG14010, FBgn0031725
Polypeptide: CG14010-PB , FBpp0292059

Conserved in dere, dwil and dyak.

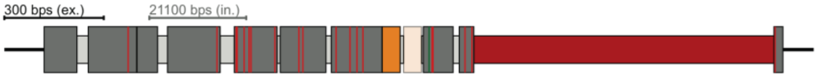


1 2L (10518bp)

For clarity introns have been scaled down by a factor of 7.08

```
                  10
              ....|....|....|...
exonA         RIPPNAVNYVENFEARHK
exonB         RIMGNIKLVANEWKARKK
dwilExonA     RIPPNAVNYVENFEARHK
dwilExonB     KAANISLIFVFVYQTRHK
```

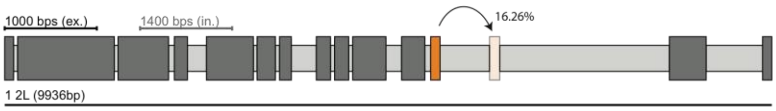
Cross-species search in dwil



1 scf2_1100000004521 (78736bp)

For clarity introns have been scaled down by a factor of 72.18

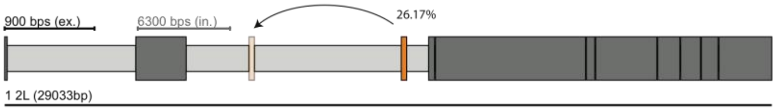
Gene: tim, timeless, FBgn0014396
Polypeptide: tim-PB , FBpp0077256



For clarity introns have been scaled down by a factor of 1.40

```
                 10        20        30
       ....|....|....|....|....|....|....|..
exonA  Y--TPDPTP-PVPNWLQLVMRSKCNHRTGPSGDPSDC
exonB  FGPTPSPTPSPTPSTSQDPTRSDAAHPLAELAAPSIF
```

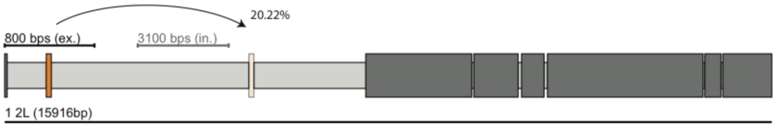
Gene: IA-2, IA-2 ortholog, FBgn0031294
Polypeptide: IA-2-PC , FBpp0290630



For clarity introns have been scaled down by a factor of 6.81

```
                 10
       ....|....|....|....
exonA  ATEIIFLLC-PYSHVCCFD
exonB  GCQFVRTLCIPHSEV-CYD
```

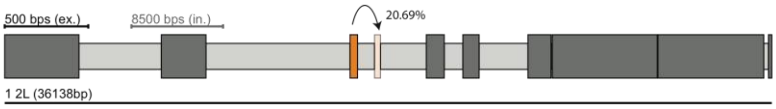
Gene: ush, u-shaped, FBgn0003963
Polypeptide: ush-PA, FBpp0077723



For clarity introns have been scaled down by a factor of 3.82

```
                 10
       ....|....|....|
exonA  GDCSDTAEEMTVDSR
exonB  GWTTETVEVHIIELQ
```

Gene: CG32982, FBgn0052982
Polypeptide: CG32982-PE, FBpp0290262

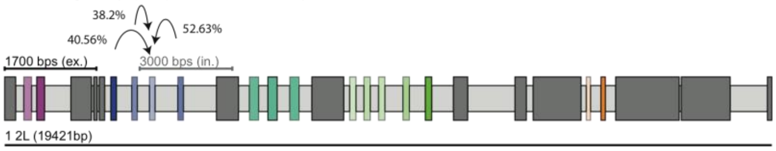


For clarity introns have been scaled down by a factor of 15.52

```
                 10
       ....|....|....|
exonA  VSCNKQTNWLNFKQD
exonB  IEC---TMWLDRRES
```

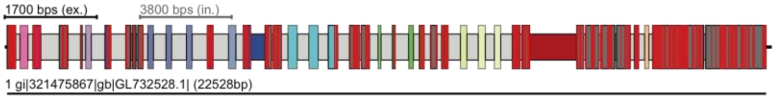
Gene: Mhc, Myosin heavy chain, FBgn0264695
Polypeptide: Mhc-PA, FBpp0080453

RNA-Seq data supports 3'-end of exon C.
Verified by literature.
Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex, Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 1.75

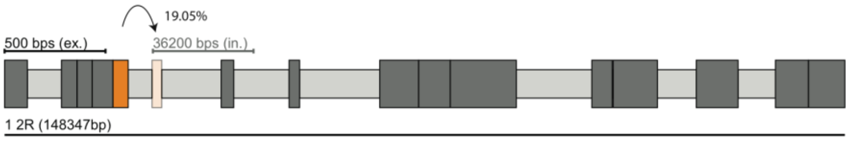
Cross-species search in Daphnia pulex



For clarity introns have been scaled down by a factor of 2.25

```
                    10        20        30
           ....|....|....|....|....|....|....|
exonA      DICLLTDNIYDYHIVSQGKVTVASIDDAEEFSLTD
exonB      EYCLLSNNIYDYRIVSQGKTTIPSVNDGEEWVAVD
exonC      EMVFLGQHIGDYPGICQGKTRIPGVNDGEEFELTD
exonD      EMCFLSDNIYDYYNVSQGKVTVPNMDDGEEFQLAD
DapExonA   ADCCLVDDIYQYNFVSQGKITIPSMDDSEEMALTD
DapExonB   ADCSLVDDIYTYNFVSQGKITIPSMDDSEEMGLTN
DapExonC   ADCRLVDDIYTYNYVSQGKITIPSMDDNEEMGLTD
DapExonD   AMCSLSDNIYDYPFVSQGKVTVPSIDDSEEMQMAD
```

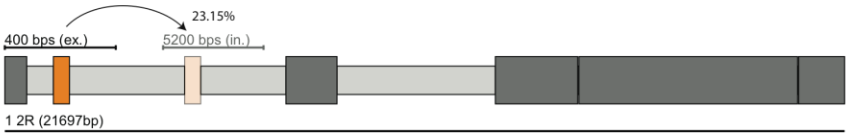
Gene: Gprk1, G protein-coupled receptor kinase 1, FBgn0260798
Polypeptide: Gprk1-PA, FBpp0110413



For clarity introns have been scaled down by a factor of 72.60

```
               10        20
       ....|....|....|....|....|
exonA  EYSKHAVASVQKYLLKNEVPVDLFE
exonB  ---------CKIFLLKNEVLVDLFE
```

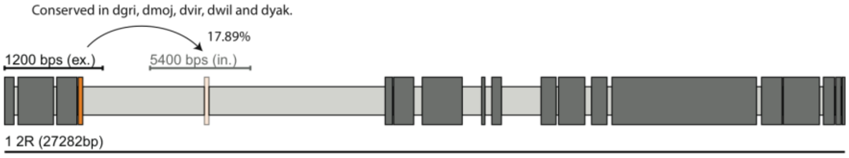
Gene: CG30438, FBgn0050438
Polypeptide: CG30438-PB, FBpp0085404



