Transposons contribute to the functional diversification of the head, gut, and ovary transcriptomes across *Drosophila* natural strains

Marta Coronado-Zamora[1] and Josefa González[1*]

[1]Institute of Evolutionary Biology, CSIC, UPF.

Marta Coronado-Zamora, marta.coronado@ibe.upf-csic.es

*Corresponding author

josefa.gonzalez@csic.es

ABSTRACT

Transcriptomes are dynamic, with cells, tissues, and body parts expressing particular sets of transcripts. Transposons are a known source of transcriptome diversity, however studies often focus on a particular type of chimeric transcript, analyze single body parts or cell types, or are based on incomplete transposon annotations from a single reference genome. In this work, we have implemented a method based on *de novo* transcriptome assembly that minimizes the potential sources of errors while identifying a comprehensive set of gene-TE chimeras. We applied this method to head, gut and ovary dissected from five *Drosophila melanogaster* natural strains, with individual reference genomes available. We found that ~19% of body part-specific transcripts are gene-TE chimeras. Overall, chimeric transcripts contribute a mean of 43% to the total gene expression, and they provide DNA binding, catalytic activity, and DNA polymerase activity protein domains. Our comprehensive dataset is a rich resource for follow-up analysis. Moreover, because transposable elements are present in virtually all species sequenced to date, their relevant role in spatially restricted transcript expression is likely not exclusive to the species analyzed in this work.

INTRODUCTION

In contrast to the genome, an animal's transcriptome is dynamic, with cell types, tissues and body parts expressing particular sets of transcripts (Pan et al. 2008; Barash et al. 2010; Brown et al. 2014; Söllner et al. 2017). The complexity and diversity of the transcriptome arises from the combinatorial usage of alternative promoters, exons and introns, and polyadenylation sites. A single gene can, therefore, encode a rich repertoire of transcripts that can be involved in diverse biological functions, and contribute to adaptive evolution and disease (*e.g.*, Kiyose et al. 2022; Verta and Jacobs 2022; Marasca et al. 2020; Singh and Ahi 2022). The potential contribution of transposable element (TE) insertions to the diversification of the transcriptome was analyzed soon after the first whole-genome sequences were available (Franchini et al. 2004; Ganko et al. 2003; Jordan et al. 2003; Lipatov et al. 2005; van de Lagemaat et al. 2003). TEs are present in virtually all genomes studied to date, are able to insert copies of themselves in the genome and, although their mutation capacity is often harmful, they also represent an important source of adaptive genetic variation (Casacuberta and González 2013; Cowley and Oakey 2013; Schrader and Schmitz 2019; Volff 2006). While TEs are a known source of transcriptome diversity, the majority of studies so far rely on incomplete transposon annotations from a single reference genome (*e.g.*, Lipatov et al. 2005). Moreover, methodologies are often specifically designed for particular types of chimeric gene-TE transcripts, *e.g.*, TE-initiated transcripts (Babaian et al. 2019), particular types of TEs, *e.g.,* L1 chimeric transcripts (Pinson et al. 2018), or have been applied to individual cell types or body parts, (*e.g.*, Babarinde et al. 2021; Treiber and Waddell 2020). As such, our knowledge on the contribution of TEs to transcriptome diversification is still partial.

Two of the most studied mechanisms by which TEs can generate chimeric transcripts are by providing alternative promoters and protein domains. In human and mouse, 2.8% and 5.2% of the total transcript start sites occur within retrotransposons, respectively (Faulkner et al. 2009). In *D. melanogaster*, over 40% of all genes are expressed from two or more promoters, with at least 1,300 promoters contained in TEs (Batut et al. 2013). As well as individual

examples of TEs providing protein domains (Cordaux et al. 2006; Newman et al. 2008; Tipney et al. 2004), a comparative genomic analysis of tetrapod genomes revealed that capture of transposase domains is a recurrent mechanism for novel gene formation (Cosby et al. 2021). There is also evidence for the retrotransposon contribution to protein novelty. Approximately 9.7% of endogenous retrovirus open reading frames across 19 mammalian genomes evolve under purifying selection and are transcribed, suggesting that they could have been co-opted as genes (Ueda et al. 2020). Across insects, and depending on the methodology used, the percentage of newly emerged domains (<225 mya) due to TEs was estimated to be from 1.7% to 6.6% (Klasberg et al. 2018). However, studies that identify and characterize a comprehensive set of gene-TE chimeras to provide a complete overview of their contribution to both transcriptome and protein diversification are still missing.

Besides describing the diverse contributions of TEs to the transcriptome, analyzing the relative contribution of gene-TE chimeras to the total gene expression is highly relevant, as it is informative of the potential functional relevance of the transcripts identified. Studies performed so far suggest that this contribution is related to the position of the TE in the transcript. Transcripts with a TE inserted in the 5'UTR or internal coding exons show significantly lower mean levels of expression compared with non-chimeric TE-gene transcripts (Babarinde et al. 2021). TEs inserted in 3'UTRs were associated with reduced gene expression both in humans and mice, but with increased gene expression in human pluripotent stem cells (Babarinde et al. 2021; Faulkner et al. 2009). In addition, whether specific TE types contribute to tissue-specific expression has been explored in mammals, where retrotransposons were found to be overrepresented in human embryonic tissues (Conley et al. 2008; Faulkner et al. 2009). In *D. melanogaster*, the contribution of TEs to tissue-specific expression has only been assessed in the head, with 833 gene-TE chimeric genes described (Treiber and Waddell 2020). Thus, whether the contribution of chimeric gene-TE transcripts is more relevant in the *D. melanogaster* head compared with other body parts is still an open question.

Within genes, TEs could also affect expression by changing the epigenetic status of their surrounding regions. In Drosophila, repressive histone marks enriched at TEs spread beyond TE sequences, which is often associated with gene down-regulation (Lee and Karpen 2017). However, there is also evidence that TEs containing active chromatin marks can lead to nearby gene overexpression (Guio et al. 2018). Genome-wide, the joint assessment of the presence of repressive and active chromatin marks has been restricted so far to the analysis of four TE families (Rebollo et al. 2012) and has never been carried out in the context of chimeric gene-TE transcripts.

In this work, we performed a high-throughput analysis to detect, characterize, and quantify chimeric gene-TE transcripts in RNA-seq samples from head, gut, and ovary dissected from the same individuals belonging to five natural strains of *D. melanogaster* (Figure 1A; Rech et al. 2022). We implemented a method based on *de novo* transcriptome assembly that (i) minimizes the potential sources of errors when detecting chimeric gene-TE transcripts; and (ii) allows to identify a comprehensive dataset of transcripts rather than focusing on particular types (Figure 1B; Lanciano and Cristofari 2020). Additionally, we assessed the coding potential and the contribution of chimeric transcripts to protein domains and gene expression as proxies for their integrity and functional relevance. Finally, we took advantage of the availability of ChIP-seq data for an active and a repressive histone mark, H3K9me3 and H3K27ac, respectively obtained from the same biological samples to investigate whether the TEs that are incorporated into the transcript sequences also affect their epigenetic status.

RESULTS

9% of *D. melanogaster* transcripts, across body parts and strains, are gene-TE chimeras

We performed a high-throughput analysis to detect and quantify chimeric gene-TE transcripts in RNA-seq samples from head, gut, and ovary, in five *D. melanogaster* strains collected from natural populations (Figure 1A). The three body parts were dissected from the same individuals, and an average coverage of 32x (22x to 43x) per RNA-seq sample was obtained (3 replicates per body part and strain, Supplemental Table S1; Coronado-Zamora et al. 2023). We *de novo* assembled transcripts in which we annotated TE insertions using a new *D. melanogaster* manually curated TE library (Rech et al. 2022). We only considered *de novo* transcripts that overlap with a known transcript obtained from a reference-guided assembly (Figure 1B), and we removed falsely-fusioned transcripts applying the algorithm described in Lima et al. (2017). We also excluded chimeric transcripts with TE fragments whose sequence contains simple repeats in more than >80% of their length (see *Methods*; Benson 1999; Rech et al. 2022). We then used the reference genome of each strain to define the exon-intron boundaries of each transcript and to identify the position of the TE in the transcript (Figure 1B). The alignment with the reference genome and the accurate TE annotation also allowed us to discard single-unit transcripts, indicative of pervasive transcription when the sequence matched a multi-exonic transcript, and TE autonomous expression, which are two important sources of errors when quantifying the contribution of TEs to gene novelty (Figure 1B; Lanciano and Cristofari 2020).

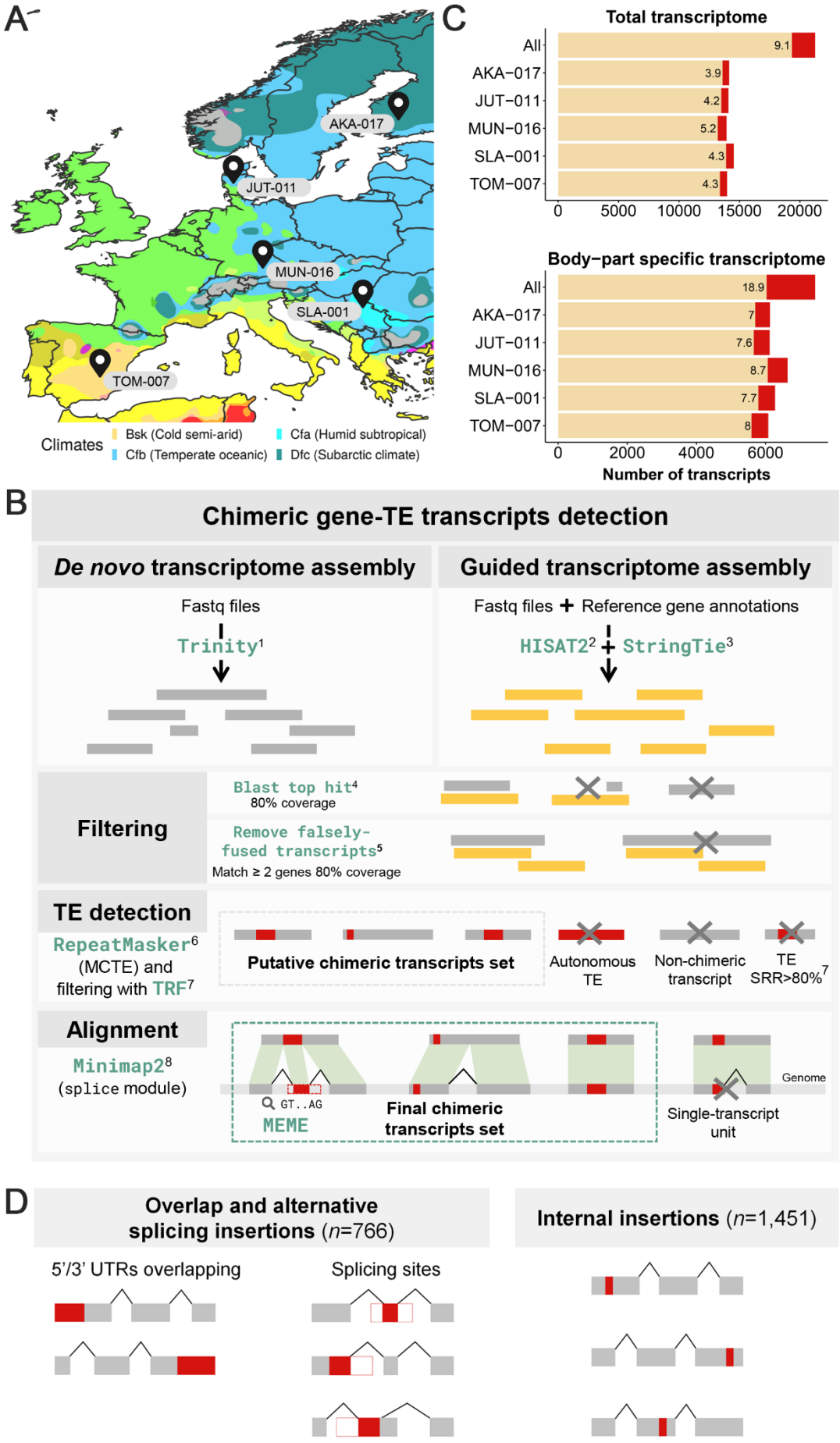
Overall, considering all the transcripts assembled in the three body parts and the five strains, we identified 1,931 chimeric gene-TE transcripts belonging to 826 genes (Supplemental Table S2A). Thus, approximately 9% (1,931/21,270) of *D. melanogaster* transcripts contain exonic sequences of TE origin. In individual strains, this percentage ranged between 3.9% to 5.2% (549-725 chimeric transcripts per genome) indicating that most of the chimeric gene-TE transcripts are strain-specific (1,312/1,931, 68%), as expected given that the majority of TEs

are present at low population frequencies (Figure 1C; Rech et al. 2022). Almost half of the strain-specific chimeric transcripts (48%, 634/1,312) were generated through strain-specific TE insertions, while the other 52% through polymorphic (141) or fixed (599) insertions, indicating that differential exon usage is an important mechanism in generating chimeric transcripts. While the overall contribution of TEs to the transcriptome is 9%, TEs contribute approximately 19% (1,406/7,435) of the total amount of body part-specific transcripts (Figure 1C).

