Supplemental Material

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.

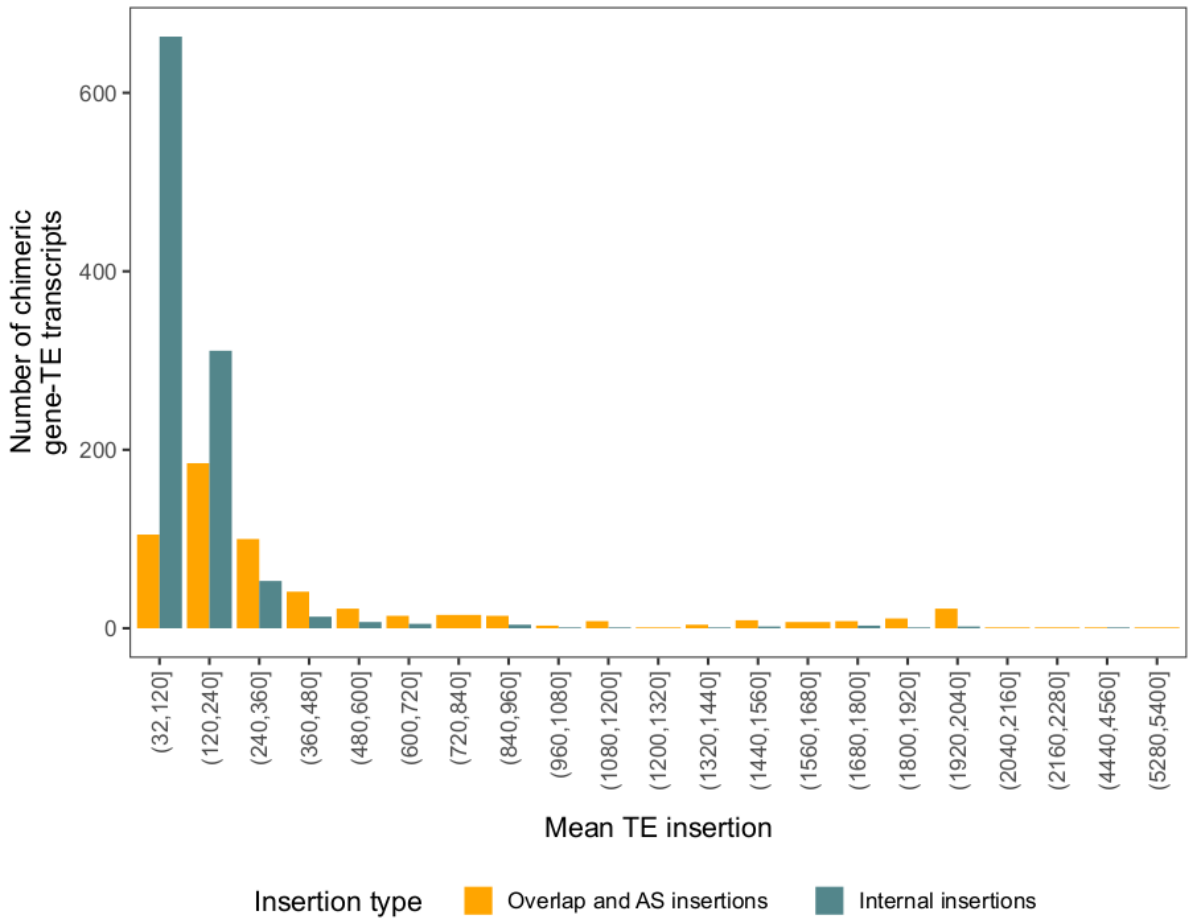
Marta Coronado-Zamora, marta.coronado@ibe.upf-csic.es

