

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

3

4

5 Marta Coronado-Zamora¹ and Josefa González^{1*}

6 ¹Institute of Evolutionary Biology, CSIC, UPF.

7

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

9

10 *Corresponding author

11 josefa.gonzalez@csic.es

12 **ABSTRACT**

13 Transcriptomes are dynamic, with cells, tissues, and body parts expressing particular sets of
14 transcripts. Transposons are a known source of transcriptome diversity, however studies often
15 focus on a particular type of chimeric transcript, analyze single body parts or cell types, or are
16 based on incomplete transposon annotations from a single reference genome. In this work,
17 we have implemented a method based on *de novo* transcriptome assembly that minimizes the
18 potential sources of errors while identifying a comprehensive set of gene-TE chimeras. We
19 applied this method to head, gut and ovary dissected from five *Drosophila melanogaster*
20 natural populations, with individual reference genomes available. We found that 18.6% of body
21 part specific transcripts are gene-TE chimeras. Overall, chimeric transcripts contribute a
22 median of 38% to the total gene expression, and they provide both DNA binding and catalytic
23 protein domains. Our comprehensive dataset is a rich resource for follow-up analysis.
24 Moreover, because transposable elements are present in virtually all species sequenced to
25 date, their relevant role in spatially restricted transcript expression is likely not exclusive to the
26 species analyzed in this work.

27 **INTRODUCTION**

28 In contrast to the genome, an animal's transcriptome is dynamic, with cell types, tissues and
29 body parts expressing particular sets of transcripts^{1–4}. The complexity and diversity of the
30 transcriptome arises from the combinatorial usage of alternative promoters, exons and introns,
31 and polyadenylation sites. A single gene can, therefore, encode a rich repertoire of transcripts
32 that can be involved in diverse biological functions, and contribute to adaptive evolution and
33 disease (e.g., ^{5–8}). The potential contribution of transposable element (TE) insertions to the
34 diversification of the transcriptome was analyzed soon after the first whole-genome sequences
35 were available^{9–13}. TEs are present in virtually all genomes studied to date, are able to insert
36 copies of themselves in the genome and, although their mutation capacity is often harmful,
37 they also represent an important source of genetic variation^{14–17}. While transposable elements
38 are a known source of transcriptome diversity, the majority of studies so far rely on incomplete
39 transposon annotations from a single reference genome (e.g., ¹²). Moreover, methodologies
40 are often specifically designed for particular types of chimeric gene-TE transcripts, e.g. TE-
41 initiated transcripts¹⁸, particular types of TEs, e.g. L1 chimeric transcripts¹⁹, or have been
42 applied to individual cell types or body parts, (e.g., ^{20,21}). As such, our knowledge on the
43 contribution of TEs to gene novelty is still partial.

44 Two of the most studied mechanisms by which TEs can generate chimeric transcripts are by
45 providing alternative promoters and protein domains. In human and mouse, 2.8% and 5.2%
46 of the total transcript start sites occurred within retrotransposons²². In *D. melanogaster*, over
47 40% of all genes are expressed from two or more promoters, with at least 1,300 promoters
48 contained in TEs²³. As well as individual examples of TEs providing protein domains^{24–26}, a
49 comparative genomic analysis of tetrapod genomes revealed that capture of transposase
50 domains is a recurrent mechanism for novel gene formation²⁷. There is also evidence for the
51 retrotransposon contribution to protein novelty. Approximately 9.7% of endogenous retrovirus
52 open reading frames across 19 mammalian genomes evolve under purifying selection and are
53 transcribed, suggesting that they could have been co-opted as genes²⁸. Across insects, and
54 depending on the methodology used, the percentage of newly emerged domains (<225 mya)
55 due to TEs was estimated to be 1.7% to 6.6%²⁹. However, studies that identify and
56 characterize a comprehensive set of gene-TE chimeras to provide a complete overview of
57 their contribution to both transcriptome and protein diversification are still missing.

58 Besides describing the diverse contributions of TEs to the transcriptome, analyzing the relative
59 contribution of gene-TE chimeras to the total gene expression is highly relevant, as it is
60 informative of the potential functional relevance of the transcripts identified. Studies performed
61 so far suggest that this contribution is related to the position of the TE in the transcript.

62 Transcripts with a TE inserted in the 5'UTR or internal coding exons show significantly lower
63 mean levels of expression compared with non-chimeric TE-gene transcripts²⁰. TEs inserted in
64 3'UTRs were associated with reduced gene expression both in humans and mice, but with
65 increased gene expression in human pluripotent stem cells^{20,22}. In addition, whether specific
66 TE types contribute to tissue-specific expression has been explored in mammals, where
67 retrotransposons were found to be overrepresented in human embryonic tissues^{22,30}. In *D.*
68 *melanogaster*, the contribution of TEs to tissue specific expression has only been assessed in
69 the head, with 833 gene-TE chimeric genes described²¹. Thus, whether the contribution of
70 chimeric gene-TE transcripts is more relevant in the *D. melanogaster* head compared with
71 other body parts is still an open question.

72 Within genes, TEs could also affect expression by changing the epigenetic status of their
73 surrounding regions. In *Drosophila*, repressive histone marks enriched at TEs spread beyond
74 TE sequences, which is often associated with gene down-regulation³¹. However, there is also
75 evidence that TEs containing active chromatin marks can lead to nearby gene
76 overexpression³². Genome-wide, the joint assessment of the presence of repressive and
77 active chromatin marks has been restricted so far to the analysis of four TE families³³ and has
78 never been carried out in the context of chimeric gene-TE transcripts.

79 In this work, we performed a high-throughput analysis to detect, characterize, and quantify
80 chimeric gene-TE transcripts in RNA-seq samples from head, gut, and ovary dissected from
81 the same individuals belonging to five natural strains of *D. melanogaster* (Figure 1A³⁴). We
82 implemented a method based on *de novo* transcriptome assembly that (i) minimizes the
83 potential sources of errors when detecting chimeric gene-TE transcripts; and (ii) allows to
84 identify a comprehensive dataset of transcripts rather than focusing on particular types (Figure
85 1B³⁵³⁵). Additionally, we assessed the coding potential and the contribution of chimeric
86 transcripts to protein domains and gene expression as proxies for their integrity and functional
87 relevance. Finally, we took advantage of the availability of ChIP-seq data for an active and a
88 repressive histone mark, H3K9me3 and H3K27ac, respectively obtained from the same
89 biological samples to investigate whether the TEs that are incorporated into the transcript
90 sequences also affect their epigenetic status.

91

92 **RESULTS**

93 **10% of *D. melanogaster* transcripts, across body parts and strains, are gene-TE**
94 **chimeras**

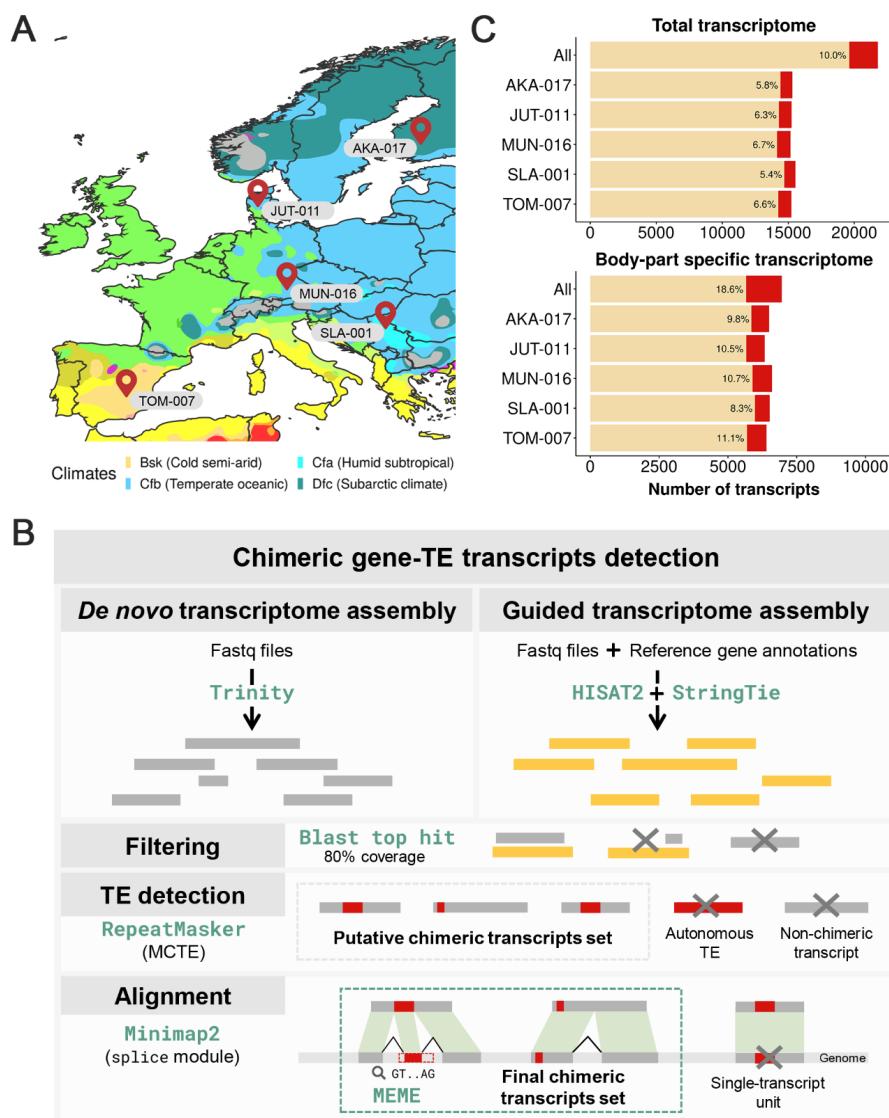
95 We performed a high-throughput analysis to detect and quantify chimeric gene-TE transcripts
96 in RNA-seq samples from head, gut, and ovary, in five *D. melanogaster* strains collected from
97 natural populations (Figure 1A). The three body parts were dissected from the same
98 individuals, and an average of 32x (22x to 43x) per RNA-seq sample was obtained (3
99 replicates per body part and strain, Table S1³⁶). We *de novo* assembled transcripts in which
100 we annotate TE insertions using the new *D. melanogaster* manually curated TE library³⁴. We
101 only considered *de novo* transcripts that overlap with a known transcript obtained from a
102 reference guided assembly (Figure 1B). We then used the reference genome of each strain
103 to define the exon-intron boundaries of each transcript and to identify the position of the TE in
104 the transcript (Figure 1B). The alignment with the reference genome and the accurate TE
105 annotation also allowed us to discard single-unit transcripts, indicative of pervasive
106 transcription, and TE autonomous expression, which are two important sources of errors when
107 quantifying the contribution of TEs to gene novelty (Figure 1B³⁵).

108 Overall, considering all the transcripts assembled in the three body parts and the five strains,
109 we identified 2,169 chimeric gene-TE transcripts belonging to 1,250 genes (Table S2A). Thus,
110 approximately 10% (2,169/21,786) of *D. melanogaster* transcripts contain exonic sequences
111 of TE origin. In individual strains, this percentage ranged between 5.4% to 6.7% (842-1,013
112 chimeric transcripts per genome) indicating that most of the chimeric gene-TE transcripts are
113 strain-specific, as expected given that the majority of TEs are present at low population
114 frequencies (Figure 1C³⁴). While the overall contribution of TEs to the transcriptome is 10%,
115 TEs contribute 18.6% (1,295/6,959) of the total amount of body part specific transcripts (Figure
116 1C).

117 We identified two groups of chimeric gene-TE transcripts (Figure 1D). The first group contains
118 chimeric transcripts which have a TE overlapping with the 5'UTR, the 3'UTR, or introducing
119 alternative splice (AS) sites (*overlap and AS insertions* group: 977 chimeric transcripts from
120 655 genes). While TEs have been reported to introduce non-canonical splice motifs²¹, we
121 found that the majority of the TEs in the *overlap and AS insertions* group were adding a
122 canonical AS motif (65.2%: 172/264) (Table S2B). The second group contains chimeric gene-
123 TE transcripts in which the TE is annotated completely inside the UTRs or internal exons
124 (*internal insertions* group: 1,587 transcripts from 890 genes) (Figure 1D). We hypothesized
125 that this group could be the result of older insertions that have been completely incorporated
126 into the transcripts. Indeed, we found that TEs in this group are shorter than those of the

127 *overlap and AS insertion* group, as expected if the former are older insertions (75.99% vs.
128 23.75%; test of proportions, *p*-value < 0.001; Figure S1; see *Methods*). Additionally, while the
129 majority of gene-TE transcripts in the *overlap and AS insertions* group were strain-specific, we
130 found more transcripts shared between strains than strain-specific in the *internal insertions*
131 group (test of proportions, *p*-value < 0.001; Figure S2A and Table S2C). This observation is
132 also consistent with this group being enriched for older insertions, and remained valid when
133 we removed the shorter insertions (test of proportions, *p*-value < 0.001; Table S2C).