For clarity introns have been scaled down by a factor of 14.36

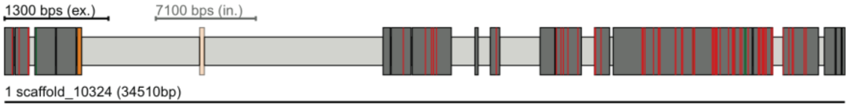
```
               10        20
       ....|....|....|....|
exonA  GGTKSHKIPFWELAKGLISR
exonB  GGLPEETTRKWRVQKGQWSQ
```

Gene: brp, bruchpilot, FBgn0259246
Polypeptide: brp-PD, FBpp0289193



For clarity introns have been scaled down by a factor of 4.40

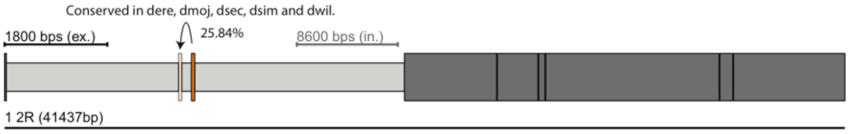
Cross-species search in dvir



For clarity introns have been scaled down by a factor of 5.67

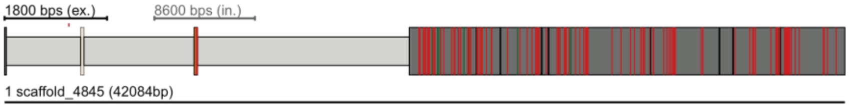
```
                    10        20
           ....|....|....|....|..
exonA      G--KEEERQMFQQMQAMA-QKQ
exonB      ---EQEQNRTFDSIQKSISQKA
dvirExonA  G--KEEERQMFQQMQAMA-QKQ
dvirExonB  GVKREKERRSRRQMQPCA--KQ
```

Gene: shn, schnurri, FBgn0003396
Polypeptide: shn-PD, FBpp0089118



For clarity introns have been scaled down by a factor of 4.86

Cross-species search in dvir

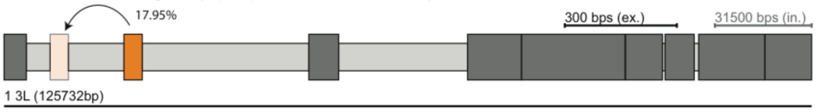


For clarity introns have been scaled down by a factor of 4.87

```
                    10        20
           ....|....|....|....|
exonA      KTTIVIKC-SKWVTSRHQEK
exonB      KSTVN-SRKSALETAREKTK
dereExonA  KQQIKATHK-ANRNKTQKIK
dereExonB  KSTVN-SRKSALESVREKPK
dmojExonA  KSTVN-SRKNTLESTREKLK
dmojExonB  KQQQRLSKKKCLSSALESSK
```

Gene: bru-3, bruno-3, FBgn0264001
Polypeptide: bru-3-PB, FBpp0303379

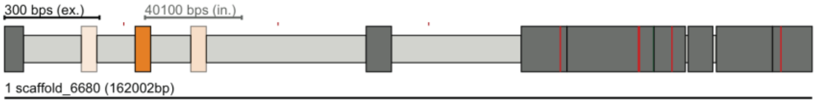
Conserved in dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.



For clarity introns have been scaled down by a factor of 121.65

```
                   10
            ....|....|....|.
exonA       IHKAGHSKPGNSSSFV
exonB       MNRALQLKPAENESRS
dmojExonA   SSQVLSVKCCSNIIES
dmojExonB   MNRALQLKPAENESRS
dmojExonC   MRAALDVLPISSLNSS
```

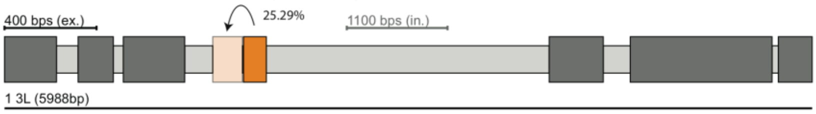
Cross-species search in dvir



For clarity introns have been scaled down by a factor of 131.27

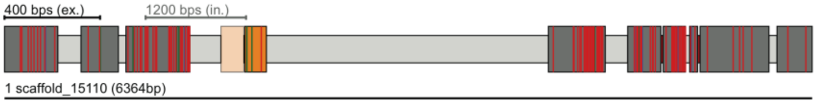
Gene: ect, ectodermal, FBgn0000451
Polypeptide: ect-PA, FBpp0076034

Conserved in dana, dere, dgri, dsim and dyak.



For clarity introns have been scaled down by a factor of 2.51

Cross-species search in dgri

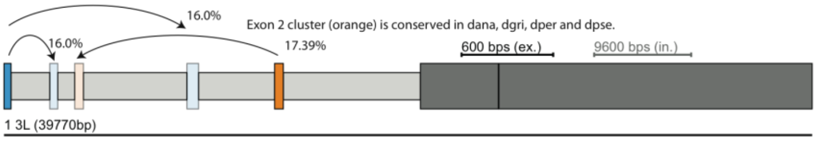


For clarity introns have been scaled down by a factor of 2.86

```
                     10        20        30        40
              ....|....|....|....|....|....|....|....|....
exonA         GAVAPGNVAAGADDTDNDDDDYDEDDETDDDDDDDIDDGVDEI-
exonB         G-------IAGDDDEEADDDD---DDDDDDIIGDDIIEARREA-
dgriExonA     ---------ASDDDDYDDEDD---EYDDDDYSDDDIDEGVDEIT
dgriExonB     ----------GDDEEADDDDD---DDDDDDIIGDDVIEARREA-
```

Gene: CG7991, FBgn0035260
Polypeptide: CG7991-PB, FBpp0292221

Exon 1 cluster (blue) is conserved in dere, dsec, dsim, dvir and dyak.

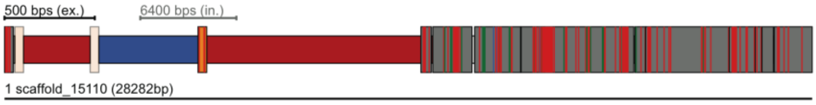


Exon 2 cluster (orange) is conserved in dana, dgri, dper and dpse.

