# Transposable elements as evolutionary driving force to host shift 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] Universite Claude Bernard Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, 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

## ABSTRACT

### *Background*

The host shift in insects has been considered a key process with potential to collaborate with reproductive isolation and 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 evolution, since they use a wide-range of cacti as hosts, and many species have different preferences. Transposable elements are engines of genetic novelty between populations and species, driving rapid adaptive evolution. However, the extent of TEs' contribution to host shift remains unexplored.

### *Results*

We performed genomic and transcriptomic analysis in seven genomes of cactophilic species/subspecies to investigate how TEs interact with genes associated with host shift. Our results revealed enrichment of TEs at promoter regions of host shift-related genes, with *Helitrons* representing ~60% of the cases, demonstrating an unprecedented putative *cis*-regulatory role of *Helitrons* in *Drosophila*. Differential expression analysis between species with different preferred hosts demonstrated divergence on gene expression in head and larvae tissues. Although TEs' presence does not affect overall gene expression, we observed 1.31% of genes generating gene-TE chimeric transcripts, including those with function affecting host preference.

### *Conclusions*

Our combined genomic and transcriptomic approaches provide evidence of TE-driven divergence between species, highlighting the evolutionary role of TEs in the context of host shift, a key adaptive process that can cause reproductive isolation.

**Keywords**: comparative transcriptomics, adaptation, chimeric transcripts.

## Background

The molecular consequences of host shift evolution can be closely related to ecological divergence, specialization and ultimately ecological speciation (1). Due to the wide variety of hosts, successful plant-insect interactions require adaptations from the insect. In many species, reproductive isolation arises as a result of successive adaptations driven by host shift (HS), which results from divergent selection between different ecological environments (1–4). Thus, natural selection increases the frequency of alleles that confer fitness benefits when HS occurs. *Drosophila* species use a wide range of necrotic host tissues as feeding and breeding sites, providing a suitable model to study the HS process and its role in insect speciation. Among them, the *repleta* group comprises a clade with ~100 cactophilic *Drosophila* species (5). The specialization in cacti represents an important ecological challenge to the flies, due to the lack of nitrogen and phosphorus compared to other hosts, as well as the presence of toxic compounds (6). In the *repleta* species, the HS was initially to the *Opuntia sp*. cacti, and posteriorly several independent shifts happened to columnar cacti (6), considered a host with higher toxicity than *Opuntia sp.* (7). Such HS is observed between sibling species with recent divergence in two clusters of the subgroup *mulleri*: cluster *buzzatii* and cluster *mojavensis*. In the cluster *buzzatii*, *Drosophila buzzatii* and *D. koepferae* are sibling species (divergence 4-5 Mya (8)) with different primary hosts. The former species use preferentially *Opuntia sp.* cacti, whereas the latter is a columnar cacti dweller (9). In the cluster *mojavensis*, *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* (10). Therefore, this set of species from *buzzatii* and *mojavensis* clusters provide an excellent model to investigate the molecular mechanisms underlying HS evolution.

The radiation of cactophilic *Drosophila* species across different hosts is an evolutionary outcome that gave rise to specific adaptations (11), which may be summarized into three steps: location, acceptation and host usage (12). In the location step, odorant-binding proteins (OBPs) and odorant receptors (ORs) are key proteins to distinguish specific host in the environment, through chemoreceptors that integrate the visual and olfactory systems. Subsequently, in the acceptance step occurs the evaluation of the nutritional compounds present on the host, as well as the presence of competitors, predators, pathogens and parasites (12). Several receptors are associated with this process, such as gustatory receptors (GRs) (13) and ionotropic receptors (IRs). (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 (ESTs) and ATP binding-cassette transporters (ABCs). Altogether, these nine gene families associated with the Host Location, Acceptance and Usage will henceforth be referred to as HLAU.

A previous study with 14 *Drosophila* genomes, representing both *Sophophora* and *Drosophila* subgenera, has proposed that 82 genes, including GSTs, CYPs, and UDPs, are under positive selection (17). Signatures of positive selection were also found between *D. mojavensis* subspecies and *D. arizonae* in subsets of GRs and ORs (16). In addition to the selective pressure acting into HLAU gene families, modifications in the gene expression also play an important role in the HS adaptation in *Drosophila* species. It has been shown that CYPs, GSTs and OBPs are differentially expressed between *D. m. wrigleyi* and *D. m. sonorensis* (18), indicating that gene expression in the nervous system might play a role into detoxification and olfactory systems, hence contributing to HS adaptation in *D. mojavensis*.

Transposable elements (TEs) are mobile sequences with the ability to create structural variants in the genomes. In most cases, mutations caused by TEs are likely to be deleterious or neutral. Throughout evolutionary time, TEs that remain in the genome tend to be silenced

by epigenetic control, accumulating mutations, and losing their transposition ability (19). Despite they become degraded, the remaining TE sequences may still contain either regulatory motifs or protein domains (20). These sequences might be co-opted by the cell machinery modifying gene expression or protein sequences of the nearby genes (reviewed in (21)). Depending upon the positive outcome of the cooption into the individual fitness, these TEs can increase their frequency in the population, contributing to populational evolution. For instance, in *D. melanogaster*, the gene *CHKov1* has a truncated mRNA derived from a *DOC* TE located into the intron, resulting in resistance to viral infection and insecticides (22). Such exaptation and domestication events can be often identified by the occurrence of chimeric transcripts, which are mRNAs with both gene and TE-derived sequences (23). A recent transcriptome-wide study has identified that ~1.12% of *D. melanogaster* genes produce chimeric transcripts in different populations (24). Among all genes, 76 generate 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 *Drosophila* ecotypes.

Many aspects involving HS adaptation in *repleta* species have been shown previously, such as detoxification pathways (25), morphology (26), life history traits (27), behavior (28), and genomic and transcriptomic differences (18,29–32). However, the contribution of TEs to the evolution of 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 analysis. We tested the hypothesis regarding the contribution of TEs to the HS in cactophilic species. The genome-wide analysis revealed that HLAU gene families are enriched with TEs in regulatory regions, particularly in ORs. Furthermore, the results show that approximately 35% of ORs with TEs in their promoter regions contain transcription factor binding sites (TFBSs) within the TE sequence, of which around 60% are Helitrons, highlighting their unprecedented potential regulatory role in *Drosophila*. The transcriptomic analysis revealed that HLAU genes are differentially expressed

between cactophilic species. Additionally, we also show that 1.31% of genes produce chimeric transcripts in head and larvae tissues, with HLAU genes are also being involved.

## 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 on species that have different preferential cacti as hosts: *D. buzzatii* (*Opuntia sp.*), *D. koepferae* (columnar cacti), *D. arizonae* (*Opuntia sp.* or 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, and 95.98% in *D. arizonae*. The *de novo* gene annotation in *D. buzzatii* and *D. koepferae* revealed a total of 18,050 and 17,848 coding genes, for which 63.77%, and 64.13% were successfully assigned as one-to-one orthologs with *D. mojavensis*, respectively. These results represent the higher gene repertoire identified for *D. buzzatii* and *D. koepferae* species, altogether with its ortholog relationship with *D. mojavensis* and *D. arizonae*. 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 usage: 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. In *D. buzzatii* and *D. koepferae* the number of annotated HLAU genes covered ~95% in both species compared to *D. mojavensis* (Table 1).

| 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* | 18,050 | 25 | 52 | 34 | 21 | 27 | 16 | 27 | 20 | 68 |
| *D. koepferae* | 17,848 | 25 | 45 | 34 | 24 | 27 | 13 | 25 | 23 | 62 |

**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 usage: ATP-binding cassette (ABC), carboxylesterase (EST), glutathione S-Transferase (GST), glycosiltransferase (UDP), and cytochrome P450 (CYP).

### *Molecular evolution of HLAU genes between species with preference for Opuntia sp. 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 (6). To investigate whether species with preferential columnar cacti usage have HLAU genes that undergone positive selection, we used a branch-site approach to assess their evolutionary rates compared to species that use *Opuntia sp*. For this analysis, we removed *D. arizonae* due to its non-preferential usage between columnar and *Opuntia sp.* cacti (6). Moreover, we added *D. navojoa*, which is a basal species of the *mojavensis* cluster that use solely *Opuntia* sp. (33). 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 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, we did not observe any gene under positive selection with FDR < 0.05. To further explore the evolutionary forces driving gene HLAU genes evolution, we also performed a test for relaxed selection. We observed the genes *Obp56d* (LOC6580688) and *Cyp304a1* (LOC6572829) with intensified selection strength (k > 1 and FDR < 0.05). Other six HLAU genes had significant relaxed selection, but with k < 1 (Sup. Table 4).

***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 TE content in the genomes ranged from 14.07% in *D. buzzatii* to 20.4% in *D. m. mojavensis* (Fig. 1A). In addition, we observed a positive correlation between TE content and genome size estimated from the assemblies (*Pearson* correlation: $p = 0.013$; $r = 0.86$) (Sup. Fig. 3). Following the correlation, both species from the *buzzatii* cluster had the lower genome sizes and TE contents compared to species of the *mojavensis* cluster (Fig. 1A). 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 to *Helitrons* and LINEs (Fig. 1B).