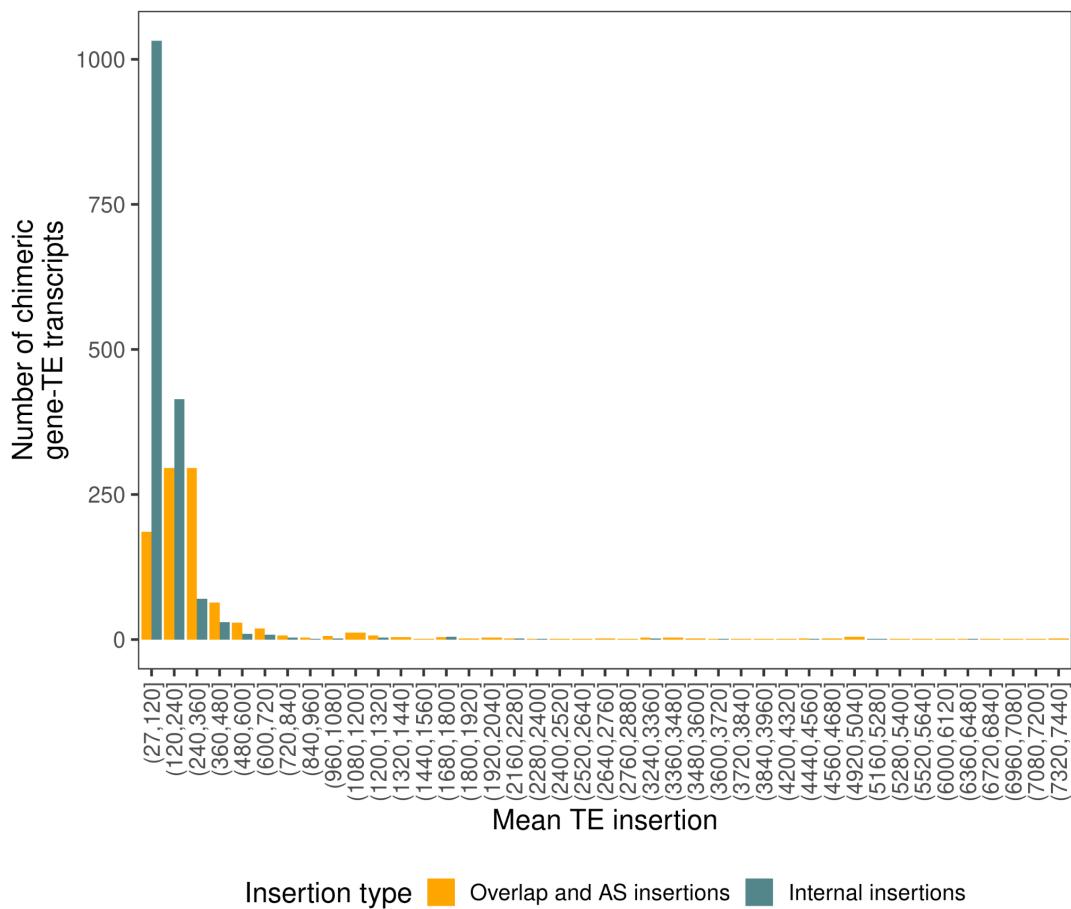
134 To test whether the *overlap and AS insertions* and the *internal insertions* groups contribute
135 differently to the diversification of the transcriptome, we performed all the subsequent analyses
136 considering all the chimeric transcripts together, and the two groups separately. In addition,
137 because shorter insertions might be enriched for false positives, *i.e.*, not corresponding to real
138 TE sequences due to the difficulty of annotating these repetitive regions, we also performed
139 the analysis with the subset of chimeric gene-TE transcripts that contains a fragment of a TE
140 insertion that is ≥ 120 bp (831/977 and 628/1587 for the *overlap and AS insertions* and the
141 *internal insertions* groups, respectively; see *Methods*).



142

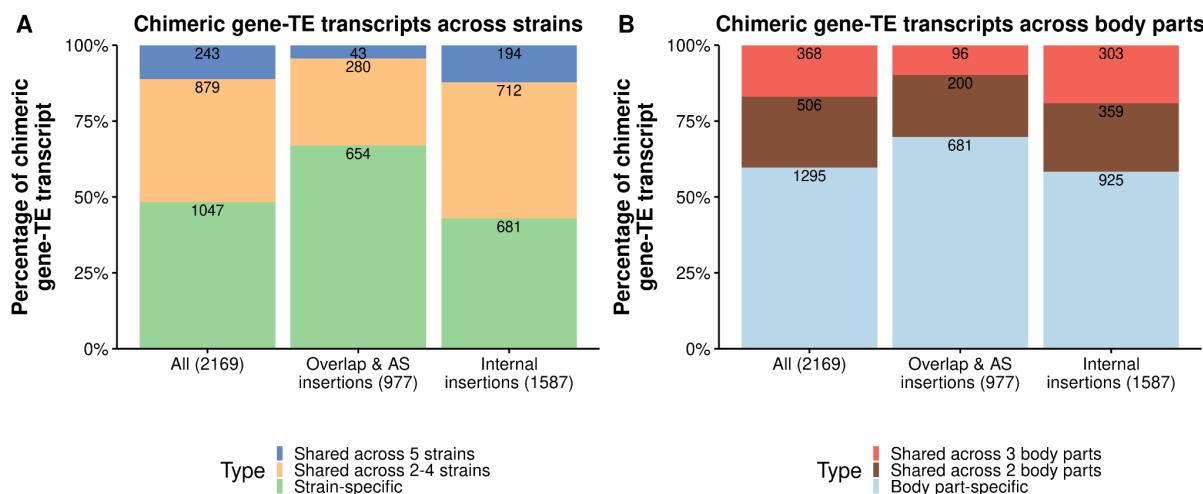
143 **Figure 1. Detection of chimeric gene-TE transcripts in five strains of *D. melanogaster*. A.** Map
 144 showing the sampling locations of the five European strains of *D. melanogaster* used in this study.
 145 TOM-007: Tomelloso, Spain (BSk); MUN-016: Munich, Germany (Cfb); JUT-011: Jutland, Denmark
 146 (Cfb); SLA-001: Slankamen, Serbia (Cfa); and AKA-017: Akaa, Finland (Dfc). Colors represent the
 147 climate zones according to the Köppen-Geiger climate distribution³⁷. **B.** Pipeline to detect chimeric
 148 transcripts. Two types of transcriptome assembly were performed: a *de novo* assembly using Trinity³⁸
 149 and a genome-guided transcriptome assembly using HISAT2³⁹ and StringTie⁴⁰. We only considered *de*
 150 *novo* transcripts that had a minimum 80% coverage with a known transcript to be screened for TEs

151 insertions. RepeatMasker⁴¹ is used with a manually curated TE library³⁴ to detect TEs in the *de novo*
152 assembled transcripts. An alignment against the reference genome of each strain is used to define the
153 exon-intron boundaries of transcripts and to identify the position of the TE in the transcript⁴². Transcripts
154 fully annotated as a TE or detected as single-transcript units are discarded. **C.** Contribution of chimeric
155 gene-TE transcripts to the total transcriptome and the body parts specific transcriptome globally and by
156 strain. *All* includes all the transcripts assembled in the three body parts and the five strains. **D.**
157 Schematic of the two groups of chimeric transcripts identified. *Overlap and alternative splicing (AS)*
158 *insertions* group, and *internal insertions* group. Note that these numbers total more than 2,169 because
159 some chimeric transcripts can have different insertions in different samples. Gray boxes represent
160 exons, red boxes represent a TE fragment incorporated in the mRNA, white boxes represent a TE
161 fragment that is not incorporated in the final mRNA. The black lines connecting the exons represent the
162 splicing events.



163

164 **Supplementary Figure 1. Histogram of the mean TE insertion length (bp) in chimeric gene-TE**
165 **transcripts of the overlap and AS insertions and internal insertions group.** 232 out of 977 (23.75%)
166 chimeric transcripts from the *overlap and AS insertions* group contain a fragment of a TE insertion <
167 120bp. 1,206 out of 1,587 (75.99%) chimeric transcripts from the *internal insertions* group contain a
168 fragment of a TE insertion < 120bp.



169

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

175

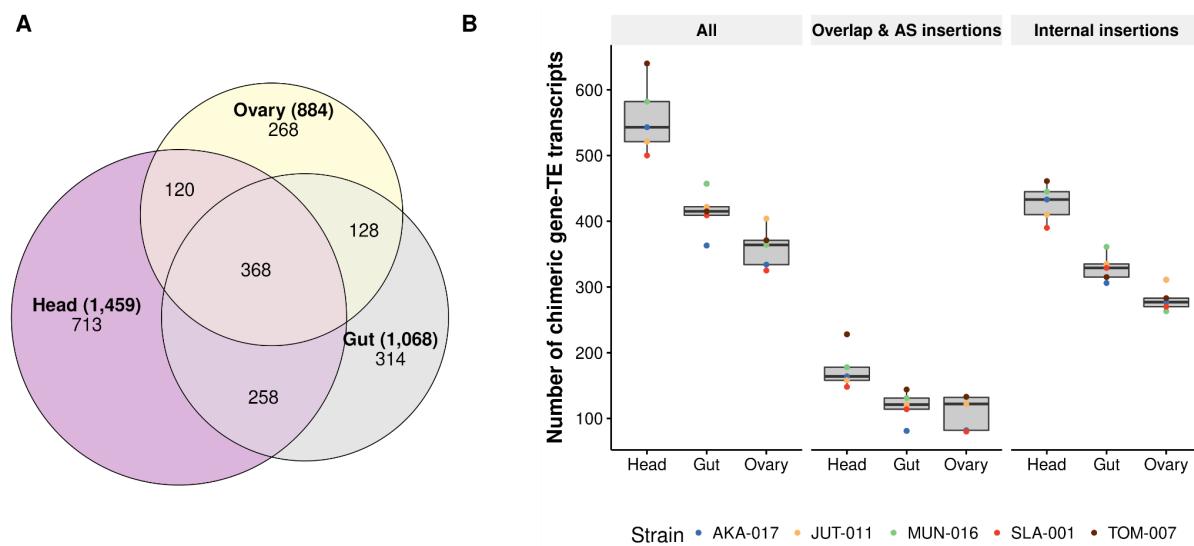
176 **Gene-TE chimeric transcripts are more abundant in the head**

177 Using high-throughput methodologies 833 chimeric genes were identified in the *D.*
178 *melanogaster* head²¹, however, the relative amount of chimeric gene-TE transcripts across
179 body parts has never been assessed before. We found that the majority of the assembled
180 chimeric gene-TE transcripts across the five strains analyzed were body part specific (60%:
181 1,295/2,169), with only 17% (368) shared across all three body parts (Figure 2A and Table
182 S3A). The same pattern was found for the *overlap and AS insertions* group and for the *internal*
183 *insertions* group, when considering all insertions and those ≥ 120 bp (Figure S2B and Table
184 S3A).

185 Head was the body part expressing the most chimeric transcripts (1,459) followed by gut
186 (1,068) and ovary (884) (Figure 2A and Table S3A). Note that 208 of the chimeric transcripts
187 identified in this work were previously described by Treiber and Waddell (2020)²¹. After
188 accounting for differences in the total number of transcripts assembled in each body part, we
189 still observed that the head was expressing more chimeric transcripts compared to gut and
190 ovary (8.54% head vs. 6.61% gut and 7% ovary; test of proportions, *p*-value = 3.89×10^{-11} and
191 2.14×10^{-7} , respectively; Table S3B). On the other hand, the proportion of total transcripts that
192 are chimeric was similar between gut and ovary (test of proportions, *p*-value = 0.337) (Table
193 S3C). A higher proportion of chimeric transcripts in head compared with gut and ovary was

194 also found when the *overlap and AS insertions* and the *internal insertions* groups were
195 analyzed separately, although in this last group the proportion across body parts is similar if
196 we focus on ≥ 120 bp insertions (Figure 2B and Table S3C). Overall, the same patterns were
197 also found at the strain level, except for JUT-011 and MUN-016, where some comparisons
198 were not significant (Table S3C).

