Genetics and Molecular Biology, 46, 3(suppl 1), e20230143 (2023)

DOI: https://doi.org/10.1590/1678-4685-GMB-2023-0143

Research Article
60 years of the PPGBM UFRGS – Special Issue

Horizontal transfer and the widespread presence of *Galileo* transposons in Drosophilidae (Insecta: Diptera)

Henrique R.M. Antoniolli[1], Sebastián Pita[2], Maríndia Deprá[1] and Vera L.S. Valente[1]

[1]*Universidade Federal do Rio Grande do Sul (UFRGS), Laboratório de Drosophila, Programa de Pós-Graduação em Genética e Biologia Molecular, Porto Alegre, RS, Brazil.*
[2]*Universidad de la República (UdelaR), Facultad de Ciencias, Sección Genética Evolutiva, Montevideo, Uruguay.*

Abstract

Galileo is a transposon notoriously involved with inversions in *Drosophila buzzatii* by ectopic recombination. Although widespread in *Drosophila*, little is known about this transposon in other lineages of Drosophilidae. Here, the abundance of the canonical *Galileo* and its evolutionary history in Drosophilidae genomes was estimated and reconstructed across genera within its two subfamilies. Sequences of this transposon were masked in these genomes and their transposase sequences were recovered using BLASTn. Phylogenetic analyses were employed to reconstruct their evolutionary history and compare it to that of host genomes. *Galileo* was found in nearly all 163 species, however, only 37 harbored nearly complete transposase sequences. In the remaining, *Galileo* was found highly fragmented. Copies from related species were clustered, however horizontal transfer events were detected between the *melanogaster* and *montium* groups of *Drosophila*, and between the latter and the *Lordiphosa* genus. The similarity of sequences found in the *virilis* and *willistoni* groups of *Drosophila* was found to be a consequence of lineage sorting. Therefore, the evolution of *Galileo* is primarily marked by vertical transmission and long-term inactivation, mainly through the deletion of open reading frames. The latter has the potential to lead copies of this transposon to become miniature inverted-repeat transposable elements.

Keywords: DNA transposon, MITEs, *P* superfamily.

Received: May 08, 2023; Accepted: January 30, 2024.

Introduction

Transposable elements (TEs) belong to the repetitive fraction of genomes, and are linear sequences of DNA that have the ability to move within or between genomes (Wells and Feschotte, 2020). Classifications divide these sequences firstly into two classes, based on the intermediate molecule in their transposition process (Finnegan, 1989). Class I is composed of retrotransposons as their mobilization involves the synthesis of an RNA molecule, which are retrotranscribed into DNA and then inserted elsewhere in the genome (Wicker *et al*., 2007). On the other hand, the majority of Class II elements – or DNA transposons – are directly excised by their transposase (TPase), and then reinserted in another site in the genome (Wicker *et al*., 2007).

In addition, TEs can be either autonomous or nonautonomous (Wicker *et al*., 2007). The first are those that present their structures preserved, encoding all necessary enzymes to be transposed. The latter comprise defective TEs that no longer encode nor produce their own proteins, and move only if recognized by the enzymes of a closely related autonomous TE; such as the Miniature Inverted-repeat Transposable Elements (MITEs). MITEs are non-autonomous TEs, derived from autonomous Class II transposons, and present a few structural characteristics: (i) small size, ranging from 50 to 500 base pairs (bp); (ii) AT-rich sequences; and (iii) a lack of a functional TPase (Deprá *et al*., 2012; Fattash *et al*., 2013).

Transposable elements are often referred to as "parasites" (Colonna Romano and Fanti, 2022), given their ability to invade new genomes and increase their copy number (Loreto *et al*., 2008). Horizontal transposon transfer (HTT) is the phenomenon in which a given TE "jumps" to the genome of a non-closely related species, i.e., sexually isolated organisms (Panaud, 2016). The role of HTT in shaping diversity as an endogenous source of evolution is widely recognized (Pace *et al*., 2008; Gilbert and Feschotte, 2018; Carvalho *et al*., 2023), and its frequency is much higher than previously thought (Schaack *et al*., 2010; Panaud, 2016; Peccoud *et al*., 2017; Melo and Wallau, 2020).

