

Article

Not peer-reviewed version

Recurrent Duplication, Testis-Biased Expression, and Functional Diversification of *Esf2/ABT1* Family Genes in *Drosophila*

[Elizaveta D. Davydova](#)[†], [Alexei A. Kotov](#)[†], Alina V. Chernizova, [Ekaterina Yu. Yakovleva](#), [Ludmila V. Olenina](#)^{*}

Posted Date: 19 August 2025

doi: 10.20944/preprints202508.1302.v1

Keywords: gene duplication; sex-biased expression; *Drosophila*; piRNA cluster; ABT1-like domain; Intrinsically Disordered Regions



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Article

Recurrent Duplication, Testis-Biased Expression, and Functional Diversification of *Esf2/ABT1* Family Genes in *Drosophila*

Elizaveta D. Davydova^{1,†}, Alexei A. Kotov^{1,†}, Alina V. Chernizova², Ekaterina Yu. Yakovleva¹, Ludmila V. Olenina^{1,*}

¹ Laboratory of Functional Genomics, Koltzov Institute of Developmental Biology of Russian Academy of Sciences, 26 Vavilov str., 119334 Moscow, Russia

² Laboratory of Cellular Neurobiology of Learning, Institute of Higher Nervous Activity and Neurophysiology, of Russian Academy of Sciences, 5A Butlerov str., 117485 Moscow, Russia

[†] These authors contributed equally to this work.

* Correspondence: olenina_ludmila@mail.ru

Simple Summary

To investigate the evolutionary history of *Esf2/ABP1* gene duplicates in *Drosophila* species, we performed a search and subsequent characterization of *Esf2/ABP1* copies using comparative genomic analysis, phylogenetic analysis, functional domain analysis, and tissue-specific transcriptomic analysis. We determined that despite the single copy of the *Esf2/ABP1* gene per genome harbored in the majority of *Drosophila* species, recurrent gene duplication events occurred in four species of the *melanogaster* subgroup and independently in three species of the *suzukii* subgroup. DNA-mediated duplications took place on the X chromosome in close proximity to the parental copy or involved translocation and amplification of duplicated copies within the introns of unrelated genes or within heterochromatic piRNA cluster. We found that novel functions of duplicated copies emerged through various evolutionary processes, leading to the development of new adaptations, including testis-enriched expression for the beneficial increase of gene dosage and protein domain acquisition and diversification.

Abstract

Gene duplications are considered to be the major evolutionary resource of novel functions. The gene family *Esf2/ABP1* is conserved in metazoan organisms from yeast to humans. Here we performed a search and characterization of *Esf2/ABP1* homologs in the *Drosophila* genus. Whereas in the majority of *Drosophila* species this gene family is represented by only a single gene, in the *melanogaster* and *suzukii* subgroups recurrent gene duplications arose, providing 47 homologous genes located on the X chromosome. To study the evolutionary history of duplicates, we performed phylogenetic, functional domain, and tissue-specific expression analyses. We revealed a male-specific and testis-biased transcription pattern of duplicated copies in *Drosophila melanogaster* and *Drosophila sechellia* compared to ubiquitous expression of the parental gene. The amplification of 21 repeated paralogs within the heterochromatic piRNA cluster resulted in the ovarian-specific transformation of these repeats into piRNAs in *D. melanogaster*. In three species of the *suzukii* subgroup, *Esf2/ABP1* genes evolved with domain diversification: in addition to RNA-binding ABT1-like domain preservation, all homologous proteins acquired expanded intrinsically disordered regions. By studying the duplicated copies of the *Esf2/ABP1* family in *Drosophila*, we offer insight into how novel gene functions emerge and are maintained, contributing to life's diversity and complexity.

Keywords: gene duplication; sex-biased expression; *Drosophila*; piRNA cluster; ABT1-like domain; Intrinsically Disordered Regions

1. Introduction

Gene duplication followed by divergence is now considered one of the fundamental mechanisms driving the evolution of new functions in metazoan organisms. Duplicated genes provide substrates from which evolution can generate novel functions, and thus duplications contribute to processes of environmental adaptation, speciation, and increasing organismic complexity. Gene duplication can be raised in different ways, including tandem gene duplication, whole-genome and segmental chromosomal duplication, and also retroposition or transposition. Following gene duplication, a newly duplicated gene copy can realize different fates. Most often it is a pseudogenization with accumulation of degenerative changes, e.g., mutations, frameshifting deletions, or premature stop codons owing its functional redundancy [1,2]. However, another part of duplicated genes could acquire new functions and be retained in the genome.

According to the modern view, multiple processes can functionally preserve duplicated genes, including conservation, neofunctionalization, subfunctionalization, specialization, dosage benefit, duplication-degeneration-complementation, and others [3-6]. When increased gene dosage is beneficial, ancestral functions are conserved in both gene copies. During neofunctionalization, one gene copy carries out its ancestral functions, and the other one acquires a novel function and expression pattern [2,3,6]. Subfunctionalization process means that copies are mutated differently with a maintenance of ancestral gene functions by both genes [2,4]. Concerted action of subfunctionalization and neofunctionalization processes provides a specialization of gene copies that functionally evolved from each other and from the ancestral gene.

The gene family *Esf2/ABP1* includes genes that are phylogenetically conserved in metazoan organisms from yeast to humans. These genes encode the nuclear protein ABT1 (activator of basal transcription 1) and its yeast homolog, pre-rRNA-processing protein Esf2 (eighteen S factor 2). The molecular function of ABT1 is described as a TBP-binding regulator of basal transcription for PolIII genes in mice [7]. In *Saccharomyces cerevisiae*, Esf2 is known as a nucleolar RNA-binding protein involved in pre-rRNA processing and the biogenesis of small ribosomal subunit [8]. The functions of Esf2 were found to be essential in yeast [7]. Despite the molecular functions of Esf2/ABP1 proteins in different metazoan organisms not being clearly understood yet, in *Drosophila melanogaster* the *Esf2/ABP1* family gene CG32708 has been found to be essential for fly development and associated with the Notch signaling pathway and cell proliferation [9-12].

The *Drosophila* represents a clear and attractive model system to study the evolutionary forces acting on the evolutionary dynamics of gene copies [13-17]. The recent appearance of high-quality genome assemblies of *Drosophila* species, including challenging heterochromatin genomic regions with arranged genomic repeats [18-20], provides a valuable opportunity to understand genetic and evolutionary mechanisms that drive the maintenance and divergence of *Esf2/ABP1* family genes. In this study we report our initial functional and evolutionary characterization of *Esf2/ABP1* gene family duplications in genomes of *Drosophila*. We found that in the majority of *Drosophila* species, this gene family is represented by only a single gene maintaining original functions. However, in the *melanogaster* subgroup as well as in the *suzukii* subgroup, we identified recurrent gene duplications providing, in sum, at least 47 homologous genes expanding to distinct genomic locations on the X chromosome. Investigating the origin, divergence, phylogenetic relationships, and expression pattern of *Esf2/ABP1* genes, we revealed a sex-specific expression bias of duplicated copies in *Drosophila melanogaster* and *Drosophila sechellia* and functional diversification of homologs in the *suzukii* subgroup species, indicating evolutionary adaptation of newly raised gene copies.

2. Materials and Methods

2.1. Homolog search and identification

We performed a search of *Esf2/ABP1* homologs and duplicates in the *Drosophila* genus using Ensemble Metazoa tool (<https://metazoa.ensembl.org/>) and CG327708 gene as a query. The analysis of the identity of homologs was performed using the coding and protein sequences identified by Ensembl (<https://ftp.ebi.ac.uk/pub/databases/ensembl/genomes/release-60/metazoa/fasta/>). Sequences were pairwise aligned using MUSCLE v3.8.1551 [21]. The percent identity was calculated as the number of matching bases/amino acids divided by the number of alignment columns, excluding gaps. The data were visualized as the heat map using the Python libraries, Matplotlib and Seaborn. For microsynteny estimation we analyzed up to three upstream and three downstream surrounding genes for each location using NCBI Genome Data Viewer (<https://www.ncbi.nlm.nih.gov/gdv?org>) and corresponding genome assemblies. Genes are considered to be in a syntenic location if they share more than 15% flanking homology.

2.2. Phylogenetic analysis

For search *Esf2/ABP1* family homologous genes 23 whole genome assemblies of *Drosophila* species (NCBI RefSeq) were used with BLAST (blastn) algorithm and CG327708 gene as a query. Corresponding genome assemblies are listed in Table S1. For multiple alignment we retrieved the CDS for homologous genes of closely related species with reference genome assemblies include *Drosophila simulans* Prin_Dsim_3.1 GCA_016746395.2, *Drosophila mauritiana* ASM438214v1 GCA_004382145.1, *Drosophila sechellia* ASM438219v2 GCA_004382195.2, *D. yakuba* Prin_Dyak_Tai18E2_2.1 GCA_016746365.2 and *Drosophila melanogaster* Release 6 plus ISO1 MT GCA_000001215.4. The nucleotide sequence alignment was carried out using MEGA 12 software [22] by the MUSCLE method [21] for codons. The alignments required minimal manual editing and G-blocks 0.91.1 [23] treatment with following parameters: data type: codon; max number of contiguous non-conserved positions: 8; min length of a block: 10; allowed gap position: with half. Phylogenetic tree constructions were performed using the MEGA 12 software. The Maximum Likelihood method was used with the Tamura-Nei model of nucleotide substitutions [24] for the tree generation with highest log likelihood (-1933.81) and *Drosophila yakuba* sequence of LOC6525244 gene as the outgroup. The percentage of replicate trees in which the associated taxa clustered together (500 replicates) is shown below the branches. Gamma distribution was used to model evolutionary rate differences across 5 categories (+G, parameter = 1.9616), with 0.00% of sites deemed evolutionarily invariant (+I). The partial deletion option was applied to eliminate all positions with less than 95% site coverage resulting in a final data set comprising 351 positions. The analytical procedure encompassed 42 coding nucleotide sequences using 1st, 2nd, 3rd, and non-coding positions. Evolutionary analyses were conducted in MEGA 12 [22] utilizing up to 4 parallel computing threads.

2.3. Z-test of neutrality

To estimate the type of selection for the *Esf2/ABP1* family genes in *D. melanogaster*, we applied the codon-based test of neutrality, the Z-test, using MEGA 11 software [22, 25]. Analyses were conducted using the Nei-Gojobori method [26]. This analysis involved two aligned nucleotide sequences. All ambiguous positions were removed for each sequence pair (pairwise deletion option). The test statistic ($dN - dS$) is represented as a Z-score, where dS and dN are the numbers of synonymous and nonsynonymous substitutions per site, respectively.

2.4. Search for protein domains

For the determination of known protein domains of *Esf2/ABP1* family proteins, NCBI's Conserved Domain Database (CDD) tool (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>) [27] was used. The search was performed against database CDD v3.21 with an expected value threshold of 0.01. Only specific hits were considered. Intrinsically disordered regions (IDRs) were predicted with the Critical Assessment of Intrinsic Protein Disorder (CAID) Prediction Portal

(<https://caid.idpcentral.org/>) [28] with a default threshold of 0.5. The overall disorder scores for the proteins expressed as 50% of IDRs in the whole protein length were considered as significant.

2.5. Fly stocks

To analyze transcript level in different fly tissues and copy number estimation we used flies of wild-type strains of *D. melanogaster* Batumi L, Harwich, GH10, and GH16. Batumi L and Harwich strains were obtained from the Collection of laboratory and natural strains of Drosophila of Institute of Developmental Biology, RAS, Russia. Strains from African population (Ghana, Africa), GH10 and GH16 [29], were obtained from the Drosophila collection of Institute of Cytology and Genetics, SB RAS, Russia. *D. sechellia* stock was obtained from the Gifsur-Yvette CNRS Center collection, France. All flies were reared at 23°C on a standard medium. The flies were subjected to a laboratory light-dark cycle of 12:12 h.