We identified two groups of chimeric gene-TE transcripts (Figure 1D). The first group contains chimeric transcripts which have a TE overlapping with the 5'UTR, the 3'UTR, or introducing alternative splice (AS) sites (*overlap and AS insertions* group: 766 chimeric transcripts from 415 genes). We found that the majority of the TEs in the *overlap and AS insertions* group were adding canonical splice sites motifs (84.1%: 132/157) (Supplemental Table S2B). The second group contains chimeric gene-TE transcripts in which the TE is annotated completely inside the UTRs or internal exons (*internal insertions* group: 1,451 transcripts from 638 genes) (Figure 1D). We hypothesized that this group could be the result of older insertions that have been completely incorporated into the transcripts. Indeed, we found that TEs in this group are shorter than those of the *overlap and AS insertion* group, as expected if the former are older insertions (78% *vs*. 22.7% proportion of chimerics with TEs < 120 bp; χ^2 test, *p*-value < 0.001; Supplemental Fig. S1; see *Methods*). Additionally, while the majority of gene-TE transcripts in the *overlap and AS insertions* group were strain-specific, we found more transcripts shared between strains than strain-specific in the *internal insertions* group (χ^2 test, *p*-value < 0.001; Supplemental Fig. S2A and Supplemental Table S2C). This observation is also consistent with this group being enriched for older insertions, and remained valid when we removed the shorter insertions (χ^2 test, *p*-value < 0.001; Supplemental Table S2C).

To test whether the *overlap and AS insertions* and the *internal insertions* groups contribute differently to the diversification of the transcriptome, we performed all the subsequent analyses considering all the chimeric transcripts together, and the two groups separately. In addition,

because shorter insertions might be enriched for false positives, *i.e.,* not corresponding to real TE sequences due to the difficulty of annotating these repetitive regions, we also performed the analysis with the subset of chimeric gene-TE transcripts that contains a fragment of a TE insertion that is ≥120bp (638/766 and 672/1,451 for the *overlap and AS insertions* and the *internal insertions* groups, respectively; see *Methods*).


A
AKA-017
JUT-011
MUN-016
SLA-001
TOM-007
Climates
Bsk (Cold semi-arid)
Cfb (Temperate oceanic)
Cfa (Humid subtropical)
Dfc (Subarctic climate)
C
Total transcriptome
All
AKA−017
JUT−011
MUN−016
SLA−001
TOM−007
9.1
3.9
4.2
5.2
4.3
4.3
0
5000
10000
15000
20000
Body−part specific transcriptome
18.9
7
7.6
8.7
7.7
8
0
2000
4000
6000
Number of transcripts
B
Chimeric gene-TE transcripts detection
De novo transcriptome assembly
Guided transcriptome assembly
Fastq files
Trinity1
Fastq files + Reference gene annotations
HISAT2² + StringTie³
Filtering
Blast top hit⁴
80% coverage
Remove falsely-fused transcripts⁵
Match ≥ 2 genes 80% coverage
TE detection
RepeatMasker⁶
(MCTE) and filtering with TRF⁷
Putative chimeric transcripts set
Autonomous TE
Non-chimeric transcript
TE SRR>80%⁷
Alignment
Minimap2⁸
(splice module)
GT..AG
MEME
Final chimeric transcripts set
Genome
Single-transcript unit
D
Overlap and alternative splicing insertions (n=766)
Internal insertions (n=1,451)
5'/3' UTRs overlapping
Splicing sites

159

Figure 1. Detection of chimeric gene-TE transcripts in five strains of *D. melanogaster*. **A.** Map showing the sampling locations of the five European strains of *D. melanogaster* used in this study. TOM-007: Tomelloso, Spain (*BSk*); MUN-016: Munich, Germany (*Cfb*); JUT-011: Jutland, Denmark (*Cfb*); SLA-001: Slankamen, Serbia (*Cfa*); and AKA-017: Akaa, Finland (*Dfc*). Colors represent the climate zones according to the Köppen-Geiger climate distribution (Peel et al. 2007). **B.** Pipeline to detect chimeric transcripts. Two types of transcriptome assembly were performed: a *de novo* assembly using Trinity[1] (Grabherr et al. 2011) and a genome-guided transcriptome assembly using HISAT2[2] (Kim et al. 2019) and StringTie[3] (Pertea et al. 2015). Only *de novo* transcripts that had a minimum 80% coverage with a known transcript were considered to be screened for TEs insertions[4]. Falsely-fusioned transcripts derived from gene-dense compact genomes were removed[5]. RepeatMasker[6] (Smit et al. 2013) is used with a manually curated TE library (Rech et al. 2022) to detect TEs in the *de novo* assembled transcripts. Transcripts with TE fragments having simple repeats in over 80% of their length were excluded[7] (Benson 1999). An alignment against the reference genome of each strain is used to define the exon-intron boundaries of transcripts and to identify the position of the TE in the transcript (Li 2018)[8]. Transcripts fully annotated as a TE or detected as single-transcript units when the sequence matched a multi-exonic transcript are discarded. **C.** Contribution of chimeric gene-TE transcripts to the total transcriptome and the body part-specific transcriptome globally and by strain. *All* includes all the transcripts assembled in the three body parts and the five strains. **D.** Schematic of the two groups of chimeric transcripts identified. *Overlap and alternative splicing (AS) insertions* group, and *internal insertions* group. Note that these numbers total more than 1,931 because some chimeric transcripts can have different insertions in different samples. Gray boxes represent exons, red boxes represent a TE fragment incorporated in the mRNA, white boxes represent a TE fragment that is not incorporated in the final mRNA. The black lines connecting the exons represent the splicing events.

Gene-TE chimeric transcripts are more abundant in the head and ovary

Using high-throughput methodologies, 833 chimeric genes were identified in the *D. melanogaster* head (Treiber and Waddell 2020), however, the relative amount of chimeric gene-TE transcripts across body parts has never been assessed before. We found that the majority of the assembled chimeric gene-TE transcripts across the five strains analyzed were body part-specific (72.8%: 1,406/1,931), with only 9.3% (180) shared across all three body parts (Figure 2A and Supplemental Table S3A). The same pattern was found for the *overlap and AS insertions* group and for the *internal insertions* group, when considering all insertions and those ≥120bp (Supplemental Fig. S2B and Supplemental Table S3A).

Head was the body part expressing the most chimeric transcripts (1,001) followed by gut (863) and ovary (772) (Figure 2A and Supplemental Table S3A). Note that 145 of the chimeric genes identified in this work were previously described by Treiber and Waddell (2020), an expected low number given that many of the detected gene-TE chimeras are strain-specific and that the applied methodologies are different. After accounting for differences in the total number of transcripts assembled in each body part, we observed that the head was expressing the same proportion of gene-TE chimeras as the ovary (6.2% head *vs*. 6.2% ovary; χ^2 test, *p*-value = 1); while head and ovary express more than the gut (5.6% gut; χ^2 test, *p*-values = 0.032 and 0.048, respectively; Supplemental Table S3C). A similar proportion of chimeric transcripts in head and ovary was also found when the *overlap and AS insertions* and the *internal insertions* groups were analyzed separately, also if we focus on ≥120bp insertions (Figure 2B and Supplemental Table S3C). Overall, the same pattern was also found at the strain level, except for some comparisons (Supplemental Table S3C).

Finally, the head was the body part that expressed the most body part-specific chimeric transcripts (62% head *vs.* 50% gut; χ^2 test, *p*-value = 0.005, and *vs.* 46.5% ovary, *p*-value < 0.001), while no differences were found between gut and ovary (46.5% ovary *vs.* 50% gut; χ^2 test, *p*-value = 0.8; Figure 2A). In the three body parts, these proportions were higher than the

total proportion of body part-specific transcripts (23.2%, 15.14% and 11.25%, for head, gut and ovary respectively; χ^2 test, *p*-values < 0.001 for all comparisons; Supplemental Table S3B).

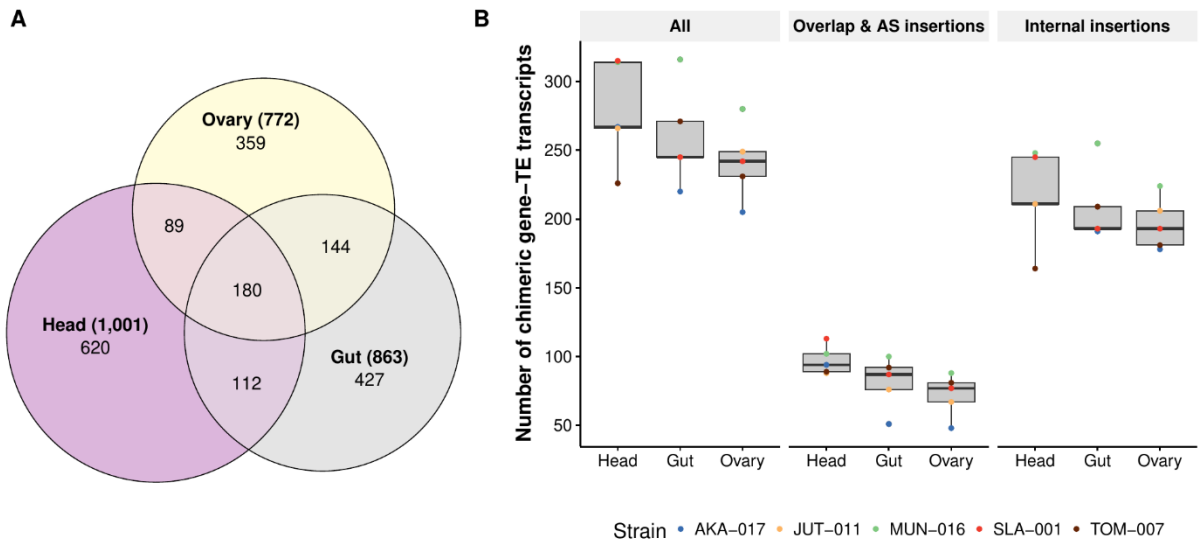


Figure 2. Distribution of chimeric transcripts across body parts and insertion groups. **A.** Venn diagram showing the overall number of chimeric transcripts shared across body parts in the five strains. **B.** Number of chimeric gene-TE transcripts detected by body part, strain and insertion group. *All* includes all chimeric transcripts detected in all body parts and strains.