In this sense, several evolutionary events have been proposed as a direct consequence of TEs mobilization and/or recombination. For instance, in several taxa the variation and evolution of genome size are directly related to the amplification or contraction in TEs copy number (Canapa *et al*., 2015; Antoniolli *et al*., 2023). Nucleotide polymorphisms are also frequently produced after transposition events (Bourque *et al*., 2018). Transposable elements are also known to be related to changes in gene expression, either by silencing or enhancing them (Finnegan, 1989), and chromosomal rearrangements – i.e., deletions, duplications, translocations and inversions by ectopic recombination (Kidwell and Lisch,

Send correspondence to Maríndia Deprá. Universidade Federal do Rio Grande do Sul (UFRGS), Laboratório de Drosophila, Avenida Bento Gonçalves, 9500, Agronomia, 91509-900, Porto Alegre, RS, Brazil. E-mail: marindiadepra@gmail.com.

1997). In the latter, distant loci in a genome carry highly similar TE copies, which allows homologous recombination to occur (see review in Bourque *et al*., 2018), thus resulting in a drastic modification in the chromosome architecture (Ren *et al*., 2018). Documented cases of a TE as a mediator of ectopic recombination include the families of retrotransposons *Bel-Pao*, *Doc*, *I* element and *roo*, as well as the transposons *foldback*, *Galileo* and *hobo* (Lim and Simmons, 1994; Delprat *et al*., 2009).

Galileo is a family of Class II transposons, and encodes its own TPase flanked by terminal inverted repeats (TIRs). Initially described as a *foldback*-like element, its TIRs and THAP domains exhibit similarities with those of the *P* element, leading to the classification of *Galileo* within the *P* superfamily (Marzo *et al*., 2008). However, unlike the *P* element, *Galileo* does not present introns (Marzo *et al*., 2008). *Galileo* was discovered by Cáceres *et al*. (1999) due to its association with the breakpoints of the *2j* inversion in wild specimens of *Drosophila buzzatii*. In fact, *Galileo* is the only TE known to induce chromosomal rearrangements in natural populations of *Drosophila* (Marzo *et al*., 2008), as most others have been observed in laboratory populations (Lim and Simmons, 1994). Besides the *2j* inversion, *Galileo* was involved with two other rearrangements described in *D. buzzatii* (Casals *et al*., 2003; Delprat *et al*., 2009). This makes this transposon as one of the most well-documented examples of a natural TE-induced chromosomal rearrangement.

Studies have shown the widespread presence of *Galileo* in the *Drosophila* genus (Marzo *et al*., 2008; Acurio, 2015). The main focus of the present study was to characterize the evolutionary history of the *Galileo* family and evaluate its main transmission mode in Drosophilidae. This transposon was masked in genome assemblies of 163 species available at online databases, and TPase sequences found were employed for reconstructing a phylogeny and testing putative cases of HTT.

Material and Methods

Masking *Galileo* in the genome assemblies

Representative genome assemblies of 163 Drosophilidae species (see details on taxonomy and accession numbers in Table S1) were retrieved from GenBank (NCBI) with a Python package written by Blin (2021). These species belong to the *Chymomyza*, *Drosophila*, *Lordiphosa*, *Scaptodrosophila*, *Scaptomyza*, and *Zaprionus* genera of the Drosophilinae subfamily; and *Leucophenga* and *Phortica* of Steganinae subfamily (Table S1). BUSCO v.5 (Manni *et al*., 2021) was employed to assess the completeness of each assembly with the Diptera orthologous database.

The nucleotide sequence of seven *Galileo* copies characterized by Marzo *et al*. (2008) in *D. ananassae* (Dana*Galileo* – BK006363), *D. buzzatii* (Dbuz*Galileo* – EU334682 and EU334685), *D. mojavensis* (Dmoj*Galileo* – BK006357), *D. persimilis* (Dper*Galileo* – BK006361), *D. virilis* (Dvir*Galileo* – BK006359) and *D. willistoni* (Dwil*Galileo* – BK006360) were downloaded from GenBank, and used as queries in our workflow. Firstly, the queries were input as the repeat library in RepeatMasker (Smit *et al*., 2023) for masking canonical *Galileo* sequences in each genome assembly. The script 'One code to find them all' (Bailly-Bechet *et al*., 2014) was then employed to parse the output, recovering the nucleotide sequence of each identified copy in an assembly with at least 80% identity to its best query and a minimum length of 80 base pairs.

