bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Transposable elements as evolutionary driving force to ecological speciation in cactophilic *Drosophila* species

Oliveira D. S.[1,2], Larue A.[2], Nunes W. V. B.[2], Sabot F.[3], Bodelón A.[4], García Guerreiro M. P.[4], 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] Laboratoire de Biométrie et Biologie Évolutive, Universite Claude Bernard Lyon 1, CNRS, UMR5558, Villeurbanne 69100, France.

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

[4] 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: daniel.siqueira-de-oliveira@etu.univ-lyon1.fr

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

ABSTRACT

Background

The host shift in insects has been considered a key process with potential to collaborate with ecological speciation. Both genomics and transcriptomics variation has been attributed to such process, in which gene families with functions for host location, acceptance and usage have been proposed to evolve. In this context, cactophilic *Drosophila* species are an excellent model to study host shift effects, since they use a wide-range of cacti as hosts, and many species have cacti-hosts preference. Despite the potential adaptive role of TEs by generating genetic variability between species and populations, the extent of TEs' contribution to host shift remains unexplored.

Results

Here, we performed genomics and transcriptomics analyses in seven genomes of cactophilic species/subspecies to investigate how TEs interact with genes likely to be associated with host shift. Our results revealed transposition bursts between species, and an enrichment of TEs at promoter regions of host shift-related genes. Pairwise differential expression analysis between species with different preferential hosts in larvae and head tissues demonstrated divergence on gene expression associated with host location in head, whereas for the larvae we found higher differential expression of genes related to usage/detoxification. Although TEs' presence does not affect overall gene expression, we observed 2.1% of genes generating gene-TE chimeric transcripts, including those with function affecting host preference. In addition, *Helitrons* were often observed interacting with genes as a *cis*-regulatory element.

Conclusions

Our combined genomics and transcriptomics approaches provide new insights regarding the evolutionary role of TEs on the context of ecological speciation.

Keywords: Host shift, transposable elements, adaptation, ecological speciation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Background

Ecological speciation refers to the evolutionary phenomenon when the divergence between populations is driven by adaptations to different niches. In many species, it might be an outcome of successive adaptations driven by a host shift (HS), as a result of divergent selection between different ecological environments (1–3) . Due to the broad variety of hosts, the successful plant-insect interaction requires adaptations from the insect. Hence, natural selection increases the frequency of alleles that bring benefits to the fitness when the HS happens. *Drosophila* species use a wide range of host necrotic tissues as feeding and breeding sites, providing a suitable model to investigate the HS process and its role in the insects' evolution. Among them, the *repleta* group comprises a clade with ~100 cactophilic *Drosophila* species. The specialization in cacti represents an important ecological challenge to the flies, due to the lack of nitrogen and phosphorus in comparison to other hosts, as well as the presence of toxic compounds (4). In the *repleta* species, the HS was initially to the *Opuntia sp*. cacti, and posteriorly several independent shifts happened to columnar cacti (4), considered a host with high toxicity compared to *Opuntia sp (5)*. In *repleta* group, there are cases of recent HS between sibling species of two complexes of the subgroup *mulleri* (*buzzatii* and *mulleri* complexes). In the complex *buzzatii*, *Drosophila buzzatii* and *D. koepferae* are sibling species that diverged approximately 4-5 Mya (6), and they have different primary hosts. The former species use preferentially *Opuntia sp.* cacti, whereas the latter is a columnar cacti dweller (7). In the complex *mulleri*, *D. arizonae* and the four allopatric subspecies of *D. mojavensis* have a high diversity regarding cacti hosts' preference. *D. arizonae* use both *Opuntia sp.* and columnar cacti, whereas *D. m. mojavensis* uses *Ferocactus cylindraceus* as preferential host, *D. m. wrigleyi* uses only *Opuntia sp*., *D. m. baja* uses *Stenocereus gummosus*, and *D. m. sonorensis* uses *S. thurberi* (8). Therefore, this set of species from *buzzatii* and *mulleri* group provide an excellent model to investigate the genetic background underlying HS, providing a temporal evolutionary frame regarding the impact of HS on ecological speciation.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The radiation of *Drosophila* species across different hosts is an evolutionary outcome that gave rise to specific adaptations (9). Such adaptations may be summarized into three steps: location, acceptation and detoxification (10). The evolution in any of these steps may involve several morphological and sensorial structures, with genetic variability underlying each of them. Location is the step in which the fly will distinguish its specific host among others, through chemoreceptors that integrate the visual and olfactory systems. The odorant-binding proteins (OBPs) are encoded by a gene family with approximately ~50 genes and they are mostly expressed in the olfactive papillae (11). These globular proteins have the ability to bind odorant molecules and transport them to the odorant proteins (ORs), encoded by a gene family with ~60 genes, mostly expressed in the olfactory sensilla and antennae (12). Subsequently to the host location, acceptance is the second step. This is the step where the fly will evaluate the nutritional compounds present on the host, as well as the presence of competitors, predators, pathogens and parasites (10). Several receptors are associated with this process, such as gustatory receptors (GRs), that correspond to a gene family with ~60 genes expressed in the legs and wings, as well as head organs, proboscis, pharynx, and antennae (13). The recognition of a proper host as a feeding and breeding site is also performed by ionotropic receptors, encoded by a gene family with ~55 genes, with functions related to chemosensation, thermosensation, and hygrosensation (14–16). Finally, the flies must be able to complete the feeding and breeding processes, but it depends upon the overcoming of toxic substances that might be present on the host. Several gene families are associated with detoxification, such as P450s (CYPs), Glutathione S-Transferases (GSTs), UDP-Glicosiltransferases (UGTs), esterases and ATP binding-cassette (ABC transporters) (10,17). CYP genes comprise a gene family of ~80 genes, expressed in several organs, as antennae, maxillary palps, labellum, legs, wings and larvae (18). It has been shown that they play an important role into the insecticides resistance, through detoxification pathways (19). The GST family has ~40 genes, expressed in antennae, maxillary palps and larvae (20). The UGTs family comprises ~30 genes that are able to produce hydrophobic molecules that are eliminated with more efficiency (21). Esterases is a gene family constituted by ~30 genes, expressed in

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

larvae, pupae, head organs, testis and ovary (22,23), with function associated to several metabolic pathways, including detoxification (24). ABC transporters are a wide family transmembrane proteins, expressed in the *Drosophila* brain and eyes, with an important role into the transport of several substrates, including xenobiotics (17). These genes associated with the location, acceptance and use of hosts will henceforth be referred to as HLAU.

A previous study with 14 *Drosophila* genomes, representing both *Sophophora* and *Drosophila* subgenus, has proposed that 82 genes, including GSTs, CYPs, and UDPs, were under positive selection (25). Signatures of positive selection were also found between *D. mojavensis* subspecies and *D. arizonae* in subsets of gustatory and odorant receptor genes (16). In addition to the selective pressure acting into HLAU gene families, modifications in the gene expression should also play an important role in the HS adaptation in *Drosophila* species. Although detoxification occurs predominantly in digestive organs, it has been shown that genes related to stress response, olfaction and detoxification were differentially expressed in the brain of African and non-african populations (26). Similar results were found between *D. m. wrigleyi* and *D. m. sonorensis*, in which CYPs, GSTs and OBPs were found differentially expressed (27), indicating that gene expression in the nervous system might play a role into detoxification and olfactory systems, being able to potentially participate to HS adaptations in *D. mojavensis*.

The genetic variability is a major key to the adaptive process associated to HS. Transposable elements (TEs) are ancestrally mobile sequences with the ability to create structural variants and thus genetic variability. In most of the cases, mutations induced by TE mobility are likely to be deleterious or neutral. Throughout evolutionary time, TEs that remain in the genome tend to be silenced through epigenetic control, accumulating mutations, and losing their transposition ability (28). Despite they become degraded, the remaining TE sequences may still contain either regulatory motifs or protein domains/motifs (29). These sequences might be co-opted by the cell machinery modifying gene expression or protein sequences of the nearby genes (reviewed in (30)). Depending upon the positive outcome of

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

the cooption into the individual fitness, these TEs might give rise to the phenomenon called TE exaptation and domestication (31). For instance, in *D. melanogaster*, the *CHKov1* gene has a truncated mRNA derived from a *DOC* TE located into the intron, resulting in resistance to viral infection and insecticides (32). Exaptation and domestication events can be often identified by the occurrence of chimeric transcripts, which are mRNAs with both gene and TE-derived sequences (33). A recent transcriptome-wide study has identified that ~1.12% of *D. melanogaster* genes produce chimeric transcripts (34). Among all genes, 76 produce polymorphic chimeric transcripts from TE insertions that were present in one strain, but absent in another, highlighting the potential of TEs as a source of genetic novelty between different natural *Drosophila* ecotypes.

Many aspects of the HS adaptation in *repleta* species have been described, such as detoxification pathways (35), morphology (36), life history traits (37), behavior (38), and genomics and transcriptomics (27,39–42). However, the contribution of TEs to HLAU genes has not yet been assessed. Here we aimed to uncover the extension of variability derived from TEs in cactophilic *Drosophila* species, using both genome- and transcriptome-wide analyses. We tested the hypothesis regarding the potential contribution of TEs to the HS in cactophilic species, and therefore to the observed ongoing ecological speciation. The genome-wide analysis revealed that TEs have different evolutionary trajectories between cactophilic species, even between the recently diverged *D. mojavensis* subspecies. In addition, we also found specific TE families inserted preferentially at regulatory regions of OR genes, in six out of seven analyzed species. Our results demonstrate that ~35% of the ORs with TEs in the promoter have TFBSs within the TE copy. In addition, on average between species, 50% of all TE insertions carrying TFBSs are *Helitrons*, demonstrating their unprecedent putative role in *Drosophila*. The transcriptomics analysis revealed that HLAU genes are differentially expressed between cactophilic species, as expected due to their differential host preference. We also show that 2.1% of genes produce chimeric transcripts in head and larvae, for which HLAU genes are also involved. Taken together, our results

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

demonstrate the substantial contribution of TEs to both genomics and transcriptomics variability of cactophilic *Drosophila* species. We suggest that the involvement of TEs in the evolution of HS in species with differential host preferences can be the driving force of ecological speciation.

Results

Genome assemblies and gene annotation

To investigate the potential role of TEs into HS of cactophilic *Drosophila* species, we performed Nanopore long-reads sequencing for species that have a wide-range of preferential cacti as host: *D. buzzatii* (*Opuntia sp.*; columnar cacti), *D. koepferae* (columnar cacti), *D. arizonae* (*Opuntia sp.*; columnar cacti), *D. m. mojavensis* (*F. cylindraceus*), *D. m. wrigleyi* (*Opuntia sp.*), *D. m. sonorensis* (*S. thurberi*), and *D. m. baja* (*S. gummosus*). We obtained high resolution assemblies for all genomes, with an average of 567 scaffolds, 13.8 Mb of N50 (Sup. Table 1), and ~98% of BUSCO genes (Sup. Fig. 1). Furthermore, the gene annotation for these genomes covered 98.63% of all reference genes in the *D. mojavensis* subspecies, whereas for *D. arizonae*, *D. buzzatii* and *D. koepferae* we obtained 95.98%, 75.42%, and 76.03% of the reference genes, respectively. The resolution of gene annotation for *D. arizonae* was similar to *D. mojavensis* subspecies due to the recent (1.5 Mya) divergence between them (43), whereas for *D. buzzatii* and *D. koepferae* it was ~23% smaller compared to *D. mojavensis*, as expected since they are more distant related species (44). Similar amount of genes had been annotated previously in the *D. buzzatii* (45,46) and *D. koepferae* (46,40) genomes. Since analysis on HLAU genes is the focus of this study, their efficient annotation is relevant. They are represented by nine gene families associated with the three steps on the use of host resources: (1) Location: OBPs, ORs; (2) Acceptance: GRs, IRs; and (3) Host use: ABCs, ESTs, GSTs, UDPs, CYPs (Sup. Table 2). In *D. arizonae* and in the *D. mojavensis* subspecies, we recovered ~98.21% of the total HLAU genes from the reference genome, whereas in *D. koepferae* and *D. buzzatii* the proportion of annotated HLAU genes was 76.56% and 74.91%, respectively (Table 1). This decrease in number of genes is likely to be associated

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

with the genomic annotation resolution for these species than cases of gene loss, since for these species the total HLAU genes compared to the reference was ~23% less.

Species/subspecies	Total genes	Host location, acceptance and usage genes								
		Location		*Acceptance*		*Host use*				
		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. baja	14,797	25	57	44	22	27	17	27	19	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	11,347	18	39	29	16	21	15	25	17	52
D. koepferae	11,440	19	40	26	16	22	16	21	16	51

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

Signatures of positive selection on HLAU genes between species with Opuntia sp. preference and columnar cacti

The use of *Opuntia sp.* has been proposed to be an ancestral state in cactophilic *Drosophila* species, whereas columnar cacti is the derived state (4). Therefore, we used branch-site approach to investigate whether species with preferential columnar cacti usage have HLAU genes that undergone positive selection, compared to species that use *Opuntia sp*. For this analysis, we removed *D. arizonae* due to its general use of both columnar and *Opuntia* cacti (4). Moreover, we added *D. navojoa*, which is a basal species of the *mulleri* complex that use solely *Opuntia* sp. (48). Therefore, we carried out the branch-site selection test with *D. buzzatii*, *D. m. wrigleyi*, and *D. navojoa* as *Opuntia*-related branches (background), and *D. m. baja*, *D. m. mojavensis*, *D. sonorensis*, and *D. koepferae* as columnar-related

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

branches (foreground). It is important to highlight that due to the need of one-to-one ortholog genes between the seven genomes, our analysis have decreased the total number of HLAU genes to ~42% of the initial number depicted on Table 1. From 128 genes, six were found under positive selection (Fig. 1, Sup. Table 3). Two of them were associated with location: *Obp56d* (*p-value = 0.014*) and *Or1a* (*p-value = 0.0444*); one with acceptance: *Ir25a* (*p-value = 0.0303*); and three with detoxification: *Cyp310a1* (*p-value = 0.0218*), *Cyp4s3* (*p-value = 0.0356*), and *Udp-ugt5* (*p-value = 0.0216*).

