Srrm234, but not canonical SR and hnRNP proteins drive inclusion of *Dscam* exon 9 variable exons

PINAR USTAOGLU[1†], IRMGARD U. HAUSSMANN[1,2], HONGZHI LIAO[1‡], ANTONIO TORRES-MENDEZ[3], ROLAND ARNOLD[4], MANUEL IRIMIA[3,5,6] AND MATTHIAS SOLLER[1,7]

1School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

2Department of Life Science, School of Health Sciences, Birmingham City University, Birmingham, B5 3TN, United Kingdom

3 Centre for Genomic Regulation, Barcelona Institute of Science and Technology (BIST), Barcelona 08003, Spain

4Institute of Cancer and Genomics Sciences, College of Medical and Dental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

5Universitat Pompeu Fabra (UPF), Barcelona 08003, Spain

6ICREA, Barcelona 08010, Spain.

Key Words: Alternative splicing, *Dscam*, RNA binding proteins, SR proteins, SRm300, hnRNP proteins.

Running title: Srrn234 mediates *Dscam* variable exon 9 selection

[6] Corresponding author

Tel: +44 121 414 5905

FAX: +44 121 414 5925

m.soller@bham.ac.uk

Present address: † MRC Centre for Molecular Bacteriology and Infection, and Department of Life Sciences, Imperial College London, Ground Floor, Flowers Building, South Kensington Campus, London SW7 2AZ, UK

‡ Zhongkai University of Agriculture and Engineering, Haizhu District, Guangzhou, China

Abstract

Alternative splicing of pre-mRNA is a major mechanism to diversify protein functionality in metazoans from a limited number of genes. In the *Drosophila melanogaster Down Syndrome Cell Adhesion Molecule (Dscam)* important for neuronal wiring up to 38,016 isoforms can be generated by mutually exclusive alternative splicing in four clusters of variable exons. However, it is not understood how a specific exon is chosen from the many variables and how variable exons are prevented from being spliced together. A main role in the regulation of *Dscam* alternative splicing has been attributed to RNA binding proteins, but how they impact on exon selection is not well understood. Serine-arginine-rich (SR) proteins and hnRNP proteins are the two main types of RNA binding proteins with major roles in exon definition and splice site selection. Here, we analyzed the role of SR and hnRNP proteins in *Dscam* exon 9 alternative splicing in mutant *Drosophila melanogaster* embryos because of their essential function for development. Strikingly, loss or overexpression of canonical SR and hnRNP proteins even when multiple proteins are depleted together, does not affect *Dscam* alternative exon selection very dramatically. Conversely, non-canonical SR protein Serine-arginine repetitive matrix 2/3/4 (Srrm234) is a main determinant of exon inclusion in *Dscam* exon 9 cluster. Since long-range base-pairings are absent in the exon 9 cluster, our data argue for a small complement of regulatory factors as main determinants of exon inclusion in the *Dscam* exon 9 cluster.

Introduction

During alternative splicing the combination of exons can be varied to generate multiple different transcripts and proteins from one gene (Soller 2006; Nilsen and Graveley 2010; Fiszbein and Kornblihtt 2017). In humans, 95% of genes, and in *Drosophila melanogaster* 63% of genes are alternatively spliced, respectively (Wang et al. 2008; Fu and Ares 2014). Among the genes where alternative splicing generates the greatest diversity of isoforms is the *Drosophila melanogaster* homolog of human *Down Syndrome Cell Adhesion Molecule* (*Dscam*), which encodes a cell surface protein of the immunoglobulin superfamily. The *Dscam* gene comprises 95 alternatively spliced exons that are organized into four clusters, namely 4, 6, 9 and 17 which contain 12, 48, 33 and 2 variables, respectively. Hence, the *Dscam* gene can generate up to 38,016 different proteins (Schmucker et al. 2000; Neves et al. 2004; Hemani and Soller 2012; Sun et al. 2013). *Dscam* is functionally required for neuronal wiring in the nervous system, but also for phagocytosis of invading pathogens in the immune system (Schmucker et al. 2000; Watson et al. 2005). Interestingly, *Dscam* in mosquitos changes its splicing pattern upon pathogen exposure to produce isoforms with higher binding affinity for binding pathogen (Dong et al. 2006). However, despite intense research, relatively little is known about how *Dscam* alternative splicing is regulated in flies.

Pre-mRNA splicing is a multistep process catalyzed by the spliceosome sequentially assembled from five U snRNPs together with numerous proteins. Spliceosome assembly initiates by the recognition of the 5' splice site by U1 snRNP and of the 3' splice site by U2 snRNP together with U2AFs recognizing the branchpoint and the polypyrimidine tract and the AG of the 3' splice site, respectively. Then the U4/5/6 tri-snRNP is recruited and upon several structural

rearrangements where U4 snRNP leaves the spliceosome, catalysis takes place by two transesterification reactions (Luhrmann and Stark 2009).

Alternative splicing is to a large degree regulated at the level of splice site recognition involving base-pairing of U1 snRNP to the 5' splice site YAG/GURAGU and U2 to the branchpoint WNCUAAU (W: A or U, *Drosophila melanogaster* consensus(Lim and Burge 2001)) whereby splice sites closer to the consensus are preferably used (Soller 2006). Splice site selection is critically assisted by RNA binding proteins (RBPs) that support or inhibit recognition of splice sites.

Serine-arginine rich (SR) and heterogenous nuclear ribonucleoproteins (hnRNPs) are two prominent classes of RBPs involved in alternative splicing regulation (Busch and Hertel 2012; Fu and Ares 2014; Bradley et al. 2015). Humans have twelve and flies have eight SR proteins each having one or two RNA Recognition Motifs (RRM) and RS domain rich in serines and arginines (Busch and Hertel 2012). In addition, RS domains are present in some other splicing factors lacking RNA binding domains such as Tra2, SRRM1 (SRm160), and SRRM2 (SRm300), SRRM3, and SRRM4 (nSR100), which are termed non-canonical SR proteins (Blencowe et al. 1999; Long and Caceres 2009; Busch and Hertel 2012; Best et al. 2014). In contrast, hnRNP proteins are more diverse in their modular assembly containing RNA binding domains (e.g. RRM, KH or RGG domains) and various auxiliary domains (Geuens et al. 2016). In humans, the most prominent hnRNPs are the abundantly expressed hnRNP A and C family (Busch and Hertel 2012; Geuens et al. 2016).

SR proteins mostly bind to exonic splicing enhancers (ESEs) and recruit spliceosomal components to splice sites through their RS domains. SR protein binding sites are present in both alternatively spliced and constitutive exons to promote exon inclusion, but they can also repress

inclusion of alternative exons (Black 2003; Shen et al. 2004; Wang et al. 2006; Chen and Manley 2009; Pandit et al. 2013). Although SR proteins recognize similar sequences, distinct functions have been shown either by binding distinct sites or when bound to the same site through differential regulation mediated by combinatorial interactions with other splicing regulators (Gabut et al. 2007; Anko et al. 2012; Pandit et al. 2013; Bradley et al. 2015).

In contrast to SR proteins, hnRNP proteins mostly bind to intronic splicing silencers (ISSs) and repress inclusion of alternative exons (House and Lynch 2006; Wang et al. 2008). In addition, they can act antagonistically to SR proteins by binding to exonic splicing silencers (ESSs) and compete with SR proteins for binding (Soller 2006; Long and Caceres 2009). However, a more comprehensive analysis revealed that SR and hnRNP proteins can also act coordinately in many instances in exon inclusion or repression (Brooks et al. 2015).

With regard to the alternative splicing in the *Dscam* gene a model has been proposed for the exon 6 cluster involving long-range base-pairing. Here, a conserved docking sequence in the first intron of the exon 6 cluster can base-pair with complementary selector sequences in front of each variable exon to bring a chosen variable exon into the proximity of the proximal constant flanking exon for splicing (Graveley 2005). This model also requires that the entire cluster is maintained in a repressed state and variable exons are selected under the control of RBPs. However, the architecture in the exon 4 an 9 cluster is different and conserved “docking site” sequences are found at the end of the exon 4 and 9 clusters, but the support for this model based on evolutionary sequence conservation is weak (Yang et al. 2011; Haussmann et al. 2019).

RNAi knock-down of RBPs in cell culture in *Drosophila melanogaster* S2 cells revealed little changes for most variable exons in the *Dscam* exon 4 cluster, but this result could be due to residual protein left (Park et al. 2004). Hence, we wanted to investigate the role of SR and

hnRNP proteins in *Drosophila melanogaster Dscam* alternative splicing more comprehensively at an organismal level using knock-out mutants and overexpression. Most SR and hnRNP proteins are essential for development to adult flies, but embryonic development proceeds such that *Dscam* alternative splicing could be analysed in late stage embryos in SR and hnRNP mutants or when over-expressed. Unexpectedly, we find that inclusion of *Dscam* exon 9 variables is affected little in loss or gain of function conditions. Likewise, even upon removal even of multiple factors *Dscam* exon 9 alternative splicing changes little. However, the non-canonical SR protein Serine-arginine repetitive matrix 2/3/4 (Srrm234) is required for inclusion of most exon 9 variables. We further find that long-range base-pairing is not supported as a general model. Hence, our results argue that a small complement of RBPs are main regulators of *Dscam* exon 9 alternative splicing.

Results

Analysis of *Dscam* exon 9 alternative splicing by restriction digests

All 33 exons in the *Drosophila melanogaster Dscam* exon 9 variable exon cluster have about the same length and run as a single band on an agarose gel (Figure 1A and B). The sequences in variable exons, however, differ enough such that a complement of restriction enzymes can digest the complex mix of PCR products to identify a majority of isoforms on sequencing type gels using one ^{32}P labelled primer after reverse transcription of the mRNA (Figure 1C and D) (Haussmann et al. 2019). Using a combination of SacII, ClaI, PshAI, HaeIII, MseI, BsrI and BstNI yields 26 fragments of unique size identifying 20 variable exons (Fig 1D).

Alteration of SR proteins has little impact on *Dscam* exon 9 alternative splicing

SR proteins are organized into five families and representative orthologues are present in *Drosophila melanogaster* (Fig 2) (Busch and Hertel 2012). To determine the role of SR proteins in *Drosophila melanogaster Dscam* exon 9 alternative splicing we obtained mutants and over-expression lines for most of the canonical SR proteins as well as for general splicing factor SF1 and non-canonical SR protein Srrm1 (SRm160) and Srrm234 (SRm300, CG7971) (Fig 2A, Supplementary Fig S1). For loss of function (LOF) alleles of SR genes, five from the ten analysed, which are *X16* ^{GS1678}, *SF2*^{GS22325}, *B52*^{29}, *Srrm1* ^{B103} and *Srrm234*$^{\Delta N}$, are required to reach adulthood (Fig 2A and B). Gain of function (GOF) conditions by pan-neural *elavGAL4* mediated over-expression via *UAS* was lethal in larval instars for *UAS GFP-X16, UAS RSF1, UAS GFP-SC35, UAS GFP-SF2* and *UAS GFP-B52*, while over-expression of *UAS SF1* and *UAS Srrm1* from *EP* lines did not result in a phenotype (Fig 2A and C).

Next, we analysed inclusion levels of exon 9 variables in embryos for LOF and GOF conditions of canonical SR proteins and SF1. For LOF alleles *X16* ^{GS1678}, *RBP1* ^{HP37044}, *RBP1-like* ^{NP0295}, *Srp54*^{GS15334}, *SC35*^{KG02986}, *SF2*^{GS22325}, *B52*^{28} and *SF1*^{G14313} the *Dscam* exon 9 splicing pattern unexpectedly remained largely unchanged and significant changes are prominent in exon 9.4, 9.21 and 9.5/9.11 for mutants compared to wild type (marked in red for decrease and blue for increase, Figure 3A and B). Similar results were obtained for over-expression of *UAS GFP-X16, UAS RSF1, UAS GFP-SC35, UAS GFP-SF2, UAS GFP-B52* and *UAS-Srrm1* with significant changes only prominent in exon 9.19 and 9.21 for GOF conditions compared to wild type (marked in red for decrease and blue for increase, Fig 3C and D).

Non-canonical SR protein Srrm234 is required for selection of *Dscam* exon 9 variables

The canonical SR proteins contain an RRM and bind to RNA. In contrast, the large Srrm1 and Srrm234 proteins contain RS domains, but seem not to bind RNA and exert splicing enhancing functions through association with proteins bound to ESEs (Fig 2) (Blencowe et al. 1998; Eldridge et al. 1999; Blencowe et al. 2000; Szymczyna et al. 2003).

We obtained null mutants for both genes, *Srrm1* ^{B103} and *Srrm234*$^{\Delta N}$, which are late embryonic lethal (Fan et al. 2014). While loss of Srrm1 had little effect on *Dscam* exon splicing, loss of Srrm234 resulted in significant reduction in inclusion for many variable exons (marked in red, 9.4, 9.7, 9.10, 9.16/9.27, 9.17, 9.18, 9.20, 9.23/9.25, 9.29 and 9.33) that is compensated by increased inclusion of a few variable exons (marked in blue, 9.3/9.32, 9.6/9.8, 9.12 and 9.30) (Figure 4A and B). Consistent with a role in *Dscam* exon 9 alternative splicing regulation, *Srrm234* is also expressed in the nervous system of *Drosophila melanogaster* embryos in the same pattern as *Dscam* (Supplementary Fig S2).

Alteration of hnRNP proteins has little impact on *Dscam* exon 9 alternative splicing

Since alterations of individual canonical SR proteins had little impact on selection of *Dscam* exon 9 variables, we focused on hnRNP proteins as candidates for repressing inclusion of exon 9 variables (Olson et al. 2007; Chen and Manley 2009; Fu and Ares 2014). *Drosophila melanogaster* has four members of the highly expressed hnRNP A family (Hrp36, Hrp38, Rb97D and Hrp48) and one member of the hnRNP C, although the *Drosophila melanogaster* orthologue is considerably longer (Fig 5)(Appocher et al. 2017). Other highly expressed hnRNP proteins are Hrp40 (hnRNP D), Glorund (hnRNP F/H), Hrb57A (hnRNP K) and Hrb59 (hnRNP M).

To determine the role of hnRNP proteins in *Drosophila melanogaster Dscam* exon 9 alternative splicing we could obtain null-mutants for all major hnRNP proteins (*Hrp36*^{BG02743}, *Hrp38*^{MI1059},

RB97D^{1}, *Hrp48* ^{GS14498}, *Hrp40* ^{GS18188}, *glo* ^{f02674}, *Hrb57* ^{G13574} and *Hrp59*^{GS6029}) and gene-switch or *EP* over-expression lines for most (Hrp36 ^{GS15926}, Hrp38 ^{GS12795}, Hrp48 ^{EY12571}, and Hrp59 ^{GS6029}, Fig 5, Fig 7A, Supplementary Fig S3). Half of the tested major hnRNPs are required for viability (*Rb97D*, *Hrp40*, *glo* and *Hrp59*), while only over-expression of Hrp36 was lethal (Fig 5).