Most chimeric transcripts contain TE insertions in the 3'UTRs

Chimeric gene-TE transcripts are enriched for TE insertions located in the 3'UTRs in *D. melanogaster* and in mammals (Babarinde et al. 2021; Lipatov et al. 2005; van de Lagemaat et al. 2003). Consistently, we also found that most of the chimeric gene-TE transcripts contain a TE in the 3'UTR (963 transcripts from 417 genes) followed by internal exons (569 transcripts from 266 genes) and insertions in the 5'UTRs (474 transcripts from 273 genes). We also detected 125 chimeric transcripts belonging to 95 genes consisting of a single-unit transcript. Note that 21 of the 5'UTR insertions detected in this work were experimentally validated in a previous analysis that estimated the promoter TE usage across developmental stages in *D. melanogaster* (Batut et al. 2013). Indeed, the number of chimeric genes with a TE inserted in

the 3' and 5'UTRs is higher than expected when taking into account the proportion of the genome that is annotated as UTRs, while there is a depletion of TEs in internal exons (χ^2 test, *p*-value < 0.001 in the three comparisons; Supplemental Table S4A). It has been hypothesized that the higher number of insertions in 3'UTRs could be explained by lack of selection against insertions in this gene compartment (Jordan et al. 2003; Lipatov et al. 2005). We thus tested whether 3'UTR chimeric transcripts were enriched for TE insertions present in more than one genome, but we found that this was not the case (χ^2 test, *p*-value = 0.9731; Figure 3A and Supplemental Table S4B).

In both the *overlap and AS insertions* and *internal insertions* groups, TE insertions were also mainly located in the 3'UTRs (53.9%: 316/586; 42%: 503/1199, respectively). In the *internal insertions* group there were more insertions in internal exons than in the *overlap and AS* insertions (414 *vs.* 42; χ^2 test, *p*-value < 0.001). This pattern also holds when we only consider ≥120bp insertions (25 *vs.* 143; χ^2 test, *p*-value < 0.001; Supplemental Table S4C). Figure 3B shows the number of chimeric gene-TE transcripts globally and by insertion group, body part and strain (Supplemental Table S4D) where it can be observed that, overall, the previous patterns hold at the body part level.

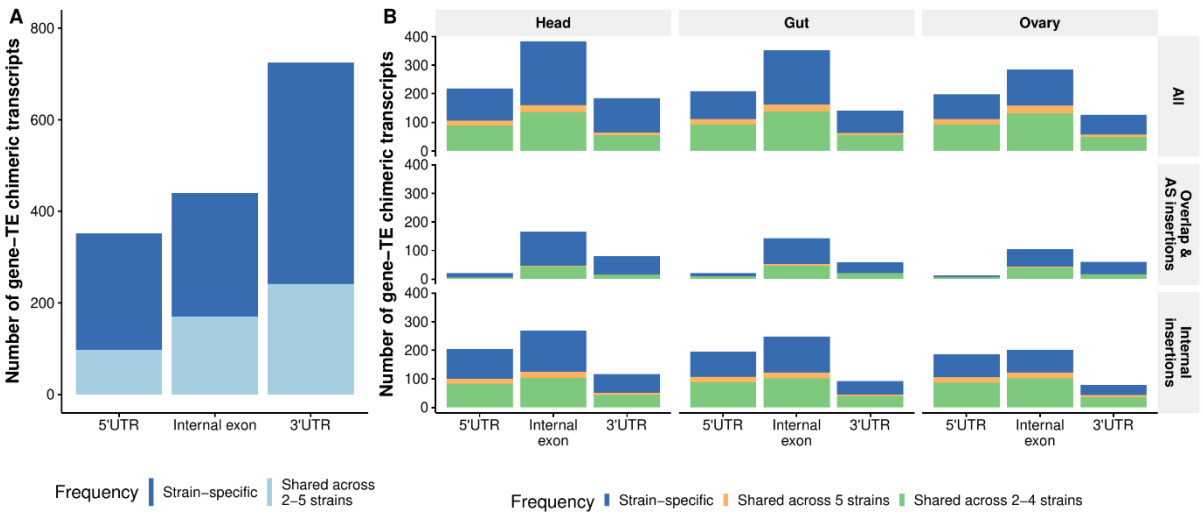


Figure 3. Position and frequency distribution of TEs in chimeric transcripts. A. Number of gene-TE chimeric transcripts by position and frequency**. B.** Number of chimeric gene-TE

transcripts by insertion group and body part, according to the insertion position (5'/3'UTRs or internal exons) and frequency. Each dot represents the number of chimeric gene-TE transcripts according to the frequency: strain-specific (blue), shared across two to four strains (orange) and shared across all five strains (green). These analyses were performed with the subset of chimeric transcripts with only one TE annotated (1,634).

Retrotransposons and DNA transposons contribute to chimeric gene-TE transcripts

We assessed the contribution of TE families to chimeric gene-TE transcripts. We found that many TE families, 90/146 (62%), were detected in chimeric gene-TE transcripts, as has been previously described in head chimeric transcripts (Supplemental Table S5A; Rech et al. 2022; Treiber and Waddell 2020). Although retrotransposons are more abundant than DNA transposons (60% on average in the five genomes analyzed (Rech et al. 2022), the contribution of retrotransposons to the chimeric gene-TE transcripts was not higher than expected (78%: 70/90; χ^2 test, *p*-value = 0.14; Supplemental Table S5B). The number of families contributing to the *overlap and AS insertions* group and the *internal insertions* group is similar (75 *vs.* 68, respectively, χ^2 test, *p*-value = 0.745; Supplemental Table S5C), and both groups were enriched for retrotransposons (χ^2 test, *p*-values < 0.05; Supplemental Table S5C). Half of these families (46: 51%) contribute to chimeric transcripts in all body parts, while 20 families were body part-specific, with 10 being head-specific, 6 gut-specific and 4 ovary-specific (Supplemental Table S5A).

The most common TE families found were *roo* (38.1%) and INE-1 (26%) (Figure 4). Indeed, these two families were over-represented in the chimeric transcripts dataset when compared to their abundance in the genome: *roo* in the five strains (χ^2 test, *p*-value < 0.0001 for all comparisons) and INE-1 in SLA-001 (χ^2 test, *p*-value = 0.0098, respectively) (Supplemental Table S5D). The *roo* and INE-1 were also the most common families both in the *overlap and AS insertions* group (17.3 and 29%, respectively) and in the *internal insertions* group (48.2%

and 28.4%, respectively; Figure 4). The same pattern was found when we analyzed only those chimeric transcripts with TEs ≥120bp (Supplemental Fig. S3 and Supplemental Table S5E).

Because *roo* insertions were enriched in all the strains analyzed, we further investigate these TE sequences. We found mainly two types of *roo* insertions: solo LTRs (18 insertions, 16 chimeric transcripts), that most of them belonged to the *overlap and AS insertions* group (16/18), and a short (45bp-192bp) low complexity sequence mapping to the positions 1,052-1,166 of the canonical *roo* element (see *Methods*). This short *roo* sequence is more common in the *internal insertions* group than in the *overlap and AS insertions* group (2,152 *vs.* 262 insertions, respectively). Note that a recent analysis by Oliveira et al. (2022) also found this same region of the *roo* consensus sequence to be the most abundant in chimeric gene-TE transcripts across four *D. melanogaster* strains. The authors discarded that these short sequences were widespread repeats across the genome as the majority of them (97.45%) were found to have a single blast hit in the genome. We argued that if these low-complexity regions have a *roo* origin, we should find that at least some of them should also have a blast hit with a *roo* insertion. To test this, we used less strict blast parameters compared with Oliveira et al.[43] and found that 51 of the low complexity regions have a *roo* element insertion as the second best hit and 168 have a *roo* insertion in the top 5 hits, suggesting that indeed some of these sequences have a clear *roo* origin (Supplemental Table S5F). Furthermore, we also tested whether this low complexity region was present in the *roo* consensus sequence from a closely related species, *D. simulans*, and found that this was the case strongly suggesting that this low complexity sequence is an integral part of the *roo* element.

We further investigated why this *roo* low complexity region was incorporated into genes. Because TEs can contain *cis*-regulatory DNA motifs, we performed a motif scan of the low complexity sequence from the canonical *roo* element. We found a C2H2 zinc finger factor motif repeated six times in this region (MA0454.1). Note that this specific motif is only found once in the *roo* consensus sequence outside the low complexity region. A scan in the *roo* sequences from the chimeras revealed that 90% (778/866) of the transcripts with the low

complexity *roo* sequence contain at least one sequence of this zinc finger motif, with 25% (196/778) containing 3 or more (Supplemental Table S5G). Additionally, we also look for motifs on *roo* solo LTRs (16 chimeric transcripts) and found that 10 out of 16 transcripts presented at least one motif, all related to DNA-binding domains found in transcription factors, including other C2H2 zinc fingers factor motifs (Supplemental Fig. S4A).

Finally, we also performed a motif scan on INE-1 elements (600 transcripts), the second most common family, and on the 20 body part-specific TE families (29 transcripts). 471 out of 600 of the analyzed transcripts contained motifs provided by the INE-1 fragment, all of them related to DNA-binding motifs, including the C2H2 zinc fingers factors (Supplemental Fig. S4B). 24 out of 29 transcripts with body-specific TE families contained at least one motif related to 16 different classes, including again C2H2 zinc finger factors (Supplemental Fig. S4). Of the 16 motifs, 62.5% (10/16) were body-part specific, with 6 detected in head-chimeric transcripts, 2 in gut and 2 in ovary (Supplemental Fig. S4C).

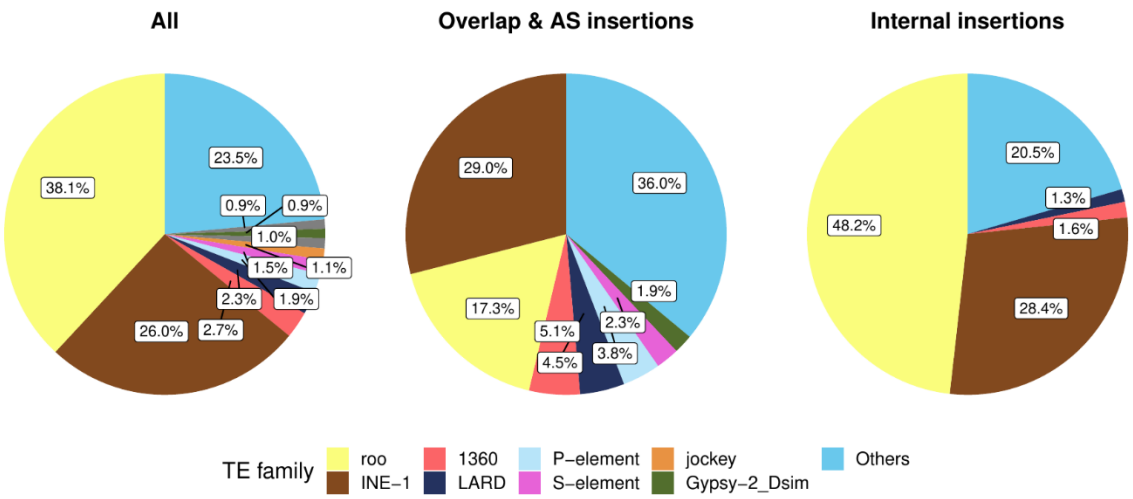


Figure 4. TE families distribution in gene-TE chimeras, globally and by insertion group. Percentage of TE families contributing to gene-TE chimeras in the global dataset (*All*), in the *overlap and AS insertions* group and in the *internal insertions* group. Only TE families found in more than 9 chimeric genes are depicted, otherwise they are grouped in *Others*.

Chimeric gene-TE transcripts contribute a median of ~43% of the total gene expression

Besides identifying and characterizing chimeric gene-TE transcripts, we quantified the level of expression of both chimeric and non-chimeric transcripts genome-wide. We focused on transcripts with ≥1 TPM in at least one of the samples analyzed (1,909 out of 1,931 chimeric transcripts, corresponding to 819/826 of the genes; see *Methods*, Supplemental Table S6A). We found that chimeric gene-TE transcripts have lower expression levels than non-chimeric transcripts (19,228; Wilcoxon's test, *p*-value < 0.001, Figure 5A). This is in contrast with previous observations in human pluripotent stem cells that reported no differences in expression between chimeric and non-chimeric transcripts (Babarinde et al. 2021). We dismissed the possibility that the lower expression of chimeric gene-TE transcripts was driven by the *roo* low complexity region identified in 829 of the chimeric transcripts (Wilcoxon's test, *p*-value < 0.0001; Figure 5A). Lower expression of the chimeric gene-TE transcripts compared to non-chimeric transcripts was also found when we analyzed the *overlap and AS insertions* and *internal insertions* groups separately and at the body part and strain levels (Wilcoxon's test, *p*-value < 0.001 for all comparisons; Figure 5A and Supplemental Table S6B).