199 Finally, the head was also the body part that expressed the most body part specific chimeric
200 transcripts (48% head vs. 29% gut; test of proportions, p -value < 0.001, and vs. 30% ovary,
201 p -value < 0.001), while no differences were found between gut and ovary (30% ovary vs. 29%
202 gut; test of proportions, p -value = 0.7; Figure 2A). In the three body parts, these proportions
203 were higher than the total proportion of body part specific transcripts (21.3%, 13.1% and 9.4%,
204 for head, gut and ovary respectively; test of proportions, p -values < 0.001 for all comparisons;
205 Table S3B).



206

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

211

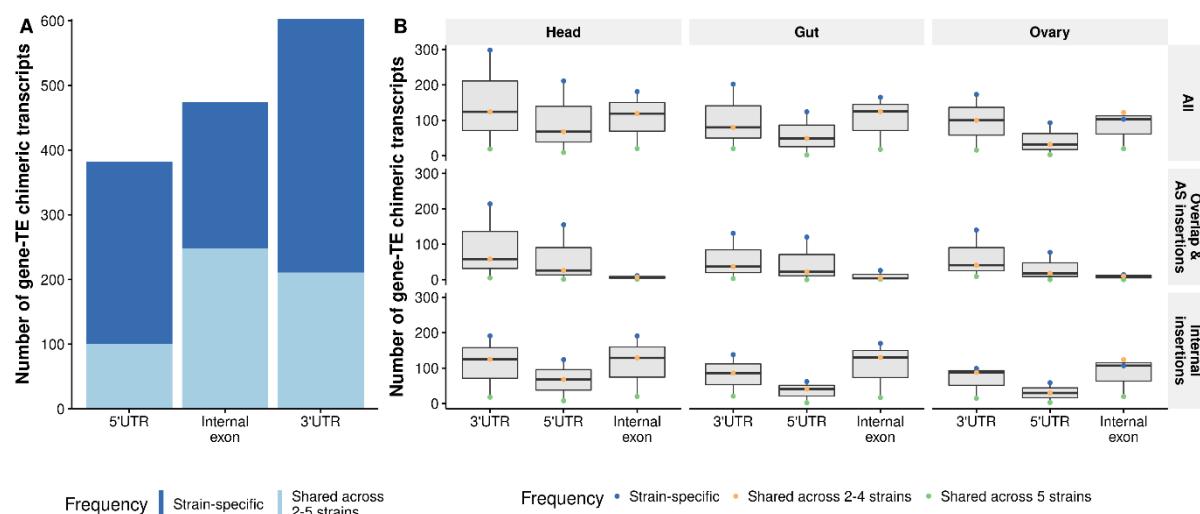
212 **Most chimeric transcripts contain TE insertions in the 3'UTRs**

213 Chimeric gene-TE transcripts are enriched for TE insertions located in the 3'UTRs in *D. melanogaster* and in mammals^{12,13,20}. Consistently, we also found that most of the chimeric
214 gene-TE transcripts contain a TE in the 3'UTR (1,084 transcripts from 662 genes) followed by
215 internal exons (924 transcripts from 529 genes) and insertions in the 5' UTRs (703 transcripts
216 from 411 genes).

217 from 499 genes). Note that 34 of the 5' UTR insertions detected in this work were
218 experimentally validated in a previous analysis that estimated the promoter TE usage across
219 developmental stages in *D. melanogaster*²³. Indeed, the number of chimeric genes with a TE
220 inserted in the 3' and 5' UTRs is higher than expected when taking into account the proportion
221 of the genome that is annotated as UTRs, while there is a depletion of TEs in internal exons
222 (test of proportions, *p*-value < 0.001 in the three comparisons; Table S4A). It has been
223 hypothesized that the higher number of insertions in 3' UTRs could be explained by lack of
224 selection against insertions in this gene compartment^{11,12}. We thus tested whether 3'UTR
225 chimeric transcripts were enriched for TE insertions present in more than one genome.
226 However, we found an enrichment of unique insertions in 3'UTR chimeric transcripts
227 suggesting that they might be under purifying selection (test of proportions, *p*-value = 0.033;
228 Figure 3A and Table S4B).

229 While in the *overlap and AS insertions* group, TE insertions were also mainly located in the 3'
230 UTRs (53.4%: 260/487), in the *internal insertions* group there were more chimeric transcripts
231 with TE insertions found in internal exons than in the 3'UTR (448 vs. 343; test of proportions,
232 *p*-value < 0.001). This pattern still holds when we only consider ≥ 120 bp insertions (166 vs.
233 125; test of proportions, *p*-value = 0.047; Table S4C). Figure 3B shows the number of chimeric
234 gene-TE transcripts globally and by insertion group, body part and strain (Table S4D) where
235 it can be observed that, overall, the previous patterns hold at the body part level.

236



237

238 **Figure 3. Position and frequency distribution of TEs in chimeric transcripts. A.** Number of gene-
239 TE chimeric transcripts by position and frequency. **B.** Number of chimeric gene-TE transcripts by
240 insertion group and body part, according to the insertion position (5'/3'UTRs or internal exons) and
241 frequency. Each dot represents the number of chimeric gene-TE transcripts according to the frequency:
242 strain-specific (blue), shared across two to four strains (orange) and shared across all five strains
243 (green). These analyses were performed with the subset of chimeric transcripts with only one TE
244 annotated in the same position across strains.

245

246 **Chimeric gene-TE transcripts are enriched for retrotransposon insertions**

247 We assessed the contribution of TE families to chimeric gene-TE transcripts. We found that
248 the majority of TE families, 111/146 (76%), were detected in chimeric gene-TE transcripts, as
249 has been previously described in head chimeric transcripts (Table S5A^{21,34}). Although
250 retrotransposons are more abundant than DNA transposons (61% on average in the five
251 genomes analyzed³⁴, the contribution of retrotransposons to the chimeric gene-TE transcripts
252 was higher than expected (81%: 90/111; test of proportions, *p*-value < 0.001; Table S5B).
253 There were slightly more families contributing to the *overlap and AS insertions* group than to
254 the *internal insertions* group (98 vs. 82, respectively, test of proportions, *p*-value = 0.01), but
255 both groups were enriched for retrotransposons (test of proportions, *p*-value < 0.001 and *p*-
256 value = 0.0179, respectively; Table S5C). More than half of these families (64: 57.7%)
257 contribute to chimeric transcripts in all body parts, while 24 families were body part-specific,
258 with 12 being head-specific, 6 gut-specific and 6 ovary-specific (Table S5A).

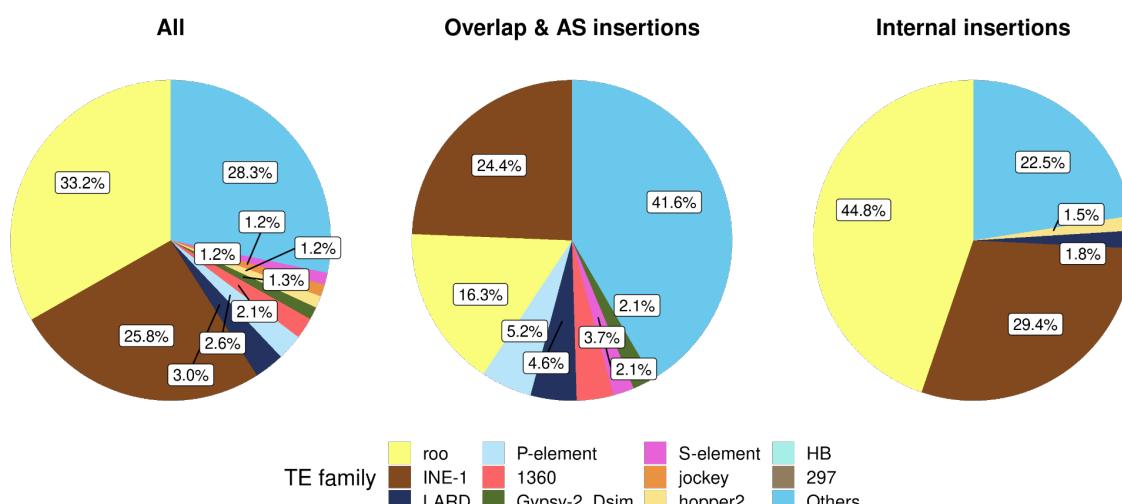
259 The most common TE families found were *roo* (33.2%) and *INE-1* (25.8%) (Figure 4). Indeed,
260 these two families were over-represented in the chimeric transcripts dataset when compared
261 to their abundance in the genome: *roo* in the five strains (test of proportions, *p*-value < 0.0001
262 for all comparisons) and *INE-1* in AKA-017 and SLA-001 (test of proportions, *p*-value = 0.004,
263 and *p*-value < 0.0001, respectively) (Table S5D). *Roo* and *INE-1* were also the most common
264 families both in the *overlap and AS insertions* group (16.3 and 24.4%, respectively) and in the
265 *internal insertions* group (44.8% and 29.4%, respectively). The same pattern was found when
266 we analyzed only those chimeric transcripts with TEs \geq 120bp (Figure S3 and Table S5E).

267 Because *roo* insertions were enriched in all the strains analyzed, we further investigate these
268 TE sequences. We found only two types of *roo* insertions: solo LTRs (23 insertions), that all
269 belong to the *overlap and AS insertions* group, and a short (45bp-217bp) low complexity
270 sequence mapping to the positions 1,052-1,166 of the canonical *roo* element (see *Methods*).
271 This short *roo* sequence is more common in the *internal insertions* group than in the *overlap*

272 and AS insertions group (911 vs. 61 insertions, respectively). Note that a recent analysis by
273 Oliveira et al. ⁴³ also found this same region of the *roo* consensus sequence to be the most
274 abundant in chimeric gene-TE transcripts across four *D. melanogaster* strains⁴³. The authors
275 evaluated whether these short sequences were widespread repeats across the genome. They
276 found that the majority of the *roo* fragments they identified (97.45%) have only one blast hit in
277 the genome, suggesting that they are not. We argued that if these low complexity regions have
278 a *roo* origin, we should find that at least some of them should also have a blast hit with a *roo*
279 insertion. To test this, we used less strict blast parameters compared with Oliveira et al.⁴³ and
280 found that 57 of the low complexity regions have a *roo* element insertion as the second best
281 hit and 148 have a *roo* insertion in the top 5 hits, suggesting that indeed some of these
282 sequences have a clear *roo* origin (Table S5F). Furthermore, we also tested whether this low
283 complexity region was present in the *roo* consensus sequence from a closely related species,
284 *D. simulans*, and found that this was the case strongly suggesting that this low complexity
285 sequence is an integral part of the *roo* element.

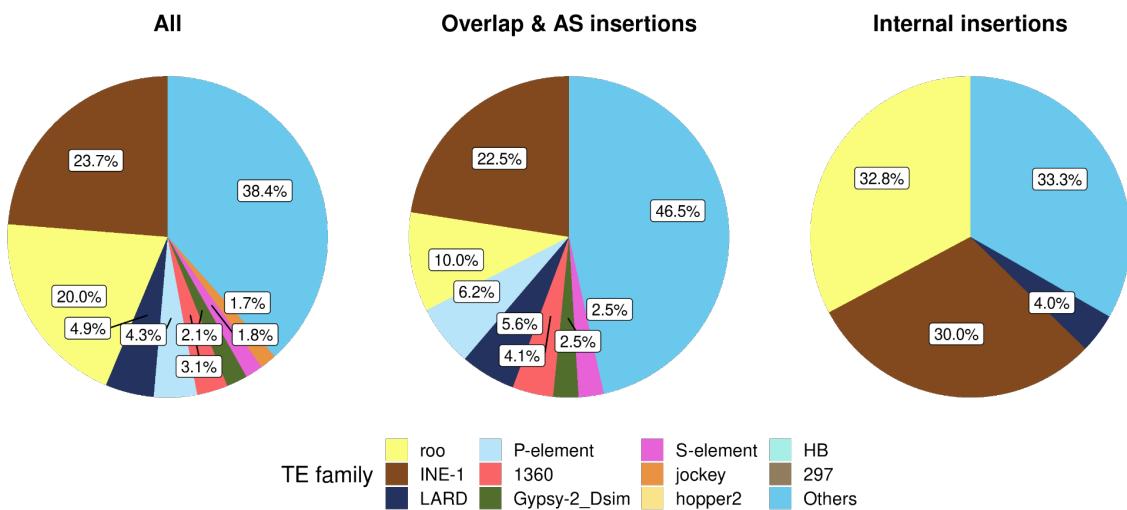
286 We further investigated why this *roo* low complexity region was incorporated into genes.
287 Because TEs can contain *cis*-regulatory DNA motifs, we performed a motif scan of the low
288 complexity sequence from the canonical *roo* element. We found a C2H2 zinc finger factor
289 motif repeated six times in this region. Note that this motif is only found once in the *roo*
290 consensus sequence outside the low complexity region. A scan in the *roo* sequences from the
291 chimeras revealed that 78% (753/972) of the transcripts with the low complexity *roo* sequence
292 contains at least one sequence of this zinc finger motif, with 26% (196/753) containing 3 or
293 more (Table S5G).