In the *mojavensis* cluster (*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 (34). *D. m. mojavensis* has the TE-richest genome (20.45%), with the abundance following from the highest to the lowest: LTR > LINE > Helitron > DNA > Unknown (Fig. 1B). 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, but followed

by Helitrons, LINEs, DNA, and Unknown (Fig. 1B). The observed differences are often associated with the divergence on the evolutionary patterns of the TE classes between the genomes. Our analysis revealed that *Helitrons* had a recent burst on *D. m. mojavensis* and *D. m. wrigleyi*, followed by a decline (Fig. 1C). In *D. arizonae* the patterns reveals high genome content derived from TEs with low Kimura 2-Parameters divergence, which indicates conserved and potentially active copies. The antagonistic patterns of TE dynamics observed in the seven analyzed genomes suggest that TEs contribute differently to genome evolution across these species.

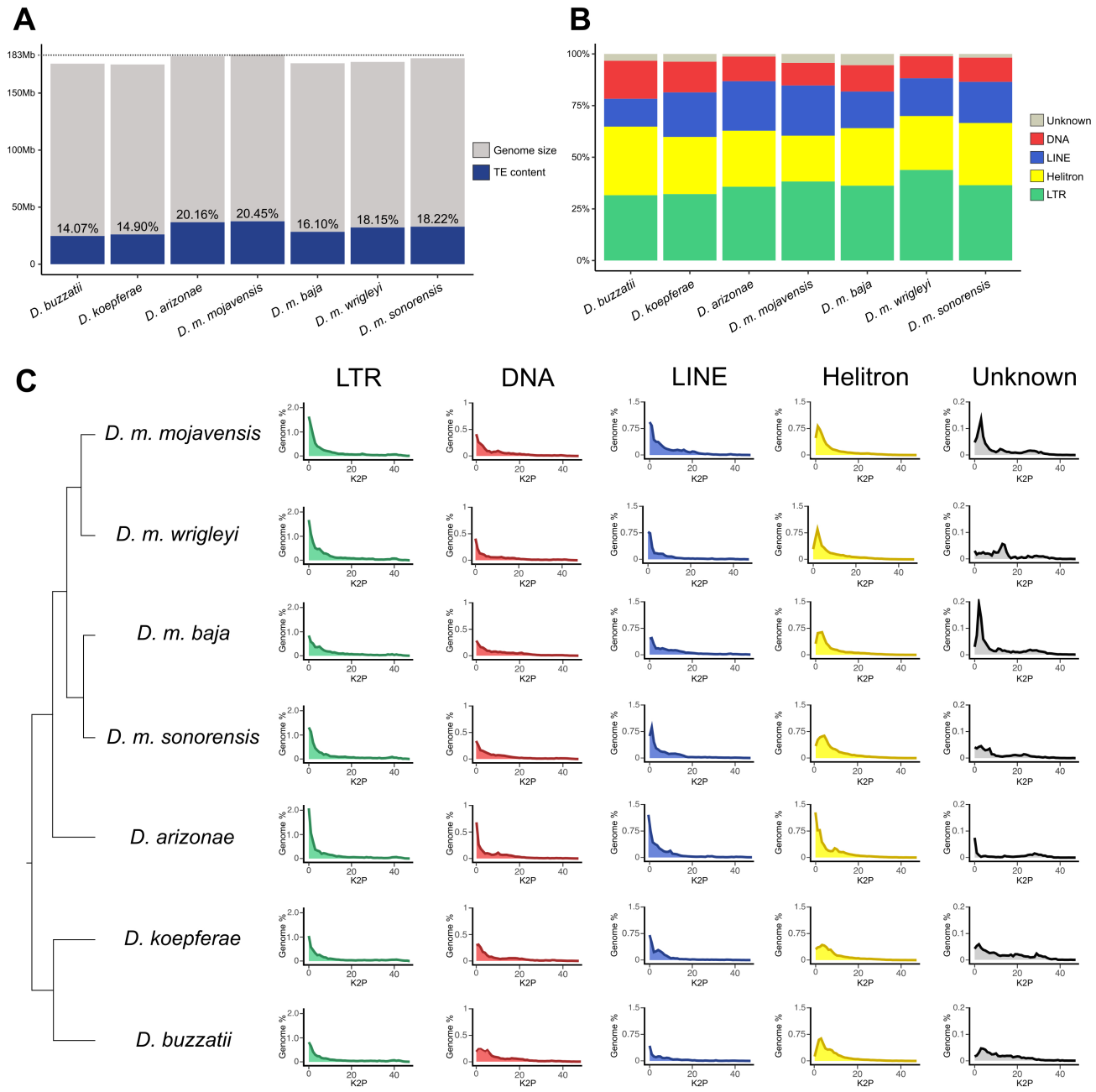


**Figure 1. A**) Genome size of the seven cactophilic *Drosophila* species sequenced in this work, and their respective TE content. **B**) Proportional distribution of the total TE content annotated in the genomes, divided by DNA transposon, LINE, Helitron, LTR, and Unknown. **C**) Phylogenetic relationship among species (adapted from (25,35)), alongside with TE families landscapes representing TE content in genome % (y-axis), and the divergence of TE insertions compared to the consensuses (x-axis), based on Kimura 2-parameters (K2P) 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.

***HLAU genes are enriched in TEs at regulatory regions***

To investigate whether TEs may have been maintained over time in a non-random distribution in the genomes relative to genes, we compared their positions and analyzed their frequencies in each genome. We analyzed whether specific TE orders are preferentially located within genes, which would suggest possible co-evolution. 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 did not observe any TE order enriched within genes, which is expected with the hypothesis of purifying selection against TEs disrupting gene sequences (Sup. Fig. 4).

Despite TEs were not enriched on gene regions overall, a few insertions may be selected in the promoter of genes if they provide beneficial outcome to the individual fitness, through donation of histone marks enrichment, or *cis*-regulatory elements (36,37). 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. To make a comparison of our findings with non-host shift related genes, we also included zinc-finger and serine/threonine kinases as two non-HLAU gene families. Regarding genes associated with location (OBPs and ORs), we observed that ORs have enrichment of TEs at the promoters in all genomes from species of the *mojavensis* cluster (p-value < 0.05), (Fig. 2A). For acceptance genes, *D. koepferae* had enrichment of TEs in the promoters of GRs (Fig. 2A). Finally, enrichment of TEs upstream to detoxification genes was observed in *D. arizonae* and *D. m. baja*. The latter had the enrichment for the CYP and Esterase families, whereas the latter had ABCs enriched by TEs (Fig. 2A). Overall, detoxification genes had z-score towards depletion of TEs on promoters, for which we observed significant depletion on *D. buzzatii* (Fig. 2A). In the non-HLAU gene families, we observed the expected result by chance: either z-score indicated the expected number of genes with TEs on promoters in a genome-wide comparison (z-score near 0), or it indicated depletion of TEs, as observed in *D. buzzatii* and

*D. koepferae*. This result reinforces our findings indicating the significant and evolutionary unexpected enrichment of TEs, especially on the promoters of ORs.

The results depicting the histograms with frequency of genes with TEs for each permutation test, and p-values are shown in the Sup. Fig. 5 and Sup. Table 5. 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 significant enrichment, indicating that the frequency of TEs downstream to HLAU genes follows the genomic distribution in these species. However, in the species of the *mojavensis* cluster, ORs had enrichment for TEs in the downstream region, as we have also observed for the 2 kb upstream. 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*. Our results suggest a possible ancient evolutionary path of TE cooption in species of the *mojavensis* cluster, since enrichment of TEs in ORs was observed in the five genomes.

The enrichment of TEs upstream and downstream of ORs indicates a possible coevolution of this gene family with TEs. Alternatively, it may also indicate that ORs are located in TE-rich regions of the genomes. To investigate this hypothesis, we identified TE-rich regions (Sup. Fig. 6) and analyzed the frequency of ORs. We observed on average 6 and 3 ORs in such regions in species of the *mojavensis* and *buzzatii* clusters, respectively (Sup. Table 6). 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 nearby ORs, such as the cooption of regulatory motifs.