Phylogenetic analysis of *Galileo* potentially autonomous copies

The complete nucleotide sequence encoding the transposase (TPase) of six copies (Dana*Galileo*, Dbuz*Galileo*, Dmoj*Galileo*, Dper*Galileo*, Dvir*Galileo*, and Dwil*Galileo*) served as queries for local BLASTn searches in each FASTA file containing the *Galileo* copies of each genome. Hits with at least 80% identity and coverage of at least 70% for any of the queries were used in downstream analyses. Additionally, a *P* element from the genome of *Drosophila buzzatii* (GenBank accession No. KC690135) and two copies of the *1360* element (GenBank accession Nos. AF533772 and AY138841) were included in the nucleotide matrix as outgroups. This matrix was aligned with MACSE v2 (Ranwez *et al*., 2018) in two steps: (i) using the option *alignSequences*, which aligns nucleotide sequences based on their underlying codon structure, accounting for frameshifts and stop codons; (ii) the resulting alignment was edited with the option *exportAlignment*, replacing codons containing frameshifts and internal stop codons with "N" (e.g., TG! was replaced by NNN). The codon alignment was then processed with Gblocks (Castresana, 2000) to remove poorly aligned regions, allowing the presence of gaps.

The final codon alignment was translated to amino acids and used for a Bayesian phylogenetic inference (BI) analysis, performed in MrBayes 3.2.7 (Ronquist *et al*., 2012). The majority-rule consensus tree was built under the best amino acid substitution model, as estimated by ModelTest-NG (Darriba *et al*., 2020). Metropolis-coupled Markov chain Monte Carlo (MCMCMC) analysis was run with two parallel runs with four chains each for 1,000,000 generations, sampling every 100. Convergence was reached when the average standard deviation of split frequencies was below 1%. A burn-in of 25% was applied to the sampled trees before obtaining the consensus tree. The tree was visualized and edited in FigTree (Rambaut, 2018).

Analysis of abundance and repeat profile

Forward short-reads of high-throughput whole genome sequencing were downloaded from the Sequence Read Archive of NCBI (see SRA accession No. in Table S1) for those species with positive hits for the TPase queries. These were submitted to the RepeatProfiler pipeline (Negm *et al*., 2021), an analysis in which sequencing reads are mapped against queries to build coverage graphs, allowing to infer which regions of a given query have a higher or lower abundance.

Quality trim was performed with fastp (Chen *et al*., 2018), when reads had their adaptor removed while keeping only reads with no N base. The total reads were downsampled to 3 million, achieving near 1x coverage for all genomes (assuming a genome size mean of 200 megabases for species of the family Drosophilidae). In addition, five single-copy

genes were randomly selected in the Diptera orthologous genes dataset of BUSCO 5 (Manni *et al*., 2021) to normalize the results (Table S2). The six complete copies of *Galileo* used in BLASTn searches were used as queries (Dbuz*Galileo* – EU334685 was excluded because it was shorter than Dbuz*Galileo* – EU334682). RepeatProfiler (Negm *et al*., 2021) was executed with default parameters.

Inference of HTT events

Possible cases of HTT were determined based on incongruences between the phylogeny of host genomes and the phylogeny of *Galileo*. Validation of such cases was performed with the *vhica* R package (Wallau *et al*., 2015), implemented on the HTT-DB platform (Dotto *et al*., 2015). This method relies on discrepancies in the evolutionary rates of synonymous positions (dS), which considers codon usage bias (CUB), between nuclear genes (vertically transmitted) and transposable elements (TEs). Wallau *et al*. (2015) demonstrated that dS and CUB are correlated, and low values for both are indicative of inconsistencies with vertical transmission.

