Transposable elements contribute to the evolution of host shift-related genes in cactophilic *Drosophila* species

Oliveira D. S.[1,2,3], Larue A.[2], Nunes W. V. B. [2], Sabot F.[4], Bodelón A.[5], García Guerreiro M. P.[5], Vieira C.[2,*], Carareto C. M. A.[1,*]

[1] São Paulo State University (Unesp), Institute of Biosciences, Humanities and Exact Sciences, São José do Rio Preto, São Paulo, 15054-000, Brazil.

[2] Universite Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS, UMR5558, Villeurbanne 69100, France.

[3] Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14850

[4] DIADE, University of Montpellier, Cirad, IRD, 911 avenue Agropolis BP64501, 34394 Montpellier, France.

[5] Grup de Genòmica, Bioinformática i Biologia Evolutiva, Departament de Genètica i Microbiologia (Edifici C), Universitat Autònoma de Barcelona, Barcelona, Spain

* Corresponding authors

Daniel Siqueira de Oliveira: ds2497@cornell.edu

Anaïs Larue: anais.larue@univ-lyon1.fr

William Vilas Boas Nunes: william.vilas-boas-nunes@univ-lyon1.fr

François Sabot: francois.sabot@ird.fr

Alejandra Bodelón: Alejandra.Bodelon@uab.cat

María Pilar García Guerreiro: MariaPilar.Garcia.Guerreiro@uab.cat

Cristina Vieira: cristina.vieira@univ.lyon1.fr

Claudia Marcia Aparecida Carareto: claudia.carareto@unesp.br

ABSTRACT

Host shifts in insects have been considered a key process with the potential to contribute to reproductive isolation and speciation. Both genomics and transcriptomics variation have been attributed to such a process, in which gene families with functions associated with host localization, acceptance, and usage have been proposed to evolve. In this context, cactophilic *Drosophila* species are an excellent model to study host shift evolution, since they use a wide range of cacti as hosts, and many species have different preferences. Transposable elements are engines of genetic novelty between populations and species, driving rapid adaptive evolution. However, the extent of TEs' contribution to host shift remains unexplored. We performed genomic and transcriptomic analyses in six genomes of cactophilic species/subspecies to investigate how TEs interact with genes associated with host shift. Our results revealed enrichment of TEs at promoter regions of host shift-related genes, with ~39% of the odorant receptors containing their transcription factor binding sites within TEs. We observed that ~50% of these TEs are *Helitrons*, demonstrating an unprecedented putative *cis*-regulatory role of *Helitrons* in *Drosophila*. Differential expression analysis between species with different preferred hosts revealed divergence in gene expression in head and larval tissues. Although TEs' presence does not affect overall gene expression, we observed 6.27% of the expressed genes generating gene-TE chimeric transcripts, including those with function affecting host preference. Our combined genomic and transcriptomic approaches provide evidence of TE-driven divergence between species,

highlighting the evolutionary role of TEs in the context of host shift, a key adaptive process that can cause reproductive isolation.

Keywords: comparative transcriptomics, adaptation, chimeric transcripts.

60 **Introduction**

61 The molecular consequences of host shift evolution can ultimately cause
62 specialization and ecological speciation (1). In many species, reproductive isolation arises as
63 a result of successive adaptations driven by host shift, which results from divergent
64 selection between different environments (1–4). Thus, natural selection increases the
65 frequency of alleles that confer higher fitness benefits when a host shift occurs. *Drosophila*
66 species use a wide range of necrotic host tissues as feeding and breeding sites, providing a
67 suitable model to study the host shift process and its role in insect speciation. Among
68 them, the *repleta* group comprises a clade with ~100 cactophilic *Drosophila* species (5). The
69 specialization in cacti represents an important ecological challenge to the flies, mainly due
70 to the presence of toxic compounds (6). In the *repleta* species, the host shift was initially to
71 the *Opuntia sp*. cacti, and secondarily, several independent shifts happened to columnar
72 cacti (6), considered a host with higher toxicity than *Opuntia sp*. (7). Such host shift is
73 observed between sibling species with recent divergence in two clusters of the subgroup
74 *mulleri*: cluster *buzzatii* and cluster *mojavensis*. In the cluster *buzzatii*, *Drosophila buzzatii*
75 and *D. koepferae* are sibling species (divergence 4-5 Mya (8)) with different primary hosts.
76 The former species preferentially use *Opuntia sp*. cacti, whereas the latter is a columnar
77 cacti dweller (9). In the cluster *mojavensis*, *D. arizonae* and the three allopatric subspecies of
78 *D. mojavensis* have different cacti preferences. *D. arizonae* uses both *Opuntia sp*. and
79 columnar cacti, whereas *D. m. mojavensis* uses the columnar cacti *Ferocactus cylindraceus* as
80 preferential host; *D. m. wrigleyi* uses only *Opuntia sp*., *D. m. baja* uses the columnar cacti

81 *Stenocereus gummosus*, and *D. m. sonorensis* is also a columnar cacti dweller, using *S.*
82 *thurberi* (10). Therefore, this set of species from *buzzatii* and *mojavensis* clusters provides
83 an excellent model to investigate the molecular mechanisms underlying host shift
84 evolution.

85 The radiation of cactophilic *Drosophila* species across different hosts is an
86 evolutionary outcome that gave rise to specific adaptations (11). They can be summarized
87 into three steps: localization, acceptance, and host usage (12). In the localization step,
88 odorant-binding proteins (OBPs) and odorant receptors (ORs) are key proteins to
89 distinguish specific hosts in the environment, through the integration of the visual and
90 olfactory systems. Subsequently, in the acceptance step, the insect evaluates the nutritional
91 compounds present on the host, as well as the presence of competitors, predators,
92 pathogens and parasites (12). Several receptors are associated with this process, such as
93 gustatory receptors (GRs) (13) and ionotropic receptors (IRs) (14–16). Finally, the use of
94 metabolites derived from the host is essential for feeding and to complete the breeding
95 process. Although not always associated with detoxification (17), cactophilic flies must
96 overcome the presence of toxic substances from columnar cacti. Several gene families are
97 associated with detoxification, such as Cytochrome P450s (CYPs), Glutathione S-
98 Transferases (GSTs), UDP-Glycosyltransferases (UGTs), esterases (ESTs), and ATP binding-
99 cassette transporters (ABCs). Altogether, these nine gene families associated with the Host
100 Localization, Acceptance, and Usage will henceforth be referred to as HLAU.

101 Transposable elements (TEs) may play a role in environmental adaptation because of
102 their ability to generate mutations. In most cases, mutations caused by TEs are likely to be
103 deleterious or neutral. Throughout evolutionary time, TEs that remain in the genome tend
104 to be silenced by epigenetic control and/or small RNA pathways (18), accumulating
105 mutations, and losing their transposition ability (19). Despite the majority of TEs becoming
106 silenced, the remaining TE sequences may still contain regulatory motifs or protein
107 domains (20). These sequences can be co-opted by the cell machinery, modifying gene
108 expression or protein sequences of the nearby genes (reviewed in (21)). Depending on the
109 outcome of the cooption for the individual fitness, these TEs can increase their frequency in
110 the population, contributing to populational evolution and hence being determined as
111 adaptive insertions. For instance, in *D. melanogaster*, the gene *CHKov1* has a truncated
112 mRNA derived from a *DOC* TE located into the intron, resulting in resistance to viral
113 infection and insecticides (22). Such exaptation and domestication events can be often
114 identified by the occurrence of chimeric transcripts, which are mRNAs with both gene and
115 TE-derived sequences (23). A recent transcriptome-wide study has identified 327 genes of
116 *D. melanogaster* genes produce chimeric transcripts in different populations (24). Among all
117 genes, 76 generate chimeric transcripts from TE insertions that were present in one strain
118 but absent in another, highlighting the potential of TEs as a source of genetic novelty
119 between different *Drosophila* ecotypes.

120 Many aspects involving host shift adaptation in *repleta* species have been shown
121 previously, such as detoxification pathways (25), morphology (26), life history traits (27),

122 behavior (28), and genomic and transcriptomic differences (29–33). However, the
123 contribution of TEs to the evolution of HLAU genes has not yet been assessed. Here, we
124 aimed to uncover the extension of the genetic variability derived from TEs in cactophilic
125 *Drosophila* species, using both genome- and transcriptome-wide analysis. We tested the
126 hypothesis whether TEs have contributed to the evolution of HLAU genes, and
127 consequently to the host shift in cactophilic species.

128 **Results**

129 ***Genome assemblies and gene annotation***

130 To investigate the potential role of TEs in host shift of cactophilic *Drosophila* species,
131 we performed Nanopore long-read sequencing on species that have different preferential
132 cacti as hosts: *D. buzzatii* (*Opuntia sp.*), *D. koepferae* (columnar cacti), *D. arizonae* (*Opuntia*
133 *sp.* or columnar cacti), *D. m. mojavensis* (*F. cylindraceus*), *D. m. wrigleyi* (*Opuntia sp.*), and *D.*
134 *m. sonorensis* (*S. thurberi*). We obtained high-resolution assemblies for all genomes, with an
135 average of 567 scaffolds, 13.8 Mb of N50 (Supplemental Table 1), and ~98% of BUSCO
136 genes (Supplemental Fig. S1). Furthermore, the gene annotation for these genomes
137 covered 98.63% of all reference genes in the *D. mojavensis* subspecies, and 95.98% in *D.*
138 *arizonae*. The *de novo* gene annotation in *D. buzzatii* and *D. koepferae* revealed a total of
139 18,050 and 17,848 coding genes, for which 63.77%, and 64.13% were successfully assigned
140 as one-to-one orthologs with *D. mojavensis*, respectively. These results represent the higher
141 gene repertoire identified for *D. buzzatii* and *D. koepferae* species, altogether with their

142 ortholog relationship with *D. mojavensis* and *D. arizonae*. Since analysis of HLAU genes is
143 the focus of this study, their efficient annotation is relevant. They are represented by nine
144 gene families associated with the three steps in the use of host resources: (1) Localization:
145 OBPs, ORs; (2) Acceptance: GRs, IRs; and (3) Host usage: ABCs, ESTs, GSTs, UDPs, CYPs
146 (Supplemental Table 2). In *D. arizonae* and *D. mojavensis* subspecies, we recovered
147 ~98.21% of the total HLAU genes from the reference genome. In *D. buzzatii* and *D.*
148 *koepferae*, the number of annotated HLAU genes covered ~95% in both species compared
149 to *D. mojavensis* (Table 1).

Species/subspecies	Total genes	Host localization, acceptance and usage genes								
		Localization		***Acceptance***		***Host usage***				
		OBP	**OR**	**GR**	**IR**	**ABC**	**EST**	**GST**	**UDP**	**CYP**
Ref. *D. m. wrigleyi*	**15,045**	**25**	**59**	**45**	**22**	**27**	**17**	**28**	**21**	**59**
D. m. mojavensis	14,810	25	59	44	22	27	17	27	20	59
D. m. wrigleyi	14,948	25	59	45	22	27	17	27	20	59
D. m. sonorensis	14,808	25	57	44	22	27	17	27	20	58
D. arizonae	14,441	25	56	41	21	27	17	27	20	59
D. buzzatii	18,050	25	52	34	21	27	16	27	20	68
D. koepferae	17,848	25	45	34	24	27	13	25	23	62

150 **Table 1**. Total number of genes annotated in six *Drosophila* cactophilic species. The "Ref. *D. m. wrigleyi*" represents the
151 annotation in the reference genome used to perform the gene annotation for our long-read Nanopore genomes. HLAU
152 genes are separated according to their functions: Localization: odorant-binding proteins (OBP), odorant receptor (OR);
153 Acceptance: gustatory receptor (GR), ionotropic receptor (IR); Host usage: ATP-binding cassette (ABC), carboxylesterase
154 (EST), glutathione S-Transferase (GST), glycosyltransferase (UDP), and cytochrome P450 (CYP).

155 ***Minor effects of positive selection in HLAU genes***

156 The use of *Opuntia sp.* has been proposed as the ancestral state in cactophilic
157 *Drosophila* species, whereas columnar cacti are the derived state (6). To investigate whether
158 species that prefer columnar cacti as hosts exhibit signs of positive selection in HLAU
159 genes, we employed a branch-site approach to compare their evolutionary rates with those
160 of species that use *Opuntia sp*. For this analysis, we removed *D. arizonae* due to its non-
161 preferential usage between columnar and *Opuntia sp.* cacti (6). Moreover, we added *D.*
162 *navojoa*, which is a basal species of the *mojavensis* cluster that uses solely *Opuntia* sp. (34).
163 Therefore, we carried out the branch-site selection test with *D. buzzatii*, *D. m. wrigleyi*, and
164 *D. navojoa* as *Opuntia*-related branches (background), and *D. m. mojavensis*, *D. sonorensis*,
165 and *D. koepferae* as columnar-related branches (foreground). It is important to highlight
166 that due to the requirement for one-to-one orthologous genes between the six genomes,
167 our analysis has reduced the total number of HLAU genes to 42.9% of the initial number
168 presented in Table 1. From 130 genes, we observed six genes under positive selection (FDR
169 < 0.05): *Obp56d*, *Obp19d*, *Gr5a*, *Ir25a*, *Cyp4s3*, and *esterase S*. This result suggests that
170 positive selection in HLAU genes is not a major evolutionary driving force in the species
171 under study.

172 ***Evolutionary dynamics of TEs in cactophilic Drosophila species***

173 We used a combination of automatic and manual methods to build a curated TE
174 library for each species, based on the classification of TEs (35). The method was designed to
175 perform polishing steps in both TE consensus and copies to remove artifacts (see Methods).

176 The total number of TE consensuses (LTRs separated from internal sequences) ranged from
177 255 in *D. buzzatii* to 372 in *D. m. mojavensis*. The TE content in the genomes ranged from
178 15.04% in *D. buzzatii* to 22.01% in *D. m. mojavensis* (Fig. 1A). In addition, we observed a
179 positive correlation between TE content and genome size estimated from the assemblies
180 (*Pearson* correlation: p = 0.0082; R = 0.93) (Supplemental Fig. S2). Both species from the
181 *buzzatii* cluster had the lower genome sizes and TE contents compared to species of the
182 *mojavensis* cluster (Fig. 1A). Regarding the proportional load of TE orders in the genomes,
183 LTRs and Helitrons contribute on average with ~70% of total TE content in all species (Fig.
184 1B).