For clarity introns have been scaled down by a factor of 15.12

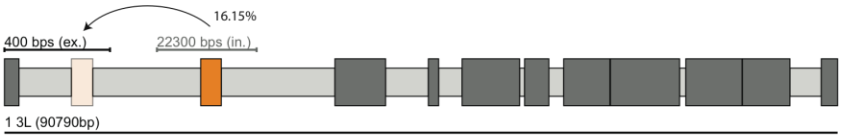
```
                       10        20
               ....|....|....|....|....|.
exon1A         MLF----------WKRRLQRSSSSS-
exon1B         MII---------DWRANMRKERSLST
exon1C         MAFILWRQGVASCWKNRRHKSSSLR-
exon2A         KRMQKRLFFSTFCHNMAKK
exon2B         SAIQERRFFGILRSAKRKD
dgriExon2A     ---QQQTEAVVATVGRRKM
dgriExon2B     ---QTIRFIDFRFAALRNN
dgriExon2C     ---QQRRFFGILRAGKRKD
```

Cross-species search in dgri



For clarity introns have been scaled down by a factor of 12.03

Gene: Eip63E, Ecdysone-induced protein 63E, FBgn0264001
Polypeptide: Eip63E-PD, FBpp0072990

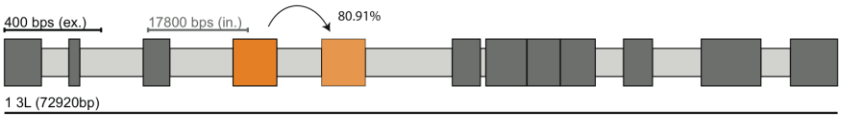


For clarity introns have been scaled down by a factor of 59.04

```
                10        20
        ....|....|....|....|....|..
exonA   GSTKIEKSDLKIQVIYMQMSNKYGQRG
exonB   GVTMREKKGGALQKLKKRLSHSFG-RL
```

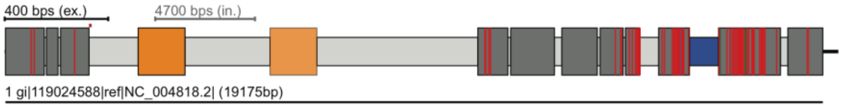
Gene: nAcRalpha-80B, nicotinic Acetylcholine Receptor alpha 80B, FBgn0037212
Polypeptide: nAcRalpha-80B-PC, FBpp0289395

Conserved in Aedes aegypti, Anopheles gambiae, Apis mellifera, Atta cephalotes, Daphnia pulex
Pediculus humanus corporis, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dsec and dwil.



For clarity introns have been scaled down by a factor of 43.15

Cross-species search in Anopheles gambiae



For clarity introns have been scaled down by a factor of 12.19

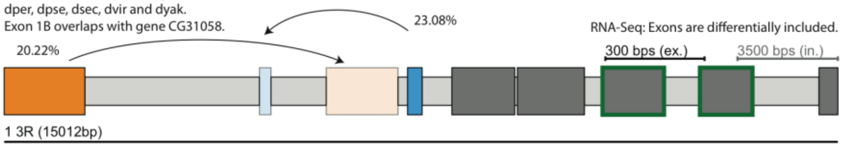
```
                   10        20        30        40        50        60
           ....|....|....|....|....|....|....|....|....|....|....|....|
exonA      ADGNFEVTLATKATIYSEGLVEWKPPAIYKSSCEIDVEYFPFDEQTCVLKFGSWTYDGFK
exonB      ADGHYEVTLMTKAIVYNNGLVIWQPPAVYKSSCSIDVEYFPYDVQTCILKLGSWTYDGFK
AngExonA   ADGNFEVTLATKATIYSEGLVEWKPPAIYKSSCEIDVEYFPFDEQTCVLKFGSWTYDGFK
AngExonB   ADGHYEVTLMTKATVYNNGMVIWQPPAVYKSSCSIDVEYFPYDVQTCVLKLGSWTYDGFK
```

Gene: tau, FBgn0051057
Polypeptide: tau-PA, FBpp0084567

Conserved in dana, dere, dgri,
dper, dpse, dsec, dvir and dyak.
Exon 1B overlaps with gene CG31058.

Conserved in dere, dper, dpse, dsec and dyak.

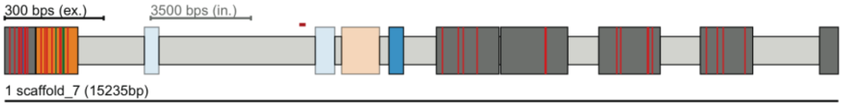
RNA-Seq: Exons are differentially included.



For clarity introns have been scaled down by a factor of 11.53

```
                   10        20
             ....|....|....|....|
exonA        VGDSDS---ESAQVA
exonB        EGDNDSGVDESTQEK
dperExonA    ELSNGFGPSQSQSQA
dperExonB    EQSDNGSAADEAGNAATAES
dperExonC    EGDNDSGVDESTQEK
```

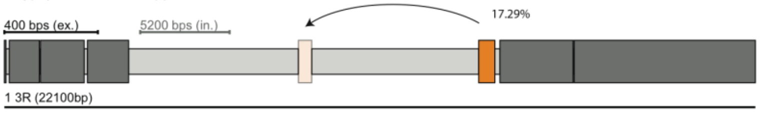
Cross-species search in dper



For clarity introns have been scaled down by a factor of 11.52

Gene: twin, FBgn0011725
Polypeptide: twin-PA, FBpp0083951

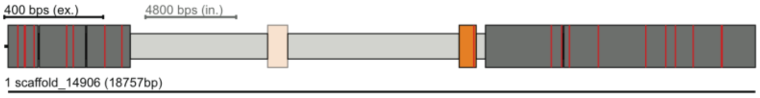
Conserved in dgri, dmoj and dper.



For clarity introns have been scaled down by a factor of 13.60

```
                10        20
       ....|....|....|....|....|..
exonA      FFQAP----PPL--WVP--ENNPSEPW
exonB      FTVNP----PPQRPWLPLAKPNKTRPA
dgriExonA  FAVTPSLPTPPPLPSSPLSQAGHNRRP
dgriExonB  FTVNP----PPQRPWLPLAKPNKSRPA
dmojExonA  FFYLPVRSTRPIA-QLQMRKPNKSRLH
dmojExonB  FTVNP----PPQRPWLPLAKPNKSRPA
```

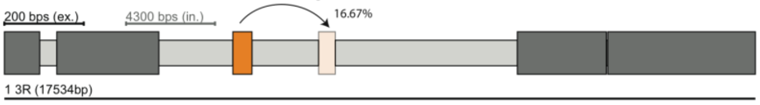
Cross-species search in dgri



For clarity introns have been scaled down by a factor of 13.08

Gene: abd-A, abdominal A, FBgn0000014
Polypeptide: abd-A-PA, FBpp0082828

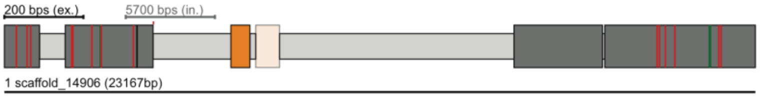
Conserved in dgri.