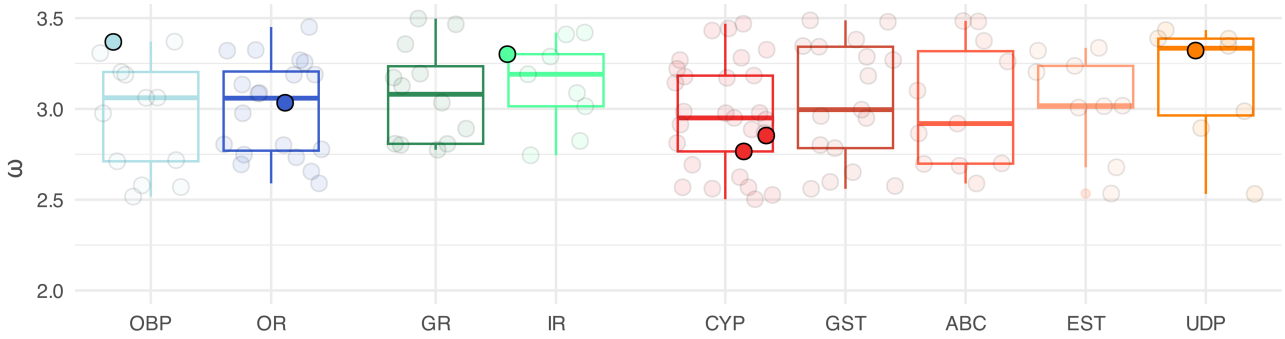

Figure 1. Branch-site signatures of selection per HLAU gene family between *Opuntia*-related *Drosophila* species and columnar-related species. X axis has the nine HLAU gene families: OBP: odorant-binding protein, OR: odorant receptor; GR: gustatory receptor, IR: ionotropic receptor, CYP: cytochrome P450, GST: glutathione S-transferase, ABC: ATP-binding cassette transporter, EST: esterase, UDP: UDP-glycosyltransferase. Distribution of ω (dN/dS) in y axis, each dot represents a gene; filled dots are genes under positive selection with p-value < 0.05 for the likelihood ratio test.

For each gene with signatures of positive selection, we searched in the seven genomes if they have TEs inserted within or nearby (upstream and downstream regions). In *D. mojavensis*, the gene *Cyp4s3* has a common *Unknown* TE inserted upstream in the four subspecies, and *D. m. baja* and *D. m. wrigleyi* have a LINE located upstream to the *Or1a* gene. Finally, in *D. buzzatii* we observed an *Helitron* insertion downstream to the *Obp56d*, and no TEs associated with these genes in *D. koepferae*. Our results suggest that HLAU genes under positive selection are not likely to have TE-derived motifs, and there is no association between diversifying selection and TEs enrichment on regulatory regions.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Evolutionary dynamics of TEs in cactophilic Drosophila species

We used a combination of automatic and manual methods to build a curated TE-library for each species. The method was designed to perform polishing steps in both TE consensus and copies to remove artefacts (see Methods). The total number of TE consensuses (LTRs separated of internal sequences) ranged from 243 in *D. m. baja* to 313 in *D. m. mojavensis*. The total TE content in the genomes ranged from 14.54% in *D. buzzatii* to 22.10% in *D. m. mojavensis* (Fig. 2A). In addition, we observed a positive correlation between TE content and genome size estimated from the assemblies (*Pearson* correlation: $p = 0.009$; $r = 0.88$) (Sup. Fig. 2). Following the correlation, both species from the *buzzatii* complex had the lower genome sizes and TE contents compared to species of the *mulleri* complex (Fig. 2A). Regarding the proportional load of TE orders in the genomes, LTRs contribute with 31.21% and 31.68% in the sister species *D. buzzatii* and *D. koepferae*, respectively. The main differences in the mobilomes are due *Helitrons* and LINEs (Fig. 2B). *D. buzzatii* has 5.44% more genomic content derived from *Helitrons* compared to *D. koepferae*, whereas *D. koepferae* has 8.16% more LINEs. DNA elements have a slight difference between them, corresponding to 3.64% more TE content in *D. buzzatii*.

In the *mulleri* complex (*D. arizonae* and the four *D. mojavensis* subspecies), the abundance of TEs in the genome varied substantially (Fig. 2B), despite their recent divergence time of 1.5 Mya (43). *D. m. mojavensis* has the TE-richest genome (22.1%), with the abundance following from the highest to the lowest: LTR > LINE > Helitron > DNA > Unknown (Fig. 2B). The distribution of TEs in *D. arizonae* and the other three *D. mojavensis* does not follow such order. All of them have LTRs with the highest abundance, followed by Helitrons, LINEs, DNA, and Unknown (Fig. 2B). In these genomes, *Helitrons* represents ~28% of the mobilome, whereas only 23.12% in *D. m. mojavensis*. On the other hand, LINEs comprise 24.34% in *D. m. mojavensis*, and ~20% in the other species (Fig. 2B). The antagonistic evolution of LINEs and Helitrons in *D. m. mojavensis* indicates a potential recent burst of LINEs, while Helitrons might be silenced.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The evolutionary dynamics of TEs in the genomes was assessed with the TE landscape approach, in which we can observe the genomic contribution of TE insertions, and their respective divergence compared to the full-length consensus. In *D. buzzati* and *D. koepferae*, the higher content of *Helitrons* in the former species is likely to be associated with recent transposition burst followed by silencing. Such pattern is observed in *D. buzzatii* due to the peak derived from conserved copies (K2P near to zero), followed by a quick depletion, which is not observed on *D. koepferae* (Fig. 2C). In addition, LINEs in *D. koepferae* had a short transposition wave that has been triggered before silencing (Fig. 2C), which did not happen in *D. buzzatii*. This evolutionary landscape explains why *D. koepferae* has 21% of the mobilome derived from LINEs, while *D. buzzatii* has 13%.

In the species of the *mulleri* complex, the higher distinctions of the mobilomes are associated with Helitrons and LINEs. LTRs did not show signatures of differential transposition rates over time, despite the observed differential accumulation, with low divergence elements ranging from ~0.9% in *D. m. baja* to ~2.2% in *D. arizonae* (Fig. 2C). Regarding the order Helitron, the evolutionary dynamics of this order in the five genomes suggests a highly common trajectory between the pair of subspecies *D. m. mojavensis* and *D. m. wrigleyi*; and the pair *D. m. baja* and *D. m. sonorensis*. It is noteworthy that the former two species are in the northern region of the *D. mojavensis* distribution, whereas the latter subspecies are located the southern region. The landscape highlights a common accumulation of *Helitrons* among the four subspecies, but a narrow and recent decrease in the northern subspecies, whereas a longer persistence in the southern. The sister species, *D. arizonae*, had another evolutionary dynamics, in which Helitrons overcame selection barriers, increasing in frequency and remaining active in the genome (Fig. 2C). Finally, Unknown TEs share narrow peaks (quick transposition/expansion followed by silencing) between *D. m. mojavensis* and *D. m. baja*. Overall, these results reveal that transposition dynamics might be associated with geographical location and interaction with environment, as we observed for Helitrons, but it may also be stochastic, as it has been shown for Unknown TEs.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

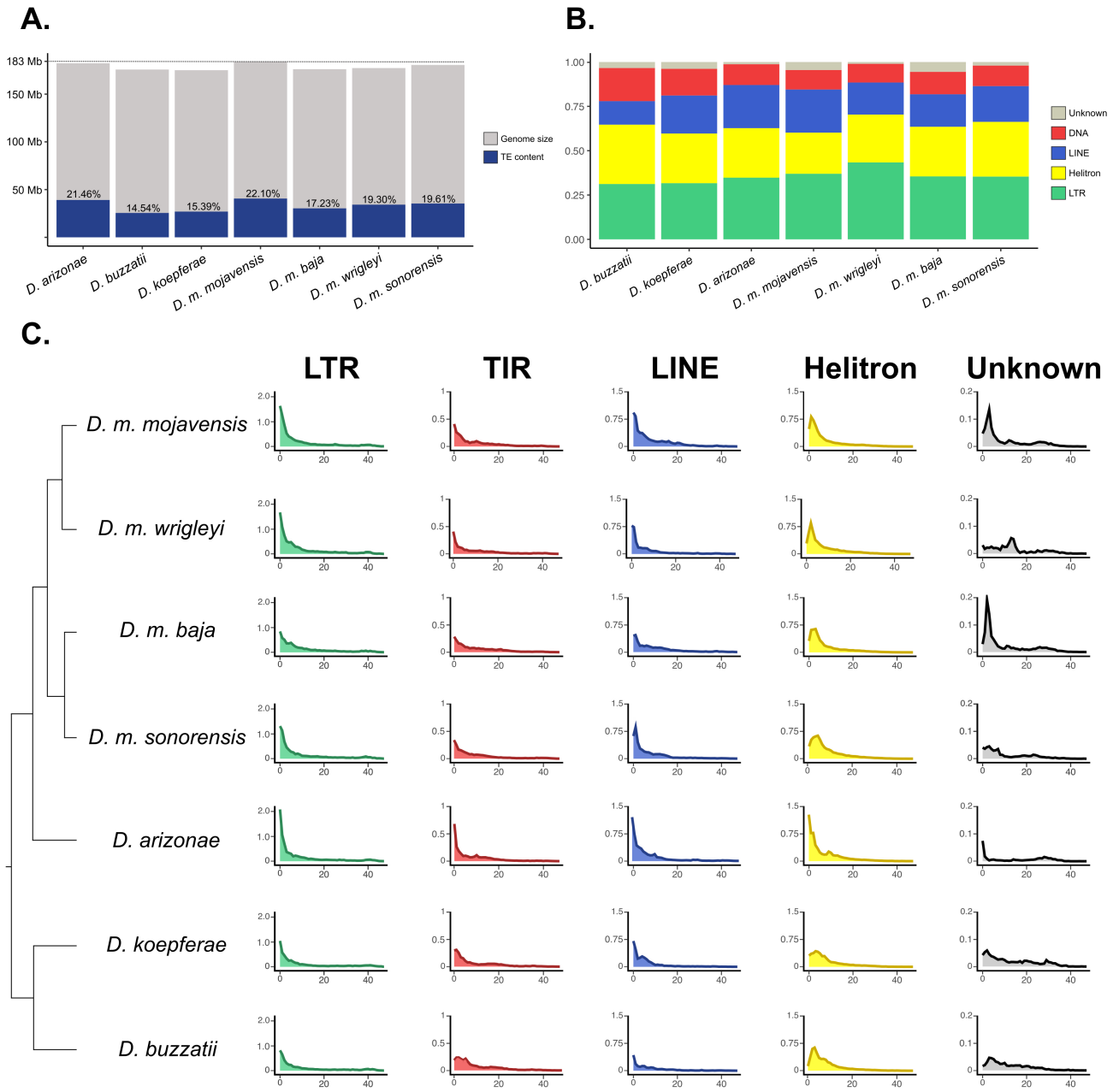

Figure 2. A) Genome size of the seven cactophilic *Drosophila* species sequenced in this work, and their respective TE content. **B**) Proportional distribution of the total TE content annotated in the genomes, divided by DNA transposon, LINE, Helitron, LTR, and Unknown. **C**) Phylogenetic relationship among species (adapted from (35,49)), alongside with TE families landscapes representing the distribution of TE content (y-axis), and the divergence of TE insertions compared to the consensuses (x-axis), based on Kimura 2-parameters corrected with CpG. The TE content located on the left of the graph is represented by TE insertions with low divergence with their respective consensus, hence inferring that they might be conserved/recent/active, whereas TE content on the right part represents copies with high divergence due to the accumulation of mutations, corresponding to old/inactive copies.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

HLAU genes are enriched in TEs at regulatory regions

To investigate whether TEs might have been maintained overtime in a non-random distribution in the genomes relative to genes, we compared their positions and analyzed their frequencies in each genome. In the absence of insertion bias and selection, TEs would have a common distribution inside and outside genes. Then, we analyzed whether some TE orders are preferentially intragenic. To do so, we compared the frequency of TE families that are preferentially inserted within gene regions (>50% of all insertions overlapping with genes), with a random distribution of TE insertions from the same order. We observed that there is no TE order enriched within genes, which is expected with the hypothesis of purifying selection against TEs disrupting gene sequences (Sup. Fig. 3).

Despite TEs were overall not enriched on gene regions, a few insertions may be selected in the promoter of genes if they provide histone marks enrichment, or *cis*-regulatory elements beneficial to the individual fitness (50,51). We performed permutation tests to investigate whether TEs are enriched in the promoters (2 Kb upstream of the TSS) from the nine HLAU gene families in the seven genomes. Regarding genes associated with location (OBPs and ORs), we observed that ORs have enrichment of TEs at the promoters in all species, except *D. buzzatii* (*p-value = 0.0719*) (Fig. 3A). For acceptance genes, *D. buzzatii* and *D. koepferae* have enrichment of TEs in the promoters of GRs, whereas *D. m. baja* species has enrichment for the IRs. (Fig. 3A). Finally, enrichment of TEs upstream to detoxification genes was observed only in *D. arizonae*, for the CYP and Esterase families (Fig. 3A). The results depicting the histograms with frequency of genes with TEs for each permutation test, and p-values are available in the Sup. Fig. 4 and Sup. Table 4. To obtain further insights in the potential regulatory role of TEs, we also analyzed the enrichment at the 2 kb downstream region. In *D. buzzatii* and *D. koepferae*, we did not observe neither significant enrichment nor depletion, indicating that the frequency of TEs downstream to HLAU genes follows the genomic distribution in these species. In the species of the *mulleri* complex, ORs had enrichment for TEs in the downstream region, as we have observed also for the 2 kb

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

upstream. In addition, OBPs had a significant depletion of TEs in all genomes from *mulleri* species, except in *D. m. baja*. Finally, in the four gene families associated to detoxification, we observed that esterases are enriched by TEs in *D. arizonae*, while GSTs are depleted by TEs in *D. m. wrigleyi* and *D. m. sonorensis*. Our results suggest a possible ancient evolutionary path of TE cooption in cactophilic species, since enrichment of TEs in ORs was observed in almost all genomes. The enrichment of TEs upstream and downstream of ORs indicates a possible coevolution of this gene family with TEs in cactophilic species, or that ORs are located in TE-rich regions of the genomes. To investigate this hypothesis, we identified TE-rich regions (Sup. Fig. 5) and analyzed the frequency of ORs. We observed on average 6 and 3 ORs in such regions in species of the complex *mulleri* and *buzzatii,* respectively (Sup. Table 5). Therefore, we rejected the hypothesis that OR genes are enriched by TEs in their vicinity due to their TE-rich genomic location. Other factors might be involved in the evolutionary maintenance of TEs in the OR's vicinity, such as the cooption of regulatory motifs.