2.6. RT-qPCR analysis

For RT-qPCR analysis, total RNA was isolated from gonads, heads and carcasses of 3-5-day adult male and female flies with TRIzol Reagent (Invitrogen). DNA-free Kit DNA Treatment & Removal (Invitrogen) kit was used for removal of genomic DNA from preparations. Reverse transcription was performed using Mint kit (Eurogen) and Oligo(dT) primers. cDNA samples were analyzed by real-time quantitative PCR using the incorporation of SYTO13 (Invitrogen). All experiments were performed with at least three independent RNA samples; each sample was analyzed in triplicate. Fold change ratios of the average expression level were calculated. We used *rp49* (*rpL32*) as a loading control. The following primers were used: *rp49* fw 5'-ATGACCATCCGCCAGCATAC-3', rev 5'-GCTTAGCATATCGATCCGACTGG-3'; *cuckoo* fw 5'-CATGGAAGTAGAGGAGGCTGAG-3', rev 5'-GGTATGTTGGATATGTAGATGATCCC-3'; *chaffinch* (shared) fw 5'-AAATGACGATAAAAAGGAGCTGG-3', rev 5'-TGTCCTTGGGTATGTTGGATATTA-3'; *brambling* fw 5'-CTGAGATGCCGTTCTCTTATAAGACG-3', rev 5'-CAAGGTCATGTGCTTGGGTAGGT-3'; *woodpecker* (shared) fw 5'-GGAGGAGATTGAGACCTCGG-3', rev 5'-GTGCTTGGGTATGTTGGATATGT-3'; *D. sechellia* parental copy LOC6619759 fw 5'-GGTGACAAGCAACGCCTA-3', rev 5'-CAGAGCCTTTCCGCTTTCTT-3'; *D. sechellia* duplicated copies of cluster 1 (LOC6621826, LOC116802136, LOC116802134, LOC116802120, LOC6619761) (shared) fw 5'-CTGAGAAACCCGAGAAGATGCA-3', rev 5'-TTAACAGCTCTTTCGGCTGCAC-3'; *D. sechellia* cluster 2 duplicates (LOC6618463, LOC116801944, LOC6618464, LOC6618462) (shared) fw 5'-AACGAACAAAATGAATTCTGAGCC-3', rev 5'-CATTCGCAGACTCACCTTCCTC-3'.

2.7. Copy number estimation of woodpecker genes

To estimate the number of *woodpecker* gene copies in the *D. melanogaster* genome, quantitative PCR was performed with genomic DNA of Batumi L, Harwich, GH10, and GH16 wild-type strains to determine the concentration level of amplification products of each of the genes of the family. We used highly specific *woodpecker* primers (see above), the efficiency of which was previously tested and compared with the efficiency of *rp49* primers (normalization control). Measurements for each strain were carried out in three replicates.

2.8. Transcriptomic analyses of Esf2/ABP1 family genes using deep sequencing library data

PolyA+ RNA-seq libraries from *w¹¹¹⁸* strain of *D. melanogaster* embryos, third instar larval gonads and sex-separated adult tissues/body parts and gonads of *w¹¹¹⁸* and *aub* mutant flies [30-35] (Table S2) were used for expression analysis (from duplicate to quadruplicate of libraries for tissue). The libraries were pseudo-aligned using Kallisto v.0.50.0 with following parameters --single -l 100 -s 0.05 [36]. The *D. melanogaster* transcriptome set obtained from NCBI was used as a reference (GCF_000001215.4). Transcript abundances were used to assess expression in terms of transcripts per million (tpm). The data were visualized using the Python libraries, Pandas, Matplotlib, and Seaborn.

2.9. Bioinformatics analysis of piRNAs

For piRNA distribution analysis in the *AT-chX* cluster, small RNA libraries from *Oregon R* ovaries [37] and *w¹¹¹⁸* (*yw*) testes [38] of *D. melanogaster* were used (Table S2). Adapter sequences from small RNA libraries were removed by cutadapt 2.8. The FastQC tool was used for quality control of the sequenced libraries. Reads were filtered by length (23-30) and quality (Phred quality score > 30) using fastq-filter (<https://github.com/LUMC/fastq-filter?tab=readme-ovfile#readme>). rRNA, snRNA, snoRNA, microRNA and tRNA were filtered off from the libraries. To analyze the reads generated from the *AT-chX* and *woodpecker* repeat containing piRNA cluster, locally unique reads (reads could only be mapped to genomic region with coordinates chrX:21631000-22440000 containing *AT-chX* repeats, *woodpecker* repeats, and neighboring *flamenco* piRNA cluster) were used. For this, libraries were mapped to the *dm6* genome assembly without contigs and chrX:21631000-22440000 region. Unmapped reads were remapped to the chrX:21631000-22440000 region using bowtie with -n 1 -l 30 parameters. The mapped reads were visualized using Integrative Genomics Viewer (IGV) (<https://www.broadinstitute.org/scientific-community/software/integrative-genomics-viewer>). To evaluate the potential silencing of *Esf2/ABP1* family genes by piRNAs, the *Oregon R* ovarian library was independently mapped to gene sequences using bowtie (-n 2). The mapped reads were counted and normalized. Ping-pong signatures for sense and antisense piRNA pairs were calculated using the signature.py script [39].

3. Results

3.1. Survey of recurrent duplications of *Esf2/ABP1* family genes in *Drosophila*

Multiple animal species represent genes belonging to the *Esf2/ABP1* family in their genomes as a single homologous copy per genome. Using Ensemble Metazoa tool (<https://metazoa.ensembl.org/>), we identified a single *Esf2/ABP1* gene per genome for the majority of *Drosophila* genus species; however, we revealed gene duplication events for several species of the *melanogaster* group (Figure S1). Among them there are four species of the *melanogaster* subgroup, *Drosophila melanogaster*, *Drosophila simulans*, *Drosophila mauritiana*, and *Drosophila sechellia*, and species of the *suzukii* subgroup, *Drosophila biarmipes* and *Drosophila subpulchrella*. Some of the *Esf2/ABP1* gene duplications for *D. melanogaster*, *D. simulans*, and *D. mauritiana* have been found earlier [40]. With pairwise sequence alignments between homologs, we revealed that high sequence homology is preserved across the genus *Drosophila* both for nucleotide sequences of their CDSs and for translated amino acid sequences (Figure 1).

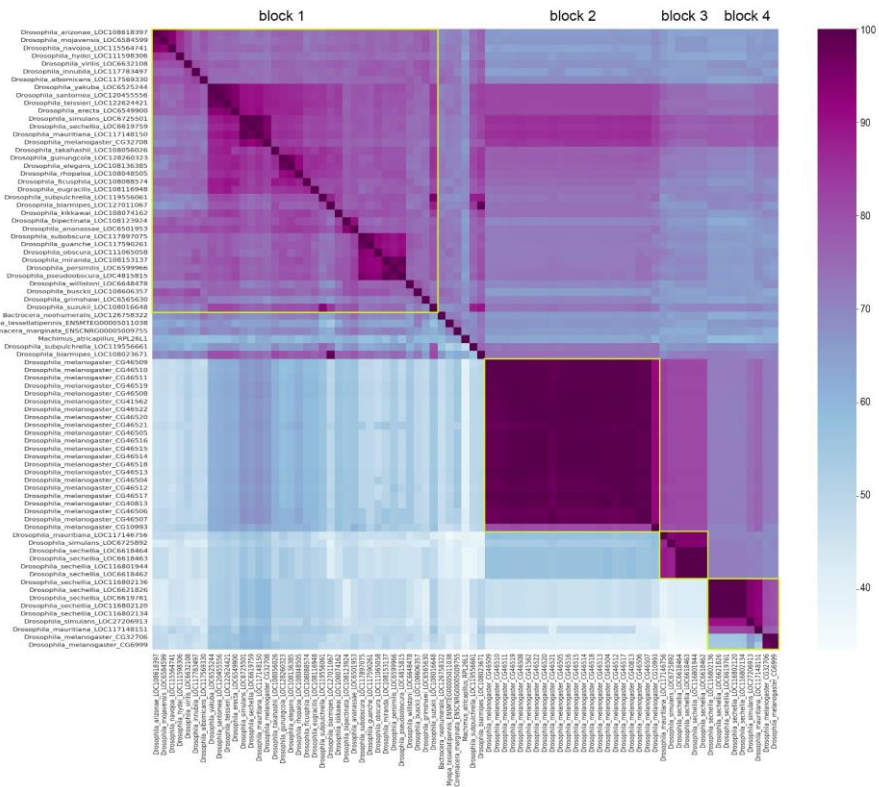


Figure 1. Heatmap profile of pairwise alignment by MUSCLE v3.8.1551 of sequences of *Esf2/ABP1* family homologs from species across the genus *Drosophila* (see Figure S1 for the phylogenetic tree). Data above the diagonal line show the nucleotide alignment identity of the CDS sequences. Data below the diagonal line show the protein alignment identity of the amino acid sequences. Identity percent for each pairwise alignment was calculated as the ratio of the number of matching bases/amino acids to the number of alignment columns, excluding gaps. *Machimus atricapillis* gene RPL26L1 was used as an unrelated control. Several data blocks are highlighted with rectangular yellow frames and designated on the top. Heatmap scale is represented on the right.

To characterize *Drosophila* *Esf2/ABP1* homologs in detail, we focused on 16 species of the *melanogaster* group of the subgenus *Sophophora* (~25 MIA divergence between species). By leverage of high-quality *Drosophila* genome assemblies (<https://www.ncbi.nlm.nih.gov/datasets/genome/>), we used blastn-based approach to identify *Esf2/ABP1* orthologs and paralogs across the *Sophophora* species and subjected all found genes to the synteny analysis (Figure 2, Table S1). Our analysis revealed that duplicated gene copies in *D. melanogaster* and species of the *simulans* clade are clustered at three distinct genomic locations on the X chromosome. The first region (cluster 1) is located within a gene-rich euchromatin environment between *APC4* (*Anaphase Promoting Complex subunit 4*) and *CCT2* (*Chaperonin containing TCP1 subunit 2*) genes (Figure 2). This region contains at least one copy of the *Esf2/ABP1* family genes in most analyzed genomes. This region of *D. melanogaster* (cytolocation 8C7) contains three paralogous *Esf2/ABP1* genes, CG32708, CG32706, and CG6999, as has been described earlier [40]. Among them, the upstream CG32708 gene has a high nucleotide identity with homologous genes of other *Drosophila* species (Figure 1 block 1). Orthologous genes of the *melanogaster* group that are predominantly located between *APC4* and *CCT2* genes on the X chromosome have nucleotide identity to CG32708 gene for their CDSs in the range of 75.2-95.4% (Figure 1, Table S1, Figure 2). CG32708 is characterized by ubiquitous expression in different fly tissues (according to FlyAtlas2 data) and is found to be the essential gene for fly development and cell proliferation [9-12]. Thus, the most parsimonious explanation is that CG32708 is the progenitor gene copy maintaining original functions in *D. melanogaster*. Taking all this into account, we assume that CG32708 and its orthologs in this location appear to be parental copies from which the following

paralogous genes originated in the *melanogaster* group. We found the parental copy and one duplicated copy in this genomic location in *D. simulans* and *D. mauritiana* genomes. However, in *D. sechellia* genome we observed in this region the parental copy with the highest homology 94.7% to CG32708 and downstream five almost identical tandem gene duplications with only 66.8-67.3% homology to CG32708 (Figure 1 block 4, Figure 2, Table S1). Considering that *D. melanogaster* has split from an ancient precursor with the *simulans* clade about 4.3-6.5 MYA [41-45], we proposed that first duplication event in this genomic region can occurred before *D. melanogaster* divergence from sibling species.

The second X chromosome region (cluster 2) containing *Esf2/ABP1* copies is the intron of the *InaE* (inactivation no afterpotential E) gene of the *melanogaster* subgroup (cytolocation 12C5 for *D. melanogaster*). Whereas, only single *Esf2/ABP1* orthologs were found for *D. melanogaster*, *D. simulans*, and *D. mauritiana* species in this genomic location; for *D. sechellia*, we observed four highly identical tandem duplicated paralogs (Figure 1 block 3, Figure 2, Table S1). The strong syntenic location of these genes also indicates that the initial gene insertion into the *InaE* intron and subsequent fixation occurred before the splitting of *D. melanogaster* from the species of the *simulans* clade.

We uncovered the third region (cluster 3) only in *D. melanogaster* genome, containing at least 21 *Esf2/ABP1* gene copies (Figure 2, Table S1, Figure 3A). This region is located in pericentromeric heterochromatin of the X chromosome (cytolocation 20B). It has been identified earlier as the expanded germline-specific piRNA cluster containing *AT-chX* repeats [33,38]. piRNA clusters are specialized genomic regions responsible for producing the majority of small non-coding piRNAs in germinal tissues [46,47]. The *Esf2/ABP1* paralogs in this region of *D. melanogaster* are arranged tandemly, with blocks of fused-together transposon fragments of repeating patterns placed between neighboring paralogous copies (Figure 3A). Despite their location within the active piRNA cluster, these duplicated copies are annotated as genes in the reference genome assembly; each copy contains one intron and possesses the highest homology among themselves (more than 99%) (Figure 1 block 2). We designated these repeats as *woodpecker* with corresponding numbers (*woodpecker 1*, *woodpecker 2*, ..., *woodpecker 21*) (Figure 3A). We estimated *woodpecker* gene copy number in four wild-type *D. melanogaster* strains (*Batumi*, *Harwitch*, *GH10*, and *GH16*) using quantitative PCR of genomic DNA (Figure 3B). There are 18-21 *woodpecker* copies in the analyzed genomes. The absence of these repeats in the genomes of other *Drosophila* species, including closely related ones of the *simulans* clade, indicates their recent arising, fixation, and maintenance only in *D. melanogaster*.