For clarity introns have been scaled down by a factor of 19.19

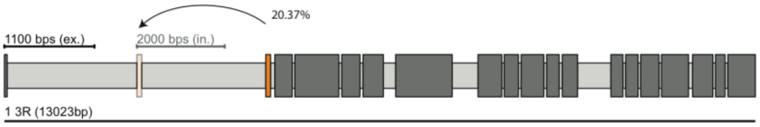
```
                10        20
           ....|....|....|....|
exonA      D----WMGSPFERVVCGDFN
exonB      D----W--RDFSSVVVGRQT
dgriExonA  D----WMGSPFERVVCGDFN
dgriExonB  DTSGNWQPMPFSSLIVDPCN
```

Cross-species search in dgri



For clarity introns have been scaled down by a factor of 25.25

Gene: CG14741, FBgn0037989
Polypeptide: CG14741-PC, FBpp0297858

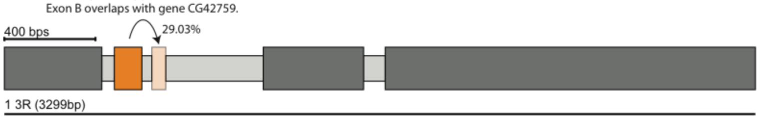


For clarity introns have been scaled down by a factor of 1.86

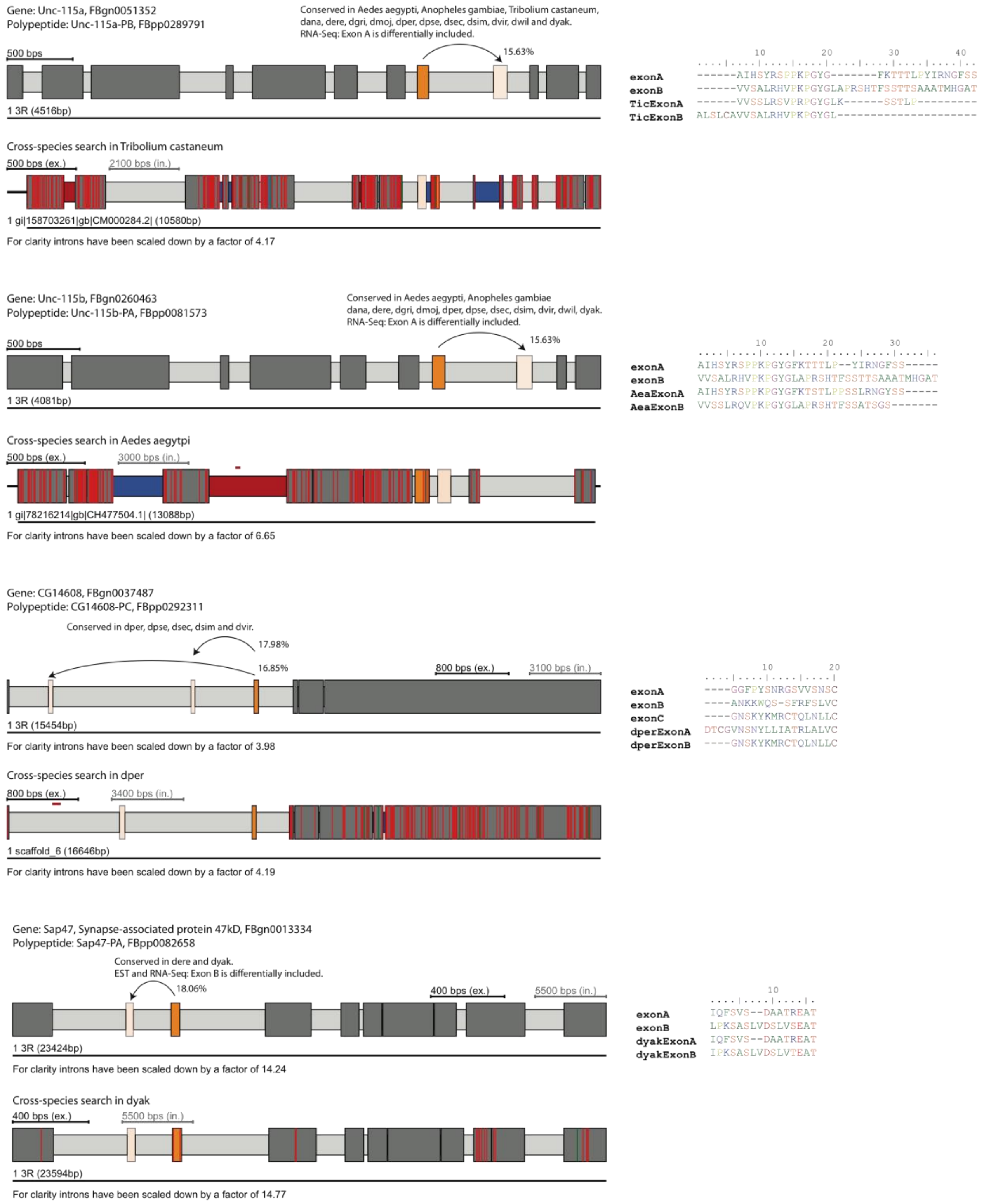
```
                10        20
       ....|....|....|....|
exonA  EGEGPKRDHDDAFGTWHRKH
exonB  ENERRIRANDKEFNAQFKYH
```

Gene: CG6241, FBgn0037792
Polypeptide: CG6241-PA, FBpp0081663

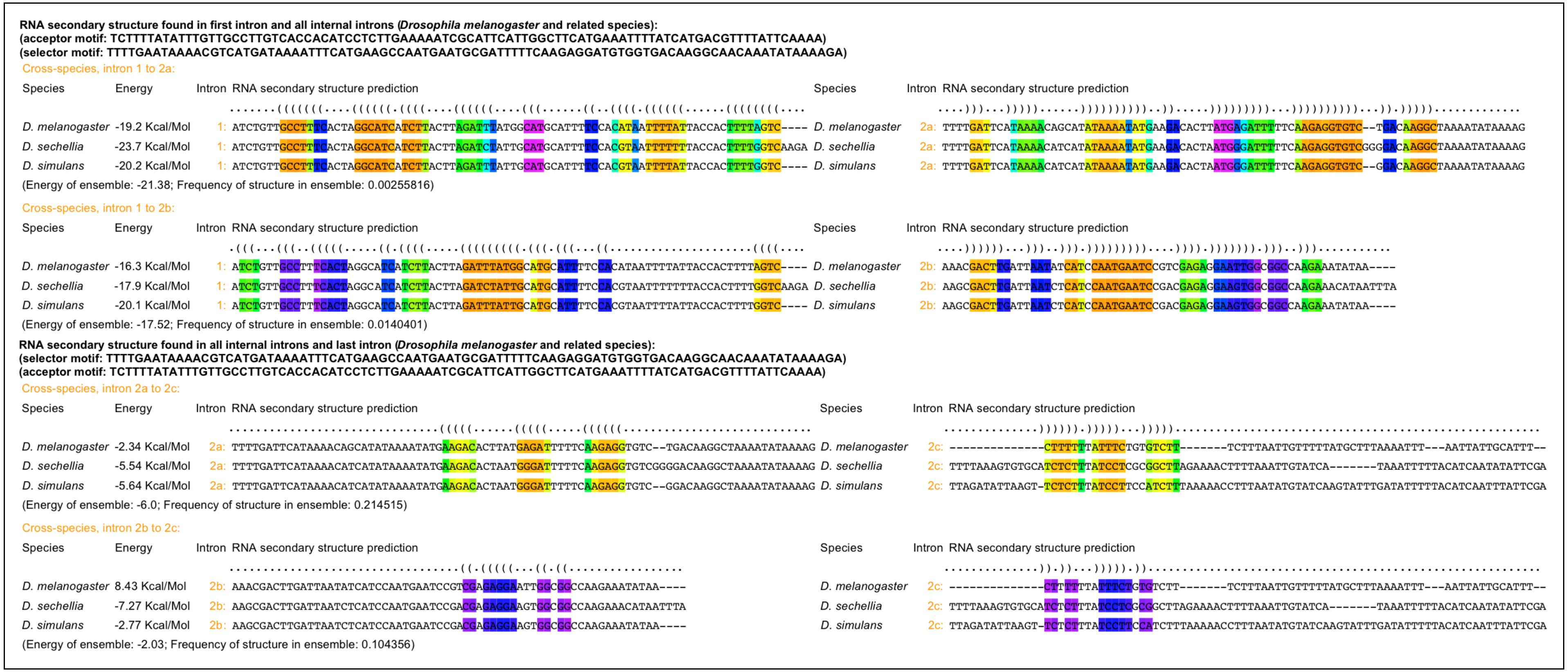
RNA-Seq data supports 3'-end of exon B.
Exon B overlaps with gene CG42759.



```
                10        20        30        40
       ....|....|....|....|....|....|....|....|
exonA  DARIIYNHKTFKKGKKGKKSTLTGDPNDERAKFRLWNRTK
exonB  --------------------TLTGDPNDERAKFRLWNRTK
```


