J Mol Evol (2014) 78:50–65
DOI 10.1007/s00239-013-9599-3

ORIGINAL ARTICLE

Evolution of the *Sex-lethal* Gene in Insects and Origin of the Sex-Determination System in *Drosophila*

Zhenguo Zhang · Jan Klein · Masatoshi Nei

Received: 16 September 2013 / Accepted: 12 November 2013 / Published online: 24 November 2013

Abstract *Sex-lethal* (*Sxl*) functions as the switch gene for sex-determination in *Drosophila melanogaster* by engaging a regulatory cascade. Thus far the origin and evolution of both the regulatory system and SXL protein's sex-determination function have remained largely unknown. In this study, we explore systematically the *Sxl* homologs in a wide range of insects, including the 12 sequenced *Drosophila* species, medfly, blowflies, housefly, *Megaselia scalaris*, mosquitoes, butterfly, beetle, honeybee, ant, and aphid. We find that both the male-specific and embryo-specific exons exist in all *Drosophila* species. The homologous male-specific exon is also present in *Scaptodrosophila lebanonensis*, but it does not have in-frame stop codons, suggesting the exon's functional divergence between *Drosophila* and *Scaptodrosophila* after acquiring it in their common ancestor. Two motifs closely related to the exons' functions, the SXL binding site poly(U) and the transcription-activating motif TAGteam, surprisingly exhibit broader phylogenetic distributions than the exons. Some previously unknown motifs that are restricted to or more abundant in *Drosophila* and *S. lebanonensis* than in other insects are also identified. Finally, phylogenetic analysis suggests that the SXL's novel sex-determination function in *Drosophila* is more likely attributed to the changes in the N- and C-termini rather than in the RNA-binding region. Thus, our results provide a clearer picture of the phylogeny of the *Sxl*'s *cis*-regulatory elements and protein sequence changes, and so lead to a better understanding of the origin of sex-determination in *Drosophila* and also raise some new questions regarding the evolution of *Sxl*.

Keywords *Sex-lethal* · Sex-determination · *Drosophila* · Regulatory elements

Abbreviations

CDS	Coding sequence
HMM	Hidden Markov model
AA	Amino acid
RRM	RNA recognition motif
RBD	RNA-binding domain

Introduction

The gene *Sex-lethal* (*Sxl*) occupies the pivotal position in the hierarchical network (Fig. 1a) controlling the sexual development and dosage compensation in *D. melanogaster* (Bell et al. 1988; Salz 2011; Penalva and Sanchez 2003; Salz et al. 1989). In the network, *Sxl*'s activation depends on the ratio of the number of X chromosome to that of each autosome (the X:A ratio). In females, there are two X chromosomes so the ratio is 1 and *Sxl* is activated (reviewed in Salz 2011; Penalva and Sanchez 2003). Once activated, *Sxl* can maintain its activation by employing an auto-regulatory loop which does not depend on the number of X chromosomes any more. Then the SXL proteins further regulate the splicing and translation of downstream genes, such as *transformer* (*tra*) and *male-specific lethal 2* (*msl-2*), which direct the sexual development and dosage

Electronic supplementary material The online version of this article (doi:10.1007/s00239-013-9599-3) contains supplementary material, which is available to authorized users.

Z. Zhang (✉) · J. Klein · M. Nei
Institute of Molecular Evolutionary Genetics and Department of Biology, Pennsylvania State University, 328 Mueller Laboratory, University Park, State College, PA 16802, USA
e-mail: zuz17@psu.edu; zhangzg.sci@gmail.com

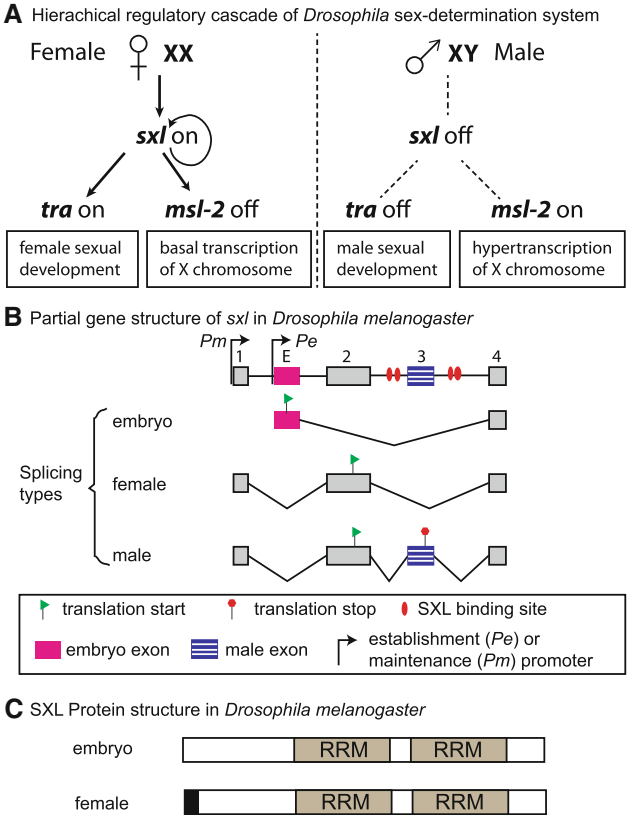


Fig. 1 The diagram of the sex-determination system in *Drosophila melanogaster*. **a** The hierarchical structure of the sex-determination system. The *solid arrows* indicate the flow of signals from the upstream genes to downstream genes either by activating or repressing the downstream gene expression. The *dashed lines* mean that there are no signals from the upstream genes. **b** The exon–intron organization for the first five exons in the gene *Sxl* of *Drosophila melanogaster*. Different splicing isoforms are given below. The embryo splicing isoform is expressed in a very early embryo stage of the females while the female and male isoforms are expressed in late embryonic stage and adults. **c** The schematic diagrams of the proteins from the embryo and from the female splicing isoforms. The *black segment* in the female protein marks the different part from the embryo protein. *RRM* represents the RNA recognition motif. The male splicing isoform does not produce any protein

compensation pathways, respectively. In males, *Sxl* is silenced because there is only one X chromosome, and the downstream genes act in an opposite direction to affect male development (Fig. 1a).

The establishment of the *Sxl* auto-regulatory loop in females involves a series of steps. In the early embryo stage, the early establishment promoter *Pe* (Fig. 1b) is transiently activated in females in response to the two-X-chromosome signal (Salz 2011; Penalva and Sanchez 2003). The transcription starts from an embryo-specific exon (exon E in Fig. 1b) splicing it directly to exon 4, which produces embryo SXL protein isoforms (Fig. 1c). The embryonic SXL is female-specific, but to distinguish it from the following adult female SXL (Fig. 1c) we refer to it as the embryonic SXL. Later on, the establishment promoter shuts off and the upstream maintenance promoter *Pm* (Fig. 1b) becomes active in both sexes. In females, the male-specific exon (exon 3 in Fig. 1b) is always skipped because the accumulated embryonic SXL proteins bind to the poly(U) tracts in the flanking introns of this exon and hinder its inclusion into final mRNAs. The female splicing isoform encodes the female SXL protein which can bind to the *Sxl* pre-mRNAs and regulate the female-specific splicing. In this way, a positive auto-regulatory loop is established and maintained in females by *Sxl* itself through the life cycle of *D. melanogaster*. By contrast, in males due to lack of the embryonic SXL proteins from the *Pe* promoter, the male-specific exon is always included in transcripts and introduces in-frame stop codons in the exon (Fig. 1b), which inhibits the protein production. Therefore, in males there are neither functional SXL proteins nor the positive auto-regulatory loop. Through this mechanism, the *Sxl* is sex-specifically expressed, and regulates sex-specifically downstream genes controlling the sexual development and dosage compensation (Salz 2011; Penalva and Sanchez 2003).

After the *Sxl*'s sex-determination role was discovered in *D. melanogaster*, several studies investigated its orthologues in other species and reported no sex-determining role in non-*Drosophila* species. For example, the *Sxl* orthologues in medfly *Ceratitis capitata*, housefly *Musca domestica* and scuttlefly *Megaselia scalaris* encoded proteins similar to the female SXL of *D. melanogaster*, but did not show sex-specific expression (Saccone et al. 1998; Meise et al. 1998; Sievert et al. 2000). Furthermore, transgenic expressions of these orthologous proteins in *D. melanogaster* did not trigger feminization in males (Saccone et al. 1998; Meise et al. 1998), indicating an alteration of protein functions. These results indicate that not only the gene expression but also an altered protein function may have contributed to the gain of sex-determination function in *Drosophila*.

The functional difference of *Sxl* orthologues between *Drosophila* and other species triggered several studies to investigate how *Sxl* acquired this novel function (Mullon et al. 2012; Cline et al. 2010; Traut et al. 2006). Traut et al. (2006) reported a duplicated gene of *Sxl*, namely *ssx*, in *Drosophila* species. The duplication probably took place after the divergence between the Drosophilidae and the Tephritidae families (Cline et al. 2010). Based on these facts, it was proposed that the new *ssx* gene serves the ancestral *Sxl* function and frees the *Sxl* to evolve the novel sex-determination function (Traut et al. 2006). Under the assumption that *ssx* took upon itself the ancestral function, Cline et al. disrupted the *ssx* in *D. melanogaster* in vivo and found that there was no significant effect on the viability and fecundity, even in combination with the loss of *Sxl*

(Cline et al. 2010). This result suggests that either *ssx* does not serve the ancestral function or that the *Sxl*'s ancestral function is not essential in the standard laboratory conditions (Cline et al. 2010).