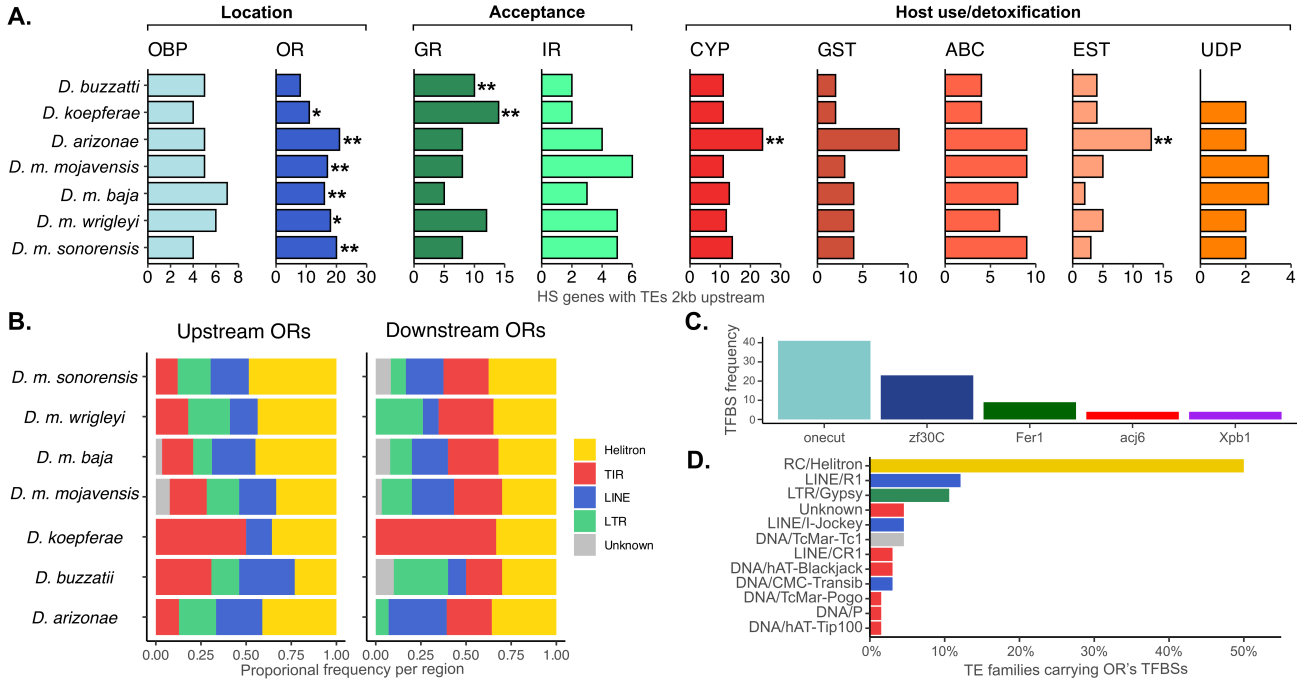

Figure 3. A) Total number of genes from each HLAU gene family with TEs located at 2kb upstream. * = p-value < 0.05; ** = p-value < 0.01 for the permutation tests; the absence of stars stands for non-significant result. **B**) Proportion of TE orders located upstream and downstream ORs. **C**) Transcription factor binding sites (TFBSs)

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

frequency on TEs located upstream OR genes in all species under study. **D**) Proportional frequency of TE families with TFBSs embedded on their sequences from TE copies located upstream OR genes.

A few TE families have been shown to be more prone to contribute with *cis*-elements than others, such as retroelements in *Drosophila* (52). Therefore, we investigated whether the observed enrichment of TEs in the HLAU gene families is due to specific TEs. Strikingly, considering all species, 40.29% and 33.53% of the insertions located upstream and downstream of ORs correspond to *Helitrons*, respectively (Fig 3B). Helitrons have been documented to act on gene evolution of bats by carrying gene sequences (53), mainly 5' UTRs and the first exons, within their own internal regions (54). Thus, we investigated whether these *Helitron* insertions upstream to HLAU genes, mainly ORs, could be part of the first exons from genes. By intersecting exons from genes enriched with their respective upstream *Helitrons*, we observed that only *Or13a* and *Or22c* have their first exons overlapping with the TEs. The former was observed only in *D. m. wrigleyi*, which indicates a subspecies-specific insertion, whereas the latter is observed in the four *D. mojavensis* subspecies, suggesting that this insertion has been maintained from their common ancestor (Sup. table 6). At the OR's downstream region, we observed five genes with *Helitrons* overlapping exons: *Or71a*, *Or10a*, *Or13a*, *Or82a* and *Or30a*. Interestingly, the same *Helitron* in the *Or82a* was observed in the four *D. mojavensis* subspecies, highlighting the inheritance from the common ancestor and maintenance of the insertion overtime (Sup. table 6). Taken together, our results rejected the hypothesis regarding the possible role of Helitrons carrying HLAU genes fragments. However, these elements could still act on gene regulation if they carry *cis*-elements such as TFBSs.

TEs nearby gene regions have been shown to modulate gene expression, through donation of TFBSs or polyA sites to the ancestral regulatory motifs. To investigate the potential functional role of the observed enrichment of TEs in the upstream region of ORs, we carried out TFBSs identification for all TE insertions located in such position. We used the TFs previously described to participate on OR's transcription in *D. melanogaster*: acj6, onecut,

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

xbp1, fer1 and zf30C (55). Our analysis revealed that ~35% of the ORs with TEs at the promoter region have the TFBS located within the insertion (Sup. table 7). In *D. koepferae*, 7 out of the observed 11 ORs (63.63%) had TEs carrying OR's TFs. Taking all species together, the most frequent TFBSs found was for the TF onecut (50.61%), followed by zf30C (28.39%), and Fer1 (11.11%) (Fig. 3C). *Helitrons* were the TEs with the highest frequency, representing 50% of all insertions carrying ORs' TFBSs (Fig. 3D). A few cases of ancestral inheritance were observed, for instance, the *Or22c* and *Or82a* have a conserved *Helitron* insertion carrying the onecut TFBS in the five genomes from the *mulleri* complex (Sup. table 7). We also observed a case of loss, where the *Or85d* has the TFBS Fer1 in the three *D. mojavensis* subspecies that uses primarily columnar cacti as host, but absent in *D. m. wrigleyi*, which uses preferentially *Opuntia sp.* These results highlight a putative functional role of TEs as donators of *cis*-elements to ORs in cactophilic *Drosophila* species.

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

Differential expression of genes, including HS ones, has been associated with adaptation to new hosts in insects (41). Here, aiming to understand how gene expression divergence is associated with adaptation to different cacti hosts, we performed differential expression analysis between species with preference for *Opuntia sp.* and columnar cacti, in head and larvae tissues. Overall, the principal component analysis (PCA) revealed remarkable differences between each species in both tissues (Fig. 4A-B). In the head tissue, species of the *buzzatii* complex are clustered together and separated from those of the *mulleri* complex based on the PC1 (45%), as we observed in larvae (Fig. 4B). In *mulleri* complex, differentially from larvae, we found *D. mojavensis* subspecies in a distribution following the PC2 (19%), for which *D. m. mojavensis* and *D. m. sonorensis* exhibit the highest variance (Fig. 4B). In larvae, the variance does not follow the same trend. In this tissue, the variance of gene expression resembles the phylogenetic divergence of the species, with the highest variance between the complexes *mulleri* and *buzzatii*, followed by divergence between *D. buzzatii* and *D. koepferae*.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

The species of the *mulleri* complex were clustered together with a slight separation between *D. arizonae* and *D. m. wrigleyi* from the other three *D. mojavensis* subspecies (Fig. 4A), whereas *D. buzzatii* and *D. koepferae* were separated from *mulleri* species based on the PC1 variance (50%), and from each other based on PC2 (18%) (Fig. 4A). Interestingly, *D. arizonae* and *D. m. wrigleyi* had the lower variance in head, as it has also been observed in larvae.

The first pairwise analyses were between species with preference for *Opuntia sp.* versus species with non-preferential cacti. Therefore, we compared the transcriptome of *D. m. wrigleyi* vs *D. arizonae* and *D. buzzatii* vs *D. koepferae*. The result demonstrates that HLAU genes have divergent differential expression in both head and larvae tissues (Fig. 4C). In head, the *Opuntia* sp. dweller *D. m. wrigleyi* had 31 HLAU genes with higher expression compared to its counterpart *D. arizonae*; whereas *D. buzzatii* had 24 HLAU genes up-regulated compared to *D. koepferae*. Among them, the *Obp99a*, *Or49b*, and *Cyp4p1* were found in both species, demonstrating a possible adaptive role for *Opuntia sp.* preference. On the other hand, the cacti-generalists *D. arizonae* and *D. koepferae* had 31 and 59 HLAU genes over-expressed compared to their counterparts. The genes *UDP-Ugt5*, *Gst1-1*, *Gr94a*, and *Obp19d* were found in both species, thus suggesting a possible molecular mechanism for convergent evolution of columnar cacti preference between species. In larvae, *D. m. wrigleyi* and *D. buzzatii* had 43 and 48 HLAU genes with higher expression compared to *D. arizonae* and *D. koepferae*, respectively. Eight genes were shared between them: *Obp56a*, *EstB1*, *UDP-Ugt4*, *Gst1*, *Gst1-1*, *Cyp9f2*, *Cyp6a14*, *Cyp4ad1*, and *Cyp4d14*. Except by *Obp56a*, all genes with high expression compared to the species with non-preferential cacti were associated with detoxification. Noteworthy, the *Gst1-1* has been found up-regulated in both head and larvae of species with preference for *Opuntia sp*. Finally, *D. arizonae* and *D. koepferae* had 49 and 22 differentially expressed HLAU genes compared to their counterparts. Only *Or85f* and *UDP-Ugt5* were observed in both species.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

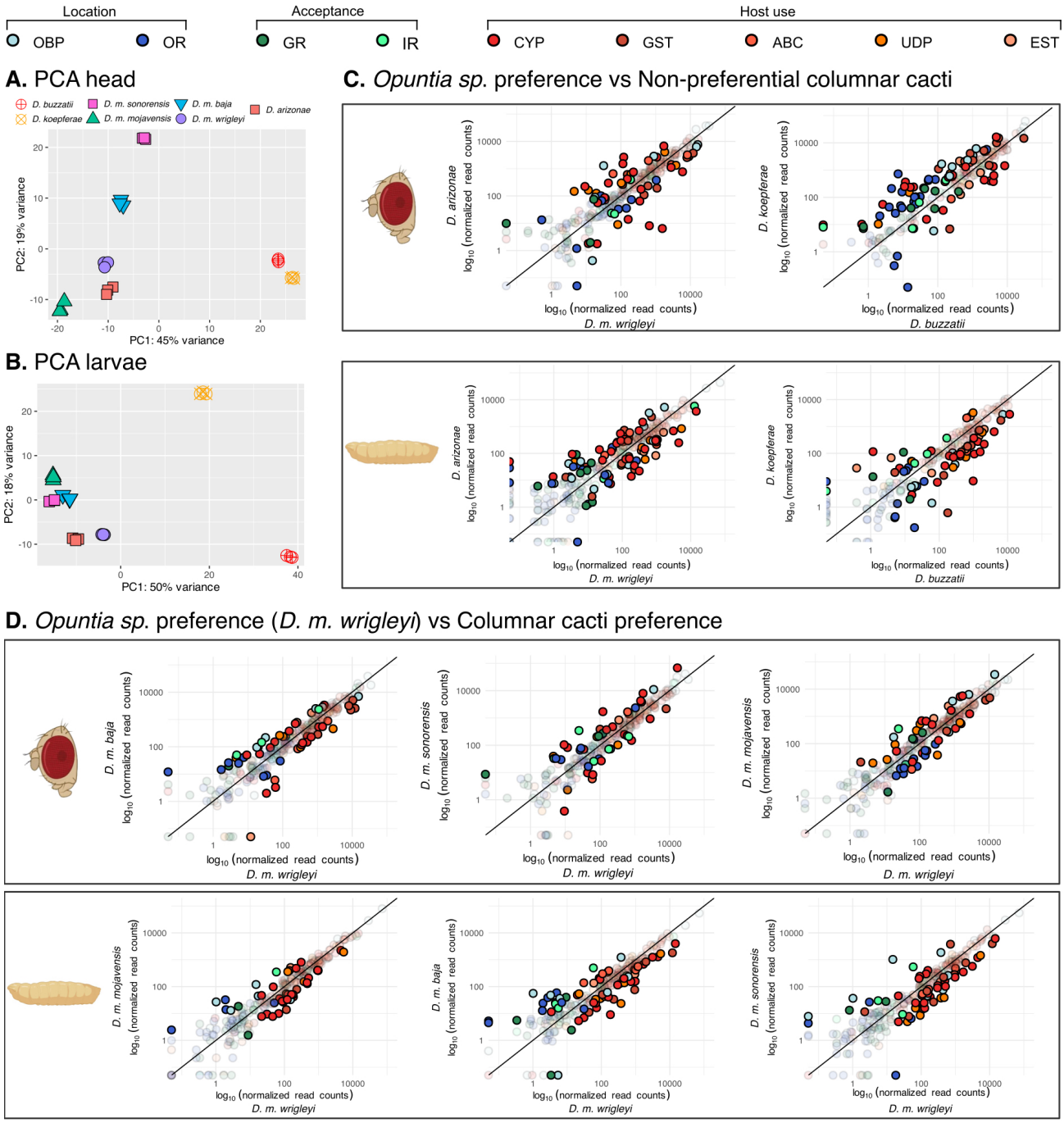

Figure 4. Principal component analysis of larvae and head transcriptomes in the seven *repleta* genomes. **a**) Head transcriptome: The variance of *buzzatii* species compared to *mulleri* corresponds to the PC1 (45%), with strong similarity between *D. buzzatii* and *D. koepferae*. In *mulleri* species, we observed a broader distribution compared to larvae corresponding to the PC2, except for *D. arizonae* and *D. m. wrigleyi*. **b**) Larvae transcriptome: As observed in head, the species of the complexes *mulleri* and *buzzatii* are separated by the PC1 with maximum 50% of variance. *D. buzzatii* and *D. koepferae* have their variance on gene expression separated by the PC2 (18%), whereas the species of the *mulleri complex* are clustered together, with two subclusters from *D. arizonae* with *D. m. wrigleyi*, and the other three *D. mojavensis* subspecies. **c**) Scatter plots depicting normalized expression of HLAU genes between *Opuntia sp.* dwellers (*D. m. wrigleyi*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

and *D. buzzatii*) and non-preferential cacti dwellers (*D. arizonae* and *D. koepferae*). **d**) Differential expression analysis of the *Opuntia sp.* dweller *D. m. wrigleyi* versus the three *D. mojavensis* with preference for columnar cacti. In C and D: Filled dots represent HLAU genes with log2 fold-change > |1| and FDR < 0.05; transparent dots are non-significant gene expression.