We further tested whether TEs inserted in different gene locations differed in their levels of expression compared with non-chimeric TE transcripts. We found that chimeric transcripts had significantly lower expression than non-chimeric transcripts regardless of the insertion position (Wilcoxon's test *p*-value < 0.001 for all comparisons; Figure 5A). Furthermore, insertions in the 3'UTR appeared to be more tolerated than those in 5'UTR and internal exons, as their expression level was higher (Wilcoxon's test, *p*-value < 0.001 for both comparisons; Figure 5A). Our results are consistent with those reported by Faulkner et al.[22] who also found that 3'UTR insertions reduced gene expression compared to non-chimeric transcripts and contrast the findings of Babarinde et al. (2021) who found that 3'UTR chimeric transcripts have significantly higher levels of expression compared to non-chimeric transcripts or with insertions in 5'UTR and internal exons.

If we focus on the chimeric genes, 38% of them (314 genes) only expressed the chimeric gene-TE transcript (in all the genomes and body parts where expression of the gene was detected). Most of these genes (76%) contain short TE insertions and accordingly, most of them belong to the *internal insertions* group (66%). For the other 62% (505) of the genes, we calculated the average contribution of the chimeric gene-TE transcript to the total gene expression across samples. While some transcripts contributed on average only ~0.1% of the total gene expression, others accounted for >90% (mean = 27%) (Figure 5B). The mean contribution to gene expression of the *internal insertions* group was higher than that of the *overlap and AS insertions* group, when considering all the insertions (28.5% *vs.* 15.7%, respectively; Wilcoxon's test, *p*-value < 0.001), and when analyzing only those transcripts with ≥120bp insertions (28.5 *vs.* 15.7%, respectively; Wilcoxon's test, *p*-value > 0.0001). Considering only the transcripts that do not contain the *roo* low complexity sequence, the mean contribution to gene expression of the *internal insertions* group was still 25.3%. Overall, taking all chimeric genes into account (819), the average chimeric gene-TE transcripts' expression contribution to the total gene expression was 43%.

Finally, we evaluated whether there are differences between the expression levels of body part-specific and body part-shared chimeric transcripts. The breadth of expression, measured as the number of tissues in which a gene is expressed, is significantly and positively correlated with the level of expression in *Drosophila* (Larracuente et al. 2008) and humans (Park and Choi 2010). Consistent with this, we found that body part-shared chimeric transcripts have significantly higher expression levels than chimeric transcripts expressed in only one body part (Wilcoxon's test, *p*-value < 0.001; Supplemental Table S6C), when considering the whole dataset and for chimeric transcripts with insertions ≥120bp (Wilcoxon's test, *p*-value < 0.001; Supplemental Table S6C). Since we observed that the head was expressing more chimeric transcripts (Figure 2A), we next assessed if head-specific chimeric transcripts were also expressed at higher levels. However, we observed that the median expression of head-specific chimeric transcripts was similar to those specific to gut ($median_{head}$= 9.7 TPM [*n* =

620], median$_{gut}$= 10 TPM [*n* = 427]; Wilcoxon's test, *p*-value = 0.054), but lower than ovary-specific chimeric transcripts (median$_{ovary}$= 12.6 TPM [*n* = 359]; Wilcoxon's test, *p*-value > 0.0001). However, this is similar to the expression level of transcripts in these body parts (median of gene expression in ovary>gut>head: 21.2>15.10>12.47).

Interestingly, chimeric transcripts found in all five strains also have significantly higher expression levels than strain-specific chimeric transcripts (Wilcoxon's test, *p*-value < 0.001; Supplemental Table S6D).

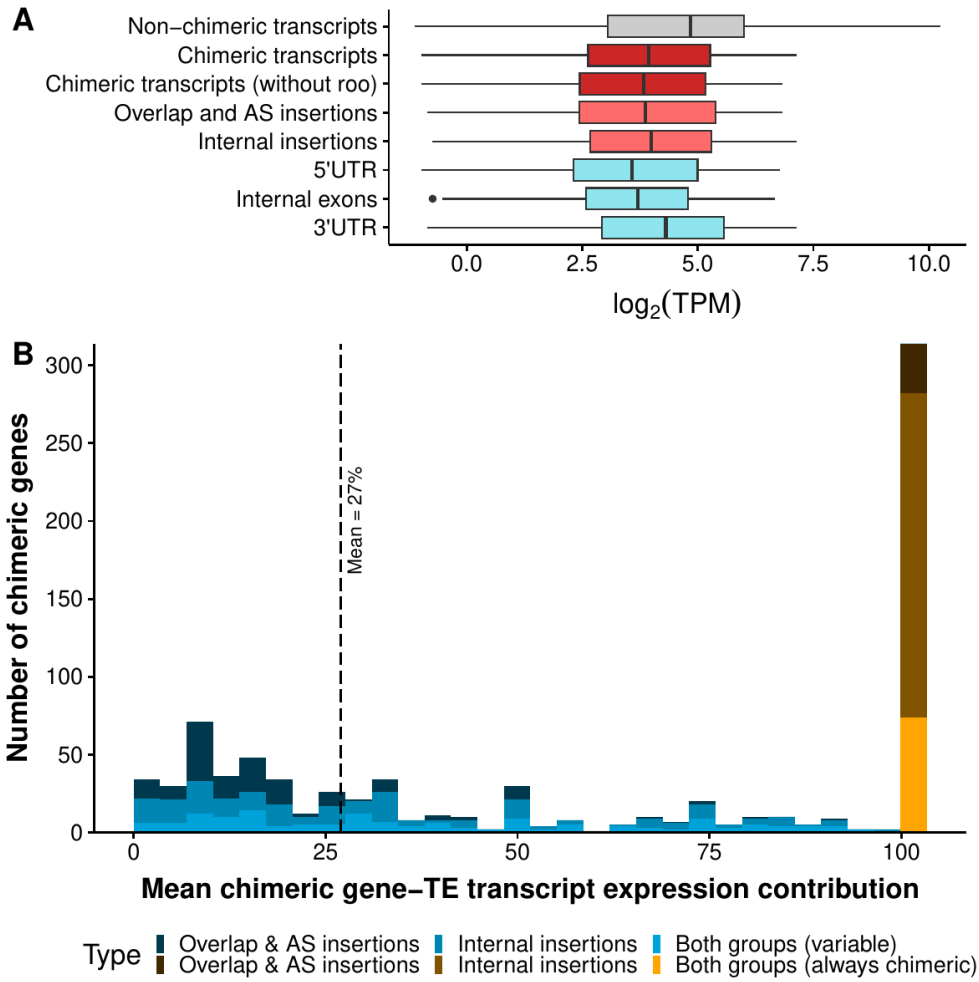


Figure 5. TE insertions within genes affect gene expression. A. Boxplots for the expression levels, measured as the logarithm of TPM: for all non-chimeric transcripts of the genome (19,228, in gray), all chimeric transcripts detected in the present study with TPM ≥1 (1,909, in dark red), chimeric transcripts without the short internal *roo* insertion (1,073, dark

red), all chimeric transcripts belonging to the *overlap and AS insertions* group (472, light red) and *internal insertions* group (1,156, light red), and chimeric transcripts divided by position of the insertion (5'UTR: 354, internal exons: 445, 3'UTR: 799, cadet blue). **B.** Histogram showing the expression contribution of chimeric transcripts to the total gene expression. Blue bars represent the contribution of variable chimeric genes (505 genes), ranging from ~0.1% to >90% (median: 27%) and the orange/brown bar represents the genes that always produced chimeric transcripts in all the genomes and body parts where expression was detected (314 genes).

17.1% of the TEs within chimeric gene-TE transcripts could also be affecting gene expression via epigenetic changes

We tested whether TEs that produce chimeric transcripts could also be affecting gene expression by affecting the epigenetic marks. We used ChIP-seq experiments previously performed in our lab (Coronado-Zamora et al. 2023) for the three body parts, in each of the five strains analyzed, for two histone marks: the silencing mark H3K9me3 (Choi and Lee 2020; Yin et al. 2011) and the activating mark H3K27ac, related to active promoters and enhancers (Buecker and Wysocka 2012; Koenecke et al. 2016). We tested whether strains expressing the gene-TE chimeric transcript differed in the gene body epigenetic marks compared with strains that do not express the chimeric transcript (430 genes). For the majority of these genes (266), we did not observe consistent epigenetic patterns across samples expressing or not expressing the chimeric transcripts, and these genes were not further analyzed. Additionally, 42 genes did not harbor any epigenetic marks while 93 genes contained the same epigenetics mark(s) (H3K27ac, H3K9me3, or both marks) in strains with and without that particular chimeric transcript (Supplemental Table S7A). Overall, for 17.1% (72/420) of the genes, we observed a consistent change in the epigenetic status associated with the presence of the gene-TE chimera. This percentage is similar for the *overlap and AS* group and the internal insertion group (16.9% and 18.4%, respectively). The majority of TEs showing consistent changes in their epigenetic status were associated with gene down-regulation (68%; Table 1)

and this percentage increases when considering chimeric transcripts with insertions in 5'UTR: 88% (16/18) genes are associated with down regulation (Supplemental Table S7B).

Table 1. Expression fold-change associated with epigenetic status of strains expressing or not expressing the chimeric transcript. Highlighted in bold, genes showing the expected change in expression according to the gained/lost histone mark.

Fold change											
	Non-chimeric gene	No mark	No mark	No mark	H3K27ac	H3K9me3 and H3K27ac	H3K9me3 and H3K27ac	H3K27ac	H3K27ac	H3K9me3	H3K9me3 and H3K27ac
	Chimeric gene	H3K27ac	H3K9me3	H3K27ac and H3K9me3	No mark	H3K27ac	H3K9me3	H3K9me3	H3K9me3 and H3K27ac	H3K9me3 and H3K27ac	No mark
FC > 1		**4**	1	6	1	**3**	0	2	4	1	2
FC < 1		4	**2**	10	**10**	1	**2**	**1**	14	1	5

Gene-TE chimeric transcripts are enriched for DNA-binding molecular functions involved in regulation and development

To get insight into the biological processes and molecular functions in which the gene-TE chimeric transcripts are involved, we performed a gene ontology (GO) clustering analysis (Huang et al. 2009). We analyzed the chimeric genes detected in each body part separately, using as a background the total genes assembled in the corresponding body part. We found that chimeric genes are enriched in general cell functions, such as regulation, and development (Figure 6A and Supplemental Table S8A). Some functions are particular to a body part, *e.g.*, *transcription from RNA polymerase II promoter* in the head, *regulation of RNA biosynthetic process* in the gut, and *morphogenesis and cell communication and signaling* in the ovary. Note that the *overlap and AS insertions* group is enriched for biological processes that are body-part specific: *communication signaling and stimulus* in the head, *cellular amide and organitrogen compound metabolic processes* in the gut *and cell projection, assembly and organization* in the ovary (Figure 6A and Supplemental Table S8C).

Finally, regarding the molecular function, chimeric genes are enriched for *DNA binding* processes across body parts (Figure 6B and Supplemental Table S8B), while in head they are also enriched for *regulatory region nucleic acid binding* and in ovary for *transcription factor activity* and *actin binding*.