Although efforts have been made to understand the evolution of *Sxl* and its relationship with the sex-determination in *Drosophila*, there still remain two main unsolved problems. First, how and when was the current regulatory system of *Sxl* for the sex-specific expression established? As explained above, the sex-specific expression pivots on the accurate regulation needing a set of sequence elements (Fig. 1b), including the male-specific exon, the *Pe* promoter and the downstream embryo-specific exon, and the sequence motifs residing in the nearby regions such as poly(U) tracts in the flanking introns of the male exon (Horabin and Schedl 1993b) and the TAGteam motif in the *Pe* promoter (ten Bosch et al. 2006). So far, only a few of these elements have been examined in a small number of *Drosophila* species and the results implied that the regulation system is probably conserved in the *Drosophila* genus (Cline et al. 2010; Penalva et al. 1996; Bopp et al. 1996). The newly sequenced insect genomes provide the opportunity to examine the phylogenetic distributions of these regulatory elements in a wider range of species and infer their origins. Second, it remains unknown what is the molecular change responsible for the functional difference of the orthologous SXL proteins between *Drosophila* and other insects such as medfly and housefly (Saccone et al. 1998; Meise et al. 1998). A typical SXL protein has two RRM-type RNA binding domains (RBDs) with a seven-amino acid (AA) linker region between them, and two terminal regions before and after the two RBDs (Fig. 1c). Whether the two RBDs or the two termini or both are involved for SXL's functional change is unknown. In this study, we aim to answer the above two questions.

Materials and Methods

Evolutionary Analyses

The Detection of the Sxl Homologs

To detect the *Sxl* homologs in the surveyed species, we searched with BLASTP the proteomes of species for which the whole genomes had been released. The protein sequences of the 12 *Drosophila* species were downloaded from the FlyBase (http://flybase.org/) (Drysdale 2008) and for other species the sequences were from the Ensembl (http://metazoa.ensembl.org/) (Flicek et al. 2012). We used the protein of *D. melanogaster* female-specific isoform-MS3 (uniprot ID: P19339-1) as a query. The *E* value cutoff was set as $\leq$10 to include all the potential homologs. After the BLAST search, each hit sequence was examined to find out whether the high-score segment alignment had a $\geq$40 amino acid overlap with at least one of the two RRMs in the query sequence and the hits with the overlap smaller than 40 AAs were eliminated. In this way, we obtained 2,065 sequences with at least partial putative RRMs. Then, a neighbor-joining tree was constructed using the sequence alignment of the RRM region, which was generated by aligning the above sequences onto the RRM model (downloaded from http://pfam.sanger.ac.uk/family/PF00076) using the tool hmmalign. On the tree, we identified the smallest clade which had the root corresponding to the common ancestor of all the examined species and contained the *D. melanogaster Sxl* query sequence. In this way, we identified 32 homologous sequences which included the 12 paralogous *ssx* genes of *Drosophila*. To add more sequences from the closely related species, we downloaded the SXL protein sequences and its corresponding mRNAs from the UniProt (Reorganizing the protein space at the Universal Protein Resource (UniProt) 2012) and NCBI databases for the following species: the medfly *C. capitata*, the housefly *M. domestica*, *Lucilia cuprina*, *Chrysomya rufifacies*, and the scuttle fly *M. scalaris*. The *Sxl* genomic sequence was requested from the medfly genome project at http://www.ars.usda.gov/research/projects/projects.htm?accn_no=414542 and now is available at NCBI (accession ID: NW_004522756.1). The partial genomic sequences of *Scaptodrosophila lebanonensis* (accession ID: EU670259) and housefly (accession ID: HM776132) were downloaded from the NCBI database. All the sequence accession numbers used in this study are listed in the Table S1 (online ESM1) and the genomic sequences are stored in Dataset S1 (online ESM2).

Evolutionary Rates and Phylogenetic Trees

The protein sequences were aligned using the program ClustalW (Thompson et al. 2002). The RRMs region in the alignment was determined by mapping the UniProt annotated RRM locations in *D. melanogaster Sxl* (RRM1: 117–195, RRM2: 203–283) onto the alignment. The alignment was then divided into three parts for the calculation of the evolutionary rates: the two RRMs and the linker between them, the region preceding the two RRMs (referred to as the N-terminus), and the region following the two RRMs (the C-terminus). The CDS alignments were then obtained by mapping the codons onto the amino acid alignments. For calculating the pairwise synonymous and nonsynonymous substitution rates, the CDS alignments were inputted into MEGA 5 (Tamura et al. 2011) using the parameters: pairwise gap deletion and *p* distance (proportion of difference). We chose *p* distance because it is model-free and worked well for our purpose. To determine

the evolutionary rates for a pair of sequences, their p distance was divided by two times their divergence time (unit: million years). For constructing the phylogenetic trees, the neighbor-joining method implemented in MEGA5 (Tamura et al. 2011) was used with p distance.

Inferring the Amino Acid Substitutions in the Common Ancestor of Drosophila

To reconstruct the sites which changed in the *Drosophila* ancestor, we used the *codeml* program in the PAML package (Yang 2007) for inferring the ancestral sequences in the internal nodes, which is based on the maximum likelihood model of amino acid changes. The parameter 'aaRatefile' was set to *wag.dat*. The analyses were done separately for the RBDs, the N-terminal and the C-terminal regions. In the analyses, only SXL orthologues from the following species are used: the 12 *Drosophila* species, the medfly *C. capitata*, the housefly *M. domestica*, *L. cuprina*, *C. rufifacies*, and *M. scalaris*. The sequences of other species and of the paralogous *ssx* genes were far from the *Drosophila Sxl* group on the tree so they did not provide more information to the inference.

The Identification of Homologous Exons and Introns from the CDS Alignment

The CDS alignment for all the sequences was obtained as described above. The exon border positions in the alignment were determined by mapping the first and the last base of each exon onto the alignment. When the positions of the two border bases (one from the preceding exon and the other from the succeeding exon) in a sequence were aligned with the corresponding exon border positions of the *D. melanogaster Sxl*, this exon border was regarded as homologous to that of *D. melanogaster Sxl* and the intron between them was also assumed to be homologous. The exons between two matching exon borders were taken for homologous exons.

The Detection of the Homologous Male-Specific and Embryo-Specific Exons

For male-specific exons, we used the annotated male-specific exon in *D. melanogaster Sxl* as query in BLASTN search of the candidate genomic regions in other species. The candidate genomic regions were selected based on the CDS alignment, that is, the genomic regions between the homologous exons 2 and 4 as defined above. We used several steps to get the final set of the male-specific exons. First, we searched the candidate sequences with the *D. melanogaster* male-specific exon using BLASTN. Second, in each candidate sequence we extracted the best high-score region and its flanking 500 bps from both sides. In this step, we used the E value cutoff ≤ 10 to obtain all the potential hits. Third, we made a global multiple alignment of the extracted regions and cut out the alignment region matching the male-specific exon of *D. melanogaster*. Fourth, we re-aligned the cut region and again selected the region matching the male-specific exon of *D. melanogaster*. Finally, a hidden Markov model (HMM) was built with this alignment and the candidate sequences were searched again for potentially missed homologous exons using the tool hmmsearch in the HMMER package (http://hmmer.org/). The last step did not find any extra exons. The final alignments also include sequence fragments from some non-*Drosophila* species just because very short sequence fragments were hit in BLAST search of those species. We include them here to indicate that the male-specific exon is really absent in those species.

As a complementary method, we also searched for the male homologous exons using the information of the conserved flanking sequences. Briefly, the sequence segments flanked by two oligo-nucleotides with the regular expressions T[AG]T{1,2}[AGT]TAG and GTAAGTAA were cut from the candidate genomic regions and regarded as putative male-specific exons. This method did not add more exons.

Using a similar strategy of BLAST search as that for the male-specific exon, we tried to identify homologous embryo-specific exons in other species with the embryo exon of *D. melanogaster* as the query. The candidate regions used were the 5,000 nucleotides upstream of the homologous exon 2 because for most species the upstream introns of exon 2 were not annotated and the 5,000 nucleotides approximate the intron size in *D. melanogaster*. For the medfly, the upstream intron of the exon 2 was used as the candidate sequence.

3D Structure Analysis

The 3D structure for the RRM domain (PDB ID: 1B7F) was downloaded from the PDB database (http://www.rcsb.org/pdb/explore/explore.do?structureId=1B7F) (Sussman et al. 1998). The protein sequence of *D. mel Sxl* was then mapped onto the structure with the software Jalview (Waterhouse et al. 2009). The display of the structure and the labeling of selected residues were done with the software Jmol (Jmol: an open-source Java viewer for chemical structures in 3D. http://www.jmol.org/).

Motif Search

The MEME tool suite was used to search for the motifs in a given set of input sequences. Briefly, we extracted flanking introns of the homologous male-specific exons and inputted

them to the program *meme* (Bailey and Elkan 1994) with the parameters 'meme -mod anr -dna -nmotifs 50 -evt 1 -minsites 5 -maxsites 60'. The parameter '-evt 1' told the search to stop when motif *E* value >1. We searched for the motifs in the upstream and downstream introns separately. To match these identified motifs to the known RNA-binding motifs, we downloaded 72 position frequency matrices from the RBPDB database (Cook et al. 2011). The motifs were compared to these matrices with the tool *tomtom* in the MEME suite (Bailey et al. 2009).

Similarly, we searched for the motifs in the 5,000 nt upstream regions of the homologous embryo-specific exons. The identified motifs were compared to all the known *Drosophila* transcriptional binding sites with the tool tomtom. The known transcriptional binding sites were downloaded from the MEME website http://ebi.edu.au/ftp/software/MEME/index.html.

Calculation of Splice Site Score

First, we obtained the 5′ and 3′ splice site motif matrices from the Table 2 in (Mount et al. 1992). For the 5′ splice site, the motif consisted of two bases from the 3′ end of exon and seven bases from the downstream intron. The 3′ splice site included 14 bases from the upstream intron and 2 bases from the 5′ end of the exon. The matrices contained counts for each type of nucleotides (A, T, C and G) at each position. Then we calculated splice site scores with the method (Miyasaka 1999) similar to that of Codon Adaptation Index (CAI). Briefly, the relative adaptiveness w_{ij} for the nucleotide i at position j was determined by its count divided by the maximum count of that position so that the maximum adaptiveness for each position is always one. For a given splice site sequence, the score was a geometric mean of the w_{ij} values, For example, the score for a 5′ splice site of CTGTAAGTA was calculated as $(w_{C1}{}^{*}\ w_{T2}{}^{*}\ w_{G3}{}^{*}\ w_{T4}{}^{*}\ w_{A5}{}^{*}\ w_{A6}{}^{*}\ w_{G7}{}^{*}\ w_{T8}{}^{*}\ w_{A9})^{1/9}$. For 3′ splice site, it was calculated in a similar way.