294



296 **Figure 4. TE families distribution in gene-TE chimeras, globally and by insertion group.**
297 Percentage of TE families contributing to gene-TE chimeras in the global dataset (*All*), in the *overlap*
298 and *AS insertions* group and in the *internal insertions* group. Only TE families found in more than 15
299 chimeric genes are depicted, otherwise they are grouped in *Others*.

300
301



303 **Supplementary Figure 3. TE families distribution in gene-TE chimeras, globally and by insertion**
304 **group considering insertions ≥ 120 bp.** Percentage of TE families contributing to gene-TE chimeras
305 considering insertions ≥ 120 bp in the global dataset (*All*), in the *overlap and AS insertions* group and in
306 the *internal insertions* group. Only TE families found in more than 15 chimeric genes are depicted,
307 otherwise they are grouped in *Others*.

308

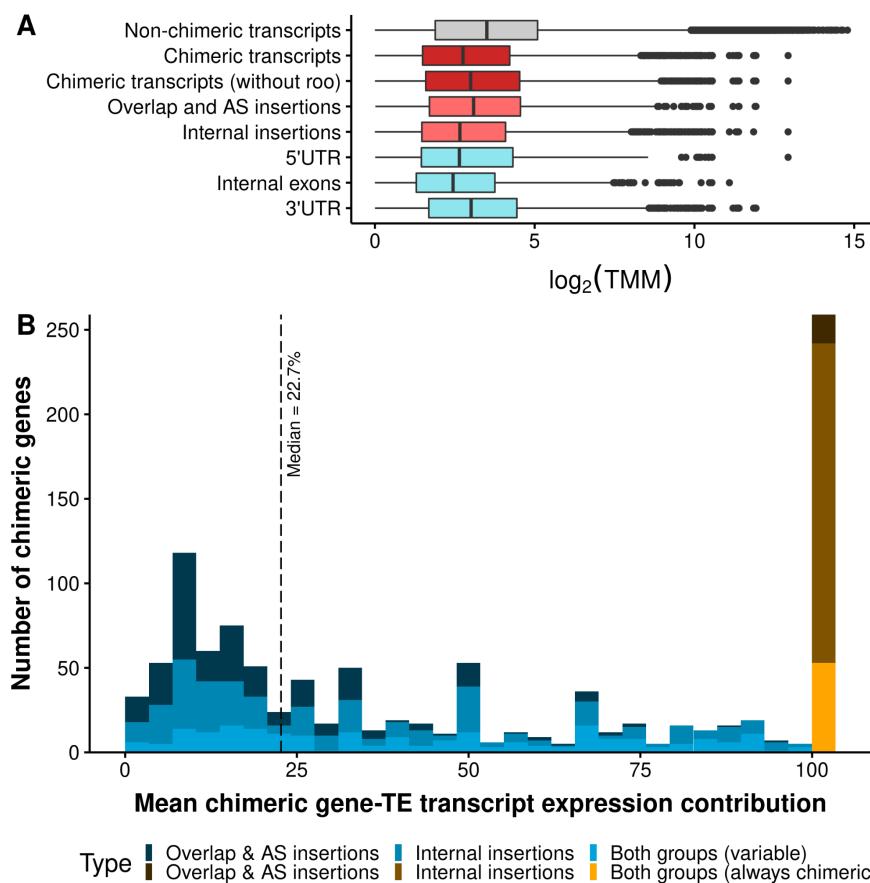
309 **Chimeric gene-TE transcripts contribute a median of ~38% of the total gene expression**

310 Besides identifying and characterizing chimeric gene-TE transcripts, we quantified the level of
311 expression of both chimeric and non-chimeric transcripts genome-wide. We focused on
312 transcripts with ≥ 1 TMM in at least one of the samples analyzed (1,779 out of 2,169 chimeric
313 transcripts, corresponding to 86% (1,074/1,250) of the genes (see *Methods*). We found that
314 chimeric gene-TE transcripts have lower expression levels than non-chimeric transcripts
315 (17,777; Wilcoxon's test, *p*-value < 0.001, Figure 5A). This is in contrast with previous
316 observations in human pluripotent stem cells that reported no differences in expression
317 between chimeric and non-chimeric transcripts²⁰. We dismissed the possibility that the lower
318 expression of chimeric gene-TE transcripts was driven by the *roo* low complexity region
319 identified in 995 of the chimeric transcripts (Wilcoxon's test, *p*-value < 0.0001; Figure 5A).
320 Lower expression of the chimeric gene-TE transcripts was also found at the body part and

321 strain levels and when we analyzed the *overlap and AS insertions* and *internal insertions*
322 groups separately (Wilcoxon's test, p -value < 0.001 for all comparisons; Figure 5A and Table
323 S6A).

324 We further tested whether TEs inserted in different gene locations differed in their levels of
325 expression compared with the non-chimeric TE transcripts. We found that chimeric transcripts
326 had significantly lower expression than non-chimeric transcripts regardless of the insertion
327 position (Wilcoxon's test p -value < 0.001 for all comparisons; Figure 5A). Furthermore,
328 insertions in the 3'UTR appeared to be more tolerated than those in 5'UTR and internal exons,
329 as their expression level was higher (Wilcoxon's test, p -value < 0.005 for both comparisons;
330 Figure 5A). Our results are consistent with those reported by Faulkner et al.²² who also found
331 that 3'UTR insertions reduced gene expression.

332 If we focus on the chimeric genes, 24% of them (259 genes) only expressed the chimeric
333 gene-TE transcript (in all the genomes and body parts where expression was detected). Most
334 of these genes (70%) contain short TE insertions and accordingly most of them belong to the
335 *internal insertions* group (93%) (test of proportions, p -value < 0.001). For the other 76% (815)
336 of the genes, we calculated the average contribution of the chimeric gene-TE transcript to the
337 total gene expression per sample. While some genes contributed only ~4% of the total gene
338 expression, others accounted for >90% (median = 22.7%) (Figure 5B). The median
339 contribution to gene expression of the *internal insertions* group is higher than that of the
340 *overlap and AS insertions* group, when considering all the insertions (25% vs. 14.3%,
341 respectively; Wilcoxon's test, p -value < 0.001), and when analyzing only those transcripts with
342 ≥ 120 bp insertions (20% vs. 14.29%, respectively; Wilcoxon's test, p -value = 0.0015).
343 Considering only the transcripts that do not contain the *roo* low complexity sequence, the
344 median contribution to gene expression of the *internal insertions* group was still 20%. Overall,
345 taking all chimeric genes into account (1,074), the median of the chimeric gene-TE transcripts'
346 expression contribution to the total gene expression was 38%.



348 **Figure 5. TE insertions within genes affect gene expression. A.** Boxplots for the expression levels,
 349 measured as the logarithm of TMM: for all non-chimeric transcripts of the genome (17,777, in gray), all
 350 chimeric transcripts detected in the present study with $\text{TMM} \geq 1$ (1,779, in dark red), chimeric transcripts
 351 without the short internal *roo* insertion (963, dark red), all chimeric transcripts belonging to the *overlap*
 352 and *AS insertions* group (758, light red) and *internal insertions* group (1,302, light red), and chimeric
 353 transcripts divided by position of the insertion (5'UTR: 546, internal exons: 741, 3'UTR: 906, cadet blue).
 354 **B.** Histogram showing the expression contribution of chimeric transcripts to the total gene expression.
 355 Blue bars represent the contribution of variable chimeric genes (815 genes), ranging from ~4% to >90%
 356 (median: 22.7%) and the orange/brown bar represents the genes that always produced chimeric
 357 transcripts in all the genomes and body parts where expression was detected (259 genes).

358 Finally, we evaluated whether there are differences between the expression levels of body
 359 part-specific and body part-shared chimeric transcripts. The breadth of expression, measured
 360 as the number of tissues in which a gene is expressed, is significantly and positively correlated
 361 with the level of expression in *Drosophila*⁴⁴ and humans⁴⁵. Consistent with this, we found that
 362 body part-shared chimeric transcripts have significantly higher expression levels than chimeric
 363 transcripts expressed in only one body part (Wilcoxon's test, p -value < 0.001; Table S6B),
 364 when considering the whole dataset and for chimeric transcripts with insertions $\geq 120\text{bp}$
 365 (Wilcoxon's test, p -value < 0.001; Table S6B). Since we observed that the head was

366 expressing more chimeric transcripts (Figure 2A), we next assessed if head-specific chimeric
367 transcripts were also expressed at higher levels. We observed that the median expression of
368 head-specific chimeric transcripts was higher than those specific of gut (median_{head}= 5.18
369 TMM [$n = 527$], median_{gut}= 3.8 TMM [$n = 205$]; Wilcoxon's test, p -value = 0.0021), but lower
370 than ovary-specific chimeric transcripts (median_{ovary}= 8.52 TMM [$n = 210$]; Wilcoxon's test, p -
371 value = 1.35×10^{-5}). However, this is similar to the expression level of genes in these tissues
372 (median of gene expression in ovary>head>gut: 20.2>9.7>8.5).

373 Interestingly, strain-shared chimeric transcripts (expressed in the five strains) also have
374 significantly higher expression levels than strain-specific chimeric transcripts (Wilcoxon's test,
375 p -value < 0.001; Table S6C).

376

377 **11.4% of the TEs within chimeric gene-TE transcripts could also be affecting gene
378 expression via epigenetic changes**

379 We tested whether TEs that are part of chimeric transcripts could also be affecting gene
380 expression by affecting the epigenetic marks. We used ChIP-seq experiments previously
381 performed in our lab for the three body parts in each of the five strains analyzed for two histone
382 marks: the silencing mark H3K9me3^{46,47} and H3K27ac, related to active promoters and
383 enhancers^{48,49}. We focused on polymorphic TEs because for these insertions we can test
384 whether strains with and without the insertion differed in the epigenetic marks (755 genes).
385 For the majority of these genes (534), we did not observe consistent epigenetic patterns
386 across samples with and without the TE insertion, and these genes were not further analyzed.
387 Additionally, 86 genes did not harbor any epigenetic marks while 49 genes contained the same
388 epigenetics mark(s) (H3K27ac, H3K9me3, or both marks) in strains with and without that
389 particular TE insertion (Table S7). Overall, only for 11.4% (86/755) of the genes, we observed
390 a consistent change in the epigenetic status associated with the presence of the TE. This
391 percentage is similar for the *overlap and AS* group and the internal insertion group (10.4% and
392 11.8%, respectively). The majority of TEs showing consistent changes in their epigenetic
393 status were associated with gene down-regulation (50/86; Table 1). While 70% (534/ 755) of
394 the genes analyzed were expressed in the head, only 57% (49/86) differed in their epigenetic
395 marks (test of proportions, p -value = 0.03).

396 **Table 1. Expression changes associated with epigenetic status of strains with and without the**
397 **TE insertion.** Highlighted in bold, genes showing the expected change in expression according to the
398 gained histone mark.