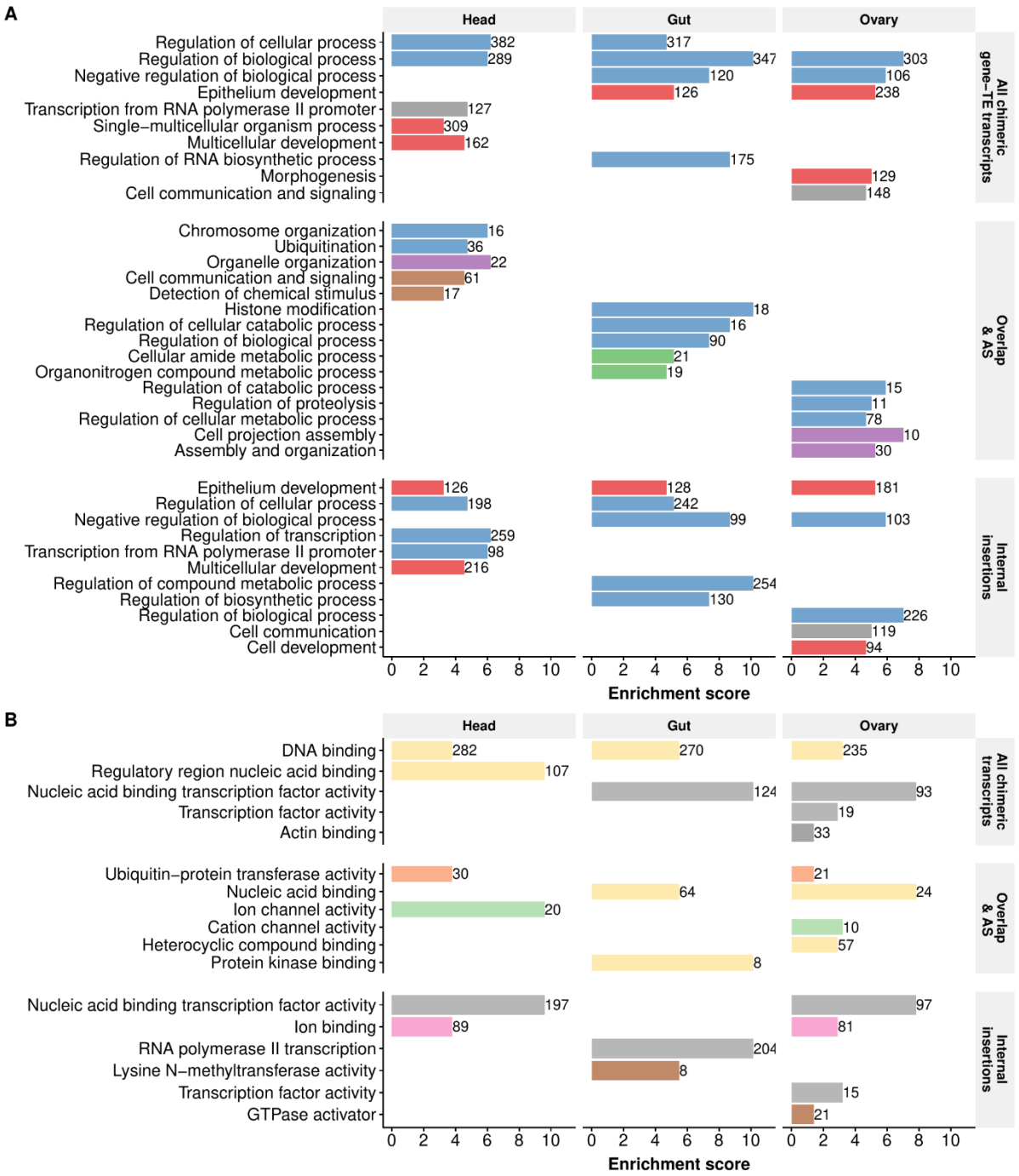


Figure 6. Biological processes and molecular functions of chimeric gene-TE transcripts. A. Biological processes clustering. **B.** Molecular functions clustering. The length

of the bars represents the cluster enrichment score. The number in the bars represents the number of genes in each cluster. Names of the annotation clusters are manually processed based on the cluster's GO terms. Colors represent similar annotation clusters. Detailed GO terms of each cluster are given in Supplemental Table S8.

Both DNA transposons and retrotransposons add functional protein domains

We next assessed whether TE sequences annotated in internal exons provided functional protein domains. We first confirmed, using the Coding Potential Assessment Tool (CPAT; Wang et al. 2013) software, that the majority of chimeric protein-coding gene-TE transcripts that have a TE annotated have coding potential (94.3%: 1,687/1,789; Supplemental Table S9A). Using PFAM (Mistry et al. 2021), we identified a total of 23 PFAM domains in 33 different chimeric transcripts from 24 genes (Table 2 and Supplemental Table S9B and S9C). These 23 domains were identified in 16 TE families, with 9 TE families providing more than one domain. The size of these domains ranged from 9bp to 610bp (median of 72bp; Supplemental Table S9B). Note that 9 of these 24 chimeric genes have been previously described in the literature (Table 2). We observed similar numbers of transcripts belonging to the *Overlap and AS insertions* group and *Internal insertions* group (12 and 16, respectively, with 5 belonging to both). Finally, we found chimeric transcripts adding domains in the three body parts analyzed (Table 2), with an even distribution across them (head: 15, ovary: 13, gut: 12).

Both DNA transposons and retrotransposons provide domains (13/24 and 11/24, respectively) and most TEs provided a nearly-full domain (22/24, ≥50% coverage), including 6 TEs adding a full-size domain (Table 2). Chimeric genes were related, among others, to transporter functions (11/24), enzyme functions (9/24) and response to stimulus (5/24) (Supplemental Table S9C). All these chimeric genes have evidence of expression, ranging from 1.43 to 61.69 TPM (Table 2, median = 18.16 TPM). We did not observe statistical differences between the level of expression of transcripts with complete domains compared to partially/uncompleted domains (Wilcoxon's test, *p*-value = 0.535). The majority of TEs for which the population TE

frequency has been reported (Rech et al. 2022), are fixed or present at high frequencies (8/14 TEs; Table 2).

We assessed if the domains detected in the TE fragment of the gene-TE chimera were also found in the consensus sequence of the TE family. Because most TE families were providing more than one domain, in total we analyzed 47 unique domains. We were able to find the domain sequence for 44 unique domains from 21 TE consensus sequences (Supplemental Table S9D). Note that for seven of these domains (from seven TEs), we had to lower PFAM detection thresholds to detect them (see *Methods*). The three domains that were not identified in the consensus sequences, were not detected in the chimeric fragments as full domain sequences.

We performed a PFAM domain enrichment analysis considering domains annotated with nearly-full domains and in transcripts expressed with a minimum of 1 TPM using dcGO (Fang and Gough 2013). Overall, seven domains were enriched for the molecular function *catalytic activity, acting on RNA* (4 domains, FDR = 3.44×10^{-4}), nucleic acid binding (6 domains, FDR = 3.44×10^{-4}), DNA polymerase activity (3 domains, FDR = 6.81×10^{-5}) and RNA-directed DNA polymerase activity (3 domains, FDR = 4.66×10^{-7}) (Table 3). Five out of seven enriched domains were found in retrotransposon insertions.

Table 2. Description of the 24 chimeric genes containing a TE providing a protein domain.

Superscript numbers in the *Gene* column represent literature describing these chimeric genes: [1] Treiber and Waddell 2020, [2] Lipatov et al. 2005, [3] Magwire et al. 2011, [4] Aminetzach et al. 2005, [5] Oliveira et al. 2022, [6] Ellison et al. 2020 [7] Bogwitz et al. 2005, [8] Guio et al. 2018 [9] Marsano et al. 2005. TPM is the gene-TE chimera expression level and it is the average across body parts and strains. NA in the *splicing motifs* column represents cases in which there are not splicing signals because the TE was found inside an exon (*internal insertion* group) while NC stands for non-canonical splicing motif. TE frequency (*Freq.*) was retrieved from Rech et al. (2022).

Gene	TE class: family	PFAM domains (%coverage)	TPM	Splicing motifs	Freq.	Body parts
CHKov1[1,2,3,4,5]	RNA: Doc	Exo_endo_phos_2 (98.32%); PRE_C2HC (98.53%); RVT_1 (100%)	34.57	NA	0.85	Head, Ovary, Gut
nxf2[5,6]	RNA: TART-A	TAP_C (89.8%)	39.16	NA	1.00	Head, Ovary, Gut
Cyp12a4[1,7,8,9]	DNA: Bari1	DDE_3 (89.38%); HTH_28 (98.08%); HTH_Tnp_Tc3_2 (100%)	9.13	GT/AG	1.00	Head, Ovary, Gut
CG7465	RNA: NewFam16	GYR (97.22%); YLP (88.89%)	19.24	NC/AG	1.00	Head, Gut
l(2)09851	DNA: Tc3	DDE_3 (28.77%)	52.35	NA	NA	Head, Ovary
stw[1]	RNA: F-element	RVT_1 (53.15%)	22.39	NA	NA	Head
Atf6[1]	RNA: BS2	RVT_1 (59.23%)	16.22	GT/NC	NA	Head
CG14043	DNA: S-element	HTH_Tnp_Tc3_2 (50%)	61.69	NA	0.98	Head
Mrp4[1]	DNA: P-element	THAP (90.7%); Tnp_P_element (37.44%)	23.12	GT/AG	0.64	Head
FASN1	DNA: P-element	THAP (90.7%); Tnp_P_element (37.44%)	10.50	GT/AG	0.42	Head
Oseg6	DNA: pogo	HTH_23 (80%); HTH_Tnp_Tc5 (95.45%)	1.43	NA	0.02	Head
Brf	RNA: jockey	PRE_C2HC (98.53%)	20.62	NA	0.02	Head
Mal-A4	DNA: S-element	DDE_3 (21.23%); HTH_Tnp_Tc3_2 (69.44%)	16.62	GT/AG	1.00	Gut
Ppcs[1]	DNA: Bari1	DDE_3 (35.62%); HTH_28 (98.08%); HTH_Tnp_Tc3_2 (100%)	12.93	NA	1.00	Gut

CG32032	RNA: jockey	Exo_endo_phos_2 (99.16%); PRE_C2HC (98.53%); RVT_1 (100%)	10.04	NA	0.06	Gut
RanBPM	DNA: pogo	HTH_23 (80%); HTH_Tnp_Tc5 (95.45%)	4.36	NA	0.02	Gut
CG3635	RNA: Blastopia	RT_RNaseH_2 (97%)	10.00	NA	NA	Gut
CG42748	RNA: gypsy6	Gypsy (99.79%)	15.83	NA	NA	Gut
l(2)05714^{5}	DNA: S-element	HTH_Tnp_Tc3_2 (50%)	5.33	NA	0.98	Ovary
ValRS-m	RNA: Doc	RVT_1 (100%)	34.64	NA	0.23	Ovary
Piezo	DNA: Bari1	DDE_3 (89.73%); HTH_28 (98.08%); HTH_Tnp_Tc3_2 (100%)	5.96	GT	0.02	Ovary
CG17883^{1}	RNA: Quasimodo	Baculo_F (22.79%)	15.80	NA	NA	Ovary
eIF4B	RNA: Invader2	Integrase_H2C2 (94.83%); RT_RNaseH (98.1%); RVT_1 (99.1%); rve (97.06%); zf-CCHC (88.89%)	5.56	NA	NA	Ovary
CG41099	RNA: Copia2	DUF4219 (96.3%); Retrotran_gag_2 (97.83%); gag_pre-integrs (86.57%); rve (95.1%)	1.99	NA	NA	Ovary

Table 3. PFAM domain enrichment analysis

dcGO enrichment results using 'Gene Ontology (GO)' with FDR < 0.01.