Results

To reveal the phylogenetic distributions of *Sxl*'s regulatory elements and the protein sequence changes, we systematically examined the *Sxl* homologues in a broad range of insects (see Fig. S1 (online ESM1) for the phylogenetic tree of all the species). For *S. lebanonensis* (a species in the sister genus of *Drosophila*) and housefly *M. domestica*, only partial genomic sequences of *Sxl* were available. For the blow-flies *C. rufifacies* and *L. cuprina* and the scuttlefly *M. scalaris*, only protein-coding regions were known. These sequences were also used when they could provide applicable information. All the sequence accession IDs are in Table S1 (online ESM1) and all the genomic sequences of the genes are stored in Dataset S1 (online ESM2). For brevity, we use the Flybase abbreviation format (one letter from the genus name and three letters from the species name) to refer to these species in figures and tables, for example, *D. mel* for *D. melanogaster*.

Conserved Exon Borders of Exons 2 and 4 from *Drosophila* to Mosquitoes

The *Sxl*'s exon–intron organization in *D. melanogaster* plays a crucial role in the sex-specific expression. In particular, the exons E, 2, 3, and 4 are the basis for forming the stage- and sex-specific isoforms (Fig. 1b). Here we examine the presence of exons 2 and 4 in other species. The embryo-specific exon E and the male-specific exon 3 will be discussed in the succeeding sections. For convenience, we use the exon numbering system in *D. melanogaster*'s *Sxl* to refer to the homologous exons and introns of other species (Fig. 2).

The end of the exon 2 and the start of the exon 4 are conserved for the *Sxl* orthologues from *Drosophila* to mosquitoes (Fig. 2, S2 (online ESM1)). The conservation disappears in more distant species and in the *ssx* paralogs. In *ssx*, the borders of exons 2 and 4 do not match at all those of *Drosophila Sxl* exons. The start of the exon 4 is shifted and the intron between the exons 4 and 5 is absent in all the *ssx* genes. The intron was probably lost during or after the gene duplication. These results imply that *ssx* may have lost the sex-specific splicing regulation. Consistent with this supposition, the *ssx* introns between exons 2 and 4 are very short, ~80 bps, ruling out the possibility of containing the 190 nt long male-specific exon. Interestingly, the exon border conservation extends to the RBD region. In this region, all the exon borders (the intron between exons 6 and 7 was lost in some species) match for all the surveyed species as well as in the *ssx* paralogs. In light of the fact that the known function of this region is to encode the two RBDs, it is intriguing why these introns have been maintained for more than 300 million years.

Although the borders of exons 2 and 4 of *Sxl* are conserved from *Drosophila* to mosquitoes, the size of the intron between them varies considerably. The genomic distances between the end of exon 2 to the start of exon 4 are about 3.5–5.2 kb for *Drosophila Sxl* genes with a very few exceptions (Fig. 2, S2 (online ESM1)), but the distances vary tremendously in other species. For example, in the medfly the distance is longer than 30 kb. The distances in mosquitoes are also quite different, being 12,139 in *Aedes aegypti* and 1,102 in *Anopheles gambiae*. The biological significance of the length variation is not clear, but it might affect the splicing efficiency of flanking exons

(Kandul and Noor 2009). The variation suggests relaxed constraint on splicing in those species.

The Male-Specific Exon is Restricted to *Drosophila* and *Scaptodrosophila*

The inclusion of the male-specific exon (exon 3) is the hallmark for the male-specific splicing isoforms in *D. melanogaster* (Fig. 1b). So far this exon was only reported in *Drosophila* species and in *S. lebanonensis* (Cline et al. 2010). The conservation of the exon borders between exon 2 and exon 4 in the medfly, housefly, and mosquitoes raises the possibility that the homologous male-specific exons might reside in these species but have not been identified.

Since the complete sequences of the male-specific exons in the reported species were not given by the earlier study (Cline et al. 2010), we searched for the homologous exons in the candidate regions (the genomic sequences between exons 2 and 4) of all the species from *Drosophila* to mosquitoes. We used two complementary methods. First, we searched for the homologous male-specific exons based on the similarity to the *D. melanogaster*'s *Sxl* male exon (see "Materials and Methods" section for details). We identified one exon for each *Sxl* orthologue in *Drosophila* species and *S. lebanonensis* (Table S2 (online ESM1)), shown in Fig. 3. The extended sequence fragments of BLAST best hits from the medfly and housefly are very different from these exons and they lack the splice sites, supporting the male exon's absence in these species. The exons show great divergence

Fig. 3 The nucleotide sequence alignment of the male-specific ▶ exons. The *red box* encloses the sequences of the *Drosophila* species. The extended BLAST best hit regions from the medfly and housefly as well as the contiguous ~10 nt intronic sequences (*lowercase*) upstream and downstream of the male exons are also shown. The in-frame stop codons of the male transcripts are marked in the *purple background* while the start codons of potential downstream translation are marked by the *violet background*. The position of the downstream alternative 3′ splice site is marked by *asterisk* and a *vertical black line* is given to indicate the starts of the exons from this splice site. Sleb: *Scaptodrosophila lebanonensis*; Ccap: *Ceratitis capitata*; Mdom: *Musca domestica*

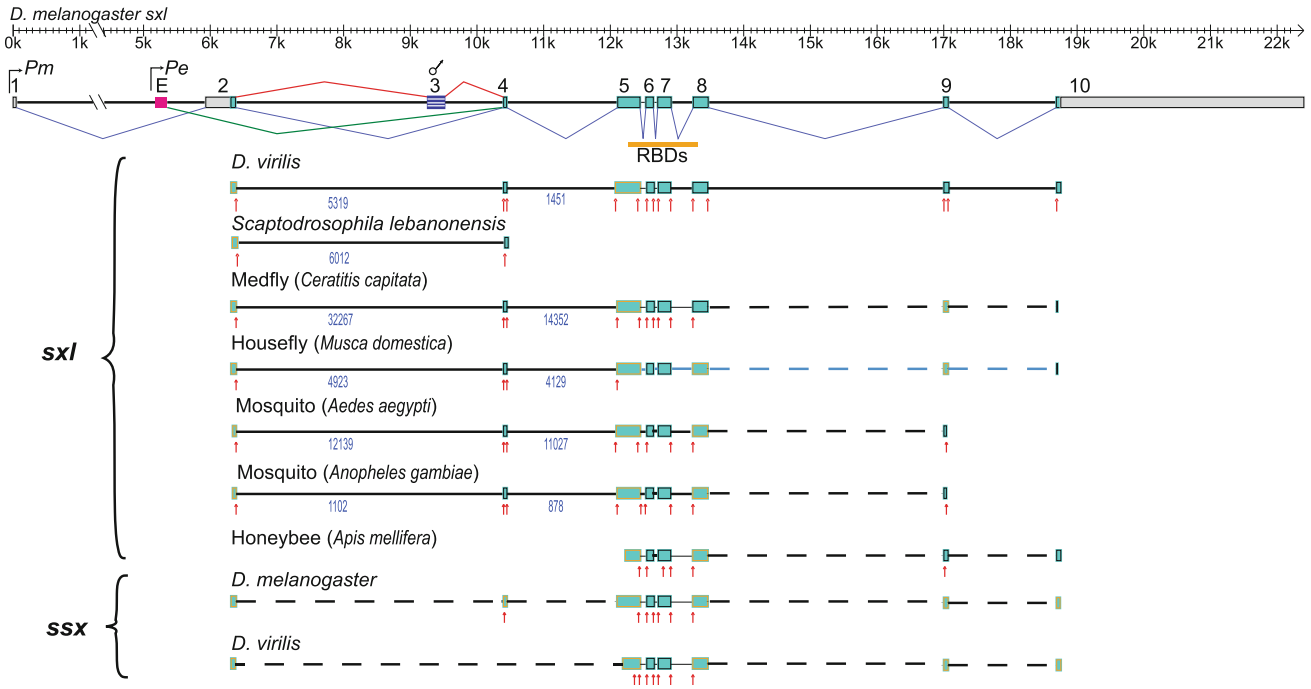
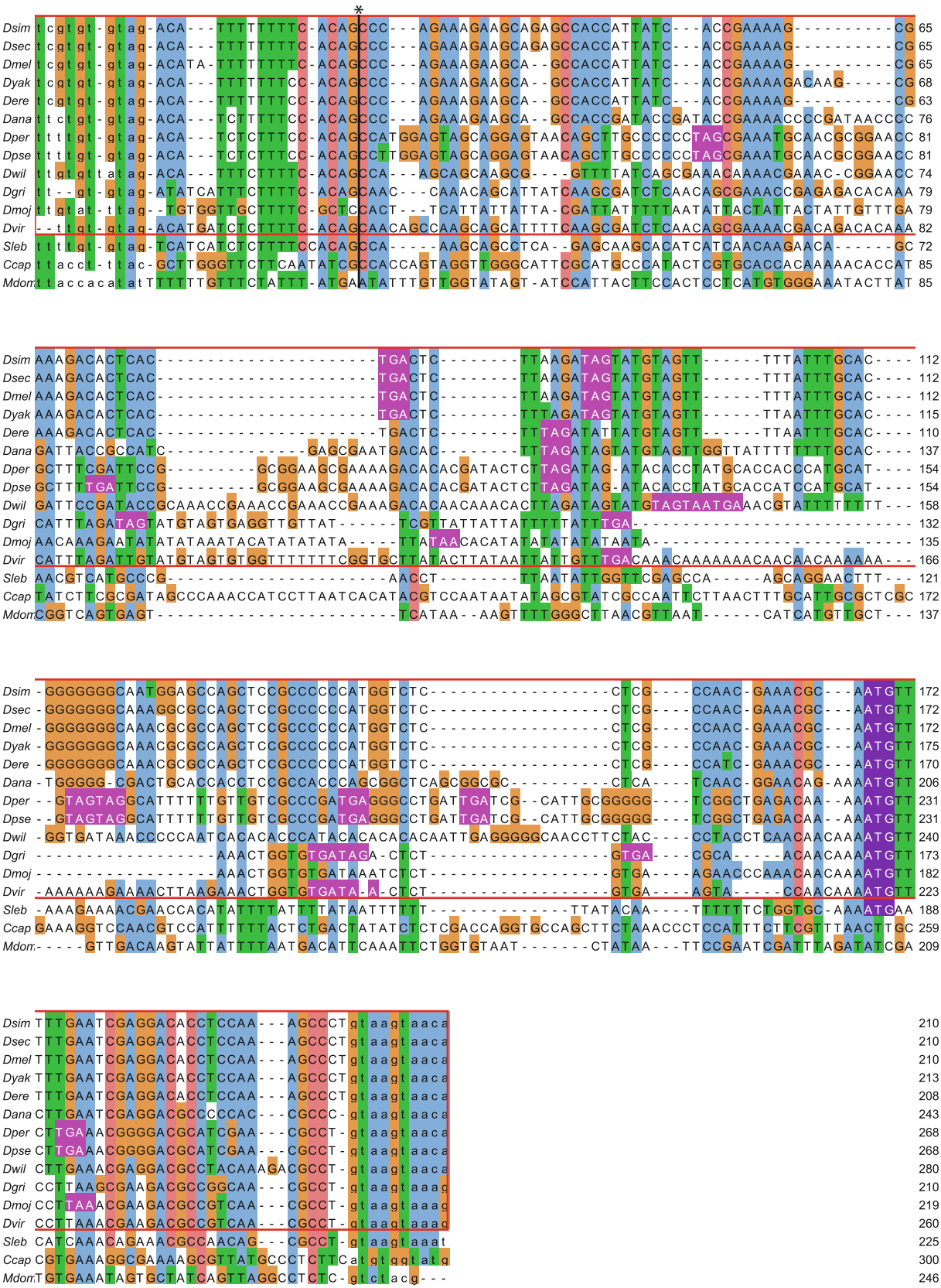