Then, we analysed *Dscam* exon 9 inclusion levels for LOF and GOF of hnRNPs. For LOF alleles, the *Dscam* exon 9 splicing pattern also unexpectedly remained largely unchanged and significant changes are prominent in exon 9.17, 9.19 and 9.21 for mutants compared to wild type (marked in red for decrease and blue for increase, Fig 6A and B). Similar results were obtained for GOF conditions with significant changes only prominent in exon 9.2, 9.4, 9.17, 9.19, 9.21, and 9.30 (marked in red for decrease and blue for increase, Fig 6C and D).

Binding of Hrp36, Hrp38, Hrp48 and Hrp40 to RNA has been analysed globally in *Drosophila melanogaster* and binding motifs have been established by SELEX (Blanchette et al. 2009). When we reanalysed these data we did find significantly increased binding for all four Hrp proteins ($p<0.05$) to the *Dscam* exon 9 variable cluster compared to averaged binding, but the binding curves did not overlap with changes in exon inclusion in LOF or GOF conditions (Supplementary Figure S4). Likewise, we did not find an overlap of SELEX motif enrichments with changes in exon inclusion in LOF or GOF conditions (Supplementary Figure S4).

Removal of multiple SR and hnRNP proteins has little impact on selection of *Dscam* exon 9 variables

hnRNP36 and SR protein B52 genes lie next to each other in the *Drosophila melanogaster a* genome and potentially could cross-regulate to compensate for each other (Fig 7A). Therefore, we generated a double knock-out of *hnRNP36* and *B52* genes (*hnRNP36/B52*$^{\Delta 1}$). In addition, we

also combined this double knock-out with the *hnRNP38* mutant as *hnRNP36* and *hnRNP38* are closely related and since they are highly expressed, they could act redundantly.

Surprisingly, even in *hnRNP36/B52*$^{\Delta 1}$ *hnRNP38*^{d05172} triple mutants, *Dscam* exon 9 was robustly spliced with only differences in inclusion levels of variables 9.4, 9.16/9.27, 9.17, 9.18, 9.19, and 9.21 compared to controls (marked in red for decrease and blue for increase, Fig 7B and C.

Variable exon selection is not explained by long-range base-pairing in *Dscam* exon 9

In the exon 4 cluster, we could not detect conserved sequences adjacent to every variable exon that could mediate long-range base-pairing arguing that such a mechanism is not involved in variable exon 4 selection (Haussmann et al. 2019). A sequence alignment between *Drosophila melanogaster* and *Drosophila virilis* showed strong conservation in the coding sequences and the architecture of the exon 9 cluster with only few insertions and deletions of exons (Supplementary Fig S5). There are conserved sequence elements in the intron before constant exon 10 that potentially could serve as a docking site (Yang et al. 2011), but we did not find conserved sequence elements between every variable exon by manual inspection (Supplementary Fig S6). A systematic bioinformatics comparison of this potential docking site with arbitrary sampled sequences of the same length and sequence complexity further did not reveal a special propensity of the docking site to form more or stronger complementary alignments within *Dscam* intronic sequences in the variable exon 9 cluster. The same picture arises when comparing predicted energies of the best secondary duplex structures formed by the reverse intronic sequence, the docking site, and sampled sequences (Supplementary Fig S7). In addition, we didn't find other instances of the docking site in any other gene of *D. melanogaster* (except in the anti-sense RNA *CR45129* of *Dscam*), indicating that the sequence is not recurrent in other genes.

Discussion

Although the sequence determinants that direct the spliceosome to its correct position are very degenerate, splicing needs to occur precisely and with high accuracy to prevent disease (Cooper et al. 2009; Zaharieva et al. 2012). It is therefore thought that RNA binding proteins play key roles in localizing functional splice sites. In particular, the abundant SR and hnRNP proteins have been attributed key roles in this process by forming RNA-protein complexes co-transcriptionally to recruit early splicosomal components for defining splice sites. Although SR proteins were initially viewed as binding ESEs to activate splicing, and hnRNP proteins to bind ISSs for antagonizing SR proteins, a number of genome-wide studies draw a more complex picture for both SR and hnRNP binding and function (Blanchette et al. 2009; Anko et al. 2010; Anko et al. 2012; Huelga et al. 2012; Pandit et al. 2013; Brooks et al. 2015). In fact, both SR and hnRNP proteins can have very specific functions in one context, but also redundant functions in another context.

In this regard, we hypothesised that an array of similar exons as found in the *Drosophila melanogaster Dscam* gene would provide a platform for SR and hnRNP proteins to evolve exon specific functions to regulate their inclusion. Surprisingly, however, *Dscam* exon 9 alternative splicing is exactly the opposite and the splicing pattern is very robustly maintained when SR and hnRNP proteins were either removed or overexpressed. In particular, the advances of *Drosophila melanogaster* genetics allowed us to use complete knock-outs of most canonical SR and general hnRNP proteins and thus avoid the ambiguity of RNAi that would leave residual protein. Hence, despite complete loss of individual SR and hnRNP proteins, or combinations thereof, the *Dscam* splicing pattern is robustly maintained.

Two explanations are possible for this scenario. First, SR and hnRNP proteins act redundantly at the very extreme such that fluctuations in many would need to occur to impact on *Dscam* alternative splicing. However, whether this model applies will be difficult to test as removal of many general splicing factors will likely lead to global perturbations affecting many genes.

A second scenario could be that *Dscam* alternative splicing is fairly independent of general splicing factors. This would imply a more specific mechanism. Initially, it has been though that long-range base-pairing would provide such a mechanism as conserved sequences have been found in the *Dscam* exon 6 cluster (Graveley 2005). Our previous analysis of the exon 4 cluster, and the in depth analysis of the exon 9 cluster in this paper, however, rule out such mechanism in these two clusters (Haussmann et al. 2019). Hence, the question remains whether two independent mechanisms arose to regulate mutually exclusive alternative splicing in *Dscam* variable clusters. Potentially, the conserved sequences present in the variable clusters could provide binding sites for RBPs that have adopted cluster specific roles. In this context it is interesting to note that deletion of the docking site in exon 6 leads to inclusion of mostly the first exon in the cluster (May et al. 2011). This is unexpected as removal of the splicing activating mechanism should result in skipping of the entire variable cluster, because the proposed repressor hnRNP36 would still be present. Accordingly, the docking site in the exon 6 cluster also exerts a repressive role in maintaining the entire cluster in a repressed state.

Likewise, our finding that non-canonical Srrm234 regulates inclusion of many variables in the *Dscam* exon 9 cluster suggests a mechanism in *Dscam* mutually exclusive alternative splicing that differs from more general splicing rules directed by canonical SR and general hnRNP proteins. One of the human homologs of Srrm234, SRRM4 has key roles in the regulation of microexons (Irimia et al. 2014). Due to their small size, microexons cannot be defined through

standard mechanism of exon definition. Hence, a distinct mechanism must apply, that can accurately direct splicing of microexons. Because microexons are often found in large introns, a robust process must underlie their selection and involves a newly described enhancer of microexons domain (eMIC), present in human SRRM3 and SRRM4. Intriguingly, in vertebrates, the ancestral pro-homologue Srrm234 has duplicated into three genes to adopt distinct functions through dedicated protein domains, but these features are maintained by alternative mRNA processing in *Drosophila melanogaster* Srrm234 to include the eMIC at the C-terminus of the protein in neuronal tissue (Torres-Mendez et al. 2019).

Dscam alternative exons comply with the general average length of exons and thus the mechanism of their regulation is likely distinct from microexons. *Dscam* exon 9 cluster regulation by Srrm234 seems to involve its Cwf21 domain, which is not required for microexon inclusion (Torres-Mendez et al. 2019). Transposon inserts in the middle of the *Srrm234* gene resulting in a truncated protein, that contains the Cwf21 domain do not affect exon 9 diversity (*Mi{ET1}CG7971*^{MB07314} and *Mi(Liang et al.)CG7971*^{MI04068}, data not shown). The Cwf21 domain, which is homologous to the yeast Cwc21 domain, has been attributed key roles in splicing due to co-purification of the human ortholog SRRM2 with active spliceosomes and its localization in the catalytic centre of the spliceosome (Bessonov et al. 2008). Furthermore, the Cwc21 domain interacts with the U5 snRNP core components Snu114 and Prp8 involved in key structural rearrangements in the spliceosome during catalysis (Grainger et al. 2009; Gautam et al. 2015). Interestingly, Cwc21 has been attributed roles in splicing of meiotic genes which are regulated differently of general intron containing genes (Gautam et al. 2015). Srrm2 forms a complex with Srrm1 to promote alternative splicing of *Drosophila melanogaster doublesex* required for sex determination, but the mechanism in *Dscam* is different as loss of Srrm1 does

not impact on alternative splicing in the exon 9 cluster (Blencowe et al. 1998; Eldridge et al. 1999). In addition, Srrm234 also interacts with U1 70K, so potentially could act to activate 5' splice site and initiate the splicing process from a repressed state of the variable cluster (Guruharsha et al. 2011).

A *Dscam* cluster-specific role has been suggested for Hrp36 acting as a repressor preventing splicing together of exon 6 variables in S2 cells, but whether the results are the same in flies remains to be tested (Olson et al. 2007). Such a factor has not been identified for exon 4 and 9 clusters, and for the factors tested, we did not observe splicing together of variable exons. Intriguingly, a newly annotated exon 4.0 in the beginning of the bee *Dscam* exon 4 cluster is spliced to exon 4.6 (Decio et al., 2019). This unexpected finding, however, is not compatible with the model described for Hrp36 in the exon 6 cluster, but might involve a repressive sequence element around exon 4.0 similar the element discovered in the beginning of the exon 6 cluster, which when deleted results only in inclusion of exon 6.1 (May et al. 2011). Likewise, factors like Srrm234 would then act as activators to drive inclusion of variables.

Taken together, *Dscam* exon 9 mutually exclusive alternative splicing is robust against fluctuations of in canonical SR and general hnRNP proteins arguing for a specific mechanism regulating inclusion levels of variable exons. Indeed, non-canonical SR protein Srrm234 plays a key role increasing inclusion of many exon 9 variables. However, since Srrm234 does not have not have one of the classic RNA binding domains, additional RNA binding proteins likely connect Srrm234 to *Dscam* exon 9. Hence, our data obtained from knock-outs of general splicing factors indicate that a small complement of RNA binding proteins are likely key regulators of *Dscam* mutually exclusive alternative splicing.

The gene structure of invertebrate *Dscam* harbouring arrays of variable exons for mutually exclusive splicing is at the extreme, but arrays of several alternative exons are common to many genes in metazoans. Likely, a yet to be discovered feature of the spliceosome has been exploited in mutually exclusive alternative splicing of *Dscam* such that only on exon is chosen. Likewise, since sequences that look like splice site are common in large introns, such a mechanism could be broadly relevant for robust select of isolated exons.

Materials and Methods

Fly genetics

Flies were maintained on standard cornmeal agar food as described (Haussmann et al. 2013). *CantonS* was used as a wild type control. The following loss of function mutants were used as depicted in Supplementary Figs S1 and S2, and Fig 7A: *X16 (P{GSV6}Hrb27C*^{GS16784}, DGRC Kyoto #206763*), RBP1 (P{EPg}mRpL37*^{HP37044}, BDSC #22011*), RBP1-like (P{GawB}Rbp1-like*^{NP0295}, DGRC Kyoto #103580*), Srp54 (P{GSV6}Srp54*^{GS15334},'DGRC Kyoto #206174*), SC35 (P{SUPor-P}SC35*^{KG02986}, BDSC #12904*), SF2 (P{GSV7}SF2*^{GS22325}, DGRC Kyoto #203903*), B52*^{28}*(Gabut et al. 2007), SF1 (P{EP}SF1*^{G14313} BDSC #30203*), Hrp36 (P{GT1}Hrb87F*^{BG02743}, BDSC #12869*), Hrp38 (Mi(Liang et al.)Hrb98DE*^{MI10594}, BDSC #55509*), Rb97D (P{PZ}Rb97D*^{1}, BDSC #11782*), Hrp48 (P{GSV6}Hrb27C*^{GS14498}, DGRC Kyoto #205836*), Hrp40 (P{GSV6}sqd*^{GS18188}, DGRC Kyoto #201020*), Glo (PBac{WH}glo*^{f02674}, BDSC #18576*), Hrb57A (P{EP}HnRNP-K*^{G13574}, BDSC #29672*), Hrp59 (P{GSV3}rump*^{GS6029}, DGRC Kyoto #200852*), Srrm1*^{B103} (*SRm160*^{B103}, (Fan et al. 2014) and *Srrm2340*$^{\Delta N}$. Since the transposon used for mutagenesis are large (~10 kb), inserts in the transcribed part were considered to be null alleles, while inserts in promoter regions were considered hypomorphic alleles. If inserted in an

intron transposons disrupt splicing, if inserted in the 5'UTR they will prevent translation of the ORF, or if inserted in the ORF lead to a truncated non-functional protein. If inserted in the promotor region, transposon inserts reduce transcription. Whether lethality of mutants mapped to the locus was tested by crossing to chromosomal deficiencies.

The null-allele *Srrm234*$^{\Delta N}$ (*CG7971*) was generated by *GenetiVision* CRISPR gene targeting services. The 3.2kb deletion at the N-terminus of the gene was generated using sgRNAs AGTCTGCTGGGGACACTGCT and CGCCGCAGGACATATAACAG together with donor template harbouring two homology arms flanking a *loxP 3xP3-GFP loxP* cassette. Left and right homology arms of the donor were amplified using primers CG7971-LAF1 (GTTCCGGTCTCTTAGCCCTGCAGCAGCTTCTGCTTG) and CG7971-LAR1 (TCCAAGGTCTCACAGTTTATATGTCCTGCGGCGCTGC), and CG7971-RAF2 (GTTCCGGTCTCTGTCAGCTGGGAGCCGGCAGTGC) and CG7971-RAR2 (TCCAAGGTCTCAATCGAGTGGAGAACCCATACGTACTTAGATCC), respectively. Successful deletion and integration of the cassette was validated by PCR and Sanger sequencing using primers CG7971-outF1 (CATCGATTGTGTTGCATGAAGTTCAC) and CG7971-outR2 (GGGGAGTATCTGTGAGCAGTTGTATC), and LA-cassette-R (AAGTCGCCATGTTGGATCGACT) and Cassette-RA-F (CCTGGGCATGGATGAGCTGT), respectively. For the analysis of *Dscam* alternative splicing the *3xP3-GFP* marker was removed by Cre mediated recombination using an insert on a third chromosome balancer (*TM6B, P{w[+mC]=Crew}DH2, Tb*, BDSC #1501) and the resulting chromosome was rebalanced with a zygotically YFP- expressing balancer (*TM6B, P{Dfd-EYFP}3 Sb, Tb*, BDSC #8704) to collect the embryonic lethal homozygous mutants.