According to our analysis, duplication events for the *Esf2/ABP1* gene family have occurred also in the *suzukii* subgroup. Both *D. biarmipes* and *D. subpulchrella* genomes contain parental *Esf2/ABP1* copies with 76.2% and 81.5% nucleotide identity to CG32708, respectively, in the genomic location flanked upstream by *APC4* gene, and, besides that, *D. biarmipes* has a second paralog in this location (Figure S1; Figure 2 cluster 1, Table S1). *D. subpulchrella* genome also contains a single *Esf2/ABP1* copy in a non-syntenic region of the X chromosome, between *eIF3g1* and *Smyd3* genes (Figure 2 cluster 2, Table S1). Our search revealed that in their closely related species *D. suzukii*, the parental *Esf2/ABP1* gene near *APC4* gene has been apparently lost, but instead we identified one copy between *eIF3g1* and *Smyd3* genes, and additionally two duplicated copies, a gene and a pseudogene, in the intron of X-linked LOC108016561 gene encoding the ortholog of *D. melanogaster* CG42594 gene (Figure 2 clusters 2 and 3, Table S1).

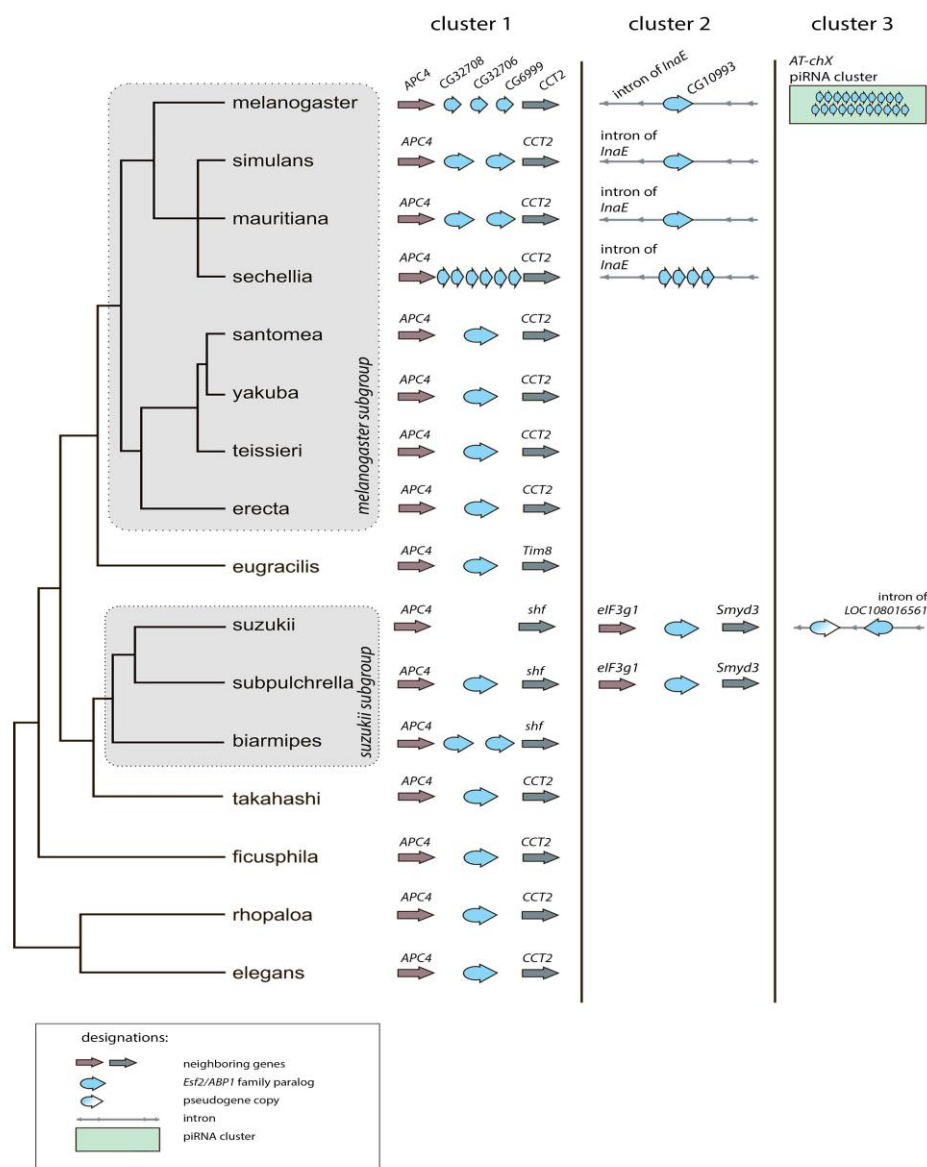


Figure 2. Survey of duplication events and synteny analysis of *Esf2/ABP1* homologs. The presence, copy number, and absence of homologous genes across the phylogeny of the *melanogaster* group are represented. The number of blue figures in each location indicates the copy number of *Esf2/ABP1* homologs. Synteny analysis showed that a single *Esf2/ABP1* copy is located on the X chromosome, mainly flanked by *APC4* and *CCT2* genes (see Table S1 for additional info). We found that four species of the *melanogaster* subgroup, *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia*, and three species of the *suzukii* subgroup, *D. suzukii*, *D. biarmipes*, and *D. subpulchrella*, are subjected to gene duplication events. We subdivided the genomic location of duplicated copies into three clusters for each subgroup as indicated on the top of the figure. Note that, except for cluster 1, the genomic locations of the clusters do not coincide for these subgroups except for their distribution on the X chromosome.

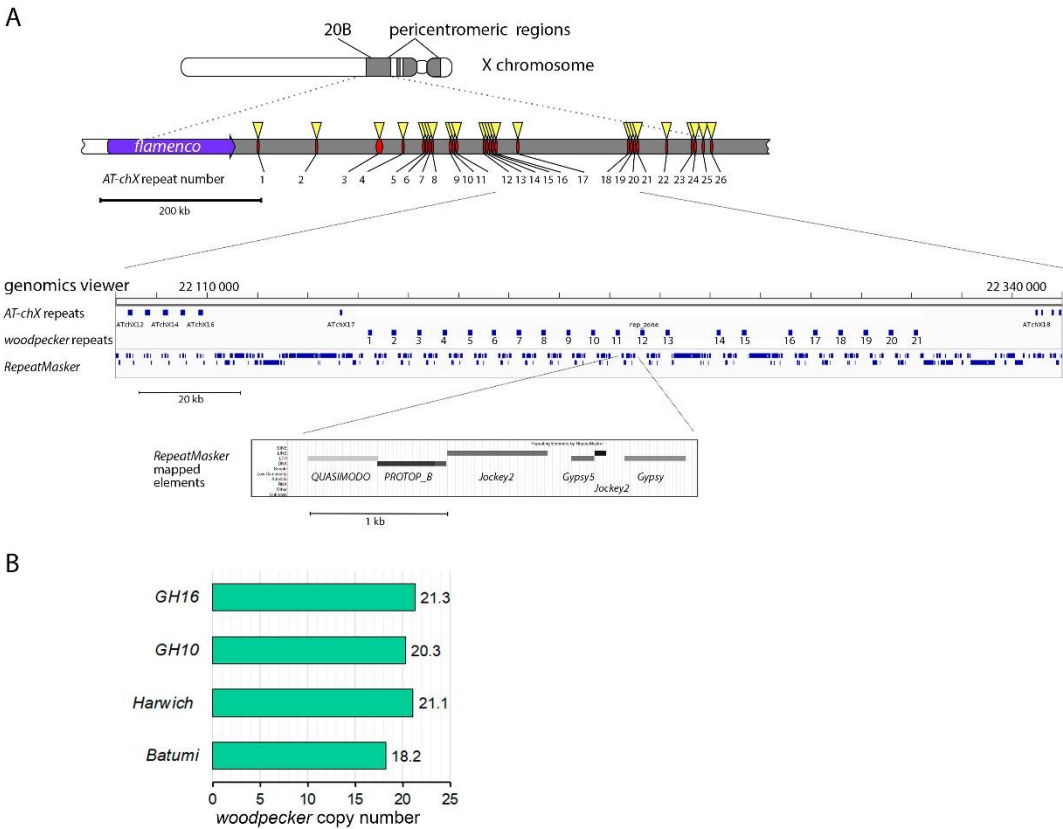


Figure 3. (A) Distribution of *Esf2/ABP1* paralog repeats (*woodpeckers*) within the *AT-chX* piRNA cluster on the pericentromeric region of the *D. melanogaster* X chromosome. On the top: the scheme of the germline-specific *AT-chX* piRNA cluster is shown according to [38]. In the middle: Integrative Genomics Viewer (IGV) was used for repeat visualization in the *D. melanogaster* genome assembly. At least 21 *woodpecker* repeats are located between *AT-chX* 17 and *AT-chX* 18 repeats within a 109.58-kb region in cytolocation 20B. The view encompassing the inner part of the *AT-chX* piRNA cluster is shown. Several *AT-chX* repeats (upper track) upstream and downstream of the *woodpecker* repeat-containing region (middle track) accompanied by *RepeatMasker* mapping (bottom track) are indicated in the browser view. Using the *RepeatMasker* tool, we identified a repeating pattern of fused transposon fragments that separate individual *woodpecker* copies from each other (on the bottom square, one magnified unit of fused transposon fragments is presented). (B) *woodpecker* copy number estimation in four wild-type *D. melanogaster* strains (see Materials and Methods for details).

The pattern of *Esf2/ABP1* homolog distribution in the *suzukii* subgroup species indicates that gene arising and loss is a rather independent process compared with species of the *melanogaster* subgroup. However, for both subgroups we observed that the parental copy and all duplicated copies are located exclusively on the X chromosome. It should be noted that the Y chromosome is not consistently represented in the available genome assemblies; for this reason, we were unable to systematically search of gene duplication on the Y chromosome, except for *D. melanogaster* and *D. suzukii* assemblies (see Table S1). However, for the *simulans* clade, we used gene annotation data of recently improved Y chromosome long-read assemblies [48] and also did not observe Y-linked *Esf2/ABP1* homologs. Beyond the *melanogaster* group, we did not detect *Esf2/ABP1* duplications in the *obscura*, *virilis*, and *repleta* groups considering only RefSeq annotated genome assemblies (data are shown in Table S1). Whereas in the *obscura* group we observed a partial synteny of the gene location; in the species of the *virilis* group of the distinct subgenus *Drosophila*, *Drosophila virilis*, the single *Esf2/ABP1* copy was found on the X chromosome at a non-syntenic location (Table S1). This suggests that the ancestral *Esf2/ABP1* gene changed its location after the divergence about 40-60 MYA [42,43] from a most recent common ancestor of this species and *D. melanogaster*. Taken together, we revealed several gene duplication events for this gene family in *Drosophila*, with the parental copy located

nearby *APC4* gene being the *Esf2/ABP1* gene that remains a single-copy gene per genome across the majority of the *melanogaster* group species.

3.2. Phylogenetic analysis for homologous genes of *melanogaster* subgroup

To evaluate the evolutionary relationship between homologs, we performed the collecting, multiple alignment, and phylogenetic analysis of coding sequences for all *Esf2/ABP1* family genes of *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* species. For generation of the phylogenetic tree, we used the Maximum Likelihood method and the Tamura-Nei model [24] with the orthologous *D. yakuba* gene LOC6525244 as the outgroup (Figure 4). We have built the tree with the highest log likelihood (-1 933.81) (Figure 4). Based both on tree characteristics and synteny analysis (Figure 2, Figure 4), we proposed that duplication events of *Esf2/ABP1* genes in cluster 1 and cluster 2 regions of the X chromosome have occurred before the splitting of *D. melanogaster* from common ancestor species with the *simulans* clade. Orthologous genes from cluster 2 (embedded in the intron of *InaE* gene in all four analyzed genomes) turn out to be in two monophyletic groups: one group contains all genes from cluster 2 of the *simulans* clade species, whereas another group includes the homologous CG10993 gene from cluster 2 and all 21 tandem *woodpecker* genes of cluster 3 of *D. melanogaster* (Figure 4). The close location of the branches of the last group on the tree indicates a potential origin of cluster 3 duplications from the single copy CG10993 after speciation of *D. melanogaster*.

We evidently revealed a low level of divergence between the first upstream orthologous parental copies from cluster 1 for all four analyzed species. This information is consistent with our analysis of nucleotide and amino acid identity (Figure 1 block 1, Figure 2, Table S1). The branch distribution for the following duplicated genes from cluster 1 inferred from the tree topology confirms high evolutionary rates for these copies; moreover, most divergence events appeared to occur before the splitting of the species. Thus, duplicated copies from cluster 1 have diverged considerably from their parental copies; however, within individual species, *D. melanogaster* and *D. sechellia*, all duplicated paralogs in cluster 1 are found to be extremely close to each other (Figure 4, Figure 1 block 4). This branch forms a monophyletic group, indicating that the generation of the first duplicated copy occurred in a common ancestor of these species, whereas second gene duplications in these regions in *D. melanogaster* and *D. sechellia* appear to emerge from the first duplicated copies after the splitting of the species.