GO term	Z-score	FDR	Annotated domains
Catalytic activity, acting on RNA	5.99	3.44×10^{-4}	PF00078 (RVT_1); PF00098 (zf-CCHC); PF00665 (rve); PF13976 (gag_pre-integrs)
Nucleic acid binding	4.62	3.44×10^{-4}	PF00098 (zf-CCHC); PF00665 (rve); PF03221 (HTH_Tnp_Tc5); PF13976 (gag_pre-integrs); PF03943 (TAP_C); PF05485 (THAP)
DNA polymerase activity	11.40	6.81×10^{-5}	PF00078 (RVT_1); PF00665 (rve); PF13976 (gag_pre-integrs)
RNA-directed DNA polymerase activity	27.54	4.66×10-7	PF00078 (RVT_1); PF00665 (rve); PF13976 (gag_pre-integrs)

DISCUSSION

TEs contribute to genome innovation by expanding gene regulation, both of individual genes and of gene regulatory networks, enriching transcript diversity, and providing protein domains (*e.g.*, reviewed in Chuong et al. 2017 and Modzelewski et al. 2022). While the role of TEs as providers of regulatory sequences has been extensively studied, their contribution to transcriptome diversification and protein domain evolution has been less characterized. In this work, we have identified and characterized chimeric gene-TE transcripts across three body parts and five natural *D. melanogaster* strains, and we have quantified their contribution to total gene expression and to protein domains. While previous studies were hindered by the incomplete annotation of TEs in the genome studied (Lipatov et al. 2005; Treiber and Waddell 2020), in this work, we took advantage of the availability of high-quality genome assemblies and genome annotations for five natural strains to carry out an in-depth analysis of gene-TE chimeric transcripts (Rech et al. 2022). We found that TEs contribute ~9% to the global transcriptome and ~19% to the body part-specific transcriptome (Figure 1). Contrary to other studies that mostly focus on a single type of chimeric gene-TE transcript, we investigated a comprehensive dataset of chimeras. Indeed, we found that besides insertions affecting the transcription start site, transcript termination, and adding spliced sites (*overlap and alternative splicing insertions*), we also identified a substantial number of TE sequences that were completely embedded within exons (*internal insertions;* Figure 1D). These two types of chimeric gene-TE transcripts shared many properties, *e.g.,* they were enriched for body part-specific transcripts (Supplemental Fig. S2B), and they showed lower expression levels than non-chimeric transcripts (Figure 5A), suggesting that they both should be taken into account when analyzing the contribution of TEs to gene novelty. The *internal insertions* group contributed more to total gene expression (Figure 5B), however, we dismissed the possibility that this increased expression was due to shorter TE insertions, which are more likely to be enriched for false annotations compared with longer insertions (Rech et al. 2022). We found, both based on size and frequency, that the *internal insertions* group is likely to be enriched for

older insertions. As such, a higher level of expression of these likely older TEs is consistent with previous observations in tetrapods suggesting that over time gene-TE chimeric transcripts often become the primary or sole transcript for a gene (Cosby et al. 2021). Overall, and taking only into account those gene-TE chimeric transcripts with evidence of expression, we found 102 (5.9%) insertions disrupting the coding capacity, 435 (25.2%) affecting the coding capacity, 316 (18.3%) and 717 (41.6%) affecting the 5' and the 3' end of the gene, respectively, while 155 (9%) affected multiple transcript positions (Supplemental Table S6A).

Our finding that TEs contribute to the expansion of the head transcriptome supports the results of Treiber and Waddell (2020) and suggests that ~6% of genes produce chimeric transcripts in the head due to exonization of a TE insertion. However, because we also analyzed gut and ovary, we further show that TEs can significantly contribute to the expansion of other body parts' transcriptomes as well (Figure 2). The observation that there are more chimeric transcripts in the head is consistent with a higher transcriptional complexity in the *Drosophila* nervous system tissues (Brown et al. 2014). The fact that chimeric gene-TE transcripts tend to be tissue-specific could be especially relevant for adaptive evolution as tissue-specific genes can free the host from pleiotropic constraints and allow the exploration of new gene functions (Park and Choi 2010; Rogers and Hartl 2012; Salvador-Martínez et al. 2018).

Finally, we identified a total of 23 TE protein domains co-opted by 24 genes (Table 2 and Supplemental Table S9). Nine of these genes have been previously described as chimeric based on high-throughput screenings or individual gene studies, with some of them, *e.g. CHKov1* and *nxf2,* having functional effects (Aminetzach et al. 2005; Ellison et al. 2020; Magwire et al. 2011) (Table 2). The majority of the domains were present in the TE consensus sequences (Supplemental Table S9D). Furthermore, the 23 domains identified were enriched for *nucleic acid binding*, *catalytic activity*, and *DNA polymerase activity* molecular functions (Table 3). Although there is evidence for DNA binding domains being recruited to generate new genes, previous data comes from a comparative genomic approach across tetrapod genomes that focused on DNA transposons as a source of new protein domains (Cosby et al.

2021). The available data for the genome-wide contribution of retrotransposons to protein domains so far is restricted to endogenous retroviruses in mammals (Ueda et al. 2020). In our dataset, which includes both DNA transposons and retrotransposons, the enrichment for DNA binding domains and for catalytic activity is indeed driven by the retrotransposon insertions (Table 2). Although most of the TEs providing protein domains identified in this work for the first time were present at low population frequencies, three were fixed and two were present at high population frequencies and are thus good candidates for follow-up functional analysis (Table 2 and 3).

Although we have detected more chimeric transcripts than any prior *D. melanogaster* study to date, our estimate of the potential contribution of TEs to the diversification of the transcriptome is likely to be an underestimate. First, and as expected, we found that the contribution of TEs to the transcriptome is body part-specific (72.8%, Supplemental Fig. S2B; Conley et al. 2008; Faulkner et al. 2009) and strain-specific (68%, Supplemental Fig. S2A). Thus analyzing other body parts and increasing the number of genomes analyzed will likely identify more chimeric gene-TE transcripts. And second, although our estimate is based on the highly accurate annotations of TE insertions performed using the REPET pipeline (Rech et al. 2022), highly diverged and fragmented TE insertions are difficult to be accurately annotated by any pipeline and as such might go undetected (Gotea and Makałowski 2006; Rodriguez and Makałowski 2022). Still, the combination of an accurate annotation of chimeric gene-TE transcripts, with expression data across tissues, and the investigation of protein domain acquisition carried out in this work, not only significantly advances our knowledge on the role of TEs in gene expression and protein novelty, but also provides a rich resource for follow-up analysis of gene-TE chimeras.

METHODS

Fly stocks

Five *D. melanogaster* strains obtained from the European Drosophila Population Genomics Consortium (DrosEU), were selected according to their different geographical origins: AKA-017 (Akaa, Finland), JUT-011 (Jutland, Denmark), MUN-016 (Munich, Germany), SLA-001 (Slankamen, Serbia) and TOM-007 (Tomelloso, Spain). Further information on collection dates and localities can be found in Rech et al. (2022).

RNA-seq and ChIP-seq data for three body parts

RNA-seq and ChIP-seq data for the five strains were obtained from Coronado-Zamora et al. (2023). A full description of the protocols used to generate the data can be found in Coronado-Zamora et al. (2023). Briefly, head, gut and ovary body parts of each strain were dissected at the same time. Three replicates of 30 4-6 old-day females each were processed per body part and strain. RNA-seq library preparation was performed using the TruSeq Stranded mRNA Sample Prep kit from Illumina, and sequenced using Illumina 125bp paired-end reads (26.4M-68.8M reads; Supplemental Table S1). For ChIP-seq, libraries were performed using TruSeq ChIP Library Preparation Kit. Sequencing was carried out in an Illumina HiSeq 2500 platform, generating 50bp single-end reads (22.2M-59.1M reads; Supplemental Table S1). RNA-seq and ChIP-seq raw data are available in the NCBI Sequence Read Archive (SRA) database under BioProject PRJNA643665. The DrosOmics genome browser (http://gonzalezlab.eu/drosomics) compiles all data used in this work and allows for its visual exploration (Coronado-Zamora et al. 2023).

Transcriptome assembly

Reference-guided transcriptome assembly

To perform reference-guided transcriptome assemblies for each body part and strain (15 samples), we followed the protocol described in Pertea et al. (2015) using HISAT2 (v2.2.1,

Kim et al. 2019) and StringTie (v2.1.2, Pertea et al. 2015) . We used *D. melanogaster* r6.31 reference gene annotations (Larkin et al. 2021, available at: ftp://ftp.flybase.net/releases/FB2019_06/dmel_r6.31/gtf/dmel-all-r6.31.gtf.gz, last accessed: October 2020). We first used *extract_splice_sites.py* and *extract_exons.py* python scripts, included in the HISAT2 package, to extract the splice sites and exon information from the gene annotation file. Next, we build the HISAT2 index using *hisat2-build* (argument: *-p 12*) providing the splice sites and exon information obtained in the previous step in the *-ss* and *-exon* arguments, respectively. We performed the mapping of the RNA-seq reads (from the fastq files, previously analyzed with FastQC, Andrews 2010) with HISAT2 (using the command *hisat2 -p 12 --dta -x*). The output sam files were sorted and transformed into bam files using *samtools* (v1.6, Li et al. 2009). Finally, we used StringTie for the assembly of transcripts. We used the optimized parameters for *D. melanogaster* provided in Yang et al. (2018) to perform an accurate transcriptome assembly: *stringtie -c 1.5 -g 51 -f 0.016 -j 2 -a 15 -M 0.95*. Finally, *stringtie --merge* was used to join all the annotation files generated for each body part and strain. We used *gffcompare* (v0.11.2) from the StringTie package to compare the generated assembly with the reference *D. melanogaster* r.6.31 annotation, and the sensitivity and precision at the locus level were 99.7 and 98.5, respectively.

De novo transcriptome assembly

A *de novo* transcriptome assembly for each sample was performed using Trinity (v2.15.1, Grabherr et al. 2011) on the fastq files previously processed with *fastp* (v.0.23.2; Chen et al. 2018). Trinity used the following parameters: *--seqType fq --samples_file <txt file with fastq directory> --CPU 12 --max_memory 78 G --jaccard_clip*. Following Trinity recommended implementation (Haas et al. 2013), we combined all reads across technical replicates, in order to assemble them into a single reference assembly per sample (15 samples: 5 strains × 3 body-parts). The *--jaccard_clip* parameter was used to split falsely fused transcripts derived from gene-dense compact genomes. Next, to keep reliable near full-length transcripts, we used *blastn* (v2.11.0, Camacho et al. 2009) to assign each *de novo* transcript to a known *D.*

melanogaster transcript obtained from the *Reference-guided transcriptome assembly*. Next, the script *analyze_blastPlus_topHit_coverage.pl* from Trinity toolkit was used to evaluate the quality of the BLAST results, and we followed a conservative approach that only kept a transcript with a coverage higher than 80% with a known *D. melanogaster* transcript, thus, keeping 150,699 assembled transcripts across all samples. Then, we applied the algorithm proposed by Lima et al. (2017), to detect and remove falsely-fusioned transcripts that were not correctly assembled by Trinity. Briefly, we used *blastn* (Camacho et al. 2009) again to assign each transcript to a known *D. melanogaster* transcript and removed it if it had two or more matches with transcripts from different genes with a coverage of at least 80% and more than 100bp. As expected, most of the fusion transcripts were due to genes overlapping by their UTRs. Finally, we tried to minimize other possible sources of confounding errors by excluding transcripts that were not overlapping a known transcript (tagged by StringTie as *possible polymerase run-on; intron match on the opposite strand [likely a mapping error]; fully contained in a reference intron*; or *intergenic*). The set of final assembled transcripts by Trinity was 131,840.

Identification and characterization of chimeric gene-TE transcripts

We focused on the set of assembled *de novo* transcripts that passed the previous filters to identify putative chimeric gene-TE transcripts. To annotate TEs in the *de novo* assembled transcripts, we used RepeatMasker (v4.1.2, Smit et al. 2013, with parameters *-norna -nolow -s -cutoff 250 -xsmall -no_is -gff*) with a manually curated TE library (Rech et al. 2022). Note that RepeatMasker states that a cutoff of 250 will guarantee no false positives (Smit et al. 2013). We excluded transcripts for which the entire sequence corresponded to a transposable element, indicative of the autonomous expression of a TE, or TE fragments whose sequence contains simple repeats in more than >80% of their length, as determined by Tandem Repeats Finder (v4.09; Benson 1999), with the exception of *roo*-containing transcripts, that were analyzed separately. To infer the exon-intron boundaries of the transcript, we used *minimap2* (v2.24, Li 2018, with arguments *-ax splice --secondary=no --sam-hit-only -C5 -t4*) to align the

transcript to the gene region obtained from the genome of the corresponding strain from which it was assembled. We excluded single-transcript unit transcripts when they matched a multi-exon reference transcript, as it could be indicative of pervasive transcription or non-mature mRNAs. If a single-transcript unit transcript matches a single-exon reference transcript, it is considered as valid. With this process, we obtained the full-length transcript from the genome sequence.