For gain of function experiments the following UAS lines, gene switch vector inserts and EP lines were used: *UAS GFP* control line (Gabut et al. 2007), *UAS GFP-X16* (Gabut et al. 2007), *UAS RSF1* (Labourier et al. 1999), *UAS GFP-SC35* (Gabut et al. 2007), *UAS GFP-SF2* (Gabut et al. 2007), *Srrm1 (P{EP}Srrm1*^{G18603}, BDSC #26938*), UAS GFP-B52* (Gabut et al. 2007), *Hrp36 (P{GSV6}*^{GS15926}, DGRC Kyoto #206416), *Hrp38 (P{GSV6}*^{GS12795}, DGRC Kyoto #204283), *Hrp48 ({EPgy2}Hrb27C*^{EY12571}, BDSC #20758), *Hrp59 (P{GSV3}*^{GS6029}, DGRC Kyoto #200852).

Hrp36 and *B52* genes lie next to each other. The *Hrp36/B52*$^{\Delta 1}$ double mutant was generated by FRT/FLP mediated recombination using *PBac{RB}e01378* and *PBac{WH}f01884* transposon insertions as described (Zaharieva et al. 2015). The lethal *Hrp36/B52*$^{\Delta 1}$ allele was balanced and validated using primers e01378 Rev (GCCACATTTAGATGATTCAGCATTAT), f01884 Rev (GATTCCAATAGATCCCAACCGTTTCG) and RB 3' MINUS (TCCAAGCGGCGACTGAGATG).

Lethal lines were rebalanced with balancers expressing YFP zygotically, but not maternally under a *Dfd* promoter (*CyO, P{Dfd-EYFP}2*, BDSC #8578, *TM6B, P{Dfd-EYFP}3 Sb, Tb*, BDSC #8704) to allow for selection of homozygous lethal mutants. Non-GFP expressing 14-18 h embryos were further selected according to the morphology of the auto-fluorescing gut to distinguish them from homozygous balancer carrying animals, which die before they express GFP (Haussmann et al. 2008). For over-expression a third chromosomal *elavGAL4* insert was used (*P{w[+mmC]=GAL4-elav.L}3*, BDSC #8760).

RNA extraction, RT-PCR, restriction digestion, denaturing acrylamide gels

Total RNA was extracted using Tri-reagent (SIGMA) and reverse transcription was done with Superscript II (Invitrogen) as described (Koushika et al. 1999) using primer *Dscam* 11RT1 (CGGAGCCTATTCCATTGATAGCCTCGCACAG, 1 pmol/ 20 µl reaction). PCR to amplify *Dscam* exon 9 cluster was done using primers 8F1 (GATCTCTGGAAGTGCAAGTCATGG) and 10R1ΔST (GGCCTTATCGGTGGGCACGAGGTTCCATCTGGGAGGTA) for 37 cycles with 1 µl of cDNA. Primers were labeled with ^{32}P gamma-ATP (6000 Ci/ mmol, 25 mM, Perkin Elmer) with PNK (NEB) to saturation and diluted as appropriate. From a standard PCR reaction with a ^{32}P labelled forward primer, 10–20% were sequentially digested with a mix of restriction enzymes (NEB) according to their buffer requirements and temperatures. PCR reaction and restriction digests were phenol/$CHCl_3$ extracted, ethanol precipitated in the presence of glycogen (Roche) and analyzed on standard 6% sequencing type denaturing polyacrylamide gels. After exposure to a phosphoimager (BioRad), individual bands were quantified using ImageQuant (BioRad) and inclusion levels for individual variable exons were calculated from the summed up total of all variables. Statistical analysis was done by one-way ANOVA followed by Tukey–Kramer post-hoc analysis using Graphpad prism. Percent inclusion levels of exon 9 variables of embryos were calculated from the total sum of variables. RNA in situs were obtained from flybase as described (Haussmann et al. 2008).

Sequence analysis

The in silico analysis for SELEX motif occurrence was done by plotting the scores from sliding window using a position weight matrix of the *Drosophila melanogaster* consensus sequence for a given RNA binding protein to the *Dscam* exon 9 variable cluster ($p<0.05$, Korhonen et al. 2009, Blanchette et al. 2009). The CLIP data was downloaded from GEO as deposited

in Blanchette et al. 2009, and the TiMAT window-scores have been used to generate the plot based on the version 4 of the genome assembly. Tiling arrays scores within the *Dscam* variable exon 9 cluster were plotted against all scores to determine whether the four Hrp proteins show significantly increased binding.

Vista alignments were generated as described (Haussmann et al. 2011). The exon 9 docking site was scanned against all gene sequences (as downloaded from FlyBase on the 26th of February 2019) using Blat (Version 35, parameters: -stepSize=1 tileSize=6 -minScore=0 -minIdentity=0, filtering for hits of at least bit-score of 30 and a coverage over the docking sequence of at least 20 residues) (Kent 2002).

In order to compare the docking site's propensity to form potential long-range base-pairing with other intronic *Dscam* sequences of the same length, we sampled 100 strictly intronic sequenced of the same or higher sequence complexity as determined by Shannon entropy. We then used the BioPython Bio.pairwise2 local alignment implementation to get pairwise alignments against all (excluding the sampled sequenced and the docking sequence) reverse and reverse-complement sub-sequences of the same length in any *Dscam* intron, using a custom substitution matrix to allow for G-U pairings (scoring scheme: G->U 0.8, A->U: 1, G->C: 1.2, gap opening penalty 0.1, gap extension penalty 0.1) (Cock et al. 2009). We then retained all hits with less than 3 gaps and normalized the alignment scores by each queries' potential highest score. To assess the best potential secondary structure formed by the docking sequence with any intronic sequence, we run the same sampled sequenced against the concatenated reverse and reverse complement intronic sequences of *Dscam* (masking each query) using RNAduplex from the Vienna package and obtained for each sequence the predicted lowest energy secondary structure prediction (Lorenz et al. 2011).

Acknowledgments

We thank Bloomington, Kyoto and Harvard stock centres, and J. Tazi and L. Rabinow for fly lines, and P. Grzechnik for comments on the manuscript. We acknowledge funding from the Sukran Sinan Memory Fund to P.U, BBSRC and the European Research Council (ERC-StG-LS2-637591) to M.I.

Author contributions: M.S. and P.U. designed and directed the project. P.U., M.S., I.H., H.L. performed experiments. A. T-M and M.I. generated and validated the *Srrm234*$^{\Delta N}$ allele. R.A. performed bioinformatic analysis. P.U. and M.S. wrote the manuscript. All authors read and approved the final manuscript.

References

Anko ML, Morales L, Henry I, Beyer A, Neugebauer KM. 2010. Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells. *Nat Struct Mol Biol* **17**: 962-970.

Anko ML, Muller-McNicoll M, Brandl H, Curk T, Gorup C, Henry I, Ule J, Neugebauer KM. 2012. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes. *Genome Biol* **13**: R17.

Appocher C, Mohagheghi F, Cappelli S, Stuani C, Romano M, Feiguin F, Buratti E. 2017. Major hnRNP proteins act as general TDP-43 functional modifiers both in Drosophila and human neuronal cells. *Nucleic Acids Res* **45**: 8026-8045.

Bessonov S, Anokhina M, Will CL, Urlaub H, Luhrmann R. 2008. Isolation of an active step I spliceosome and composition of its RNP core. *Nature* **452**: 846-850.

Best A, Dalgliesh C, Kheirollahi-Kouhestani M, Danilenko M, Ehrmann I, Tyson-Capper A, Elliott DJ. 2014. Tra2 protein biology and mechanisms of splicing control. *Biochem Soc Trans* **42**: 1152-1158.

Black DL. 2003. Mechanisms of alternative pre-messenger RNA splicing. *Annu Rev Biochem* **72**: 291-336.

Blanchette M, Green RE, MacArthur S, Brooks AN, Brenner SE, Eisen MB, Rio DC. 2009. Genome-wide analysis of alternative pre-mRNA splicing and RNA-binding specificities of the Drosophila hnRNP A/B family members. *Mol Cell* **33**: 438-449.

Blencowe BJ, Bauren G, Eldridge AG, Issner R, Nickerson JA, Rosonina E, Sharp PA. 2000. The SRm160/300 splicing coactivator subunits. *Rna* **6**: 111-120.

Blencowe BJ, Bowman JA, McCracken S, Rosonina E. 1999. SR-related proteins and the processing of messenger RNA precursors. *Biochem Cell Biol* **77**: 277-291.

Blencowe BJ, Issner R, Nickerson JA, Sharp PA. 1998. A coactivator of pre-mRNA splicing. *Genes Dev* **12**: 996-1009.

Bradley T, Cook ME, Blanchette M. 2015. SR proteins control a complex network of RNA-processing events. *Rna* **21**: 75-92.

Brooks AN, Duff MO, May G, Yang L, Bolisetty M, Landolin J, Wan K, Sandler J, Booth BW, Celniker SE et al. 2015. Regulation of alternative splicing in Drosophila by 56 RNA binding proteins. *Genome Res* **25**: 1771-1780.

Busch A, Hertel KJ. 2012. Evolution of SR protein and hnRNP splicing regulatory factors. *Wiley Interdiscip Rev RNA* **3**: 1-12.

Chen M, Manley JL. 2009. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol* **10**: 741-754.

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**: 1422-1423.

Cooper TA, Wan L, Dreyfuss G. 2009. RNA and disease. *Cell* **136**: 777-793.

Decio P, Ustaoglu P, Roat TC, Malaspina O, Devaud JM, Stöger R, Soller M. 2019 Acute Thiamethoxam toxicity in honeybees is not enhanced by common fungicide and herbicide and lacks stress-induced changes in mRNA splicing. BioRxi doi: 10.1101/641407.

Dong Y, Taylor HE, Dimopoulos G. 2006. AgDscam, a hypervariable immunoglobulin domain-containing receptor of the Anopheles gambiae innate immune system. *PLoS Biol* **4**: e229.

Eldridge AG, Li Y, Sharp PA, Blencowe BJ. 1999. The SRm160/300 splicing coactivator is required for exon-enhancer function. *Proc Natl Acad Sci U S A* **96**: 6125-6130.

Fan YJ, Gittis AH, Juge F, Qiu C, Xu YZ, Rabinow L. 2014. Multifunctional RNA processing protein SRm160 induces apoptosis and regulates eye and genital development in Drosophila. *Genetics* **197**: 1251-1265.

Fiszbein A, Kornblihtt AR. 2017. Alternative splicing switches: Important players in cell differentiation. *Bioessays* **39**.

Fu XD, Ares M, Jr. 2014. Context-dependent control of alternative splicing by RNA-binding proteins. *Nat Rev Genet* **15**: 689-701.

Gabut M, Dejardin J, Tazi J, Soret J. 2007. The SR family proteins B52 and dASF/SF2 modulate development of the Drosophila visual system by regulating specific RNA targets. *Mol Cell Biol* **27**: 3087-3097.

Gautam A, Grainger RJ, Vilardell J, Barrass JD, Beggs JD. 2015. Cwc21p promotes the second step conformation of the spliceosome and modulates 3' splice site selection. *Nucleic Acids Res* **43**: 3309-3317.

Geuens T, Bouhy D, Timmerman V. 2016. The hnRNP family: insights into their role in health and disease. *Hum Genet* **135**: 851-867.

Gibson DG, Young L, Chuang RY, Venter JC, Hutchison CA, 3rd, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat Methods* **6**: 343-345.

Grainger RJ, Barrass JD, Jacquier A, Rain JC, Beggs JD. 2009. Physical and genetic interactions of yeast Cwc21p, an ortholog of human SRm300/SRRM2, suggest a role at the catalytic center of the spliceosome. *Rna* **15**: 2161-2173.

Graveley BR. 2005. Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures. *Cell* **123**: 65-73.

Guruharsha KG, Rual JF, Zhai B, Mintseris J, Vaidya P, Vaidya N, Beekman C, Wong C, Rhee DY, Cenaj O et al. 2011. A protein complex network of Drosophila melanogaster. *Cell* **147**: 690-703.

Haussmann IU, Hemani Y, Wijesekera T, Dauwalder B, Soller M. 2013. Multiple pathways mediate the sex-peptide-regulated switch in female Drosophila reproductive behaviours. *Proc Biol Sci* **280**: 20131938.

Haussmann IU, Li M, Soller M. 2011. ELAV-mediated 3'-end processing of ewg transcripts is evolutionarily conserved despite sequence degeneration of the ELAV-binding site. *Genetics* **189**: 97-107.

Haussmann IU, Ustaoglu P, Brauer U, Hemani Y, Dix TC, Soller M. 2019. Plasmid-based gap-repair recombineered transgenes reveal a central role for introns in mutually exclusive alternative splicing in Down Syndrome Cell Adhesion Molecule exon 4. *Nucleic Acids Res* **47**: 1389-1403.

Haussmann IU, White K, Soller M. 2008. Erect wing regulates synaptic growth in Drosophila by integration of multiple signaling pathways. *Genome Biol* **9**: R73.

Hemani Y, Soller M. 2012. Mechanisms of Drosophila Dscam mutually exclusive splicing regulation. *Biochem Soc Trans* **40**: 804-809.

House AE, Lynch KW. 2006. An exonic splicing silencer represses spliceosome assembly after ATP-dependent exon recognition. *Nat Struct Mol Biol* **13**: 937-944.

Huelga SC, Vu AQ, Arnold JD, Liang TY, Liu PP, Yan BY, Donohue JP, Shiue L, Hoon S, Brenner S et al. 2012. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell Rep* **1**: 167-178.

Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallieres M, Tapial J, Raj B, O'Hanlon D et al. 2014. A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**: 1511-1523.

Kent WJ. 2002. BLAT--the BLAST-like alignment tool. *Genome Res* **12**: 656-664.

Korhonen J, Martinmäki P, Pizzi C, Rastas P, Ukkonen E. 2009. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics* **25**: 3181-3182.

Koushika SP, Soller M, DeSimone SM, Daub DM, White K. 1999. Differential and inefficient splicing of a broadly expressed Drosophila erect wing transcript results in tissue-specific enrichment of the vital EWG protein isoform. *Mol Cell Biol* **19**: 3998-4007.

Labourier E, Bourbon HM, Gallouzi IE, Fostier M, Allemand E, Tazi J. 1999. Antagonism between RSF1 and SR proteins for both splice-site recognition in vitro and Drosophila development. *Genes Dev* **13**: 740-753.