Sequences of single-copy orthologous genes were searched in the assemblies with positive hits of *Galileo* using BUSCO 5 (Manni *et al*., 2021) and the Diptera orthologous database. Nucleotide sequences of 30 randomly selected genes (see Table S3) were aligned based on codons using the ClustalW algorithm (Thompson *et al*., 1994) implemented in MEGA 11 (Tamura *et al*., 2021). These alignments were used to compare the dS-CUB between the host nuclear genome and *Galileo* sequences. A substitution rate of 0.016 per million years (Sharp and Li, 1989) was applied to estimate the time of divergence between *Galileo* sequences.

To provide an evolutionary context for the results, a phylogenetic tree of the 37 host genomes was reconstructed using the entire set of BUSCO genes shared among them. Their amino acid sequences were aligned with MUSCLE (Edgar, 2004) and refined with trimAl (Capella-Gutiérrez *et al*., 2009), implemented in a pipeline written by McGowan (2020). *Scaptodrosophila lebanonensis* was included in this analysis as an outgroup. Their phylogenetic relationships were reconstructed under maximum likelihood in IQ-TREE 2 (Minh *et al*., 2020), with the best substitution model selected based on AIC scores (flags *--m* and *--merit*). Branch supports were estimated by applying 1,000 replicates of ultrafast bootstrap.

Results

Search for canonical *Galileo* copies

Sequences of *Galileo* were masked in all analyzed genomes (Table S1), except for *D. ercepeae* and *D. nannoptera* – which belong to the *melanogaster* and *nannoptera* groups, respectively. Assemblies showed satisfactory levels of completeness, with the majority having more than 90% of single-copy orthologous genes (S). The exception was eight species, with S percentages ranging from 70% to 90% (see Table S4). In the second round of searches, conducted using local BLASTn with TPases as queries, 37 species yielded positive hits after the filtering process (Table S1). The positive results in the BLASTn search were limited to species within the *Drosophila* and *Lordiphosa* genera (Drosophilinae subfamily, Drosophilini tribe). All identified TPase sequences exhibited mutations, including stop codons, coding frame shifts, or both.

Phylogenetic analysis and abundance of *Galileo* sequences

The final sizes of nucleotide and amino acid alignments were 1,035 bp and 345 amino acids, respectively. The best amino acid substitution model was JTT+G4+F, based on the Akaike Information Criterion (AIC). Every copy of *Galileo* found in the genomes was placed in the same clade as its query. Major clades exhibited strong node support (PP > 0.95), with exceptions mainly observed among intraspecific sequences.

The query Dana*Galileo* recovered three clades: the first two (yellow, Figure S1) containing sequences found in genomes of the *melanogaster* group (in which *D. ananassae* is phylogenetically placed); and the third (orange clade, Figure S1) containing sequences found in species of the *montium* group, along with *Lordiphosa collinella* and *L. stackelbergi* (pink sequences, Figure S1). Dwil*Galileo* clustered homologous sequences found in the *willistoni* group (light pink sequences, Figure S1), along with its sister *saltans* group (blue sequences, Figure S1). On the other hand, the sequences of *Galileo* found by Dvir*Galileo* (green clade, Figure S1) in species of the *virilis* group formed a sister clade (PP = 1.0) to those of the *willistoni* and *saltans* groups. Finally, Dper*Galileo* recovered *Galileo* from species belonging to the *obscura* group (red clade, Figure S1), and Dmoj*Galileo* retrieved sequences in *D. mojavensis* (purple clade, Figure S1). The abundance of *Galileo* sequences in these species, as assessed by the coverage analysis in RepeatProfiler, showed that the TPase region had lower coverage than that of TIRs in all cases (Figures 1 and S2-S7).

Inference of HTT events

Two major incongruences were found between host species (Figure 2A) and *Galileo* phylogenies. The first is the similarity of elements found in *Lordiphosa collinella* and *Lordiphosa stackelbergi* with species of the *montium* group (Figure 2B). The second incongruence (Figure 2C) is the clade formed by *virilis* (*Drosophila* subgenus) and *willistoni* plus *saltans* groups (*Sophophora* subgenus). No signals of HTT events were detected (p-value > 0.05) between the species of the *virilis* group and the *willistoni* and *saltans* groups (Figure 2D). However, HTT was detected (p-value < 0.05) between *L. collinella* and *L. stackelbergi* and species of the *montium* group. Signals were also detected between the *melanogaster* and *montium* groups, both belonging to the *Sophophora* subgenus (Figure 2E). Estimates of divergence times (Table S5) span from ~679 thousand years ago (*D. auraria* x *L. stackelbergi*) to ~6 million years ago (*D. carrolli* × *D. watanabei*).