We also assessed differential expression of HLAU genes between *D. mojavensis* subspecies, comparing *D. m. wrigleyi*, which has preference for *Opuntia sp.*, to the other three subspecies with preference for columnar cacti (Fig. 4D). In head, *D. m. wrigleyi* had 27, 24, and 19 HLAU genes with higher expression compared to *D. m. mojavensis*, *D. m. baja*, and *D. m. sonorensis*, respectively (Fig. 4D). Four genes were found with higher expression in the head of *D. m. wrigleyi* in the three comparisons: *UDP-ugt4*, *Gst-theta3*, *Abc-a3*, *Cyp6a21*. The *UDP-ugt4* has also been observed previously in the differential expression against *D. arizonae* and between *D. buzzatii* versus *D. koepferae*, indicating conservation of higher expression on *Opuntia sp.* dwellers also in shorter evolutionary scales. In the subspecies with preference for columnar cacti, we observed on average 26 HLAU genes with higher expression compared to *D. m. wrigleyi*. Among them, seven were recurrent in the three subspecies: *Ir21a, EstB1*, *UDP2*, *UDP-Ugt5*, *Cyp9b2*, *UDP-Ugt5-1*, *Cyp12a2*. Finally, the same analysis on larvae revealed that *D. m. wrigleyi* has three genes with recurrent higher expression compared to the subspecies: *Or82a*, *Obp59a*, and *Ir21a*; whereas in *D. mojavensis* subspecies we observed 17 common up-regulated genes compared to *D. m. wrigleyi*, which comprise seven CYP genes, three GSTs, four UDPs, one Esterase and one OBP. Taken together, the quantitative results suggest that adaptations that have allowed the shift to columnar cacti may arise from the evolution of gene expression in both head and larvae.

The enrichment of TEs on HLAU genes is not associated with changes on gene expression

Since we observed enrichment of TEs on the regulatory regions of a few HS families, mainly ORs, and we also found differentially expressed HLAU genes between species, we tested whether the presence of TE insertions is associated with differential expression. To do so, we selected expressed HLAU genes (see Methods) and performed X^2 independency test, splitting them into four categories: differentially expressed, non-differentially expressed, with

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

TEs at 2kb upstream, without TE at 2kb upstream. Considering all pairwise analysis (Fig. 4C-D), the X^2 results demonstrated that the differential expression of HLAU genes is not associated with the presence of TEs on the promoter region (Sup. Table 8), except in *D. m. sonorensis* compared to *D. m. wrigleyi* ($X^2 = 8.18$, *p-value = 0.004*). Since we did not observe overall effect of TEs on the expression of all HLAU genes together, we investigated the association of TEs with differential expression of HLAU genes from families with TE enrichment (Fig. 3A). None of the enriched gene families had significant association between the presence of TEs and the expression level (Sup. Table 8). Thus, enrichment of TEs at the regulatory region is not likely to be associated with the differential expression of HLAU genes.

Gene-TE chimeras are produced from 298 and 285 genes in the head and larvae transcriptomes

TEs inserted within and nearby genes might be incorporated on gene transcripts, either acting as regulatory elements, or donating protein motifs to the CDS. Despite highly deleterious, a small fraction of TEs generating chimeric transcripts can be positively selected, giving rise to domestication and exaptation events (31). We analyzed the transcriptome of head and larvae tissues in the seven cactophilic species, aiming to identify the extent of the transcriptome variability generated by chimeric transcripts. Considering the whole transcriptome, our results revealed that 298 (2.15%) and 285 (2.06%) genes produce chimeric transcripts in all species, in head and larvae, respectively. In both tissues, TE-exonized transcripts represented the highest frequency, with 67.56% in head and 70.37% in larvae, followed by TE-terminated transcripts with 21.55% and 17.79%, and TE-initiated transcripts with 11.12% and 12%.

To investigate whether specific TE families could be more prone to produce chimeric transcripts, we analyzed their frequencies in chimeras. Regarding TE-initiated transcripts, we observed that the most frequent families are not the same between species. *D. koepferae* had TIRs with the highest frequency, whereas *D. buzzatii* had *Helitrons* (Sup. Fig. 6). Among species of the *mulleri* complex, *Helitrons* were the most frequent, generating in average

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

~27% of TE-initiated transcripts in *D. arizonae*, *D. m wrigleyi* and *D. m. sonorensis*, whereas the family *Gypsy* was the most frequent (~28%) in *D. m. mojavensis* and *D. m. baja* (Sup. Fig. 6). These results demonstrate that co-option of TE-derived sequences as promoters is not family-specific on cactophilic *Drosophila* species. In addition, our results highlight the rapid turnover of TEs acting as cis-elements, mainly demonstrated by *Helitron* and *Gypsy* between recently diverged *D. mojavensis* subspecies.

In TE-exonized transcripts, we observed high conservation among TE families between *mulleri* and *buzzatii* complexes. In all genomes from species of the *mulleri* complex, we found *Gypsy* as the most frequent, followed by *Helitrons* and *R1*. Despite differences in the proportion, such as *R1* with ~10% in *D. m. sonorensis* and ~22% in *D. arizonae*, the common presence of these families on TE-exonized transcripts indicate that either they are old insertions, shared through common ancestors, or these superfamilies are preferentially inserted within gene regions. In *D. buzzatii* and *D. koepferae*, *Helitrons* have the higher frequency, producing ~44% and ~28% of all TE-exonized chimeras in these species, respectively. Due to the high prevalence of certain families in these chimeras, we analyzed whether TE insertion may contain protein domains, which provides evidence for potential domestication events. Chimeras derived from insertions in introns or overlapped with exons might still produce TE-initiation or TE-termination as alternative isoforms. Therefore, we focused our analyses only on TE insertions embedded in CDSs.

TE-exonized transcripts represent 68.96% of all chimeras from both head and larvae, but we observed only 13 chimeras with TE copies carrying conserved TE-related protein domains within CDSs (Sup. table 9). *D. arizonae* and *D. m. mojavensis* had only one chimeric transcript; *D. m. sonorensis, D. m. wrigleyi,* and *D. m. baja* had two; *D. koepferae* had five; and *D. buzzatii* had no one. Interestingly, *D. arizonae*, *D. m. mojavensis*, *D. m. sonorensis*, *D. m. baja*, and *D. m. wrigleyi* had the same gene *E3 SUMO-protein ligase ZBED1-like* (*LOC116803923*) producing a chimeric transcript with one or two *hAT-Ac* insertions. These copies provide a *Dimer_Tnp_hAT* and/or *zf-BED* protein domains to the CDS (Sup. table 9).

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

In *D. arizonae*, only *Dimer_Tnp_hAT* was observed; whereas *D. m. baja* and *D. m. mojavensis* only *zf-BED* was found; and finally, *D. m. wrigleyi* and *D. m. sonorensis* have both protein domains. The common presence of the same TE-exonized transcript in all *mulleri* species, and the absence on *buzzatii* species, indicates that these *hAT-Ac* copies have their origin in the common ancestor of *D. arizonae* and *D. mojavensis*. The other protein domains found were *Peptidase_A17* and *Peptidase_C1*, in which the former was observed in *D. m. baja* with the uncharacterized gene *LOC122757269*, and the latter in *D. koepferae* with *cathepsin L* (LOC6583090), *procathepsin L* (LOC6580571), and putative *cysteine proteinase* (LOC6574357) genes.

TE-terminated transcript was the only category of chimeras with the same TE family having the highest frequency among species (Sup. Fig. 6). *Helitrons* ranged from 22% (*D. m. mojavensis*) to 55% (*D. koepferae*). Therefore, we propose that *Helitrons* may have preference to insert themselves downstream to genes, and additionally they might provide polyadenylation sites (PASs) to the gene transcript, giving rise to TE-terminated chimeras. Our results demonstrate unprecedented evidence of *Helitrons* potentially playing a role as signaling sequences to stop Pol II transcription in 259 genes from head and larvae tissues. The recurrence of such characteristic in the seven species studied here supports this possible evolutionary role of *Helitrons*.

Conserved chimeric transcripts between species suggests potential domestications

Somatic chimeric transcripts might be produced by new insertions without any change in the offspring fitness. Moreover, even when TEs are present in the germline, part of the chimeras detected in the transcriptome are likely to be produced by alternative splicing malfunctioning in a small set of cells, producing TE exonizations with no functional value. To infer the functional relevance of the chimeric transcripts, we analyzed the recurrence of chimeras between species, and we observed that most of them are species/subspecies-specific in both head and larvae tissues (Sup. Fig. 7). These chimeric transcripts might

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

represent either recent insertions retained in the populations through positive selection or genetic drift, or they are product of pervasive transcription. Therefore, we used chimeras shared between species as a proxy of potential adaptive insertions. We selected three nodes in the species phylogeny: 1) *buzzatii* subgroup; 2) *mulleri* subgroup, and 3) *D. mojavensis* subspecies (Sup. Fig. 8). We did not find any ancestral chimera shared between all species. In the other hand, in the *buzzatii* subgroup, we found 50 common chimeras in head and 27 in larvae; whereas 84 (head) and 86 (larvae) between *mulleri* species; and 94 (head) and 90 (larvae) between *D. mojavensis* subspecies (Sup. Table 10). This result demonstrates that chimeric transcripts are likely to be lost or undetectable over longer evolutionary times (9.6 Mya between *mulleri* and *buzzatii*), but still detectable in shorter scale (1.5 Mya within *mulleri* and 4.5 Mya within *buzz*atii). In addition, it is noteworthy the relationship between higher number of common chimeras and lower divergence time of each node.

To identify chimeras possibly associated with the host usage of *Opuntia sp.* and columnar cacti, we selected genes generating common chimeric transcripts solely on species that use preferentially *Opuntia sp.* (*D. buzzatii* and *D. m. wrigleyi*), species using preferentially columnar cacti (*D. m. sonorensis*, *D. m. mojavensis* and *D. m. baja*), and species with non-preferential cacti usage (*D. koepferae* and *D. arizonae*). Regarding *Opuntia sp.* dwellers, we observed 12 common chimeras. There is no common TE-initiated transcript, and only one TE-terminated transcript (*LOC6585106*: *rho GTPase-activating protein 190 – LTR/Pao*). In TE-exonized transcripts, we observed five in head and six in larvae (Table 2). In the group of species using preferentially columnar cacti, we observed 93 chimeras from head and larvae, in which 89.24% corresponds to TE-exonized transcripts (Table 2). Finally, species with non-preferential cacti usage had 20 chimeric transcripts, with the vast majority (80%) also from TE-exonized transcripts. The low proportion of common chimeras observed in *Opuntia sp.* dwellers and species with non-preferential host are expected, since these groups are composed by species from *buzzatii* and *mulleri*. In addition, we did not find any common chimeras between all species from the two complexes, the 12 and 20 chimeras found in

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Opuntia sp. dwellers and non-preferential usage groups are either produced by evolutionary convergence, in which the same TEs have been coopted independently, or they have been shared through the common ancestor and have been lost in the other species overtime.

	TE-initiated		**TE-exonized**		**TE-terminated**	
	Head	***Larvae***	***Head***	***Larvae***	***Head***	***Larvae***
Opuntia sp. dwellers	0	0	5	6	1	0
Preferentially columnar cacti	3	1	38	45	5	1
Non-preferential usage	1	1	7	9	2	0

Table 2: Number of chimeric transcripts shared solely between group of species with the same preferential cacti host. "*Opuntia sp.* dwellers" comprise *D. buzzatii* and *D. m. wrigleyi*; "Preferentially columnar cacti" stands for *D. m. mojavensis*, *D. m. baja* and *D. m. sonorensis*; "Non-preferential usage" represents the group of *D. arizonae* and *D. koepferae*.