Gene: Unc-115a, FBgn0051352
Polypeptide: Unc-115a-PB, FBpp0289791
Conserved in Aedes aegypti, Anopheles gambiae, Tribolium castaneum, dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil and dyak.
RNA-Seq: Exon A is differentially included.
500 bps
15.63%
1 3R (4516bp)
Cross-species search in Tribolium castaneum
500 bps (ex.)
2100 bps (in.)
1 gi|158703261|gb|CM000284.2| (10580bp)
For clarity introns have been scaled down by a factor of 4.17
exonA ------AIHSYRSPPKPGYG-------FKTTTLPYIRNGFSS
exonB ------VVSALRHVPKPGYGLAPRSHTFSSTTSAAATMHGAT
TicExonA ------VVSSLRSVPRPGYGLK------SSTLP---------
TicExonB ALSLCAVVSALRHVPKPGYGL---------------------
Gene: Unc-115b, FBgn0260463
Polypeptide: Unc-115b-PA, FBpp0081573
Conserved in Aedes aegypti, Anopheles gambiae dana, dere, dgri, dmoj, dper, dpse, dsec, dsim, dvir, dwil, dyak.
RNA-Seq: Exon A is differentially included.
500 bps
15.63%
1 3R (4081bp)
Cross-species search in Aedes aegytpi
500 bps (ex.)
3000 bps (in.)
1 gi|78216214|gb|CH477504.1| (13088bp)
For clarity introns have been scaled down by a factor of 6.65
exonA AIHSYRSPPKPGYGFKTTTLP--YIRNGFSS-----
exonB VVSALRHVPKPGYGLAPRSHTFSSTTSAAATMHGAT
AeaExonA AIHSYRSPPKPGYGFKTSTLPPSSLRNGYSS-----
AeaExonB VVSSLRQVPKPGYGLAPRSHTFSSATSGS-------
Gene: CG14608, FBgn0037487
Polypeptide: CG14608-PC, FBpp0292311
Conserved in dper, dpse, dsec, dsim and dvir.
17.98%
16.85%
800 bps (ex.)
3100 bps (in.)
1 3R (15454bp)
For clarity introns have been scaled down by a factor of 3.98
Cross-species search in dper
800 bps (ex.)
3400 bps (in.)
1 scaffold_6 (16646bp)
For clarity introns have been scaled down by a factor of 4.19
exonA ----GGFPYSNRGSVVSNSC
exonB ----ANKKWQS-SFRFSLVC
exonC ----GNSKYKMRCTQLNLLC
dperExonA DTCGVNSNYLLIATRLALVC
dperExonB ----GNSKYKMRCTQLNLLC
Gene: Sap47, Synapse-associated protein 47kD, FBgn0013334
Polypeptide: Sap47-PA, FBpp0082658
Conserved in dere and dyak.
EST and RNA-Seq: Exon B is differentially included.
18.06%
400 bps (ex.)
5500 bps (in.)
1 3R (23424bp)
For clarity introns have been scaled down by a factor of 14.24
Cross-species search in dyak
400 bps (ex.)
5500 bps (in.)
1 3R (23594bp)
For clarity introns have been scaled down by a factor of 14.77
exonA IQFSVS--DAATREAT
exonB LPKSASLVDSLVSEAT
dyakExonA IQFSVS--DAATREAT
dyakExonB IPKSASLVDSLVTEAT

Supplementary Figure S23. Genes containing newly predicted mutually exclusive spliced exons (MXEs) which were not annotated in Flybase release 5.36 nor in release 5.48. All transcripts are represented 5' to 3'. The color coding is explained in the legend. Colored big bars represent MXEs. The darkest colored bar is the exon that was included in the query sequence, while the lighter colored bars represent identified MXEs. The higher the similarity between the candidate and the query exon the darker the color of the candidate (100% identity would result in the same color). The opacity of the colors of each alternative exon corresponds to the alignment score of the alternative exon to the original one.



RNA secondary structure found in first intron and all internal introns (*Drosophila melanogaster* and related species):
(acceptor motif: TCTTTTATATTTGTTGCCTTGTCACCACATCCTCTTGAAAAATCGCATTCATTGGCTTCATGAAATTTTATCATGACGTTTTATTCAAAA)
(selector motif: TTTTGAATAAAACGTCATGATAAAATTTCATGAAGCCAATGAATGCGATTTTTCAAGAGGATGTGGTGACAAGGCAACAAATATAAAAGA)

Cross-species, intron 1 to 2a:

```
Species          Energy          Intron  RNA secondary structure prediction                                                          Species          Intron  RNA secondary structure prediction
                                         .......(((((((....((((((.((((....((((((....(((......((..((((.((((((......((((((((....                           ....)))...)))))......))))))))))..))......)))))))))...))))))))))...)).)))))............. 
D. melanogaster  -19.2 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATGGCATGCATTTTCCACATAATTTTATTACCACTTTTAGTC----  D. melanogaster  2a:     TTTTGATTCATAAAACAGCATATAAAATATGAAGACACTTATGAGATTTTTCAAGAGGTGTC--TGACAAGGCTAAAATATAAAAG
D. sechellia     -23.7 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATCTATTGCATGCATTTTCCACGTAATTTTTTTACCACTTTTGGTCAAGA  D. sechellia     2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTCGGGGACAAGGCTAAAATATAAAAG
D. simulans      -20.2 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATTGCATGCATTTTCCACGTAATTTTATTACCACTTTTGGTC----  D. simulans      2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTC--GGACAAGGCTAAAATATAAAAG
```