While most *Drosophila* species harbor a single *Esf2/ABP1* gene copy, the *D. melanogaster* genome contains multiple paralogs according to our and previously published data [40]. To facilitate subsequent analysis, we introduced new designations for them (Table 1). We defined the parental paralog CG32708 as *cuckoo* because its daughter copies have been scattered at different locations of the X chromosome. Both duplicated copies in cluster 1, CG23706 and CG6999, were named as *chaffinch 1* and *chaffinch 2*, correspondingly. We defined the single paralog CG10993 in cluster 2 as *brambling*. Tandem amplified copies in cluster 3 were named as *woodpeckers* (*woodpecker 1*, *woodpecker 2*, ..., *woodpecker 21*), as mentioned above.

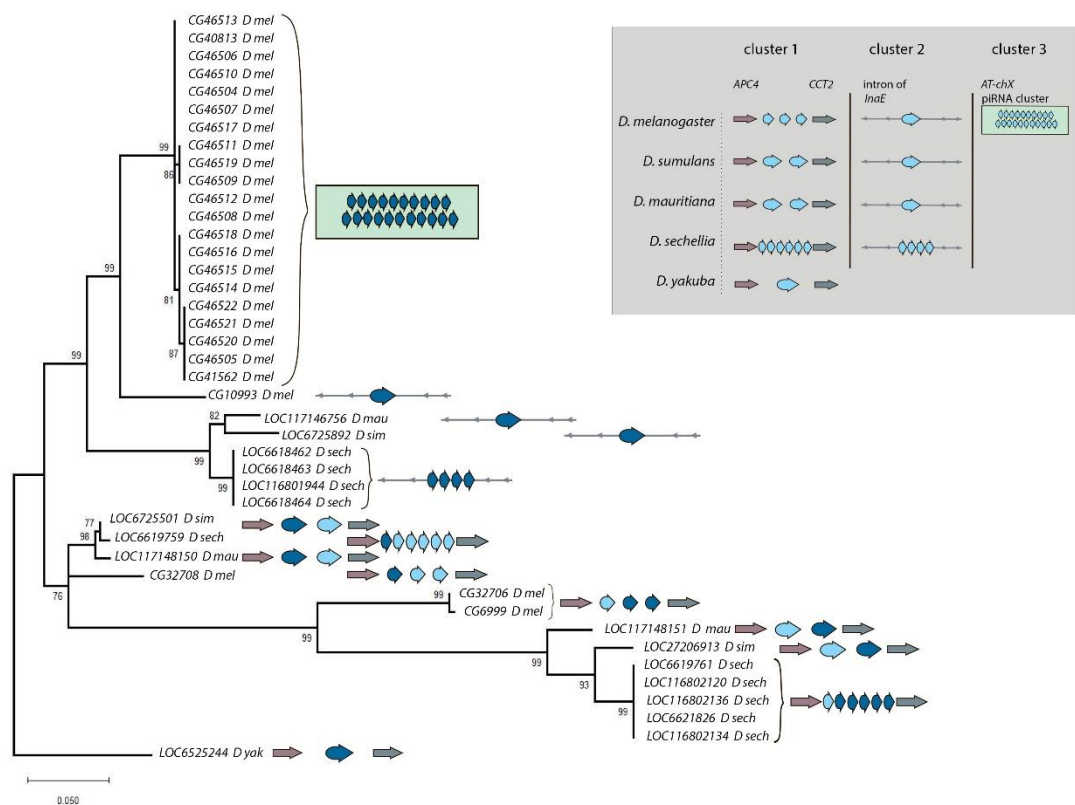


Figure 4. Maximum Likelihood phylogenetic tree for resolving the evolutionary relationship of *Esf2/ABP1* homologs in *D. melanogaster* and sibling species of the *simulans* clade with the highest log likelihood (-1933.81). The percentage of replicate trees in which the associated taxa clustered together (500 replicates) is shown near the branches. For the outgroup, *LOC6525244* gene of *Drosophila yakuba* was used. For all genes, their genomic locations are indicated by the icons according to Figure 2. Deeply colored blue figures on the icons mark genes corresponding to certain nodes. In the grey box on the top right, the common scheme of duplicated copy distribution in the analyzed genomes is represented according to Figure 2.

To detect signatures of sequence evolution in *Esf2/ABP1* homologs of *D. melanogaster*, we performed the codon-based Z-test that estimates the ratios of synonymous and nonsynonymous divergence between analyzed genes and close orthologs of *D. simulans* (Table S3). We found that purifying selection maintains the parental copy, *cuckoo*. In relation to other duplicated copies in all three genomic locations, we did not find significant deviations from neutral evolution.

Table 1. The list of new designations for *Esf2/ABP1* family genes in *D. melanogaster*.

Gene number	New gene designation
CG32708	cuckoo
CG32706	chaffinch 1
CG6999	chaffinch 2
CG10993	brambling
CG40813	woodpecker 1
CG46504	woodpecker 2
CG46505	woodpecker 3
CG46506	woodpecker 4
CG46507	woodpecker 5
CG46508	woodpecker 6

CG46509	woodpecker 7
CG46510	woodpecker 8
CG46511	woodpecker 9
CG46512	woodpecker 10
CG46513	woodpecker 11
CG46514	woodpecker 12
CG46515	woodpecker 13
CG46516	woodpecker 14
CG46517	woodpecker 15
CG46518	woodpecker 16
CG46519	woodpecker 17
CG46520	woodpecker 18
CG46521	woodpecker 19
CG46522	woodpecker 20
CG41562	woodpecker 21

3.3. Protein domain search in *Esf2*/ABP1 proteins

To detect potential innovations and functional sequence divergence in duplicated copies, we performed a search and comparison of known protein domains across *Esf2*/ABP1 homologs. Domain prediction for *Esf2*/ABP1 proteins of the melanogaster subgroup revealed that they shared with high e-value the RRM (RNA recognition motif) ABT1-like domain (domain found in activator of basal transcription 1) (Figure 5, Figure S2) that has been firstly described for the *Esf2*/ABP1 family [7]. The most striking divergence occurred within the species of the simulans clade with duplicated copies embedded in the intron of *InaE* gene (cluster 2). The insertion of the duplicated copy into the intron of the unrelated gene was supplemented by duplication of extended internal fragment, providing a second RRM ABT1-like domain in these proteins (Figure 5, Figure S2). We did not observe an intragenic duplication for the orthologous Brambling protein of *D. melanogaster*, indicating that the emergence and fixation of this whole-domain duplication occurred after the splitting of *D. melanogaster* from their common ancestor. All proteins analyzed also contain predicted Intrinsically Disordered Regions (IDRs) (with length from 39 aa to 109 aa); however, the overall disorder score for these proteins does not reach a high level (less than 50% of total length, from 19.5% to 41.7%) (Figure 5).

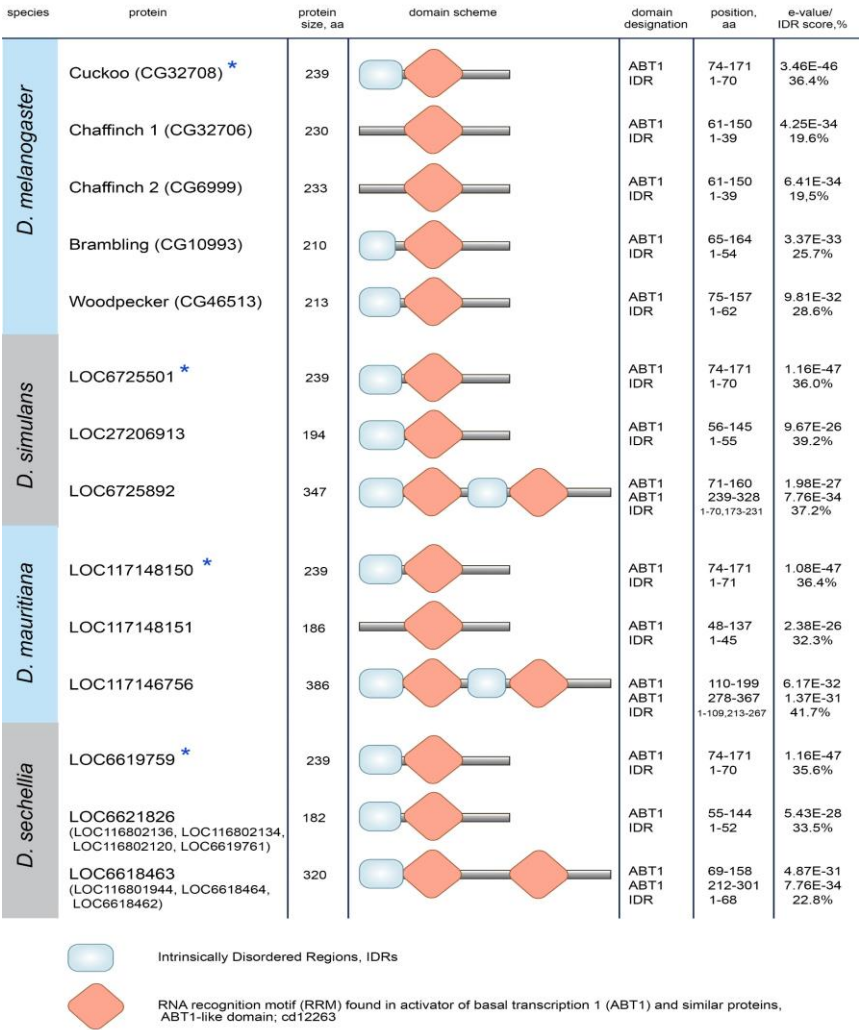


Figure 5. Protein domain determination for Esf2/ABP1 proteins of the melanogaster subgroup species. Blue asterisks mark proteins encoded by the parental gene copy in each species. All Esf2/ABP1 homologs, as predicted, retain the RRM ABT1-like domain with high e-value (last column). The overall disorder scores for the proteins are indicated as percentages in the last column. Only predicted Intrinsically Disordered Regions (IDRs) more than 50 aa in length are specified on the scheme. Note that we provided domain data for only Woodpecker 11 protein, CG46513, but the same properties were found for the other 20 Woodpecker proteins.

Protein domain determination for Esf2/ABP1 proteins of the *suzukii* subgroup revealed that all Esf2/ABP1 homologs retain the RRM ABT1-like domain near their C-termini (Figure 6, Figure S2). In addition, they contain extended continuous (in the range of 88 aa to 759 aa) IDRs at their N-termini (Figure 6). All of them have the overall disorder score equal to or exceeding 50% of the total length. Two proteins, LOC119556661 of *D. subpulchrella* and LOC108023671 of *D. biarmipes*, achieved strongly increased relative size (up to 80% of the whole length) of their continuous IDRs due to multiple intragenic tandem duplications in these regions (Figure 6, Figure S3). Thus, the divergence pattern of duplicated *Esf2/ABP1* genes in the *suzukii* subgroup is significantly different from that of the *melanogaster* subgroup duplicated orthologs. A gain of IDRs for duplicated copies is currently considered as a mechanism of the acquisition of new interaction partners and can reflect a putative new functional role [49,50].

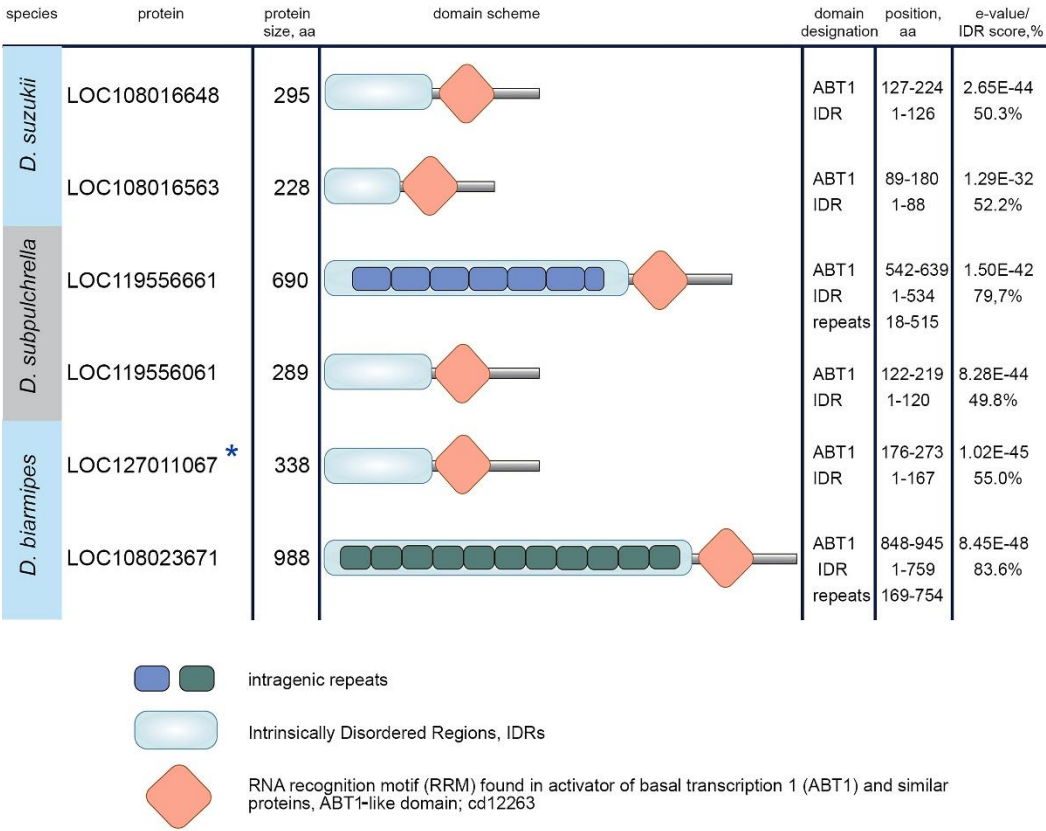


Figure 6. Protein domain determination for *Esf2*/*ABP1* homologous proteins of the *suzukii* subgroup. A blue asterisk marks the protein encoded by the parental gene copy. All *Esf2*/*ABP1* homologs retain the RRM ABT1-like domain. In addition, all proteins contain extended Intrinsically Disordered Regions (IDRs) at the amino terminus as indicated on the scheme. The overall disorder scores for the whole proteins are indicated as percentages in the last column. We did not specify predicted continuous IDRs less than 50 aa on the scheme. We identified multiple intragenic duplications for LOC119556661 of *D. subpulchrella* (6.5 repeats of 83 aa) and LOC108023671 of *D. biarmipes* (11 repeats of 54 aa) that significantly impact IDR expansion in these proteins.