HLAU genes generate chimeric transcripts

To investigate whether TEs produce chimeric transcripts with HLAU genes, and their possible role in the host preference between species, we focused our analysis on the nine HLAU gene families. In the head and larvae tissues, we observed an average of ~15 and ~10 genes producing chimeric transcripts in the seven cactophilic species/subspecies, respectively. Indeed, we found more chimeric transcripts in head than in larvae for all HLAU gene families, except for UDPs (Fig. 5A). For instance, only head tissue had chimeras in the OBP and GR gene families (Fig. 5A). TE-initiated transcripts constitute 12.03% of all chimeras found in head, and 13.88% in larvae (Fig. 5B-C). TE-exonized transcripts had a slightly higher proportion in larvae (76.38%) compared to head (69.44%). Finally, TE-terminated transcripts were the ones with the biggest difference between tissues, with 18.51% in head and 9.72% in larvae. Taken together, these results demonstrate unprecedented transcriptional interaction between TEs and HLAU genes in a tissue-specific manner.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

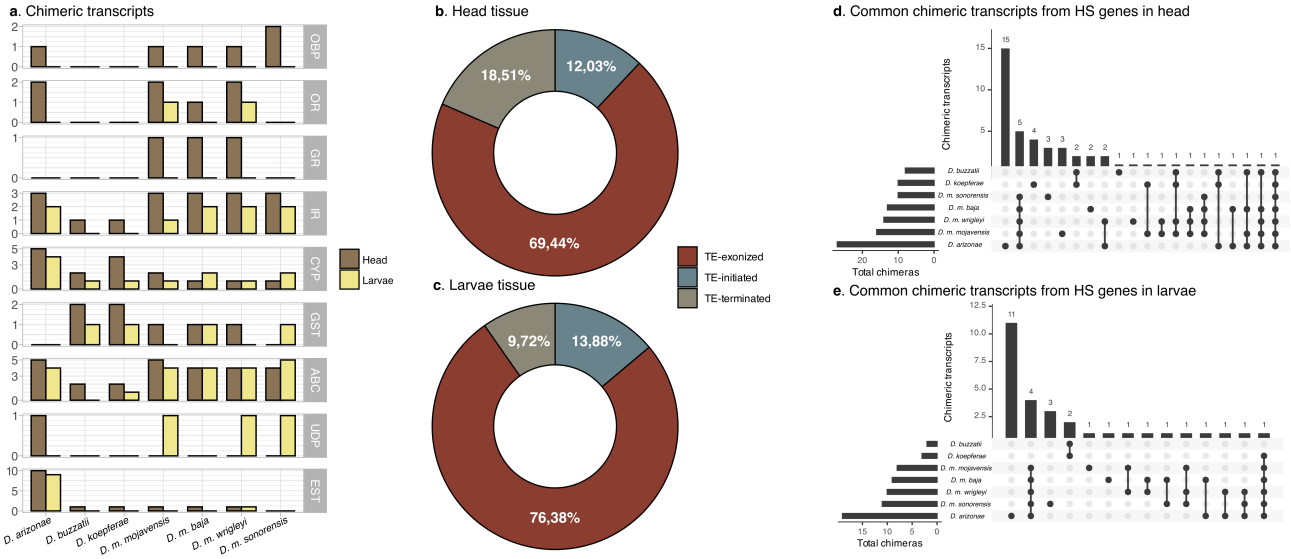

Figure 5: **A**) Number of HLAU genes generating chimeric transcripts in each genome. **B**-**C**) Categories of chimeric transcripts in head (B) and larvae (C) tissues. **D-E**) Common chimeric transcripts between the seven genomes in head (D) and larvae (E) tissues.

The quantitative analysis on chimeric transcripts revealed a clear transcriptional polymorphism on HLAU genes due to TEs. Then, we investigated chimeras present in only one species, possibly associated with recent insertions, as well as common chimeras present in more than one species, highlighting their retention over evolutionary time. We observed that 58.33% and 53.33% of the chimeric transcripts derived from HLAU genes are species-specific in head and larvae tissues, respectively (Fig. 5D-E). In head, from the total of 48 HLAU genes producing chimeric transcripts, *D. arizonae* and *D. koepferae* sum up 19 expressed uniquely in these species. In larvae, *D. arizonae* also has the highest number of uniquely expressed chimeras from HLAU genes, with 11 genes out of 30. Following *D. arizonae*, *D. m. sonorensis* has 3 unique chimeras, whereas *D. koepferae* does not have any uniquely HLAU gene-derived chimera (Fig. 5E). Regarding common chimeric transcripts, we observed in the head tissue only one HLAU gene producing chimera in the seven genomes (Fig. 5D). The gene *Abc10* produces a TE-exonized transcript in species from the *mulleri* complex, and a TE-initiated transcript in the species from *buzzatii*. All chimeras are derived from a *TIR* element, from the *P* family, which is embedded in 5′ UTR in the TE-exonized cases and located

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

upstream for the TE-initiated transcripts. Although the insertion generates the same chimera in the seven genomes, the element is not conserved between species. In *D. m. sonorensis*, it is a 3.9kb insertion overlapping the 5' UTR of the *Abc10*, whereas in the other subspecies the TE is fragmented in 3 small insertions, all embedded into the 5' UTR. We also observed five HLAU genes producing chimeras only in species from the *mulleri* complex: *AbcA3 – LINE/R1; AbcF2 – LINE/R1; AbcD – Helitron and DNA/TcMar-Tc1; Ir-k2 – Helitron; and Obp28a – LTR/Pao*. In the head tissue we also found the *Cyp* gene (*LOC6576681*) producing chimeric transcripts in *D. buzzatii*, *D. koepferae* and *D. arizonae*. Although the same gene, each species produce a chimera from a different element. In *D. buzzatii* the gene produces a TE-exonized transcript in which a *hAT-Blackjack* is located at the intron, in *D. koepferae* the *Cyp* produces a TE-terminated transcript with a *Helitron*, and finally in *D. arizonae* another *Helitron* insertion is embedded into the exon producing another TE-exonized transcript.

Discussion

Insect HS and its respective role in speciation has been subject of research to understand the evolution of reproductive isolating barriers and adaptation to local environments. Indeed, many insects have finely tuned adaptations to locate, feed, reproduce, and develop on the host tissue. Adaptations to new hosts can lead to ecological speciation, wherein speciation arises as a consequence of evolving reproductive barriers within divergent natural populations inhabiting distinct environments. In this context, we investigated whether TEs could play a role contributing to genetic novelty of HLAU genes facilitating rapid adaptive radiation to new hosts. We check if TE were providing new regulatory motifs or protein domains when inserted within gene regions as well as their impact in the transcriptome. To test this hypothesis, we selected seven *Drosophila* species/subspecies that belong to the *repleta* group and have different primary cacti hosts.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Robustness of genome assemblies allow high-resolution analyses

The repetitive nature of TEs makes their identification challengeable with short reads, leading to underestimations of genomic TE content, and preventing high-resolution genome-wide analysis concerning TEs. In recent years, high-throughput sequencing technology has provided advances in genomics studies, such as long reads Nanopore sequencing method, with high-quality genome assemblies and allowing deeper genomics and transcriptomics analysis (56–60). In our study, we performed Nanopore sequencing for seven cactophilic *Drosophila* species/subspecies. Both genomes of *D. buzzatii* and *D. koepferae* had 175 Mb of estimated genome sizes, which is similar with 160 Mb and 169 Mb obtained on previously assemblies (46,61), performed through the combination of short and long reads technologies. Furthermore, we obtained 11,347 and 11,440 fully annotated genes in *D. buzzatii* and *D. koepferae*, which is very similar with the previous 10,983 and 11,804 obtained for the same species. The genome assembly of *D. arizonae* had 182 Mb of estimated genome size, which is different from the 142 Mb previously assembled for the same species (43). The difference might be due to the sequencing and assembly method used previously, that has been relied only on Illumina short reads. Furthermore, we obtained 14,441 orthologous genes with *D. mojavensis*, whereas the previous genome assembly had 12,129 genes (43). Regarding the assemblies of *D. mojavensis* subspecies, we obtained very similar genome size between them, from 177 Mb to 183 Mb, demonstrating the consistency between the depth of our assemblies. The reference genome of *D. m. wrigleyi* has 163 Mb of estimated genome size (62). We attribute the difference in genome size estimations to different sequencing and assembly approach, since the reference genome was assembled by merging Nanopore and Illumina reads. Nevertheless, our gene annotation covered in average 14,840 genes, representing ~98% of all reference genes from the reference genome. Taken together, these results highlight the robustness of our assemblies, showing that all downstream analyses performed in this work were assessed with high-quality genomes.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Paucity of HLAU genes under selection

The necrotic tissue of each cacti-host used by cactophilic *Drosophila* species has different composition in terms of yeasts and alkaloids (63,64). Therefore, species have adapted to feed and breed in different hosts, which has been demonstrated by behavioral and physiological response when species feed and breed in non-preferential hosts. The involvement of positive selection in the HS process is an evolutionary force that favors mutations linked to amino acid alterations that allow the recognition of a broader spectrum of odors, metabolization of nutrients, and detoxification pathways (65). Indeed, specialist species are likely to lose part of their genes related to sensory pathways, whereas those retained are subject to adaptive selection (66,67). A preliminary genome-wide analysis of signatures of positive selection in the four *D. mojavensis* subspecies revealed high ω rates on genes associated with chemosensory, perception, immunity, behavior, detoxification, and reproduction (68). Although a previous genome-wide analysis has identified 1,294 genes under positive selection (45), here we narrowed down the analysis to investigate specifically HLAU gene families. We observed six genes under positive selection: *Obp56d*, *Or1a*, *Ir25a*, *Cyp310a1*, *Cyp4s3* and *Udp-ugt5*. Genes associated with chemosensory function with high positive ω values have been associated previously as an important step to host specialization in *Drosophila* species (67–69). In addition, positive selection on detoxification genes have also been shown previously in *D. buzzatii* and *D. aldrichi* (45,70). Our results demonstrate that HLAU genes under positive selection is not a common phenomenon when considering the host preference of the species. We argue that despite we found less genes, these might contribute to the convergent evolution for the shift to columnar cacti, which happened independently multiple times in species of the *repleta* group.

We performed a strict branch-site analysis, in which we selected *Opuntia* dwellers as our phylogenetic background. Therefore, the unprecedent method used in this work focused on the common differences of species adapted to columnar cacti, rather than independent evolutionary events. We propose that the six genes observed under selection in this work

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

might have high relevance to the HS and adaptation that allowed the transition from *Opuntia* to columnar cacti. The paucity of genes (4.68% of the total) is also associated with technical limitations, since we based our analysis on 1:1 orthologs, and a certain proportion of the genes evolve too fast to be correctly assigned during orthologs identification (71). Additionally, the ω values are higher than 1.5, for instance as *Gr59d* with ω = 3.49, which despite not significant still indicates ongoing divergence between species/subspecies. Moreover, it has been proposed that gene families enriched by TEs evolve under positive selection, regardless of the eukaryotic taxonomic group (72). Interestingly, these gene families are involved with sensory perception and environmental interaction, among others. However, our molecular analysis does not support such hypothesis since the genes under positive selection are not enriched by TEs.

TE landscapes in cactophilic Drosophila species

Rapid local adaptation and the colonization of a new habitats has been associated with TEs activity (73,74). The hypothesis proposes that populations invading new habitats might have undergone genetic bottleneck and founder effects, due to the selective pressure under stressful environments. Since stressful conditions can increase TEs expression/activity (75–78), although highly TE family-dependent, the exploration of new niches could facilitate transposition bursts and rapid adaptation by modifications in the genome structure and, increasing plasticity due to TEs mobilization. Therefore, the occurrence of transposition bursts among diverging populations is an important biological phenomenon that may cause genetic variability between populations, facilitating the adaptive process and perhaps speciation. For instance, transposition bursts followed by periods of species radiation have occurred in insects, and primates (79–81).

In this work, we built genome-specific TE libraries based on sequence homology and structure-based approaches, combining a set of automatic analysis and manual curation. This approach provides *de novo* TE consensus for each genome, avoiding the bias of using TE

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

libraries from databases that might have under and overestimations depending upon the species. The life-cycle of TEs in a given species can last for a long time and it goes through three stages: replication dynamic, silencing, and degradation (82). Thus, the global evolutionary history of TEs in a genome can be assessed with landscape analysis, in which the divergence of TE insertions with its respective consensus provides the insertion's age. In species from the *buzzatii* complex, represented in this work by *D. buzzatii* and *D. koepferae* we found 14.54% and 15.39% of TE content, respectively. The previous annotated mobilome of *D. buzzatii* was described with up to 11.16% of TE content (83). We assign the difference with our result to the genome assembly quality, since the previous genome had an estimated genome size of 153 Mb, divided into 10,529 scaffolds, whereas ours has 175 Mb and 719 scaffolds. In addition, in our TE annotation, we observed that Helitrons are the most abundant TE class, followed by LTR, DNA, and LINEs. The previous *D. buzzatii* mobilome had also *Helitrons* with highest TE content, but followed by LINEs, LTR, and DNA (83). The difference into the TE classes' abundances might exist because of the difference in the methods to build the TE library. Our method consists in the identification of consensus in a genomic-specific manner, whereas the previous *D. buzzatii* mobilome was assessed with merged consensus from Flybase and Repbase, as well as built-in consensus with *RepeatModeler2*. For *D. koepferae*, we reported the first mobilome, representing 13.19% of TE content. The TE landscapes from *D. buzzatii* and *D. koepferae* revealed that TE proliferation has occurred differentially between them. For instance, although Helitrons are the most abundant TE class in both species, comprising 6.1% and 3.58% in *D. buzzatii* and *D. koepferae*, they have substantially less full-length TE insertions in *D. buzzatii*, than in *D. koepferae*. Studies with these species have shown incomplete post-zygotic isolation, potentially caused by TEs' activity, since interspecific F1 hybrids have transposition bursts of specific TE families in the hybrid progeny, for which the male is sterile (84,85). The phenomenon is attributed to divergence between TE families among both parentals, affecting the secondary piRNA pathway (ping-pong loop), resulting into inefficient post-transcriptional control for certain elements (86). Our results are in agreement with these previous observations, since the two

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

mobilomes revealed that TEs from *D. buzzatii* and *D. koepferae* have different TEs abundances and proliferation waves.

In the species of the *mulleri* complex, TE content ranged from 15.07% in *D. m. baja* to 16.12% in *D. arizonae*. This is the first mobilome description of the *D. mojavensis* subspecies complex and *D. arizonae*, except for *D. m. wrigleyi* that was already published. Previous mobilomes for *D. m. wrigleyi* have reported 8.9% and 15.35% of TE content (83,87). The TE content in the four subspecies was on average 15.57%, close to what has been described for *D. m. wrigleyi*. However, the abundance of TE classes that we obtained for *D. m. wrigleyi* followed LTR > Helitron > DNA > LINE, whereas the previous mobilome it was Helitron > LTR > LINE > DNA (83). Despite the difference, Helitron and LTR had 5.13% and and 5.06% of TE content, respectively, whereas we reported 4.68% and 5.09%. Therefore, these minor differences between the two classes might be associated with different assembly strategies.