Fig. 2 The exon–intron organization of the *Sxl* and the *ssx* homologous genes of selected species. The exon–intron organization of *Drosophila melanogaster* is shown at the *top*. The protein-coding region is shown in *turquoise* while the UTR is in *gray*. The introns are displayed as *black solid lines*. The embryo-specific exon (exon E) is colored in *pink*, and the male-specific exon (exon 3) is colored in *violet* and has *white horizontal lines* in it. The female splicing is ligated in the *blue lines*. The different parts of the splicing for the male isoform and for the embryo isoform are shown in *red* and *green lines*, respectively. The genomic region encoding the two RRM-type RNA-binding domains and the linker sequence (RBDs) are also marked below the gene structure with an *orange bar*. *Arrows* indicate the establishment promoter *Pe* and maintenance promoter *Pm*. The gene structures from other species are aligned based on the CDS alignment of all *Sxl* sequences. Below each gene structure, the *red arrows* mark the exon–intron borders. The *black dashed lines* indicate the gaps, usually because one exon is split into two parts to align with the exon in *D. melanogaster Sxl*. The *blue dashed line* in the medfly *Musca domestica* sequence means unknown gene structure because the genomic sequence of that region is unavailable. The gene structure in the partial *Scaptodrosophila lebanonensis* genomic sequence is also shown. The sizes for introns 2 and 4 are given below the introns


Dsim
Dsec
Dmel
Dyak
Dere
Dana
Dper
Dpse
Dwil
Dgri
Dmoj
Dvir
Sleb
Ccap
Mdom

among the *Drosophila* species, which is not surprising in view of the fact that the known function is to introduce stop codons in male splicing isoforms thus making the coding sequence free to mutate as long as stop codons are retained. In contrast to the exon sequence itself, the sequences at the 5′ and 3′ splice sites are conserved. In the 5′ splice site, the ten nucleotides downstream of the exon have the consensus sequence GTAAGTAA[C/A]R, and the two nucleotides at the end of the exon are CT for all species (except for CC in *D. ananassae*). In the 3′ splice site, the seven nucleotides upstream of the exon are TGTGTAG for all exons (except for TGTtaTAG in *D. willistoni*) followed by the four exonic nucleotides ACAT for all species (except for TGTG in *D mojavensis* and TCAT in *S. lenanonensis*). The high consistency at these splice sites implies strong selection on splicing regulation. However, these splice sites do not match well the known *Drosophila* consensus sequences (Mount et al. 1992), and in terms of the extents of matching the consensus sequence these splice sites are weaker than the corresponding upstream 5′ splice site of exon 2 and the downstream 3′ splice site of exon 4 in most species (Fig. 4a). These splice site patterns suggest that the male-specific exons are under selection for maintaining weak splice sites and are generally not preferred in splicing compared to two neighboring exons. This observation contradicts the known regulation model in which the male exon is included in the male transcript by default while in females some other factors are needed to skip this exon (Horabin and Schedl 1993a). It suggests that there might be some other unknown factors which promote the splicing of male exons.

Second, we used this information of conserved nucleotides at two ends of the male exons to search for the putative homologous exons (see "Materials and Methods" section). This method bypasses the obstacle of great sequence divergence among the homologous exons. The search, however, failed to find more exons.

To check whether all the homologous exons have the capacity to introduce stop codons when translated, we construct a putative male-specific transcript for each species by joining this exon to the corresponding exons 2 and 4 and translated it with the FlyBase (Drysdale 2008) annotated frames. All the exons in *Drosophila* contain more than one in-frame stop codon, though their positions vary among species (Fig. 3). Surprisingly, in *S. lebanonensis* this exon does not have any in-frame stop codon. What is the function of this exon in *S. lebanonensis* (see "Discussion" section)?

Conserved Sequence Motifs in the Flanking Introns of the Male-Specific Exon

Accurate splicing regulation of male-specific exon requires sequence motifs located in the upstream and downstream introns (Horabin and Schedl 1993b). The best characterized

Fig. 4 The splice sites of the male-specific exons and the conserved sequence motifs in the flanking introns. **a** The splice sites of the male-specific exons, the 5′ splice sites of the upstream exons 2 and 3′ splice sites of the downstream exons 4. The scores measure the match degrees with the *Drosophila* splice site consensus sequences (CONS.). **b** and **c** The distributions of conserved motifs in the upstream and downstream introns of the *D. melanogaster* male-specific exon, respectively. The distributions in the introns of medfly (*C. cap*) and housefly (*M. dom*) are shown for comparison

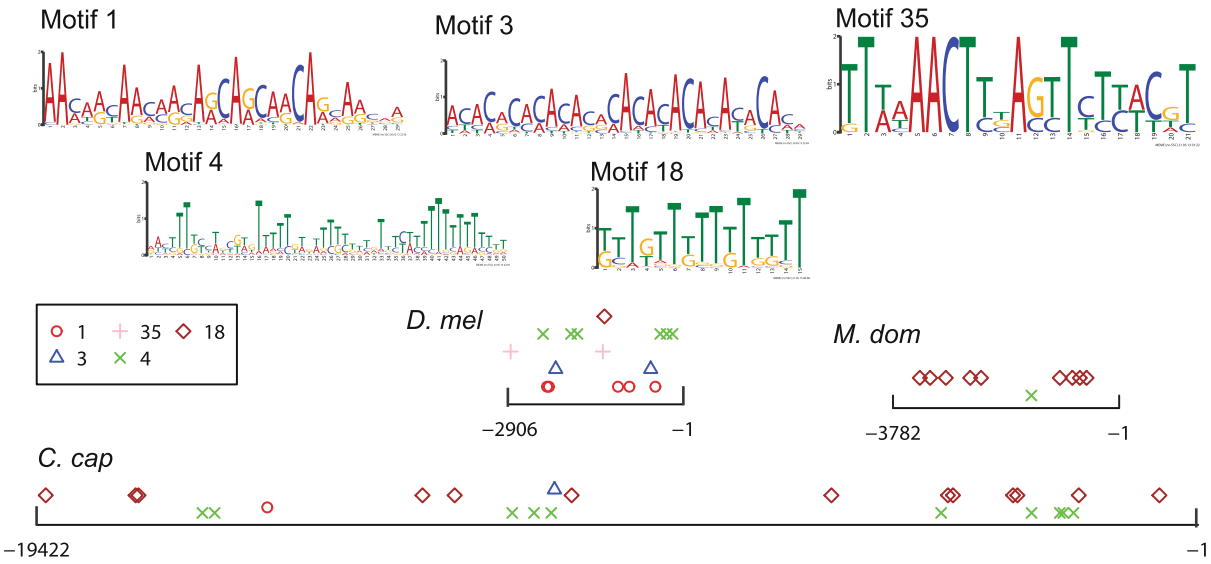
of these motifs is the poly(U) [poly(T) in DNA] tract to which the SXL protein binds and thus initiates the skipping of this exon during splicing. Deletion of these motifs from the introns destroys the splicing regulation in *D. melanogaster* (Horabin and Schedl 1993b). To find out whether this motif exists in all the *Drosophila* species and thus obtain a good indication that the identified homologous male exons are functional, as well as to look for other previously unidentified motifs, we searched for all the potential motifs in the upstream and downstream introns of the male-specific exons separately using the MEME tool (Bailey and Elkan 1994). As negative controls, we included the corresponding introns in medfly and housefly in the input sequences. Tens of motifs were found showing diverse phylogenetic distributions (see Dataset S2 (online ESM3)). Here we focus on the motifs which are shared by all the *Drosophila* species but absent or fewer in the medfly and housefly because this kind of motifs can be supposed to be more relevant to the common sex-specific regulation in *Drosophila*. The motifs are referred to by their identification numbers (ID) given by the MEME tool which do not have any biological meaning. In the upstream introns, we found two motifs (motifs 1 and 3) being more abundant in the *Drosophila* and one motif (motif 35) being unique to the *Drosophila* (Table 1). These motifs disperse in the intron (Fig. 4b, c) and do not match any known motifs in the database RBPDB (Cook et al. 2011) (see "Materials and Methods" section). For the downstream intron, there are no motifs unique or more abundant for the *Drosophila*. However, the motif 8 is shown here because it is shared by all the *Drosophila* species (Table 1) and near the splice site (Fig. 4c), arguing for potential functions common to *Drosophila*.