(Energy of ensemble: -21.38; Frequency of structure in ensemble: 0.00255816)

Cross-species, intron 1 to 2b:

```
Species          Energy          Intron  RNA secondary structure prediction                                                          Species          Intron  RNA secondary structure prediction
                                         .(((...(((..(((((.....((.((((.....(((((((((.(((.(((...((.....................((((....                           ....))))))...)))..))).)))))))))....)))).))))))).)))..)))...........
D. melanogaster  -16.3 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATGGCATGCATTTTCCACATAATTTTATTACCACTTTTAGTC----  D. melanogaster  2b:     AAACGACTTGATTAATATCATCCAATGAATCCGTCGAGAGGAATTGGCGGCCAAGAAATATAA----
D. sechellia     -17.9 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATCTATTGCATGCATTTTCCACGTAATTTTTTTACCACTTTTGGTCAAGA  D. sechellia     2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAACATAATTTA
D. simulans      -20.1 Kcal/Mol  1:      ATCTGTTGCCTTTCACTAGGCATCATCTTACTTAGATTTATTGCATGCATTTTCCACGTAATTTTATTACCACTTTTGGTC----  D. simulans      2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAATATAA----
```

(Energy of ensemble: -17.52; Frequency of structure in ensemble: 0.0140401)

RNA secondary structure found in all internal introns and last intron (*Drosophila melanogaster* and related species):
(selector motif: TTTTGAATAAAACGTCATGATAAAATTTCATGAAGCCAATGAATGCGATTTTTCAAGAGGATGTGGTGACAAGGCAACAAATATAAAAGA)
(acceptor motif: TCTTTTATATTTGTTGCCTTGTCACCACATCCTCTTGAAAAATCGCATTCATTGGCTTCATGAAATTTTATCATGACGTTTTATTCAAAA)

Cross-species, intron 2a to 2c:

```
Species          Energy          Intron  RNA secondary structure prediction                                                               Species          Intron  RNA secondary structure prediction
                                         ...............................(((((......(((((.....((((((............................                            ..............)))))).)))))...)))))......................................................
D. melanogaster  -2.34 Kcal/Mol  2a:     TTTTGATTCATAAAACAGCATATAAAATATGAAGACACTTATGAGATTTTTCAAGAGGTGTC--TGACAAGGCTAAAATATAAAAG  D. melanogaster  2c:     --------------CTTTTTTATTTCTGTGTCTT-------TCTTTAATTGTTTTTATGCTTTAAAATTT---AATTATTGCATTT--
D. sechellia     -5.54 Kcal/Mol  2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTCGGGGACAAGGCTAAAATATAAAAG  D. sechellia     2c:     TTTTAAAGTGTGCATCTCTTTATCCTCGCGGCTTAGAAAACTTTTAAATTGTATCA-------TAAATTTTTACATCAATATATTCGA
D. simulans      -5.64 Kcal/Mol  2a:     TTTTGATTCATAAAACATCATATAAAATATGAAGACACTAATGGGATTTTTCAAGAGGTGTC--GGACAAGGCTAAAATATAAAAG  D. simulans      2c:     TTAGATATTAAGT-TCTCTTTATCCTTCCATCTTTAAAAACCTTTAATATGTATCAAGTATTTGATATTTTTACATCAATTTATTCGA
```

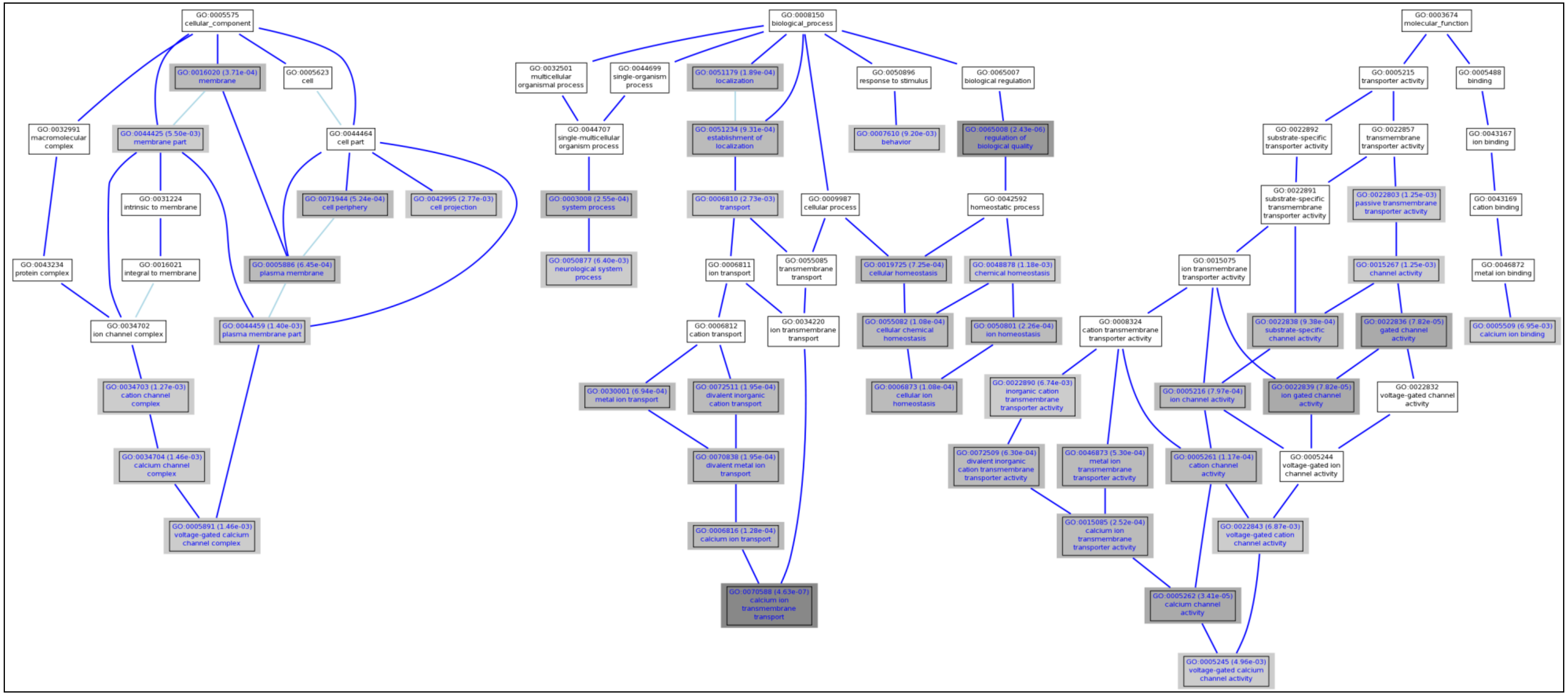
(Energy of ensemble: -6.0; Frequency of structure in ensemble: 0.214515)