Liang XH, Haritan A, Uliel S, Michaeli S. 2003. trans and cis splicing in trypanosomatids: mechanism, factors, and regulation. *Eukaryot Cell* **2**: 830-840.

Lim LP, Burge CB. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**: 11193-11198.

Long JC, Caceres JF. 2009. The SR protein family of splicing factors: master regulators of gene expression. *Biochem J* **417**: 15-27.

Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**: 26.

Luhrmann R, Stark H. 2009. Structural mapping of spliceosomes by electron microscopy. *Curr Opin Struct Biol* **19**: 96-102.

May GE, Olson S, McManus CJ, Graveley BR. 2011. Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster. *RNA* **17**: 222-229.

Neves G, Zucker J, Daly M, Chess A. 2004. Stochastic yet biased expression of multiple Dscam splice variants by individual cells. *Nat Genet* **36**: 240-246.

Nilsen TW, Graveley BR. 2010. Expansion of the eukaryotic proteome by alternative splicing. *Nature* **463**: 457-463.

Olson S, Blanchette M, Park J, Savva Y, Yeo GW, Yeakley JM, Rio DC, Graveley BR. 2007. A regulator of Dscam mutually exclusive splicing fidelity. *Nat Struct Mol Biol*.

Pandit S, Zhou Y, Shiue L, Coutinho-Mansfield G, Li H, Qiu J, Huang J, Yeo GW, Ares M, Jr., Fu XD. 2013. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Mol Cell* **50**: 223-235.

Park JW, Parisky K, Celotto AM, Reenan RA, Graveley BR. 2004. Identification of alternative splicing regulators by RNA interference in Drosophila. *Proc Natl Acad Sci U S A* **101**: 15974-15979.

Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, Dixon JE, Zipursky SL. 2000. Drosophila Dscam is an axon guidance receptor exhibiting extraordinary molecular diversity. *Cell* **101**: 671-684.

Shen H, Kan JL, Green MR. 2004. Arginine-serine-rich domains bound at splicing enhancers contact the branchpoint to promote prespliceosome assembly. *Mol Cell* **13**: 367-376.

Soller M. 2006. Pre-messenger RNA processing and its regulation: a genomic perspective. *Cell Mol Life Sci* **63**: 796-819.

Sun W, You X, Gogol-Doring A, He H, Kise Y, Sohn M, Chen T, Klebes A, Schmucker D, Chen W. 2013. Ultra-deep profiling of alternatively spliced Drosophila Dscam isoforms by circularization-assisted multi-segment sequencing. *The EMBO journal* **32**: 2029-2038.

Szymczyna BR, Bowman J, McCracken S, Pineda-Lucena A, Lu Y, Cox B, Lambermon M, Graveley BR, Arrowsmith CH, Blencowe BJ. 2003. Structure and function of the PWI motif: a novel nucleic acid-binding domain that facilitates pre-mRNA processing. *Genes Dev* **17**: 461-475.

Torres-Mendez A, Bonnal S, Marquez Y, Roth J, Iglesias M, Permanyer J, Almudi I, O'Hanlon D, Guitart T, Soller M et al. 2019. A novel protein domain in an ancestral splicing factor drove the evolution of neural microexons. *Nat Ecol Evol*.

Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470-476.

Wang Z, Xiao X, Van Nostrand E, Burge CB. 2006. General and specific functions of exonic splicing silencers in splicing control. *Mol Cell* **23**: 61-70.

Watson FL, Puttmann-Holgado R, Thomas F, Lamar DL, Hughes M, Kondo M, Rebel VI, Schmucker D. 2005. Extensive diversity of Ig-superfamily proteins in the immune system of insects. *Science* **309**: 1874-1878.

Yang Y, Zhan L, Zhang W, Sun F, Wang W, Tian N, Bi J, Wang H, Shi D, Jiang Y et al. 2011. RNA secondary structure in mutually exclusive splicing. *Nat Struct Mol Biol* **18**: 159-168.

Zaharieva E, Chipman JK, Soller M. 2012. Alternative splicing interference by xenobiotics. *Toxicology* **296**: 1-12.

Zaharieva E, Haussmann IU, Brauer U, Soller M. 2015. Concentration and localization of co-expressed ELAV/Hu proteins control specificity of mRNA processing. *Mol Cell Biol* **35**: 3104-3115.

Figure legends

Figure 1. Analysis of *Dscam* exon 9 alternative splicing by restriction digestion of PCR products. (A) Schematic of the *Dscam* exon 9 variable cluster gene region. Constitutive exons are shown in orange and variable exons in blue. Primers to amplify the variable part are shown on top of the exons. (B) RT-PCR product for variable exon cluster 9 shown on a 3% agarose gel. (C) Schematic of the method used to resolve inclusion levels of variable exons using a ^{32}P labelled forward primer in combination with a set of restriction enzymes followed by separation of a denaturing polyacrylamide gel. (D)Denaturing acrylamide gel (6%) showing a restriction digest (SacII, ClaI, PshAI, HaeIII, MseI, BsrI, BstNI) of *Dscam* exon 9 variables amplified with a ^{32}P labelled forward primer from 14-18 h *Drosophila* embryos. Single enzyme reference digests are shown on the left (lanes 2-8) and the combination of all enzymes on the right (lanes 9-11).

Figure 2. Protein domain structure of *Drosophila* SR proteins and phenotype of loss of function (LOF) and gain of function (GOF) mutants. (A) Evolutionary relationship of *Drosophila* SR proteins with human homologues is shown on the left and the domain structure is indicated by coloured boxes. Arginine-serine-rich domain (RS, green), RNA recognition domain (RRM, light blue), hnRNP K homology domain (KH, purple), zinc finger domain (ochre), Proline-Tryptophan-Isoleucine domain (PWI, yellow) and CWF21 domain (red). The type of allele obtained, and the loss (LOF) and gain of function (GOF) phenotypes are indicated on the right. (B) Viability was determined from stocks that harbour a zygotically expressing GFP marked balancer chromosome, which contains a set of inversions to supress recombination and a recessive lethal mutation. If a gene is essential, only heterozygous flies will survive.

Homozygous mutant embryos were identified in the progeny of these stocks, by the lack of GFP and advanced development as homozygous balancer carrying embryos die early before GFP expression. (C) To obtain embryos overexpressing SR proteins flies carrying the yeast GAL4 transcription factor under the control of the pan-neuronal *elav* promoter were crossed with lines harbouring SR proteins under the control of the yeast *UAS* promoter.

Figure 3. Analysis of *Dscam* exon 9 alternative splicing in canonical SR protein loss of function (LOF) and gain of function (GOF) mutants. (A, C) Denaturing acrylamide gel showing restriction digests of *Dscam* exon 9 variables amplified with a ^{32}P labelled forward primer from 14-18 h *Drosophila* embryos for canonical SR protein LOF (A) and *elavGAL4 UAS* GOF mutants (C). Quantification of inclusion levels are shown as means with standard error from three experiments for canonical SR protein LOF (B) and GOF mutants (D). Prominent changes in inclusion levels in mutants compared to wild type are marked with red letters for a decrease and in blue for an increase, and statistically significant differences are indicated by asterisks (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$).

Figure 4. Analysis of *Dscam* exon 9 alternative splicing in non-canonical SR proteins Srrm1 and Srrm234. (A) Denaturing acrylamide gel showing restriction digests of *Dscam* exon 9 variables amplified with a ^{32}P labelled forward primer from 14-18 h *Drosophila* embryos for Srrm1 and Srrm234 protein LOF mutants. Quantification of inclusion levels are shown as means with standard error from three experiments (B). Red arrows point towards exons with reduced inclusion levels in the *Srrm234*$^{\Delta N}$ mutant compared to wild type. Prominent changes in inclusion levels are marked with red letters for a decrease and in blue for an increase in the *Srrm234*$^{\Delta N}$

mutant compared to wild type. Statistically significant differences are indicated by asterisks (***p≤ 0.001, **p≤ 0.01, *p≤ 0.05).

Figure 5. Protein domain structure of *Drosophila* hnRNP proteins and phenotype of loss of function (LOF) and gain of function (GOF) mutants. Evolutionary relationship of *Drosophila* hnRNP proteins with human homologues is shown on the left and the domain structure is indicated by coloured boxes. RNA recognition domain (RRM, light blue) and hnRNP K homology domain (KH, purple). The type of allele obtained, and the loss (LOF) and gain of function (GOF) phenotypes are indicated on the right.

Figure 6. Analysis of *Dscam* exon 9 alternative splicing in general hnRNP protein loss of function (LOF) and gain of function (GOF) mutants. (A, C) Denaturing acrylamide gel showing restriction digests of *Dscam* exon 9 variables amplified with a ^{32}P labelled forward primer from 14-18 h *Drosophila* embryos for general hnRNP protein LOF (A) and *elavGAL4 UAS* GOF mutants (C). Quantification of inclusion levels are shown as means with standard error from three experiments for canonical SR protein LOF (B) and GOF mutants (D). Prominent changes in inclusion levels are marked with red letters for a decrease and in blue for an increase, and statistically significant differences are indicated by asterisks (***p≤ 0.001, **p≤ 0.01, *p≤ 0.05).

Figure 7. Analysis of *Dscam* exon 9 alternative splicing in double and triple mutants for general splicing regulators. (A) Schematic of the genomic region of *Hrp36* and *B52* genes. Transposon inserts are shown with triangles and the deletion Hrp36/B52$^{\Delta 1}$ is indicated at the

bottom. (B) Denaturing acrylamide gel showing restriction digests of *Dscam* exon 9 variables amplified with a ^{32}P labelled forward primer from 14-18 h *Drosophila* embryos for Hrp36^{BG02743}, Hrp36/B52$^{\Delta 1}$, Hrp38^{d05172} and Hrp36/B52$^{\Delta 1}$ Hrp38^{d05172} mutants. (C) Quantification of inclusion levels are shown as means with standard error from three experiments. Prominent changes in inclusion levels are marked with red letters for a decrease and in blue for an increase, statistically significant differences are indicated by asterisks (***$p \leq 0.001$, **$p \leq 0.01$, *$p \leq 0.05$).