Discussion

The 163 genomes analyzed in this study provided a broader sampling across Drosophilidae when compared to previous studies (Marzo *et al*., 2008; Acurio, 2015), including many different taxonomic levels. We were able to search for *Galileo* in the genomes of the two subfamilies – Drosophilinae and, for the first time, Steganinae. Furthermore, our sampling included two tribes of the first (Colocasiomyini

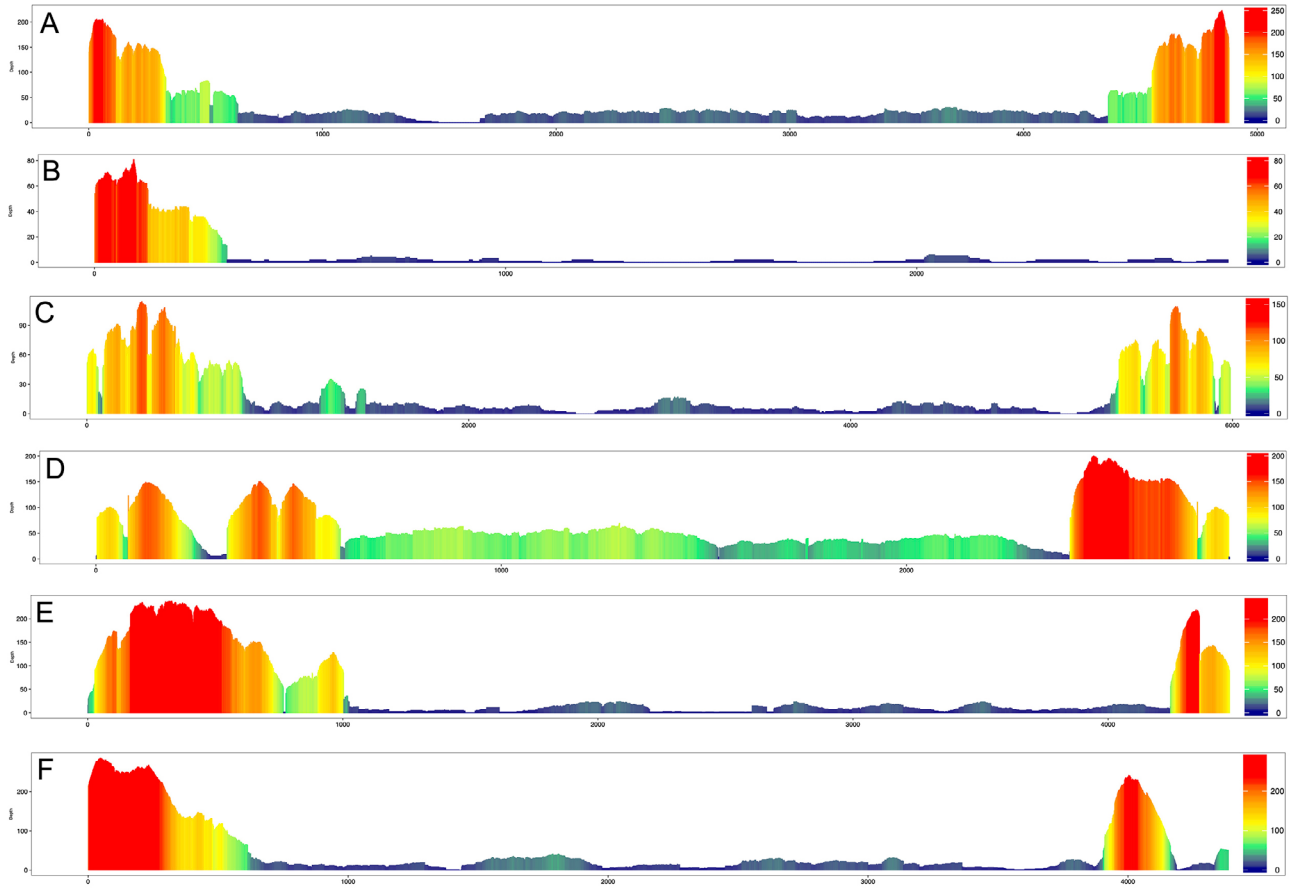

Figure 1 – Coverage graphs for six queries of *Galileo* against its corresponding species: (A) Dana*Galileo* in *Drosophila ananassae*; (B) Dbuz*Galileo* in *D. buzzatii*; (C) Dmoj*Galileo* in *D. mojavensis*; (D) Dper*Galileo* in *D. persimilis*; (E) Dvir*Galileo* in *D. virilis*; and (F) Dwil*Galileo* in *D. willistoni*. Colors correspond to the coverage scale on the right side of each graph. Axis X corresponds to base pairs positions.

and Drosophilini) and two tribes of the latter (Gitonini and Steganini). Indeed, the *Drosophila* genus is a paraphyletic lineage due to the offshoot of several genera within its phylogenetic tree (Suvorov *et al*., 2022); e.g., the *Lordiphosa* genus is placed within the *Sophophora* subgenus as a sister lineage to the Neotropical clade, which includes the *saltans* and *willistoni* groups (Figure 2D).

Galileo is fragmentally widespread in Drosophilidae

The majority of *Galileo* sequences recovered in our study consisted of fragments. Indeed, high levels of structural dynamism in *Galileo* have been described both within and between genomes, as TIRs presented variable sizes (see review in Marzo *et al*., 2008). Therefore, our results suggest that the canonical *Galileo* is widespread and abundant in the genomes of Drosophilidae, although its copies are potentially defective. Given the lack of coding for a transposase, these copies would be incapable of autonomous transposition, remaining as relics—as in the case of Miniature Inverted-repeat Transposable Elements (MITEs).