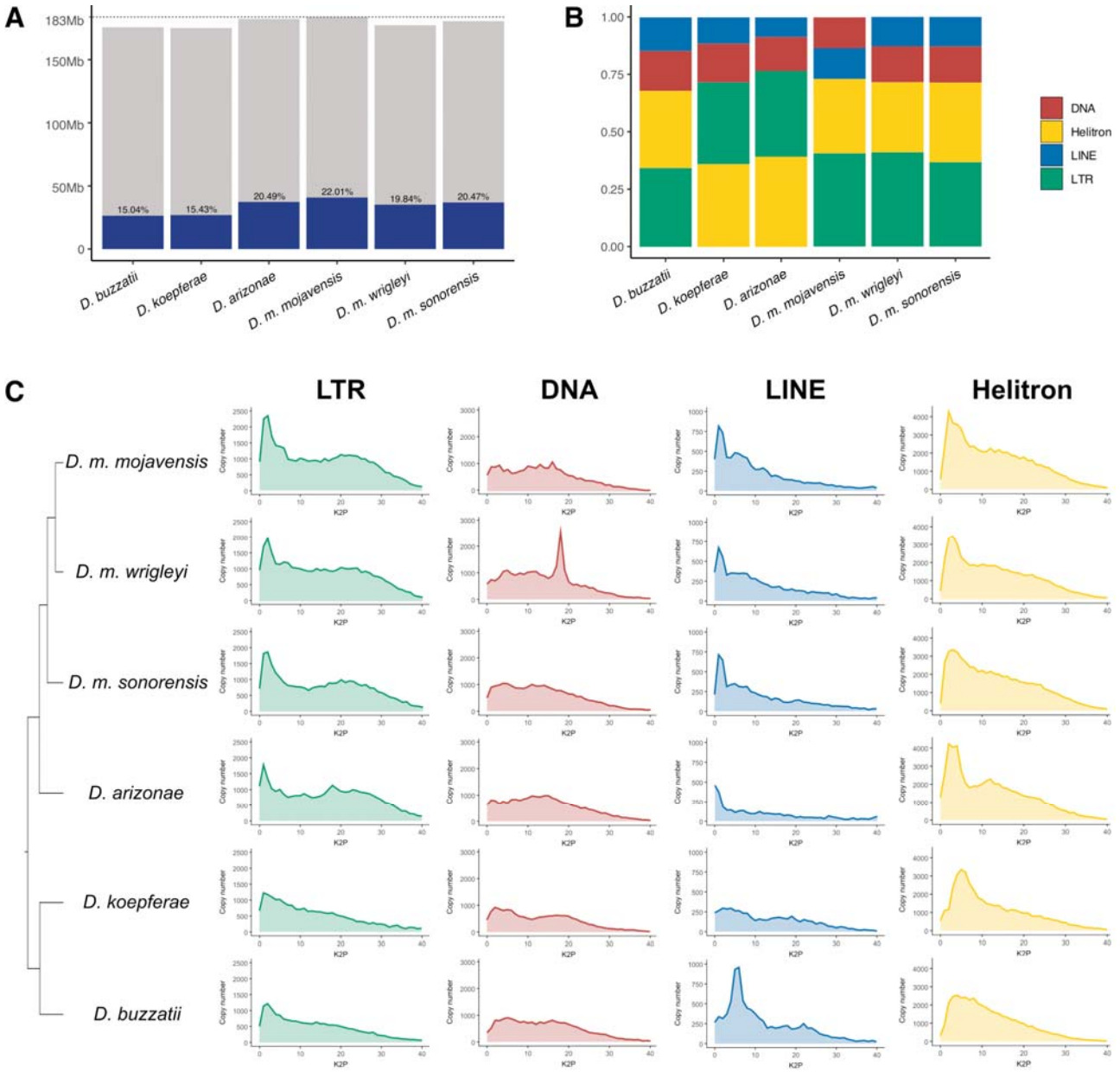


Figure 1. A) Genome size of the six cactophilic *Drosophila* species sequenced in this work, and their respective TE content. **B**) Proportional distribution of the total TE content in the genomes, in crescent order. **C**) Phylogenetic relationship among species (adapted from (25,37)), alongside TE order landscapes representing total copy number (y-axis), and the divergence of TE insertions compared to the consensuses (x-axis), based on Kimura 2-parameters (K2P) corrected with CpG. The copies located on the left of the graph have low divergence with their respective consensus, hence inferring that they might be

192 conserved/recent/active, whereas copies on the right side represent copies with high divergence due to the
193 accumulation of mutations, corresponding to old/inactive copies.

194 In the *mojavensis* cluster (*D. arizonae* and the three *D. mojavensis* subspecies), the
195 abundance of TEs in the genome varied substantially (Fig. 1B), despite their recent
196 divergence time of 1.5 Mya (36). *D. m. mojavensis* has the TE-richest genome (22.01%), with
197 the abundance following from the highest to the lowest: LTR > Helitron > LINE > DNA (Fig.
198 1B). The distribution of TEs in the other two *D. mojavensis* subspecies and *D. buzzatii* does
199 not follow such order. LTRs and Helitrons have the highest abundance, followed by DNA
200 and LINEs (Fig. 1B). In *D. arizonae* and *D. koepferae*, Helitrons are the most abundant,
201 following LTR, DNA, and LINEs (Fig. 1B). The observed differences are often associated with
202 the divergence in the evolutionary patterns of the TE orders between the genomes. We
203 used copy number as a proxy to measure TE expansion in the genomes, and divergence
204 Kimura 2-parameters (K2P) as age measurement. In the *mojavensis* cluster, we observed
205 that LTRs and Helitrons had a higher expansion in *D. m. mojavensis* compared to the other
206 genomes (Fig. 1C). The higher copy number in *D. m. mojavensis* is mostly from young TEs
207 (K2P 0-5), which might explain the higher TE content in this species. DNA elements have a
208 constant evolutionary dynamics in the genomes, except in *D. m. wrigleyi*. In this species, we
209 observed a peak in copy number with K2P between 16 and 20, considered old insertions.
210 LINEs had a similar evolutionary landscape in *D. mojavensis* subspecies, with higher copy
211 number from young copies compared to *D. arizonae*. In the *buzzatii* cluster, Helitrons show
212 higher copy number from young insertions in *D. koepferae* than in *D. buzzatii* (Fig. 1C),

213 whereas LINEs show a past transposition burst in *D. buzzatii* compared to *D. koepferae*. The
214 antagonistic patterns of TE dynamics observed in the six genomes suggest that TEs had
215 different successful lineages across these species.

216 ***HLAU genes are enriched in TEs within regulatory regions***

217 TE insertions may be selected in the promoter of genes if they provide a benefit to
218 the individual fitness, through the spreading of histone marks, or *cis*-regulatory elements
219 (38,39). To investigate whether TEs are enriched in the promoters (2 Kb upstream of the
220 TSS) from the nine HLAU families, we performed permutation tests in the six genomes (see
221 Methods). To compare our findings with other non-host shift related genes, we also
222 included zinc-finger and serine/threonine kinases as two non-HLAU gene families. We
223 observed that ORs have enrichment of TEs at the promoters in all genomes from species of
224 the *mojavensis* cluster (p-value < 0.05), (Fig. 2A). For acceptance genes, *D. koepferae* had
225 enrichment of TEs in the promoters of GRs (Fig. 2A). Finally, the enrichment of TEs within
226 promoters of detoxification genes was observed in *D. arizonae* (CYPs), *D. buzzatii* (GSTs),
227 and *D. koepferae* (ABCs). The results depicting the histograms with frequency of genes with
228 TEs for each significant permutation test, and p-values are shown in the Supplemental Fig.
229 S3 and Supplemental Table 3. Overall, detoxification genes had z-scores towards depletion
230 of TEs on promoters, with significant depletion for UGTs in *D. buzzatii* (p-value = 0.033) and
231 *D. arizonae* (p-value = 0.0135), GSTs in *D. koepferae* (p-value = 0.0205) (Fig. 2A). In the non-
232 HLAU gene families, we observed the expected result by chance: either z-score indicated

the expected number of genes with TEs on promoters in a genome-wide comparison (z-score near 0), or it indicated depletion of TEs, as observed for SER genes in *D. buzzatii*. This result reinforces our findings, indicating the significant and evolutionary unexpected enrichment of TEs in HLAU genes, especially on the promoters of ORs.

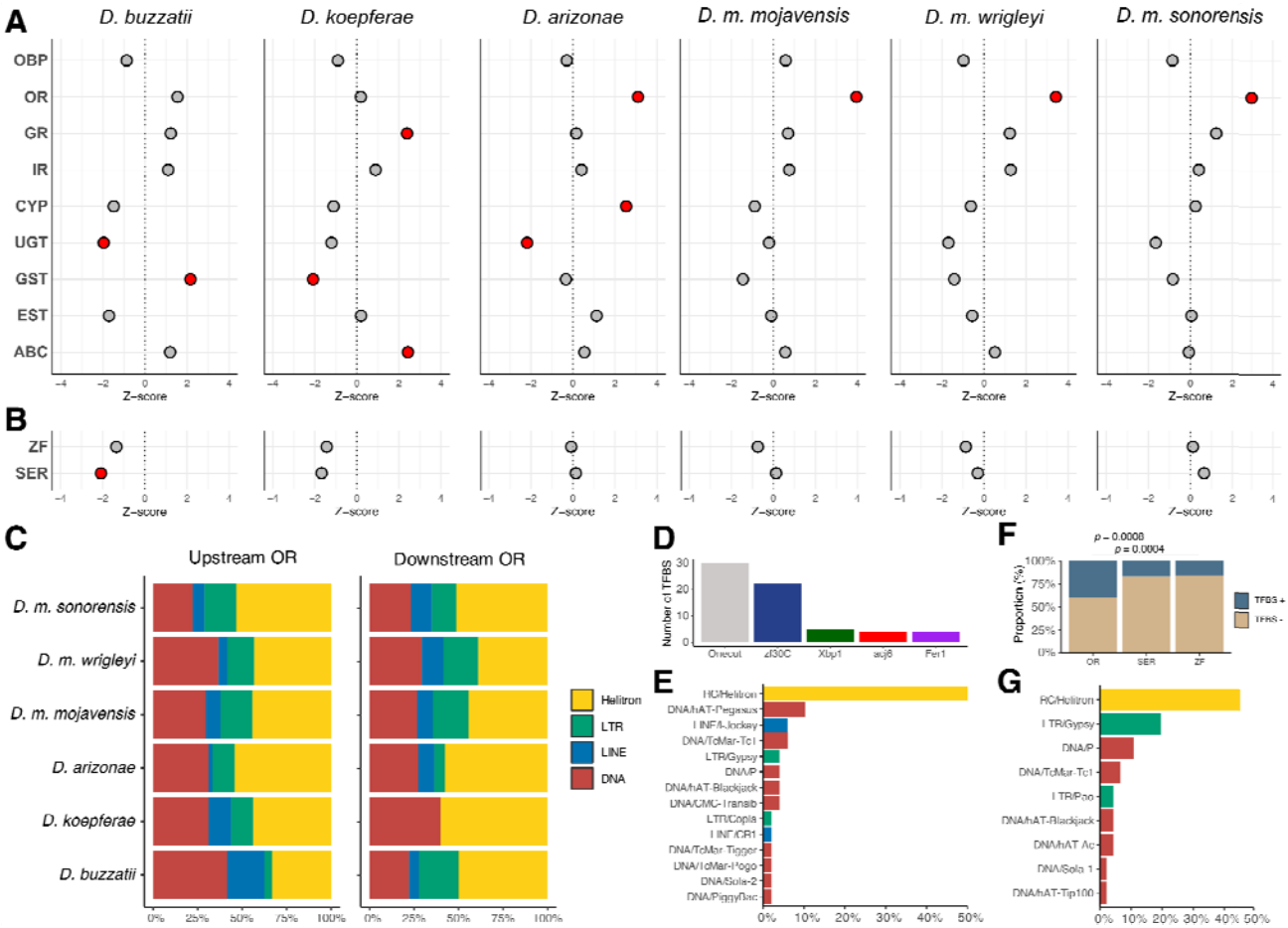


Figure 2. A) Enrichment of TEs in the promoter region (2kb upstream) of HLAU gene families. Z-score > 0 indicates enrichment, and Z-score < 0 indicates depletion. Red dots: p-value < 0.05, grey dots: p-value > 0.05. **B**) The same enrichment analysis as depicted on A, but for non-HLAU gene families, represented by zinc-finger proteins (ZF) and Serine/Threonine Kinases (SER). **C**) Proportion of TE orders located upstream and downstream ORs. **D**) Transcription factor binding sites (TFBSs) frequency on TEs located upstream of OR genes in all species under study. **E**) Proportional frequency of TE families with TFBSs embedded on their

244 sequences from TE copies located upstream of OR genes. **F**) Significant Fisher's exact test between OR genes
245 with OR-related TFBSs from TE insertions compared to the frequency observed in the non-HLAU gene families
246 Serine/Threonine Kinases (SER) and zinc-finger proteins (ZF). **G**) Proportion of TE orders carrying OR-related
247 TFBSs in non-HLAU gene families.

248 To obtain further insights into the potential regulatory role of TEs, we also analyzed
249 the enrichment at the 2 kb downstream region. Among all species, the only significant
250 enrichments were observed in ORs from species of the *mojavensis* cluster (Supplemental
251 Table 3), as we observed in the promoters. The TE enrichment upstream and downstream
252 of ORs indicates a possible coevolution of this gene family with TEs. Alternatively, it may
253 also suggest that ORs are located in TE-rich regions of the genomes. To investigate this
254 hypothesis, we identified TE-rich regions (Supplemental Fig. S4) and analyzed the frequency
255 of ORs. We observed a single OR in TE-rich regions in *D. arizonae*, *D. m. mojavensis*, and *D.*
256 *m. wrigleyi*. (Supplemental Table 4). Therefore, we rejected the hypothesis that ORs are
257 enriched with TEs in their vicinity due to their location in TE-rich regions. Other factors
258 might be involved in the evolutionary maintenance of TEs near ORs, such as the co-option
259 of regulatory motifs.

260 A few TE families are more prone to contribute to *cis*-elements than others, such as
261 retroelements in *D. melanogaster* (40). Therefore, we investigated whether the observed
262 enrichment of TEs in the HLAU gene families is TE-specific. Considering all species, 45.36%
263 and 50.36% of the insertions located upstream and downstream of ORs correspond to
264 *Helitrons*, respectively (Fig. 2C). We decided to analyze whether these *Helitron* insertions

265 upstream to HLAU genes, mainly ORs, could be part of the first exons from genes, since
266 such overlap has been observed previously in other species than *Drosophila* (41,42). By
267 intersecting exons from genes enriched with their respective upstream *Helitrons*, we
268 observed that *Or85c* (overlap 772 nt) and *Or22c* (52 nt) had overlap with *Helitrons* in *D. m.*
269 *mojavensis*, whereas *D. m. sonorensis* had only *Or22c* (58 nt). While the insertion in *Or85c*
270 suggests a subspecies-specific insertion, the insertion in *Or22c* is likely to be inherited from
271 the species' common ancestor (Supplemental Table 5). At the ORs' downstream region, we
272 observed four genes with *Helitrons* overlapping exons: *Or71a*, *Or10a*, *Or13a*, and *Or82a*.
273 Interestingly, the same *Helitron* in the *Or82a* was observed in the three *D. mojavensis*
274 subspecies, highlighting the inheritance from the common ancestor and maintenance of
275 the insertion over time. Taken together, our results rejected the hypothesis regarding the
276 possible role of *Helitrons* carrying HLAU gene fragments. However, these elements could
277 still act on gene regulation if they carry *cis*-elements as TFBSs.

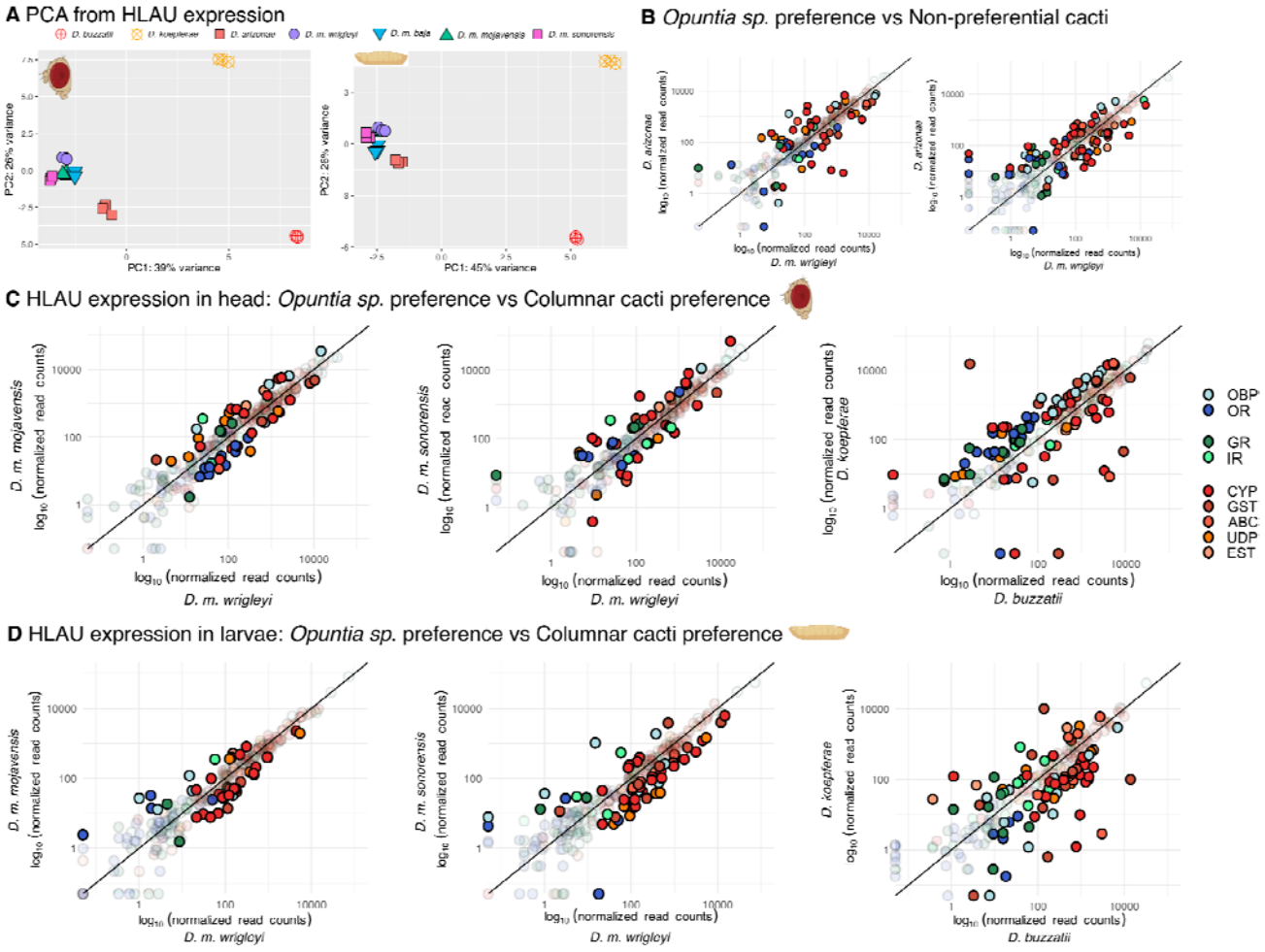
278 TEs located near genes can modulate gene expression by donating TFBSs or polyA
279 signals to the ancestral regulatory motifs. To investigate the potential functional role of the
280 observed enrichment of TEs in ORs, we performed TFBS identification on all TE insertions
281 located in OR promoters. We used the transcription factors (TFs) previously described to
282 participate in OR's transcription in *D. melanogaster*: acj6, onecut, xbp1, fer1, and zf30C (43).
283 Taking all species together, the most frequent TFBS found was for the TF onecut (46.15%),
284 followed by zf30C (33.84%), and Xbp1 (7.69%) (Fig. 2D). *Helitrons* were the TEs with the
285 highest frequency, representing 50% of all insertions carrying ORs' TFBSs (Fig. 2E). Our