3.4. Diversification of the expression pattern of *Esf2*/*ABP1* paralogs in *D. melanogaster* and *D. sechellia*

Using RT-qPCR and RNAseq data (Figure 7A,B), we revealed a sex-dimorphic pattern of *Esf2*/*ABP1* paralog expression in selected tissues of *D. melanogaster*. For differential expression analysis we used PolyA + RNAseq libraries of embryos, third instar larval gonads, and adult tissues/body parts for each sex. All *Esf2*/*ABP1* family genes are transcribed in the testes, with the highest transcript level found for the sum of *woodpecker* genes. In the ovaries we observed a noticeable level of transcripts only for *cuckoo*, whereas other paralogs were not expressed. Parental gene *cuckoo* was found to be expressed ubiquitously in all selected tissues (Figure 7B) with maximal transcript level in 2-4 h embryos, larval testes, and adult ovaries. Duplicated copies from cluster 1, *chaffinch 1*, and *chaffinch 2*, taken together, show maximal transcript level in larval and adult testes, heads, and thoraxes of adult males, demonstrating strong male-biased expression divergence from the parental gene. *Brambling* and *woodpeckers*, duplicated genes from clusters 2 and 3, correspondingly, exhibit a pronounced testis-specific expression pattern, likely associated with beneficial male germline activities (Figure 7B). Our RT-qPCR analysis of gonads, carcasses, and heads of adult flies of the wild-type *Batumi* strain of *D. melanogaster* generally confirmed the results obtained using RNAseq expression analysis (Figure 7A). Observed male germline enriched expression of young duplicated genes is interesting in the context of their acquisition and evolution. We perform the analogous RT-qPCR analysis of transcripts for *Esf2*/*ABP1* genes in selected tissues of *D. sechellia* adult males and females. In whole we found a similar expression pattern to *D. melanogaster*: a ubiquitous expression

of the older parental gene, but male-specific and testis-biased expression of more young duplicated copies (Figure S4).

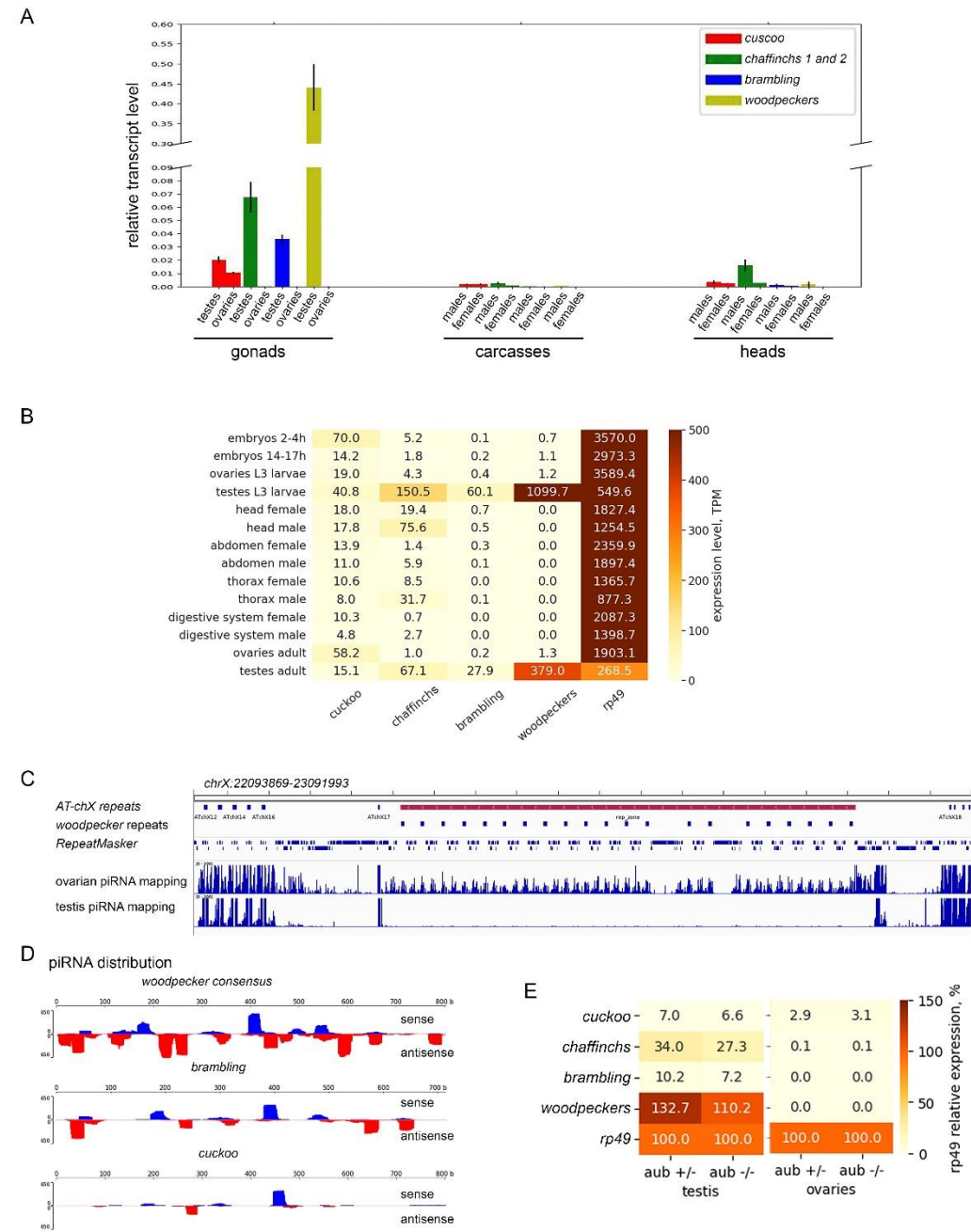


Figure 7. (A) RT-qPCR analysis of transcript levels of *Esf2/ABP1* family paralog in the gonads, carcasses, and heads of adult male and female flies of *w¹¹¹⁸* strain of *D. melanogaster*. The aggregated expression levels present for *chaffinch* and *woodpecker* genes. The expression levels of mRNAs were normalized to *rp49* transcripts. Error bars represent standard errors of the mean. (B) Expression analysis of *Esf2/ABP1* family genes from *w¹¹¹⁸* *D. melanogaster* in PolyA + RNA-seq libraries of embryos, third instar larval gonads, and adult tissues/body parts for each sex. The heatmap shows the expression of *cuckoo*, *brambling*, the collective gene expression of *chaffinch 1* and *chaffinch 2* genes, and *woodpecker* genes. *RpL32* (*rp49*) was used as the endogenous control (last column). Data are represented in TPM (transcripts per million). (C) Distribution of “locally unique” reads mapped to genomic region containing *woodpecker* repeats. The inner fragment of the *AT-chX* piRNA cluster encompassing *woodpecker* repeats (marked as *rep_zone* on the top (red bar)) is visualized using the Integrative Genomics Viewer (IGV). On the bottom: piRNAs map on the *woodpecker* repeat-containing region only in wild-type *D. melanogaster* ovaries but not in testes. The y-axis at the left represents read coverage of mapped piRNAs (number of reads per nucleotide position). (D) piRNA distribution across *woodpecker* consensus, *cuckoo*, and *brambling* coding

sequences was performed, permitting 0–2 mismatches. The *y*-axis shows read coverage (read number per nucleotide position). The sense read density is shown in blue; the antisense read density is shown in red (in relation to *woodpecker*, *cuckoo*, and *brambling* transcripts). (E) Comparative analysis of differential expression of *Esf2/ABP1* family genes in gonads of *D. melanogaster* of piRNA mutants (*aub* ^{-/-}) and their heterozygous siblings (*aub* ^{+/-}) using RNA-seq library data. The expression levels were normalized to *rp49* transcripts taken as 100%. Up-regulation of *Esf2/ABP1* family genes in the case of disruption of the piRNA pathway is not observed.

As we have shown above (Figure 3A), *woodpecker* repeats have amplified in the *D. melanogaster* genome in the pericentromeric heterochromatin region, which is located within the expanded germline-specific *AT-chX* piRNA cluster, one of the specialized genomic regions providing long transcript precursors for producing a bulk of small piRNAs in gonads [38,47]. It is shown that the *AT-chX* is one of the major piRNA clusters both in the testes and ovaries of *D. melanogaster* [33,38]. However, *woodpecker* repeats are annotated as genes, containing one intron and possessing the highest nucleotide and amino acid homology among themselves (more than 99%) (Figure 1 block 2) that likely resulted from gene conversion homogenizing these gene duplicates or from very recent tandem duplication. Taking into account their testis-specific expression (Figure 7B), we asked whether the amplification of *woodpecker* genes within active piRNA cluster leads to potential piRNA production in the gonads.

To gain insights into peculiarities of *woodpecker* repeat expression, we analyzed profiles of small RNAs in the libraries from gonads of wild-type lines of *D. melanogaster*. Here we analyzed “locally unique” reads that map only to the region containing *AT-chX* and *woodpecker* repeats (see Materials and Methods for details). We found that multiple piRNA reads selected by size (23–29 nt) from the ovarian small RNA library were mapped to genome coordinates corresponding to the region. Surprisingly, piRNA reads mapped to *woodpecker*-containing region were not practically observed in the testis small RNA library (Figure 7C). Ovarian sense and antisense *woodpecker*-mapped piRNAs are distributed along the whole length of the *woodpecker* repeat consensus sequence with permitting 0–2 mismatches (Figure 7D), indicating that this genomic region is actually a part of the *AT-chX* piRNA cluster in the ovaries. Thus, *woodpecker* repeats function as a part of the double-stranded piRNA cluster in a sex-dependent manner only in the ovaries of *D. melanogaster*.

The repression of transposon activity is considered a main function of the piRNA pathway in *Drosophila* [46]; however, a cohort of piRNAs has been shown to take part in the repression of protein-coding genes [33,38,51]. Does the presence of antisense *woodpecker*-derived piRNAs indicate the possibility of silencing of *Esf2/ABP1* family genes in the ovaries of *D. melanogaster* due to high complementarity to their transcripts? To evaluate the potential feasibility of piRNA-mediated silencing, we independently mapped ovarian piRNA reads to transcript sequences of *woodpecker*, *cuckoo*, *brambling*, and *chaffinch* genes with the indicated number of mismatches, from 0 to 2 (Table 2). Evidently, the most read numbers were found to be mapped to *woodpecker* repeats themselves (201.6 rpm for sense, 568.6 rpm for antisense). The calculated ping-pong signature value (*z*₁₀ score) suggested that the ping-pong amplification mechanism [52] is active in the biogenesis of these piRNAs in the ovaries. We also found that sense and antisense piRNA reads can be mapped to *brambling* and *cuckoo* transcripts (Table 2; Figure 7). Only a small number of piRNA reads was perfectly mapped (with 0-1 mismatch) to these transcripts, indicating piRNA origin from the distinct region. For transcripts of both *chaffinch* genes, we did not observe perfectly mapped piRNAs above background level (Table 2). Considering that target silencing requires a significant number of antisense piRNAs with a high level of complementarity to the target [33,38], we assumed that these piRNAs could presumably repress expression of *woodpecker* and *brambling* genes in the ovaries. However, piRNA silencing potential for repression of *cuckoo*, which is the only *Esf2/ABP1* paralog expressed in the ovaries, appears to be not enough, because the level of mapped antisense piRNAs is less than 100 rpm (Table 2). Our comparative analysis of polyA + RNAseq libraries of gonads of piRNA mutants (with disrupted piRNA biogenesis) and their heterozygous siblings revealed no significant change in the expression of all *Esf2/ABP1* paralogs in the gonads, suggesting that they are

not regulated by piRNAs (Figure 7E). Note that *woodpecker* repeats are transcribed as protein-coding genes only in the testes, whereas in the ovaries they are transcribed only as bidirectional piRNA precursors within the large piRNA cluster.