Figure 1
Ustaoglu et al

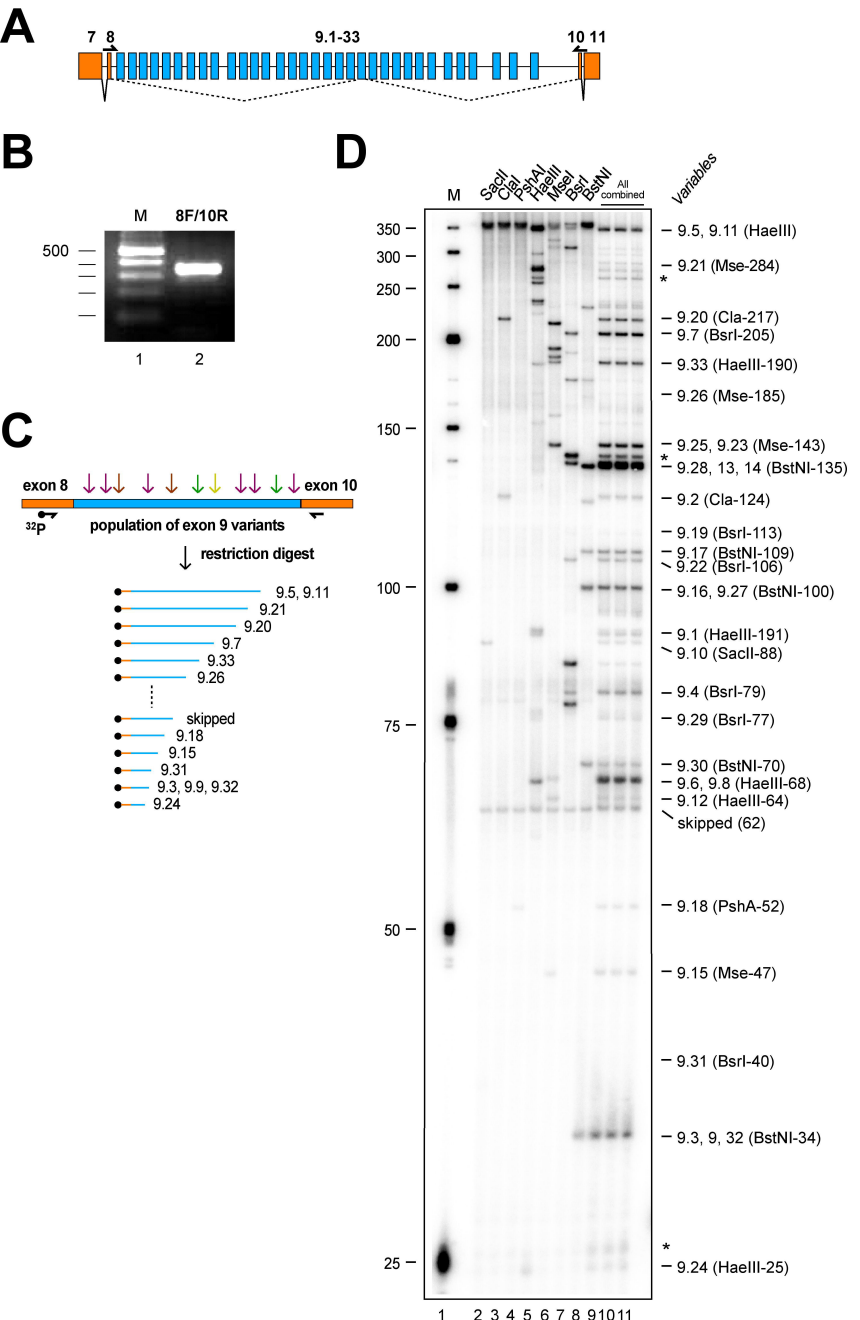


Figure 2
Ustaoglu et al

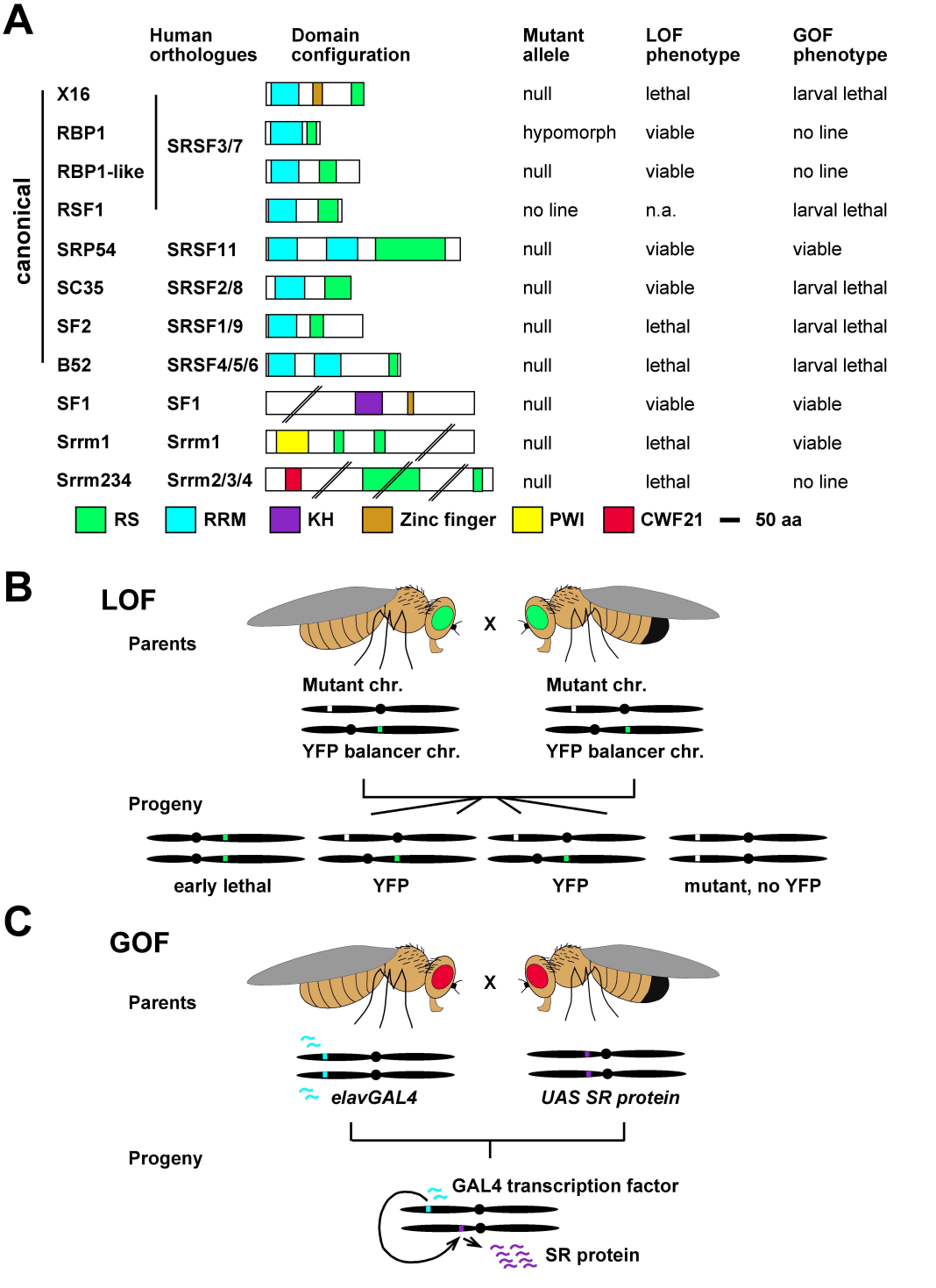


Figure 3
Ustaoglu et al

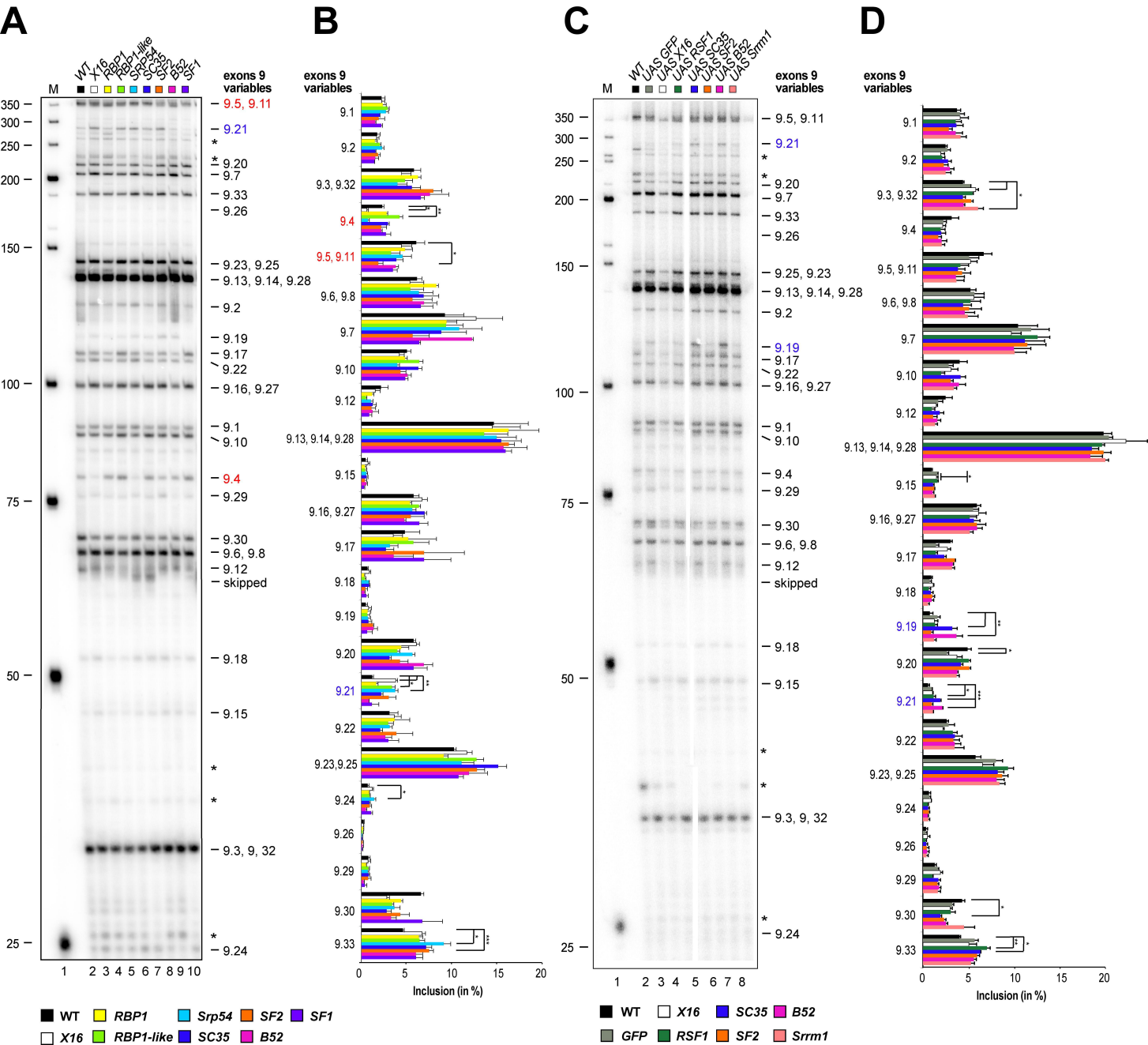


Figure 4
Ustaoglu et al

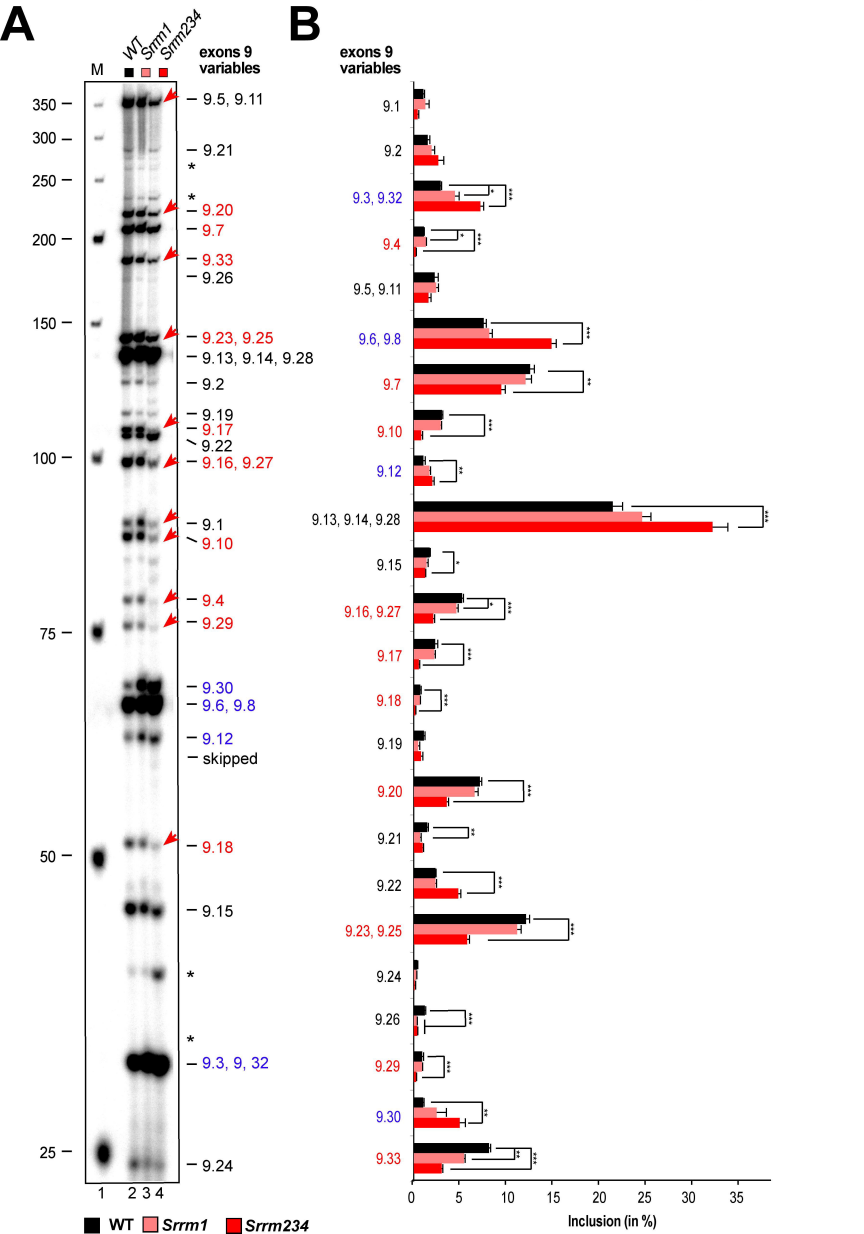


Figure 5
Ustaoglu et al

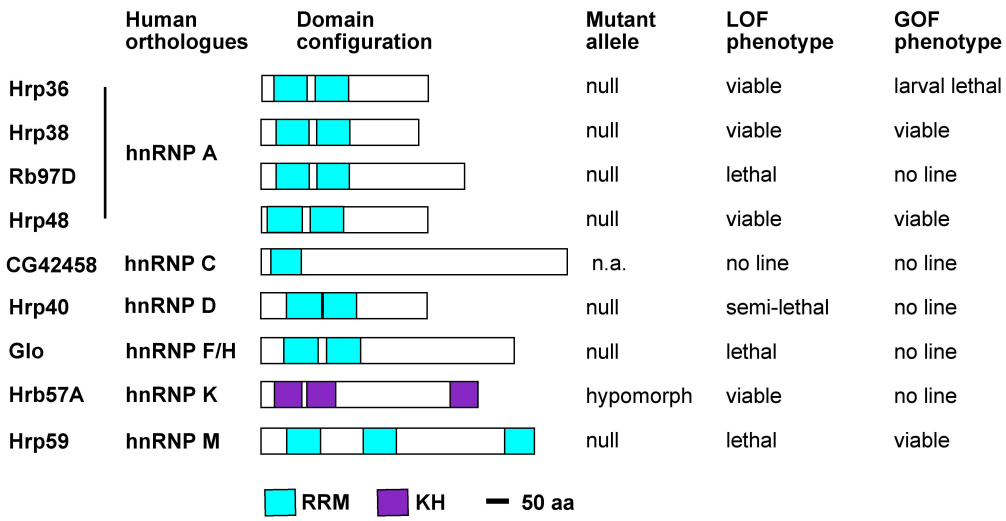


Figure 6
Ustaoglu et al

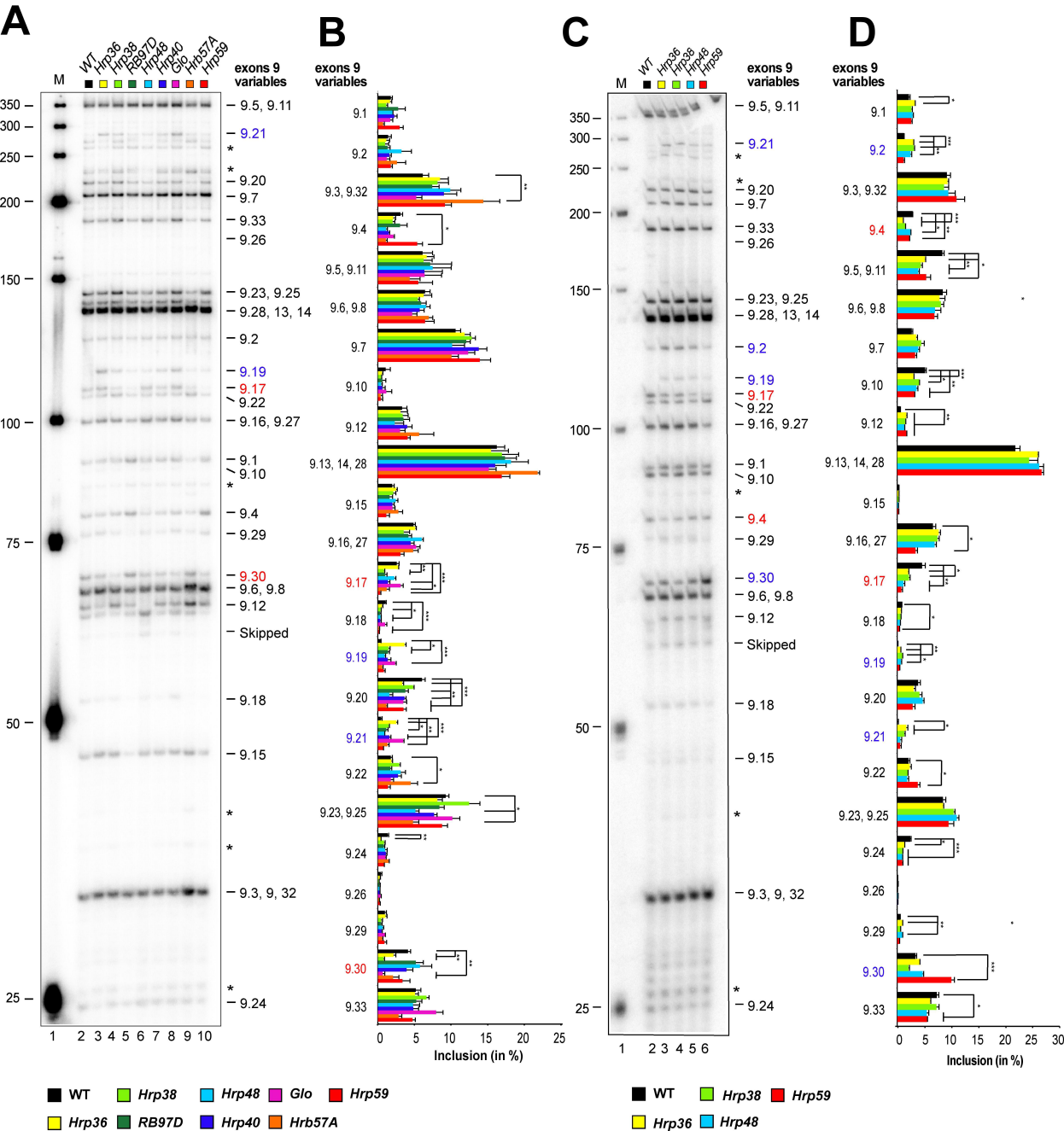
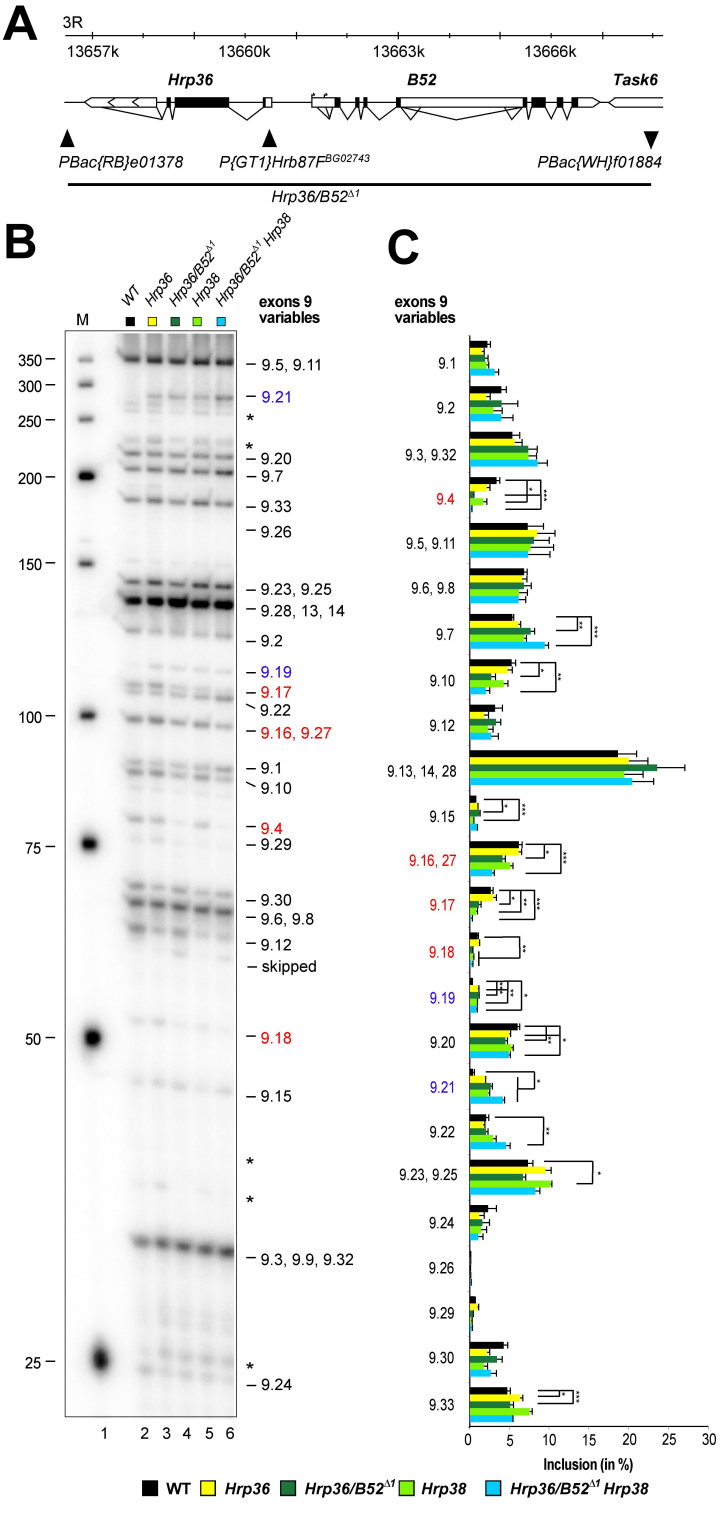


Figure 7
Ustaoglu et al



Srrm234, but not canonical SR and hnRNP proteins drive inclusion of *Dscam* exon 9 variable exons

PINAR USTAOGLU[1†], IRMGARD U. HAUSSMANN[1,2], HONGZHI LIAO[1‡], ANTONIO TORRES-MENDEZ[3], ROLAND ARNOLD[4], MANUEL IRIMIA[3,5,6] AND MATTHIAS SOLLER[1,7]

Figure legends for Supplementary Figures

Supplementary Figure S1. Transposon inserts in SR protein genes. The genomic locus is depicted for the genes where transposon insertion alleles were analysed. The gene span is shown on top as blue box and below transcribed parts are shown as boxes for exons and lines for introns. Transposon insertions are shown as green triangles. Note that transposon have a size of ~10 kb and disrupt gene expression.

Supplementary Figure S2. Expression of *Dscam* (A) and *Srrm234* (B) in the nervous system of *Drosophila* late stage embryos as determined by RNA in situ hybridization.

Supplementary Figure S3. Transposon inserts in hnRNP protein genes. Transposon inserts in SR protein genes. The genomic locus is depicted for the genes where transposon insertion alleles were analysed. The gene span is shown on top as blue box and below transcribed parts are shown as boxes for exons and lines for introns. Transposon insertions are shown as green triangles. Note that transposon have a size of ~10 kb and disrupt gene expression.

Supplementary Figure S4. Hrp36, Hrp38, Hrp48 and Hrp40 SELEX motif occurrence is shown for the *Dscam* variable cluster 9 with the line indicating maximal amount of significant hits ($p<0.05$) of the SELEX motif in a sliding window (12 for Hrp36, 12 for Hrp38, 8 for Hrp48 and 9 for Hrp40).