286 analysis revealed that 39.28% of the ORs with TEs in promoters have the TFBS located
287 within the insertion (Supplemental Table 6). To test whether these results are specific to
288 ORs, we also performed the same detection of TFBSs and calculation of TEs frequency in
289 the non-HLAU families ZF and SER. Our results show that ZF and SER have, on average,
290 15.55% and 17.14% of the genes with OR-related TFBSs, respectively. The detection of
291 these TFBSs in ZF and SER is likely due to chance, given the short length of the motifs.
292 Comparing these proportions with the ones found in ORs, we observed significantly more
293 TFBSs compared to ZF (*Fisher's exact test: p-value = 0.0004341*) and SER (*Fisher's exact test:*
294 *p-value = 0.0008519*) (Fig. 2F and Supplemental Table 7). In addition, *Helitrons* also had
295 higher prevalence than other TEs upstream of SER and ZF genes (Fig. 2G), demonstrating a
296 potential insertion preference for promoter regions. Although similar *Helitron* frequencies
297 between OR, SER, and ZF genes, our data suggest that *Helitrons* might have been retained
298 in OR promoters due to the presence of functional TFBSs (Fig. 2F).

299 ***HLAU genes are differentially expressed between Opuntia sp. and columnar cacti***
300 ***dwellers***

301 Differential expression of genes has been associated with adaptation to new hosts in
302 insects (32). Here, aiming to test whether gene expression divergence is associated with
303 adaptation to different cactus hosts, we performed differential expression analysis between
304 species with *Opuntia sp.* preference vs columnar cacti preference, in head and larval tissues.
305 Since our fly stocks were reared in the lab with corn media, the observed transcriptome

306 response is constitutive rather than triggered by the native cacti. Overall, the principal
307 component analysis from HLAU gene expression (normalized read counts) revealed
308 remarkable differences between species in both tissues (Fig. 3A). The variance of gene
309 expression resembles the phylogenetic divergence of the species, with the highest variance
310 between the *buzzatii* and *mojavensis* clusters (PC1), followed by divergence between *D.*
311 *buzzatii* and *D. koepferae* (PC2). The species of the *mojavensis* cluster were clustered
312 together with a slight separation between *D. arizonae* and the subspecies of *D. mojavensis*.



313

314 **Figure 3**: **A**) Principal component analysis of HLAU genes' expression in head and larval tissues. Both tissues
315 had HLAU genes' variance representing the phylogenetic distance between species: *D. buzzatii* and *D.*

316 *koepferae*, compared to the *mojavensis* cluster corresponds to the PC1 (45%), and 26% between them in PC2.
317 Species from the *mojavensis* cluster had strong similarity, with *D. arizonae* slightly separated from *D.*
318 *mojavensis*. **B**) HLAU genes' expression in the *Opuntia sp.* dweller *D. m. wrigleyi* versus its sister species *D.*
319 *arizonae*, which does not have preference between *Opuntia sp.* and columnar cacti. **C**) HLAU expression in
320 head: differential expression analysis of HLAU genes from head in the *Opuntia sp.* dweller *D. m. wrigleyi*
321 versus the three columnar dwellers *D. m. mojavensis,* and *D. buzzatii* versus *D. koepferae*. In A and B: Filled
322 dots represent HLAU genes with log2 fold-change > |1| and FDR < 0.05; transparent dots are non-significant
323 differential expression. **D**) HLAU expression in larvae: The same pairwise species as in A with *Opuntia sp.*
324 dweller vs columnar dwellers, but from larvae tissue.

325 We performed pairwise analysis of HLAU gene expression between species with
326 *Opuntia sp.* preference (*D. m. wrigleyi* and *D. buzzatii*) versus species with preference for
327 columnar cacti, and *D. arizonae* as non-preferential cacti. All differentially expressed HLAU
328 genes, as well as the presence of TEs in their promoters, can be accessed in (Supplemental
329 Table 8). In head, *D. m. wrigleyi* had 27 and 19 HLAU genes with higher expression
330 compared to *D. m. mojavensis* and *D. m. sonorensis*, respectively (Fig. 3A). Four genes were
331 found with higher expression in the head of *D. m. wrigleyi* compared to the other two
332 subspecies: *UDP-Ugt4*, *Gst-theta3*, *Abc-a3*, and *Cyp6a21*. In the subspecies with columnar
333 cacti preference, we observed, on average, 26 HLAU genes with higher expression
334 compared to *D. m. wrigleyi*. Among them, seven were recurrent in the two subspecies:
335 *Ir21a*, *EstB1*, *UDP2*, *UDP-Ugt5*, *Cyp9b2*, *UDP-Ugt5-1*, and *Cyp12a2*. Since they were
336 observed in the two subspecies with columnar cacti preference, we propose that they are
337 candidates to be related to the adaptation to columnar cacti. Finally, the same analysis on

338 larvae revealed that *D. m. wrigleyi* has three genes with recurrent higher expression
339 compared to the other subspecies: *Or82a*, *Obp59a*, and *Ir21a*; whereas in the other two *D.*
340 *mojavensis* subspecies we observed 17 common up-regulated genes compared to *D. m.*
341 *wrigleyi*, which comprise seven CYP genes, three GSTs, four UDPs, one Esterase and one
342 OBP (Fig. 3B). Regarding *D. buzzatii* and *D. koepferae*, we observed 27 HLAU genes in head
343 and 54 in larvae with higher expression in *D. buzzatii*, and 63 in head and 29 with higher
344 expression in *D. koepferae*. We also compared the head expression of HLAU genes between
345 *D. m. wrigleyi* and its sister species without cacti host preference - *D. arizonae*. In *D. m.*
346 *wrigleyi*, we found 31 HLAU genes with higher expression compared to its counterpart *D.*
347 *arizonae*. In larvae, *D. m. wrigleyi* had 43 HLAU genes with higher expression compared to
348 *D. arizonae* (Fig. 3D). This result demonstrates that HLAU genes have divergent differential
349 expression in both head and larval tissues.

350 To further investigate whether differential expression of HLAU genes has a higher
351 prevalence compared to non-HLAU genes, we compared the proportion of differentially
352 expressed genes associated with localization, acceptance, and usage with the proportion of
353 differentially expressed genes from the non-HLAU gene families Ser/Thr Kinases and Zinc-
354 finger. In head, our analysis revealed that localization and usage have higher differential
355 expression than non-HLAU genes between *D. buzzatii* and *D. koepferae* (Supplemental Fig.
356 S5). In *D. mojavensis* subspecies, we observed higher differential expression for localization
357 genes between *D. m. wrigleyi* and *D. m. mojavensis*. In larvae, we did not observe any
358 significant difference between non-HLAU and HLAU genes, but the pairwise differential

359 expression between *D. m. wrigleyi* and two subspecies revealed a higher prevalence of
360 gene families associated with usage (detoxification) compared to non-HLAU (Supplemental
361 Fig. S5). In *D. m. wrigleyi*, we also observed higher differential expression of localization
362 genes than in non-HLAU. Finally, the differential expression between *D. m. wrigleyi* and *D.*
363 *arizonae* demonstrated significantly higher proportion of genes associated with localization
364 and usage compared to non-HLAU in larvae, and higher prevalence of differentially
365 expressed usage genes in head (Supplemental Fig. S5). Taken together, the quantitative
366 results suggest that adaptations that have allowed the shift from *Opuntia sp* to columnar
367 cacti may arise from the evolution of gene expression.

368 ***The enrichment of TEs on HLAU genes is not associated with changes in gene***
369 ***expression***

370 Since we observed enrichment of TEs on the regulatory regions of a few host shift-
371 related families, mainly ORs, and we also found differentially expressed HLAU genes
372 between species, we tested whether the presence of TE insertions is associated with
373 differential expression. To do so, we selected expressed HLAU genes (see Methods) and
374 performed X^2 independence test, splitting them into two variables with two categories
375 each: 1) expression level: differentially expressed, and non-differentially; 2) TE presence:
376 with TEs at 2kb upstream, and without TE at 2kb upstream. Considering all pairwise analysis
377 (Fig. 3A,B-D), the X^2 results demonstrated that the differential expression of HLAU genes is
378 not associated with the presence of TEs on the promoter region (Supplemental Table 9).

379 Since we did not observe overall effects of TEs on the expression of HLAU genes together,
380 we investigated the association of TEs with differential expression only for HLAU gene
381 families that we found TE enrichment (Fig. 2A). None of the enriched gene families had a
382 significant association between the presence of TEs and their expression level
383 (Supplemental Table 9). Thus, a global enrichment of TEs in the regulatory region is not
384 likely to be associated with the differential expression of HLAU genes.

385 ***Make or break: gene-TE chimeras contribute either predominantly or negligibly to***
386 ***gene expression***

387 Although highly deleterious, a small fraction of TEs generating chimeric transcripts
388 can be positively selected, giving rise to domestication and exaptation events (44). Aiming
389 to identify the extent of the transcriptome variability generated by gene-TE chimeric
390 transcripts, we analyzed the transcriptome of head and larval tissues in the six cactophilic
391 *Drosophila* species. The method employed in this work considered paired-end reads
392 spanning from TEs to exons. Although systematically found in all replicates, many chimeras
393 are likely to be products of pervasive transcription, generating ectopic transcripts whose
394 expression levels are insufficient to play a functional role. Therefore, we developed an
395 automatic method to assemble chimeric transcript isoforms and compute their expression
396 contribution relative to the total gene expression. This method is implemented in the
397 pipeline *ChimeraTE v2.0* (24) as an optional downstream analysis to the total chimeras
398 detection.

399 Our results revealed that, on average, 715 and 574 genes produce chimeric
400 transcripts across all species, corresponding to 7.51% and 5.04% of the expressed genes in
401 head and larvae, respectively. In both tissues, TE-exonized transcripts represented the
402 highest frequency, with 87.59% in head and 89.09% in larvae (Fig. 4A), followed by TE-
403 terminated transcripts with 10.63% and 9.45%, and TE-initiated transcripts with 1.79% and
404 1.51%. Interestingly, all species have a similar number of expressed genes (head: $\bar{x}$ = 9,508;
405 σ = 300; larvae: $\bar{x}$ = 11,375; σ = 117), but we observed on average 11-fold fewer chimeric
406 transcripts in *D. buzzatii* and *D. koepferae* compared to *D. arizonae* and *D. mojavensis*
407 subspecies. This result suggests that species from the *mojavensis* cluster are more prone to
408 generate chimeras, or that *D. buzzatii* and *D. koepferae* have a strong purifying selection
409 against gene-TE chimeric transcripts.

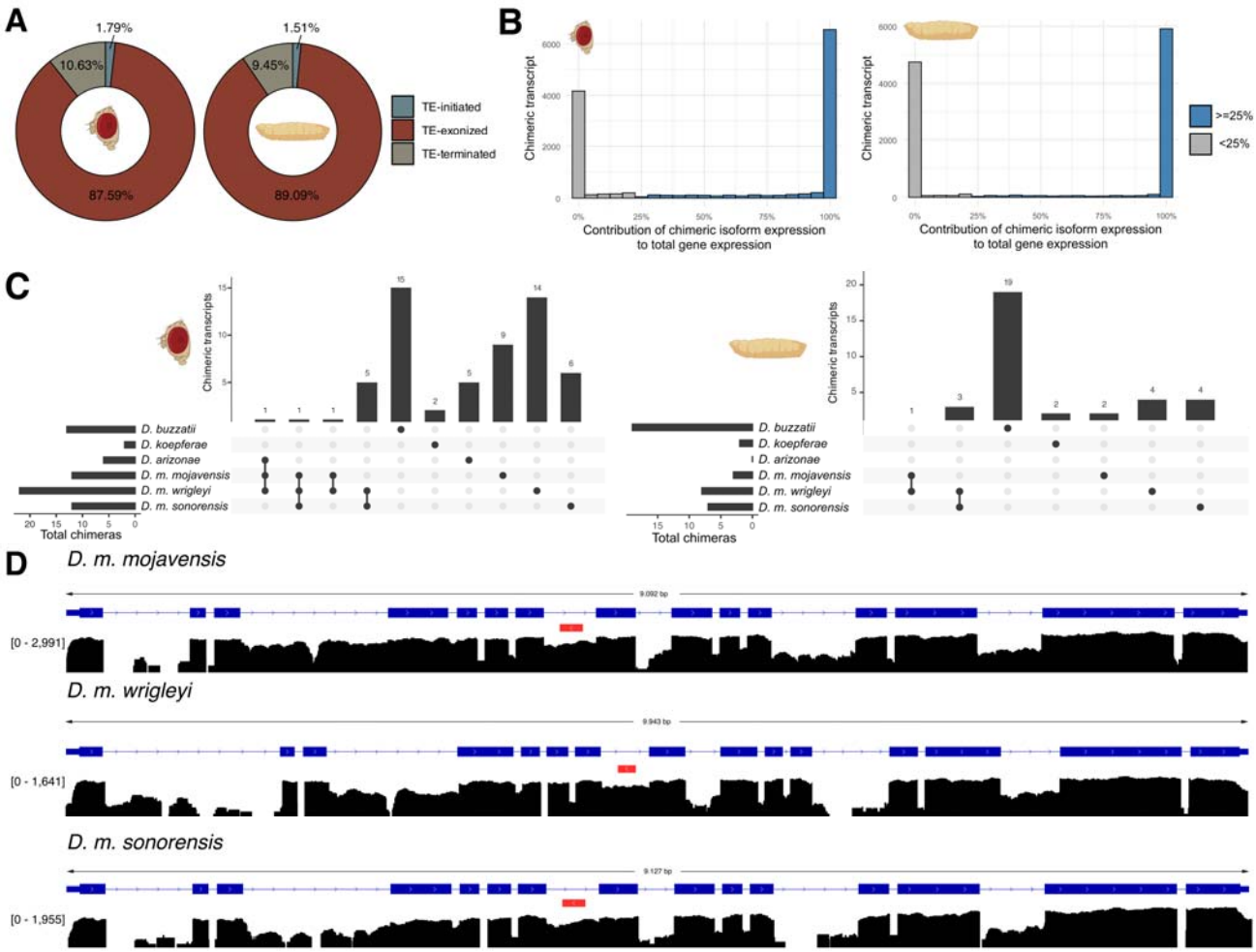


Figure 4: **A**) Proportion of each chimeric transcript category based on the position of TE insertions relative to the gene structure. **B**) Histograms with contribution of chimeric transcript isoform relative to total gene expression (x-axis). Chimeric transcripts without assembled sequence or contribution < 25% were removed (grey bars). **C**) Number of common and unique chimeric transcripts from HLAU genes detected in head and larvae. **D**) *Eato/LOC6578120* (ABC transporter) gene has a conserved TE-exonized transcript with a LINE/R1 element located in its 7th intron. The depth in exons, including the 7th intron, was assessed with RNA-seq from head tissue.