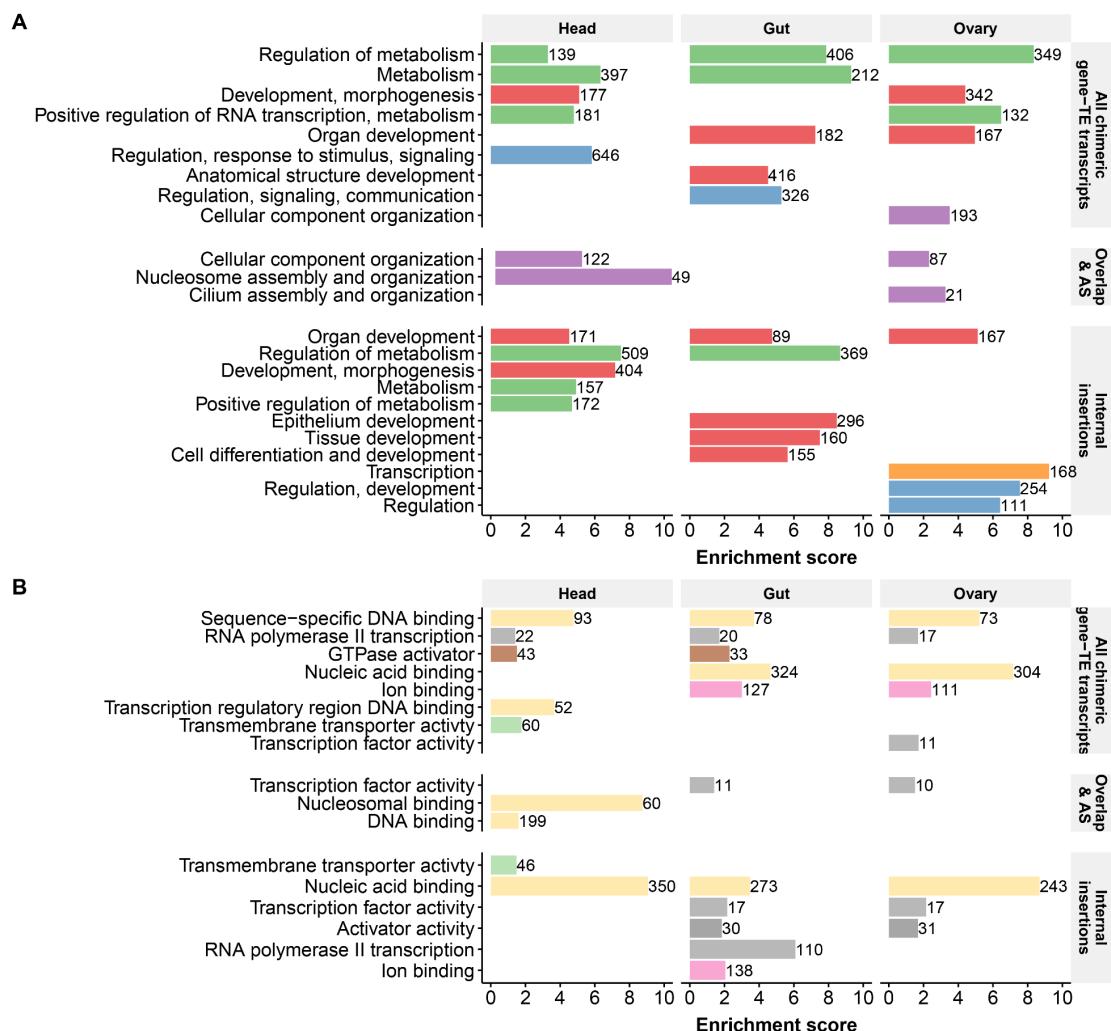
Fold change	Gain of H3K27ac	Gain of H3K9me3	Gain of both marks	Loss of H3K27ac	Loss of H3K9me3
FC > 1	15	5	13	1	2
FC < 1	26	9	14	1	0

399

400 **Gene-TE chimeric transcripts are enriched for DNA binding molecular functions**
401 **involved in metabolism and its regulation, and development**

402 To get insight on the biological processes and molecular functions in which the gene-TE
403 chimeric transcripts are involved, we performed a gene ontology (GO) clustering analysis⁵⁰.
404 We analyzed the chimeric genes detected in each body part separately, using as a
405 background the total genes assembled in the corresponding body part. We found that chimeric
406 genes are enriched in general cell functions, such as metabolism and its regulation, and
407 development (Figure 6A and Table S8A). Some functions are particular to a body part, e.g.
408 *response to stimulus and signaling* in the head, *anatomical structure development and*
409 *regulation, and signaling and communication* in the gut, and *cellular component organization*
410 in the ovary. Note that the *overlap and AS insertions* group is enriched for *cellular component*
411 *organization, and nucleosome and cilium assembly and organization*, across tissues (Figure
412 6A and Table S8C).

413 Finally, regarding the molecular function, chimeric genes are enriched for DNA binding
414 processes and *RNA polymerase II transcription* across body parts (Figure 6B and Table S8B),
415 while in head they are also enriched for *transmembrane transporter activity* and in ovary for
416 *transcription factor activity*.



417

418 **Figure 6. Biological processes and molecular functions of chimeric gene-TE transcripts. A.**
419 Biological processes clustering. **B.** Molecular functions clustering. The length of the bars represents the
420 cluster enrichment score. The number in the bars represents the number of genes in each cluster.
421 Names of the annotation clusters are manually processed based on the cluster's GO terms. Colors
422 represent similar annotation clusters. Detailed GO terms of each cluster are given in Table S8.

423

424 Both DNA transposons and retrotransposons add functional protein domains

425 We next assessed whether TE sequences annotated in internal exons provided functional
426 domains. We first confirmed, using the Coding Potential Assessment Tool (CPAT⁵¹) software,
427 that the majority of chimeric protein-coding gene-TE transcripts that have a TE annotated in
428 an internal exon have coding potential (95.12%: 858/902; Table S9A). Using PFAM⁵², we
429 identified a total of 27 PFAM domains in 36 different chimeric transcripts from 29 genes (Table
430 2 and Table S9B). These 27 domains were identified in 24 TE families, with 16 TE families
431 providing more than one domain. The size of these domains ranged from 9bp to 610bp (mean

432 of 123.5bp; Table S9B). Note that 10 of these 29 chimeric genes have been previously
433 described in the literature (Table 2). Most of the transcripts (67%: 24/36) belong to the *overlap*
434 and *AS insertions* group. Finally, we found chimeric transcripts adding domains in the three
435 tissues analyzed (Table 2), with an enrichment in ovary compared to head (test of proportions,
436 *p*-value = 0.027).

437 The majority of TEs adding domains were retrotransposons (22/29) and most TEs provided a
438 nearly-full domain (24/29, $\geq 50\%$ coverage), including 9 TEs adding a full-size domain (Table
439 2). Almost 30% (9/29) of the chimeric genes are related to gene expression functions and 20%
440 (5/29) are related to cell organization and biogenesis (Table S9C). The majority of these
441 chimeric genes (21/29) have evidence of expression, ranging from 1.05 to 47.14 TMM (Table
442 2, median = 8.26 TMM). The median expression was higher for the transcripts with complete
443 domains compared to partially/uncompleted domains (median TMM 22.16 vs. 9.03), although
444 the difference was not statistically significant (Wilcoxon's test, *p*-value = 0.08). The majority of
445 TEs for which the population TE frequency has been reported, are fixed or present at high
446 frequencies (12/22 TEs; Table 2).

447 We assessed if the domains detected in the TE fragment of the gene-TE chimera were also
448 found in the consensus sequence of the TE family. Because most TE families were providing
449 more than one domain, in total we analyzed 54 unique domains. We were able to find the
450 domain sequence for 50 unique domains from 20 TE consensus sequences (Table S9D). Note
451 that for five of these domains (from four TEs), we had to lower PFAM detection thresholds to
452 detect them (see Methods). The four domains that were not identified in the consensus
453 sequences, were smaller than the average (ranging 18bp-101bp, mean: 62.25bp) and were
454 not detected in the chimeric fragments as full domain sequences.

455 A PFAM domain enrichment analysis considering domains annotated with nearly-full domains
456 and in transcripts expressed with minimum of 1 TMM using dcGO⁵³, found enrichment of the
457 molecular function *nucleic acid binding* (6 domains, FDR = 4.12×10^{-4}) and *catalytic activity,*
458 *acting on RNA* (4 domains, FDR = 4.12×10^{-4}) (Table 3). All the enriched domains are found in
459 retrotransposon insertions. Consistent with the enrichment of the molecular functions, these
460 domains were enriched in the *nuclear body* and in *regulation of mRNA metabolic process*
461 (Table 3).

462 **Table 2. Description of the 29 chimeric genes containing a TE providing a protein domain.**

463 NA in the *splicing motifs* column represents cases in which there are not splicing signals because the
 464 TE was found inside an exon (*internal insertion* group) while NC stands for non-canonical splicing motif.
 465 TMM is the expression level and it is the average if more than one transcript was detected across body
 466 parts or strains. TE frequency (*Freq.*) was retrieved from Rech et al.³⁴. Superscript numbers in the *gene*
 467 column represent literature describing these chimeric genes: [1]²¹, [2]¹², [3]⁵⁴, [4]⁵⁵, [5]⁵⁶, [6]⁵⁷, [7]
 468³², [8]⁵⁸.

Gene	TE class: family	PFAM domains (%coverage)	TMM	Splicing motifs	Freq.	Body parts
<i>CHKov1^{1,2,3,4}</i>	RNA: Doc	Exo_endo_phos_2 (98.3%), RVT_1 (100%)	16.77	NA	0.85	Head, Ovary, Gut
<i>nxf2⁵</i>	RNA: TART-A	TAP_C (89.8%)	15.55	NA	1.00	Gut, Ovary
<i>Pld¹</i>	RNA: I-element	RNase_H (21%)	5.27	NC	1.00	Gut
<i>Smn</i>	RNA: TART-A	TAP_C (89.8%)	6.04	NA	1.00	Ovary, Gut
<i>Brf</i>	RNA: jockey	Exo_endo_phos_2 (99.2%), PRE_C2HC (98.5%), RVT_1 (100%)	4.63	AG/GT	0.02	Gut, Head
<i>Dbp45A</i>	RNA: Doc6	Exo_endo_phos_2 (84%), RVT_1 (98.7%)	0	NC/GT	0.04	Ovary
<i>Fer2LCH</i>	RNA: blood	Integrase_H2C2 (96.6%), RT_RNaseH_2 (100%), RVP (86.1%), RVT_1 (99.1%)	3.88	NC	0.04	Ovary
<i>smg</i>	RNA: rover	Baculo_F (23.4%), Integrase_H2C2 (87.9%), RT_RNaseH (99.1%), RVT_1 (98.7%)	0	NC	0.04	Ovary
<i>eIF4B</i>	RNA: Invader2	rve (78.4%)	26.99	AG/GT	NA	Gut
<i>CG7465</i>	RNA: NewFam16	GYR (98.6%), YLP (92.6%)	4.58	AG/GT	1.00	Gut
<i>CG7582</i>	RNA: jockey	PRE_C2HC (98.5%)	0	AG/GT	0.02	Head
<i>CG17883¹</i>	RNA: Quasimodo	Integrase_H2C2 (87.9%), RT_RNaseH (100%), RVT_1 (99.6%)	29	AG/GT	NA	Ovary
<i>Prat2</i>	RNA: Gypsy-2_Dsim	Integrase_H2C2 (100%), RT_RNaseH (100%), RVT_1 (100%), rve (93.1%)	0	AG/NC	0.02	Gut
<i>CG32032</i>	RNA: jockey	PRE_C2HC (98.5%)	28.14	AG/NC	0.06	Head
<i>Nlg1²</i>	RNA: Invader3	SAP (88.6%), zf-CCHC (83.3%)	0.58	AG/GT	0.17	Head
<i>CG33178</i>	RNA: mdg3	zf-CCHC (88.9%)	3.36	NC	0.02	Head
<i>stw¹</i>	RNA: F-element	Exo_endo_phos (100%), RVT_1 (100%)	3.49	AG/GT	NA	Ovary
<i>I(3)80Fj</i>	RNA: Cr1a	RVT_1 (39.2%)	29.89	AG/NC	NA	Head
<i>I(3)80Fg</i>	RNA: gypsy8	RT_RNaseH (87.6%)	8.26	AG/NC	NA	Gut
<i>Mctp</i>	RNA: Burdock	Integrase_H2C2 (94.8%), RT_RNaseH (98.1%), RVT_1 (98.7%)	1.05	NA	0.02	Gut
<i>CG2162^{1,2}</i>	RNA: diver	DUF1758 (93.3%), DUF1759 (96%), Integrase_H2C2 (86.2%), Peptidase_A17 (80.9%)	47.14	NA	0.02	Ovary
<i>pps</i>	RNA: Transpac	Integrase_H2C2 (94.8%), RT_RNaseH (100%), RVT_1 (99.1%)	22.94	NA	0.02	Ovary
<i>Gmd</i>	DNA: S-element	HTH_Tnp_Tc3_2 (50%)	15.04	NA	0.98	Ovary, Head