In the four *D. mojavensis* subspecies, we found a similar mobilome for *D. m. sonorensis*, *D. m. baja*, and *D. m. mojavensis*. LINEs and LTRs did not have substantial differences between them, but on average 6.55% of the TE content is derived from *Helitrons* and 1.17% from DNA, whereas in *D. m. wrigleyi* the same classes represent 4.68% and 3.67%, demonstrating that the latter subspecies might have had a silencing period of *Helitrons*, whereas DNA elements are likely to have undergone transposition burst. Furthermore, considering the recent divergence time between *D. mojavensis* subspecies, the TE distribution among them might suggests a rapid evolutionary dynamic of TEs that happened either in mainland subspecies or in Santa Catalina Island. The same evolutionary trend of silencing and proliferation of DNA and *Helitrons* was observed in the *D. arizonae* mobilome, which has 5.47% and 2.88% of TE content from these families. Following the TE landscape distribution as an evolutionary marker, *D. m. wrigleyi* would be the subspecies with the closest phylogenetic relationship with *D. arizonae* than the others.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

OR genes are enriched with TE insertions at 2 kb upstream regulatory region

The literature has recurrently demonstrated that TEs can create new regulatory elements, including gene promoters (88–90), enhancers (91,92), and insulators (93,94). Indeed, TE-derived sequences contribute up to 40% of genome-wide binding sites for transcription factors located in TE-rich regions (95). It has been proposed that such co-option of TE-regulatory elements for gene regulatory networks can provide substantial modification of gene expression over short time scales, contributing to generate rapid genetic variability on the transcriptome level (96–98). Because the invasion of new habitats is accompanied by modifications in behavior, which has been associated with alterations in sensory cues and environment perception (99,100), our results demonstrate the potential of TEs to change gene regulation of specific HLAU gene families on several species/subspecies, and thus contributing to the HS from *Opuntia sp*. to columnar cacti.

We observed species-specific enrichment of TEs in the promoter region for CYP, GR and IR genes, but OR gene family had the highest enrichment in six out of seven genomes from our cactophilic *Drosophila* species. In ants, ORs have also been found enriched in TEs, but in this case they were associated with tandem duplication by crossing over, allowing massive duplication events (101). In our case, since there is no expansion of ORs compared to other *Drosophila* species, the enrichment in TEs is likely to have another function, probably related with gene regulation due to its specific location in promoters. Indeed, we observed that *Helitrons* represent 50% of all cases. These elements have been extensively reported on promoter regions of other eukaryotic species (102–105), in which a few cases of embedded TFBSs were reported (106,107). Here, we demonstrated that part of the enrichment in TEs observed at the promoter region of ORs might be due to the exaptation of TFBSs mostly derived from *Helitrons*. In addition, the recurrence of the enrichment in species of the *mulleri* and *buzzatii* complexes might demonstrate potential old cases of TE co-option in the *repleta* phylogeny. Our genomics analysis provides a robust list of potential candidate genes in which *Helitrons* might play a regulatory role, an interesting hypothesis that can be

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

determined through cutting-edge technologies such CRISPR methods to delete insertions in site-specific manner (108).

Head transcriptome reveals subspecies specific patterns

The differential expression of HLAU genes in the head tissue revealed that location and host use (detoxification) gene families have the highest divergence across the subspecies, since we observed more differentially expressed genes for these families. Furthermore, in *D. m. wrigleyi*, which is the only subspecies with *Opuntia sp*. as preferential host instead of columnar cacti, we observed that CYP genes had on average 12 up-regulated genes compared to other species. Taken together, these results highlight that the divergence of the host preference between *D. mojavensis* subspecies might be explained by modifications in gene expression of HLAU genes in the head tissue. In addition, despite lower toxicity in *Opuntia sp*. compared to columnar cacti, we found a similar number of up-regulated host use genes between *D. m. wrigleyi* and the other subspecies, demonstrating that *D. m. wrigleyi* exhibits a constitutive transcriptional response for detoxification that is not driven by the host.

In the head transcriptome, the four *D. mojavensis* subspecies were clustered independently in our PCA (Fig. 4), demonstrating that gene regulation in head tissue might play a role in the life-history traits divergence that exist between them. The extension of differentially expressed genes corresponds to 6.17% from all genes analyzed. These results demonstrate that the regulation of host location genes might be an important factor that evolved and allowed host shift. Previous studies with *D. melanogaster* have shown that the divergence in the amino acid sequences due to single polymorphism mutations in ORs and OBPs have been suggested as factor of the modulation in odor sensitivity in *Drosophila* (100,109), which could drive to divergence of host location. In addition, many ORs are involved in a large repertoire of ecological relationships, influencing not only behavior, but also oviposition and feeding (110). Another study has identified 18 ORs differentially

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

expressed between heads of *D. m. wrigleyi* and *D. m. mojavensis* subspecies, demonstrating that modifications in the sensory transcriptome have evolved between these species (111), reinforcing our observations for the same tissue. Here, we extended the comprehension of such divergence with the pairwise comparison between *Opuntia sp.* and columnar cacti dwellers, and pointed out that each species has a large set of differentially expressed HLAU genes.

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

In the larvae tissue, we found a common gene expression pattern between columnar cacti dwellers, which in turn diverge from species that use *Opuntia sp.* Location and acceptance genes had mild alterations, whereas gene families associated with host use (detoxification) had on average more up-regulated genes than the others, especially due to overexpression of CYP genes. As we have observed for the head tissue, there is also an increase of CYP genes expression in *D. m. wrigleyi* larvae compared to the other subspecies. Indeed, in *D. m. wrigleyi*, we found the highest number of CYP up-regulation, with 23 CYP genes overexpressed in comparison to *D. m. sonorensis*. Among the other subspecies that preferentially use different species of columnar cacti as hosts, we also found detoxification genes with the highest numbers of differentially expressed genes. Similarly with our observations in the head tissue, these results indicate that larvae tissue exhibits constitutive gene expression that is subspecies-specific, demonstrating part of the genetic background that allows these subspecies to be specialized in specific cacti species.

The variance in gene expression in larvae revealed that species of the *buzzatii* complex have higher divergence compared to *mulleri* ones, and lower between *D. buzzatii* and *D. koepferae*, following the phylogenetic relationship between the species. Although we have species with preferential use for *Opuntia* and for columnar cacti in both subgroups, *D. koepferae* and *D. buzzatii* are adapted to different hosts for a longer time than those of the *mulleri* complex studied here. In addition, a preliminary study demonstrated that *D. buzzatii*

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

have a high divergence on larvae gene expression compared to *D. koepferae* (40), reinforcing the variance that we have observed. In *mulleri* complex, although not high, it is noteworthy that the subclusters of gene expression from *D. m. mojavensis*, *D. m. baja*, *D. m. sonorensis*, and the subcluster from *D. arizonae* and *D. m. wrigleyi*. Indeed, differences in larvae behavior and survival has been observed when *Opuntia*-related species are reared on columnar cacti (112,113). Our results show that divergence in the regulation of gene expression in larvae is one of the factors shaping the adaptation to new hosts.

The up-regulated HLAU genes were mostly represented by host use gene families, which are associated with detoxification. A previous study demonstrated that when *D. m. mojavensis*, which has *S. thurbery* as primary host, is reared on media with *S. gummosus*, the larval viability had a reduction of 60% (114). The authors also proposed that 21% of all genes would be differentially expressed between larvae reared in the primary and alternative cacti-host. Many genes were reported with detoxification, energy production, among others. Another similar study with larvae transcriptome, but with *D. buzzatii* and *D. koepferae*, proposed that *D. buzzatii* flies reared in columnar cacti media have a stronger detoxification response in comparison with *D. koepferae* flies, suggesting that *D. koepferae* has a canalized transcriptome response to the toxic alkaloids from columnar cacti (40). We expected to find a higher number of detoxification genes in the subspecies with preference for columnar cacti compared to *D. m. wrigleyi* and *D. buzzatii* that use primarily *Opuntia sp*. However, we observed a similar number of differentially expressed detoxification genes regardless of the host preference. Since CYP genes also have other functions than detoxification, and they were overrepresented in our results, we suggest that their differential expression might also be associated with developmental pathways in larvae. (115,116). Therefore, further analysis in the biological role of up-regulated CYP genes in *D. m. wrigleyi* and *D. buzzatii* might be assessed. Alternatively, despite *Opuntia sp.* has lower toxic compounds compared to columnar cacti, it also has a basal level of toxicity. Thus, any cactophilic species is expected to have evolved a certain level of detoxification ability, which explains our observations. In

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

addition, the high frequency of common HLAU genes associated with detoxification on columnar cacti dweller species is expected and already demonstrated in previous works. Here, for instance, we observed that the three *D. mojavensis* subspecies have the same seven genes with higher expression compared to *D. m. wrigleyi*. Most of them are associated with detoxification except *Ir21a*: two CYP genes, three UDPs and one Esterase.

Transposable elements are a source of transcriptome variability between cactophilic species

TEs contribute to genome and transcriptome evolution by rewiring regulatory networks, and by incorporating protein domains into gene transcripts. In both cases, TE copies that are inserted within genes can integrate their sequences into the mRNA, generating chimeric transcripts. A recent study with different ecotypes of *D. melanogaster* has shown that these chimeras have the potential to generate transcriptome novelty (34). These chimeras have also been demonstrated to be active in a tissue-specific manner (117). Although the production of chimeric transcripts derived from polymorphic TE insertions is well-known, there is no study addressing the question on how the transcriptomic variability derived from TEs could impact the HS between species/subspecies. Here, we took advantage of our unprecedented genomics and transcriptomics data to analyze the extent of genetic variability derived from chimeric transcripts in cactophilic species, as well as HLAU genes produce chimeric transcripts.

A preliminary transcriptome-wide study with *D. melanogaster* revealed that ~1.12% of the genes produce chimeric transcripts (34). Here we observed ~2.1% on average between head and larvae tissues. The higher number of genes in *repleta* species might reflect the difference between their mobilome composition compared to *D. melanogaster*, which is depleted of active *Helitrons* for example (118). Despite this difference, the proportion of TE-exonized, TE-terminated, and TE-initiated transcripts follows similar values between them, in which TE-exonized transcripts comprise ~70% of all chimeras (34,117). In our analysis, TE-initiated transcripts represent chimeric transcripts in which the TSS of the mRNA is located

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

within the TE insertion (34), for which *Helitrons* had the highest frequency in four of the seven species. *Helitrons* have been extensively reported as a source of TFBSs when located at the promoter region of genes. In the bat *Myotis lucifugus*, 1.4% of the transcripts from the salivary gland have *Helitron*-derived sequences, for which ~6% correspond to promoters (103). In the plant *Brassica napus*, an *Helitron* insertion disrupts the ancestral promoter of the gene *BnSP11-1*, altering the mating system of the populations carrying this insertion, which in turn contributes to the adaptation and speciation (119). In *Arabidopsis thaliana*, *Helitrons* have been responsible for the genome-wide distribution of the TFBS for the PHE1 TF, which is associated with crucial endosperm regulation (106). In *Caenorhabditis* species, TFBSs for the TF heat shock factor 1 have been widely spread by *Helitrons*, rewiring and diversifying heat shock response between species and strains (120). Combining our results from TFBSs from OR genes with our transcriptome-wide analysis of TE-initiated transcripts, the results reinforce that *Helitrons* have a *cis*-regulatory role in cactophilic *Drosophila* species. Regarding TE-terminated transcripts, the high prevalence of *Helitrons* also highlights their role on the termination of gene transcription. This phenomenon has been already described in bats, in which the genes *Otud3* and *Tor1b* have novel PASs from *Helitrons* (54). Another study with *A. thaliana* revealed by poly(A)-tag sequencing data that ~48% of all TE-derived PASs are from *Helitrons* (121). Although there is vast documentation of TEs acting as alternative PASs (122), here we provide the first transcriptome-wide report of *Helitrons* as a major source of putative PASs in *Drosophila*.

Chimeric transcripts reveals recent and old insertions interacting with HLAU genes

Specialist species are expected to have low genetic variability and increased population structure, since the effective population size is limited by the availability of the hosts used as feeding and breeding sites. On the other hand, generalist species have more genetic variation and less structured populations, due to the large range of resources (123). A comparative analysis between generalist and specialist *Drosophila* species revealed that niche amplitude does not play a role in TE dynamics in the genomes (124). Nevertheless,

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

several studies have proposed that the successful adaptation to new environments in invasive species might be associated with TEs activity, despite reduced genetic diversity caused by founder effects (125,126). Through TE-initiated and TE-terminated transcripts, we provide evidence for TE insertions playing a regulatory role on HLAU genes, whereas TE-exonized transcripts suggest a putative source for TE-driven coding sequence evolution. Here, we demonstrate that TEs interact with genes associated with HS, providing a wide range of transcriptomic variability in seven cactophilic species with different hosts preferences. We observed that species *D. arizonae* and *D. koepferae,* which does not have preferential usage of neither *Opuntia sp.* nor columnar cacti, are the ones with the highest production of unique chimeras. This result demonstrates a possible relationship of broader niche due to TE-driven adaptation, at least from HLAU genes expressed on head.

Regarding the extension of transcriptomic variability from chimeric transcripts, the CYP gene (LOC6576681) is a remarkable case where the same gene produces different chimeras from different TE families, depending upon the species. In addition, our results demonstrate that the polymorphism from chimeric transcripts has two sources: 1) polymorphic insertions, where TEs are gained or lost in a genome or group of species; and 2) where the same TE have been maintained over time and is shared between species, but its integration into the mRNA only happens in a few species. The *Abc10* gene is a clear example where a TIR element from the *P* family has been maintained over 9.5 My at the promoter region of species from the *repleta* group, and its function as alternative TSS is tissue-dependent, since the chimera is not observed in larvae of *D. buzzatii*. The combined genomics and transcriptomics approaches used in this study provide new insights regarding the role of TEs in the evolution of HS, and ultimately their contribution to ecological speciation.

The detection of chimeric transcripts is a reliable way to infer transcriptional interactions between genes and TEs. Nevertheless, even applying strict parameters as we did, the contribution of the chimeric isoform to the total gene transcript abundance is unknown.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Thus, perhaps the integration of TE-derived sequences into the gene transcript is not beneficial, but still maintained in the transcriptome due to its low expression. Furthermore, chimeric transcripts might be degraded by surveillance pathways that control for aberrant mRNA production, such as no-go decay (127), non-stop decay (128), and non-sense-mediated RNA (129). Our study provides a set of genes producing chimeric transcripts, in which part of them might have phenotypic impact and ultimately association with host preference. The application of techniques of RNAi aiming chimeric transcripts, or deletion of TE insertions producing chimeras with CRISPR (108) will be able to confirm the predictions of our results.