Cross-species, intron 2b to 2c:

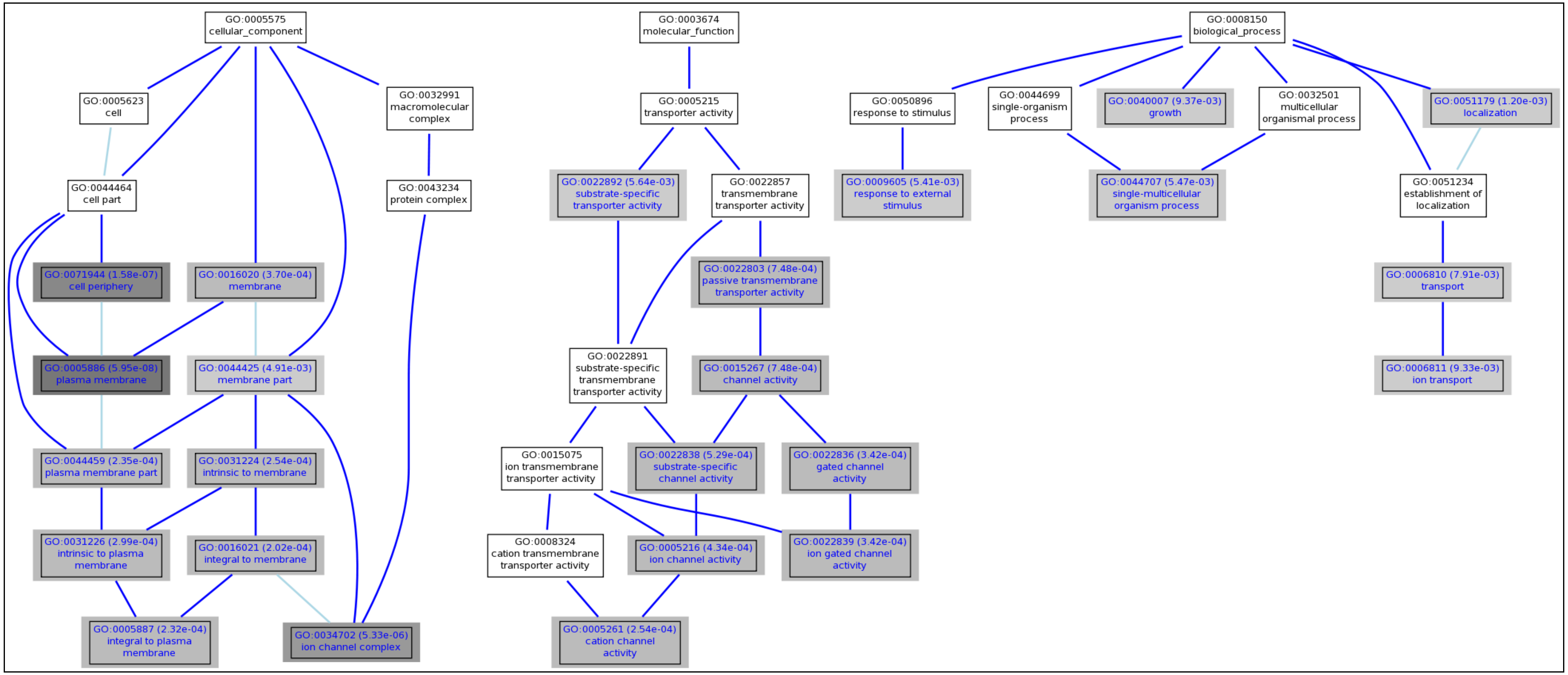
```
Species          Energy          Intron  RNA secondary structure prediction                             Species          Intron  RNA secondary structure prediction
                                         ..................................((.(((((...((.((.................                           ..............)).))...))))).))..........................................................
D. melanogaster  8.43 Kcal/Mol   2b:     AAACGACTTGATTAATATCATCCAATGAATCCGTCGAGAGGAATTGGCGGCCAAGAAATATAA----  D. melanogaster  2c:     --------------CTTTTTTATTTCTGTGTCTT-------TCTTTAATTGTTTTTATGCTTTAAAATTT---AATTATTGCATTT--
D. sechellia     -7.27 Kcal/Mol  2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAACATAATTTA  D. sechellia     2c:     TTTTAAAGTGTGCATCTCTTTATCCTCGCGGCTTAGAAAACTTTTAAATTGTATCA-------TAAATTTTTACATCAATATATTCGA
D. simulans      -2.77 Kcal/Mol  2b:     AAGCGACTTGATTAATCTCATCCAATGAATCCGACGAGAGGAAGTGGCGGCCAAGAAATATAA----  D. simulans      2c:     TTAGATATTAAGT-TCTCTTTATCCTTCCATCTTTAAAAACCTTTAATATGTATCAAGTATTTGATATTTTTACATCAATTTATTCGA
```

(Energy of ensemble: -2.03; Frequency of structure in ensemble: 0.104356)

Supplementary Figure S24. RNA secondary structure prediction for gene CG14608 of *Drosophila melanogaster*.