We ran RepeatMasker again (same parameters) on the full-length transcripts to annotate the full TEs and obtain the length of the insertion. We considered that short insertions are those shorter than 120bp (Rech et al. 2022). Finally, we used an *ad-hoc* bash script to define the TE position within the transcript and define the two insertions groups: the *overlap and AS insertions* group and the *internal insertions* group. The *overlap and AS insertions* group have a TE overlapping with the first (5'UTR) or last (3'UTR) exon, or overlap with the exon-intron junction and thus introduce alternative splice sites (see *Splice sites motif scan analysis*). The *internal insertions group* corresponds to TE fragments detected inside exons.

TE-generating strain-specific chimeras frequency in other strains

To assess the frequency of TE insertions generating strain-specific chimeras, we examined whether the genomes where the chimeric transcript was not identified contained the TE insertion. For that, we used *blastn* (v2.11.0, Camacho et al. 2009) on the gene region, using the TE sequence of the strain-specific chimeric transcript as the query. We applied the following parameters: *-qcov_hsp_perc 80 and -perc_identity 80*. We then categorized the insertions as strain-specific, polymorphic, or fixed.

Splice sites motif scan analysis

We followed Treiber and Waddell (2020) approach to detect the splice acceptors and splice donor sites in the *alternative splice (AS) insertions* subgroup of chimeric gene-TE transcripts. In brief, we randomly extracted 11-12bp of 500 known donor and acceptor splice sites from

the reference *D. melanogaster* r.6.31 genome. Using the MEME tool (v5.4.1, Bailey and Elkan 1994), we screened for the donor and acceptor motifs in these two sequences, using default parameters. The obtained motifs were then searched in the predicted transposon-intron breakpoints position of our transcripts using FIMO (v5.4.1, Grant et al. 2011, with a significant *p*-value threshold of < 0.05).

***Roo* analyses**

Identification in the *roo* consensus sequence of the location of the *roo* low complexity region incorporated into gene-TE chimeric transcripts. To determine the location of the *roo* low complexity region in the *roo* consensus sequence, we downloaded the *roo* consensus sequence from FlyBase (version FB2015_02, Larkin et al. 2021, available at https://flybase.org/static_pages/downloads/FB2015_02/transposons/transposon_sequence_set.embl.txt.gz). We extracted the *roo* fragments detected in the chimeric gene-TE transcripts using *bedtools getfasta* (v2.30.0, Quinlan and Hall 2010), and used *blastn* (v2.11.0, Camacho et al. 2009) with parameters *-dust no -soft_masking false -word_size 7 -outfmt 6 -max_target_seqs 1 -evalue 0.05 -gapopen 5 -gapextend 2* to determine the matching position in the consensus sequence.

Identification of transcription factor binding sites in *roo* sequences. We retrieved from JASPAR (v2022, Castro-Mondragon et al. 2022) the models for all the transcription factor binding sites (TFBS) motifs of *D. melanogaster* (160 motifs). We used FIMO (v5.4.1, Grant et al. 2011) to scan for TBFS in the repetitive *roo* sequence from the consensus sequence (region: 1052-1166), as well as in the fragments incorporated in the gene-TE chimeras, with a significant threshold of 1×10^{-4}.

Genome-wide BLAST analysis of *roo* low complexity sequences. We performed a BLAST search with *blastn* (v2.11.0, Camacho et al. 2009) (with parameters: *-dust no -soft_masking false -outfmt 6 -word_size 7 -evalue 0.05 -gapopen 5 -gapextend 2 -qcov_hsp_perc 85 -perc_identity 75*). Next, we used *bedtools intersect* (v2.30.0, Quinlan and Hall 2010) with the

gene and transposable elements annotations to see in which positions the matches occur. We analyzed the top 20 matches of each *blastn* search.

Identification of *D. simulans roo* consensus sequence. We obtained a superfamily level transposable elements library for *D. simulans* using REPET (Flutre et al. 2011). We used *blastn* (v2.11.0, Camacho et al. 2009) with a minimum coverage and percentage of identity of 80% (*-qcov_hsp_perc 80 -perc_identity 80*) to find the sequence corresponding to the *roo* family. Then, we used again *blastn* (with parameters *-qcov_hsp_perc 80 -perc_identity 80 -dust no -soft_masking false -word_size 7 -max_target_seqs 1 -evalue 0.05 -gapopen 5 -gapextend 2*) to check if the *roo* sequence from *D. simulans* contained the repetitive region present in the *D. melanogaster roo* consensus sequence. The *roo* consensus sequence from *D. simulans* is available in the GitHub repository (https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster).

Retrotransposons and DNA transposons enrichment

We used the percentage of retrotransposons and DNA transposons of the genome of the five strains provided in Rech et al. (2022) and performed a χ^2 test to compare this percentage to the percentage of retrotransposons and DNA transposons detected in the chimeric gene-TE transcripts dataset.

Expression level estimation

The expression level of the whole set of transcripts assembled was computed by Trinity package (v2.15.1, Grabherr et al. 2011), which used *salmon* (Patro et al. 2017) as the abundance estimation method, and applied TPM (transcripts per million) as expression normalization. *Salmon* handles multi-mapped reads and applies an *expectation-maximization* algorithm that attributes read counts to the most likely transcripts. For the expression analyses, we considered transcripts with a minimum expression level of one TPM unless specified. Genes were categorized into three groups: (i) genes that were never detected as producing chimeric isoforms, (ii) genes that always were detected as producing chimeric gene-TE

transcripts and (iii) genes producing both chimeric and non-chimeric isoforms. For the later type of genes, we calculated the fraction of the total gene expression that comes from the chimeric transcript to compute the contribution of the chimeric transcripts to the total gene expression.

Coding capacity assessment

We assessed whether protein-coding chimeric gene-TE transcripts can produce a protein by using the Coding Potential Assessment Tool (CPAT, v3.0.4) software (Wang et al. 2013) with default parameters. CPAT has been optimized for the prediction of coding and non-coding isoforms in *Drosophila*. Thus, we used the coding probability cutoff at 0.39 (Wang et al. 2013).

PFAM scan of domain analysis and enrichment

To scan for PFAM domains (Mistry et al. 2021) in the TEs detected in an internal exon, we extracted the TE sequence from the chimeric transcripts using *bedtools getfasta* (v2.29.2, Quinlan and Hall 2010) translated it to ORF using *getorf* (EMBOSS:6.6.0.0, Rice et al. 2000) and scan it using the script *pfam_scan.pl* (v1.6, Finn et al. 2014) to identify any of the known protein family domains of the Pfam database (version 34). We used dcGO enrichment online tool (Fang and Gough 2013) to perform an enrichment of the PFAM domains detected.

We scanned the consensus TE sequences for the domains present in TE fragments detected in the chimerics transcripts using *pfam_scan.pl* (v1.6, Finn et al. 2014). If the domain was not detected using *pfam_scan.pl* default parameters, we lowered the hmmscan *e*-value sequence and domain cutoffs to 0.05.

Chip-seq peak calling

ChIP-seq reads were processed in Coronado-Zamora et al. (2023). Briefly, we used *fastp* (v0.20.1, Chen et al. 2018) to remove adaptors and low-quality sequences. Processed reads were mapped to the corresponding reference genome using the *readAllocate* function (parameter: *chipThres = 500*) of the Perm-seq R package (v0.3.0, Zeng et al. 2015), with

bowtie (v1.2.2, Langmead et al. 2009) as the aligner and the CSEM program (v2.3, Chung et al. 2011) in order to try to define a single location for multi-mapping reads. In all cases, bowtie was used with default parameters selected by Perm-seq.

Then, we used the ENCODE ChIP-Seq caper pipeline (v2, available at: https://github.com/ENCODE-DCC/chip-seq-pipeline2) in *histone* mode, using *bowtie2* as the aligner, disabling pseudo replicate generation and all related analyses (argument *chip.true_rep_only = TRUE*) and pooling controls (argument *chip.always_use_pooled_ctl = TRUE*). MACS2 peak caller was used with default settings. We used the output narrowPeak files obtained for each replicate of each sample to call the histone peaks. To process the peak data and keep a reliable set of peaks for each sample, we first obtained the summit of every peak and extended it ±100bp. Next, we kept those peaks that overlapped in at least two out of three replicates (following Yang et al. 2014) allowing a maximum gap of 100bp, and merged them in a single file using *bedtools merge* (v2.30.0, Quinlan and Hall 2010). Thus, we obtained for every histone mark of each sample a peak file. We considered that a chimeric gene-TE transcript had a consistent epigenetic status when the same epigenetic status was detected in at least 80% of the samples in which it was detected.

GO clustering analysis

The Gene Ontology (GO) clustering analysis in the biological process (BP) and molecular process (MP) category was performed using the DAVID bioinformatics online tool (Huang et al. 2009). Names of the annotation clusters were manually processed based on the cluster's GO terms. Only clusters with a score >1.3 were considered (Huang et al. 2009).

Statistical analysis

All statistical analyses were performed in R (v4.1.2) statistical computing environment (R Core Team 2021). Graphics were created using *ggplot2* R package (Wickham 2016).

DATA ACCESS

The set of chimeric gene-TE transcripts detected in this work are available in GitHub (https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster). Scripts to perform analyses are available at GitHub (https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster).

COMPETING INTEREST STATEMENT

The authors declare no competing interests.

ACKNOWLEDGMENTS

We thank Carlos Vargas-Chavez and Simón Orozco for providing the *D. simulans* REPET TE library. We thank Simón Orozco and Ewan Harney for comments on the manuscript. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (H2020-ERC-2014-CoG-647900), and from grant PID2020-115874GB-I00 funded by MCIN/AEI/10.13039/501100011033.

AUTHOR CONTRIBUTIONS

JG conceived the project; JG and MCZ designed the analyses; MCZ performed the analyses; JG and MCZ wrote and revised the manuscript.

REFERENCES

Aminetzach YT, Macpherson JM, Petrov DA. 2005. Pesticide resistance via transposition-mediated adaptive gene truncation in Drosophila. *Science* **309**: 764–767.

Andrews S. 2010. FastQC: A Quality Control Tool for High Throughput Sequence Data. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Babaian A, Thompson IR, Lever J, Gagnier L, Karimi MM, Mager DL. 2019. LIONS: analysis suite for detecting and quantifying transposable element initiated transcription from RNA-seq. *Bioinforma Oxf Engl* **35**: 3839–3841.

Babarinde IA, Ma G, Li Y, Deng B, Luo Z, Liu H, Abdul MM, Ward C, Chen M, Fu X, et al. 2021. Transposable element sequence fragments incorporated into coding and noncoding transcripts modulate the transcriptome of human pluripotent stem cells. *Nucleic Acids Res* **49**: 9132–9153.

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.

Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, Blencowe BJ, Frey BJ. 2010. Deciphering the splicing code. *Nature* **465**: 53–59.

Batut P, Dobin A, Plessy C, Carninci P, Gingeras TR. 2013. High-fidelity promoter profiling reveals widespread alternative promoter usage and transposon-driven developmental gene expression. *Genome Res* **23**: 169–180.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* **27**: 573–580.

Bogwitz MR, Chung H, Magoc L, Rigby S, Wong W, O'Keefe M, McKenzie JA, Batterham P, Daborn PJ. 2005. Cyp12a4 confers lufenuron resistance in a natural population of Drosophila melanogaster. *Proc Natl Acad Sci* **102**: 12807–12812.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome. *Nature* **512**: 393–399.

Buecker C, Wysocka J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet TIG* **28**: 276–284.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. *BMC Bioinformatics* **10**: 421. doi: 10.1186/1471-2105-10-421.

Casacuberta E, González J. 2013. The impact of transposable elements in environmental adaptation. *Mol Ecol* **22**: 1503–1517.

Castro-Mondragon JA, Riudavets-Puig R, Rauluseviciute I, Berhanu Lemma R, Turchi L, Blanc-Mathieu R, Lucas J, Boddie P, Khan A, Manosalva Pérez N, et al. 2022. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res* **50**: D165–D173.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**: i884–i890.