Due to the importance of the poly(U) motif in the regulation, we intentionally looked for alike motifs. Three motifs, the motifs 4 and 18 in the upstream introns and the motif 1 in the downstream introns, resemble the poly(U) motif (Fig. 4b, c). As expected, all the *Drosophila Sxl* sequences have the two or all of the three motifs (Table 1). However, these motifs are more abundant in the medfly and the housefly than in the *Drosophila* species (Table 1; Dataset S2 (online ESM3)). In *D. melanogaster*, the poly(U)-like motifs more often cluster near the male-specific exon while in the medfly and housefly they scatter

A Splice sites flanking the male-specific exon

	Exon 2 5'ss	score	Male exon 3'ss	score	Male exon 5'ss	score	Exon 4 3'ss	score
CONS.	AGGTRAGTA	1.0	TTTYYYTYYTNCAGRT	1.0	AGGTRAGTA	1.0	TTTYYYTYYTNCAGRT	1.0
Dsim	AGGTAAATT	0.759	TTAATCGTGTGTAGAC	0.637	CTGTAAGTA	0.717	GATTCATCGTACAGTG	0.674
Dsec	AGGTAAATT	0.759	TTAATCGTGTGTAGAC	0.637	CTGTAAGTA	0.717	GATTCATCGTACAGTG	0.674
Dmel	AGGTAAATT	0.759	TTAATCGTGTGTAGAC	0.637	CTGTAAGTA	0.717	GATTCATCGTACAGTG	0.674
Dyak	AGGTAAATT	0.759	TTAATCGTGTGTAGAC	0.637	CTGTAAGTA	0.717	GATTCATCGAACAGCG	0.674
Dere	AGGTAAATT	0.759	TTAATCGTGTGTAGAC	0.637	CTGTAAGTA	0.717	GATACATCCTATAGCG	0.674
Dana	AGGTAAATT	0.759	TTAATTCTGTGTAGAC	0.677	CCGTAAGTA	0.686	CGATCGACGAACAGCG	0.876
Dper	AGGTAAATT	0.759	TTAATTTTGTGTAGAC	0.689	CTGTAAGTA	0.717	GTCGAATCGTGCAGCG	0.674
Dpse	AGGTAAATT	0.759	TTAATTTTGTGTAGAC	0.689	CTGTAAGTA	0.717	GTCGAATCGTGCAGCG	0.674
Dwil	AGGTAAATT	0.759	TAATTGTGTTATAGAC	0.646	CTGTAAGTA	0.717	TCTTTAATTTATAGCG	0.674
Dgri	AGGTAAATT	0.759	ATTTAATTGTGTAGAT	0.688	CTGTAAGTA	0.717	CGGTCAATCAACAGCG	0.721
Dmoj	AGGTAAATT	0.759	TCGATTGTATTTAGTG	0.582	CTGTAAGTA	0.717	AATCAATGAAACAGCG	0.737
Dvir	AGGTAAATT	0.759	ATTTAATTGTGTAGAC	0.658	CTGTAAGTA	0.717	CGTTCAATCAACAGCG	0.721
Sleb	AGGTAAATT	0.759	TTAATTTTGTGTAGTC	0.664	CTGTAAGTA	0.717	TCAATCAACAACAGCG	0.619

B Motifs upstream of the male-specific exon



C Motifs downstream of the male-specific exon

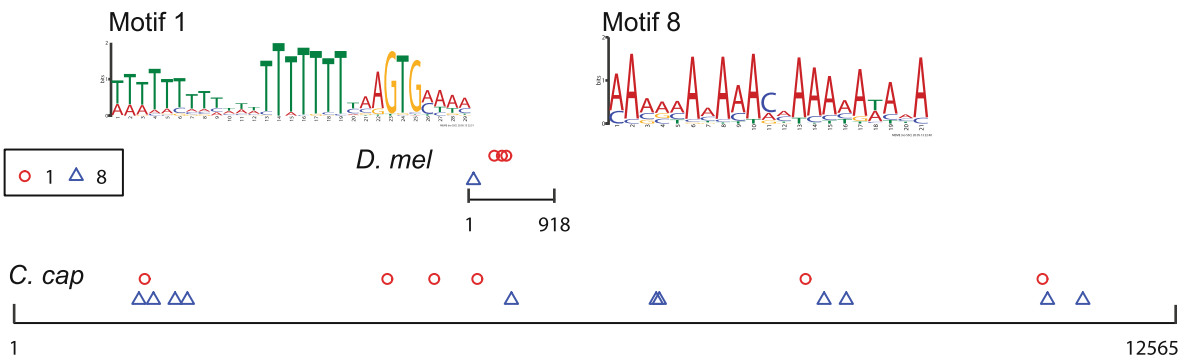


Table 1 Conserved motifs in the flanking introns of the male-specific exons

	Motif ID	Motif E-value	Number of *Drosophila* species with this motif	Average count per *Drosophila* species	Count in *S. lebanonensis*	Count in medfly	Count in housefly
Upstream intron	1	2.4e−229	12	4.7	2	1	0
	3	2.8e−205	12	4.6	4	1	0
	35	3.5e−18	11	1.4	0	0	0
	4	1.7e−166	12	3.7	2	10	1
	18	1.1e−44	9	3.4	4	13	9
Downstream intron	1	8.5e−175	12	4.2	2	6	0
	8	1.0e−31	12	2.5	1	11	0

more evenly in the introns. These results indicate that poly(U) motifs are not restricted to *Drosophila*. Whether their presence in other species is related to some functions is unknown.

The Embryo-Specific Exon is Unique to *Drosophila*

Similarly, we also searched for the presence of embryo-specific exon in the genomes of all the surveyed species. Using the same strategy as that in the male-specific exon search, we found one homologous exon in each *Drosophila* species (Fig. 5; Table S2 (online ESM1)). No such exon could be identified in other species. Due to the lack of genomic sequence upstream of exon 2 in *S. lebanonensis* we do not know whether this exon exists there. The extended best BLAST hit regions from other species (*A. aegypti*, *Apis mellifera* and *Tribolium castaneum*) are included for comparison. Compared to the male-specific exon, this exon is more conserved. The 3′ end region shows relatively higher conservation because it starts encoding the embryo protein. Intriguingly, we also observed a ~50 nucleotide conserved segment at the 5′ end in *Drosophila*. Because of its proximity to the transcription start it might have some regulatory function in transcription.

Then we checked whether all the homologous embryo-specific exons could be correctly translated. The lengths of these exon coding regions have a remainder 2 after division by 3 except the one in *D. mojavensis* which is divisible by 3. The remainder 2 is the same as that for the exon 2 coding region so it will not result in frameshift when joining with the downstream exon 4 (Fig. 1b). The unusual case in *D. mojavensis* suggests that there might be sequencing errors in this exon's sequence. Supporting this idea is the observation that when translated, this coding region has three in-frame stop codons and leads to many frameshifts in the downstream exons. One potential sequencing error is a 1-nucleotide deletion after 14 nucleotides from the translation starting point where it is a T in a stretch of 4 Ts in *D. virilis*, the closest species of *D. mojavensis* in the dataset. After replacing that gap with a T, the coding sequence within this exon is translated without any stop codon, but it still results in many stops in the downstream sequence. Therefore, simple sequencing errors may not be responsible for the different coding length. Actually, there is an ATG codon at the position −8 near the 3′ end. If translation starts from it, there is no problem in translation, but the protein product would be truncated compared to that in other species. To completely exclude the possibility of sequencing errors, re-sequencing this region in *D. mojavensis* is needed.

Conserved Sequence Motifs in the Putative Upstream Promoters of the Embryo-Specific Exons

To check whether there are conserved motifs in the upstream regions of the homologous embryo-specific exons, we searched the upstream 5,000-nucleotide region of each identified embryo-specific exon. As negative controls, we also scanned the motifs in the intron upstream of the first coding exon in the medfly and 5,000-nucleotide upstream regions of best hit regions (Fig. 5) in *A. aegypti*, *A. mellifera*, and *T. castaneum*. We found three motifs (motifs 1, 2, and 6) which are abundant in *Drosophila* but absent or much fewer in other examined species (Table 2, Fig. 6; Dataset S2 (online ESM3)). The motif 6 is unique to *Drosophila* and matches with the motif E-box CANNTG which is bound by the *sisterlessB* and *daughterless* proteins to activate *Sxl*'s establishment promoter *Pe* (Yang et al. 2001). For the other two motifs we do not find any matching known motifs.

We also searched for the TAGteam motif known to be involved in the activation of transcription of genes (including *Sxl*) in the early embryonic development stage (Ten Bosch et al. 2006; Satija and Bradley 2012). This motif is present in all the surveyed *Drosophila* species and three copies of the motif clusters near the embryo-specific exon in *D. melanogaster* (Fig. 6). Interestingly, this motif is also present in the examined regions of medfly and mosquitoes, but the number of copies is lower and its distribution is more even along the examined region (Fig. 6).