The hypothesis of classifying these fragmented copies as MITEs of *Galileo* in *D. mojavensis* was considered by Marzo *et al*. (2013a), but was discarded by those authors because the sequences were longer and had a lower copy number compared to typical MITEs. However, our analysis of normalized coverage suggested the opposite; highly amplified short segments of *Galileo* TIRs were detected (Figures 1 and S2-S7), consistent with the size of MITEs. In *D. virilis*, for example, the TPase segment of Dvir*Galileo* had a low coverage (~10X) while its TIRs had a coverage of < 200X (Figure 1E). Although strong evidence was found, further characterization is still needed to assist in the classification of these short canonical sequences as MITEs.

Interestingly, *Galileo* seems to be highly amplified in Neotropical species. Among the 15 species with the highest copy number (Figure 3; Table S1), eight are endemic to the Neotropical region: *D. mojavensis*, *D. sturtevanti*, *D. willistoni*, *D. paulistorum*, *D. navojoa*, and *D. buzzatii*, *D. tropicalis*, and *D. montana* (listed from the highest to the lowest copy number). In fact, the heterogeneity found across the Neotropical region provides innumerous distinct environments, challenging the survival of species (Miranda *et al*., 2022). Such environments also impact genomes, as expanding into new areas may relieve the epigenetic silencing or control of TEs, leading to their mobilization and amplification (Gregory, 2001; Rebollo *et al*., 2010; Antoniolli *et al*., 2023). For instance, *D. willistoni* – which harbors an exceptional diversity of *Galileo* (Gonçalves *et al*., 2014) – is distributed throughout the Neotropical region, and TEs differentially populate its genomes (Bertocchi *et al*., 2022).