<i>Ppcs</i> ¹	DNA: Bari1	DDE_3 (89.7%), HTH_28 (98.1%), HTH_Tnp_Tc3_2 (100%)	40.96	NA	1.00	Gut
<i>CG2225</i> ¹	DNA: HB	HTH_Tnp_Tc3_2 (84.7%)	0.44	AG/GT	NA	Gut
<i>CG1671</i>	DNA: pogo	DDE_1 (98.9%), HTH_23 (80%), HTH_Tnp_Tc5 (95.5%)	0	NC	0.11	Head
<i>Cyp12a4</i> ^{1,6,7,8}	DNA: Bari1	DDE_3 (89.0%), HTH_28 (98.1%), HTH_Tnp_Tc3_2 (100%)	0.29	AG/GT	1.00	Ovary
<i>Idh3b</i>	DNA: P-element	THAP (90.7%), Tnp_P_element (38.4%)	1.46	AG/GT	0.11	Ovary
<i>CG14043</i>	DNA: S-element	HTH_Tnp_Tc3_2 (50%)	1.28	NA	0.98	Ovary

469

470

471

Table 3. PFAM domain enrichment analysis

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

GO term	Z-score	FDR	Annotated domains
Molecular function			
Nucleic acid binding	4.62	4.12×10^{-4}	PF00098 (zf-CCHC); PF00665 (rve); PF02037 (SAP); PF03372 (Exo_endo_phos); PF03943 (TAP_C); PF05485 (THAP)
Catalytic activity, acting on RNA	5.99	4.12×10^{-4}	PF00078 (RVT_1); PF00098 (zf-CCHC); PF00665 (rve); PF03372 (Exo_endo_phos)
Cellular component			
Nuclear body	7.61	1.11×10^{-3}	PF02037 (SAP); PF03372 (Exo_endo_phos); PF03943 (TAP_C)
Biological process			
Regulation of mRNA metabolic process	9.26	1.57×10^{-3}	PF00098 (zf-CCHC); PF02037 (SAP); PF03372 (Exo_endo_phos)

473

474

475

DISCUSSION

476 TEs contribute to genome innovation by expanding gene regulation, both of individual genes
 477 and of gene regulatory networks, enriching transcript diversity, and providing protein domains
 478 (e.g., reviewed in Chuong et al.⁵⁹ and Modzelewski et al.⁶⁰). While the role of TEs as providers
 479 of regulatory sequences has been extensively studied, their contribution to transcriptome
 480 diversification and protein domain evolution has been less characterized. In this work, we have
 481 identified and characterized chimeric gene-TE transcripts across three body parts and five
 482 natural *D. melanogaster* strains, and we have quantified their contribution to total gene
 483 expression and to protein domains. While previous studies were hindered by the incomplete
 484 annotation of TEs in the genome studied^{12,21}, in this work, we took advantage of the availability
 485 of high-quality genome assemblies and genome annotations for five natural strains to carry
 486 out an in depth analysis of gene-TE chimeric transcripts³⁴. We found that TEs contribute 10%

487 to the global transcriptome and 18% to the body part specific transcriptome (Figure 1).
488 Contrary to other studies that mostly focus on a single type of chimeric gene-TE transcript, we
489 investigated a comprehensive dataset of chimeras. Indeed, we found that besides insertions
490 affecting the transcription start site, transcript termination, and adding spliced sites (*overlap*
491 and *alternative splicing insertions*), we also identified a substantial number of TE sequences
492 that were completely embedded within exons (*internal insertions*; Figure 1D). These two types
493 of chimeric gene-TE transcripts shared many properties, e.g. they were enriched for body part
494 specific transcripts and for retrotransposons (Figure S2B and Figure 4), and they showed
495 lower expression levels than non-chimeric transcripts (Figure 5A), suggesting that they both
496 should be taken into account when analyzing the contribution of TEs to gene novelty. The
497 *internal insertions* group contributed more to total gene expression (Figure 5B), however, we
498 dismissed the possibility that this increased expression was due to shorter TE insertions,
499 which are more likely to be enriched for false annotations compared with longer insertions³⁴.
500 We found, both based on size and frequency, that the *internal insertions* group is likely to be
501 enriched for older insertions. As such, a higher level of expression of these likely older TEs is
502 consistent with previous observations in tetrapods suggesting that over time gene-TE chimeric
503 transcripts often become the primary or sole transcript for a gene²⁷. Overall, and taking only
504 into account those gene-TE chimeric transcripts with evidence of expression, we found 155
505 (8.6%) insertions disrupting the coding capacity, 415 (22.9%) affecting the coding capacity,
506 314 (17.3%) and 591 (32.6%) affecting the 5' and the 3' end of the gene, respectively, while
507 338 (18.6%) affected multiple transcript positions.

508 Our finding that TEs contribute to the expansion of the head transcriptome supports the results
509 of Treiber and Wadell (2020)²¹ suggesting that ~6% of genes produce chimeric transcripts in
510 the head due to exonization of a TE insertion. However, because we also analyzed gut and
511 ovary, we further show that TEs can significantly contribute to the expansion of other body
512 parts transcriptomes as well (Figure 2). The observation that there are more chimeric
513 transcripts in the head is consistent with a higher transcriptional complexity in the *Drosophila*
514 nervous system tissues³. The fact that chimeric gene-TE transcripts tend to be tissue-specific
515 could be especially relevant for adaptive evolution as tissue-specific genes can free the host
516 from pleiotropic constraints and allow the exploration of new gene functions^{45,61,62}.

517 Finally, we identified a total of 27 TE protein domains co-opted by 29 genes (Table 2 and
518 Table S24). Ten of these genes have been previously described as chimeric based on high-
519 throughput screenings or individual gene studies, with some of them, e.g. *CHKov1* and *nxf2*,
520 having functional effects^{54–56} (Table 2). The majority of the domains were present in the TE
521 consensus sequences (Table S9D). Furthermore, the 27 domains identified were enriched for

522 *nucleic acid binding and catalytic activity, acting on RNA* molecular functions (Table 3).
523 Although there is evidence for DNA binding domains being recruited to generate new genes,
524 previous data comes from a comparative genomic approach across tetrapod genomes that
525 focused on DNA transposons as a source of new protein domains²⁷. The available data for the
526 genome-wide contribution of retrotransposons to protein domains so far is restricted to
527 endogenous retroviruses in mammals²⁸. In our dataset, that includes both DNA transposons
528 and retrotransposons, the enrichment for DNA binding domains and for catalytic activity is
529 indeed driven by the retrotransposon insertions (Table 2). Although most of the TEs providing
530 protein domains identified in this work for the first time were present at low population
531 frequencies, four were fixed and two present at high population frequencies and are thus good
532 candidates for follow-up functional analysis (Table 2).

533 Although we have detected more chimeric transcripts than any prior *D. melanogaster* study to
534 date, our estimate of the potential contribution of TEs to the diversification of the transcriptome
535 is likely to be an underestimate. First, and as expected, we found that the contribution of TEs
536 to the transcriptome is body part specific^{22,30} (60%, Figure S2B) and strain-specific³⁴ (48%
537 Figure S2A). Thus analyzing other body parts and increasing the number of genomes
538 analyzed will likely identify more chimeric gene-TE transcripts. And second, although our
539 estimate is based on the highly accurate annotations of TE insertions performed using the
540 REPET pipeline³⁴, highly diverged and fragmented TE insertions are difficult to be accurately
541 annotated by any pipeline and as such might go undetected^{63,64}. Still, the combination of an
542 accurate annotation of chimeric gene-TE transcripts, with expression data across tissues, and
543 the investigation of protein domain acquisition carry out in this work, not only significantly
544 advances our knowledge on the role of TEs in gene expression and protein novelty, but also
545 provides a rich resource for follow-up analysis of gene-TE chimeras.

546

547

548 MATERIAL AND METHODS

549 Fly stocks

550 Five *D. melanogaster* strains obtained from the European Drosophila Population Genomics
551 Consortium (DrosEU), were selected according to their different geographical origins: AKA-
552 017 (Akaa, Finland), JUT-011 (Jutland, Denmark), MUN-016 (Munich, Germany), SLA-001
553 (Slankamen, Serbia) and TOM-007 (Tomelloso, Spain).

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

555 RNA-seq and ChIP-seq data for the five strains were obtained from ³⁶. A full description of the
556 protocols used to generate the data can be found in ³⁶. Briefly, head, gut and ovary body parts
557 of each strain were dissected at the same time. Three replicates of 30 4-6 old-day females
558 each were processed per body part and strain. RNA-seq library preparation was performed
559 using the TruSeq Stranded mRNA Sample Prep kit from Illumina, and sequenced using
560 Illumina 125bp paired-end reads (26.4M-68.8M reads; Table S1). For ChIP-seq, libraries were
561 performed using TruSeq ChIP Library Preparation Kit. Sequencing was carried out in a
562 Illumina HiSeq 2500 platform, generating 50bp single-end reads (22.2M-59.1M reads; Table
563 S1).

564 **Transcriptome assembly**

565 **Reference-guided transcriptome assembly**

566 To perform reference-guided transcriptome assemblies for each body part and strain (15
567 samples), we followed the protocol described in Pertea et al.⁴⁰ using HISAT2³⁹ (v2.2.1) and
568 StringTie⁴⁰ (v2.1.2). We used *D. melanogaster* r6.31 reference gene annotations⁶⁵ (available
569 at: ftp://ftp.flybase.net/releases/FB2019_06/dmel_r6.31/gtf/dmel-all-r6.31.gtf.gz, last
570 accessed: October 2020). We first used *extract_splice_sites.py* and *extract_exons.py* python
571 scripts, included in the HISAT2 package, to extract the splice sites and exon information from
572 the gene annotation file. Next, we build the HISAT2 index using *hisat2-build* (argument: *-p 12*)
573 providing the splice sites and exon information obtained in the previous step in the *-ss* and *-exon*
574 arguments, respectively. We performed the mapping of the RNA-seq reads (from the
575 fastq files, previously analyzed with FastQC⁶⁶) with HISAT2 (using the command *hisat2 -p 12*
576 *--dta -x*). The output sam files were sorted and transformed into bam files using samtools⁶⁷
577 (v1.6). Finally, we used StringTie for the assembly of transcripts. We used the optimized
578 parameters for *D. melanogaster* provided in⁶⁸ to perform an accurate transcriptome assembly:
579 *stringtie -c 1.5 -g 51 -f 0.016 -j 2 -a 15 -M 0.95*. Finally, *stringtie --merge* was used to join all
580 the annotation files generated for each body part and strain. We used *gffcompare* (v0.11.2)
581 from the StringTie package to compare the generated assembly with the reference *D.*
582 *melanogaster* r.6.31 annotation, and the sensitivity and precision at the locus level was 99.7
583 and 98.5, respectively.

584 **De novo transcriptome assembly**

585 A *de novo* transcriptome assembly was performed using Trinity³⁸ (v2.11.0 with the following
586 parameters: *--seqType fq --samples_file <txt file with fastq directory> --CPU 12 --*

587 *max_memory 78 G --trimmomatic*. To keep reliable near full-length transcripts, we used
588 *blastn*⁶⁹ (v2.2.31) to assign each *de novo* transcript to a known *D. melanogaster* transcript
589 obtained from the *Reference-guided transcriptome assembly*. Next, the script
590 *analyze_blastPlus_topHit_coverage.pl* from Trinity toolkit was used to evaluate the quality of
591 the BLAST results, and we followed a conservative approach that only kept a transcript with a
592 coverage higher than 80% with a known *D. melanogaster* transcript, thus, keeping 144,099
593 transcripts across all samples.

594 **Identification and characterization of chimeric gene-TE transcripts**

595 We focused on the set of assembled *de novo* transcripts that passed the coverage filtering to
596 identify putative chimeric gene-TE transcripts. We tried to minimize the possible sources of
597 confounding errors by excluding transcripts that were not overlapping a known transcript
598 (tagged by StringTie as *possible polymerase run-on* or *intergenic*). To annotate TEs in the *de*
599 *novo* assembled transcripts, we used RepeatMasker⁴¹ (v4.1.1⁴¹ with parameters *-norma* -
600 *nolow* -*s* -*cutoff* 250 -*xsmall* -*no_is* -*gff* with a manually curated TE library³⁴. Note that
601 RepeatMasker states that a cutoff of 250 will guarantee no false positives⁴¹. We excluded
602 transcripts for which the entire sequence corresponded to a transposable element, indicative
603 of the autonomous expression of a TE. To infer the exon-intron boundaries of the transcript,
604 we used minimap2⁴² (v2.17⁴² with arguments *-ax splice* --*secondary=no* --*sam-hit-only* -C5 -
605 *t4* to align the transcript to the genome of the corresponding strain from which it was
606 assembled. We excluded single-transcript unit transcripts, that could be indicative of pervasive
607 transcription or non-mature mRNAs. With this process, we obtained the full-length transcript
608 from the genome sequence.