GO:0005575 cellular_component
GO:0016020 (3.71e-04) membrane
GO:0005623 cell
GO:0032991 macromolecular complex
GO:0044425 (5.50e-03) membrane part
GO:0044464 cell part
GO:0031224 intrinsic to membrane
GO:0071944 (5.24e-04) cell periphery
GO:0042995 (2.77e-03) cell projection
GO:0043234 protein complex
GO:0016021 integral to membrane
GO:0005886 (6.45e-04) plasma membrane
GO:0034702 ion channel complex
GO:0044459 (1.40e-03) plasma membrane part
GO:0034703 (1.27e-03) cation channel complex
GO:0034704 (1.46e-03) calcium channel complex
GO:0005891 (1.46e-03) voltage-gated calcium channel complex
GO:0008150 biological_process
GO:0032501 multicellular organismal process
GO:0044699 single-organism process
GO:0051179 (1.89e-04) localization
GO:0050896 response to stimulus
GO:0065007 biological regulation
GO:0044707 single-multicellular organism process
GO:0051234 (9.31e-04) establishment of localization
GO:0007610 (9.20e-03) behavior
GO:0065008 (2.43e-06) regulation of biological quality
GO:0003008 (2.55e-04) system process
GO:0006810 (2.73e-03) transport
GO:0009987 cellular process
GO:0042592 homeostatic process
GO:0050877 (6.40e-03) neurological system process
GO:0006811 ion transport
GO:0055085 transmembrane transport
GO:0019725 (7.25e-04) cellular homeostasis
GO:0048878 (1.18e-03) chemical homeostasis
GO:0006812 cation transport
GO:0034220 ion transmembrane transport
GO:0055082 (1.08e-04) cellular chemical homeostasis
GO:0050801 (2.26e-04) ion homeostasis
GO:0030001 (6.94e-04) metal ion transport
GO:0072511 (1.95e-04) divalent inorganic cation transport
GO:0006873 (1.08e-04) cellular ion homeostasis
GO:0070838 (1.95e-04) divalent metal ion transport
GO:0006816 (1.28e-04) calcium ion transport
GO:0070588 (4.63e-07) calcium ion transmembrane transport
GO:0003674 molecular_function
GO:0005215 transporter activity
GO:0005488 binding
GO:0022892 substrate-specific transporter activity
GO:0022857 transmembrane transporter activity
GO:0043167 ion binding
GO:0022891 substrate-specific transmembrane transporter activity
GO:0022803 (1.25e-03) passive transmembrane transporter activity
GO:0043169 cation binding
GO:0015075 ion transmembrane transporter activity
GO:0015267 (1.25e-03) channel activity
GO:0046872 metal ion binding
GO:0008324 cation transmembrane transporter activity
GO:0022838 (9.38e-04) substrate-specific channel activity
GO:0022836 (7.82e-05) gated channel activity
GO:0005509 (6.95e-03) calcium ion binding
GO:0022890 (6.74e-03) inorganic cation transmembrane transporter activity
GO:0005216 (7.97e-04) ion channel activity
GO:0022839 (7.82e-05) ion gated channel activity
GO:0022832 voltage-gated channel activity
GO:0072509 (6.30e-04) divalent inorganic cation transmembrane transporter activity
GO:0046873 (5.30e-04) metal ion transmembrane transporter activity
GO:0005261 (1.17e-04) cation channel activity
GO:0005244 voltage-gated ion channel activity
GO:0015085 (2.52e-04) calcium ion transmembrane transporter activity
GO:0022843 (6.87e-03) voltage-gated cation channel activity
GO:0005262 (3.41e-05) calcium channel activity
GO:0005245 (4.96e-03) voltage-gated calcium channel activity

Supplementary Figure S25. Gene Ontology (GO) term enrichment analysis of genes containing mutually exclusive spliced exons (MXEs), which are annotated and reconstructed.



Supplementary Figure S26. Gene Ontology (GO) term enrichment analysis of genes containing mutually exclusive spliced exons (MXEs), which were predicted but not annotated.

2 Supplementary Tables

Supplementary Table S1. The table shows the numbers of mutually exclusive spliced exons (MXEs), which were annotated in Flybase release r5.36, which were predicted, and which were annotated and also predicted. For the prediction criteria a maximal length difference of 20 aa, a minimal similarity score of 15% and a minimal original exon length of 15 aa was used. Predicted MXE candidates, which overlap (and not exactly match) an already annotated exon, were filtered out.

		Matching prediction criteria of MXEs					
	Annotated MXEs	Annotated AND predicted MXEs		Predicted MXEs		Annotated as constitutive or differentially included	
			Cross / EST evidence		Cross / EST evidence		Annotated as MXEs in r5.48
Initial	660	31	28 / 17	65	47 / 20	2	0
3’-terminal	376	42	36 / 22	55	45 / 25	8	0
Internal	261	218	206 / 56	419	321 / 88	159	5
Sum	1297	291	270 / 95	539	413 / 133	169	5

Supplementary Table S2. The table shows the numbers of mutually exclusive spliced exons (MXEs), which were annotated in Flybase release r5.36, which were predicted, and which were annotated and also predicted. In contrast to Table S1, the minimal similarity score was set to 10% instead of 15% (the maximal length difference of 20 aa and the minimal original exon length of 15 aa are unchanged). Predicted MXE candidates, which overlap (and not exactly match) an already annotated exon, were filtered out.

		Matching prediction criteria of MXEs		
	Annotated MXEs	Annotated AND predicted MXEs	Predicted MXEs	Annotated as constitutive or differentially included
Initial	660	42	205	4
3’-terminal	376	48	106	10
Internal	261	228	844	198
Sum	1297	318	1155	212

Supplementary Table S3. The table shows the versions and sources of the genome sequence, protein annotation and EST datasets.

Species	**Dataset release**	**URL**
Drosophila melanogaster	dmel_r5.36_FB2011_04	ftp://ftp.flybase.net/genome
Drosophila ananassae TSC#14024-0371.13	dana_r1.3_FB2011_07	
Drosophila erecta TSC#14021-0224.01	dere_r1.3_FB2011_08	
Drosophila grimshawi TSC#15287-2541.00	dgri_r1.3_FB2010_02	
Drosophila mojavensis TSC#15081-1352.22	dmoj_r1.3_FB2011_05	
Drosophila persimilis MSH-3	dper_r1.3_FB2010_02	
Drosophila pseudoobscura MV2-25	dpse_r2.25_FB2011_10	
Drosophila sechellia Rob3c	dsec_r1.3_FB2011_08	
Drosophila simulans str. Mosaic	dsim_r1.3_FB2011_08	
Drosophila virilis TSC#15010-1051.87	dvir_r1.2_FB2011_07	
Drosophila willistoni TSC#14030-0811.24	dwil_r1.3_FB2010_02	
Drosophila yakuba Tai18E2	dyak_r1.3_FB2011_08	
Daphnia pulex	V1.0	http://www.ncbi.nlm.nih.gov/nuccore/ACJG00000000.1
Anopheles gambiae str. PEST	AgamP3	http://ftp.ncbi.nih.gov/genomes/Anopheles_gambiae
Aedes aegypti str. Liverpool	AaegL1	http://www.ncbi.nlm.nih.gov/nuccore/AAGE00000000.2
Atta cephalotes	Attacep1.0	http://www.ncbi.nlm.nih.gov/nuccore/ADTU00000000.1
Apis mellifera str. DH4	Amel_4.5	http://www.ncbi.nlm.nih.gov/nuccore/AADG00000000.6
Tribolium castaneum str. Georgia GA2	Tcas_3.0	http://www.ncbi.nlm.nih.gov/nuccore/AAJJ00000000.1
Pediculus humanus corporis str. USDA	JCVI_LOUSE_1.0	http://www.ncbi.nlm.nih.gov/nuccore/AAZO00000000.1
Caenorhabditis elegans	WS230	ftp://ftp.wormbase.org/pub/wormbase/releases
Homo sapiens	Build 37.3	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens
Arabidopsis thaliana	TAIR10_genome_release	ftp://ftp.arabidopsis.org/home/tair/Genes
Drosophila melanogaster (EST data)	v2010_11_11	http://www.ncbi.nlm.nih.gov/nucest/?term=txid7227%5BOrganism%5D

3 Supplementary References

49. Zhang, Z., Zhu, X., Stevens, L. M. & Stein, D. Distinct functional specificities are associated with protein isoforms encoded by the Drosophila dorsal-ventral patterning gene pipe. *Development* **136,** 2779–2789 (2009).

50. Wilkin, M. B. *et al.* Drosophila dumpy is a gigantic extracellular protein required to maintain tension at epidermal-cuticle attachment sites. *Curr. Biol.* **10,** 559–567 (2000).