Hrp36, Hrp38, Hrp48 and Hrp40 CLIP binding data from Blanchette et al. (2009) is shown for the *Dscam* variable cluster 9 with top and bottom lines indicating the binding score per tiling oligo probe (±0.54 for Hrp 36 and Hrp38, and ±49 for Hrp48 and Hrp40) compared to averaged binding score for all tiling oligos as determined by TiMAT.

Changes in exon inclusion as $\log_2$ fold change in LOF and GOF conditions are shown at the bottom.

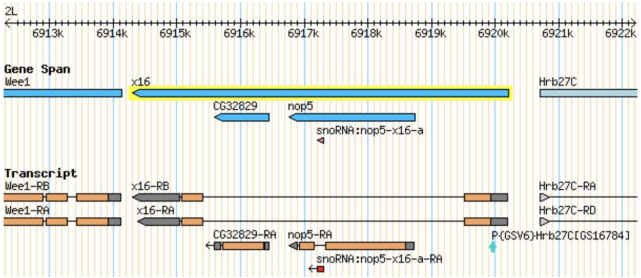
Supplementary Figure S5. Vista plot alignment of *Dscam* exons 7 to 11 of the *D. melanogaster* sequence compared to the *D. virilis* sequence. Exons are shown as boxes on top and indicated in pink on the alignment. Constant exons are shown in orange and variable exons in light blue. Dark blue exons are absent in *D. virilis*, and exons only present in *D. virilis* are shon on top.

Supplementary Figure S6. Alignment of the short introns between variable exons from the exon 9 cluster of *D. melanogaster* (A) and *D. virilis* (D), and analysis for sequences complementary to the conserved sequence from intron 9.33 termed docking site (B, Yang et al, 2011). Longer introns with postulated selector sequences are shown at the bottom of the alignment. Sequences with base-pairing capacity to the intron 9.33 sequence termed docking site are shown for *D. melanogaster* (C) and *D. virilis* (E). Note that sequences with base-pairing capacity to the intron 9.33 sequence termed docking site are absent from most introns.

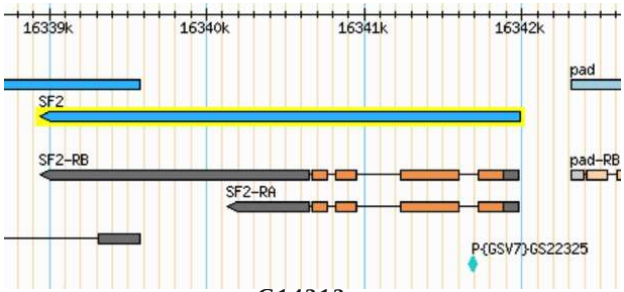
Supplementary Figure S7. Comparison of long-range base pairing potential of the docking site compared to 100 randomly chosen *Dscam* intronic sequences of the same length. (A) The distribution density of normalised alignment scores (filtered to have at maximum two gaps) against reverse and reverse complement sequences of the same length (docking: alignment scores for the docking sequence, randomly selected: 100 other sequences, see Material and Methods). (B) The same data, but showing the amount of sequences that pass the gap filtering threshold in both sets. (C) Density distribution of computed optimal secondary structures for each sequence. Vertical lines indicate the result for the docking sequence against the reverse and reverse-complement concatenated intronic sequences.