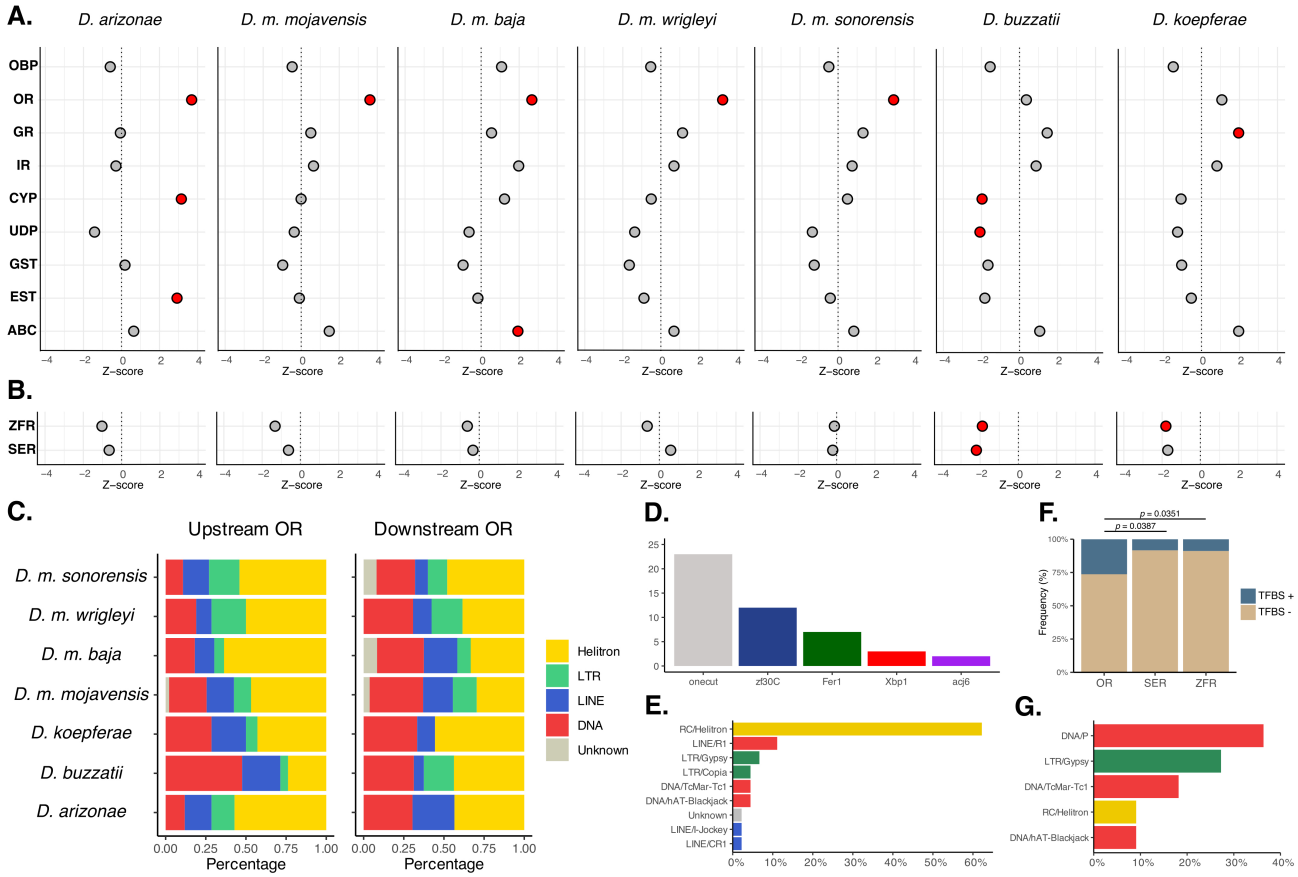


**Figure 2. A**) Enrichment of TEs in the promoter region (2kb upstream) of HLAU gene families. Z-score > 0 indicates enrichment, and Z-score < 0 depletion. Red dots: p-value < 0.05, grey dots: p-value > 0.05. **B**) The same enrichment analysis as depicted on A, but for non-HLAU gene families, represented by zinc-finger proteins (ZFR) and Serine/Threonine Kinases (SER). **C**) Proportion of TE orders located upstream and downstream ORs. **D**) Transcription factor binding sites (TFBSs) frequency on TEs located upstream OR genes in all species under study. **E**) Proportional frequency of TE families with TFBSs embedded on their sequences from TE copies located upstream OR genes. **F**) Significant Fisher's exact test between OR genes with OR-related TFBSs from TE insertions compared to the frequency observed in the non-HLAU gene families Serine/Threonine Kinases (SER) and zinc-finger proteins (ZFR). **G**) Proportion of TE orders carrying OR-related TFBSs in non-HLAU gene families. *Helitrons* represent only 10% of the copies.

A few TE families have been shown to be more prone to contribute with *cis*-elements than others, such as retroelements in *D. melanogaster* (38). Therefore, we investigated whether the observed enrichment of TEs in the HLAU gene families is TE-specific. Considering all species, 50.42% and 40% of the insertions located upstream and downstream of ORs

correspond to *Helitrons*, respectively (Fig. 3C). *Helitrons* have been documented to act on gene evolution of bats by carrying gene sequences (39), mainly 5′ UTRs and the first exons, within their own internal regions (40). Thus, we analyzed 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*, suggesting a subspecies-specific insertion, whereas the latter is observed in the four *D. mojavensis* subspecies, likely be to inherited from their common ancestor (Sup. table 7). At the ORs' 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 7). 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 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 transcription factors (TFs) previously described to participate on OR's transcription in *D. melanogaster*: acj6, onecut, xbp1, fer1 and zf30C (41). Our analysis revealed that 35.71% of the ORs with TEs at the promoter region have the TFBS located within the insertion (Sup. table 8). Taking all species together, the most frequent TFBSs found was for the TF onecut (48.93%), followed by zf30C (25.53%), and Fer1 (14.89%) (Fig. 3D). *Helitrons* were the TEs with the highest frequency, representing 59.57% of all insertions carrying ORs' TFBSs (Fig. 3E). To test whether these results are specific to ORs, we also performed the same TFBSs' detection and frequency of TEs ZFR SER genes as non-HLAU families. Our results show that ZFR and

SER have on average 9.75% and 9.09% of the genes with OR-related TFBSs, respectively, which can be considered as noise from the analysis. Comparing these proportions with the ones found in ORs, we had significantly more TFBSs compared to ZFR (*Fisher's exact test: p-value = 0.0387*) and SER (*Fisher's exact test: p-value = 0.0351*) (Fig. 2F and Sup. Table 9). In addition, *Helitrons* represented only 10% of the insertions carrying TFBSs in ZFR and SER genes (Fig. 2G). These comparative results suggest that our findings regarding the presence of OR-related TFBSs from TE insertions, and the enrichment of *Helitrons*, are unlikely to be due to chance. 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 *mojavensis* cluster (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*. These results highlight a putative functional role of TEs as donators of *cis*-elements to ORs in the cactophilic species *D. mojavensis* and *D. arizonae*.

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

Differential expression of genes has been associated with adaptation to new hosts in insects (31). Here, aiming to understand how gene expression divergence is associated with adaptation to different cactus 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) from HLAU gene expression revealed remarkable differences between species in both tissues (Fig. 3C). The variance of gene expression resembles the phylogenetic divergence of the species, with the highest variance between the *buzzatii* and *mojavensis* clusters (PC1), followed by divergence between *D. buzzatii* and *D. koepferae* (PC2). The species of the *mojavensis* cluster were clustered together with a slight separation between *D. arizonae* and the subspecies of *D. mojavensis*.

We performed pairwise analysis of HLAU genes expression between species with preference for *Opuntia sp.* (*D. m. wrigleyi* and *D. buzzatii*) versus species with preference for

columnar cacti, and *D. arizonae* as non-preferential cacti. 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. 3A). 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*. 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*. Since they were observed in all subspecies with preference to columnar cacti, the higher expression of these genes can be related to the adaptation to columnar cacti. Finally, the same analysis on larvae revealed that *D. m. wrigleyi* has three genes with recurrent higher expression compared to the other 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 (Fig. 3B). Regarding *D. buzzatii* and *D. koepferae*, we observed 27 HLAU genes in head and 54 in larvae with higher expression in *D. buzzatii*, and 63 in head and 29 with higher expression in *D. koepferae*. We also compared the expression of HLAU genes between *D. m. wrigleyi* and its sister species without preferential cacti host, *D. arizonae*. In head, the *Opuntia* sp. dweller *D. m. wrigleyi* had 31 HLAU genes with higher expression compared to its counterpart *D. arizonae*. In larvae, *D. m. wrigleyi* had 43 HLAU genes with higher expression compared to *D. arizonae* (Fig. 3D). The result demonstrates that HLAU genes have divergent differential expression in both head and larvae tissues.

To further investigate whether differential expression of HLAU genes have higher prevalence compared to non-HLAU genes, we compared the proportion of differentially expressed genes associated to location, acceptance, and usage with the proportion of differentially expressed genes from the non-HLAU gene families Ser/Thr Kinases and Zinc-finger. In head, our analysis revelaed that location and usage have higher differential expression than non-HLAU genes between *D. buzzatii* and *D. koepferae* (Sup. Figure 7). In *D. mojavensis* subspecies, we observed higher differential expression for location genes

between *D. m. wrigleyi* and *D. m. mojavensis*, and for usage between *D. m. wrigleyi* and *D. m. baja*. In larvae, we did not observe significant difference between non-HLAU and HLAU genes, but the pairwise differential expression between *D. m. wrigleyi* and three subspecies revealed higher prevalence of gene families associated with usage (detoxification) compared to non-HLAU (Sup. Figure 7). In *D. m. wrigleyi* and *D. m. baja*, we also observed higher differential expression of location genes than non-HLAU. Finally, the differential expression between *D. m. wrigleyi* and *D. arizonae* demonstrated significantly higher proportion of genes associated with location and usage compared to non-HLAU in larvae, and higher prevalence of differentially expressed usage genes in head (Sup. Figure 7). 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.