Table 2. Result of independent mapping of piRNA reads from *Oregon R* ovarian library to *woodpecker*, *cuckoo*, *brambling*, and *chaffinch* genes with the indicated mismatch number. The number of 10-nt overlapping of sense and antisense piRNA pairs and the z_{10} score are indicated in the two bottom lines, respectively. NA – data not available.

gene piRNA read type	woodpeckers		brambling		cuckoo		chaffinch1		chaffinch2	
	sense piRNAs	antisense piRNAs	sense piRNAs	antisense piRNAs	sense piRNAs	antisense piRNAs	sense piRNAs	antisense piRNAs	sense piRNAs	antisense piRNAs
Mapped reads (0-2 mm), rpm	201.6	568.6	115.8	203.8	68.8	47.2	0.6	0.2	0	0.2
Mapped reads (0 mm), rpm	150.3	422.7	1.1	55.2	15.9	0.9	0	0	0	0
Mapped reads (1 mm), rpm	40.5	129.5	25.2	51.4	10.1	3.2	0	0	0	0
Mapped reads (2 mm), rpm	10.7	16.4	89.5	97.1	45.7	43.2	0.6	0.2	0	0.2
Overlap_10, pairs	365		155		23		0		0	
z_{10} -score	2.86		3.08		1.01		NA		NA	

4. Discussion

Here we provide analysis of the evolutionary history of *Esf2/ABP1* gene duplications in *Drosophila* species of the subgenus *Sophophora*. Using high-quality genome assemblies and synteny analysis, we identified 56 homologous *Esf2/ABP1* genes in 16 species of the *melanogaster* group, separating approximately 25 MYA of evolution (Figure S1, Figure 1, Figure 2). Our analysis revealed that despite the single copy of the *Esf2/ABP1* gene per genome harbored in the majority of *Drosophila* species, gene duplication events occurred in four species of the *melanogaster* subgroup and independently in three species of the *suzukii* subgroup (Figure S1, Figure 2, Table S1). We found that recurrent DNA-mediated duplication events occurred at the same cytolocation nearby the parental copy as well as with translocation and amplification of duplicated gene copies in the introns of unrelated genes or in heterochromatic piRNA cluster (Figure 2).

Our phylogenetic and functional domain analyses (Figure 4, Figure 5) allow us to reconstruct the evolutionary history of *Esf2/ABP1* gene duplications in *D. melanogaster* and its sibling species of the *simulans* clade (Figure 8A). According to the proposed scenario, duplication events took place before speciation of *D. simulans* and *D. mauritiana*, whereas they occurred both before and after speciation in *D. melanogaster* and *D. sechellia*, providing the emergence and fixation of multiple new genes encoding proteins with RNA-binding ABT1-like domain (Figure 5). Duplicated paralogs in *D. melanogaster* do not exhibit any traces of pseudogenization and possess the characteristic features of functional copies. They contain an unbroken ORF, one conserved intron sequence, and an RNA-binding domain (ABT1-like) encoding region (Figure 5, Figure S2). We found that they acquired predominantly male-specific and testis-enriched expression pattern, unlike the parental gene, *cuckoo* (CG32708), which is expressed in different male and female tissues, with the highest expression in the ovaries and testes (Figure 7A,B). Among them, two duplicated copies in the same location with the parental gene, *chaffinch 1* (CG32706) and *chaffinch 2* (CG6999), have diverged considerably from *cuckoo*; however, they were found to maintain a high identity to each other (Figure 1 block 4, Figure 4).

The fate of most gene duplications is a rapid deletion or pseudogenization, but in some cases a duplicate copy can be fixed in the genome by drift or selection during a so-called fixation phase and subsequently maintained over time owing to acquiring a benefit function in the organism [6]. More than a dozen models were developed to explain the emergence, evolution, and maintenance of gene

duplications in genomes [2,5,6]. Note that the parental gene *cuckoo* is found to be essential for fly development and cell proliferation in *D. melanogaster* [9-12]. *cuckoo* has a high nucleotide and protein identity with orthologous genes of other species of the *melanogaster* group in the syntenic location (Figure 1 block 1, Figure 2, Table S1) and is maintained under purifying selection in the genome (Table S3). Molecular functions of *chaffinch 1* and *chaffinch 2* genes are not understood to date; however, we determined the distinct pattern of *chaffinch* gene expression, i.e., male-biased expression with a maximal level in the testes, and the absence or very low expression level in female tissues (Figures 7A,B). It allows excluding subfunctionalization or duplication–degeneration–complementation models from consideration, because in these cases expression pattern must be shared between the subfunctionalized original and duplicated copies. According to the neofunctionalization model, duplicated copies occasionally accumulate substitutions to acquire a novel gene function that will be supported by selection, whereas a parental copy will not. Due to the conservation of the whole exon-intron structure and functional domain of *chaffinch* genes (Figure 5), we do not consider the neofunctionalization model as permissible. The Z-test for positive selection for both *chaffinch* genes showed no deviation from neutrality (Table S3). We proposed that the most suitable model for the evolution of *chaffinch* genes is the “beneficial increase in dosage”. As the heterogametic sex in *Drosophila*, males possess a gene-rich X chromosome and a strongly heterochromatic Y chromosome with few protein-coding genes [53]. To correct the potentially lethal imbalance of X-linked gene expression level in males, the mechanism of dosage compensation in somatic cells of *D. melanogaster* males provides an approximately two-fold increasing expression of most X-linked genes with the aid of attraction of the MSL (Male-Specific Lethal) protein complex to the male X chromosome for acetylation of histone H4 at K16 position (acetylH4K16), a mark of active chromatin [54,55]. However, the somatic dosage compensation system appears to be not applicable to male germline cells, where MSL complex is not assembled and the chromatin of the X chromosome loses a high level of acetylH4K16 modification [56,57]. It is shown that testis germ cells in *Drosophila* are characterized by a massive wave of transcription on the spermatogonial stage but undergo meiotic X-chromosome inactivation later on the stage of mature spermatocytes [58]. Taking all these circumstances into account, we propose that owing to the extra transcript level in the period of active transcription in the testes, these X-linked gene duplications have been fixed and retained by a dosage selection, according to the beneficial increase of gene dosage model.

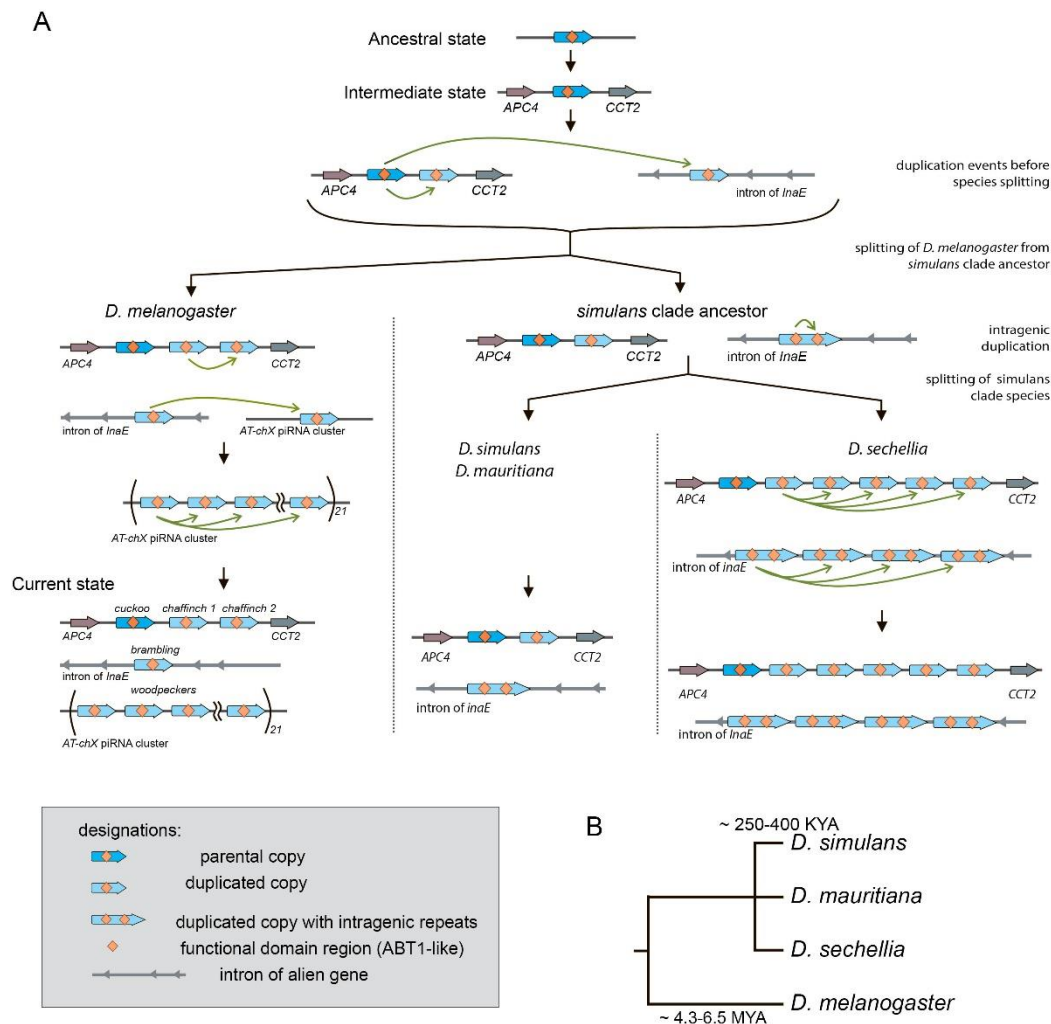


Figure 8. (A) Reconstruction of the basic events of the emergence of duplicated copies of *Esf2/ABP1* genes in the *melanogaster* subgroup genomes based on phylogenetic and functional domain analysis. A single *Esf2/ABP1* (parental) copy is located on the X chromosome flanked by *APC4* and *CCT2* genes in the genome of an intermediate ancestor fly. Duplication events indicated here by green arrows occurred both before and after speciation. Duplicated copies of the *simulans* clade species contain intragenic duplication, encoding ABT1-like domain. (B) The cladogram exhibits the relationship between *D. melanogaster* and the *simulans* clade species. *D. melanogaster* splits from the common ancestor with the *simulans* clade between 4.3 and 6.5 million years ago. *D. simulans*, *D. mauritiana*, and *D. sechellia* diverged from each other near 250 - 400 thousand years ago.

The same logic appears applicable to the remaining duplicated *Esf2/ABP1* genes in *D. melanogaster*, *brambling* and *woodpeckers*, since they all support a strong testis-restricted expression, preserve the exon-intron structure, and retain a functional ABT1-like domain (Figure 5, Figure 7A,B). *woodpecker* tandem gene amplification is unique to the *D. melanogaster* genome. According to our phylogenetic analysis, *brambling* and all *woodpecker* genes form the monophyletic group on the tree (Figure 4), confirming the origin of the last ones from the single-copy *brambling* gene after speciation of *D. melanogaster* (Figure 8A). We could assume that both *brambling* and *woodpecker* genes were fixed and preserved in the genome also owing to the beneficial increase of gene dosage. However, in the case of *woodpecker* genes, their genomic location leads to unexpected functional consequences that likely require a correction of the proposed model. The introduction of young duplicated copies into the piRNA cluster (Figure 2, Figure 3) resulted in their ovary-specific functional innovation. While in the testes these repeats are expressed as protein-coding genes, providing increased gene dosage; in the ovaries these repeats are subject to double-stranded non-canonical transcription as long

precursors for the piRNA biogenesis (Figure 7C,D). Determination of the forces ensuring the preservation of highly identical *woodpecker* genes in this unusual genomic location appears to be complex. At the same time, the mechanism of gene conversion can impact their maintenance. We found that piRNA silencing potential of *woodpecker*-derived antisense piRNAs for repression of *cuckoo*, which is the only *Esf2/ABP1* paralog with detected expression in the ovaries, is rather weak owing to a non-perfect piRNA complementarity to *cuckoo* transcripts (Table 2). A biological significance of the production of these piRNAs in the female germline is currently unclear and requires further research. In summary, here we uncovered a case of the functional adaptation of duplicated non-transposon genes with their sex-specific transformation into non-coding small RNAs. The acquisition of this potentially sex-antagonistic novel function took place owing to the insertion and amplification in a new genomic context without a significant divergence of *woodpecker* repeats from their parental *brambling* gene.