Conclusions

TEs have been demonstrated to contribute to genome evolution of cactophilic *Drosophila* species. Their enrichment on promoter regions of HLAU gene families, as well as the putative presence of TFBSs on their sequences highlights their potential role into the evolution of the adjacent genes. In addition, their frequent exonization reveals a mechanism by which TEs increase diversity between populations, with potential to drive host shift and hence contribute to ecological speciation. Our combined genomics and transcriptomics approaches provide new insights regarding the divergence between species/subspecies that use different cacti-hosts and have incipient divergence time.

Material and methods

Fly stocks and genome sequencing

Most of the strains were obtained from the UC San Diego *Drosophila* Stock Center: *[D. m. mojavensis*: Anza (01), Anza Borrego Desert, California, USA, Stock Center n°: 15081-1352.01; *D. m. wrigleyi*: CI (22), Catalina Island, California, USA, Stock Center n°: 15081-1352.22; *D. m. baja*: BC (20), Cape Region, Santiago, Baja California Sur, México, Stock Center n°: 15081-1352.20 e *D. m. sonorensis*: AG (26), Agiabampo Bay, Sonora, México, Stock Center n°: 15081-1352.26] and *D. arizonae* [HI (17), collected in Metztitlan, Hidalgo, Mexico, Stock

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Center n°: 15081-1271.17]. *D. buzzatii* (Bu28) and *D. koepferae* (Ko2) stocks correspond to two laboratory lines used in previous works (4,130). Both stocks were maintained by brother-sister mating for more than a decade and then kept by mass culturing.

The same following protocols of DNA extraction and genome sequencing were applied to all samples. DNA was extracted from 10 males and 10 females from each species using the Qiagen DNeasy Blood&Tissue kit. The genomic DNA amount and quality were evaluated with NanoDropTM One UV-Vis spectrophotometer (Thermo Fisher Scientific, Waltham, MA, USA) and Qubit ® 1.0 Fluorometer (Invitrogen, Carlsbad, CA, USA). Three micrograms of DNA were repaired using the NEBNext FFPE DNA Repair Mix (NEB M6630). End repair and dA-tailing were performed using the NEBNext End repair/dA-tailing Module (E7546, NEB). Ligation was then assessed with the Ligation Sequencing Kit 1D. MinION sequencing was performed according to the manufacturer's guidelines using R9.4.1 flow cells (FLO-MIN106, ONT) and a Nanopore MinIon Mk1b sequencer (ONT) controlled by *ONT MinKNOW v.18.3.1*. Base calling was performed after sequencing using *Guppy v.4.0.5* in high accuracy mode for isogenic wild-type strains *v.3.1.5*.

Genome assemblies and gene annotation

The quality control for Nanopore reads was performed with *Nanoplot* v.1.10.2 (https://github.com/wdecoster/NanoPlot). Reads with quality lower than 7 were removed from downstream analysis. In order to assemble contigs, *Flye v.2.8* (131) was used with default parameters, except –plasmids. All contigs were aligned with *minimap2* v2.16 (132), with –x map-ont option. The alignment was used to perform the polishing of contigs with four rounds of *RACON* v1.3.2 (133), default parameters. The assembly quality metrics was assessed with *Assembly-Stats* v1.0.1 (https://github.com/sanger-pathogens/assembly-stats). Assembly incongruences were manually visualized with *D-genies* v1.2.0 (134) and corrected with *samtools* v1.9.0. (135), function *faidx*; and *Gepard* v1.4.0(136) for determination of breaking points. The super scaffolding of all corrected assemblies was performed with

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

RaGOO v.1.1 (137), with –s and –t 4 parameters, using the respective reference genome assembly for each species. Subsequently, the chromosome-scale assemblies were submitted to a Benchmarking Universal Single-Copy Orhologs (BUSCO) v.5.4.3 analysis (138) with lineage *diptera_odb10* (*-l* parameter). The gene annotation for our Nanopore genomes was performed with the software *Liftoff v.1.6.3* (139), with the reference genome of *D. mojavensis* (GCA_018153725.1) (62). We have used the parameters *–exclude_partial*; *-a 0.5* and *–s 0.75* to keep annotated genes with at least 75% of the reference gene length. To recover HLAU genes sequenced and annotated in our study, we have used the annotation of the reference genome of *D. mojavensis*, removing genes with descriptions as "low quality protein" and "unestablished gene function".

Molecular evolution analysis

To identify HLAU genes with signatures of positive selection in species that have preference for columnar cacti compared to species with preference for *Opuntia sp.*, we selected the four *D. mojavensis* subspecies, *D. navojoa*, *D. koepferae* and *D. buzzatii*. *D. arizonae* species was removed from this analysis due to its generalist preference for both columnar and *Opuntia* cacti. The *D. navojoa* genome was retrieved from NCBI (GCF_001654015.2), which was produced with Illumina reads (140). For the other Nanopore genomes sequenced in this work, we performed a polishing with Illumina paired-end reads. We used the software *NextPolish* (141) to fix base errors in the assembled genomes. The four *D. mojavensis* subspecies were polished with genomic Illumina reads sequenced previously (142), whereas for *D. koepferae* and *D. buzzatii* we used publicly available genomic Illumina reads (44) and RNA-seq reads (45), respectively. In the latter, to increase the read coverage and optimize the polishing to the higher number of genes and exons as possible, we merged RNA-seq reads from adult male and female (45). All genomes were corrected using two rounds of polishing, as recommended by the authors (141). Then, we extracted the longest CDS sequences for all genes combining AGAT (142) *agat_extract_sequences.pl* parameters *-t mrna*, *samtools faidx* (144), and *seqtk* (https://github.com/lh3/seqtk). Posteriorly, we

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

intersected the occurrence of orthologous genes between the seven genomes, recovering a list of genes that were present in all of them. The sequences were aligned using *MAFFT v.7.487* [67] and the non-aligned codons were removed using *pal2nal* (145). To obtain a strongly supported tree for this analysis, we inferred the species phylogeny tree with 2,043 shared BUSCO genes. The phylogeny was estimated using the concatenated alignment (3,730,914 sites) with *IQ-TREE v. 2.2.2.6* (146), using the best model for each gene, 10,000 replicates of ultrafast bootstrap and 1000 replicates for Approximate Likelihood-Ratio Test (ALRT) (parameters *-m MFP -bb 10000 -alrt 1000*). Finally, we carried out analyses to detect signatures of selection analysis for each gene individually with *CODEML* from *PAML v.4.10.7* (147), using *branch-site* approach. In order to find signatures of positive selection in species that use columnar cacti , we marked the respective branches of *D. koepferae*, *D. m. baja*, *D. m. mojavensis* and *D. m. sonorensis*, with the same label (#1), to compare them with the background species that use Opuntia *sp.* (*D. buzzatii*, *D. m. wrigleyi*). For this, we tested two hypotheses: (i) H0 – Null hypothesis, in which all branches evolved under negative selection (model = 2, Nssites = 2, fix_omega = 1, ômega = 1); (ii) H1 – Alternative hypothesis, in which the labeled branches (foreground) have sites with signatures of positive selection (model = 2, NSsites = 2, fix_omega = 0, omega = 1.5). The hypotheses were tested through chi-squared test, calculated by the log-likelihood (lnL) ratio between H1 and H0 (2ΔlnL) and degrees of freedom obtained from the difference between the number of estimated parameters for H1 and H0 (Δnp). Sites under positive selection were identified by Bayes Empirical Bayes and posterior probability, implemented in *CODEML*.

RNA extraction and sequencing

We performed RNA-seq for larvae and head tissues from *D. buzzatii*, *D. koepferae*, *D. arizonae* and the four *D. mojavensis* subspecies. The RNA extraction was assessed with RNeasy Plus Kit for 10 female heads (10 days old) and 15 larvae with three technical replicates. The library preparation was done with stranded mRNA-seq/standard quantity (ligation), and 100 bp paired-end reads were produced in an Illumina Hiseq 4000. Illumina

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

adaptors and low quality reads were removed for downstream analysis with *Trimommatic v.0.39* (148), parameters*: LEADING:3*; *TRAILING:3*; *SLIDINGWINDOW:4:15*; *HEADCROP:10*; *CROP:90*.

TE annotation

For each sequenced genome, we developed a novel pipeline combining automatic and manual approaches to construct TE consensuses. First, we used *Earl Grey v.2.2* (149) to build de novo TE libraries with *RepeatModeler2 v.2.0.4* (150), which uses RECON *v.1.08* (151) and *RepeatScout v.1.0.6* (152). Then, in order to retrieve full-length TE consensus, the pipeline performs an extension of each predicted consensus sequence using a method named "Blast, Extract, Extend" (BEE) (153). Such method extracts the sequences of the 20 full-length (or longest) TE insertions for each consensus, and then it retrieves the 1 kb flanking regions from them. Posteriorly, insertions altogether with flanking regions are aligned with *MAFFT v.7.487* (154), and the alignment is trimmed with *trimAl v.1.4* (155), parameters *–gt 0.6* and *–cons 60*. Subsequently, the alignment is analyzed to check whether the flanking regions from the full-length insertions have high quality, meaning that they are part of the TE rather than genomic regions. Finally, the consensus sequence for each TE family is updated with *EMBOSS v.6.6.0.0* (156). The BEE method is repeated up to five times for each consensus. Once the extended TE libraries were obtained for the seven genomes, they were submitted to a CDS filtering step, in which we removed all consensus with high similarity with *D. mojavensis wrigleyi* CDSs using blastn, parameter *-qcov_hsp_perc 80*. A last filter step was applied to remove TE consensus with less than 3 *RepeatMasker* hits across the genome, with at least 80% of its length. A similar cutoff has been applied to *D. suzukii* TE annotation (125).

TE consensus that had passed in our filtering thresholds were then submitted to the classification analysis. We developed a pipeline named TEclassifier, which performs similarity analysis of non-reference consensus with a database provided by the user (https://github.com/OliveiraDS-hub/TEclassifier). The analysis identifies similarity analysis

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

between the non-ref. consensus and a database provided. All consensuses are masked with *RepeatMasker v.4.1.4* (157), parameters *-lib*, *-cutoff 250*, *-s*, and *-norna*. Then, the pipeline performs different steps depending on the number and variety of hits that the non-ref. consensus has: (1) only one hit is considered unique and the total length of the masked hit is compared with the total length of the non-ref insertion; (2) multiple hits, but all from the same class and family: the length of the hits is summed up, and the total length is compared with the length from the non-ref consensus; (3) multiple hit, but different classes and/or families: in this case each class is treated separately. Due to the possibility of overlapped hits from TE families of the same class in the non-ref consensus, the pipeline removes the extent of overlapped regions from two insertions, hence decreasing the length of the insertions. Once all hits from a given class have no overlap, the pipeline sum up the total extent of the hits and compare it with the non-ref consensus. The process is repeated for all classes masked. Finally, the pipeline classifies only TEs with a corresponding match in >= 50% over their total length. In this work, we have used Dfam v.3.7 (158) as reference database. Due to the "curated" library contains only *D. melanogaster* consensuses, we used the "uncurated" library from Dfam. We extracted *Drosophila* TEs with *famdb.py v.0.4.3*, parameters *-i Dfam.h5 families --include-class-in-name -f fasta_name -ad 'Diptera'*. Then, we filtered out non-Drosophila sequences, and "Unknown" TEs with the unix commandline: *grep -i 'drosophila' file.fa | grep -v -i 'unknown'*, obtaining 49,916 out of 344,562 consensuses. Finally, all obtained IDs were used with *seqtk v.1.3* (https://github.com/lh3/seqtk) to subset the complete fasta file generated by *famdb*.

The polished TE libraries constructed to the seven cactophilic *Drosophila* genomes were then used to perform TE annotation with *RepeatMasker v.4.1.4* (157), parameters *-cutoff 250*, *-s*, and *-norna*. Then, we have used *RepeatCraft* (159) with *LTR Finder v.1.2* results to merge LTRs and their respective internal regions. Subsequently, we removed all TE insertions shorter than 80 nucleotides, following the classification rule 80-80-80 (160). Finally, a last filtering step in the TE annotation was performed to remove insertions comprised by simple

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

sequence repeat (SSR). We used *TRF v.4.09* (161) in all TE insertions from each genome, with parameters *2*, *5*, *6*, *75*, *20*, *50*, *500*, *-m*, *-h*, and removed TEs that had SSRs over more than 50% of their sequences. Such method has been previously used to remove TE copies that are indistinguishable from simple repeats (34,162).

TE enrichment at regulatory regions

To identify TE insertions within and nearby genes for each genome, we created *bed* files with the genomic positions of all insertions. Then, we created *bed* files for the gene regions corresponding to the 2 Kb upstream to the transcription start-site (TSS). In order to test whether the prevalence of TEs in each gene region was significantly higher than expected (enriched) compared to random genomic distributions of TEs in the same region of other genes, we performed permutation tests with the R package *regioneR* (163), parameter *randomization* = *resampleRegions*, with 1,000 random samples. Therefore, for each gene family with *n* genes, the number of genes with TEs was compared to other random *n* genes 1,000 times. Such analysis allowed us to infer whether there are HLAU gene families with the observed frequency of genes with TEs at the promoter region higher than expected compared to the rest of the genome. Taking into account that TEs might harbor regulatory and/or coding motifs in both strands (98), our analysis considered TEs within the investigated gene regions regardless the strand.

Identification of TE-rich genomic regions

We split the genomes in bins of 200 kb, generating a bed file with such intervals across the genomes. Then, we used the function *intersect* from *bedtools v.2.30.0* (164) to identify the number of TE insertions located in each genomic bin. Finally, we calculated the 3rd quartile with the frequency of TEs in each region and defined as TE-rich genomic regions those with more TE insertions than the 3rd quartile. We used the same function from *bedtools* to identify whether genes were present in the TE-rich regions, or in the TE-poor (less TEs than the 3rd quartile).

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Transcription factor binding sites on TE insertions

The identification of transcriptions factor binding sites (TFBSs) was assessed with *FIMO* (165), from the *MEME v. 5.0.5* suite tools (166). All TE insertions located upstream ORs had their sequences extracted from the genome with *bedtools v. 2.30.0*, function *getfasta*, using parameter *-s* (164). We obtained the TFBS motifs described for *D. melanogaster* from Jaspar database (167), for the transcription factors (TFs) acj6, zf30c, fer1, onecut, and xbp1, which have been demonstrated to activate ORs (55). We only considered TEs with motifs identified with p-value < 0.01.