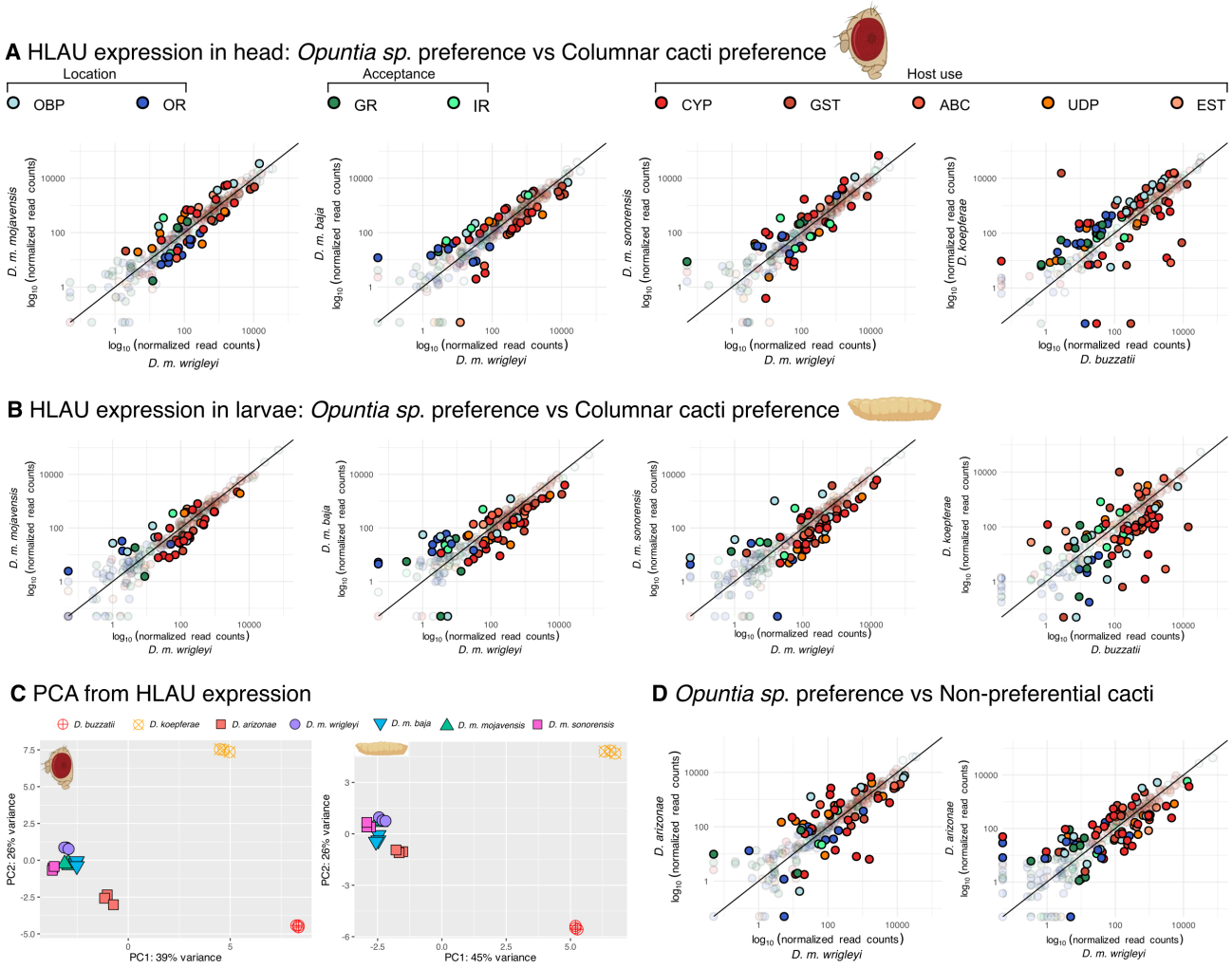

**Figure 3**: **A**) HLAU expression in head: differential expression analysis of HLAU genes from head in the *Opuntia sp.* dweller *D. m. wrigleyi* versus the three columnar dwellers *D. mojavensis,* and *D. buzzatii* versus *D. koepferae*. In A and B: Filled dots represent HLAU genes with log2 fold-change > |1| and FDR < 0.05; transparent dots are non-significant differential expression. **B**) HLAU expression in larvae: The same pairwise species as in A with *Opuntia sp.* dweller vs columnar dwellers, but from larvae tissue. **C**) Principal component analysis of HLAU genes' expression in head and larvae tissues. Both tissues had HLAU genes variance representing the phylogenetic distance between species: *D. buzzatii* and *D. koepferae* compared to *mojavensis* cluster corresponds to the PC1 (45%), and 26% between them in PC2. Species from the *mojavensis* cluster had strong similarity, with *D. arizonae* slightly separated from *D. mojavensis*. **D**) HLAU genes' expression in the *Opuntia sp.* dweller *D. m. wrigleyi* versus its sister species *D. arizonae*, which does not have preference between *Opuntia sp.* and columnar cacti.

### ***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 two variables with two categories each: 1) expression as columns: differentially expressed, and non-differentially; 2) TE presence as rows: with TEs at 2kb upstream, and without TE at 2kb upstream. Considering all pairwise analysis (Fig. 3A,B-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 10), 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. 2A). None of the enriched gene families had significant association between the presence of TEs and their expression level (Sup. Table 10). Thus, enrichment of TEs at the regulatory region is not likely to be associated with the differential expression of HLAU genes.

### *HLAU genes generate chimeric transcripts*

To investigate further cases where 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 5 and 3 HLAU genes producing chimeric transcripts per species/subspecies, respectively. We found more chimeric transcripts in head than in larvae for all HLAU gene families, except for UDPs (Fig. 4A). For instance, only head tissue had chimeras in the OBP, GR and GST gene families (Fig. 4A). TE-initiated transcripts constitute 11.36% of all chimeras found in head, and 12% in larvae (Fig. 4B). TE-exonized transcripts had a slightly higher proportion in larvae (80%) compared to head (81.81%). Finally, TE-terminated transcripts were the ones with the biggest difference between tissues, with 6.81% in head and 8% in larvae. Taken together, these results demonstrate unprecedented transcriptional interaction between TEs and HLAU genes in a tissue-specific manner.

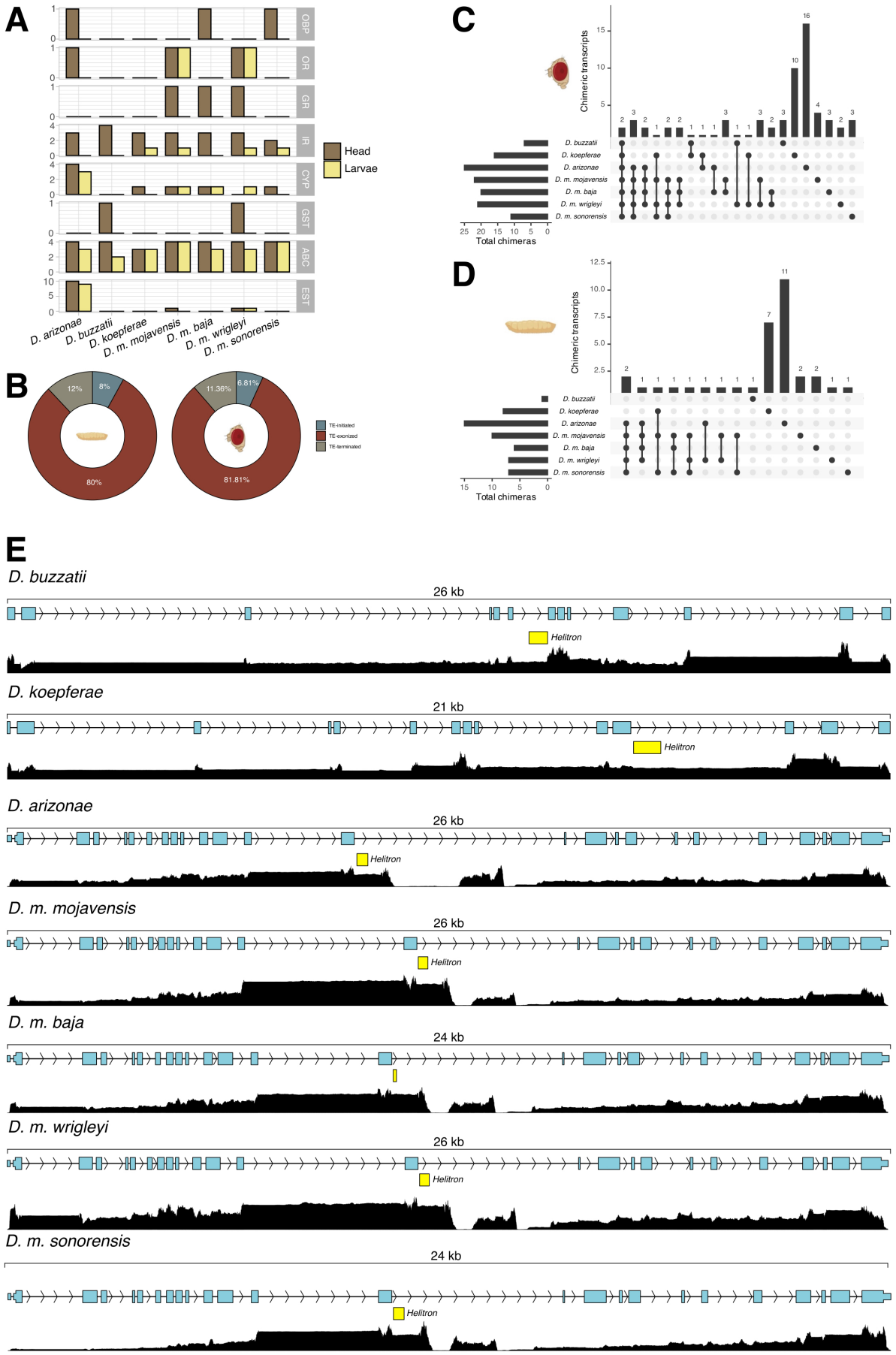
A
OBP
OR
GR
IR
CYP
GST
ABC
EST
Head
Larvae
D. arizonae
D. buzzatii
D. koepferae
D. m. mojavensis
D. m. baja
D. m. wrigleyi
D. m. sonorensis
B
12%
8%
80%
11.36%
6.81%
81.81%
TE-initiated
TE-exonized
TE-terminated
C
Chimeric transcripts
Total chimeras
D
Chimeric transcripts
Total chimeras
E
D. buzzatii
26 kb
Helitron
D. koepferae
21 kb
Helitron
D. arizonae
26 kb
Helitron
D. m. mojavensis
26 kb
Helitron
D. m. baja
24 kb
D. m. wrigleyi
26 kb
Helitron
D. m. sonorensis
24 kb
Helitron

**Figure 4**: **A**) Number of HLAU genes generating chimeric transcripts in each genome. **B**) Categories of chimeric transcripts in head and larvae tissues. **C-D**) Common chimeric transcripts between the seven genomes in head (D) and larvae (E) tissues. **E**) TE-exonized transcript from Glutamate IR, kainate 2 (LOC6585822) gene, with an *Helitron* insertion located in the intron. The same chimera was found in the seven genomes analyzed in this study.