In head and larvae tissues, ~49% of the detected chimeric transcript isoforms were not assembled or had contribution to gene expression lower than 25% (Fig. 4B). Among these discarded chimeras, the vast majority was not detected by transcriptome assembly

421 approach (0% contribution) (Fig. 4B), which is likely to be associated with negligible
422 expression. Considering chimeric transcripts with contribution >= 25%, more than 95%
423 have contribution equal to 100% (Fig. 4B). This result indicates that almost all genes
424 producing chimeras are dependent on the TE sequence to constitute their complete mRNA
425 sequence. Since ~88% of all chimeras detected with > 25% contribution are from TE-
426 exonized transcripts, for which 72.35% are from TEs embedded in exons, a high frequency
427 of chimeras with total gene expression could be expected. Notably, this pattern is also
428 observed in TE-initiated and TE-terminated transcripts (Supplemental Fig. S6), despite the
429 TEs being located upstream and downstream of the gene body. Our results reveal a clear
430 binomial trend: chimeric transcripts either have low expression and are unlikely to be
431 biologically relevant, or they account for the total gene expression, suggesting a functional
432 role.

433 ***HLAU genes generate chimeric transcripts***

434 To investigate cases where TEs produce chimeric transcripts from HLAU genes, and
435 their possible role in the evolution of these genes, we identified all chimeras derived from
436 HLAU genes in head and larvae transcriptomes. We found a total of 28 (head) and 24
437 (larvae) HLAU genes producing chimeric transcripts. None of them are the six genes under
438 positive selection. Notably, ABCs, GSTs, and IRs were the most frequent gene families with
439 chimeric transcripts (Supplemental Table 10). Although *D. buzzatii* had 11-fold fewer
440 chimeras than species from the *mojavensis* cluster, we observed 29 gene-TE (a few genes

441 with more than one chimera) chimeras derived from HLAU genes in both tissues, the
442 highest number among all species. Its sister species *D. koepferae* had only four chimeras,
443 which are the same two ABC genes in head and larvae (Supplemental Table 10). In *D.*
444 *buzzatii*, ~93% of all chimeras are from GSTs, expressed in both larvae and head tissue.
445 Interestingly, all of them represent cases of *Helitron* exonizations (Supplemental Table 10).
446 We did not find any GST with chimeric transcripts in the other transcriptomes, revealing an
447 unprecedented recent evolutionary novelty in GSTs' evolution in *D. buzzatii*.

448 Chimeric transcripts from HLAU genes are represented by 14.81% TE-terminated
449 transcripts and 85.19% TE-exonized transcripts. We did not observe any TE-initiated
450 transcript from HLAU genes. Despite the compelling evidence of TEs enrichment in the
451 promoter of OR genes from species of the *mojavensis* cluster, we did not observe chimeric
452 transcripts from ORs at high frequency. Only *Or45b* (*LOC6578440*), *Or13a* (*LOC26527676*),
453 and *Or85a* (*LOC26528431*) had chimeric transcripts in *D. m. sonorensis*, *D. m. wrigleyi*, and
454 *D. arizonae*, respectively (Supplemental Table 10). This result reveals that TEs located in the
455 promoter of ORs are not likely to be co-opted as alternative transcription start sites. In the
456 species of the *mojavensis* cluster, we observed ABCs as the gene family with the highest
457 frequency of chimeras (Supplemental Table 10), despite not having enrichment (Fig. 2A).
458 We observed, on average, 5 ABC genes with exonization of multiple TE copies in this group
459 of species. This result indicates that the genomic distribution of TEs is not necessarily
460 associated with their transcriptional interaction with genes.

461 The quantitative analysis on chimeric transcripts revealed transcriptional
462 polymorphism in HLAU genes due to TEs. We investigated common chimeras present in
463 more than one species, highlighting their retention over evolutionary time, as well as
464 species-specific chimeras that may be associated with recent insertions. Our results
465 revealed that 94.44% and 88.57% of the chimeric transcripts derived from HLAU genes are
466 species-specific in head and larvae tissues, respectively (Fig. 4C). In head, the *Opuntia sp.*
467 dwellers *D. buzzatii* and *D. m. wrigleyi* were the ones with the highest number of HLAU
468 genes with species-specific chimeric transcripts (Fig. 4C). The 15 chimeras uniquely found in
469 *D. buzzatii* are derived from 12 genes (2 ABCs and 10 GSTs); whereas the 14 chimeras from
470 *D. m. wrigleyi* were observed from 10 genes (5 ABCs, 3 IRs, 1 CYP, and 1 OBP)
471 (Supplemental Table 10). In larvae, *D. buzzatii* follows a similar result as observed in head,
472 15 chimeras derived from 14 genes (2 ABCs and 12 GSTs), while *D. m. wrigleyi* had only four
473 unique chimeras (2 ORs, 1 IR, and 1 CYP).

474 Cases of common chimeric transcripts between populations/species may indicate
475 exaptation/domestication events. In the six transcriptomes analyzed here, we found 8
476 genes in the head and four in larvae present in more than one genome (Supplemental
477 Table 11). These common chimeric transcripts occurred only between *D. mojavensis*
478 subspecies, revealing that TEs interacting with HLAU genes are likely to be lost over higher
479 evolutionary scales. Their inclusion into the mRNA sequence and high contribution can be
480 observed. For instance, the ABC transporter (*Eato* gene) has an exonized LINE/R1, which is
481 located in the 7th intron (Fig. 4D), with an average contribution of 84.25% to the gene

482 expression. Taken together, these results demonstrate unprecedented transcriptional
483 interaction between TEs and HLAU genes in a tissue-specific manner.

484 **Discussion**

485 Insect host shift and its role in speciation have been the subject of research to
486 understand the evolution of reproductive isolating barriers and adaptation to local
487 environments. Indeed, many insects have finely tuned adaptations to locate, feed,
488 reproduce, and develop on the host tissue. Adaptations to new hosts can lead to
489 reproductive isolation, wherein ecological speciation arises as a consequence within
490 divergent natural populations inhabiting distinct environments. In this context, we
491 investigated whether TEs could play a role in contributing to the evolution of HLAU genes,
492 with the potential to facilitate rapid adaptive radiation to new hosts. To test this hypothesis,
493 we selected six *Drosophila* species/subspecies from the *repleta* group with different primary
494 cacti hosts.

495 ***Weak signal of positive selection in HLAU genes through branch-site model***

496 The necrotic tissue of each cactus-host used by cactophilic *Drosophila* species differs
497 in its composition in terms of yeasts and alkaloids (45,46). Therefore, species have adapted
498 to feed and breed on different hosts, as demonstrated by behavioral and physiological
499 response when species feed and breed in non-preferential hosts. Positive selection in HLAU
500 genes have been proposed to allow the recognition of a broader spectrum of odors,

501 metabolize nutrients, and activate detoxification pathways (12). Indeed, specialist species
502 are likely to lose some of their genes related to sensory pathways, while the genes that are
503 retained are subject to positive selection (47,48). A preliminary genome-wide analysis of
504 signatures of positive selection in the four *D. mojavensis* subspecies revealed high ω rates
505 on genes associated with chemosensation, perception, immunity, behavior, detoxification,
506 and reproduction (49). Although a previous genome-wide analysis has identified 1,294
507 genes under positive selection (50), here we focused the analysis to investigate specifically
508 HLAU gene families. Our results with brach-site model had only six HLAU genes under
509 positive selection: *Obp19d, Obp56d, Gr5a*, *Ir25a*, *Cyp4s3*, *esterase S*. The unprecedent
510 method used in this work focused on the common differences of species adapted to
511 columnar cacti, rather than independent evolutionary events, making it more stringent. The
512 lack of genes with statistical support for positive selection may be also associated with
513 technical limitations, since we based our analysis on 1:1 orthologs, and a certain proportion
514 of the genes evolve too fast to be correctly assigned during orthologs identification (51).
515 Additionally, we observed ω values higher than 1.5, for instance as *Gr59d* with ω = 3.49.
516 Although we did not observe signficiant positive selection, this result indicates ongoing
517 divergence between species/subspecies.

518 ***OR promoters are enriched with TE insertions***

519 The literature has recurrently demonstrated that TEs can create new regulatory
520 elements, including gene promoters (52–54), enhancers (55,56), and insulators (57,58). In

521 mammals, TE-derived sequences contribute up to 40% of genome-wide binding sites for
522 transcription factors located in TE-rich regions (59). It has been proposed that such co-
523 option of TE-regulatory elements for gene regulatory networks can provide substantial
524 modification of gene expression over short time scales, contributing to genetic variability at
525 the transcriptome level (60–62). The colonization of new habitats and hosts is accompanied
526 by behavioral changes, mainly through alterations in environmental perception (63,64).
527 Here, we observed that ORs are enriched by TEs in the four species from the *mojavensis*
528 cluster. In ants, ORs have also been found enriched in TEs. But in ants, TEs are associated
529 with tandem duplication by crossing over, allowing massive duplication events (65). Since
530 there is no expansion of ORs in cactophilic species, the enrichment in TEs is likely to have
531 another function, probably related to gene regulation due to its specific location in
532 promoters. Although we did not observe a significant enrichment of TEs in differentially
533 expressed genes, we argue that TEs might contribute to the regulation of specific HLAU
534 genes. In ORs, ~39% of the genes have OR's TFBSs located within TEs. We propose that this
535 subset of genes is not sufficient to have a significant signal when tested with ORs without
536 TEs, but they may have functional TE-derived *cis*-regulatory elements.

537 *Helitrons* have been extensively reported on the promoter of other eukaryotic
538 species (41,66–68), in which a few cases of embedded TFBSs were reported (69,70). Here,
539 we demonstrated that ~50% of the enrichment in TEs observed at the promoter region of
540 ORs might be due to the exaptation of TFBSs derived from *Helitrons*. In addition, the
541 recurrence of the enrichment in species of the *mojavensis* cluster might demonstrate

542 potential cases of TE co-option inherited by the common ancestor. Our genomic analysis
543 provides a robust list of potential candidate genes in which *Helitrons* might play a
544 regulatory role. This is an interesting hypothesis that can be determined through cutting-
545 edge technologies such as CRISPR methods to delete insertions with site-specific resolution
546 (71). Our results suggest a putative role of TEs in the regulation of HLAU genes across
547 multiple species/subspecies.

548 ***Head transcriptome reveals subspecies-specific patterns***

549 The differential expression of HLAU genes in the head tissue revealed that
550 localization and host usage (detoxification) gene families have the highest divergence
551 across the subspecies, since we observed more differentially expressed genes for these
552 families. In *D. m. wrigleyi*, which is the only *D. mojavensis* subspecies with a preference for
553 *Opuntia sp*., we observed that CYP genes had on average 12 up-regulated genes compared
554 to other species. This result highlights that the divergence of the cacti preference between
555 *D. mojavensis* subspecies might be explained by modifications in gene expression of HLAU
556 genes in the head tissue. Although *Opuntia sp.* has lower toxicity than columnar cacti (72),
557 we found a similar number of up-regulated detoxification genes between *D. m. wrigleyi*
558 and the other subspecies. This suggests that *D. m. wrigleyi* has a constitutive detoxification
559 response that is independent of host toxicity. This is a plausible explanation since *Opuntia*
560 *sp.* produces fewer benzaldehyde volatiles than the columnar cacti *S. thurberi*, but both
561 cacti are lethal to *D. melanogaster* after 48 hours (72). Therefore, the adaptation to use

562 *Opuntia sp*. must have a constitutive gene expression of detoxification pathways in *D. m.*
563 *wrigleyi*.

564 Comparative transcriptomic studies have revealed divergence in gene expression
565 between *D. buzzatii* and *D. koepferae*, particularly in head tissues associated with sensory
566 perception and detoxification. For instance, Guillén et al. (2014) reported significant
567 differences in the expression of genes involved in olfaction and xenobiotic metabolism,
568 suggesting that these species have adapted to their respective cactus hosts (*Opuntia* for *D.*
569 *buzzatii* and columnar cacti for *D. koepferae*). Similarly, divergence in expression patterns
570 has been reported in these species in ORs, OBPs, and cytochrome P450 genes (73). These
571 reported differences are consistent with our results in head and larval tissue, reinforcing the
572 hypothesis that ecological specialization and divergence in head transcriptomes play a role
573 in host shift between *D. buzzatii* and *D. koepferae*.

574 Previous studies with *D. melanogaster* have shown that sequence divergence in ORs
575 and OBPs significantly modulates odor sensitivity (64,74), potentially leading to divergence
576 of host location. In addition, many ORs are involved in a broad range of ecological
577 interactions, influencing behaviors associated with oviposition and feeding (75). For
578 instance, OR evolution has mediated the herbivorous nature of leaf-mining specialist
579 *Scaptomyza flava* (76). Furthermore, transcriptomic analyses have identified 18 ORs that are
580 differentially expressed between heads of *D. m. wrigleyi* and *D. m. mojavensis*,
581 demonstrating that modifications in the sensory transcriptome have evolved between these

582 species (77). In this study, we expanded the understanding of this divergence by comparing
583 species with a preference for *Opuntia sp.* and columnar cacti, and showed that each species
584 has a distinct set of differentially expressed HLAU genes.

585 ***Gene expression in larvae is one of the factors shaping the adaptation to new hosts***

586 Differences in larval behavior and survival have been observed when *Opuntia*-related
587 species are reared in columnar cacti (78,79). A previous study showed that *D. m. mojavensis*,
588 which primarily uses *S. thurberi* as its host, has a 60% reduction in larval viability when
589 reared on *S. gummosus*-based media (80). The authors also proposed that 21% of the
590 genes are differentially expressed between larvae reared in the primary and alternative cacti
591 hosts. Many genes were reported with detoxification, energy production, among others.
592 Another similar study with larvae transcriptome, but with *D. buzzatii* and *D. koepferae*,
593 proposed that *D. buzzatii* flies reared in columnar cacti media have a stronger
594 detoxification response in comparison with *D. koepferae* flies, suggesting that *D. koepferae*
595 has a canalized transcriptome response to the toxic alkaloids from columnar cacti (81).

596 Here, we observed that the majority of differentially expressed HLAU genes are from
597 gene families associated with detoxification. Since *D. m. wrigleyi* and *D. buzzatii* use
598 primarily *Opuntia sp.*, we expected to find a higher number of detoxification genes in the
599 other species/subspecies with columnar cacti preference. However, we observed a similar
600 number of differentially expressed detoxification genes regardless of the host preference.
601 CYPs were overrepresented in our results, but they represent a diverse gene family that can

602 have functions other than detoxification, such as those related to developmental pathways
603 in larvae (82,83). Thus, further analysis of the biological role of up-regulated CYP genes in
604 *D. m. wrigleyi* and *D. buzzatii* might be assessed. Alternatively, as we observed in the head
605 transcriptome, although *Opuntia sp.* has lower toxic compounds compared to columnar
606 cacti, *D. m. wrigleyi* and *D. buzzatii* may also have a resistance to toxicity (72). Thus, any
607 cactophilic species is expected to have evolved a certain level of detoxification ability, which
608 explains our observations. In this study, we propose a set of common HLAU genes in
609 columnar cacti dweller species associated with detoxification. We observed that the three
610 *D. mojavensis* subspecies have the same seven genes with higher expression compared to
611 *D. m. wrigleyi*. Most of them are associated with detoxification (*EstB1*, *UDP2*, *UDP-Ugt5*,
612 *Cyp9b2*, *UDP-Ugt5-1*, *Cyp12a2*), except *Ir21a*. It is noteworthy that other evolutionary
613 forces than gene expression might be associated with host shift. For instance, gene
614 duplication events have been reported in CYPs and GSTs in cactophilic species (84). These
615 duplications are thought to enable flies' resistance to different toxic environments.
616 Similarly, *D. buzzatii* and *D. koepferae* show both copy number variation and expression
617 divergence in detoxification genes, indicating that gene duplication followed by regulatory
618 or coding sequence divergence contributes to host-specific physiological adaptations
619 (50,84). These findings support the view that gene duplication serves as a key evolutionary
620 mechanism facilitating ecological specialization in cactophilic *Drosophila*. Our results show
621 that divergence in the regulation of gene expression in larvae is one of the factors likely to
622 shape the adaptation to new hosts.