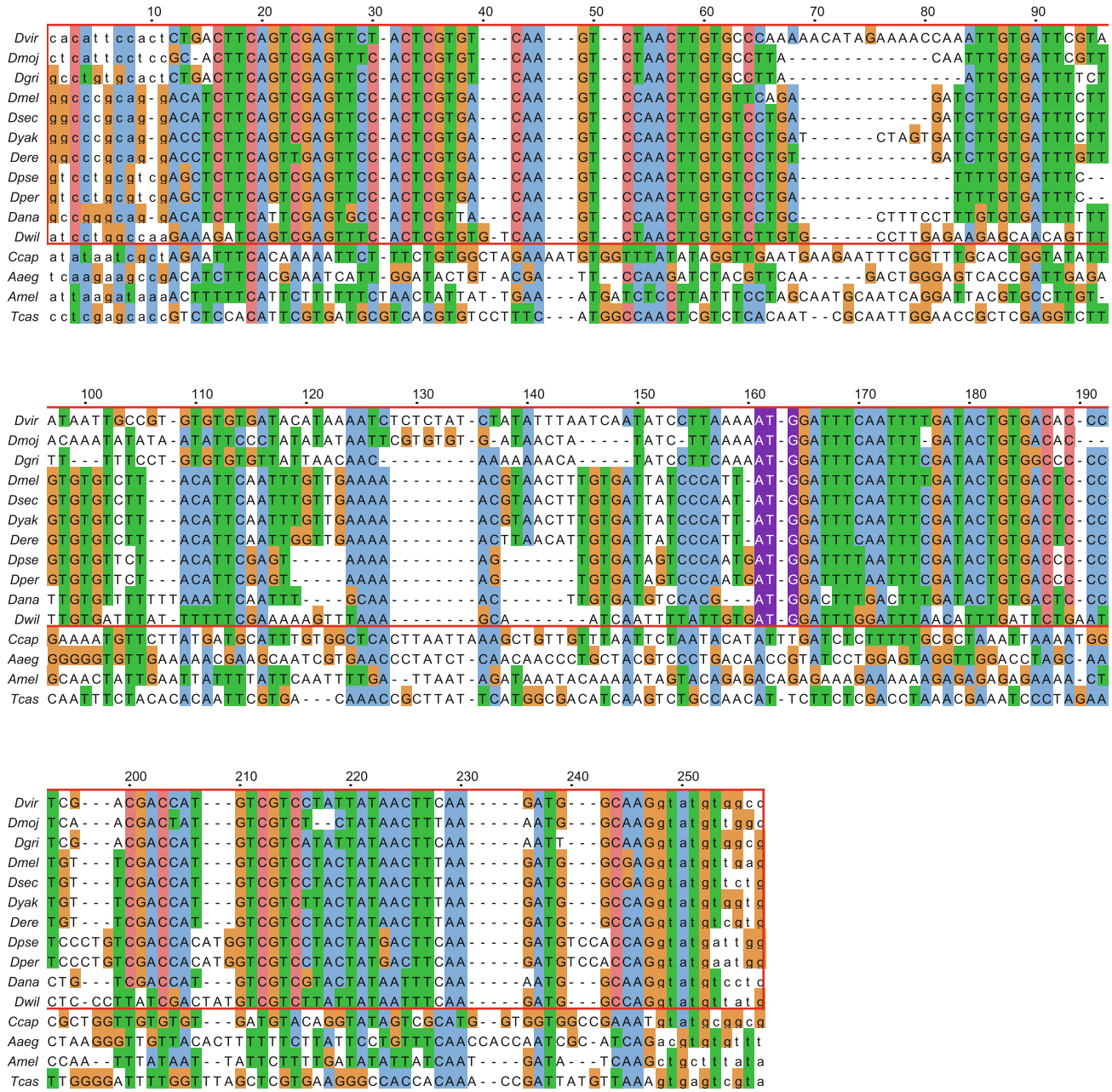


Fig. 5 The nucleotide sequence alignment of the homologous embryo-specific exons. The sequences from the *Drosophila* species are enclosed in the *red box*. Due to the un-sequenced bases in the corresponding region of *D. simulans*, the embryo exon is not shown. In the alignment, the putative start codons at the 3′ end are marked in violet. Approximately ten nucleotides upstream and downstream of these exons for each species as well as the extended BLAST best hit regions from the medfly (Ccap: *Ceratitis capitata*), the mosquito (Aaeg: *Aedes aegypti*), the honeybee (Amel: *Apis mellifera*), and the beetle (Tcas: *Tribolium castaneum*) are also shown

Different Evolutionary Patterns in the RNA-Binding Domains and the Two Termini of the SXL Protein

Finally, aiming to understand the molecular basis of the functional difference of the SXL proteins between *Drosophila* and the medfly or the housefly (Saccone et al. 1998; Meise et al. 1998), we aligned all the surveyed protein sequences (Fig. S3 (online ESM1)), including the SXLs and SSXs. The alignment reveals that the RNA-binding domains in the middle region of the proteins are quite conserved while the two termini show greater divergence among species. This fact is particularly evident for the SSXs whose sequences have numerous insertions, deletions, and substitutions (Fig. S3 (online ESM1)). However, conserved segments in the two termini of the *Drosophila Sxl* orthologues are observed. To quantify this variation, we plot the nonsynonymous substitution rate between *D. melanogaster* and each of another species against their species divergence

Table 2 Conserved motifs in the putative promoter upstream of the embryo-specific exons

Morif ID	Motif E-value	Number of *Drosophila* species with this motif[a]	Average count per *Drosophila* species	Count in medfly
1	2.8e−141	11	4.73	0
2	1.1e−109	11	6.27	1
6	5.4e−55	11	5.55	0
TAGteam	N/A[b]	11	6.27	2

[a] The sequence of *D. simulans* is not included here as the embryo exon is not identified due to the unknown bases in the genomic region

[b] The TAGteam motif consensus is from the literatures, not by the software MEME

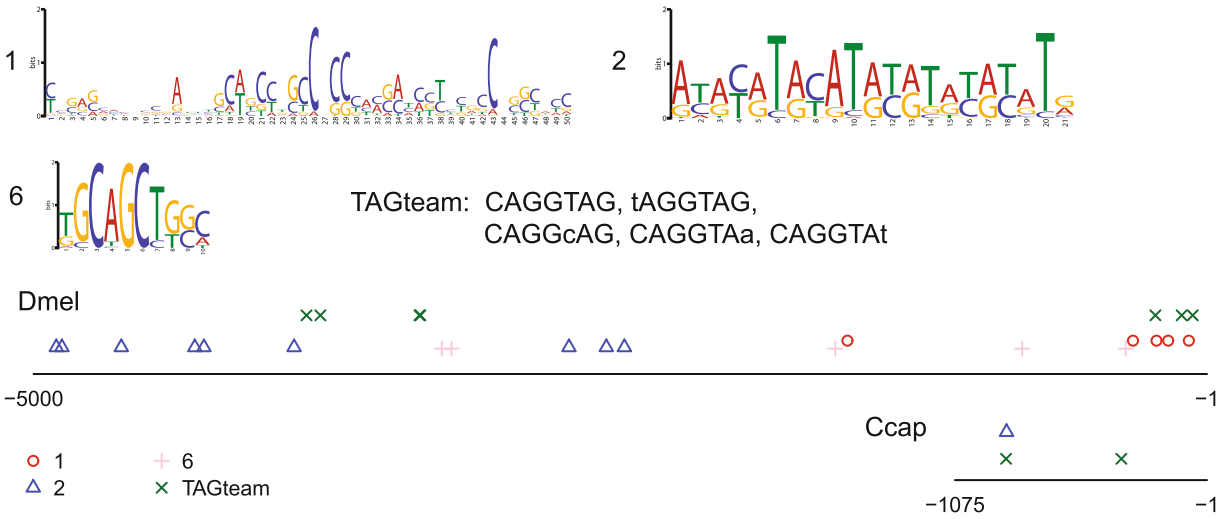


Fig. 6 The distributions of the conserved motifs in the 5,000-nucleotide region upstream of the embryo-specific exon in *D. melanogaster* and in the 1,075-nucleotide intron upstream of exon 2 in the medfly (Ccap: *Ceratitis capitata*). Each motif is represented by *one different symbol*

time. As shown in Fig. 7a–c, the substitution rates are low between *Drosophila* species but increase rapidly when compared to the medfly and *Calyptratae*. After that, the rates increase slowly and linearly with the divergence time. This pattern suggests that there are some large changes in the *Sxl* after the split of *Drosophila* and the close outgroup species. The increases of substitution rates from *Drosophila* to other species are more dramatic in the two termini than in the RNA-binding domains (Fig. S4 (online ESM1); Wilcoxon rank sum test, $P = 0.0027$). The synonymous substitution rates do not show the same pattern (Fig. 7d–f), suggesting that the observed patterns of non-synonymous substitution rates are not caused by different mutation rates among regions and species. These results indicate that the N- and C-termini of SXL proteins are much more different between *Drosophila* and other species than between different *Drosophila* species.

To see what kind of amino acid changes happened after the split between *Drosophila* and other insects, we inferred amino acid substitutions on the branch leading to the ancestor of *Drosophila* in the phyologenetic tree (see "Materials and Methods" section). Six sites in the RNA-binding domain region changed in the *Drosophila* ancestor (Fig. S3 (online ESM1)). The replaced amino acids of these sites have similar properties with the original ones except for the two at positions 228 (Q → S) and 290 (H → A) which involve charged amino acids, arguing that the function of this region may have largely been conserved. To further explore the potential functional effect of these changes (for example, directly altering RNA-binding), five of the six sites were mapped onto the 3D structure of the two RBDs of *D. melanogaster* (Handa et al. 1999). As shown in Fig. S5 (online ESM1), only one of these sites is located in the beta-sheets (known to interact with target RNAs) and is far from the backbone of the RNA (Handa et al. 1999). The absence of the interaction between these sites and RNA is further confirmed by checking the known RNA-interacting sites from a previous study (Handa et al. 1999). These data argue that the RNA-binding capacity of SXL may not have changed after divergence from the medfly/housefly. The effect of these changes on protein interactions remains to be evaluated experimentally.