The quantitative analysis on chimeric transcripts revealed a clear transcriptional polymorphism on HLAU genes due to TEs. We investigated common chimeras present in more than one species, highlighting their retention over evolutionary time, as well as species-specific chimeras, possibly associated with recent insertions. Our results revealed that 62.12%and 73.52% of the chimeric transcripts derived from HLAU genes are species-specific in head and larvae tissues, respectively (Fig. 4C-D). *D. arizonae* and *D. koepferae* were the ones with the highest number of HLAU genes with species-specific chimeric transcripts. In head, *D. arizonae* had 16 and *D. koepferae* 10; whereas in larvae we found 11 and 7, respectively. Regarding common chimeras, there are two expressed in head observed in all analyzed transcriptomes (Fig. 4E). The *AbcD (LOC6585769*) and *Glutamate IR, kainate 2* (*LOC6585822*) have TE-exonized transcripts expressed in the head with *Helitron* insertions. In *AbcD*, the copy is located in the intron, ~100nt away from the nearby exon, representing a case where the chimeras generated through an exon retention process. In the IR, the same process is observed since the TE is also located in the intron. Regarding the *buzzatii* cluster, we observed the gene *AbcA3* expressed only in *D. buzzatii* and *D. koepferae* with multiple insertions generating chimeras (Supp. Table 12). In the *mojavensis* cluster, two genes from the ABC gene family produced chimeras in the five transcriptomes. *Abc10* had a chimeric transcript with an *DNA/P* insertion, which is embedded in the 5′ UTR. 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. The second gene is *AbcF2*, which has a TE-exonized transcript derived from LINE/R1 insertion in all species, except *D. m. sonorensis*, which has an *Helitron*. Although not the same TE in all species, the

copies are located partially in the intron and partially in the *AbcF2*'s exon. Both recurrent chimeras in the species of the *mojavensis* cluster were observed in larvae and head. When narrowing down our analysis for chimeras observed only among *D. mojavensis* subspecies, we did not find any.

## Discussion

Insect host shift and its role in speciation have been the 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 reproductive isolation, wherein ecological speciation arises as a consequence 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. To test this hypothesis, we selected seven *Drosophila* species/subspecies that belong to the *repleta* group and have different primary cacti hosts.

### *Relaxed selection on HLAU genes*

The necrotic tissue of each cactus-host used by cactophilic *Drosophila* species differs in its composition in terms of yeasts and alkaloids (42,43). Therefore, species have adapted to feed and breed on different hosts, as demonstrated by behavioral and physiological response when species feed and breed in non-preferential hosts. Positive selection in the HS process acts as an evolutionary force, favoring mutations that lead to amino acid substitutions. These substitutions enhance the ability to recognize a broader spectrum of odors, metabolize nutrients, and activate detoxification pathways (44). Indeed, specialist species are likely to lose some of their genes related to sensory pathways, while the genes that are retained are subject to positive selection (45,46). 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 (47). Although a previous genome-wide analysis has identified 1,294 genes under positive selection (48), here we focused the analysis to investigate specifically HLAU gene families. Our results with brach-site model did not have any HLAU gene under positive selection when considering the species' host preference. We argue that perhaps the branch-site approach was too stringent. The unprecedent method used in this work focused on the common differences of species adapted to columnar cacti, rather than independent evolutionary events. The lack of genes with statistical support for positive selection may be 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 (49). 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.

Due to the lack of positive selection from our branch-site test, we decided to chose to apply the RELAX model to identify evidence of relaxed selection across HLAU genes. This model is particularly suited for detecting shifts from strong purifying selection to relaxed selection (50), which may not be identified by conventional methods focusing solely on positive selection, such as the one employed by CODEML. This is critical because evolutionary processes often involve not only episodes of adaptive evolution but also phases of selection relaxation, where selective pressures diminish but still influence gene evolution (51). The two HLAU genes (*Obp56d* and *Cyp304a1*) with significant relaxed selection suggest undergoing evolutionary transitions in response to changes in host preference. This result demonstrate that selective pressures on certain genes may diminish either during HS or after, leading to a period of relaxed selection as the species adapt to novel hosts. This relaxation of selection could indicate a functional change or reduced evolutionary constraint as these species exploit new ecological niches (52). Given that the species analyzed in this study have different primary hosts, the observed relaxed selection across these genes could reflect their

involvement in the physiological or behavioral adaptations required for successfully exploiting new hosts.

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

The literature has recurrently demonstrated that TEs can create new regulatory elements, including gene promoters (55–57), enhancers (58,59), and insulators (60,61). In mammals, TE-derived sequences contribute up to 40% of genome-wide binding sites for transcription factors located in TE-rich regions (62). 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 (63–65). Because the invasion of new habitats is accompanied by modifications in behavior, which has been associated with alterations in sensory cues and environment perception (66,67), 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, EST, and ABC genes, but OR gene family had the highest enrichment in five 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 (68). 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. *Helitrons* have been extensively reported on promoter regions of other eukaryotic species (69–72), in which a few cases of embedded TFBSs were reported (73,74). Here, we demonstrated that ~60% 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 *mojavensis* cluster might demonstrate

potential cases of TE co-option inherited by common ancestor. Our genomic analysis provides a robust list of potential candidate genes in which *Helitrons* might play a regulatory role, an interesting hypothesis that can be determined through cutting-edge technologies such CRISPR methods to delete insertions in site-specific manner (75).

***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 preference for *Opuntia sp*., 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 (76), 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.

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 polymorphic mutations in ORs and OBPs have been suggested as factor of the modulation in odor sensitivity in *Drosophila* (67,77), 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 (78). Another study has identified 18 ORs differentially expressed between heads of *D. m. wrigleyi* and *D. m. mojavensis* subspecies, demonstrating that modifications in the sensory transcriptome have evolved between these

species, as previously reported for the same tissue (79). 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. *Opuntia sp.* produces less benzadelhyde volatiles than the columnar cacti *S. thurberi*, but both cacti are lethal to *D. melanogaster* after 48 hours (76). Therefore, the adaptation to use *Opuntia sp.* must have a constitutive gene expression of detoxification pathways in *D. m. wrigleyi*. Similarly with our observations in the head tissue, these results indicate that larvae tissue exhibits subspecies-specific expression, 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* cluster have higher divergence compared to *mojavensis* cluster, and lower between *D. buzzatii* and *D. koepferae*, following the phylogenetic relationship between the species. Although we have species with preference for *Opuntia sp.* and for columnar cacti in both clusters, *D. koepferae* and *D. buzzatii* are adapted to different hosts for a longer time than those of the *mojavensis*

cluster. In addition, a preliminary study demonstrated that *D. buzzatii* have a high divergence on larvae gene expression compared to *D. koepferae* (80), reinforcing the variance that we have observed. Indeed, differences in larvae behavior and survival has been observed when *Opuntia*-related species are reared on columnar cacti (81,82). 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% (83). 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 (80). 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 (84,85). 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 (76). Thus, any cactophilic species is expected to have evolved a certain level of detoxification ability, which explains our observations. In addition, the high frequency of common HLAU genes associated with

detoxification on columnar cacti dweller species is expected and already demonstrated in previous works.

***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 novelties (24). These chimeras have also been demonstrated to be active in a tissue-specific manner (86). 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 genomic and transcriptomic data to analyze the extent of genetic variability derived from chimeric transcripts in cactophilic species, as well as HLAU genes producing chimeric transcripts.

A preliminary transcriptome-wide study with *D. melanogaster* revealed that ~1.12% of the genes produce chimeric transcripts (24). Here we observed ~1.31% on average between head and larvae tissues. The slightly 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 (87). Despite this difference, the proportion of TE-exonized, TE-terminated, and TE-initiated transcripts follows similar values between them, in which TE-exonized have the higher frequency with ~70% of all chimeras (24,86). In our analysis, TE-initiated transcripts represent chimeric transcripts in which the TSS of the mRNA is located within the TE insertion (24). *Helitrons* were the most frequent TEs falling in the category in four of the seven species. They 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 (70). 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 (88). 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 (73). 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 (89). 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* in six transcriptomes 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* (40). Another study with *A. thaliana* revealed by poly(A)-tag sequencing data that ~48% of all TE-derived PASs are from *Helitrons* (90). Although there is vast documentation of TEs acting as alternative PASs (91), 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***

A comparative analysis between generalist and specialist *Drosophila* species revealed that niche amplitude is not likely to play a role in TE dynamics in the genomes (92). Nevertheless, several studies have proposed that the successful adaptation to new environments in invasive species might be associated with TE activity, despite reduced genetic diversity caused by founder effects (93,94). 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 transcriptomic variability in seven cactophilic species with different hosts preferences. We observed that species *D. arizonae,* which does not have preferential usage between *Opuntia sp.* and columnar cacti, are the ones with the highest production of unique HLAU chimeras in both head and larvae. This result demonstrates a possible relationship of broader niche and TE-driven adaptation.