609 We ran RepeatMasker again (same parameters) on the full-length transcripts to annotate the
610 full TEs and obtain the length of the insertion. Finally, we used an *ad-hoc* bash script to define
611 the TE position within the transcript and define the two insertions groups: the *overlap and AS*
612 *insertions* group and the *internal insertions* group. The *overlap and AS insertions* group have
613 a TE overlapping with the first (5'UTR) or last (3'UTR) exon, or overlap with the exon-intron
614 junction and thus introduce alternative splice sites (see *Splice sites motif scan analysis*). The
615 *internal insertions group* corresponds to TE fragments detected inside exons.

616 **TE insertion length**

617 As mentioned above, for each chimeric gene-TE transcript, we obtained the length of the TE
618 insertion from the TE annotation in the full-length transcript. We considered that short
619 insertions are those shorter than 120bp³⁴.

620 **Splice sites motif scan analysis**

621 We followed Treiber and Waddell (2020)²¹ approach to detect the splice acceptors and splice
622 donor sites in the *alternative splice (AS) insertions* subgroup of chimeric gene-TE transcripts.
623 In brief, we randomly extracted 11-12bp of 500 known donor and acceptor splice sites from
624 the reference *D. melanogaster* r.6.31 genome. Using the MEME tool⁷⁰ (v5.3.0), we screened
625 for the donor and acceptor motifs in these two sequences, using default parameters. The
626 obtained motifs were then searched in the predicted transposon-intron breakpoints position of
627 our transcripts using FIMO⁷¹ (v5.3.0 with a significant *p*-value threshold of < 0.05).

628 **Roo analyses**

629 **Identification of the position of the *roo* sequences incorporated into gene-TE chimeric**
630 **transcripts in the *roo* consensus.** To determine the position of the *roo* insertions, we
631 downloaded the *roo* consensus sequence from FlyBase⁶⁵ (version FB2015_02, available at
632 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
633 using *bedtools getfasta*⁷² (v2.29.2), and used *blastn*⁶⁹ with parameters *-dust no -soft_masking*
634 *false -word_size 7 -outfmt 6 -max_target_seqs 1 -evalue 0.05 -gapopen 5 -gapextend 2*
635 (v2.2.31) to determine the matching position in the consensus sequence.

637 **Identification of transcription factor binding sites in *roo* sequences.** We retrieved from
638 JASPAR⁷³ (v2022) the models for 160 transcription factor binding sites (TFBS) motifs of *D.*
639 *melanogaster*. We used FIMO⁷¹ (v5.3.0) to scan for TBFS in the repetitive *roo* sequence from
640 the consensus sequence (region: 1052-1166), as well as in the fragments incorporated in the
641 gene-TE chimeras, with a significant threshold of 1×10^{-4} .

642 **Genome-wide BLAST analysis of *roo* low complexity sequences.** We performed a BLAST
643 search with *blastn*⁶⁹ (v2.2.31) (with parameters: *-dust no -soft_masking false -outfmt 6 -*
644 *word_size 7 -evalue 0.05 -gapopen 5 -gapextend 2 -qcov_hsp_perc 85 -perc_identity 75*).
645 Next, we used *bedtools intersect*⁷² (v2.29.2) with the gene and transposable elements
646 annotations to see in which positions the matches occur. We analyzed the top 20 matches of
647 each blastn search.

648 **Identification of *D. simulans* *roo* consensus sequence.** We obtained a superfamily level
649 transposable elements library for *D. simulans* using REPET. We used *blastn*⁶⁹ (v2.2.31) with
650 a minimum coverage and percentage of identity of the 80% (*-qcov_hsp_perc 80 -perc_identity*
651 *80*) to find the sequence corresponding to the *roo* family. Then, we used again *blastn*⁶⁹ (with
652 parameters *-qcov_hsp_perc 80 -perc_identity 80 -dust no -soft_masking false -word_size 7 -*

653 *max_target_seqs 1 -eval 0.05 -gapopen 5 -gapextend 2*) to check if the *roo* sequence from
654 *D. simulans* contained the repetitive region present in the *D. melanogaster* *roo* consensus
655 sequence. The *roo* consensus sequence from *D. simulans* is available in the GitHub repository
656 (<https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster>).

657 **Retrotransposons and DNA transposons enrichment**

658 We used the percentage of retrotransposons and DNA transposons of the genome of the five
659 strains provided in Rech et al (2022)³⁴ and performed a test of proportions to compare this
660 percentage to the percentage of retrotransposons and DNA transposons detected in the
661 chimeric gene-TE transcripts dataset.

662 **Expression level estimation**

663 To estimate the level of expression of the whole set of transcripts assembled we used the
664 script *align_and_estimate_abundance.pl* from the Trinity package³⁸ (v2.11.0), using salmon⁷⁴
665 as the estimation method. We next used the script *abundance_estimates_to_matrix.pl* from
666 the Trinity package to obtain the level of expression of transcripts using the TMM normalization
667 (Trimmed Means of M values). For each transcript, the expression levels of the three replicates
668 were averaged. For the analyses, we considered transcripts with a minimum expression level
669 of one TMM. Genes were categorized in three groups: (i) genes that were never detected as
670 producing chimeric isoforms, (ii) genes that always were detected as producing chimeric gene-
671 TE transcripts and (iii) genes producing both chimeric and non-chimeric isoforms. For the later
672 type of genes, we calculated the fraction of the total gene expression that comes from the
673 chimeric transcript.

674 **Coding capacity assessment**

675 We assessed whether protein-coding chimeric gene-TE transcripts can produce a protein by
676 using the Coding Potential Assessment Tool (CPAT) software⁵¹ with default parameters.
677 CPAT has been optimized for the prediction of coding and non-coding isoforms in *Drosophila*.
678 Thus, we used the coding probability cutoff at 0.39⁵¹.

679 **PFAM scan of domain analysis and enrichment**

680 To scan for PFAM domains⁵² in the TEs detected in an internal exon, we extracted the TE
681 sequence from the chimeric transcripts using *bedtools getfasta*⁷² (v2.29.2⁷², translated it to the
682 longest ORF using *getorf*⁷⁵ (EMBOSS:6.6.0.0⁷⁵ and scan it using the script *pfam_scan.pl*^{52,76}
683 (v1.6) to identify any of the known protein family domains of the Pfam database (version 34).
684 We used dcGO enrichment online tool⁵³ to perform an enrichment of the PFAM domains
685 detected.

686 We scanned the consensus TE sequences for the domains present in TE fragments detected
687 in the chimerics transcripts using *pfam_scan.pl*^{52,76} (v1.6). If the domain was not detected
688 using pfam default parameters, we lowered the hmmscan e-value sequence and domain
689 cutoffs to 0.05.

690 **Chip-seq peak calling**

691 ChIP-seq reads were processed using fastp⁷⁷ (v0.20.1) to remove adaptors and low-quality
692 sequences. Processed reads were mapped to the corresponding reference genome using the
693 *readAllocate* function (parameter: *chipThres* = 500) of the Perm-seq R package⁷⁸ (v0.3.0), with
694 *bowtie*⁷⁹ (v1.2.2) as the aligner and the CSEM program⁸⁰ (v2.3) in order to try to define a single
695 location for multi-mapping reads. In all cases bowtie was performed with default parameters
696 selected by Perm-seq.

697 Then, we used the ENCODE ChIP-Seq caper pipeline (v2, available at:
698 <https://github.com/ENCODE-DCC/chip-seq-pipeline2>) in *histone* mode, using *bowtie2* as the
699 aligner, disabling pseudo replicate generation and all related analyses (argument
700 *chip.true_rep_only* = *TRUE*) and pooling controls (argument *chip.always_use_pooled_ctl* =
701 *TRUE*). MACS2 peak caller was used with default settings. We used the output narrowPeak
702 files obtained for each replicate of each sample to call the histone peaks. To process the peak
703 data and keep a reliable set of peaks for each sample, we first obtained the summit of every
704 peak and extended it ± 100 bp. Next, we kept those peaks that overlapped in at least 2 out of
705 3 replicates (following⁸¹) allowing a maximum gap of 100bp, and merged them in a single file
706 using *bedtools merge*⁷² (v2.30.0). Thus, we obtained for every histone mark of each sample a
707 peak file. We considered that a chimeric gene-TE transcript had a consistent epigenetic status
708 when the same epigenetic status was detected in at least 80% of the samples in which it was
709 detected.

710 **GO clustering analysis**

711 The Gene Ontology (GO) clustering analysis in the biological process (BP) and molecular
712 process (MP) category was performed using the DAVID bioinformatics online tool⁵⁰. Names
713 of the annotation clusters were manually processed based on the cluster's GO terms. Only
714 clusters with a score >1.3 were considered⁵⁰.

715 **Statistical analysis**

716 All statistical analyses were performed in R (v3.6.3) statistical computing environment⁸².
717 Graphics were created using *ggplot2* R package⁸³.

718 **Data availability**

719 RNA-seq and ChIP-seq raw data is available in the NCBI Sequence Read Archive (SRA)
720 database under BioProject PRJNA643665. The set of chimeric transcripts detected are
721 available in GitHub (<https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster>).
722 DrosOmics genome browser³⁶ (<http://gonzalezlab.eu/drosomics>) compiles all data generated
723 in this work.

724 **Code availability**

725 Scripts to perform analyses are available at GitHub
726 (<https://github.com/GonzalezLab/chimerics-transcripts-dmelanogaster>).

727 **ACKNOWLEDGMENTS**

728 We thank Carlos Vargas-Chavez and Simón Orozco for providing the *roo* consensus
729 sequence of *D. simulans*. We thank Simón Orozco and Ewan Harney for comments on the
730 manuscript.

731 **FUNDING**

732 This project has received funding from the European Research Council (ERC) under the
733 European Union's Horizon 2020 research and innovation programme (H2020-ERC-2014-
734 CoG-647900), and from grant PID2020-115874GB-I00 funded by MCIN/AEI/
735 10.13039/501100011033.

736
737 **REFERENCES**

- 738 1. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative
739 splicing complexity in the human transcriptome by high-throughput sequencing. *Nat. Genet.* **40**, 1413–1415 (2008).
- 741 2. Barash, Y. *et al.* Deciphering the splicing code. *Nature* **465**, 53–59 (2010).
- 742 3. Brown, J. B. *et al.* Diversity and dynamics of the Drosophila transcriptome. *Nat. 2014*
743 **5127515** **512**, 393–399 (2014).
- 744 4. Söllner, J. F. *et al.* An RNA-Seq atlas of gene expression in mouse and rat normal tissues.
745 *Sci. Data* **4**, 170185 (2017).
- 746 5. Kiyose, H. *et al.* Comprehensive analysis of full-length transcripts reveals novel splicing
747 abnormalities and oncogenic transcripts in liver cancer. *PLOS Genet.* **18**, e1010342
748 (2022).

749 6. Marasca, F. *et al.* The Sophisticated Transcriptional Response Governed by Transposable
750 Elements in Human Health and Disease. *Int. J. Mol. Sci.* 2020 Vol 21 Page 3201 **21**, 3201
751 (2020).

752 7. Singh, P. & Ahi, E. P. The importance of alternative splicing in adaptive evolution. *Mol.*
753 *Ecol.* (2022) doi:10.1111/MEC.16377.

754 8. Verta, J.-P. & Jacobs, A. The role of alternative splicing in adaptation and evolution.
755 *Trends Ecol. Evol.* **37**, 299–308 (2022).

756 9. Franchini, L. F., Ganko, E. W. & McDonald, J. F. Retrotransposon-gene associations are
757 widespread among *D. melanogaster* populations. *Mol. Biol. Evol.* **21**, 1323–1331 (2004).

758 10. Ganko, E. W., Bhattacharjee, V., Schliekelman, P. & McDonald, J. F. Evidence for the
759 contribution of LTR retrotransposons to *C. elegans* gene evolution. *Mol. Biol. Evol.* **20**,
760 1925–1931 (2003).

761 11. Jordan, I. K., Rogozin, I. B., Glazko, G. V. & Koonin, E. V. Origin of a substantial fraction
762 of human regulatory sequences from transposable elements. *Trends Genet. TIG* **19**, 68–
763 72 (2003).