*Corresponding author

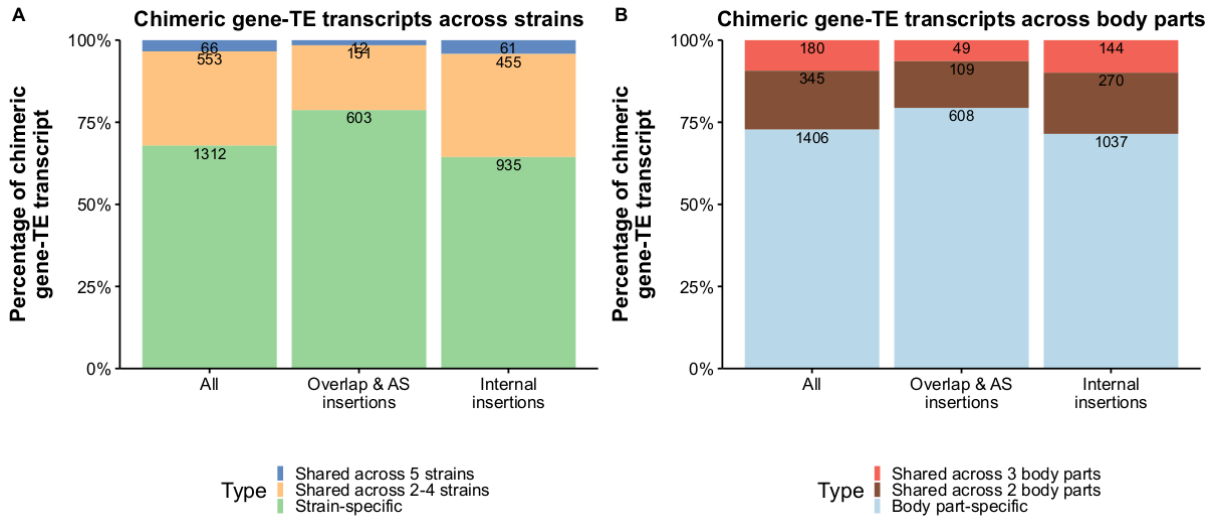
josefa.gonzalez@csic.es

Table of contents:

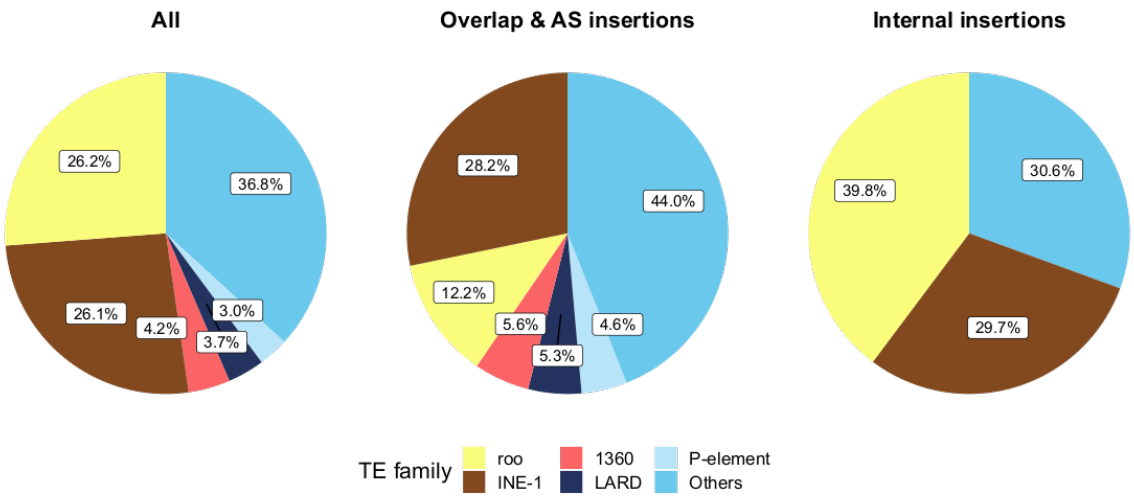
Supplemental Figures



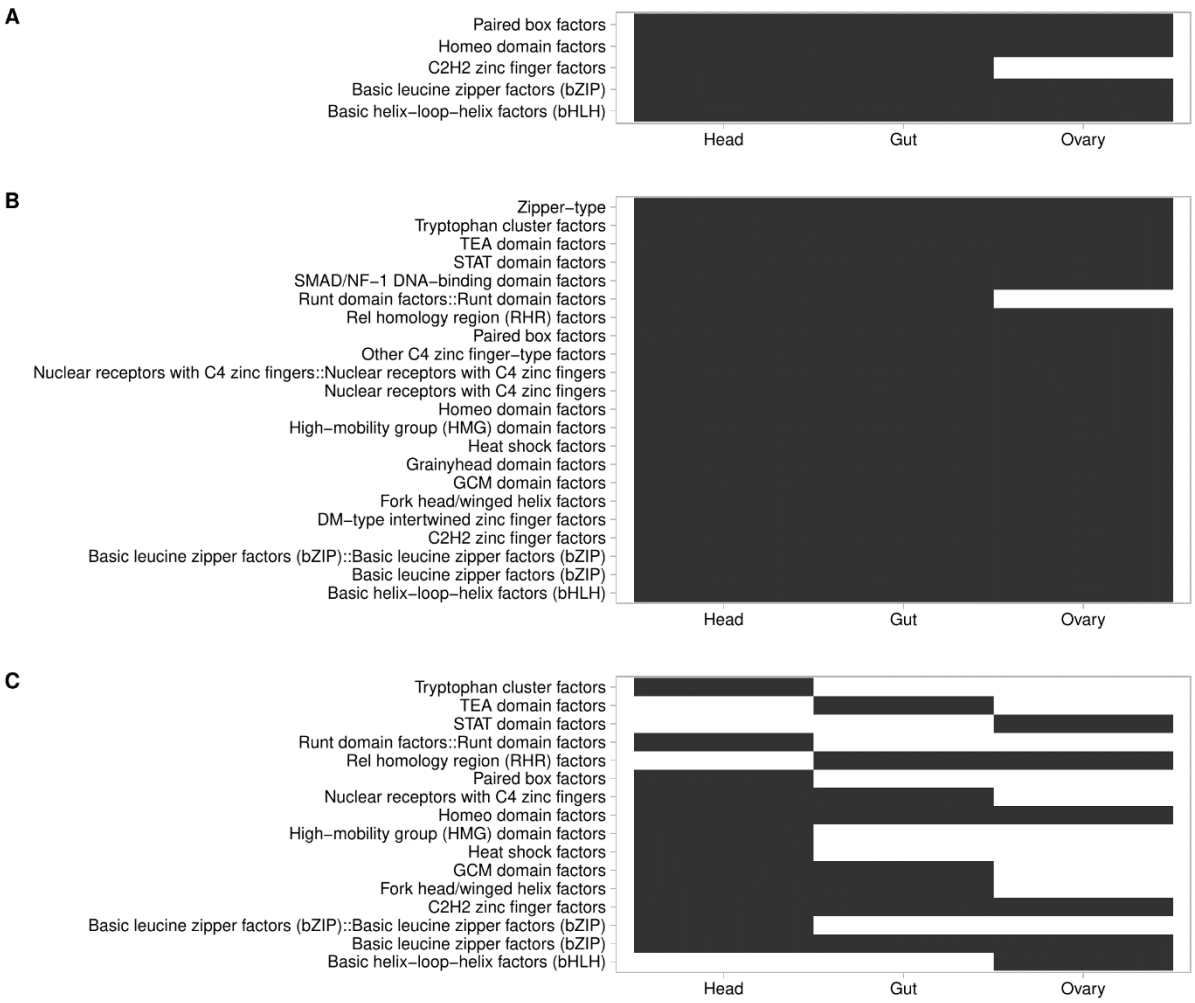
Supplemental Figure 1. Histogram of the mean TE insertion length (bp) in chimeric gene-TE transcripts of the *overlap and AS insertions* and *internal insertions* group. 174 out of 766 (22.7%) chimeric transcripts from the *overlap and AS insertions* group contain a fragment of a TE insertion < 120bp. 1,131 out of 1,451 (78%) chimeric transcripts from the *internal insertions* group contain a fragment of a TE insertion < 120bp.



Supplemental Figure 2. Percentage of chimeric gene-TE transcripts strains and body parts. **A.** Bar plot showing the percentage of chimeric transcripts detected across strains. In the global set of chimeric transcripts (*All*), in the *Overlap and AS insertions* group, and the *Internal insertions* group. **B.** Bar plot showing the percentage of chimeric transcripts detected across body parts. In the global set of chimeric transcripts (*All*), in the *Overlap and AS insertions* group, and the *Internal insertions* group.



Supplemental Figure 3. TE families distribution in gene-TE chimeras, globally and by insertion group considering insertions ≥ 120 bp. Percentage of TE families contributing to gene-TE chimeras considering insertions ≥ 120 bp in the global dataset (*All*), in the o*verlap 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*.



Supplemental Figure 4. Scan of transcription factor binding site (TFBS) motifs on the sequences of TE fragments. Black boxes represent the presence of the TFBS motifs while empty boxes represent the absence of motifs in a TE fragment from a chimeric transcripts detected in each of the three body parts. **A.** Results of the TFBS motif scan of *roo* solo LTR TE sequences incorporated in chimeric transcripts. **B.** Results of the scan of INE-1 fragments incorporated in chimeric transcripts. **C.** Results of the scan of TE sequences of body-part specific chimeric transcripts. Data available in Supplemental Table S5H.

Supplemental Methods

Reference-guided transcriptome assembly

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.

ChIP-seq peak calling

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 *Bowtie 2* 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.

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.

Supplemental Bibliography

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

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.

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

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.

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.

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

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.

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.

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.

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

26: 841–842.

Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, Fiston-Lavier A-S, Luyten I, Venkataram S, Quesneville H, et al. 2019. Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLOS Genet* **15**: e1007900.

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.

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.

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.

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.