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. 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 (95), non-stop decay (96), and non-sense-mediated RNA (97). Our study provides a set of HLAU genes producing chimeric transcripts, in which part of them might have phenotypic impact and ultimately association with host preference. However, it is fundamental to perform comparative phenotypic studies of host adaptation cues with mutant strains for the TEs interacting with HLAU genes identified in this study. The application of techniques of RNAi aiming chimeric transcripts, or deletion of TE insertions producing chimeras with CRISPR (75) will be able to confirm the predictions of our results.

***Transposable elements driving host shift***

Here, we demonstrated that HLAU genes are enriched by TEs in their regulatory regions, and produce different chimeric transcripts across species with different primary hosts. Host specificity is a key mechanism of reproductive isolation in plant pathogens, and is regulated by the repertoire of effector genes within each pathogen. Effector proteins modify host physiology, facilitating colonization by specific pathogens (98). In *Phytophthora sp*., the host specificity is partly regulated by RXLR class effectors that facilitate host exploitation. Notably, synthetic chimeras of a short interspersed element (SINE) linked to an

effector gene in *P. infestans* induce silencing of effector genes (99). This silencing likely occurs naturally, as transcriptional inactivation of effectors is known to occur, and over half of RXLR effectors in the *P. infestans* genome are located in TE-rich regions (100). Consequently, TEs inserted near these genes may have influenced HS in *P. infestans*. Our findings demonstrate multiple interactions between genes and TEs, highlighting the contribution of TEs to the genome evolution of cactophilic species. Through the TE-cooption process, we postulate that TEs can play a role in the HS in these species. Although many TEs located on promoters of HLAU genes can have low frequency in natural populations, or chimeric transcripts identified in this work can be a result of ectopic expression instead of adaptive, our results demonstrate the first transcriptome-wide evidence of TE co-option in cactophilic species. The genes producing chimeras with TEs must be further studied, since the lack of gene-phenotype association between HLAU genes is not fully reported, preventing us to pinpoint specific relation between chimeras and possibly reproductive isolation.

The existence of reproductive barriers between populations that undergone a host shift is the key process to ecological speciation. In our model, with species from the *buzzatii* and *mojavensis* clusters, reproductive isolation has been reported between *D. m. sonorensis* and *D. m. baja* (26,101,102). The premating isolation between them occurs due to the difference on cuticular hydrocarbons, which participate in the signaling pathway of mate recognification (103). Importantly, artificially shifting the cactus-host was the main cause of changes on the hydrocarbon profiles. Thus, the source of reproductive barriers are dependent of the cacti species on which the flies feed and breed (104,105). The close relationship between HLAU genes to the HS process, and the observed contribution of TEs to the evolution of these genes, supports our proposal that TEs may play a role in the HS of cactophilic species. Hence, such process rises the hypothesis that TEs can play a role in HS. Hence, such process rises the hypothesis that TEs could drive reproductive isolation via adaptation to different environments, ultimately driving ecological speciation.

## Conclusions

TEs have been shown to contribute to genome evolution of cactophilic *Drosophila* species. Their enrichment in promoter regions of HLAU gene families, along with potential presence of TFBSs within their sequences highlights their potential role in the evolution of adjacent genes. Additionally, their frequent exonization reveals a mechanism by which TEs increase genetic diversity between populations, potentially contributing to host shifts. Taken together, our combined genomic and transcriptomic approaches provide new insights regarding TE-related mechanisms that promote genome evolution between species and subspecies with different primary cactus hosts.

## 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 Center n°: 15081-1271.17]. *D. buzzatii* (Bu28) and *D. koepferae* (Ko2) stocks correspond to two laboratory lines used in previous works (106,107). Both stocks were maintained by brother-sister mating for more than a decade and then kept by mass culturing.

We used the same following protocols of DNA extraction and genome sequencing to all samples. DNA was extracted from 10 males and 10 females from each species using the Qiagen DNeasy Blood&Tissue kit. Then, we evaluated the genomic DNA amount and quality 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). We performed end

repair and dA-tailing 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* (108) was used with default parameters, except –plasmids. All contigs were aligned with *minimap2* v2.16 (109), with –x map-ont option. The alignment was used to perform the polishing of contigs with four rounds of *RACON* v1.3.2 (110), 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 (111) and corrected with *samtools* v1.9.0. (112), function *faidx*; and *Gepard* v1.4.0(113) for determination of breaking points. The super scaffolding of all corrected assemblies was performed with *RaGOO* v.1.1 (114), 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 (115) with lineage *diptera_odb10* (-*l* parameter). The gene annotations for *D. mojavensis* subspecies and *D. arizonae* were performed with the software *Liftoff v.1.6.3* (116), with the reference genome of *D. mojavensis* (GCA_018153725.1) (117). 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. We removed from our gene annotation HLAU genes flagged as "low quality protein" and "unestablished gene function" in the reference *D. mojavensis* genome.

*D. buzzatii* and *D. koepferae* genomes were annotated with a de novo strategy with the software *BRAKER3 v.3.0.8* (118). To optimize the prediction of gene structures and repertoire of isoforms, we used a mix of RNA-seq data from different tissues publicly available and produced in this work (Sup. Table 11), in addition to protein sequences from *D. mojavensis* as reference. Ortholog genes between *D. buzzatii, D. koepferae*, *D. mojavensis* and *D. arizonae* were obtained with *Orthofinder v.2.5.4* (119) using *DIAMOND v.0.9.14* (120) as search engine and default parameters. One-to-one orthology were firstly obtained directly from orthofinder, and remaining proteins without an ortholog pair were analyzed with *blastp v.2.5.0+* (121), for which we considered as one-to-one proteins only those with a single homologous with parameters *-qcov_hsp_perc 95* and identity >= 90%. Since *Orthofinder* performs its search with protein sequences, any non-coding RNA gene has been identified. Then, we used *Liftoff v.1.6.3* (122) with all non-coding RNA genes from *D. mojavensis* as reference to maximize our homology-based gene annotation.

***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 (123). For the other Nanopore genomes sequenced in this work, we performed a polishing with Illumina paired-end reads. We used the software *NextPolish* (124) to fix base errors in the assembled genomes. The four *D. mojavensis* subspecies were polished with genomic Illumina reads sequenced previously (125), whereas for *D. koepferae* and *D. buzzatii* we used publicly available genomic Illumina reads (126) and RNA-seq reads (48), 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 (48). All genomes were corrected using two

rounds of polishing, as recommended by the authors (124). Then, we extracted the longest CDS sequences for all genes combining AGAT (127) *agat_extract_sequences.pl* parameters *-t mrna*, *samtools faidx* (128), and *seqtk* (https://github.com/lh3/seqtk). Posteriorly, we 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* (129). 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* (130), 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 analysis to detect signatures of selection analysis for each gene individually with *CODEML* from *PAML v.4.10.7* (131), 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-square 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*.

We used the software *HYPHY* (132) to test relaxed selection with the same phylogeny framework used for branch-site analysis, using the RELAX model (50). The phylogenetic tree had its branches corresponding to columnar cacti dwellers as foreground, and *Opuntia sp.* as

background. We used the parameters *--test Foreground* and *--model Minimal*. Multiple testing correction was performed with False discovery rate (FDR). Genes with FDR < 0.05 and k > 1 were considered under intensified selection.

***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 adaptors and low quality reads were removed for downstream analysis with *Trimommatic v.0.39* (133), 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* (134) to build de novo TE libraries with *RepeatModeler2 v.2.0.4* (135), which uses RECON *v.1.08* (136) and *RepeatScout v.1.0.6* (137). 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) (138). 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* (139), and the alignment is trimmed with *trimAl v.1.4* (140), 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* (141). 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 (93).

TE consensus that had passed in our filtering thresholds were then submitted to the classification analysis. We developed a pipeline to perform similarity analysis of non-reference consensus with a database provided by the user (https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-Drosophila-Species). The pipeline identifies similarity analysis between the non-ref. consensus and a database provided. All consensuses were masked with *RepeatMasker v.4.1.4* (142), 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 (143) 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, we used all obtained IDs with *seqtk v.1.3* (https://github.com/lh3/seqtk) to subset the complete fasta file generated by *famdb*.

We used the polished TE libraries constructed to the seven cactophilic *Drosophila* genomes to perform TE annotation with *RepeatMasker v.4.1.4* (142), parameters *-cutoff 250*, *-s*, and *-norna*. Then, we have used *RepeatCraft* (144) 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 (145). Finally, we performed a last filtering step in the TE annotation to remove insertions comprised by simple sequence repeat (SSR). We used *TRF v.4.09* (146) 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 (24,147).

***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* (148), parameter *randomization = resampleRegions*, with 2,000 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. We also selected zinc-finger proteins and serine/threonine kinases as two non-HLAU genes families to compare our findings. Taking

into account that TEs might harbor regulatory and/or coding motifs in both strands (65), 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* (149) to identify the number of TE insertions located in each genomic bin. Finally, we calculated the 3$^{rd}$ quartile with the frequency of TEs in each region and defined as TE-rich genomic regions those with more TE insertions than the 3$^{rd}$ 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 3$^{rd}$ quartile).

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

We assessed the identification of transcriptions factor binding sites (TFBSs) with *FIMO* (150), from the *MEME v. 5.0.5* suite tools (151). We extracted the sequences of all TE insertions located 2kb upstream of ORs with *bedtools v. 2.30.0*, function *getfasta*, using parameter *-s* (149). We obtained the TFBS motifs described for *D. melanogaster* from Jaspar database (152), for the transcription factors (TFs) acj6, zf30c, fer1, onecut, and xbp1, which have been demonstrated to activate ORs (41). To control our findings with OR genes, we used IR as a HLAU gene family member, and Serine/Threonine Kinases as a non-HLAU gene family. We performed The same process of TFBSs identification performed for ORs on these two gene families. We only considered TEs with TFBSs those with p-value < 0.01 and score > 250.