764 12. Lipatov, M., Lenkov, K., Petrov, D. A. & Bergman, C. M. Paucity of chimeric gene-
765 transposable element transcripts in the *Drosophila melanogaster* genome. *BMC Biol.* **3**,
766 24 (2005).

767 13. van de Lagemaat, L. N., Landry, J.-R., Mager, D. L. & Medstrand, P. Transposable
768 elements in mammals promote regulatory variation and diversification of genes with
769 specialized functions. *Trends Genet. TIG* **19**, 530–536 (2003).

770 14. Casacuberta, E. & González, J. The impact of transposable elements in environmental
771 adaptation. *Mol. Ecol.* **22**, 1503–1517 (2013).

772 15. Cowley, M. & Oakey, R. J. Transposable Elements Re-Wire and Fine-Tune the
773 Transcriptome. *PLOS Genet.* **9**, e1003234 (2013).

774 16. Schrader, L. & Schmitz, J. The impact of transposable elements in adaptive evolution.
775 *Mol. Ecol.* **28**, 1537–1549 (2019).

776 17. Wolff, J. N. Turning junk into gold: Domestication of transposable elements and the
777 creation of new genes in eukaryotes. *BioEssays* **28**, 913–922 (2006).

778 18. Babaian, A. *et al.* LIONS: analysis suite for detecting and quantifying transposable
779 element initiated transcription from RNA-seq. *Bioinforma. Oxf. Engl.* **35**, 3839–3841
780 (2019).

781 19. Pinson, M.-E., Pogorelcnik, R., Court, F., Arnaud, P. & Vaurs-Barrière, C. CLIFinder:
782 identification of LINE-1 chimeric transcripts in RNA-seq data. *Bioinformatics* **34**, 688–690
783 (2018).

784 20. Babarinde, I. A. *et al.* Transposable element sequence fragments incorporated into coding
785 and noncoding transcripts modulate the transcriptome of human pluripotent stem cells.
786 *Nucleic Acids Res.* **49**, 9132–9153 (2021).

787 21. Treiber, C. D. & Waddell, S. Transposon expression in the *Drosophila* brain is driven by
788 neighboring genes and diversifies the neural transcriptome. *Genome Res.* gr.259200.119
789 (2020) doi:10.1101/gr.259200.119.

790 22. Faulkner, G. J. *et al.* The regulated retrotransposon transcriptome of mammalian cells.
791 *Nat. Genet.* 2009 **41** 563–571 (2009).

792 23. Batut, P., Dobin, A., Plessy, C., Carninci, P. & Gingeras, T. R. High-fidelity promoter
793 profiling reveals widespread alternative promoter usage and transposon-driven
794 developmental gene expression. *Genome Res.* **23**, 169–180 (2013).

795 24. Cordaux, R., Udit, S., Batzer, M. A. & Feschotte, C. Birth of a chimeric primate gene by
796 capture of the transposase gene from a mobile element. *Proc. Natl. Acad. Sci.* **103**, 8101–
797 8106 (2006).

798 25. Newman, R. M. *et al.* Evolution of a TRIM5-CypA Splice Isoform in Old World Monkeys.
799 *PLOS Pathog.* **4**, e1000003 (2008).

800 26. Tipney, H. J. *et al.* Isolation and characterisation of GTF2IRD2, a novel fusion gene and
801 member of the TFII-I family of transcription factors, deleted in Williams–Beuren syndrome.
802 *Eur. J. Hum. Genet.* **12**, 551–560 (2004).

803 27. Cosby, R. L. *et al.* Recurrent evolution of vertebrate transcription factors by transposase
804 capture. *Science* **371**, (2021).

805 28. Ueda, M. T. *et al.* Comprehensive genomic analysis reveals dynamic evolution of
806 endogenous retroviruses that code for retroviral-like protein domains. *Mob. DNA* **11**, 29
807 (2020).

808 29. Klasberg, S., Bitard-Feildel, T., Callebaut, I. & Bornberg-Bauer, E. Origins and structural
809 properties of novel and de novo protein domains during insect evolution. *FEBS J.* **285**,
810 2605–2625 (2018).

811 30. Conley, A. B., Piriyapongsa, J. & Jordan, I. K. Retroviral promoters in the human genome.
812 *Bioinforma. Oxf. Engl.* **24**, 1563–1567 (2008).

813 31. Lee, Y. C. G. & Karpen, G. H. Pervasive epigenetic effects of *Drosophila* euchromatic
814 transposable elements impact their evolution. *eLife* **6**, e25762 (2017).

815 32. Guio, L., Vieira, C. & González, J. Stress affects the epigenetic marks added by natural
816 transposable element insertions in *Drosophila melanogaster*. *Sci. Rep.* 2018 **81** **8**, 1–10
817 (2018).

818 33. Rebollo, R. *et al.* A snapshot of histone modifications within transposable elements in
819 *Drosophila* wild type strains. *PloS One* **7**, e44253 (2012).

820 34. Rech, G. E. *et al.* Population-scale long-read sequencing uncovers transposable elements
821 associated with gene expression variation and adaptive signatures in *Drosophila*. *Nat. Commun.* **13**, 1948 (2022).

823 35. Lanciano, S. & Cristofari, G. Measuring and interpreting transposable element expression.
824 *Nat. Rev. Genet.* **21**, 721–736 (2020).

825 36. Coronado-Zamora, M., Salces-Ortiz, J. & González, J. DrosOmics: the comparative
826 genomics browser to explore omics data in natural strains of *D. melanogaster*.
827 2022.07.22.501088 Preprint at <https://doi.org/10.1101/2022.07.22.501088> (2022).

828 37. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger
829 climate classification. *Hydrol. Earth Syst. Sci.* **11**, 1633–1644 (2007).

830 38. Grabherr, M. G. *et al.* Trinity: reconstructing a full-length transcriptome without a genome
831 from RNA-Seq data. *Nat. Biotechnol.* **29**, 644–652 (2011).

832 39. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome
833 alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–
834 915 (2019).

835 40. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-
836 seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).

837 41. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0.
838 <<http://www.repeatmasker.org>>. (2013).

839 42. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–
840 3100 (2018).

841 43. Oliveira, D. S. *et al.* ChimeraTE: A pipeline to detect chimeric transcripts derived from
842 genes and transposable elements. 2022.09.05.505575 Preprint at
843 <https://doi.org/10.1101/2022.09.05.505575> (2022).

844 44. Larracuente, A. M. *et al.* Evolution of protein-coding genes in *Drosophila*. *Trends Genet. TIG* **24**, 114–123 (2008).

846 45. Park, S. G. & Choi, S. S. Expression breadth and expression abundance behave
847 differently in correlations with evolutionary rates. *BMC Evol. Biol.* **10**, 241 (2010).

848 46. Choi, J. Y. & Lee, Y. C. G. Double-edged sword: The evolutionary consequences of the
849 epigenetic silencing of transposable elements. *PLOS Genet.* **16**, e1008872 (2020).

850 47. Yin, H., Sweeney, S., Raha, D., Snyder, M. & Lin, H. A High-Resolution Whole-Genome
851 Map of Key Chromatin Modifications in the Adult *Drosophila melanogaster*. *PLOS Genet.*
852 **7**, e1002380 (2011).

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

855 49. Koenecke, N., Johnston, J., Gaertner, B., Natarajan, M. & Zeitlinger, J. Genome-wide
856 identification of *Drosophila* dorso-ventral enhancers by differential histone acetylation
857 analysis. *Genome Biol.* **17**, 196 (2016).

858 50. Huang, D. W. *et al.* Extracting Biological Meaning from Large Gene Lists with DAVID.
859 *Curr. Protoc. Bioinforma.* **27**, 13.11.1-13.11.13 (2009).

860 51. Wang, L. *et al.* CPAT: Coding-Potential Assessment Tool using an alignment-free logistic
861 regression model. *Nucleic Acids Res.* **41**, e74 (2013).

862 52. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–
863 D419 (2021).

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

866 54. Magwire, M. M., Bayer, F., Webster, C. L., Cao, C. & Jiggins, F. M. Successive Increases
867 in the Resistance of *Drosophila* to Viral Infection through a Transposon Insertion Followed
868 by a Duplication. *PLOS Genet.* **7**, e1002337 (2011).

869 55. Aminetzach, Y. T., Macpherson, J. M. & Petrov, D. A. Pesticide resistance via
870 transposition-mediated adaptive gene truncation in *Drosophila*. *Science* **309**, 764–767
871 (2005).

872 56. Ellison, C. E., Kagda, M. S. & Cao, W. Telomeric TART elements target the piRNA
873 machinery in *Drosophila*. *PLOS Biol.* **18**, e3000689 (2020).

874 57. Bogwitz, M. R. *et al.* Cyp12a4 confers lufenuron resistance in a natural population of
875 *Drosophila melanogaster*. *Proc. Natl. Acad. Sci.* **102**, 12807–12812 (2005).

876 58. Marsano, R. M., Caizzi, R., Moschetti, R. & Junakovic, N. Evidence for a functional
877 interaction between the Bari1 transposable element and the cytochrome P450 cyp12a4
878 gene in *Drosophila melanogaster*. *Gene* **357**, 122–128 (2005).

879 59. Chuong, E. B., Elde, N. C. & Feschotte, C. Regulatory activities of transposable elements:
880 from conflicts to benefits. *Nat. Rev. Genet.* **18**, 71–86 (2017).

881 60. Modzelewski, A. J., Gan Chong, J., Wang, T. & He, L. Mammalian genome innovation
882 through transposon domestication. *Nat. Cell Biol.* **24**, 1332–1340 (2022).

883 61. Rogers, R. L. & Hartl, D. L. Chimeric Genes as a Source of Rapid Evolution in *Drosophila*
884 *melanogaster*. *Mol. Biol. Evol.* **29**, 517 (2012).

885 62. Salvador-Martínez, I., Coronado-Zamora, M., Castellano, D., Barbadilla, A. & Salazar-
886 Ciudad, I. Mapping Selection within *Drosophila melanogaster* Embryo's Anatomy. *Mol.*
887 *Biol. Evol.* **35**, 66–79 (2018).

888 63. Gotea, V. & Makałowski, W. Do transposable elements really contribute to proteomes?
889 *Trends Genet.* **22**, 260–267 (2006).

890 64. Rodriguez, M. & Makałowski, W. Software evaluation for de novo detection of
891 transposons. *Mob. DNA* **13**, 14 (2022).

892 65. Larkin, A. *et al.* FlyBase: updates to the *Drosophila melanogaster* knowledge base.
893 *Nucleic Acids Res.* **49**, D899–D907 (2021).

894 66. Andrews, S. FastQC: A Quality Control Tool for High Throughput Sequence Data.
895 Available online at: <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. (2010).

896 67. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–
897 2079 (2009).

898 68. Yang, H. *et al.* Re-annotation of eight *Drosophila* genomes. *Life Sci. Alliance* **1**,
899 e201800156 (2018).

900 69. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421
901 (2009).

902 70. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover
903 motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).

904 71. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif.
905 *Bioinformatics* **27**, 1017–1018 (2011).

906 72. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic
907 features. *Bioinformatics* **26**, 841–842 (2010).

908 73. Castro-Mondragon, J. A. *et al.* JASPAR 2022: the 9th release of the open-access
909 database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173
910 (2022).

911 74. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon: fast and bias-
912 aware quantification of transcript expression using dual-phase inference. *Nat. Methods*
913 **14**, 417–419 (2017).

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

916 76. Finn, R. D. *et al.* Pfam: the protein families database. *Nucleic Acids Res.* **42**, D222–230
917 (2014).

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

920 78. Zeng, X. *et al.* Perm-seq: Mapping Protein-DNA Interactions in Segmental Duplication and
921 Highly Repetitive Regions of Genomes with Prior-Enhanced Read Mapping. *PLOS
922 Comput. Biol.* **11**, e1004491 (2015).

923 79. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
924 alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

925 80. Chung, D. *et al.* Discovering Transcription Factor Binding Sites in Highly Repetitive
926 Regions of Genomes with Multi-Read Analysis of ChIP-Seq Data. *PLOS Comput. Biol.* **7**,
927 e1002111 (2011).

928 81. Yang, Y. *et al.* Leveraging biological replicates to improve analysis in ChIP-seq
929 experiments. *Comput. Struct. Biotechnol. J.* **9**, e201401002 (2014).

930 82. R Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation
931 for Statistical Computing, 2021).

932 83. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York,
933 2016).

934