Note that a related history has been reconstructed earlier for the origin, evolution, and maintenance of the *Stellate-Suppressor of Stellate* (*Ste-Su(Ste)*) genetic system in the *D. melanogaster* genome [59,60]. Tandem amplification arrays of highly homologous genes, *Stellate* and *Su(Ste)*, arose from a chimeric Y-chromosomal intermediate precursor on the X and Y chromosomes, respectively. The getting of *Su(Ste)* repeats into heterochromatin context as well as the insertion of transposon *hoppe* in the *Su(Ste)* promoter appear to facilitate *Su(Ste)* repeat pseudogenization and the acquisition of bidirectional transcription across them, providing abundant *Su(Ste)* piRNAs for the silencing of harmful *Stellate* genes. *Su(Ste)* repeats have lost protein-coding potential but evolved to form the major testis-specific piRNA cluster in *D. melanogaster* genome [38,51], ensuring *Stellate* gene repression and maintenance of correct spermatogenesis. Thus, *woodpecker* repeats, which are proposedly at an earlier stage of their evolution than the *Ste-Su(Ste)* system, can be considered as a valuable model for studying the evolutionary forces driving the formation of piRNA clusters.

In relation to the pattern of duplications in species of the *simulans* clade, we can note that according to our hypothesis (Figure 8A), in *D. simulans* and *D. mauritiana* genomes the parental gene and both duplicated copies were inherited from an immediate ancestor of the clade. In *D. sechellia*, additional amplification events supposedly happened after speciation. Considering a short divergence time between the *simulans* clade species (Figure 8B), we can suggest that amplified copies of *D. sechellia* located in cluster 1 and cluster 2 are very young. They maintain extremely high identity between each other within their genome locations (Figure 1 blocks 3 and 4). Analyzing the expression pattern of *Esf2/ABP1* genes in the selected tissues of *D. sechellia* adult males and females, we revealed a picture strikingly similar to *D. melanogaster*, with male-specific and testis-biased expression of young duplicated copies and a ubiquitous expression of the parental gene (Figure S4). Taking into account intragenic duplication providing the acquisition of the second RNA-binding ABT1-like domain by the duplicated copies of cluster 2 (Figure 5), we can propose that testis-specific dosage selection also took place for their retention and maintenance.

We uncovered several duplication events for *Esf2/ABP1* genes in the *suzukii* subgroup species (Figure 2). Despite that in both the *melanogaster* and *suzukii* subgroup species we observed distribution of *Esf2/ABP1* family genes only on the X chromosome, the duplication events in these subgroups likely occurred through different mechanisms as independent processes (Figure 2, Table S1). A search of functional domains for *Esf2/ABP1* proteins of the *suzukii* subgroup allows us to reveal an unexpected domain diversification. These *Esf2/ABP1* proteins, in addition to ABT1-like domain preservation, contain long Intrinsically Disordered Regions, IDRs, composing 50 to 80% of the whole polypeptide length (Figure 6, Figure S2, Figure S3). They are significantly longer than those of *melanogaster* subgroup orthologs (Figure 5, Figure 6). IDRs are generally described as protein regions that do not adopt a defined three-dimensional structure under physiological conditions, but they are nevertheless functional. It is known that the majority of eukaryotic proteins contain both structured and disordered regions; however, only about 10% of proteins consist of IDRs representing half or more of their length, relating to the category of highly disordered proteins [50]. Among the *Esf2/ABP1* family proteins of the *suzukii* subgroup species, we determined two, LOC119556661 of *D. subpulchrella*

and LOC108023671 of *D. biarmipes*, containing multiple intragenic tandem repeat sequences that contribute to great extension of their continuous IDRs to near 80% of their whole length, consisting of 534 aa and 759 aa, respectively (Figure 6; Figure S3). Note that these tandem repeats are strongly enriched by acidic amino acid and serine residues, in sum consisting of more than 40% of their content (Figure S3). A common feature of long IDRs is their ability to form dynamic multivalent, flexible, and self-aggregating molecular networks, providing the assembly of liquid-liquid phase-separated droplets or hydrogels, commonly named biomolecular condensates [50,61]. The proteins with extended IDRs are mostly involved in the DNA, RNA, and protein binding to ensure molecular communication and regulation of cellular functions [50,62]. It is shown that expanded IDRs of certain RNA-binding proteins enable the formation in the cells of macromolecular granule-like assemblies, allowing localization, accumulation, and storage of RNAs [50,61]. Although the verification of molecular functions of expanded IDRs in fly tissues needs further investigations, we propose that *Esf2/ABP1* proteins in the *suzukii* subgroup species evolved with the acquisition of novel functional features mediated by their IDRs for RNA accumulation and processing. The tandem expansion of internal repeats is one of the ways by which IDR-encoding genes arose and spread during evolution [63]. According to our findings, we assume that intragenic duplications and fragment shuffling by recombination were presumable selection instruments driving the evolution of *Esf2/ABP1* family genes in a common ancestor of the *suzukii* subgroup flies. After species divergence, the expansion of IDRs owing to multiple internal repeats independently in *Esf2/ABP1* copies of *D. subpulchrella* and *D. biarmipes* appears to contribute to following evolutionary adaptation. However, the evolutionary scenario for this subgroup can hardly be accurately reconstructed owing to the complicated pattern of arising and loss of *Esf2/ABP1* genes (Figure 2).

Conclusions

To investigate the evolutionary history of *Esf2/ABP1* family duplicates in *Drosophila*, we employed a multifaceted approach, combining comparative genomics, phylogenetic analysis, functional domain analysis, and tissue-specific transcriptomic analysis. This strategy allowed us to identify and characterize 47 duplicated gene copies across seven *Drosophila* species clustered at three distinct genomic locations on the X chromosome. Recurrent gene duplication events occurred in four species of the *melanogaster* subgroup and independently in three species of the *suzukii* subgroup. In both cases duplication provided the emergence and fixation of multiple new genes encoding proteins with RNA-binding ABT1-like domain. In three species of the *suzukii* subgroup, *Esf2/ABP1* genes evolved with domain diversification: in addition to RNA-binding ABT1-like domain preservation, all homologous proteins acquired expanded Intrinsically Disordered Regions.

The first and the second duplication events in four species of the *melanogaster* subgroup probably occurred before the splitting of *D. melanogaster* from the species of the *simulans* clade, but the third one occurred after. Gene, named *cuckoo*, and its orthologs are parental, characterized by ubiquitous expression in different fly tissues, and found to be the essential gene for fly development and cell proliferation. The *chaffinch 1* and *chaffinch 2* genes, resulting from a second duplication event, have a distinct pattern of expression, i.e., male-biased expression with a maximal level in the testes. We proposed that the most suitable model for the evolution of *chaffinch* genes is the “beneficial increase in dosage.” Purifying selection maintains the parental copy, *cuckoo*. In relation to other duplicated copies in all three genomic locations, we did not find significant deviations from neutral evolution. The insertion and amplification of young duplicated *woodpecker* copies into the piRNA cluster lead to sex-specific functional innovation. Whereas in the testes these repeats are expressed as protein-coding genes; in the ovaries they undergo non-canonical transcription as long piRNA precursors.

This way, used complex approach permitted us to reconstruct duplicated gene evolutionary trajectories and revealing how their functions and expression patterns diverged over time.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Figure S1: Phylogenetic tree of *Esf2/ABP1* family orthologs and paralogs in the

genus *Drosophila*, Figure S2: Alignment of RRM ABT1 like domains predicted in *Esf2/ABP1* family proteins of the *melanogaster* subgroup and *suzukii* subgroup, Figure S3: Representation of multiple tandem intragenic repeats for proteins LOC119556661 of *D. subpulchrella* (6.5 repeats) (A) and LOC108023671 of *D. biarmipes* (11 repeats) (B), Figure S4: RT-qPCR analysis of transcript levels of *Esf2/ABP1* family genes in the gonads, carcasses, and heads of adult male and female flies of *D. sechellia*, Table S1: The table contains data for determination of orthologs and paralogs of *Esf2/ABP1* gene family in genome assemblies of *Drosophila* species, Table S2: List of RNA library resources used for the analyses, Table S3: A codon-based test of neutrality, the Z-test, for selection type for the *Esf2/ABP1* genes of *D. melanogaster* was performed using MEGA software.

Author Contributions: L.V.O.: conceptualization, methodology, writing—original draft preparation, visualization, project administration, and supervision. A.A.K.: methodology, data curation, formal analysis, writing—original draft preparation. E.D.D.: data acquisition, formal analysis, investigation, visualization, and writing—review and editing. E.Yu.Ya.: data acquisition, formal analysis, and writing—review and editing. A.V.Ch.: investigation, validation, formal analysis, and writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the IDB RAS Government basic research program in 2025 № 0088-2024-0017.

Data Availability Statement: The open-source RNAseq data resources list used in this study is included in the **Supplementary Materials, Table S2**.

Acknowledgments: We thank S. Yu. Sorokina and A. M. Kulikov for helpful advices, I. A. Kombarov and S. S. Bazylev for help with experiments. We thank the *Drosophila* collection of Institute of Cytology and Genetics, SB RAS, Russia, for fly strains. The research was done using equipment of the Core Centrum of Institute of Developmental Biology RAS and the Collection of laboratory and natural strains of *Drosophila* of Institute of Developmental Biology RAS.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Long, M.; Betrán, E.; Thornton, K.; Wang, W. The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **2003**, *4*, 865–875.
2. Conant, G.C.; Wolfe, K.H. Turning a hobby into a job: how duplicated genes find new functions. *Nat. Rev. Genet.* **2008**, *9*, 938–950.
3. Ohno, S. *Evolution by Gene Duplication*, Springer: New York, USA, 1970; pp. 1–160.
4. Force, A.; Lynch, M.; Pickett, F.B.; Amores, A.; Yan, Y.L.; Postlethwait, J. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* **1999**, *151*, 1531–1545.
5. Assis, R.; Bachtrog, D. Neofunctionalization of young duplicate genes in *Drosophila*. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17409–17414.
6. Innan, H.; Kondrashov, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat. Rev. Genet.* **2010**, *11*, 97–108.
7. Oda, T.; Kayukawa, K.; Hagiwara, H.; Yudate, H.T.; Masuho, Y.; Murakami, Y.; Tamura, T.A.; Muramatsu, M.A. A novel TATA-binding protein-binding protein, ABT1, activates basal transcription and has a yeast homolog that is essential for growth. *Mol. Cell. Biol.* **2000**, *20*, 1407–1418.
8. Hoang, T.; Peng, W.T.; Vanrobays, E.; Krogan, N.; Hiley, S.; Beyer, A.L.; Osheim, Y.N.; Greenblatt, J.; Hughes, T.R.; Lafontaine, D.L. *Esf2p*, a U3-associated factor required for small-subunit processome assembly and compaction. *Mol. Cell. Biol.* **2005**, *25*, 5523–5534.
9. Mummery-Widmer, J.L.; Yamazaki, M.; Stoeger, T.; Novatchkova, M.; Bhalerao, S.; Chen, D.; Dietzl, G.; Dickson, B.J.; Knoblich, J.A. Genome-wide analysis of Notch signaling in *Drosophila* by transgenic RNAi. *Nature* **2009**, *458*, 987–992.