Supplementary Figure S1

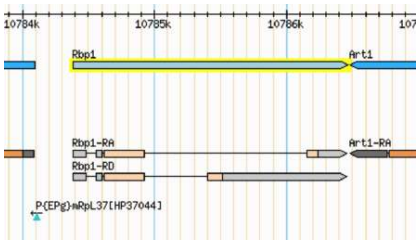
X16: P{GSV6}Hrb27C^{GS16784}



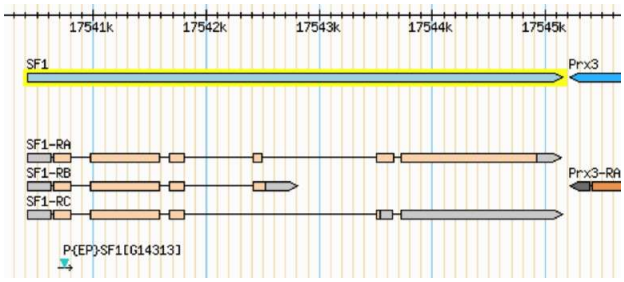
SF2: P{GSV7}SF2^{GS22325}



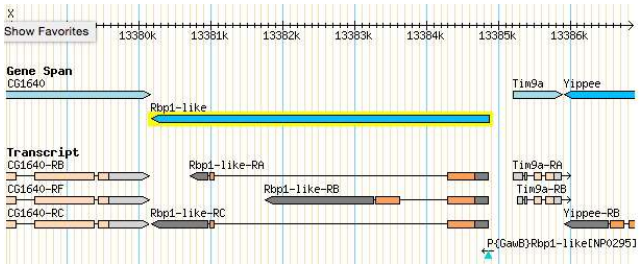
RBP1: P{EP}mRpL^{HP37044}



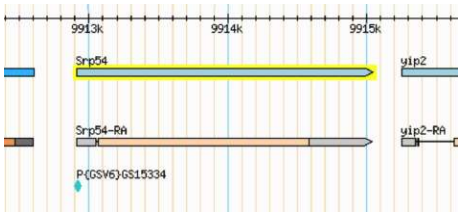
SF1: P{EP}SF1^{G14313}



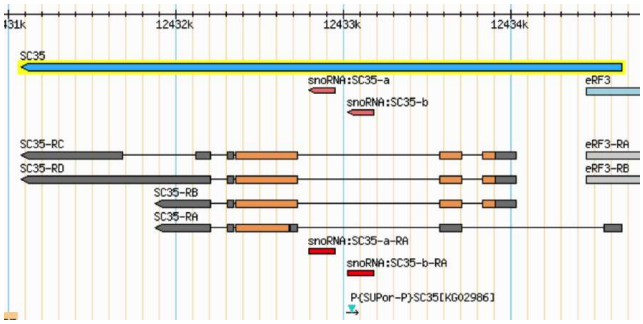
RBP1-like: P{GawB}RBP1-like^{NP0295}



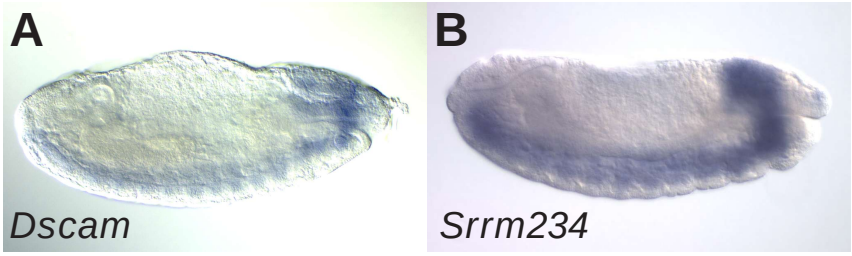
SRP54: P{GSV6}Srp54^{GS15334}



SC35: P{SUPor-P}SC35^{KG02986}

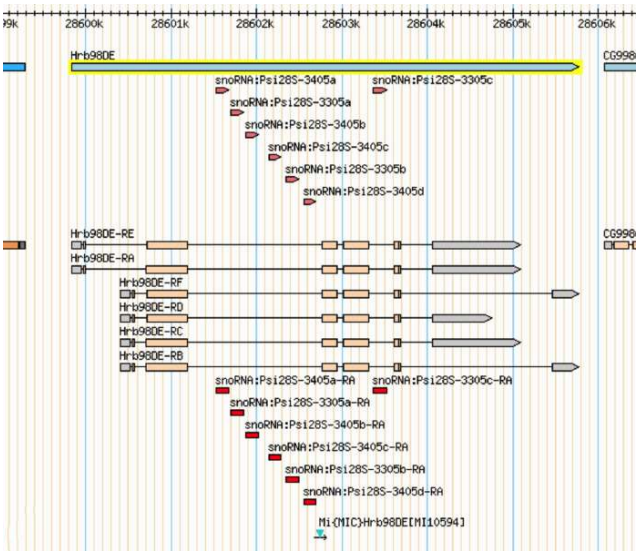


Supplementary Figure S2

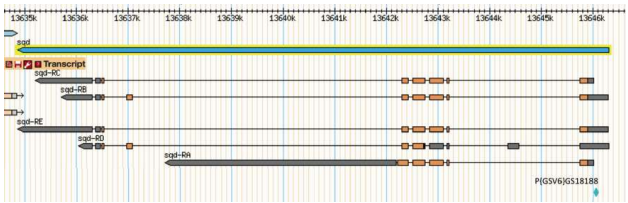


Supplementary Figure S3

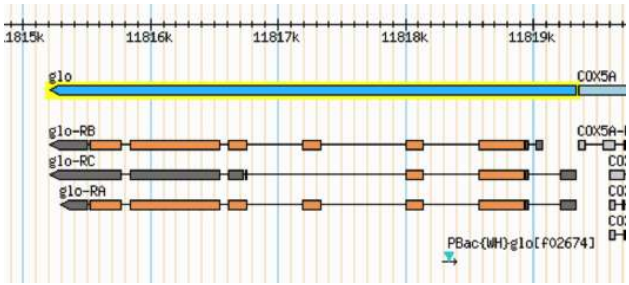
Hrp38: Mi{MIC}Hrb98DE^{MI10594}



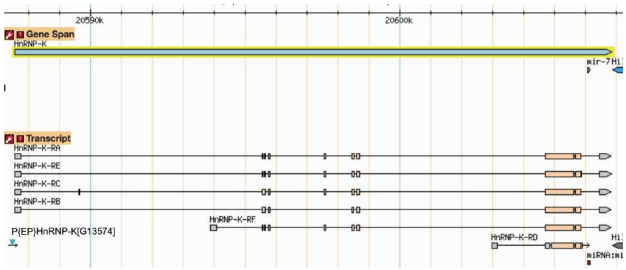
Hrp40: P{GSV6}sqd^{GS18188}



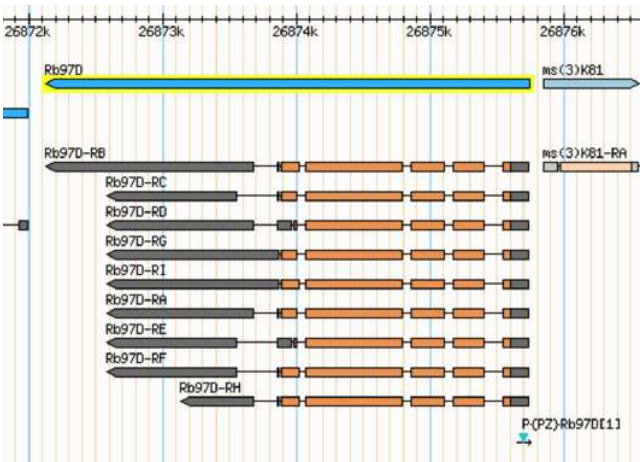
Glo: PBac{WH}glo^{f02674}



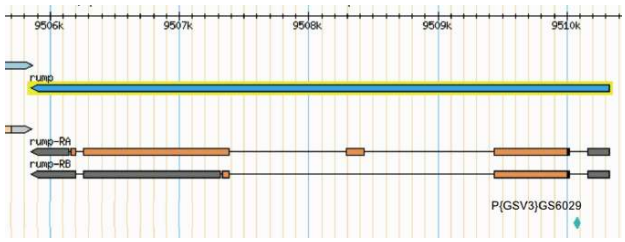
Hrb57A: P{EP}HnRNP-K^{G13574}



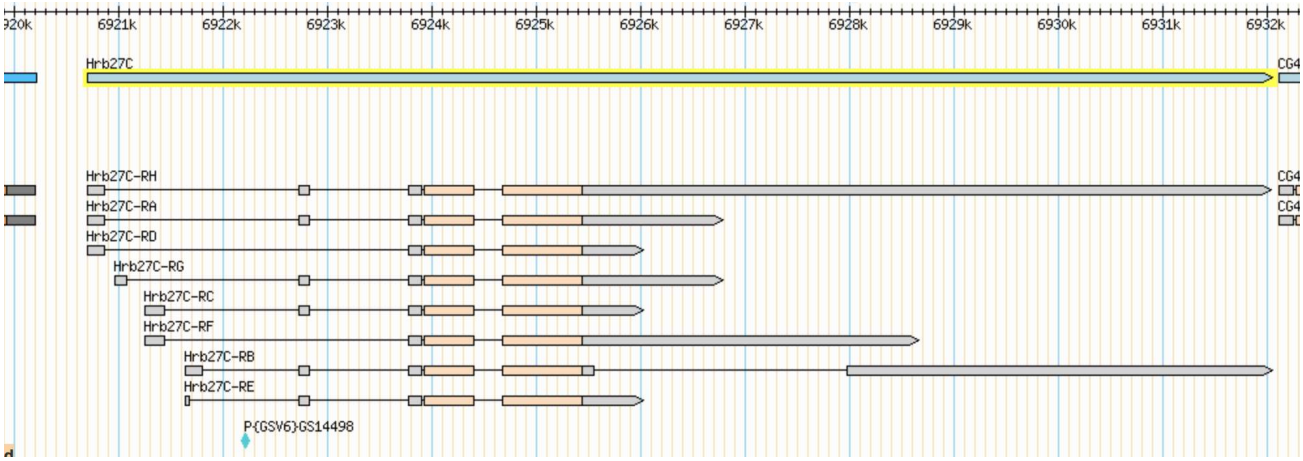
RB97d: P{PZ}Rb97D^{1}



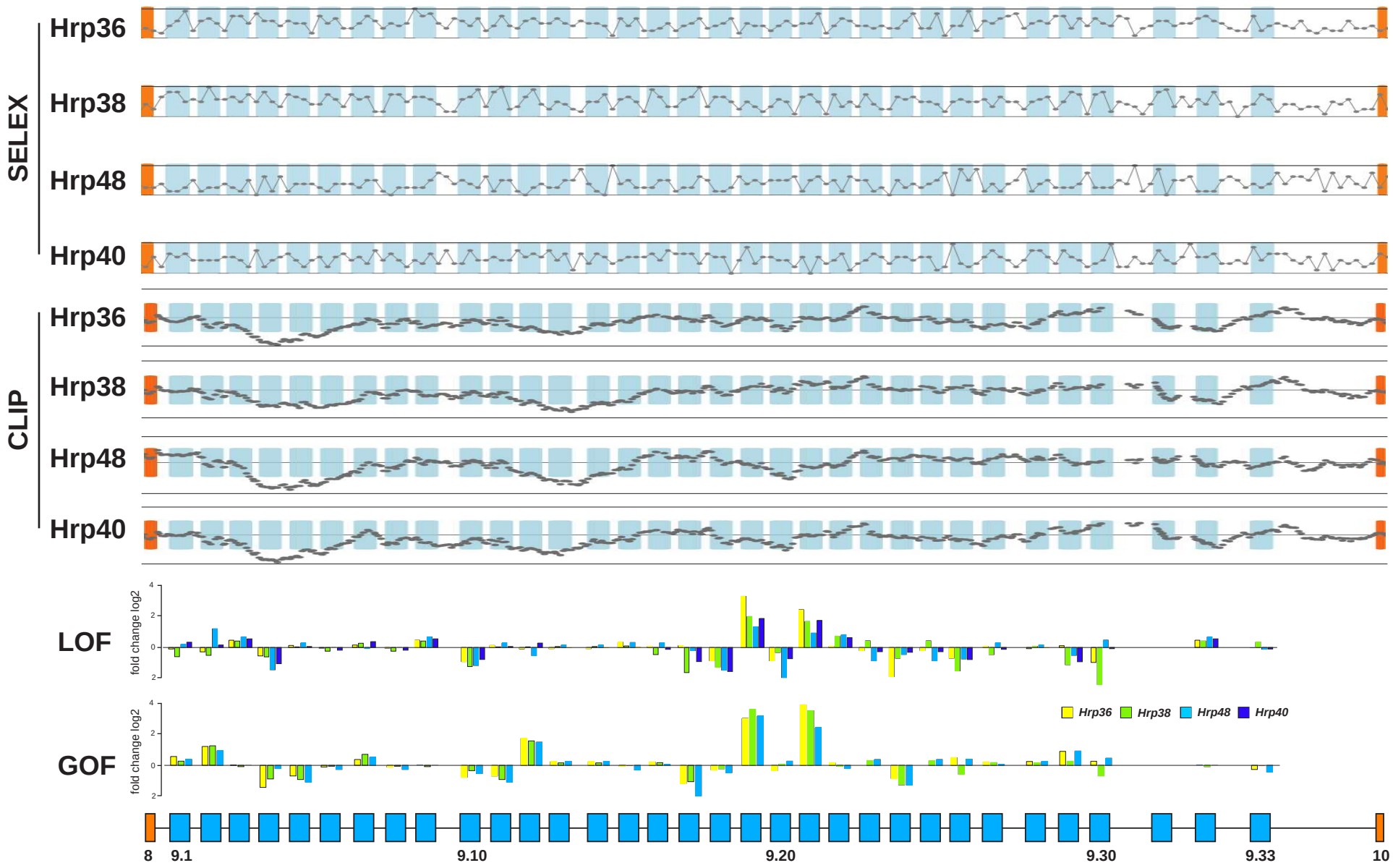
Hrp59: P{GSV3}rump^{GS6029}



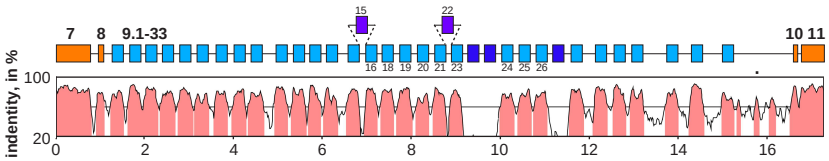
Hrp48: P{GSV6}Hrb27C^{GS14498}



Supplementary Figure S4



Supplementary Figure S5



Supplementary Figure S6

A *D. melanogaster*

```
                10        20        30        40        50        60        70        80        90       100       110       120       130
int9.1   GTT-----------CGCTTAGGCCTTC--TTCGATTCG--TAGATTGAT---TAGTTCCTTGGTT------TTTGAGTT--CTGTTTTCTTGACCCTAC----TCGCATCCTCGCCCTCCTCCGC------AG  97
int9.2   GTTA--A-----TAGCATCGAATAGCC--ACATGATCGTTTTTGATTTT---C----CCTTGGTT----CTC-----TT--CTGTT------AACCCAACCCCATC-C----------------------CAG  77
int9.3   GTTA-----------------TTGTCC--A---GCAGATACGGT-------------GTCAATTTCTTAAGTTTTCCTT--GCC-----ATC----TTTTGCTTT-TTGCCT-TCCCATCCCGGCTCAACCAG  85
int9.4   GTTA--A-----TCCCTGTTGATCTCT--TAAGATTTGCTTTTGTTTCT---TAC-ATTTTCGTTATGACCC-----TTT-TCTTT-----ACCCTCACTCCATT-CGTTCCCTC-CAATGCCAT----TCAG 102
int9.5   GT----A-----C--ATGGAC----CC--AGGGACCTCACTCTAAGCTC---CTTTATTATTATTTGCTTATCATAATT---TTATTTACTTTAATTTT-ATTTACACCTTTCGAAAT-------CTTT---AG 100
int9.6   GTTGGTGAAACTCCCTTTTGGATCTCTCTTTATATCCGCCTGTAAAGTTGCTTCCTACTAACACGATCCTGGTTTATTTGGTGGCTAAGAGCTTTGTATATGTCTCGAGCTTTAAATGGATTTCCCTTCATAG 133
int9.7   GT----T-----T--GTTTGA----CC--TTA-TCACTATTTGTAGTGC---TTTAAGTTTTGTATTTGATTTCATATT--TTACTTACTTTACTTTTGATTCTCTCCACCTCATCTC----CACTTCTAAAG 106
int9.8   GTCATTA-----TGCAAATCAT-ACTC--ATA---------------GT---TAAACGT----TTTTCTATTTTAAATT--ATTT----ATTTGAGTTTTAATTTACCACTTATTTCCTGTCCATT---TAAG  94
int9.10  GTTAATA-----TCTGAACAAGGA----------------TATGGGACT---TA--AGTTATATCTTTATTTCT---TC--TACCCCCTCCTCTCCCCTACTCCTATCCTGCCATTTATGGTC-------CAG  95
int9.11  GT---------------TAC-TACCC--ATAGATTT-CCTATA-TATA---TAT-ATTTTGGTTTACTTTTCGAA-----TGATTTTAA--AC----GAT-------ACCCCAC---AT-CCATCCCA---G  83
int9.12  GT-T--A-----CCCAT--TAATA----------TTC---------GTT----AACATTAATGTTCTGTT----GA--T--TGCGATCATTTTATACCCATCACCGT--CCCAAT-C-AC-TTCCC--A---G  82
int9.14  GTTA--G-----ACAAACGAAGAGATC--GCCTTTAGACCTAAGAATAT---TAC-ACATCATTTATGATTTTGT-GTT--TACTTACGTTGATTTGA-------GTACCCTTAC-CCCTACCTTATC-TCAG 108
int9.15  GTTA--A-----TACTCTTGAAATATT---TAAATTGTCCTTTTATTTT---T--TTTCAT-TTTCGTTTT-----ATC--TCC--TTAACCCCCAAAAAACA--------CCCT-TGATA----------AG  89
int9.16  GT----A-----C----TT----------TCA-ACAGCTTGACAATCAG---TTTTAATGTACACCATGATTTTG-ATT--TGATTTA-TTTATTCCACACTCCCACTCTTCCACATT----C---CCCGAAG  95
int9.17  GT-------ATTGTGGGGGGAATAT-------GGTATTGTGGTATTTGTGGAATATCACAGAGAGCTACGTACGTTACTCG----TTCGAATATAATTTGAATCCCTCCACATCCTTTAATTCC---CATTAAG 112
int9.18  GTTGGCA-----GGGAAACCCTTGTCC--ATA--------------AGT---TATATAA----TTGTTTGGTTTTTTTT--TTTTGAATATCGGACTTTTATTTTTACCTATTCCCTCATTCCTCTTA-ACAAG 102
int9.19  GT-------------ATTAGA----TC--AT-TTAAAATTAGTG-------TTTCGATAA-------ATATTTC--CC--TTA---ATTTTTCATATTAAATTGACTTTAATTTACC----TACCCA---AG  84
int9.20  GTAT--A-----TCGGTTAAGGGTTTT-ATCAACTAATGTTAAGTCTT---T--CGTTTTATTTAGTAAG-----ATT--TTCGTTCAATTTTACGTCTTTA--------CAAC-CCA-A----------AG  92
int9.21  GT----A-----CCCGC--AAATG----------TTTA------GAGGA----AAACTTCAAATTAGGAT----TA--T--TGTTTTTAAGTTCCTTTAATTATTAT--TACACT-CCAA-TCCCTTTA---G  87
int9.22  GT---------------ACACTTAGTT--G-AAGTATACATAATT-AGT---TGGGC--------TTTTGGATTAGCTT--ATTGTTTTGT-GTATTCTTGCGTATTTATCAACATTTTGATTAACCCAT-AG  99
int9.23  GTAC--A-----CTTGTGTTACTAG--------ATTTA---ATGGAGTT---TAGAGTTTATTTTCCTTA----AAGTT--TACTTTCTGTTCTCCCCCCCCCCCATATTGCACT-CAAT-TCACACAAA-AG 103
int9.24  GT-----------CATTAGA----TT--ATGCTTATA-----AAAATC---CGTCGCTGTT----TACCACTTTCATT--TAAGA---TTATCGTTT-------ATTTCCTCCCACTTAAAC-TGAATAAAG  90
int9.25  GT----A-----C--ACCT----------ACA-ATTTTCTTAAATTCAT---TTTTAGTGTAAGGTACATCTTTA-ATT--TTATGTA-ATT-TTTATTAAAACCACC--TTCAC----------TCCAAAG  90
int9.26  GTAC--T-----GATGGATACGTTCCC--ACCTATGTACCTACAATATT---TATTATTATCGTTTTTATTTCATAGTT--TTGTTTTGAGAACCCAAGATTATTGTAGCCTTCC-CAAT-CTATGTCACTAG 117
int9.28  GT-GGGA-----TAGAGATCCACATCC--ATATGCATATACCGTAGAGA---TAGAGATCACCTTATTTGTCTTCATTT--TTTCCTTGATTGTTCGTTGAGTTGACTGCTTTTGTTTTGTGTATCCCACAAG 120
int9.29  GT----T-----CTCATCCCA----GC--GTCTTAGTATTAAGAGAAGC---TTTCGTTGATGCGGTTATGTTTTAATT--TTATGCACCTTTTGTTTTACCATCACATCACTTTACCGAAACGCACATCCAG 113
```

```
int9.9
GTTTGTCCCCACAAGAAATATTGAATAAGGAGAAATAAGAGCACGCTTACTGTCTATGTTGTTTTCACTTAGTATTGCTTGAGGTGTTTGGTGTTGGAGCAGTCGCATTCATATGATATTTGTGCTGAATGTCATATAAATC
AGAAAAATTAGGTGTAATATGATTATTAAGACAAGCTACAATACTTTAAGAAAAAATATTCTGAGGAGCTGTTTTGTCATTTTGGTAGATAAAATACGACCGCGTTCTCTTTGCCCATTCGCATGCAG                270
int9.13
GTACTGAAATTGAGTGTGTCATGATTTAGTTTTAGTGGTTGCGGTTACTAAAGGTCTTGCTAGTTGTTGCTACTAATCAAACCCATGCCCATTTCTCTAACTTTCTATCTAACTAAGCCTGCGTGGATAACTGTGAACTTTTT
TGCATATTGCCCTTACTTATTCGAACTGCATTTTCTATCCCCCAAAAG                                                                                                191
```