For the two termini, the amino acid substitutions occurred in the *Drosophila* common ancestor often caused charge changes (Fig. S3 (online ESM1)). At present, there are no 3D structures for these regions, so their effects on protein functions cannot be inferred.

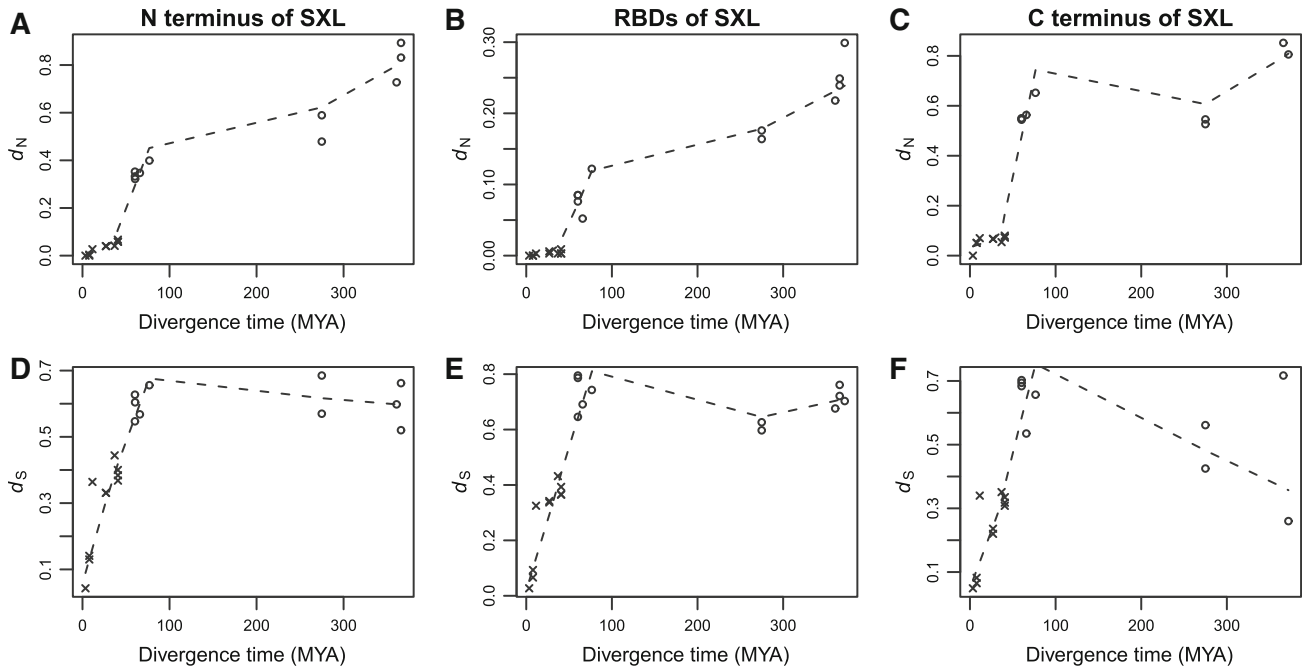


Fig. 7 The plot of the substitution rates versus the species divergence time. The *Sxl* CDS alignment is divided into three regions: the N-terminus (**a** and **d**), two RBDs and the middle linker sequence (**b** and **e**), and the C-terminus (**c** and **f**). In each *panel*, the substitution rate between *Drosophila melanogaster* and another species is plotted against their divergence time (see Fig. S1 (online ESM1) for the values and sources). In **a**–**c**, the nonsynonymous substitution rates (*p* distance, i.e., the proportion of different sites) are displayed, while in **d**–**f**, the synonymous substitution rates are used. The *cross symbols* denote the values between *D. melanogaster* and another *Drosophila* species

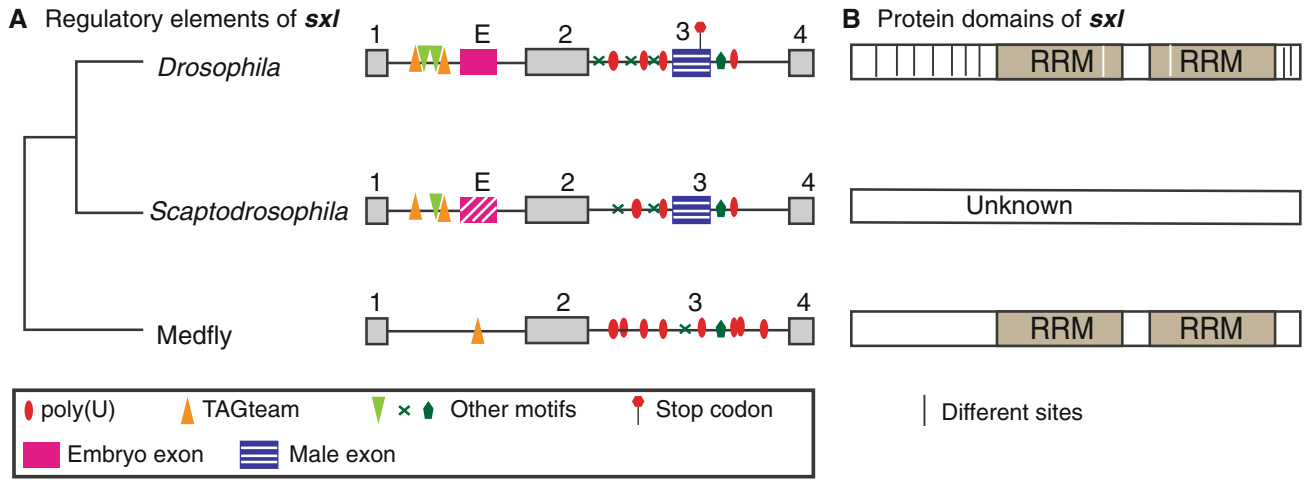


Fig. 8 The diagram of phylogenetic distributions of the *Sxl* regulatory elements (**a**) and of the protein sequence changes (**b**). The number of each *motif* in each species indicates the abundance relative to other species, not the absolute abundance. The *vertical lines* in the box representation of protein sequence indicate different sites between *Drosophila* and medfly. The embryo exon with *shading lines* indicates that it is inferred only

Discussion

In this study, we examined systematically the phylogeny of regulatory elements engaged in the *Sxl*'s sex-specific regulation (Fig. 8a). Our results reveal that: (1) the male-specific exon is present in *Drosophila* and *Scaptodrosophila* species; (2) the embryo-specific exon might have similar distribution but this conclusion needs experimental corroboration in *Scaptodrosophila*; (3) the male-specific exon in *S. lebanonesis* does not have any in-frame stop codons; (4) the poly(U) and TAGteam motifs have broader phylogenetic distributions, being present in medfly and other species; and (5) some conserved motifs are restricted to *Drosophila* and *Scaptodrosophila* (Tables 1, 2). Taken

together, a more parsimonious explanation is that the male-specific exon and probably the embryo-specific exon originated after the divergence between the Drosophilidae and Tephritidae families but before the split of *Drosophila* and *Scaptodrosophila* genera, and that the poly(U) and TAGteam may have preceded the emergence of these exons. The functions of the newly discovered motifs need closer experimental investigation in the future to unveil their functions.

Introducing translational stops in the male-specific exon of *D. melenogaster* is essential for sex-specific expression of *Sxl*. The lack of stop codons in the male-specific exon of *S. lebanonensis* questions the function of the *Sxl* orthologue. After carefully examining the sequences of male-specific exons, we find two possible ways for this *S. lebanonensis* exon to introduce translational terminations. First, each male-specific exon (except the one of *D. mojavensis*) has an alternative 3′ splice site downstream of the defined splice site (Fig. 3; Table S3 (online ESM1)) which gives a short-form male-specific exon. All the short-form exons, including the one of *S. lebanonensis*, contain in-frame stop codons (Table S3 (online ESM1)). An immediate question about the two 3′ splice sites is which is more often used. By looking at the splice site scores, it seems that the downstream splice site is preferred in splicing. The reason why have two 3′ splice sites is unclear. Second, the long-form male-specific exon of *S. lebanonensis* may introduce stop codons in the downstream exons due to frameshifts. Note that the length of this exon is 205 nucleotides, which gives a different translation frame for the downstream exons compared with the direct ligation of exon 2 to exon 4 (the first complete codon of exon 4 starts from the first base vs. from the second base). Thus it is possible that there might be in-frame stop codons in the downstream exons. However, with the available 30-nt segment of exon 4 we do not find any stop codon. The complete cDNA sequence is needed to test this possibility. Altogether these facts suggest that the male-specific exon of *S. lebanonensis* may function for the same purpose as those of *Drosophila*. However, further experiments are needed to verify these hypotheses. It would be interesting to know whether there is a functional difference of *Sxl* orthologues between *Drosophila* and *S. lebanonensis*. Since only one *Sxl* sequence of *S. lebanonensis* is available, the possibility of sequencing errors should also be taken into account.

On the other hand, the comparison of SXL protein sequences between *Drosophila* and other species revealed that the two termini of the protein had undergone more extensive changes than the RBD region after split from the medfly (Fig. 8b). This finding argues that the N- and C-termini may be responsible for the functional difference of SXL orthologues between *Drosophila* and other species. It has been reported that the *D. melanogaster* SXL N-terminus is essential for both *tra* splicing and *Sxl*'s auto-regulation. The 99 AA N-terminal fragment of *D. melanogaster* alone can promote female-specific splicing of *tra* pre-mRNAs (Deshpande et al. 1999) and the deletion of N-terminal 40 AA nearly eliminated the ability to female-specifically splice *tra* (Yanowitz et al. 1999). Moreover, the deletion of N-terminal 40 AAs severely compromised *Sxl*'s auto-regulatory function and the 99 AA long N-terminus can incorporate into the SXL:SNF splicing complex and interfere its normal function (Deshpande et al. 1999). The importance of N-terminus in splicing regulation is further emphasized by the fact that the male *Sxl* protein of *D. virilis*, which is different from the female protein at its N-terminal 25 amino acids, is unable to alter the splicing of *Sxl* and *tra* even though it can bind onto the pre-mRNAs (Bopp et al. 1996). Since the N-termini of the medfly and housefly homologs are significantly different from that of *Drosophila*, it is understandable why they do not induce feminization in *D. melanogaster* (Saccone et al. 1998; Meise et al. 1998). This also implies that the poly(U) tracts in the medfly intron, if functional, may not function in the same way as in *Drosophila* because of the different N-terminus of the medfly SXL protein. On the other hand, the N-terminal 40 AA truncation did not result in a significant impairment of dosage compensation (Yanowitz et al. 1999), raising the possibility that the N-terminus is not essential for this function. However, the transgenic expression of housefly homolog in *D. melanogaster* males did not change expression of the gene *msl-2* (Meise et al. 1998), the key regulator for dosage compensation (Fig. 1a). These two findings together argue that the partial N-terminal sequence near the RBDs may be needed for regulation of *msl-2*. Our results also suggest that the SXL C-terminus may have important functions yet to be discovered.