### *Differential expression of genes*

We aligned the RNA-seq reads with *STAR v.2.7.3* (153), 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* (154), 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 species of the *buzzatii* and *mojavensis* clusters 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 species of the *buzzatii* and *mojavensis* clusters 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 phylogenetic distance between species of the *buzzatii* and *mojavensis* clusters, we performed the differential expression analysis separately for each group. We only considered genes with adjusted *p-value* < 0.05 and *log2 fold-change* > |1| 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 >= 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. We carried out the test for all pairwise analysis between cactophilic species with preference for *Opuntia sp* and columnar cacti. The proportional differences between location, acceptance and usage genes compared to non-HLAU genes (Ser/Thr Kinases and zinc-finger gene families) were assessed with Fisher's exact test, with contingency table containing average number of genes per each group: 1) non-HLAU differentially expressed genes; 2) non-HLAU non-differentially expressed genes; 3) HLAU differentially expressed genes; and 4) HLAU non-differentially expressed genes. *Fisher's* Exact test < 0.05 were considered as significant.

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

We detected gene-TE chimeric transcripts with *ChimeraTE v.1.1.1* Mode 1 (24). 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.

## Declarations

### *Availability of data and materials*

The quality control data from both ONT and RNA-seq data, the genome assemblies and their respective gene and TE annotation (libraries and gtf files), and the data used to produce phylogenetic are available in the Zenodo repository, doi: https://zenodo.org/records/13117512. The code to reproduce the analyses are available on Github repository: https://github.com/OliveiraDS-hub/Pipelines-Cactophilic-Drosophila-Species.

### *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] and CAPES – PrInt [88887.716810/2022-00] to D. S. O.; by Ministerio de Ciencia e Innovación (Spain) [grant number PID2021-127107NB-I00], and Generalitat de Catalunya (Spain) [grant number 2021 SGR 00526] to M. P. G. G.; and by UdI-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. Whiteman NK, Pierce NE. Delicious poison: genetics of Drosophila host plant preference. Trends in Ecology & Evolution. 2008 Sep 1;23(9):473–8.
2. 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 May;71(5):1126–37.
3. Rundle HD, Nosil P. Ecological speciation. Ecology Letters. 2005;8(3):336–52.
4. Howard DJ, Berlocher SH. Endless Forms: Species and Speciation. Oxford University Press; 1998. 486 p.
5. O'Grady PM, Markow TA. Rapid morphological, behavioral, and ecological evolution in Drosophila: comparisons between the endemic Hawaiian Drosophila and the cactophilic repleta species group. In: Singh RS, Xu J, Kulathinal RJ, editors. Rapidly Evolving Genes and Genetic Systems [Internet]. Oxford University Press; 2012 [cited 2024 Jul 1]. p. 0. Available from: https://doi.org/10.1093/acprof:oso/9780199642274.003.0018
6. 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 Sep;64(3):533–44.
7. 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.
8. Gomez GA. Transpecific Polymorphisms in an Inversion Linked Esterase Locus in Drosophila buzzatii. Molecular Biology and Evolution. 2003 Mar 1;20(3):410–23.
9. 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 Jan 1;65:319–26.
10. Newby BD, Etges WJ. Host Preference Among Populations of Drosophila mojavensis (Diptera: Drosophilidae) that Use Different Host Cacti. Journal of Insect Behavior. 1998 Sep 1;11(5):691–712.
11. 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 Jan 7;110(1):80–91.
12. Markow TA. Host use and host shifts in Drosophila. Current Opinion in Insect Science. 2019 Feb;31:139–45.
13. Montell C. A taste of the Drosophila gustatory receptors. Current Opinion in Neurobiology. 2009 Aug;19(4):345–53.
14. Ni L. The Structure and Function of Ionotropic Receptors in Drosophila. Front Mol Neurosci. 2021 Feb 1;13:638839.
15. Chen Y, Amrein H. Ionotropic Receptors Mediate Drosophila Oviposition Preference through Sour Gustatory Receptor Neurons. Current Biology. 2017 Sep;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 Aug 30;12:253.
17. 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 Jan 7;110(1):80–91.
18. 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. Betran E, editor. Genome Biology and Evolution. 2020 Aug 1;12(8):1407–18.
19. Fedoroff NV. Transposable Elements, Epigenetics, and Genome Evolution. Science. 2012 Nov 9;338(6108):758–67.
20. Capy P. Taming, Domestication and Exaptation: Trajectories of Transposable Elements in Genomes. Cells. 2021 Dec 20;10(12):3590.
21. Drongitis D, Aniello F, Fucci L, Donizetti A. Roles of Transposable Elements in the Different Layers of Gene Expression Regulation. IJMS. 2019 Nov 15;20(22):5755.
22. 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 Oct 20;7(10):e1002337.
23. Lipatov M, Lenkov K, Petrov DA, Bergman CM. Paucity of chimeric gene-transposable element transcripts in the Drosophila melanogaster genome. BMC Biol. 2005 Dec;3(1):24.
24. 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 Oct 13;51(18):9764–84.
25. 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 Feb 1;178(2):1073–83.
26. 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 Apr 1;43(15–16):923–38.
27. 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.
28. 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 Aug 31;283(1837):20161562.
29. 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.
30. 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 Dec;23(1):515.
31. 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.
32. 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 [Internet]. bioRxiv; 2022 [cited 2023 Nov 2]. p. 2022.10.16.512445. Available from: https://www.biorxiv.org/content/10.1101/2022.10.16.512445v1

33. 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 Sep;64(3):533–44.
34. 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 Oct 1;6(10):3097–105.
35. 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 Jul;21(13):3293–307.
36. Guio L, Vieira C, González J. Stress affects the epigenetic marks added by natural transposable element insertions in Drosophila melanogaster. Sci Rep. 2018 Aug 15;8(1):12197.
37. 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 Sep 27;297(5590):2253–6.
38. 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 Jul;32(7):1408–23.
39. 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 Feb 6;104(6):1895–900.
40. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. Genome Biology and Evolution. 2014 Oct;6(10):2595–610.
41. 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.
42. Fogleman JC, Danielson PB. Chemical Interactions in the Cactus-Microorganism-Drosophila Model System of the Sonoran Desert1. American Zoologist. 2001 Aug 1;41(4):877–89.
43. 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 Jan;71(3):413–7.
44. Markow TA. Host use and host shifts in Drosophila. Current Opinion in Insect Science. 2019 Feb;31:139–45.
45. McBride CS. Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia. Proceedings of the National Academy of Sciences. 2007 Mar 20;104(12):4996–5001.
46. McBride CS, Arguello JR. Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily. Genetics. 2007 Nov 1;177(3):1395–416.
47. Allan CW, Matzkin LM. Genomic analysis of the four ecologically distinct cactus host populations of Drosophila mojavensis. BMC Genomics. 2019 Oct 12;20(1):732.
48. 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 Jan;7(1):349–66.
49. Kristensen DM, Wolf YI, Mushegian AR, Koonin EV. Computational methods for Gene Orthology inference. Briefings in Bioinformatics. 2011 Sep 1;12(5):379–91.
50. Wertheim JO, Murrell B, Smith MD, Kosakovsky Pond SL, Scheffler K. RELAX: Detecting Relaxed Selection in a Phylogenetic Framework. Molecular Biology and Evolution. 2015 Mar 1;32(3):820–32.
51. Lahti DC, Johnson NA, Ajie BC, Otto SP, Hendry AP, Blumstein DT, et al. Relaxed selection