623 ***Transposable elements are a source of transcriptome variability between cactophilic***
624 ***species***

625 TEs contribute to genome and transcriptome evolution by rewiring regulatory
626 networks, and by incorporating protein domains into gene transcripts. In both cases, TE
627 copies that are inserted within genes can integrate their sequences into the mRNA,
628 generating chimeric transcripts. A recent study with different ecotypes of *D. melanogaster*
629 has shown that these chimeras have the potential to generate transcriptome novelties (24).
630 These chimeras have also been demonstrated to be active in a tissue-specific manner (85).
631 Although the production of chimeric transcripts derived from polymorphic TE insertions is
632 well-known, there is no study addressing the question of how TEs contribute to the
633 evolution of genes associated with host shift. Here, we took advantage of genomic and
634 transcriptomic data to analyze the transcriptional interaction of TEs with HLAU genes
635 through gene-TE chimeric transcripts. Our finding demonstrated that around half of the
636 chimeras have low expression levels (0-5%), and are likely to represent pervasive
637 transcription of TEs. Notably, over 95% of the chimeras overpassing the 5% contribution to
638 gene expression constitute nearly the total gene expression.

639 Considering previous findings in different tissues of *D. melanogaster*, where 38% of
640 all chimeric transcripts express only the chimeric isoform (85), our result shows a higher
641 prevalence of chimeras in the repertoire of gene splicing. In tetrapods, the exonization of
642 TE-derived transposases has been shown to constitute the unique gene isoform (86). In

643 HLAU genes, all chimeric transcripts represent TE-exonized transcripts. We suggest that TEs
644 in the promoters of ORs might still recruit transcription factors at their terminal sequences,
645 which would prevent their inclusion in the gene mRNA (chimera). In addition, the
646 maintenance of TEs in the promoters of ORs might also be due to epigenetic marks. TEs
647 can have chromatin marks (87,88), making them versatile regulatory motifs acting in the
648 regulation of nearby genes (89,90). Further experimental analysis, as ChIP-seq, may be
649 performed to test the TEs' function in OR's transcription. Furthermore, TE annotation
650 remains challenging in non-model organisms. Although we performed several steps to
651 remove potentially false TE copies from the genome, we can not rule out the possibility of
652 remaining annotation errors in the exons of HLAU genes.

653 A comparative analysis between generalist and specialist *Drosophila* species revealed
654 that niche amplitude is not likely to play a role in TE dynamics in the genomes (91).
655 Nevertheless, several studies have proposed that the successful adaptation to new
656 environments in invasive species might be associated with TE activity, despite reduced
657 genetic diversity caused by founder effects (92,93). Here, we demonstrate that TEs interact
658 with genes associated with HS, providing a transcriptomic variability in six cactophilic
659 species with different host preferences. Our results show that ABC transporters produce
660 chimeric transcripts in all cactophilic species, highlighting a compelling mechanism of
661 structural innovation. In insects specifically, MITEs (non-autonomous DNA transposons)
662 have been repeatedly identified within defensome genes, including ABC transporters, such
663 as in *Helicoverpa* spp., suggesting a potential causal link to insecticide resistance (94).

664 Functionally, exonized TEs are stably expressed and translated, often producing altered
665 protein domains or localization signals, thereby expanding proteomic repertoires while
666 providing raw material for adaptive selection (95). Taken together, our findings suggest
667 that TE exonization in ABC transporters may offer cactophilic *Drosophila* a versatile
668 genomic strategy to fine-tune transporter function, potentially under the selective
669 pressures of xenobiotics. Although we found a high contribution of the chimeric isoform to
670 total gene expression, future work should be conducted to elucidate which exonized TEs
671 are translated, and whether they modulate substrate specificity, protein stability, or cellular
672 localization, thereby affecting organismal fitness.

673 Our discovery of 13 GST genes in *D. buzzatii* harboring exonized Helitrons
674 underscores an unprecedented finding of TE-driven functional diversification in
675 detoxification enzymes. *Helitrons* have been shown to supply novel exons, splice sites, and
676 promoters across eukaryotes (41). This finding, unique to *D. buzzatii,* suggests a lineage-
677 specific co-option of *Helitrons* to expand and diversify GST isoforms. GSTs are central to
678 xenobiotic metabolism, and exonization of Helitron-derived sequences could introduce
679 novel domains or alter enzymatic properties, potentially conferring adaptive advantages
680 under environmental pressures such as exposure to cactus-derived toxins. Future work
681 should be performed to determine the function of these *Helitron*-derived motifs in the host
682 preference of *D. buzzatii*.

683 It is important to state that the identification of chimeric transcripts through
684 transcriptome analysis is an important, but limited step in terms of function. These mRNA
685 molecules might be degraded by surveillance pathways that control for aberrant mRNA
686 production, such as no-go decay (96), non-stop decay (97), and nonsense-mediated RNA
687 decay (98). Our study provides a set of HLAU genes producing chimeric transcripts, in
688 which part of them might have a phenotypic impact and ultimately be associated with host
689 preference. However, it is fundamental to perform comparative phenotypic studies of host
690 adaptation cues with mutant strains for the TEs interacting with HLAU genes identified in
691 this study. The application of techniques of RNAi aiming at chimeric transcripts, or deletion
692 of TE insertions related to chimeras with CRISPR/Cas9 (71) would be crucial to confirm the
693 predictions of our results.

694 ***Transposable elements driving host shift***

695 Host specificity is a key mechanism of reproductive isolation in plant pathogens and
696 is regulated by the repertoire of effector genes within each pathogen. In *Phytophthora sp.*,
697 the host specificity is partly regulated by RXLR class effectors that facilitate host
698 exploitation. Notably, synthetic chimeras of a short interspersed element (SINE) linked to an
699 effector gene in *P. infestans* induced their silencing (99). This silencing likely occurs
700 naturally, as transcriptional inactivation of effectors is known to occur, and over half of
701 RXLR effectors in the *P. infestans* genome are located in TE-rich regions (100).
702 Consequently, TEs inserted near these genes may have influenced host shift in *P. infestans*.

703 Our findings demonstrate multiple interactions between genes and TEs, highlighting the
704 contribution of TEs to the genome and transcriptome evolution of cactophilic species.
705 Although our data did not allow the quantification of the TE insertions frequency in natural
706 populations, our results demonstrate the first transcriptome-wide evidence of TE co-option
707 in cactophilic species. The genes producing chimeras with TEs must be further studied,
708 since the lack of gene-phenotype association between HLAU genes is not fully reported,
709 preventing us from pinpointing a specific relation between chimeras and possibly host
710 preference. In addition, our study focused exclusively on nine gene families that have been
711 shown to evolve during host shift events (12). However, host shift is undoubtedly a complex
712 trait to study, and many other genes can be associated with the adaptation to new hosts.
713 For instance, differential expression of genes associated with development and
714 neurological processes has been identified in cactophilic species reared in different cacti
715 media (101). Behavioral changes are also based on host shift evolution and have been
716 reported in many systems (102). Despite focusing on nine relevant gene families, we
717 suggest further studies including other aspects of host shift than localization, acceptance,
718 and host usage.

719 The existence of reproductive barriers between populations that undergone a host
720 shift is the key process to ecological speciation. In our model, with species from the *buzzatii*
721 and *mojavensis* clusters, incipient reproductive isolation has been reported between *D. m.*
722 *sonorensis* and *D. m. baja* (26,103,104). The premating isolation between them occurs due
723 to the difference in cuticular hydrocarbons, which participate in the signaling pathway of

724 mate recognition (105). Importantly, artificially shifting the cactus host was the main cause
725 of changes in the hydrocarbon profiles. Thus, the source of reproductive barriers is
726 dependent on the cacti species on which the flies feed and breed (106,107). The close
727 relationship between HLAU genes to the host shift process, and the observed contribution
728 of TEs to the evolution of these genes, supports a possible role of TEs in the host
729 preference of cactophilic species.

730 **Methods**

731 ***Fly stocks and genome sequencing***

732 Most of the strains were obtained from the UC San Diego *Drosophila* Stock
733 Center: *[D. m. mojavensis*: Anza (01), Anza Borrego Desert, California, USA, Stock Center
734 n°: 15081-1352.01; *D. m. wrigleyi*: CI (22), Catalina Island, California, USA, Stock Center
735 n°: 15081-1352.22; and *D. m. sonorensis*: AG (26), Agiabampo Bay, Sonora, México, Stock
736 Center n°: 15081-1352.26] and *D. arizonae* [HI (17), collected in Metztitlan, Hidalgo, Mexico,
737 Stock Center n°: 15081-1271.17]. *D. buzzatii* (Bu28) and *D. koepferae* (Ko2) stocks
738 correspond to two laboratory lines used in previous works (108,109). Both stocks were
739 maintained by brother-sister mating for more than a decade and then kept by mass
740 culturing.

741 We used the same protocols of DNA extraction and genome sequencing for all
742 samples. DNA was extracted from 10 males and 10 females from each species using the
743 Qiagen DNeasy Blood&Tissue kit. Then, we evaluated the genomic DNA amount and

744 quality with NanoDropTM One UV-Vis spectrophotometer (Thermo Fisher Scientific,
745 Waltham, MA, USA) and Qubit ® 1.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Three
746 micrograms of DNA were repaired using the NEBNext FFPE DNA Repair Mix (NEB M6630).
747 We performed end repair and dA-tailing using the NEBNext End repair/dA-tailing Module
748 (E7546, NEB). Ligation was then assessed with the Ligation Sequencing Kit 1D. MinION
749 sequencing was performed according to the manufacturer's guidelines using R9.4.1 flow
750 cells (FLO-MIN106, ONT) and a Nanopore MinION Mk1b sequencer (ONT) controlled by
751 *ONT MinKNOW v.18.3.1*. Base calling was performed after sequencing using *Guppy v.4.0.5*
752 in high accuracy mode for isogenic wild-type strains *v.3.1.5*.

753 ***Genome assemblies and gene annotation***

754 The quality control for Nanopore reads was performed with *Nanoplot* v.1.10.2
755 (https://github.com/wdecoster/NanoPlot). Reads with quality lower than 7 were removed
756 from downstream analysis. In order to assemble contigs, *Flye v.2.8* (110) was used with
757 default parameters, except –plasmids. All contigs were aligned with *minimap2* v2.16 (111),
758 with –x map-ont option. The alignment was used to perform the polishing of contigs with
759 four rounds of *RACON* v1.3.2 (112), default parameters. The assembly quality metrics was
760 assessed with *Assembly-Stats* v1.0.1 (https://github.com/sanger-pathogens/assembly-stats).
761 Assembly incongruences were manually visualized with *D-genies* v1.2.0 (113) and corrected
762 with *samtools* v1.9.0. (114), function *faidx*, and *Gepard* v1.4.0(115) for determination of
763 breaking points. The super scaffolding of all corrected assemblies was performed with

764 *RaGOO* v.1.1 (116), with –s and –t 4 parameters, using the respective reference genome
765 assembly for each species. Subsequently, the chromosome-scale assemblies were
766 submitted to a Benchmarking Universal Single-Copy Orhologs (BUSCO) v.5.4.3 analysis
767 (117) with lineage *diptera_odb10* (*-l* parameter) with nucleotide sequences. The gene
768 annotations for *D. mojavensis* subspecies and *D. arizonae* were performed with the
769 software *Liftoff v.1.6.3* (118), with the reference genome of *D. mojavensis*
770 (GCA_018153725.1) (119). We have used the parameters *–exclude_partial*; *-a 0.5* and *–s 0.75*
771 to keep annotated genes with at least 75% of the reference gene length. We removed from
772 our gene annotation HLAU genes flagged as "low quality protein" and "unestablished gene
773 function" in the reference *D. mojavensis* genome.

774 *D. buzzatii* and *D. koepferae* genomes were annotated with a de novo strategy with
775 the software *BRAKER3 v.3.0.8* (120). To optimize the prediction of gene structures and
776 repertoire of isoforms, we used a mix of RNA-seq data from different tissues publicly
777 available and produced in this work (Supplemental Table 12), in addition to protein
778 sequences from *D. mojavensis* as reference. Ortholog genes between *D. buzzatii, D.*
779 *koepferae*, *D. mojavensis,* and *D. arizonae* were obtained with *Orthofinder v.2.5.4* (121) using
780 *DIAMOND v.0.9.14* (122) as search engine and default parameters. One-to-one orthologs
781 were obtained directly from *OrthoFinder v.2.5.4*, and remaining proteins without an
782 ortholog pair were analyzed with *blastp v.2.5.0+* (123), for which we considered as one-to-
783 one proteins only those with a single homologous with parameters *-qcov_hsp_perc 95* and
784 identity >= 90%. Since *Orthofinder* performs its search with protein sequences, the

785 identification of orthologs from non-coding RNA genes was assessed with *Liftoff v.1.6.3*
786 (124) using all non-coding RNA genes from *D. mojavensis* as reference to maximize our
787 homology-based gene annotation.

788 ***Molecular evolution analysis***

789 To identify HLAU genes with signatures of positive selection in species that prefer
790 columnar cacti compared to those that prefer *Opuntia sp.*, we selected the three *D.*
791 *mojavensis* subspecies, *D. navojoa*, *D. koepferae,* and *D. buzzatii*. *D. arizonae* species was
792 removed from this analysis due to its generalist preference for both columnar and *Opuntia*
793 cacti. The *D. navojoa* genome was retrieved from NCBI (GCF_001654015.2) (125). For the
794 other Nanopore genomes sequenced in this work, we performed a polishing step with
795 Illumina paired-end reads. We used the software *NextPolish* (126) to fix base errors in the
796 assembled genomes. The three *D. mojavensis* subspecies were polished with genomic
797 Illumina reads sequenced previously (127), whereas for *D. koepferae* and *D. buzzatii,* we
798 used publicly available genomic Illumina reads (128) and RNA-seq reads (50), respectively.
799 In the latter, to increase the read coverage and optimize the polishing to the highest
800 number of genes and exons as possible, we merged RNA-seq reads from adult males and
801 females (50). All genomes were corrected using two rounds of polishing, as recommended
802 (126). Then, we extracted the longest CDS sequences for all genes combining AGAT (129)
803 *agat_extract_sequences.pl* parameters *-t mrna*, *samtools faidx* (130), and *seqtk*
804 (https://github.com/lh3/seqtk). Posteriorly, we intersected the occurrence of orthologous

805 genes between the six genomes, recovering a list of genes that were present in all of them.
806 The sequences were aligned using *MAFFT v.7.487* [67] and the non-aligned codons were
807 removed using *pal2nal* (131). To obtain a strongly supported tree for this analysis, we
808 inferred the species phylogeny tree with 2,043 shared BUSCO genes. The phylogeny was
809 estimated using the concatenated alignment (3,730,914 sites) with *IQ-TREE v. 2.2.2.6* (132),
810 using the best model for each gene, 10,000 replicates of ultrafast bootstrap and 1000
811 replicates for Approximate Likelihood-Ratio Test (ALRT) (parameters *-m MFP -bb 10000 -*
812 *alrt 1000*). Finally, we carried out an analysis to detect signatures of positive selection for
813 each gene individually with *CODEML* from *PAML v.4.10.7* (133), using *branch-site* approach.
814 In order to find signatures of positive selection in species that use columnar cacti, we
815 marked the respective branches of *D. koepferae*, *D. m. mojavensis,* and *D. m. sonorensis* with
816 the same label (#1), to compare them with the background species that use *Opuntia sp.* (*D.*
817 *buzzatii*, *D. m. wrigleyi,* and *D. navojoa*). For this, we tested two hypotheses: (i) H0 – Null
818 hypothesis, in which all branches evolved under negative selection (model = 2, Nssites = 2,
819 fix_omega = 1, omega = 1); (ii) H1 – Alternative hypothesis, in which the labeled branches
820 (foreground) have sites with signatures of positive selection (model = 2, NSsites = 2,
821 fix_omega = 0, omega = 1.5). The hypotheses were tested through a chi-square test,
822 calculated by the log-likelihood (lnL) ratio between H1 and H0 (2ΔlnL) and degrees of
823 freedom obtained from the difference between the number of estimated parameters for H1
824 and H0 (Δnp). Sites under positive selection were identified by Bayes Empirical Bayes and
825 posterior probability, implemented in *CODEML*.