Finally, although all the *Drosophila* species seem to use the same regulatory system for sex-specific regulation, we observe variations in the regulatory elements. (1) The embryo-specific exon in *D. mojavensis* contains in-frame stop codons that would result in translation terminations if joining with the downstream exons, though experiments are needed to exclude the possibility of sequencing errors. The male-specific exon of *D. mojavensis* is also very different from others at the 5′ end and missing the downstream alternative 3′ splice site (Fig. 3). (2) The 5′ splice site of the male-specific exon in *D. ananassae* is weaker than the sites in other *Drosophila* species where the male exon 5′ splice sites have the same nucleotides, suggesting strong constraint on this site. The variant in *D. ananassae* might affect its regulation. (3) The copy number of each motif also varies among *Drosophila* species. For instance, *D. mojavensis* has fewer poly(U) in the upstream intron of the male exon than other species (Dataset S2 (online

ESM3)). These variations argue that although all the *Drosophila* species use the sex-specific splicing to regulate *Sxl* expression, modifications of the mechanism may have occurred in some species.

A remaining open question is the mechanisms of the origins of the male-specific exon and the embryo-specific exon. In theory, new exons can be generated by duplication or exonization of intronic sequences (Sorek 2007). If these exons are the products of exon duplication, we may find their parental copies by searching the whole genome sequence. However, we do not see any significant hit in other genomic regions for the male- and embryo-specific exons in the *D. melanogaster* genome. A challenge here is that the exons in extant species have already greatly diverged from the copy in the ancestral species of *Drosophila*. Unless the ancestral *Drosophila* genome can be accurately re-constructed, it will be difficult to identify the parental copies of the exons. With the current data the exonization remains a viable explanation.

In conclusion, our study reveals a clear picture of phylogenetic distribution of regulatory elements involved in the *Sxl*'s auto-regulation as well as improves the understanding of molecular basis for *Drosophila* SXL's functional difference from other orthologues. These results support the idea that *Drosophila Sxl* acquired the novel sex-determination function through both expansion of the sex-specific regulatory network and alteration of the protein sequence. The information lays a foundation for future studies of *Sxl*'s functional evolution and ultimately for unraveling the origin of the sex-determination system.

Acknowledgments We thank Prof. Thomas Cline for sending us the genomic sequences near the male-specific exon and Dr. Hielim Kim for comments on the early version of the manuscript. We thank Prof. Claude dePamphilis and his lab members for useful comments on our work. We thank Dr. Alfred M. Handler at the USDA-ARS for sending us the link of the medfly genomic sequence.

Conflict of interest The authors declare that they have no conflict of interest.

References

Bailey TL, Elkan C (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 2:28–36

Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202-208. doi:10.1093/nar/gkp335

Bell LR, Maine EM, Schedl P, Cline TW (1988) Sex-lethal, a *Drosophila* sex determination switch gene, exhibits sex-specific RNA splicing and sequence similarity to RNA binding proteins. Cell 55(6):1037–1046

Bopp D, Calhoun G, Horabin JI, Samuels M, Schedl P (1996) Sex-specific control of Sex-lethal is a conserved mechanism for sex determination in the genus *Drosophila*. Development 122(3):971–982

Cline TW, Dorsett M, Sun S, Harrison MM, Dines J, Sefton L, Megna L (2010) Evolution of the *Drosophila* feminizing switch gene Sex-lethal. Genetics 186(4):1321–1336. doi:10.12120210.1534/genetics.110.121202

Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39:D301–D308. doi:10.1093/nar/gkq1069

Deshpande G, Calhoun G, Schedl PD (1999) The N-terminal domain of Sxl protein disrupts Sxl autoregulation in females and promotes female-specific splicing of tra in males. Development 126(13):2841–2853

Drysdale R (2008) FlyBase : a database for the *Drosophila* research community. Methods Mol Biol 420:45–59. doi:10.1007/978-1-59745-583-1_3

Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJ, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SM (2012) Ensembl 2012. Nucleic Acids Res 40:D84–D90. doi:10.1093/nar/gkr991

Handa N, Nureki O, Kurimoto K, Kim I, Sakamoto H, Shimura Y, Muto Y, Yokoyama S (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. Nature 398(6728):579–585. doi:10.1038/19242

Horabin JI, Schedl P (1993a) Regulated splicing of the *Drosophila* Sex-lethal male exon involves a blockage mechanism. Mol Cell Biol 13(3):1408–1414

Horabin JI, Schedl P (1993b) Sex-lethal autoregulation requires multiple cis-acting elements upstream and downstream of the male exon and appears to depend largely on controlling the use of the male exon 5′ splice site. Mol Cell Biol 13(12):7734–7746

Kandul NP, Noor MA (2009) Large introns in relation to alternative splicing and gene evolution: a case study of *Drosophila* bruno-3. BMC Genet 10:67. doi:10.1186/1471-2156-10-67

Meise M, Hilfiker-Kleiner D, Dubendorfer A, Brunner C, Nothiger R, Bopp D (1998) Sex-lethal, the master sex-determining gene in *Drosophila*, is not sex-specifically regulated in *Musca domestica*. Development 125(8):1487–1494

Miyasaka H (1999) The positive relationship between codon usage bias and translation initiation AUG context in *Saccharomyces cerevisiae*. Yeast 15(8):633–637. doi:10.1002/(SICI)1097-0061(19990615)15:8<633:AID-YEA407>3.0.CO;2-O

Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C (1992) Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. Nucleic Acids Res 20(16):4255–4262

Mullon C, Pomiankowski A, Reuter M (2012) Molecular evolution of *Drosophila* Sex-lethal and related sex determining genes. BMC Evol Biol 12:5. doi:10.1186/1471-2148-12-5

Penalva LO, Sanchez L (2003) RNA binding protein Sex-lethal (Sxl) and control of *Drosophila* sex determination and dosage compensation. Microbiol Mol Biol Rev 67(3):343–359

Penalva LO, Sakamoto H, Navarro-Sabate A, Sakashita E, Granadino B, Segarra C, Sanchez L (1996) Regulation of the gene Sex-lethal: a comparative analysis of *Drosophila melanogaster* and *Drosophila subobscura*. Genetics 144(4):1653–1664

Saccone G, Peluso I, Artiaco D, Giordano E, Bopp D, Polito LC (1998) The Ceratitis capitata homologue of the *Drosophila* sex-

determining gene Sex-lethal is structurally conserved, but not sex-specifically regulated. Development 125(8):1495–1500

Salz HK (2011) Sex determination in insects: a binary decision based on alternative splicing. Curr Opin Genet Dev 21(4):395–400. doi:10.1016/j.gde.2011.03.001

Salz HK, Maine EM, Keyes LN, Samuels ME, Cline TW, Schedl P (1989) The Drosophila female-specific sex-determination gene, Sex-lethal, has stage-, tissue-, and sex-specific RNAs suggesting multiple modes of regulation. Genes Dev 3(5): 708–719

Satija R, Bradley RK (2012) The TAGteam motif facilitates binding of 21 sequence-specific transcription factors in the *Drosophila* embryo. Genome Res 22(4):656–665. doi:10.1101/gr.130682.111

Sievert V, Kuhn S, Paululat A, Traut W (2000) Sequence conservation and expression of the Sex-lethal homologue in the fly *Megaselia scalaris*. Genome 43(2):382–390

Sorek R (2007) The birth of new exons: mechanisms and evolutionary consequences. RNA 13(10):1603–1608. doi:10.1261/rna.682507

Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE (1998) Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D 54(Pt 6 Pt 1):1078–1084

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28(10):2731–2739. doi:10.1093/molbev/msr121

ten Bosch JR, Benavides JA, Cline TW (2006) The TAGteam DNA motif controls the timing of *Drosophila* pre-blastoderm transcription. Development 133(10):1967–1977. doi:10.1242/dev.02373

Thompson JD, Gibson TJ, Higgins DG (2002) Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2, Unit 2.3. doi:10.1002/0471250953.bi0203s00

Traut W, Niimi T, Ikeo K, Sahara K (2006) Phylogeny of the sex-determining gene Sex-lethal in insects. Genome 49(3):254–262. doi:10.1139/g05-107

UniProt Consortium (2012) Reorganizing the protein space at the Universal Protein Resource (UniProt). Nucleic Acids Res 40:D71–D75. doi:10.1093/nar/gkr981

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ (2009) Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 25(9):1189–1191. doi:10.1093/bioinformatics/btp033

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24(8):1586–1591. doi:10.1093/molbev/msm088

Yang D, Lu H, Hong Y, Jinks TM, Estes PA, Erickson JW (2001) Interpretation of X chromosome dose at Sex-lethal requires non-E-box sites for the basic helix-loophelix proteins SISB and daughterless. Mol Cell Biol 21(5):1581–1592. doi:10.1128/MCB.21.5.1581-1592.2001

Yanowitz JL, Deshpande G, Calhoun G, Schedl PD (1999) An N-terminal truncation uncouples the sex-transforming and dosage compensation functions of Sex-lethal. Mol Cell Biol 19(4):3018–3028