in the wild. Trends in Ecology & Evolution. 2009 Sep 1;24(9):487–96.
52. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Molecular Biology and Evolution. 1994 Sep 1;11(5):725–36.
53. 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 Jun 13;7(1):ncomms11855.
54. 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 Jan 16;20(1):52.
55. Faulkner GJ, Kimura Y, Daub CO, Wani S, Plessy C, Irvine KM, et al. The regulated retrotransposon transcriptome of mammalian cells. Nat Genet. 2009 May;41(5):563–71.
56. 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 Apr 25;9(4):e1003470.
57. 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 Jan;29(1):239–47.
58. 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 Apr 7;117(14):7905–16.
59. 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.
60. 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 Jan;148(1–2):335–48.
61. Wang J, Lunyak VV, Jordan IK. Genome-wide prediction and analysis of human chromatin boundary elements. Nucleic Acids Research. 2012 Jan;40(2):511–29.
62. 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 Dec;24(12):1963–76.
63. 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 Mar 4;40(2):157–89.
64. 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 Jun;534(7605):102–5.
65. Mateo L, Ullastres A, González J. A Transposable Element Insertion Confers Xenobiotic Resistance in Drosophila. Feschotte C, editor. PLoS Genet. 2014 Aug 14;10(8):e1004560.
66. 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 Jan;16(1):101–9.
67. 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. Noor MAF, editor. PLoS Biol. 2007 Apr 24;5(5):e118.
68. McKenzie SK, Kronauer DJC. The genomic architecture and molecular evolution of ant odorant receptors. Genome Res. 2018 Jan 11;28(11):1757–65.
69. 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 Jan 15;63(6):1577–87.
70. Thomas J, Phillips CD, Baker RJ, Pritham EJ. Rolling-Circle Transposons Catalyze Genomic Innovation in a Mammalian Lineage. Genome Biology and Evolution. 2014 Oct;6(10):2595–610.
71. 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 Jan;57(1):115–27.
72. 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 Jun;23(3):396–405.
73. 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 Dec 2;8:e50541.
74. 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 Mar;537(1):93–9.
75. Merenciano M, Aguilera L, González J. Two-step CRISPR-Cas9 protocol for transposable element deletion in *D. melanogaster* natural populations. STAR Protocols. 2023 Sep 15;4(3):102501.
76. Wright CR, William S. Characterization of Volatiles of Necrotic Stenocereus thurberi and Opuntia littoralis and Toxicity and Olfactory Preference of Drosophila melanogster, D. mojavensis wrigleyi, and D. mojavensis sonorensis to Necrotic Cactus Volatiles. Natural Product Communications. 2014;9(6):1934578X1400900833.
77. 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 Dec;17(1):648.
78. Mansourian S, Stensmyr MC. The chemical ecology of the fly. Current Opinion in Neurobiology. 2015 Oct;34:95–102.
79. 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 Aug 31;283(1837):20161562.
80. 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 Dec;23(1):515.
81. Ruiz A, Heed WB. Host-Plant Specificity in the Cactophilic Drosophila mulleri Species Complex. The Journal of Animal Ecology. 1988 Feb;57(1):237.
82. 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 Jan 7;110(1):58–67.
83. Matzkin LM. Population transcriptomics of cactus host shifts in *Drosophila mojavensis*: Population Transcriptomics of *D. mojavensis*. Molecular Ecology. 2012 May;21(10):2428–39.
84. Feyereisen R. Insect CYP Genes and P450 Enzymes. In: Insect Molecular Biology and Biochemistry [Internet]. Elsevier; 2012 [cited 2022 Sep 28]. p. 236–316. Available from: https://linkinghub.elsevier.com/retrieve/pii/B978012384747810008X
85. Tijet N, Helvig C, Feyereisen R. The cytochrome P450 gene superfamily in Drosophila melanogaster : Annotation, intron-exon organization and phylogeny. Gene. 2001 Jan;262(1–2):189–98.
86. 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 Jan

9;33(9):1541–53.
87. 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 Jun 1;19(6):926–37.
88. 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 Sep 21;6(1):33785.
89. Garrigues JM, Tsu BV, Daugherty MD, Pasquinelli AE. Diversification of the Caenorhabditis heat shock response by Helitron transposable elements. Landry CR, Wittkopp PJ, editors. eLife. 2019 Dec 11;8:e51139.
90. 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 Jun 15;6(1):28060.
91. 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 Oct 1;36(17):5581–90.
92. 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 Sep 1;27(3):203–19.
93. 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, editor. Molecular Biology and Evolution. 2021 Sep 27;38(10):4252–67.
94. 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 Sep 1;13(9):evab208.
95. Harigaya Y, Parker R. No-go decay: a quality control mechanism for RNA in translation. WIREs RNA. 2010 Jul;1(1):132–41.
96. Vasudevan S, Peltz SW, Wilusz CJ. Non-stop decay—a new mRNA surveillance pathway. BioEssays. 2002;24(9):785–8.
97. Hug N, Longman D, Cáceres JF. Mechanism and regulation of the nonsense-mediated decay pathway. Nucleic Acids Res. 2016 Feb 29;44(4):1483–95.
98. Beckerson WC, Rodríguez de la Vega RC, Hartmann FE, Duhamel M, Giraud T, Perlin MH. Cause and Effectors: Whole-Genome Comparisons Reveal Shared but Rapidly Evolving Effector Sets among Host-Specific Plant-Castrating Fungi. mBio. 2019 Nov 5;10(6):10.1128/mbio.02391-19.
99. Whisson SC, Vetukuri RR, Avrova AO, Dixelius C. Can silencing of transposons contribute to variation in effector gene expression in Phytophthora infestans? Mobile Genetic Elements. 2012 Mar 1;2(2):110–4.
100. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, et al. Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. Science. 2010 Dec 10;330(6010):1540–3.
101. Zouros E, d'Entremont CJ. Sexual Isolation Among Populations of Drosophila mojavensis: Response to Pressure from a Related Species. Evolution. 1980;34(3):421–30.
102. Markow TA. Sexual Isolation among Populations of Drosophila mojavensis. Evolution. 1991;45(6):1525–9.
103. Howard RW, Blomquist GJ. ECOLOGICAL, BEHAVIORAL, AND BIOCHEMICAL ASPECTS OF INSECT HYDROCARBONS*. Annual Review of Entomology. 2005 Jan 1;50(Volume 50, 2005):371–93.
104. Stennett MD, Etges WJ. Premating Isolation Is Determined by Larval Rearing Substrates in

CactophilicDrosophila mojavensis. III. Epicuticular Hydrocarbon Variation Is Determined by Use of Different Host Plants inDrosophila mojavensis andDrosophila arizonae. J Chem Ecol. 1997 Dec 1;23(12):2803–24.
105. Havens JA, Etges WJ. Premating isolation is determined by larval rearing substrates in cactophilic Drosophila mojavensis. IX. Host plant and population specific epicuticular hydrocarbon expression influences mate choice and sexual selection. Journal of Evolutionary Biology. 2013 Mar 1;26(3):562–76.
106. 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 Mar;8(3):556–61.
107. 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 Feb 4;14(2):evac024.
108. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.
109. Li H. Minimap2: pairwise alignment for nucleotide sequences. Birol I, editor. Bioinformatics. 2018 Sep 15;34(18):3094–100.
110. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:13033997 [q-bio] [Internet]. 2013 May 26 [cited 2021 Nov 5]; Available from: http://arxiv.org/abs/1303.3997
111. Cabanettes F, Klopp C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. PeerJ. 2018 Jun 4;6:e4958.
112. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
113. Krumsiek J, Arnold R, Rattei T. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. Bioinformatics. 2007 Apr 15;23(8):1026–8.
114. 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 Dec;20(1):224.
115. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 Oct 1;31(19):3210–2.
116. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A, editor. Bioinformatics. 2021 Jul 19;37(12):1639–43.
117. 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.
118. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully automated genome annotation using RNA-seq and protein evidence with GeneMark-ETP, AUGUSTUS, and TSEBRA. Genome Res [Internet]. 2024 Jun 12 [cited 2024 Jul 15]; Available from: https://genome.cshlp.org/content/early/2024/05/28/gr.278090.123
119. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 Dec;20(1):238.
120. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015 Jan;12(1):59–60.
121. Altschul S. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997 Sep 1;25(17):3389–402.
122. Shumate A, Salzberg SL. Liftoff: accurate mapping of gene annotations. Valencia A, editor. Bioinformatics. 2021 Jul 19;37(12):1639–43.
123. 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 Jan 7;110(1):118–23.
124. 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 Apr 1;36(7):2253–5.
125. 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 May 10;11(1):9844.
126. 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 Jan 1;178:107653.
127. Dainat J. AGAT: Another Gff Analysis Toolkit to handle annotations in any GTF/GFF format. Zenodo [Internet]. 2023 [cited 2023 Nov 9]; Available from: https://zenodo.org/records/8178877
128. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.
129. Suyama M, Torrents D, Bork P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Research. 2006 Jul 1;34(suppl_2):W609–12.
130. 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 May 1;37(5):1530–4.
131. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Molecular Biology and Evolution. 2007 Apr 18;24(8):1586–91.
132. Kosakovsky Pond SL, Poon AFY, Velazquez R, Weaver S, Hepler NL, Murrell B, et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. Molecular Biology and Evolution. 2020 Jan 1;37(1):295–9.
133. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014 Aug 1;30(15):2114–20.
134. Baril TJ, Galbraith JD, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline [Internet]. bioRxiv; 2024 [cited 2024 Feb 20]. p. 2022.06.30.498289. Available from: https://www.biorxiv.org/content/10.1101/2022.06.30.498289v3
135. 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 Apr 28;117(17):9451–7.
136. Bao Z, Eddy SR. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Res. 2002 Aug 1;12(8):1269–76.
137. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005 Jun 1;21(Suppl 1):i351–8.
138. Platt RN, Blanco-Berdugo L, Ray DA. Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biol Evol. 2016 Feb;8(2):403–10.
139. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution. 2013 Apr 1;30(4):772–80.
140. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009 Aug 1;25(15):1972–3.
141. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 16(6):276–7.
142. SMIT, Arian FA. RepeatMasker Open 3.0. 2004;
143. 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 Jan 4;44(D1):D81–9.

144. Wong WY, Simakov O. RepeatCraft: a meta-pipeline for repetitive element de-fragmentation and annotation. Hancock J, editor. Bioinformatics. 2019 Mar 15;35(6):1051–2.
145. 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 Dec;8(12):973–82.
146. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research. 1999 Jan 1;27(2):573–80.
147. Flutre T, Permal E, Quesneville H. Transposable Element Annotation in Completely Sequenced Eukaryote Genomes. In: Grandbastien MA, Casacuberta JM, editors. Plant Transposable Elements [Internet]. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012 [cited 2023 Feb 27]. p. 17–39. (Topics in Current Genetics; vol. 24). Available from: https://link.springer.com/10.1007/978-3-642-31842-9_2
148. 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 Sep 30;btv562.
149. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
150. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011 Apr 1;27(7):1017–8.
151. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Research. 2015 Jul 1;43(W1):W39–49.
152. 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 Nov 8;gkz1001.
153. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan;29(1):15–21.
154. Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics. 2015 Jan 15;31(2):166–9.
155. 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.