826 ***RNA extraction and sequencing***

827 We performed RNA-seq for larvae and head tissues from *D. buzzatii*, *D. koepferae*,
828 *D. arizonae,* and the four *D. mojavensis* subspecies. The RNA extraction was assessed with
829 the RNeasy Plus Kit for 10 female heads (10 days old) and 15 larvae, with three technical
830 replicates. The library preparation was done with stranded mRNA-seq/standard quantity
831 (ligation), and 100 bp paired-end reads were produced in an Illumina HiSeq 4000. Illumina
832 adaptors and low-quality reads were removed for downstream analysis with *Trimommatic*
833 *v.0.39* (134), parameters*: LEADING:3*; *TRAILING:3*; *SLIDINGWINDOW:4:15*; *HEADCROP:10*;
834 *CROP:90*.

835 ***TE annotation***

836 For each sequenced genome, we developed a novel pipeline combining automatic
837 and manual approaches to construct TE consensuses. First, we used *Earl Grey v.2.2* (135) to
838 build de novo TE libraries with *RepeatModeler2 v.2.0.4* (136), which uses *RECON v.1.08* (137),
839 and *RepeatScout v.1.0.6* (138). We used the following parameters: -r 32281 (NCBI taxon ID
840 for *Drosophila*), -c yes (cluster redundant consensus), -m yes (remove consensus with less
841 than 100 nt), and -e yes (enable *HELIANO v1.2.1* (139) to identify Helitrons). In short,
842 EarlGrey automatically retrieves TE consensus, and performs an extension of each sequence
843 using a method named "Blast, Extract, Extend" (BEE) (140). Such method extracts the
844 sequences of the 20 full-length (or longest) TE insertions for each consensus, and then it
845 retrieves the 1 kb flanking regions from them. Posteriorly, insertions altogether with

846 flanking regions are aligned with *MAFFT v.7.487* (141), and the alignment is trimmed with
847 *trimAl v.1.4* (142), parameters *–gt 0.6* and *–cons 60*. Subsequently, the alignment is analyzed
848 to check whether the flanking regions from the full-length insertions have high quality,
849 meaning that they are part of the TE rather than genomic regions. Finally, the consensus
850 sequence for each TE family is updated with *EMBOSS v.6.6.0.0* (143). The BEE method is
851 repeated up to 10 times for each consensus.

852 Once the extended TE libraries were obtained for the six genomes with EarlGrey, we
853 implemented our pipeline into eight steps. Firstly, we removed all sequences other than TEs
854 from EarlGrey's library: satellites, low complexity repeats, tRNA, and rRNA. At step 2, we
855 checked the presence of redundant consensus respecting the 80-80 (35) rule with all-vs-all
856 blastn (-qcov_hsp_perc 80 and -perc_identity 80). Surprisingly, ~45% of the consensuses
857 were duplicated. Some of them were identical, or nearly identical, but in opposite strands.
858 We selected the longest consensus from the ones considered redundant. At step 3, we
859 used Tandem Repeat Finder (144) to identify consensuses that were had over 50%
860 annotated as tandem repeats. Consensus falling in this cutoff were removed for being
861 indistinguishable from a tandem repeat. At step 4, we removed all consensus with high
862 similarity with *D. mojavensis wrigleyi* CDSs using blastn, parameters *-qcov_hsp_perc 80 and -*
863 *perc_identity 80*. TE consensuses that had passed our filtering thresholds were then
864 submitted to the classification analysis. At step 5 of the pipeline, TE consensus will be
865 classified at the family level based on similarity analysis of non-reference consensus with a
866 database provided by the user (https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-

867 Drosophila-Species). In our case, we used all consensuses from *Drosophila* species available
868 on Dfam v3.7 (145). Due to the "curated" Dfam library containing only *D. melanogaster*
869 consensus, we used the "uncurated" library from Dfam. We extracted *Drosophila* TEs with
870 famdb.py v.0.4.3, parameters -i Dfam.h5 families --include-class-in-name -f fasta_name -ad
871 'Diptera'. Then, we filtered out non-Drosophila sequences and "Unknown" TEs with the Unix
872 command line: grep -i 'drosophila' file.fa | grep -v -i 'unknown', obtaining 49,916 out of
873 344,562 consensus. The pipeline firstly classifies at the family level by doing a blastn (-
874 qcov_hsp_perc 80 and -perc_identity 80) between the *Drosophila* Dfam consensus and the
875 provided de novo consensus. Approximately 50% of the identified consensus were
876 classified on the family level based on this blastn. The consensus without results from this
877 stringent blastn and smaller than 200 nt were filtered out. Due to the stringency of the
878 matches with blastn, we also masked the remaining unclassified consensus with the
879 *Drosophila* Dfam consensus with RepeatMasker .v4.1.8 (146). Based on RepeatMasker
880 output, we computed the length of all matches from a given family in the unclassified
881 consensus, and how much it has been covered based on the unclassified length. If the
882 coverage is > 80% of the total length, we assume that the unclassified consensus has
883 diverged from that one of the database, but it still classified as the same matched family.
884 This step classified ~25% of our consensuses. Finally, the remaining unclassified sequences
885 that could not be classified at the family level are classified at the class/order level. We used
886 a cutoff of 50% of covered matches in the unknown consensus to classify their classes. As
887 the family name, we substituted the pattern from RepeatModeler2 as (i.e. rnd-1_family-1)

888 to a pattern representing: "species preffix", "author initials", "homology percentage" (i.e.
889 Dari_DSO_86.0275p#DNA/TcMar-Tc1). In this example, "Dari" stands for *D. arizonae*, "DSO"
890 is the author's initials, and "86.0275p" represents the percentage of the total length that
891 this consensus has been matched with a DNA/TcMar-Tc1. The remaining putative TE
892 consensuses with less than 50% of coverage with known *Drosophila* TEs were removed. At
893 the end of this step, we end up with polished TE consensus.

894 In step 6, we checked the strandness of TE consensus sequences based on stranded
895 RNA-seq data. During the clustering of similar sequences and removal of redundancy, we
896 noticed many consensuses with nearly identical sequences, but in different senses. To
897 correct TE strandness, we used our stranded RNA-seq data. We annotated TE insertions in
898 the genome with *RepeatMasker v.4.1.4* (122), parameters *-cutoff 250*, *-s*, and *-norna*. Then,
899 we mapped the reads from a mixed library from head and larvae tissues with STAR v2.7.11b
900 (147). To be certain about the strandness origin of the reads, we removed all insertions that
901 have any overlap with other TEs and/or exons. Then, for each consensus, the pipeline
902 identifies the three copies in the genome with higher expression. Based on these copies,
903 the number of uniquely mapped reads to the same strand of the copy is quantified. If >
904 50% of the reads align to the same strand of the copies, we assume the consensus
905 strandness was right. Otherwise, the pipeline converts the consensus sequence to its
906 reverse complement. This step will add a strandness tag to the consensus ID. For instance,
907 if 75% of the reads support forward strand, a tag "75st" will be added to the family name.
908 This is useful to provide a metric of certainty about the strandness of each consensus. Such

909 information can be valuable when working with piRNAs, where alignment strandness is
910 crucial to interpret the results. Then, we used *RepeatCraft* (148) with *LTR Finder v.1.2* results
911 to merge LTRs and their respective internal regions. Subsequently, we removed all TE
912 insertions shorter than 80 nucleotides, following the classification rule 80-80-80 (35).
913 Finally, we performed a last filtering step in the TE annotation to remove insertions
914 composed of simple sequence repeats. We used *TRF v.4.09* (144) in all TE insertions from
915 each genome, with parameters *2*, *5*, *6*, *75*, *20*, *50*, *500*, *-m*, *-h*, to remove TEs with masked
916 repeats over more than 50% of their sequences. This method has been previously used to
917 remove TE copies that are indistinguishable from simple repeats (24,149).

918 ***TE enrichment at regulatory regions***

919 To identify TE insertions within and nearby genes for each genome, we created *bed*
920 files with the genomic positions of all insertions. Then, we created *bed* files for the gene
921 regions corresponding to the 2 Kb upstream of the transcription start-site (TSS). To test
922 whether the prevalence of TEs in each gene region was significantly higher than expected
923 (enriched) compared to random genomic distributions of TEs in the same region of other
924 genes, we performed permutation tests with the R package *regioneR* (150), parameter
925 *randomization* = *resampleRegions*, with 2,000 random samples. For each gene family with *n*
926 genes, the number of genes with TEs was compared to other random *n* genes 2,000 times.
927 This analysis allowed us to infer whether there are HLAU gene families with the observed
928 frequency of genes with TEs at the promoter region higher than expected compared to the

929 rest of the genome. We also selected zinc-finger proteins and serine/threonine kinases as
930 two non-HLAU gene families to compare our findings. Taking into account that TEs might
931 harbor regulatory and/or coding motifs in both strands (62), our analysis considered TEs
932 within the investigated gene regions regardless of the strand.

933 ***Identification of TE-rich genomic regions***

934 We split the genomes into bins of 200 kb, generating a bed file with such intervals
935 across the genomes. Then, we used the function *intersect* from *bedtools v.2.30.0* (151) to
936 identify the number of TE insertions located in each genomic bin. Posteriorly, we merged all
937 overlapped TE insertions into single ranges. These ranges were summed to obtain the total
938 TE content in bp for each 200kb bin. By plotting the TE content of each genome, we
939 noticed that bins with more than 50kb of TE content were outliers of the normal
940 distribution. Therefore, we defined TE-rich regions as all bins with > 50kb of TE content,
941 which represents 25% of the bin size. We used the same function from *bedtools* to identify
942 whether OR genes were present in the TE-rich regions, or the TE-poor (bins with less than
943 50kb of TE content).

944 ***Transcription factor binding sites on TE insertions***

945 We assessed the identification of transcription factor binding sites (TFBSs) with *FIMO*
946 (152), from the *MEME v. 5.0.5* suite tools (153). We extracted the sequences of all TE
947 insertions located 2kb upstream of ORs with *bedtools v. 2.30.0*, function *getfasta*, using
948 parameter *-s* (151). We obtained the TFBS motifs described for *D. melanogaster* from Jaspar

949 database (154), for the transcription factors (TFs) acj6, zf30c, fer1, onecut, and xbp1, which
950 have been demonstrated to activate ORs (43). To compare our findings with OR genes, we
951 performed the same analysis with Serine/Threonine Kinases and zinc-finger as non-HLAU
952 gene families. Only TFBSs with p-value < 0.01 were considered. The enrichment of OR
953 TFBSs in OR genes compared to non-HLAU gene families was assessed with the four
954 frequencies: 1) ORs with TFBSs; 2) ORs without TFBSs; 3) non-HLAU with TFBSs; and 4) non-
955 HLAU without TFBSs.

956 ***Differential expression of genes***

957 We aligned the RNA-seq reads with *STAR v.2.7.3* (147), using the assembled
958 genomes and annotations of their respective species. The alignments were used to create
959 the count matrix reporting the expression of each gene, with *htseq v.0.13.5* (155),
960 parameters *–order pos*; *-s reverse*; *–i gene*; *-t exon*; the latter indicates that the count for
961 each gene corresponds to the sum of all reads aligned into exons (CDSs and UTRs). We
962 merged the counts from species of the *buzzatii* and *mojavensis* clusters separately,
963 maintaining only ortholog genes that have at least 90% of the length of the annotated
964 gene in the reference genome. This threshold allowed us to avoid misinterpretation of the
965 expression level by comparing genes with different lengths. Then, we performed
966 independent statistical analysis for species of the *buzzatii* and *mojavensis* clusters with
967 *DEseq2* v.1.28.1 (156), using a single-factor experimental design, which considers only the
968 factor that different species may have differentially expressed genes. Due to the

969 phylogenetic distance between species of the *buzzatii* and *mojavensis* clusters, we
970 performed the differential expression analysis separately for each group. We only
971 considered genes with adjusted *p-value* < 0.05 and |*log2 fold-change*| > 1 as differentially
972 expressed. To test for association between TE enrichment and differential expression, we
973 selected HLAU genes with the mean of normalized counts between replicates >= 100.
974 Then, we build the contingency table to apply the X^2 independence test, splitting them into
975 four categories: differentially expressed, non-differentially expressed, with TEs at the
976 promoter region, and without TEs at the promoter region. We carried out the test for all
977 pairwise analyses between cactophilic species with preference for *Opuntia sp* and columnar
978 cacti. X^2 with p-value < 0.05 was considered significant. The proportional differences
979 between localization, acceptance, and host use genes compared to non-HLAU genes
980 (Ser/Thr Kinases and zinc-finger gene families) were assessed with Fisher's exact test, with a
981 contingency table containing the average number of genes per group: 1) non-HLAU
982 differentially expressed genes; 2) non-HLAU non-differentially expressed genes; 3) HLAU
983 differentially expressed genes; and 4) HLAU non-differentially expressed genes. *Fisher's*
984 Exact test < 0.05 was considered significant.

985 ***Identification of gene-TE chimeric transcripts***

986 We detected gene-TE chimeric transcripts with *ChimeraTE v.2.0.0* Mode 1 (24). Each
987 species was analyzed using its respective genome, gene annotation, and TE annotation.
988 Shortly, *ChimeraTE* identifies chimeras based on paired-end reads spanning from exons to

989 TE sequences. Based on the TE position compared to the gene (upstream, exon, intron,
990 downstream), chimeric transcripts are classified into categories. We have used default
991 parameters, except *--replicate 3*, which considers true chimeras only those identified in the
992 three RNA-seq replicates, and *–coverage 10* to obtain chimeras with at least 10 chimeric
993 reads on average between replicates. To calculate the contribution of the chimeric
994 transcript expression relative to total gene expression, we developed a new module for
995 ChimeraTE. This module can be used downstream to ChimeraTE mode 1 analysis. In the first
996 step, all RNA-seq replicates will be merged into a single fastq library. Then, the merged
997 reads are aligned to the genome with STAR (147). Next, Stringtie2 (157) is used to assess
998 transcript expression, including both reference and non-reference isoforms. All TEs
999 generating chimeras are intersected with exons predicted by Stringtie with pyranges (158).
1000 Only overlaps with 80% of the TE within the exon are considered. Finally, the expression of
1001 the TE-containing isoforms is compared to the total gene expression. In our analysis, we
1002 removed all chimeric transcripts for which Stringtie was not able to build the chimeric
1003 isoform. In addition, aiming to keep chimeras with putative function, we selected only the
1004 ones with contribution >= 25%. This analysis can be reproduced with any output from
1005 ChimeraTE with the code "mode1_contrib_chimeras.py" (https://github.com/OliveiraDS-
1006 hub/ChimeraTE/blob/main/scripts/mode1_contrib_chimeras.py).

1007 ***Data access***

The quality control data from ONT and RNA-seq data, the genome assemblies, and their respective gene and TE annotation (libraries and gtf files) are available in the Zenodo repository: https://zenodo.org/records/16501018. The codes to reproduce differential expression analysis, positive selection tests, gene annotation, permutation tests, TE annotation, and TFBS analysis are available on the GitHub repository: https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-Drosophila-Species.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

We thank Séverine Chambeyron for useful discussions and advice. This work was performed using the computing facilities of the CC LBBE/PRABI. This study was funded by Brazilian agencies FAPESP- The São Paulo Research Foundation [grant number 2020/06238-2] and from the National Council for Scientific and Technological Development (308020/2021-9) to C. M. A. C; the Eiffel Fellowship Campus France [P769649C] and CAPES – PrInt [88887.716810/2022-00] to D. S. O.; by Ministerio de Ciencia e Innovación (Spain) [grant number PID2021-127107NB-I00], and Generalitat de Catalunya (Spain) [grant number 2021 SGR 00526] to M. P. G. G.; and by Udl-FAPESP ANR-IDEX-0005 to C. V. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001 to DSO.

1027 *Authors' contributions:* A. L., A. B, M. P. G. G, performed experiments; F.S., Performed
1028 genome assemblies; W. V. B. N. performed phylogenetic analysis; D. S. O performed
1029 computational analysis related to genomics, transcriptomics, and statistical tests; D. S. O.
1030 wrote and edited the paper; C. V. and C. M. A. C. designed the study, and reviewed the
1031 manuscript. All authors read, corrected, and approved the final manuscript.