Differential expression of genes

The RNA-seq reads were aligned with *STAR v.2.7.3* (168), using the assembled genomes sequences and annotations of their respective species. The alignments were used to create the count table reporting the expression of each gene, with *htseq v.0.13.5* (169), parameters *–order pos*; *-s reverse*; *–i gene*; *-t exon*; the latter indicates that the count for each gene corresponds to the sum of all reads aligned into exons (CDSs and UTRs). We have merged the counts from *mulleri* species and *buzzatii* species separately, maintaining only ortholog genes that have at least 90% of the length of the annotated gene in the reference genome. Such threshold allowed us to avoid misinterpretation of expression level by comparing gene with different lengths. Then, we performed independent statistical analysis for *mulleri* and *buzzatii* species with *DEseq2* v.1.28.1 (155), using single factor experimental design, which considers only the factor that different species may have differentially expressed genes. Due to the difference on gene annotation between *mulleri* and *buzzatii* species, we performed the differential expression analysis separately for each group. Only genes with adjusted *p-value* < 0.05 and *log2 fold-change* > |1| were considered as differentially expressed. To test for association between TE enrichment and differential expression, we selected HLAU genes with the mean of normalized counts between replicates

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

>= 100. Then, we build the contingency table to apply X^2 independency test splitting them into four categories: differentially expressed, non-differentially expressed, with TEs at promoter region, and without TEs at the promoter region. The test was carried out for all pairwise analysis between cactophilic species with preference for *Opuntia sp* and columnar cacti.

Identification of gene-TE chimeric transcripts

Gene-TE chimeric transcripts were detected with *ChimeraTE v.1.1.1* Mode 1 (34). Each species was analyzed using its respective genome, gene annotation, and TE annotation. Shortly, *ChimeraTE* identifies chimeras based on paired-end reads spanning from exons to TE sequences. Based on the TE position compared to the gene (upstream, exon, intron, downstream), chimeric transcripts are classified in categories. We have used default parameters, except *--replicate 3*, which consider true chimeras only those identified in the three RNA-seq replicates.

Protein domains on TEs embedded into CDSs

The sequences of TE-exonized elements embedded into CDSs were obtained *bedtools getfasta*, parameter *-s* (164). Due to the unknown reading frame of the TE sequence incorporated into transcript, we extracted the six open reading frames with *EMBOSS v.6.6.0 getorf* (171). Then, we identified the protein domains with *pfam-scan v.1.6* (172), using *Pfam-A.hmm* database from InterPro (173), with default parameters. We selected TEs with protein domains associated with TEs, following the description from InterPro, and *e-value < 1E-05*.

Declarations

Availability of data and materials

The genome assemblies and their respective gene and TE annotation generated and analyzed during the current study are available in the Zenodo repository, doi: https://zenodo.org/doi/10.5281/zenodo.10886791.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Competing interests

The authors declare that they have no competing interests.

Funding

The present study has been 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] 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.

Authors' contributions

A. L., A. B, M. P. G. G, performed experiments; F.S., Performed genome assemblies; W. V. B. N. performed phylogenetic analysis; D. S. O performed computational analysis for genomics and transcriptomics analysis; D. S. O. wrote and edited the paper; C. V. and C. M. A. C. designed the study, and reviewed the manuscript. All authors read, corrected, and approved the final manuscript.

Acknowledgements

We thank Séverine Chambeyron for useful discussions and advice. This work was performed using the computing facilities of the CC LBBE/PRABI.

References

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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

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

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

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

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

11. Larter NK, Sun JS, Carlson JR. Organization and function of *Drosophila* odorant binding proteins. eLife. 2016; 15;5:e20242.

12. Guo S, Kim J. Molecular Evolution of *Drosophila* Odorant Receptor Genes. Molecular Biology and Evolution. 2007; 24(5):1198–207.

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

14. Ni L. The Structure and Function of Ionotropic Receptors in *Drosophila*. Front Mol Neurosci. 2021; 1;13:638839.

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

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

17. Wu C, Chakrabarty S, Jin M, Liu K, Xiao Y. Insect ATP-Binding Cassette (ABC) Transporters: Roles in Xenobiotic Detoxification and Bt Insecticidal Activity. IJMS. 2019; 20(11):2829.

18. Chung H, Sztal T, Pasricha S, Sridhar M, Batterham P, Daborn PJ. Characterization of *Drosophila melanogaster* cytochrome P450 genes. Proceedings of the National Academy of Sciences. 2009; 106(14):5731–6.

19. Good RT, Gramzow L, Battlay P, Sztal T, Batterham P, Robin C. The Molecular Evolution of Cytochrome P450 Genes within and between *Drosophila* Species. Genome Biology and Evolution. 2014; 6(5):1118–34.

20. Riveron J, Boto T, Alcorta E. Transcriptional basis of the acclimation to high environmental temperature at the olfactory receptor organs of *Drosophila melanogaster*. BMC Genomics. 2013; 14(1):259.

21. Luque T, O'Reilly DR. Functional and phylogenetic analyses of a putative *Drosophila melanogaster* UDP-glycosyltransferase gene. Insect Biochemistry and Molecular Biology. 2002; 32(12):1597–604.

22. Oakeshott JG, van Papenrecht EA, Boyce TM, Healy MJ, Russell RJ. Evolutionary genetics of *Drosophila* esterases. Genetica. 1993; 90(2–3):239–68.

23. Campbell PM, de Q. Robin GC, Court LN, Dorrian SJ, Russell RJ, Oakeshott JG. Developmental expression and gene/enzyme identifications in the alpha esterase gene cluster of *Drosophila melanogaster*. Insect Mol Biol. 2003;12(5):459–71.

24. Montella IR, Schama R, Valle D. The classification of esterases: an important gene family involved in insecticide resistance - A review. Mem Inst Oswaldo Cruz. 2012; 107(4):437–49.

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

26. Catalán A, Hutter S, Parsch J. Population and sex differences in *Drosophila melanogaster* brain gene expression.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

BMC Genomics. 2012; 13(1):654.

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

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

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

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

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

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

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

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

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

36. Pfeiler E, Castrezana S, Reed LK, Markow TA. Genetic, ecological and morphological differences among populations of the cactophilic *Drosophila mojavensis* from southwestern USA and northwestern Mexico, with descriptions of two new subspecies. Journal of Natural History. 2009; 43(15–16):923–38.

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

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

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

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

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

42. Benowitz KM, Allan CW, Jaworski CC, Sanderson MJ, Diaz F, Chen X, et al. Chromosome-length genome assemblies of cactophilic *Drosophila* illuminate links between structural and sequence evolution. bioRxiv; 2022; doi: https://doi.org/10.1101/2022.10.16.512445.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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

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

47. Smith G, Lohse K, Etges WJ, Ritchie MG. Model-based comparisons of phylogeographic scenarios resolve the intraspecific divergence of cactophilic *Drosophila mojavensis*: Phylogeography of *Drosophila mojavensis*. Molecular Ecology. 2012; 21(13):3293–307.

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

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

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

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

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

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

54. Pollard MO, Gurdasani D, Mentzer AJ, Porter T, Sandhu MS. Long reads: their purpose and place. Human Molecular Genetics. 2018; 27(R2):R234–41.

55. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Molecular Cell. 2020; 80(5):915-928.e5.

56. Debladis E, Llauro C, Carpentier MC, Mirouze M, Panaud O. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. BMC Genomics. 2017; 18(1):537.

57. Shahid S, Slotkin RK. The current revolution in transposable element biology enabled by long reads. Current Opinion in Plant Biology. 2020; 54:49–56.

58. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. Nat Commun. 2022; 13(1):1948.

59. González J, Nefedov M, Bosdet I, Casals F, Calvete O, Delprat A, et al. A BAC-based physical map of the *Drosophila buzzatii* genome. Genome Res. 2005; 15(6):885–92.

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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

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

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

67. Arguello JR, Cardoso-Moreira M, Grenier JK, Gottipati S, Clark AG, Benton R. Extensive local adaptation within the chemosensory system following *Drosophila melanogaster*'s global expansion. Nat Commun. 2016; 7(1):ncomms11855.

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

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

70. Gozashti L, Corbett-Detig R. Universal signatures of transposable element compartmentalization across eukaryotic genes. bioRxiv; 2023; doi: https://www.biorxiv.org/content/10.1101/2023.10.17.562820v1.

71. Stapley J, Santure AW, Dennis SR. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. Mol Ecol. 2015; 24(9):2241–52.

72. Serrato-Capuchina A, Matute D. The Role of Transposable Elements in Speciation. Genes. 2018; 9(5):254.

73. Moschetti R, Palazzo A, Lorusso P, Viggiano L, Massimiliano Marsano R. "What You Need, Baby, I Got It": Transposable Elements as Suppliers of Cis-Operating Sequences in *Drosophila*. Biology. 2020; 9(2):25.

74. Oliveira DS, Rosa MT, Vieira C, Loreto ELS. Oxidative and radiation stress induces transposable element transcription in *Drosophila melanogaster*. J of Evolutionary Biology. 2021; 34(4):628–38.

75. Moskalev A, Zhikrivetskaya S, Krasnov G, Shaposhnikov M, Proshkina E, Borisoglebsky D, et al. A comparison of the transcriptome of *Drosophila melanogaster* in response to entomopathogenic fungus, ionizing radiation, starvation and cold shock. BMC Genomics. 2015; 16(S13):S8.

76. Moskalev A, Shaposhnikov M, Snezhkina A, Kogan V, Plyusnina E, Peregudova D, et al. Mining Gene Expression Data for Pollutants (Dioxin, Toluene, Formaldehyde) and Low Dose of Gamma-Irradiation. PLoS ONE. 2014; 9(1):e86051.

77. Craddock EM. Profuse evolutionary diversification and speciation on volcanic islands: transposon instability and amplification bursts explain the genetic paradox. Biol Direct. 2016; 11(1):44.

78. Fernández-Medina RD, Granzotto A, Ribeiro JM, Cararetto CMA. Transposition burst of mariner-like elements in the sequenced genome of *Rhodnius prolixus*. Insect Biochemistry and Molecular Biology. 2016; 69:14–24.

79. Pace JK, Feschotte C. The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage. Genome Res. 2007; 17(4):422–32.

80. Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci USA. 1997; 94(15):7704–11.

81. Rius N, Guillén Y, Delprat A, Kapusta A, Feschotte C, Ruiz A. Exploration of the *Drosophila buzzatii* transposable element content suggests underestimation of repeats in *Drosophila* genomes. BMC Genomics. 2016; 17(1):344.

82. Labrador M, Farre M, Utzet F, Fontdevila A. Interspecific hybridization increases transposition rates of Osvaldo. Molecular Biology and Evolution. 1999; 16(7):931–7.

83. Naveira H, Fontdevila A. The evolutionary history of *D. buzzatii*. XXII. Chromosomal and genic sterility in male

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

hybrids of *Drosophila buzzatii* and *Drosophila koepferae*. Heredity. 1991; 66(2):233–9.

84. Luo S, Lu J. Silencing of Transposable Elements by piRNAs in *Drosophila*: An Evolutionary Perspective. Genomics, Proteomics & Bioinformatics. 2017; 15(3):164–76.

85. *Drosophila* 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 2007; 450(7167):203–18.

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

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

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

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

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

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

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

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

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

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

96. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers Xenobiotic Resistance in *Drosophila*. Feschotte C, editor. PLoS Genet. 2014; 10(8):e1004560.

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

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

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

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

101. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze Genomic Innovation in a

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

Mammalian Lineage. Genome Biology and Evolution. 2014; 6(10):2595–610.

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

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

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

105. Carareto CMA, Hernandez EH, Vieira C. Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling *Drosophila* species. Gene. 2014; 537(1):93–9.

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

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

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

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

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

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

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

113. Feyereisen R. Insect CYP Genes and P450 Enzymes. In: Insect Molecular Biology and Biochemistry. Elsevier; 2012. p. 236–316.

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

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

116. Bartolomé C, Maside X, Charlesworth B. On the Abundance and Distribution of Transposable Elements in the Genome of *Drosophila melanogaster*. Molecular Biology and Evolution. 2002; 19(6):926–37.

117. Gao C, Zhou G, Ma C, Zhai W, Zhang T, Liu Z, et al. Helitron-like transposons contributed to the mating system transition from out-crossing to self-fertilizing in polyploid *Brassica napus* L. Sci Rep. 2016; 6(1):33785.

118. Garrigues JM, Tsu BV, Daugherty MD, Pasquinelli AE. Diversification of the *Caenorhabditis* heat shock response by Helitron transposable elements. eLife. 2019; 8:e51139.

119. Guo C, Spinelli M, Liu M, Li QQ, Liang C. A Genome-wide Study of "Non-3UTR" Polyadenylation Sites in *Arabidopsis thaliana*. Sci Rep. 2016; 6(1):28060.

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

120. Lee JY, Ji Z, Tian B. Phylogenetic analysis of mRNA polyadenylation sites reveals a role of transposable elements in evolution of the 3′-end of genes. Nucleic Acids Research. 2008; 36(17):5581–90.

121. Kelley ST, Farrell BD, Mitton JB. Effects of specialization on genetic differentiation in sister species of bark beetles. Heredity. 2000; 84(2):218–27.

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

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

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

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

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

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

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

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

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

131. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013; doi: http://arxiv.org/abs/1303.3997.

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

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

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

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

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

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

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

141. Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo. 2023; doi: https://zenodo.org/records/8178877.

142. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006; 34:W609–12.

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

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

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

147. Baril TJ, Galbraith JD, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. bioRxiv; 2024; doi: https://www.biorxiv.org/content/10.1101/2022.06.30.498289v3.

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

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

150. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005; 21:i351–8.

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

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

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

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

155. SMIT, Arian FA. RepeatMasker Open 3.0. 2004.

156. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016; 44(D1):D81–9.

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

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

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

bioRxiv preprint doi: https://doi.org/10.1101/2024.03.27.587021; this version posted March 30, 2024. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY 4.0 International license.

80.

160. Flutre T, Permal E, Quesneville H. Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In: Plant Transposable Elements. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012. p. 17–39.

161. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2015; btv562.

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

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

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

165. Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, et al. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research. 2019; gkz1001.

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

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

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

169. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. Nucleic Acids Research. 2021; 49(D1):D412–9.

170. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Research. 2023; 51(D1):D418–27.