B **Docking site**

Dmel GGCTCATGCTCCAACACCGAATAAACC
Dvir GGCTCATGCTCCAACACCGAATAAACT

- postulated acceptor site, Yang et al, 2011
- putative acceptor site

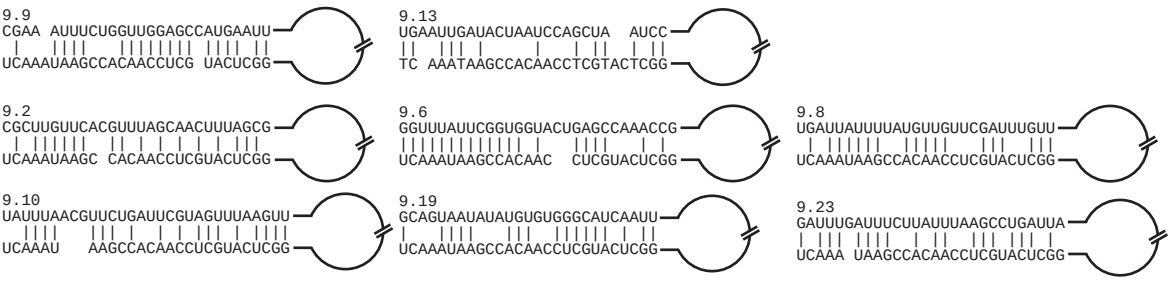
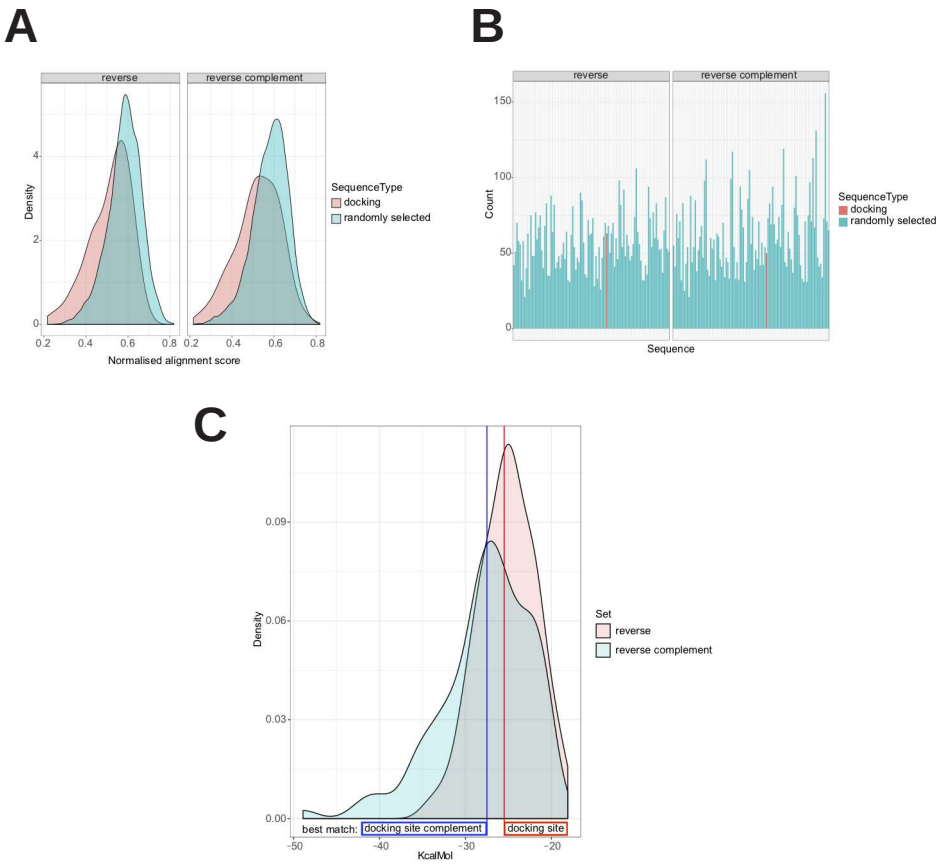
C

D *D. virilis*

```
                 10        20        30        40        50        60        70        80        90       100       110       120       130
vir int9.1  GTGATGCTTCTCTCTCTTCCACTCTCTCTCTCTCCCTCTCTCTCTCGATCCGCTTAGTAGTTGTGC-ATTGTTCTAGACCGATTGTTACGA-GTAGTCTAACGTTATGATTTGTTAAACCC--TCCAACCAATC-CAAG 134
vir int9.2  GT----------------TA--AGAGT--TTTATTAGT--TCGCTTGTTCACGT--TTAGCAACTT--TAGCGTTTCAT--TTGCTT-TGATTTTGTTCT------TCCCAAA----ACCCAATC------------AG  88
vir int9.3  GT----------------TA--ATC-C--TTTAGTCTC--ACG-TTGCTTACAT--TCTACTTTCG--TTGC-TTTTGT--TTGCTT-TTACTTGATCAAA---AATCCCAACC---ACCCATAC------------AG  89
vir int9.4  GTT---------AAGCTTTAGTTCTGCTT---G------CAGTAGCTC---CAGTTT-AATCTTTACTTTATGTTCAGTTCTGTCC-----CCCAACCCGATCCGCTAT-TATTTA--TTTG---T----------CAG  96
vir int9.5  GTAT----ATGGCAGCAT---------ATTGTGATC-CTATAACTGATATCGGCTGTTAGCTTTATA-----AATTAGTTAACCCCT-TTTAAGTTCTTCATTTATAATTTGAT-ACATT---ATT--CACA----CAG 109
vir int9.7  GTT------------CGTCA--TGTTGATAGAAATTGATTTAAGTTAT---AGT--TTATTA--TTGGCTGTTTTTATTTGATACTC-AACTCTTTTGTTGTCC-----CGACCCCCATCAAAA------------AAG 100
vir int9.8  GTTT----GTGCGGAAGTTG--TGGGACTTATTGGGATCATAACGCACCTAGCTCATATACATTTTGATTATTTTATGTTGTTCGAT-TTGTTTGATTTGTCTGAACT---AACCCCACAAATAAC--CC--CTTCAAG 125
vir int9.10 GTT------------CGAAC--TATT--ATTTAA--CGTTCTGATTCGT---AGT--TTA--A---------GTTTTTGGTTTATACCT-TGACACTTTGAT-CCC-----CACATCCCAC------------------AG  81
vir int9.11 GTG-----------ACCCTG--TGCACATTCTATTAGACAAAAATTATTAAAAT--TTCTTTTCTTGGTTTTTTTTTGTAATTCCGT-TTGTTTTGCCATGTCCACTCCCAATCCTGACCCAACTC--CT--CTGGCAG 119
vir int9.12 GTGC---ATCTAACATTTATGCAGCATTTTAGG------ATATTTTATTTGTACAATCG--TTTTGGTTCGT---AACTCCCGCCTAACACTTTACAC---TCTCTCATACATTGA-ACCCA-----GTGCTGCCGCAG 116
vir int9.14 GTTT----GTC----AGCCG--ACAGA---ATGAGCATC-TACTTAACTTAATTGATTTGC---CTGAT--TTGCAT-TTATGTTTT-TGGTTTAATTTTGTTGAATT------CCCACACATGTC--C--------AG 102
vir int9.15 GTGCGTGG-CCATAAAATTCATAGAG--CTACGAGTATAAAAA--CGCCTATTTGGAGACTCTTTCTAT------------ACCCTTGCAGTGTCGC--------------AGTTTCGC------------------AG  90
vir int9.18 GT------------ACAGAA--AGC-CTTTCTACT-GTCAA------CTGAA----TTATTTTCT--GTTTTTTTTTTTTCTACTTC-TTATTTTGT--TGCCCATTACCCGAACCAACCC--TTC--CA--AATCAAG 102
vir int9.19 GTAT----A-CGCAGTAATA--TATG---TGTGGGCATCA-ATTGCATG-ATCTATGATTTTTCTTTG---ACTTTTGCTATGATTT-TCGTTCCGTAACAC----ACAC-GATTCTACCCA-------------CAG 104
vir int9.20 GTTG----GCAAGAGCATGA--CATCTATTATATC-----TAAATCTTTTATA-AGACATCCGTTTGATT---TACGACTTAATTTT-TACTATTGTTGGATATCCCTTTA--TCCTTTCTGCTGC--TC----CCAAG 115
vir int9.21 GTACT-AAACTACGGA---------GACCTATTT-----TTAATTTATGTTTTTTACTACTATTCCGT-TG------AAAATATCTTTTTGTTATAACGC----ACATTTGAATTAT-----TTTGG--TAACCAA-AG 105
vir int9.22 GTAT----ATTGGAATATTA--AGTGGA---AATTCATTTTGA--TATTTGTAT-----TCAGTTTGGTTTGATTTGGTTAAAACTC-C--TTGCCCCCGACACATTCTTACGACACACCCGACTC--TTTGCTTAAAG 118
vir int9.23 GT-A----TTCGCATAGTTG-------ATAATGAAA-ATATATTGCATATAG--ATTTGATTTCTTA-----TTTAAGCC-TGATTA-ACATTTTGTTTTGTTGATATT--GAT-AT-----TACC--CTGA----A-G 102
vir int9.24 GTAT---ATCAAAGGTGTTAGCCTTAATTTAGG------AAATAAAACGTACAATTTCAATTTTAGTTTTATCCTAAGTTCTCCTCAAAACCCCAGACTTATTTTTAATATATTTACATTCA---TGAAATTCCTCTAG 127
vir int9.25 GTTG----TTA------------------TATAAT-----TGAATCTTTTGTC-CGTGACTTGCTGTATT---TTTGATTTTGTTTC-TAATGTTATCC-ACGCAACCTGA--CCTTAT------C--CC----ATAAG 92
vir int9.26 GTATTTGGATCATGAGAGTAGTTACGACTTAACTAGATGTTATTTGGTTCTTTCTATTTTAATTTCATATACCTCAAAAAATACCCAACTAGAGAGCCGCCTATACAGTAGTAACATGTCCCCTTGAATTCTTCACTAG 139
vir int9.28 GTTG----GCCATAGATTTG-------CTG-TAGTT-TTAGTTTTCATATAG--TAGAGCCTACATA-----TTTACGTTATCTTTT-GTGTTTTTTTTTTTGTATTTTCTGTT-GTGCTCATACC--CAAA----AAG 111
vir int9.29 GTGC----A-GCGAAGACTC--TATG---C-TAAATACCTTATTTAATT-ACGTAGAATATTTTTGCATACGCTCTCTTTTTCATTT-GCGCACTCTATTATTTCGGTTC-TTTCCCATCCAAAT------ACGTTCAG 119
```

```
int9.9
GTTTGTTTGCGCAGATAAAATACTAAATTAGTCAACATACCAAACTATGCTTAAGTAGCTAGCACTTAATTTTAACTGCTGAGTTCGAAATTTCTGGTTGGAGCCATGAATTTACTTGAAGCTTTCTAAAAAGCATTTAAGAC
TGTTGAATAGCCATGCATTGTTGTTGTGTTTTTTGTAGTTATCGAGCTGTTTTGTCTAGTTTCGAGCTTTTCTATATTTCGCCCAG                                                        229
int9.13
GTATTCTTCGCTTCGTTTCGATTCGATTAATGGGGTAGTCTAGTGTCACCTTCGGGAGTTGGTTGTGGTTACGTTTAACGAGCTAACTTCGTTGCCTAACAGACCCTGTTTGAATTGATACTAATCCAGCTAATCCAGCGTAA
TGGAAATACTTACATCTATCTACATACATACTTCTTAACCTGCCTATCTACAACTTTGCTAACAATGACTCCAACTGCTCTTAAG                                                            228
int9.6
GTGGGTCGTTCAATAGCTTCAAAGTATTCTTTTCTCCCAGCCGCCTGCTTCTATTGCTTTCCTACTTATTAACAACATTCTGGTTTATTCGGTGGTACTGAGCCAAACCGTGTGTGCTGAATAGGGCTAACCTAGTCGAAGTC
GAAGTGGAGAAACTATAAACATATTTTTTTTTTAAACTCAG                                                                                                        184
```

E



Supplementary Figure S7