1032 **References**

1033 1. Whiteman NK, Pierce NE. Delicious poison: genetics of *Drosophila* host plant
1034 preference. Trends in Ecology & Evolution. 2008;23(9):473–8.

1035 2. Forbes AA, Devine SN, Hippee AC, Tvedte ES, Ward AKG, Widmayer HA, et al.
1036 Revisiting the particular role of host shifts in initiating insect speciation. Evolution. 2017
1037 May;71(5):1126–37.

1038 3. Rundle HD, Nosil P. Ecological speciation. Ecology Letters. 2005;8(3):336–52.

1039 4. Howard DJ, Berlocher SH. Endless Forms: Species and Speciation. Oxford University
1040 Press; 1998. 486 p.

1041 5. O'Grady PM, Markow TA. Rapid morphological, behavioral, and ecological evolution
1042 in *Drosophila*: comparisons between the endemic Hawaiian *Drosophila* and the cactophilic
1043 *repleta* species group. In: Singh RS, Xu J, Kulathinal RJ, editors. Rapidly Evolving Genes and
1044 Genetic Systems. Oxford University Press; 2012. Available from:
1045 https://doi.org/10.1093/acprof:oso/9780199642274.003.0018

1046 6. Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ.
1047 Monophyly, divergence times, and evolution of host plant use inferred from a revised
1048 phylogeny of the *Drosophila repleta* species group. Molecular Phylogenetics and Evolution.
1049 2012;64(3):533–44.

1050 7. Stintzing FC, Carle R. Cactus stems (*Opuntia spp.*): A review on their chemistry,
1051 technology, and uses. Molecular Nutrition & Food Research. 2005;49(2):175–94.

1052 8. Gomez GA. Transpecific Polymorphisms in an Inversion Linked Esterase Locus in
1053 *Drosophila buzzatii*. Molecular Biology and Evolution. 2003;20(3):410–23.

1054 9. Hasson E, Naveira H, Fontdevila A. The breeding sites of the Argentinian species of
1055 the *Drosophila mulleri* complex (subgenus *Drosophila-repleta* group). Rev Chil Hist Nat.
1056 1992;65:319–26.

1057 10. Newby BD, Etges WJ. Host Preference Among Populations of *Drosophila mojavensis*
1058 (Diptera: Drosophilidae) that Use Different Host Cacti. Journal of Insect Behavior.
1059 1998;11(5):691–712.

1060 11. Rane RV, Clarke DF, Pearce SL, Zhang G, Hoffmann AA, Oakeshott JG. Detoxification
1061 Genes Differ Between Cactus-, Fruit-, and Flower-Feeding *Drosophila*. Journal of Heredity.
1062 2019;110(1):80–91.

1063 12. Markow TA. Host use and host shifts in *Drosophila*. Current Opinion in Insect
1064 Science. 2019;31:139–45.

1065 13. Montell C. A taste of the *Drosophila* gustatory receptors. Current Opinion in
1066 Neurobiology. 2009;19(4):345–53.

1067 14. Ni L. The Structure and Function of Ionotropic Receptors in *Drosophila*. Front Mol
1068 Neurosci. 2021;13:638839.

1069 15. Chen Y, Amrein H. Ionotropic Receptors Mediate *Drosophila* Oviposition Preference
1070 through Sour Gustatory Receptor Neurons. Current Biology. 2017;27(18):2741-2750.e4.

1071 16. Gomez-Diaz C, Martin F, Garcia-Fernandez JM, Alcorta E. The Two Main Olfactory
1072 Receptor Families in *Drosophila*, ORs and IRs: A Comparative Approach. Front Cell Neurosci.
1073 2018;12:253.

1074 17. Lang M, Murat S, Clark AG, Gouppil G, Blais C, Matzkin LM, et al. Mutations in the
1075 neverland Gene Turned *Drosophila pachea* into an Obligate Specialist Species. Science.
1076 2012;337(6102):1658–61.

1077 18. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. Developmentally
1078 Regulated piRNA Clusters Implicate MILI in Transposon Control. Science.
1079 2007;316(5825):744–7.

1080 19. Fedoroff NV. Transposable Elements, Epigenetics, and Genome Evolution. Science.
1081 2012;338(6108):758–67.

1082 20. Capy P. Taming, Domestication and Exaptation: Trajectories of Transposable
1083 Elements in Genomes. Cells. 2021;10(12):3590.

1084 21. Drongitis D, Aniello F, Fucci L, Donizetti A. Roles of Transposable Elements in the
1085 Different Layers of Gene Expression Regulation. IJMS. 2019;20(22):5755.

1086 22. Magwire MM, Bayer F, Webster CL, Cao C, Jiggins FM. Successive Increases in the
1087 Resistance of *Drosophila* to Viral Infection through a Transposon Insertion Followed by a
1088 Duplication. Begun DJ, editor. PLoS Genet. 2011;7(10):e1002337.

1089 23. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-
1090 transposable element transcripts in the *Drosophila melanogaster* genome. BMC Biol.
1091 2005;3(1):24.

1092 24. Oliveira DS, Fablet M, Larue A, Vallier A, Carareto CMA, Rebollo R, et al. ChimeraTE: a
1093 pipeline to detect chimeric transcripts derived from genes and transposable elements.
1094 Nucleic Acids Research. 2023;51(18):9764–84.

1095 25. Matzkin LM. The Molecular Basis of Host Adaptation in Cactophilic *Drosophila*:
1096 Molecular Evolution of a Glutathione *S* -Transferase Gene (*GstD1*) in *Drosophila*
1097 *mojavensis*. Genetics. 2008;178(2):1073–83.

1098 26. Pfeiler E, Castrezana S, Reed LK, Markow TA. Genetic, ecological and morphological
1099 differences among populations of the cactophilic *Drosophila mojavensis* from southwestern

1100 USA and northwestern Mexico, with descriptions of two new subspecies. Journal of Natural
1101 History. 2009;43(15–16):923–38.

1102 27. Etges WJ. Genetics of Host-Cactus Response and Life-History Evolution Among
1103 Ancestral and Derived Populations of Cactophilic *Drosophila mojavensis*. Evolution.
1104 1993;47(3):750–67.

1105 28. Crowley-Gall A, Date P, Han C, Rhodes N, Andolfatto P, Layne JE, et al. Population
1106 differences in olfaction accompany host shift in *Drosophila mojavensis*. Proc R Soc B. 2016
1107 Aug 31;283(1837):20161562.

1108 29. Benowitz KM, Coleman JM, Allan CW, Matzkin LM. Contributions of cis- and trans-
1109 Regulatory Evolution to Transcriptomic Divergence across Populations in the *Drosophila*
1110 *mojavensis* Larval Brain. Genome Biology and Evolution. 2020;12(8):1407–18.

1111 30. Rajpurohit S, Oliveira CC, Etges WJ, Gibbs AG. Functional genomic and phenotypic
1112 responses to desiccation in natural populations of a desert drosophilid. Molecular Ecology.
1113 2013;22(10):2698–715.

1114 31. De Panis D, Dopazo H, Bongcam-Rudloff E, Conesa A, Hasson E. Transcriptional
1115 responses are oriented towards different components of the rearing environment in two
1116 *Drosophila* sibling species. BMC Genomics. 2022;23(1):515.

1117 32. Matzkin L, Markow TA. Transcriptional Differentiation Across the Four Subspecies of
1118 *Drosophila mojavensis*. In: Speciation: Natural Processes, Genetics and Biodiversity. New
1119 York: Nova Scientific Publishers; 2013. p. 119–35.

1120 33. Benowitz KM, Allan CW, Jaworski CC, Sanderson MJ, Diaz F, Chen X, et al.
1121 Chromosome-length genome assemblies of cactophilic *Drosophila* illuminate links between
1122 structural and sequence evolution. bioRxiv; 2022. Available from:
1123 https://www.biorxiv.org/content/10.1101/2022.10.16.512445v1

1124 34. Oliveira DCSG, Almeida FC, O'Grady PM, Armella MA, DeSalle R, Etges WJ.
1125 Monophyly, divergence times, and evolution of host plant use inferred from a revised
1126 phylogeny of the *Drosophila* repleta species group. Molecular Phylogenetics and Evolution.
1127 2012;64(3):533–44.

1128 35. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified
1129 classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–
1130 82.

1131 36. Sanchez-Flores A, Peñaloza F, Carpinteyro-Ponce J, Nazario-Yepiz N, Abreu-Goodger
1132 C, Machado CA, et al. Genome Evolution in Three Species of Cactophilic *Drosophila*. G3
1133 Genes|Genomes|Genetics. 2016;6(10):3097–105.

1134 37. Smith G, Lohse K, Etges WJ, Ritchie MG. Model-based comparisons of
1135 phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila*

1136 *mojavensis*: Phylogeography of *Drosophila mojavensis*. Molecular Ecology.
1137 2012;21(13):3293–307.

1138 38. Guio L, Vieira C, González J. Stress affects the epigenetic marks added by natural
1139 transposable element insertions in *Drosophila melanogaster*. Sci Rep. 2018;8(1):12197.

1140 39. Daborn PJ, Yen JL, Bogwitz MR, Le Goff G, Feil E, Jeffers S, et al. A Single P450 Allele
1141 Associated with Insecticide Resistance in *Drosophila*. Science. 2002;297(5590):2253–6.

1142 40. Chang NC, Rovira Q, Wells J, Feschotte C, Vaquerizas JM. Zebrafish transposable
1143 elements show extensive diversification in age, genomic distribution, and developmental
1144 expression. Genome Res. 2022;32(7):1408–23.

1145 41. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze
1146 Genomic Innovation in a Mammalian Lineage. Genome Biology and Evolution.
1147 2014;6(10):2595–610.

1148 42. Pritham EJ, Feschotte C. Massive amplification of rolling-circle transposons in the
1149 lineage of the bat *Myotis lucifugus*. Proceedings of the National Academy of Sciences.
1150 2007;104(6):1895–900.

1151 43. Jafari S, Alkhori L, Schleiffer A, Brochtrup A, Hummel T, Alenius M. Combinatorial
1152 activation and repression by seven transcription factors specify *Drosophila* odorant receptor
1153 expression. PLoS Biol. 2012;10(3):e1001280.

1154 44. Capy P. Taming, Domestication and Exaptation: Trajectories of Transposable
1155 Elements in Genomes. Cells. 2021;10(12):3590.

1156 45. Fogleman JC, Danielson PB. Chemical Interactions in the Cactus-Microorganism-
1157 *Drosophila* Model System of the Sonoran Desert. American Zoologist. 2001;41(4):877–89.

1158 46. Fogleman JC, Heed WB, Kircher HW. *Drosophila mettleri* and senita cactus alkaloids:
1159 Fitness measurements and their ecological significance. Comparative Biochemistry and
1160 Physiology Part A: Physiology. 1982;71(3):413–7.

1161 47. McBride CS. Rapid evolution of smell and taste receptor genes during host
1162 specialization in *Drosophila sechellia*. Proceedings of the National Academy of Sciences.
1163 2007;104(12):4996–5001.

1164 48. McBride CS, Arguello JR. Five *Drosophila* Genomes Reveal Nonneutral Evolution and
1165 the Signature of Host Specialization in the Chemoreceptor Superfamily. Genetics.
1166 2007;177(3):1395–416.

1167 49. Allan CW, Matzkin LM. Genomic analysis of the four ecologically distinct cactus host
1168 populations of *Drosophila mojavensis*. BMC Genomics. 2019;20(1):732.

1169 50. Guillén Y, Rius N, Delprat A, Williford A, Muyas F, Puig M, et al. Genomics of
1170 Ecological Adaptation in Cactophilic *Drosophila*. Genome Biology and Evolution.
1171 2015;7(1):349–66.

1172 51. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene
1173 Orthology inference. Briefings in Bioinformatics. 2011;12(5):379–91.

1174 52. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated
1175 retrotransposon transcriptome of mammalian cells. Nat Genet. 2009;41(5):563–71.

1176 53. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al.
1177 Transposable Elements Are Major Contributors to the Origin, Diversification, and
1178 Regulation of Vertebrate Long Noncoding RNAs. Hoekstra HE, editor. PLoS Genet.
1179 2013;9(4):e1003470.

1180 54. Emera D, Casola C, Lynch VJ, Wildman DE, Agnew D, Wagner GP. Convergent
1181 Evolution of Endometrial Prolactin Expression in Primates, Mice, and Elephants Through the
1182 Independent Recruitment of Transposable Elements. Molecular Biology and Evolution.
1183 2012;29(1):239–47.

1184 55. Ye M, Goudot C, Hoyler T, Lemoine B, Amigorena S, Zueva E. Specific subfamilies of
1185 transposable elements contribute to different domains of T lymphocyte enhancers. Proc
1186 Natl Acad Sci USA. 2020;117(14):7905–16.

1187 56. Xie M, Hong C, Zhang B, Lowdon RF, Xing X, Li D, et al. DNA hypomethylation within
1188 specific transposable element families associates with tissue-specific enhancer landscape.
1189 Nat Genet. 2013;45(7):836–41.

1190 57. Schmidt D, Schwalie PC, Wilson MD, Ballester B, Gonçalves Â, Kutter C, et al. Waves
1191 of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in
1192 Multiple Mammalian Lineages. Cell. 2012;148(1–2):335–48.

1193 58. Wang J, Lunyak VV, Jordan IK. Genome-wide prediction and analysis of human
1194 chromatin boundary elements. Nucleic Acids Research. 2012;40(2):511–29.

1195 59. Sundaram V, Cheng Y, Ma Z, Li D, Xing X, Edge P, et al. Widespread contribution of
1196 transposable elements to the innovation of gene regulatory networks. Genome Res.
1197 2014;24(12):1963–76.

1198 60. Gill RA, Scossa F, King GJ, Golicz AA, Tong C, Snowdon RJ, et al. On the Role of
1199 Transposable Elements in the Regulation of Gene Expression and Subgenomic Interactions
1200 in Crop Genomes. Critical Reviews in Plant Sciences. 2021;40(2):157–89.

1201 61. Hof AE van't, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, et al. The
1202 industrial melanism mutation in British peppered moths is a transposable element. Nature.
1203 2016;534(7605):102–5.

1204 62. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers
1205 Xenobiotic Resistance in *Drosophila*. PLoS Genet. 2014;10(8):e1004560.

1206 63. Dekker T, Ibba I, Siju KP, Stensmyr MC, Hansson BS. Olfactory Shifts Parallel
1207 Superspecialism for Toxic Fruit in *Drosophila melanogaster* Sibling, *D. sechellia*. Current
1208 Biology. 2006;16(1):101–9.

1209 64. Matsuo T, Sugaya S, Yasukawa J, Aigaki T, Fuyama Y. Odorant-Binding Proteins
1210 OBP57d and OBP57e Affect Taste Perception and Host-Plant Preference in *Drosophila*
1211 *sechellia*. PLoS Biol. 2007;5(5):e118.

1212 65. McKenzie SK, Kronauer DJC. The genomic architecture and molecular evolution of
1213 ant odorant receptors. Genome Res. 2018;28(11):1757–65.

1214 66. Cultrone A, Domínguez YR, Drevet C, Scazzocchio C, Fernández-Martín R. The tightly
1215 regulated promoter of the *xanA* gene of *Aspergillus nidulans* is included in a helitron: A
1216 fungal helitron includes a regulated promoter. Molecular Microbiology. 2007;63(6):1577–87.

1217 67. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. A novel class of Helitron-
1218 related transposable elements in maize contain portions of multiple pseudogenes. Plant
1219 Mol Biol. 2005;57(1):115–27.

1220 68. Huang K, Li CF, Wu J, Wei JH, Zou Y, Han MJ, et al. Enhancer activity of *Helitron* in
1221 *sericin-1* gene promoter from *Bombyx mori*: Enhancer activity of Helitrons in sericin-1.
1222 Insect Science. 2016;23(3):396–405.

1223 69. Batista RA, Moreno-Romero J, Qiu Y, van Boven J, Santos-González J, Figueiredo DD,
1224 et al. The MADS-box transcription factor PHERES1 controls imprinting in the endosperm by
1225 binding to domesticated transposons. eLife. 2019;8:e50541.