10. Schnorrer, F.; Schönbauer, C.; Langer, C.C.; Dietzl, G.; Novatchkova, M.; Schernhuber, K.; Fellner, M.; Azaryan, A.; Radolf, M.; Stark, A.; Keleman, K.; Dickson, B.J. Systematic genetic analysis of muscle morphogenesis and function in *Drosophila*. *Nature* **2010**, *464*, 287–291.
11. Neumüller, R.A.; Richter, C.; Fischer, A.; Novatchkova, M.; Neumüller, K.G.; Knoblich, J.A. Genome-wide analysis of self-renewal in *Drosophila* neural stem cells by transgenic RNAi. *Cell Stem Cell* **2011**, *8*, 580–593.
12. Viswanatha, R.; Li, Z.; Hu, Y.; Perrimon, N. Pooled genome-wide CRISPR screening for basal and context-specific fitness gene essentiality in *Drosophila* cells. *Elife* **2018**, *7*, e36333.
13. Clifton, B.D.; Jimenez, J.; Kimura, A.; Chahine, Z.; Librado, P.; Sánchez-Gracia, A.; Abbassi, M.; Carranza, F.; Chan, C.; Marchetti, M.; Zhang, W.; Shi, M.; Vu, C.; Yeh, S.; Fanti, L.; Xia, X.Q.; Rozas, J.; Ranz, J.M. Understanding the Early Evolutionary Stages of a Tandem *Drosophila melanogaster*-Specific Gene Family: A Structural and Functional Population Study. *Mol. Biol. Evol.* **2020**, *37*, 2584–2600.
14. Chang, C.H.; Mejia Natividad, I.; Malik, H.S. Expansion and loss of sperm nuclear basic protein genes in *Drosophila* correspond with genetic conflicts between sex chromosomes. *Elife* **2023**, *12*, e85249.
15. Kuvaeva, E.E.; Cherezov, R.O.; Kulikova, D.A.; Mertsalov, I.B. The *Drosophila* toothrin Gene Related to the d4 Family Genes: An Evolutionary View on Origin and Function. *Int. J. Mol. Sci.* **2024**, *25*, 13394.
16. Brand, C.L.; Oliver, G.T.; Farkas, I.Z.; Buszczak, M.; Levine, M.T. Recurrent Duplication and Diversification of a Vital DNA Repair Gene Family Across *Drosophila*. *Mol. Biol. Evol.* **2024**, *41*, msae113.
17. Zakerzade, R.; Chang, C.H.; Chatla, K.; Krishnapura, A.; Appiah, S.P.; Zhang, J.; Unckless, R.L.; Blumenstiel, J.P.; Bachtrog, D.; Wei, K.H. Diversification and recurrent adaptation of the synaptonemal complex in *Drosophila*. *PLoS Genet.* **2025**, *21*, e1011549.
18. Clark, A.G.; et al.; Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **2007**, *450*, 203–218.
19. Miller, D.E.; Staber, C.; Zeitlinger, J.; Hawley, R.S. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3 (Bethesda)* **2018**, *8*, 3131–3141.
20. Kim, B.Y.; Wang, J.R.; Miller, D.E.; Barmina, O.; Delaney, E.; Thompson, A.; Comeault, A.A.; Peede, D.; D'Agostino, E.R.R.; Pelaez, J.; Aguilar, J.M.; Haji, D.; Matsunaga, T.; Armstrong, E.E.; Zych, M.; Ogawa, Y.; Stamenković-Radak, M.; Jelić, M.; Veselinović, M.S.; Tanasković, M.; Erić, P.; Gao, J.J.; Katoh, T.K.; Toda, M.J.; Watabe, H.; Watada, M.; Davis, J.S.; Moyle, L.C.; Manoli, G.; Bertolini, E.; Košťál, V.; Hawley, R.S.; Takahashi, A.; Jones, C.D.; Price, D.K.; Whiteman, N.; Kopp, A.; Matute, D.R.; Petrov, D.A. Highly contiguous assemblies of 101 drosophilid genomes. *Elife* **2021**, *10*, e66405.
21. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **2004**, *32*, 1792–1797.
22. Kumar, S.; Stecher, G.; Suleski, M.; Sanderford, M.; Sharma, S.; Tamura, K. MEGA12: Molecular Evolutionary Genetic Analysis version 12 for adaptive and green computing. *Mol. Biol. Evol.* **2024**, *41*, msae263.
23. Lemoine, F.; Correia, D.; Lefort, V.; Doppelt-Azeroual, O.; Mareuil, F.; Cohen-Boulakia, S.; Gascuel, O. NGPhylogeny.fr: new generation phylogenetic services for non-specialists. *Nucleic Acids Res.* **2019**, *47*, W260–W265.
24. Tamura, K.; Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **1993**, *10*, 512–526.
25. Tamura, K.; Stecher, G.; Kumar, S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Mol. Biol. Evol.* **2021**, *38*, 3022–3027.
26. Nei, M.; Gojobori, T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **1986**, *3*, 418–426.
27. Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Gonzales, N.R.; Gwadz, M.; Lu, S.; Marchler, G.H.; Song, J.S.; Thanki, N.; Yamashita, R.A.; Yang, M.; Zhang, D.; Zheng, C.; Lanczycki, C.J.; Marchler-Bauer, A. The conserved domain database in 2023. *Nucleic Acids Res.* **2023**, *51*, D384–D388.
28. Conte, A.D.; Mehdiabadi, M.; Bouhraoua, A.; Miguel Monzon, A.; Tosatto, S.C.E.; Piovesan, D. Critical assessment of protein intrinsic disorder prediction (CAID) - Results of round 2. *Proteins* **2023**, *91*, 1925–1934.

29. Bergman, C.M.; Haddrill, P.R. Strain-specific and pooled genome sequences for populations of *Drosophila melanogaster* from three continents. *F1000Res.* **2015**, *4*, 31.
30. Teixeira, F.K.; Okuniewska, M.; Malone, C.D.; Coux, R.X.; Rio, D.C.; Lehmann, R. piRNA-mediated regulation of transposon alternative splicing in the soma and germ line. *Nature* **2017**, *552*, 268–272.
31. Yang, H.; Jaime, M.; Polihronakis, M.; Kanegawa, K.; Markow, T.; Kaneshiro, K.; Oliver, B. Re-annotation of eight *Drosophila* genomes. *Life Sci. Alliance* **2018**, *1*, e201800156.
32. Maksimov, D.A.; Laktionov, P.P.; Posukh, O.V.; Belyakin, S.N.; Koryakov, D.E. Genome-wide analysis of SU(VAR)3-9 distribution in chromosomes of *Drosophila melanogaster*. *Chromosoma* **2018**, *127*, 85–102.
33. Chen, P.; Kotov, A.A.; Godneeva, B.K.; Bazylev, S.S.; Olenina, L.V.; Aravin, A.A. piRNA-mediated gene regulation and adaptation to sex-specific transposon expression in *D. melanogaster* male germline. *Genes Dev.* **2021**, *35*, 914–935.
34. Ramalingam, V.; Natarajan, M.; Johnston, J.; Zeitlinger, J. TATA and paused promoters active in differentiated tissues have distinct expression characteristics. *Mol. Syst. Biol.* **2021**, *17*, e9866.
35. Mahadevaraju, S.; Pal, S.; Bhaskar, P.; McDonald, B.D.; Benner, L.; Denti, L.; Cozzi, D.; Bonizzoni, P.; Przytycka, T.M.; Oliver, B. Diverse somatic Transformer and sex chromosome karyotype pathways regulate gene expression in *Drosophila* gonad development. *bioRxiv [Preprint]*. **2024** 2024.08.12.607556.
36. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527.
37. Srivastav, S.P.; Feschotte, C.; Clark, A.G. Rapid evolution of piRNA clusters in the *Drosophila melanogaster* ovary. *Genome Res.* **2024**, *34*, 711–724.
38. Kotov, A.A.; Adashev, V.E.; Godneeva, B.K.; Ninova, M.; Shatskikh, A.S.; Bazylev, S.S.; Aravin, A.A.; Olenina, L.V. piRNA silencing contributes to interspecies hybrid sterility and reproductive isolation in *Drosophila melanogaster*. *Nucleic Acids Res.* **2019**, *47*, 4255–4271.
39. Antoniewski, C. Computing siRNA and piRNA overlap signatures. *Methods Mol. Biol.* **2014**, *1173*, 135–146.
40. Fan, C.; Chen, Y.; Long, M. Recurrent tandem gene duplication gave rise to functionally divergent genes in *Drosophila*. *Mol. Biol. Evol.* **2008**, *25*, 1451–1458.
41. Lachaise, D.; Cariou, M.L.; David, J.R.; Lemeunier, F.; Tsacas, L.; Ashburner, M. Historical biogeography of the *Drosophila-Melanogaster* species subgroup. *Evol. Biol.* **1988**, *22*, 159–225.
42. Russo, C.A.; Takezaki, N.; Nei, M. Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **1995**, *12*, 391–404.
43. Tamura, K.; Subramanian, S.; Kumar, S. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol. Biol. Evol.* **2004**, *21*, 36–44.
44. Garrigan, D.; Kingan, S.B.; Geneva, A.J.; Andolfatto, P.; Clark, A.G.; Thornton, K.R.; Presgraves, D.C. Genome sequencing reveals complex speciation in the *Drosophila simulans* clade. *Genome Res.* **2012**, *22*, 1499–511.
45. Suvorov, A.; Kim, B.Y.; Wang, J.; Armstrong, E.E.; Peede, D.; D'Agostino, E.R.R.; Price, D.K.; Waddell, P.; Lang, M.; Courtier-Orgogozo, V.; David, J.R.; Petrov, D.; Matute, D.R.; Schrider, D.R.; Comeault, A.A. Widespread introgression across a phylogeny of 155 *Drosophila* genomes. *Curr. Biol.* **2022**, *32*, 111–123.e5.
46. Malone, C.D.; Brennecke, J.; Dus, M.; Stark, A.; McCombie, W.R.; Sachidanandam, R.; Hannon, G.J. Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **2009**, *137*, 522–535.
47. Konstantinidou, P.; Loubalova, Z.; Ahrend, F.; Friman, A.; Almeida, M.V.; Poulet, A.; Horvat, F.; Wang, Y.; Losert, W.; Lorenzi, H.; Svoboda, P.; Miska, E.A.; van Wolfswinkel, J.C.; Haase, A.D. A comparative roadmap of PIWI-interacting RNAs across seven species reveals insights into *de novo* piRNA-precursor formation in mammals. *Cell Rep.* **2024**, *43*, 114777.
48. Chang, C.H.; Gregory, L.E.; Gordon, K.E.; Meiklejohn, C.D.; Larracuente, A.M. Unique structure and positive selection promote the rapid divergence of *Drosophila* Y chromosomes. *Elife* **2022**, *11*, e75795.
49. Montanari, F.; Shields, D.C.; Khaldi, N. Differences in the Number of Intrinsically Disordered Regions between Yeast Duplicated Proteins, and Their Relationship with Functional Divergence. *PLoS ONE* **2011**, *6*, e24989.

50. van der Lee, R.; Buljan, M.; Lang, B.; Weatheritt, R.J.; Daughdrill, G.W.; Dunker, A.K.; Fuxreiter, M.; Gough, J.; Gsponer, J.; Jones, D.T.; Kim, P.M.; Kriwacki, R.W.; Oldfield, C.J.; Pappu, R.V.; Tompa, P.; Uversky, V.N.; Wright, P.E.; Babu, M.M. Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **2014**, *114*, 6589–6631.
51. Aravin, A.A.; Naumova, N.M.; Tulin, A.V.; Vagin, V.V.; Rozovsky, Y.M.; Gvozdev, V.A. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the *D. melanogaster* germline. *Curr. Biol.* **2001**, *11*, 1017–1027.
52. Brennecke, J.; Aravin, A.A.; Stark, A.; Dus, M.; Kellis, M.; Sachidanandam, R.; Hannon, G.J. Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. *Cell* **2007**, *128*, 1089–1103.
53. Kotov, A.A.; Bazylev, S.S.; Adashev, V.E.; Shatskikh, A.S.; Olenina, L.V. *Drosophila* as a Model System for Studying of the Evolution and Functional Specialization of the Y Chromosome. *Int. J. Mol. Sci.* **2022**, *23*, 4184.
54. Prestel, M.; Feller, C.; Straub, T.; Mitlöhner, H.; Becker, P.B. The activation potential of MOF is constrained for dosage compensation. *Mol. Cell* **2010**, *38*, 815–826.
55. Lucchesi, J.C.; Kuroda, M.I. Dosage compensation in *Drosophila*. *Cold Spring Harb, Perspect Biol.* **2015**, *7*, a019398.
56. Rastelli, L.; Kuroda, M.I. An analysis of *maleless* and histone H4 acetylation in *Drosophila melanogaster* spermatogenesis. *Mech. Dev.* **1998**, *71*, 107–117.
57. Meiklejohn, C.D.; Landeen, E.L.; Cook, J.M.; Kingan, S.B.; Presgraves, D.C. Sex chromosome-specific regulation in the *Drosophila* male germline but little evidence for chromosomal dosage compensation or meiotic inactivation. *PLoS Biol.* **2011**, *9*, e1001126.
58. Mahadevaraju, S.; Fear, J.M.; Akeju, M.; Galletta, B.J.; Pinheiro, M.M.L.S.; Avelino, C.C.; Cabral-de-Mello, D.C.; Conlon, K.; Dell'Orso, S.; Demere, Z.; Mansuria, K.; Mendonça, C.A.; Palacios-Gimenez, O.M.; Ross, E.; Savery, M.; Yu, K.; Smith, H.E.; Sartorelli, V.; Yang, H.; Rusan, N.M.; Vbranovski, M.D.; Matunis, E.; Oliver, B. Dynamic sex chromosome expression in *Drosophila* male germ cells. *Nat. Commun.* **2021**, *12*, 892.
59. Chang, C.H.; Larracuente, A.M. Heterochromatin-enriched assemblies reveal the sequence and organization of the *Drosophila melanogaster* Y chromosome. *Genetics* **2019**, *211*, 333–348.
60. Adashev, V.E.; Kotov, A.A.; Bazylev, S.S.; Shatskikh, A.S.; Aravin, A.A.; Olenina, L.V. *Stellate* Genes and the piRNA Pathway in Speciation and Reproductive Isolation of *Drosophila melanogaster*. *Front. Genet.* **2021**, *11*, 610665.
61. Banani, S.F.; Lee, H.O.; Hyman, A.A.; Rosen, M.K. Biomolecular condensates: Organizers of cellular biochemistry. *Nat. Rev. Mol. Cell Biol.* **2017**, *18*, 285–298.
62. Holehouse, A.S.; Kragelund, B.B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat. Rev. Mol. Cell Biol.* **2024**, *25*, 187–211.
63. Tompa, P. Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* **2003**, *25*, 847–855.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.