Choi JY, Lee YCG. 2020. Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLOS Genet* **16**: e1008872. doi: 10.1371/journal.pgen.1008872.

Chung D, Kuan PF, Li B, Sanalkumar R, Liang K, Bresnick EH, Dewey C, Keleş S. 2011. Discovering Transcription Factor Binding Sites in Highly Repetitive Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLOS Comput Biol* **7**: e1002111. doi: 10.1371/journal.pcbi.1002111.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from

conflicts to benefits. *Nat Rev Genet* **18**: 71–86.

Conley AB, Piriyapongsa J, Jordan IK. 2008. Retroviral promoters in the human genome. *Bioinforma Oxf Engl* **24**: 1563–1567.

Cordaux R, Udit S, Batzer MA, Feschotte C. 2006. Birth of a chimeric primate gene by capture of the transposase gene from a mobile element. *Proc Natl Acad Sci* **103**: 8101–8106.

Coronado-Zamora M, Salces-Ortiz J, González J. 2023. DrosOmics: A Browser to Explore -omics Variation Across High-Quality Reference Genomes From Natural Populations of Drosophila melanogaster. *Mol Biol Evol* **40**: msad075. doi: 10.1093/molbev/msad075.

Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, Feschotte C. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. *Science* **371**: eabc6405. doi: 10.1126/science.abc6405.

Cowley M, Oakey RJ. 2013. Transposable Elements Re-Wire and Fine-Tune the Transcriptome. *PLOS Genet* **9**: e1003234. doi: 10.1371/journal.pgen.1003234.

Ellison CE, Kagda MS, Cao W. 2020. Telomeric TART elements target the piRNA machinery in Drosophila. *PLOS Biol* **18**: e3000689. doi: 10.1371/journal.pbio.3000689.

Fang H, Gough J. 2013. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic Acids Res* **41**: D536–D544.

Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, Schroder K, Cloonan N, Steptoe AL, Lassmann T, et al. 2009. The regulated retrotransposon transcriptome of mammalian cells. *Nat Genet 2009 415* **41**: 563–571.

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al. 2014. Pfam: the protein families database. *Nucleic Acids Res* **42**: D222-230.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE **6**: e16526. doi: 10.1371/journal.pone.0016526.

Franchini LF, Ganko EW, McDonald JF. 2004. Retrotransposon-gene associations are widespread among D. melanogaster populations. *Mol Biol Evol* **21**: 1323–1331.

Ganko EW, Bhattacharjee V, Schliekelman P, McDonald JF. 2003. Evidence for the contribution of LTR retrotransposons to C. elegans gene evolution. *Mol Biol Evol* **20**: 1925–1931.

Gotea V, Makałowski W. 2006. Do transposable elements really contribute to proteomes? *Trends Genet* **22**: 260–267.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol* **29**: 644–652.

Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**: 1017–1018.

Guio L, Vieira C, González J. 2018. Stress affects the epigenetic marks added by natural transposable element insertions in Drosophila melanogaster. *Sci Rep 2018 81* **8**: 12197. doi: 10.1038/s41598-018-30491-w.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* **8**: 1494–1512.

Huang DW, Sherman BT, Zheng X, Yang J, Imamichi T, Stephens R, Lempicki RA. 2009. Extracting biological meaning from large gene lists with DAVID. *Curr Protoc Bioinforma*

Chapter 13: Unit 13.11.

Jordan IK, Rogozin IB, Glazko GV, Koonin EV. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet* **19**: 68–72.

Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**: 907–915.

Kiyose H, Nakagawa H, Ono A, Aikata H, Ueno M, Hayami S, Yamaue H, Chayama K, Shimada M, Wong JH, et al. 2022. Comprehensive analysis of full-length transcripts reveals novel splicing abnormalities and oncogenic transcripts in liver cancer. *PLOS Genet* **18**: e1010342. doi: 10.1371/journal.pgen.1010342.

Klasberg S, Bitard-Feildel T, Callebaut I, Bornberg-Bauer E. 2018. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J* **285**: 2605–2625.

Koenecke N, Johnston J, Gaertner B, Natarajan M, Zeitlinger J. 2016. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. *Genome Biol* **17**: 196. doi: 10.1186/s13059-016-1057–2.

Lanciano S, Cristofari G. 2020. Measuring and interpreting transposable element expression. *Nat Rev Genet* **21**: 721–736.

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**: R25. doi: 10.1186/gb-2009-10-3-r25.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, Santos G dos, Garapati PV, Goodman JL, Gramates LS, Millburn G, Strelets VB, et al. 2021. FlyBase: updates to the Drosophila melanogaster knowledge base. *Nucleic Acids Res* **49**: D899–D907.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, Sturgill D, Zhang Y, Oliver

B, Clark AG. 2008. Evolution of protein-coding genes in Drosophila. *Trends Genet TIG* **24**: 114–123.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. *eLife* **6**: e25762. doi: 10.7554/eLife.25762.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094–3100.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lima L, Sinaimeri B, Sacomoto G, Lopez-Maestre H, Marchet C, Miele V, Sagot M-F, Lacroix V. 2017. Playing hide and seek with repeats in local and global de novo transcriptome assembly of short RNA-seq reads. *Algorithms Mol Biol* **12**: 2: doi: 10.1186/s13015-017-0091–2.

Lipatov M, Lenkov K, Petrov DA, Bergman CM. 2005. Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. *BMC Biol* **3**: 24. doi: 10.1186/1741-7007-3-24.

Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. 2011. Successive Increases in the Resistance of Drosophila to Viral Infection through a Transposon Insertion Followed by a Duplication. *PLOS Genet* **7**: e1002337. doi: 10.1371/journal.pgen.1002337.

Marasca F, Gasparotto E, Polimeni B, Vadalà R, Ranzani V, Bodega B. 2020. The Sophisticated Transcriptional Response Governed by Transposable Elements in Human Health and Disease. *Int J Mol Sci* **21**: 3201. doi: 10.3390/ijms21093201.

Marsano RM, Caizzi R, Moschetti R, Junakovic N. 2005. Evidence for a functional interaction

between the Bari1 transposable element and the cytochrome P450 cyp12a4 gene in Drosophila melanogaster. *Gene* **357**: 122–128.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**: D412–D419.

Modzelewski AJ, Gan Chong J, Wang T, He L. 2022. Mammalian genome innovation through transposon domestication. *Nat Cell Biol* **24**: 1332–1340.

Newman RM, Hall L, Kirmaier A, Pozzi L-A, Pery E, Farzan M, O'Neil SP, Johnson W. 2008. Evolution of a TRIM5-CypA Splice Isoform in Old World Monkeys. *PLOS Pathog* **4**: e1000003. doi: 10.1371/journal.ppat.1000003.

Oliveira DS, Fablet M, Larue A, Vallier A, Carareto CMA, Rebollo R, Vieira C. 2022. ChimeraTE: A pipeline to detect chimeric transcripts derived from genes and transposable elements. *bioRxiv* doi: 10.1101/2022.09.05.505575.

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. 2008. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* **40**: 1413–1415.

Park SG, Choi SS. 2010. Expression breadth and expression abundance behave differently in correlations with evolutionary rates. *BMC Evol Biol* **10**: 241. doi: 10.1186/1471-2148-10-241.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. 2017. Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. *Nat Methods* **14**: 417–419.

Peel MC, Finlayson BL, McMahon TA. 2007. Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci* **11**: 1633–1644.

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol* **33**: 290–295.

Pinson M-E, Pogorelcnik R, Court F, Arnaud P, Vaurs-Barrière C. 2018. CLIFinder: identification of LINE-1 chimeric transcripts in RNA-seq data. *Bioinformatics* **34**: 688–690.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria https://www.R-project.org/.

Rebollo R, Horard B, Begeot F, Delattre M, Gilson E, Vieira C. 2012. A snapshot of histone modifications within transposable elements in Drosophila wild type strains. *PloS ONE* **7**: e44253. doi: 10.1371/journal.pone.0044253.

Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, Lindstadt H, Jamilloux V, Quesneville H, González J. 2022. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. *Nat Commun* **13**: 1948. doi: 10.1038/s41467-022-29518–8.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet TIG* **16**: 276–277.

Rodriguez M, Makałowski W. 2022. Software evaluation for de novo detection of transposons. *Mob DNA* **13**: 14. doi: 10.1186/s13100-022-00266–2.

Rogers RL, Hartl DL. 2012. Chimeric Genes as a Source of Rapid Evolution in Drosophila melanogaster. *Mol Biol Evol* **29**: 517–529.

Salvador-Martínez I, Coronado-Zamora M, Castellano D, Barbadilla A, Salazar-Ciudad I.

2018. Mapping Selection within Drosophila melanogaster Embryo's Anatomy. *Mol Biol Evol* **35**: 66–79.

Schrader L, Schmitz J. 2019. The impact of transposable elements in adaptive evolution. *Mol Ecol* **28**: 1537–1549.

Singh P, Ahi EP. 2022. The importance of alternative splicing in adaptive evolution. *Mol Ecol* **31**: 1928–1938.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Söllner JF, Leparc G, Hildebrandt T, Klein H, Thomas L, Stupka E, Simon E. 2017. An RNA-Seq atlas of gene expression in mouse and rat normal tissues. *Sci Data* **4**: 170185. doi: 10.1038/sdata.2017.185.

Tipney HJ, Hinsley TA, Brass A, Metcalfe K, Donnai D, Tassabehji M. 2004. Isolation and characterisation of GTF2IRD2, a novel fusion gene and member of the TFII-I family of transcription factors, deleted in Williams–Beuren syndrome. *Eur J Hum Genet* **12**: 551–560.

Treiber CD, Waddell S. 2020. Transposon expression in the Drosophila brain is driven by neighboring genes and diversifies the neural transcriptome. *Genome Res* **30**: 1559–1569.

Ueda MT, Kryukov K, Mitsuhashi S, Mitsuhashi H, Imanishi T, Nakagawa S. 2020. Comprehensive genomic analysis reveals dynamic evolution of endogenous retroviruses that code for retroviral-like protein domains. *Mob DNA* **11**: 29. doi: 10.1186/s13100-020-00224-w.

van de Lagemaat LN, Landry J-R, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet* **19**: 530–536.

Verta J-P, Jacobs A. 2022. The role of alternative splicing in adaptation and evolution. *Trends Ecol Evol* **37**: 299–308.

Volff JN. 2006. Turning junk into gold: Domestication of transposable elements and the creation of new genes in eukaryotes. *BioEssays* **28**: 913–922.

Wang L, Park HJ, Dasari S, Wang S, Kocher J-P, Li W. 2013. CPAT: Coding-Potential Assessment Tool using an alignment-free logistic regression model. *Nucleic Acids Res* **41**: e74. doi: 10.1093/nar/gkt006.

Wickham H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York https://ggplot2.tidyverse.org.

Yang H, Jaime M, Polihronakis M, Kanegawa K, Markow T, Kaneshiro K, Oliver B. 2018. Re-annotation of eight Drosophila genomes. *Life Sci Alliance* **1**: e201800156. doi: 10.26508/lsa.201800156.

Yang Y, Fear J, Hu J, Haecker I, Zhou L, Renne R, Bloom D, McIntyre LM. 2014. Leveraging biological replicates to improve analysis in ChIP-seq experiments. *Comput Struct Biotechnol J* **9**: e201401002. doi: 10.5936/csbj.201401002.

Yin H, Sweeney S, Raha D, Snyder M, Lin H. 2011. A High-Resolution Whole-Genome Map of Key Chromatin Modifications in the Adult Drosophila melanogaster. *PLOS Genet* **7**: e1002380. doi: 10.1371/journal.pgen.1002380.

Zeng X, Li B, Welch R, Rojo C, Zheng Y, Dewey CN, Keleş S. 2015. Perm-seq: Mapping Protein-DNA Interactions in Segmental Duplication and Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLOS Comput Biol* **11**: e1004491. doi: 10.1371/journal.pcbi.1004491.