1226 70. Carareto CMA, Hernandez EH, Vieira C. Genomic regions harboring insecticide
1227 resistance-associated *Cyp* genes are enriched by transposable element fragments carrying

1228 putative transcription factor binding sites in two sibling *Drosophila* species. Gene.
1229 2014;537(1):93–9.

1230 71. Merenciano M, Aguilera L, González J. Two-step CRISPR-Cas9 protocol for
1231 transposable element deletion in *D. melanogaster* natural populations. STAR Protocols.
1232 2023;4(3):102501.

1233 72. Wright CR, William S. Characterization of Volatiles of Necrotic *Stenocereus thurberi*
1234 and *Opuntia littoralis* and Toxicity and Olfactory Preference of *Drosophila melanogster*, *D.*
1235 *mojavensis wrigleyi*, and *D. mojavensis sonorensis* to Necrotic Cactus Volatiles. Natural
1236 Product Communications. 2014;9(6):1934578X1400900833.

1237 73. Hasson E, De Panis D, Hurtado J, Mensch J. Host Plant Adaptation in Cactophilic
1238 Species of the *Drosophila buzzatii* Cluster: Fitness and Transcriptomics. Journal of Heredity.
1239 2019;110(1):46–57.

1240 74. Hickner PV, Rivaldi CL, Johnson CM, Siddappaji M, Raster GJ, Syed Z. The making of
1241 a pest: Insights from the evolution of chemosensory receptor families in a pestiferous and
1242 invasive fly, *Drosophila suzukii*. BMC Genomics. 2016;17(1):648.

1243 75. Mansourian S, Stensmyr MC. The chemical ecology of the fly. Current Opinion in
1244 Neurobiology. 2015;34:95–102.

1245 76. Goldman-Huertas B, Mitchell RF, Lapoint RT, Faucher CP, Hildebrand JG, Whiteman
1246 NK. Evolution of herbivory in Drosophilidae linked to loss of behaviors, antennal responses,
1247 odorant receptors, and ancestral diet. Proc Natl Acad Sci USA. 2015;112(10):3026–31.

1248 77. Crowley-Gall A, Date P, Han C, Rhodes N, Andolfatto P, Layne JE, et al. Population
1249 differences in olfaction accompany host shift in *Drosophila mojavensis*. Proc R Soc B.
1250 2016;283(1837):20161562.

1251 78. Ruiz A, Heed WB. Host-Plant Specificity in the Cactophilic *Drosophila mulleri* Species
1252 Complex. The Journal of Animal Ecology. 1988;57(1):237.

1253 79. Mateus RP, Nazario-Yepiz NO, Ibarra-Laclette E, Ramirez Loustalot-Laclette M,
1254 Markow TA. Developmental and Transcriptomal Responses to Seasonal Dietary Shifts in the
1255 Cactophilic *Drosophila mojavensis* of North America. Journal of Heredity. 2019;110(1):58–67.

1256 80. Matzkin LM. Population transcriptomics of cactus host shifts in *Drosophila*
1257 *mojavensis*: Population Transcriptomics of *D. mojavensis*. Molecular Ecology.
1258 2012;21(10):2428–39.

1259 81. De Panis D, Dopazo H, Bongcam-Rudloff E, Conesa A, Hasson E. Transcriptional
1260 responses are oriented towards different components of the rearing environment in two
1261 *Drosophila* sibling species. BMC Genomics. 2022;23(1):515.

1262 82. Feyereisen R. Insect CYP Genes and P450 Enzymes. In: Insect Molecular Biology and
1263 Biochemistry. Elsevier; 2012. p. 236–316. Available from:
1264 https://linkinghub.elsevier.com/retrieve/pii/B978012384747810008X

1265 83. Tijet N, Helvig C, Feyereisen R. The cytochrome P450 gene superfamily in *Drosophila*
1266 *melanogaster*: Annotation, intron-exon organization and phylogeny. Gene. 2001;262(1–
1267 2):189–98.

1268 84. Rane RV, Pearce SL, Li F, Coppin C, Schiffer M, Shirriffs J, et al. Genomic changes
1269 associated with adaptation to arid environments in cactophilic *Drosophila* species. BMC
1270 Genomics. 2019;20(1):52.

1271 85. Coronado-Zamora M, González J. Transposons contribute to the functional
1272 diversification of the head, gut, and ovary transcriptomes across *Drosophila* natural strains.
1273 Genome Res. 2023;33(9):1541–53.

1274 86. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution
1275 of vertebrate transcription factors by transposase capture. Science.
1276 2021;371(6531):eabc6405.

1277 87. Pal D, Patel M, Boulet F, Sundarraj J, Grant OA, Branco MR, et al. H4K16ac activates
1278 the transcription of transposable elements and contributes to their cis-regulatory function.
1279 Nat Struct Mol Biol. 2023;30(7):935–47.

1280 88. Lippman Z, Gendrel AV, Black M, Vaughn MW, Dedhia N, Richard McCombie W, et
1281 al. Role of transposable elements in heterochromatin and epigenetic control. Nature.
1282 2004;430(6998):471–6.

1283 89. Guio L, Vieira C, González J. Stress affects the epigenetic marks added by natural
1284 transposable element insertions in *Drosophila melanogaster*. Sci Rep. 2018 Aug
1285 15;8(1):12197.

1286 90. Coronado-Zamora M, González J. The epigenetics effects of transposable elements
1287 are context dependent and not restricted to gene silencing. bioRxiv; 2023. Available from:
1288 https://www.biorxiv.org/content/10.1101/2023.11.27.568862v1

1289 91. Fonseca PM, Moura RD, Wallau GL, Loreto ELS. The mobilome of *Drosophila*
1290 *incompta*, a flower-breeding species: comparison of transposable element landscapes
1291 among generalist and specialist flies. Chromosome Res. 2019;27(3):203–19.

1292 92. Mérel V, Gibert P, Buch I, Rodriguez Rada V, Estoup A, Gautier M, et al. The
1293 Worldwide Invasion of *Drosophila suzukii* Is Accompanied by a Large Increase of
1294 Transposable Element Load and a Small Number of Putatively Adaptive Insertions.
1295 Molecular Biology and Evolution. 2021;38(10):4252–67.

1296 93. Marin P, Jaquet A, Picarle J, Fablet M, Merel V, Delignette-Muller ML, et al.
1297 Phenotypic and Transcriptomic Responses to Stress Differ According to Population
1298 Geography in an Invasive Species. Genome Biology and Evolution. 2021;13(9):evab208.

1299 94. Zidi M, Klai K, Confais J, Chénais B, Caruso A, Denis F, et al. Genome-Wide Screening
1300 of Transposable Elements in the Whitefly, *Bemisia tabaci* (Hemiptera: Aleyrodidae),
1301 Revealed Insertions with Potential Insecticide Resistance Implications. Insects.
1302 2022;13(5):396.

1303 95. Arribas YA, Baudon B, Rotival M, Suárez G, Bonté PE, Casas V, et al. Transposable
1304 element exonization generates a reservoir of evolving and functional protein isoforms. Cell.
1305 2024;187(26):7603-7620.e22.

1306 96. Harigaya Y, Parker R. No-go decay: a quality control mechanism for RNA in
1307 translation. WIREs RNA. 2010;1(1):132–41.

1308 97. Vasudevan S, Peltz SW, Wilusz CJ. Non-stop decay—a new mRNA surveillance
1309 pathway. BioEssays. 2002;24(9):785–8.

1310 98. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-
1311 mediated decay pathway. Nucleic Acids Res. 2016;44(4):1483–95.

1312 99. Whisson SC, Vetukuri RR, Avrova AO, Dixelius C. Can silencing of transposons
1313 contribute to variation in effector gene expression in Phytophthora infestans? Mobile
1314 Genetic Elements. 2012;2(2):110–4.

1315 100. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, et al. Genome
1316 Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. Science.
1317 2010;330(6010):1540–3.

1318 101. De Panis DN, Padró J, Furió-Tarí P, Tarazona S, Milla Carmona PS, Soto IM, et al.
1319 Transcriptome modulation during host shift is driven by secondary metabolites in desert
1320 *Drosophila*. Molecular Ecology. 2016;25(18):4534–50.

1321 102. Knolhoff LM, Heckel DG. Behavioral Assays for Studies of Host Plant Choice and
1322 Adaptation in Herbivorous Insects. Annual Review of Entomology. 2014;59(Volume 59,
1323 2014):263–78.

1324 103. Zouros E, d'Entremont CJ. Sexual Isolation Among Populations of *Drosophila*
1325 *mojavensis*: Response to Pressure from a Related Species. Evolution. 1980;34(3):421–30.

1326 104. Markow TA. Sexual Isolation among Populations of *Drosophila mojavensis*. Evolution.
1327 1991;45(6):1525–9.

1328 105. Howard RW, Blomquist GJ. Ecological, behavioral, and biochemical aspects of the
1329 insect hydrocarbons. Annual Review of Entomology. 2005;50:371–93.

1330 106. Stennett MD, Etges WJ. Premating Isolation Is Determined by Larval Rearing
1331 Substrates in Cactophilic *Drosophila mojavensis*. III. Epicuticular Hydrocarbon Variation Is
1332 Determined by Use of Different Host Plants in *Drosophila mojavensis* and *Drosophila*
1333 *arizonae*. J Chem Ecol. 1997;23(12):2803–24.

1334 107. Havens JA, Etges WJ. Premating isolation is determined by larval rearing substrates
1335 in cactophilic *Drosophila mojavensis*. IX. Host plant and population specific epicuticular

1336 hydrocarbon expression influences mate choice and sexual selection. Journal of
1337 Evolutionary Biology. 2013;26(3):562–76.

1338 108. Romero-Soriano V, Burlet N, Vela D, Fontdevila A, Vieira C, García Guerreiro MP.
1339 *Drosophila* Females Undergo Genome Expansion after Interspecific Hybridization. Genome
1340 Biol Evol. 2016;8(3):556–61.

1341 109. Bodelón A, Fablet M, Veber P, Vieira C, García Guerreiro MP. High Stability of the
1342 Epigenome in *Drosophila* Interspecific Hybrids. Genome Biology and Evolution.
1343 2022;14(2):evac024.

1344 110. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using
1345 repeat graphs. Nat Biotechnol. 2019;37(5):540–6.

1346 111. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics.
1347 2018;34(18):3094–100.

1348 112. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-
1349 MEM. arXiv:13033997 [q-bio]. 2013; Available from: http://arxiv.org/abs/1303.3997

1350 113. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient
1351 and simple way. PeerJ. 2018;6:e4958.

1352 114. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1353 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

1354 115. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating
1355 dotplots on genome scale. Bioinformatics. 2007;23(8):1026–8.

1356 116. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO:
1357 fast and accurate reference-guided scaffolding of draft genomes. Genome Biol.
1358 2019;20(1):224.

1359 117. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:
1360 assessing genome assembly and annotation completeness with single-copy orthologs.
1361 Bioinformatics. 2015;31(19):3210–2.

1362 118. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A,
1363 editor. Bioinformatics. 2021;37(12):1639–43.

1364 119. Kim BY, Wang JR, Miller DE, Barmina O, Delaney E, Thompson A, et al. Highly
1365 contiguous assemblies of 101 drosophilid genomes. eLife. 2021;10(e66405):33.

1366 120. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully
1367 automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP,
1368 AUGUSTUS, and TSEBRA. Genome Research. 2024;34(5):769-777.

1369 121. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative
1370 genomics. Genome Biol. 2019;20(1):238.

1371 122. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND.
1372 Nat Methods. 2015;12(1):59–60.

1373 123. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database
1374 search programs. Nucleic Acids Research. 1997;25(17):3389–402.

1375 124. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A,
1376 editor. Bioinformatics. 2021;37(12):1639–43.

1377 125. Vanderlinde T, Dupim EG, Nazario-Yepiz NO, Carvalho AB. An Improved Genome
1378 Assembly for *Drosophila navojoa*, the Basal Species in the mojavensis Cluster. Journal of
1379 Heredity. 2019;110(1):118–23.

1380 126. Hu J, Fan J, Sun Z, Liu S. NextPolish: a fast and efficient genome polishing tool for
1381 long-read assembly. Berger B, editor. Bioinformatics. 2020;36(7):2253–5.

1382 127. Banho CA, Mérel V, Oliveira TYK, Carareto CMA, Vieira C. Comparative
1383 transcriptomics between *Drosophila mojavensis* and *D. arizonae* reveals transgressive gene
1384 expression and underexpression of spermatogenesis-related genes in hybrid testes. Sci
1385 Rep. 2021;11(1):9844.

1386 128. Moreyra NN, Almeida FC, Allan C, Frankel N, Matzkin LM, Hasson E. Phylogenomics
1387 provides insights into the evolution of cactophily and host plant shifts in *Drosophila*.
1388 Molecular Phylogenetics and Evolution. 2023;178:107653.

1389 129. Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF
1390 format. Zenodo. 2023; Available from: https://zenodo.org/records/8178877

1391 130. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
1392 Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

1393 131. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence
1394 alignments into the corresponding codon alignments. Nucleic Acids Research.
1395 2006;34(2):W609–12.

1396 132. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A,
1397 et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the
1398 Genomic Era. Molecular Biology and Evolution. 2020;37(5):1530–4.

1399 133. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology
1400 and Evolution. 2007;24(8):1586–91.

1401 134. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina
1402 sequence data. Bioinformatics. 2014;30(15):2114–20.

1403 135. Baril TJ, Galbraith JD, Hayward A. Earl Grey: a fully automated user-friendly
1404 transposable element annotation and analysis pipeline. Molecular Biology and Evolution.
1405 2024;41(4)msae068.

1406 136. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2
1407 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA.
1408 2020;117(17):9451–7.

1409 137. Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in
1410 Sequenced Genomes. Genome Res. 2002;12(8):1269–76.

1411 138. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large
1412 genomes. Bioinformatics. 2005;21(Suppl 1):i351–8.

1413 139. Li Z, Gilbert C, Peng H, Pollet N. Discovery of numerous novel *Helitron* -like elements
1414 in eukaryote genomes using HELIANO. Nucleic Acids Research. 2024;52(17):e79–e79.

1415 140. Platt RN, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is
1416 Vital When Analyzing New Genome Assemblies. Genome Biol Evol. 2016;8(2):403–10.

1417 141. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7:
1418 Improvements in Performance and Usability. Molecular Biology and Evolution.
1419 2013;30(4):772–80.

1420 142. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated
1421 alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–
1422 3.

1423 143. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open
1424 Software Suite. Trends in Genetics. 2000;16(6):276–7.

1425 144. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic
1426 Acids Research. 1999;27(2):573–80.

1427 145. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of
1428 transposable element families, sequence models, and genome annotations. Mobile DNA.
1429 2021;12(1):2.

1430 146. SMIT, Arian FA. RepeatMasker Open 3.0. 2004.

1431 147. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast
1432 universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21.

1433 148. Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-
1434 fragmentation and annotation. Hancock J, editor. Bioinformatics. 2019;35(6):1051–2.

1435 149. Flutre T, Permal E, Quesneville H. Transposable Element Annotation in Completely
1436 Sequenced Eukaryote Genomes. In: Grandbastien MA, Casacuberta JM, editors. Plant
1437 Transposable Elements. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 17–39.
1438 (Topics in Current Genetics; vol. 24). Available from: https://link.springer.com/10.1007/978-
1439 3-642-31842-9_2

1440 150. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR:
1441 an R/Bioconductor package for the association analysis of genomic regions based on
1442 permutation tests. Bioinformatics. 2015;32(2):289-291.

1443 151. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
1444 features. Bioinformatics. 2010;26(6):841–2.

1445 152. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif.
1446 Bioinformatics. 2011;27(7):1017–8.

1447 153. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Research.
1448 2015;43(W1):W39–49.

1449 154. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et
1450 al. JASPAR 2020: update of the open-access database of transcription factor binding
1451 profiles. Nucleic Acids Research. 2019;48(D1):87-92.

1452 155. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-
1453 throughput sequencing data. Bioinformatics. 2015;31(2):166–9.

1454 156. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion
1455 for RNA-seq data with DESeq2. Genome Biol. 2014;15(12):550.

1456 157. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M. Transcriptome
1457 assembly from long-read RNA-seq alignments with StringTie2. Genome Biol.
1458 2019;20(1):278.

1459 158. Stovner EB, Sætrom P. PyRanges: efficient comparison of genomic intervals in
1460 Python. Bioinformatics. 2020;36(3):918–9.

1461