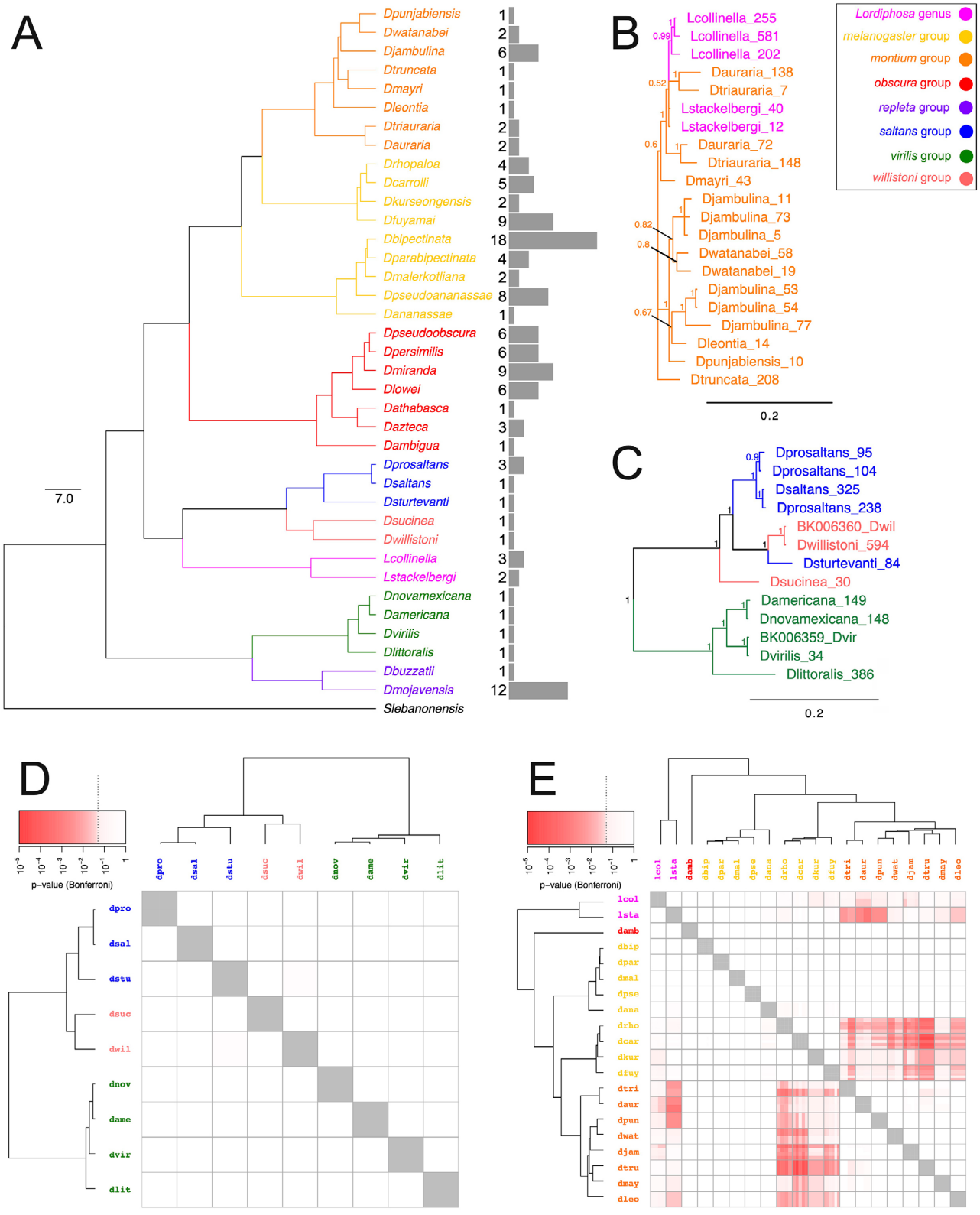

Figure 2 – (A) Ultrametric tree showing the phylogenetic relationships between species harboring nearly complete transposases, assessed through maximum likelihood. Ultrafast bootstrap (UFboot) not shown, as for all nodes UFboot = 100. (B and C) Majority-rule consensus tree showing the phylogenetic relationships between sequences of *Galileo*, (B) found in genomes of the *montium* group of *Drosophila* and species of the *Lordiphosa* genus, and (C) found in genomes of the *saltans*, *virilis* and *willistoni* groups of *Drosophila*; numbers next to each node reflect its posterior probability support. (D and E) Results of the horizontal transposon transfer (HTT) analysis in vhica, between (D) *saltans*, *virilis* and *willistoni* groups of *Drosophila*; and (E) *Lordiphosa* genus and *melanogaster* and *montium* groups of *Drosophila*. (D and E) Red squares represent statistically significant ($P < 0.05$) pairwise comparisons between sequences of *Galileo*, indicating a HTT event. Phylogenetic relationships between host genomes are shown by ultrametric trees drawn on the external sides of each graph.

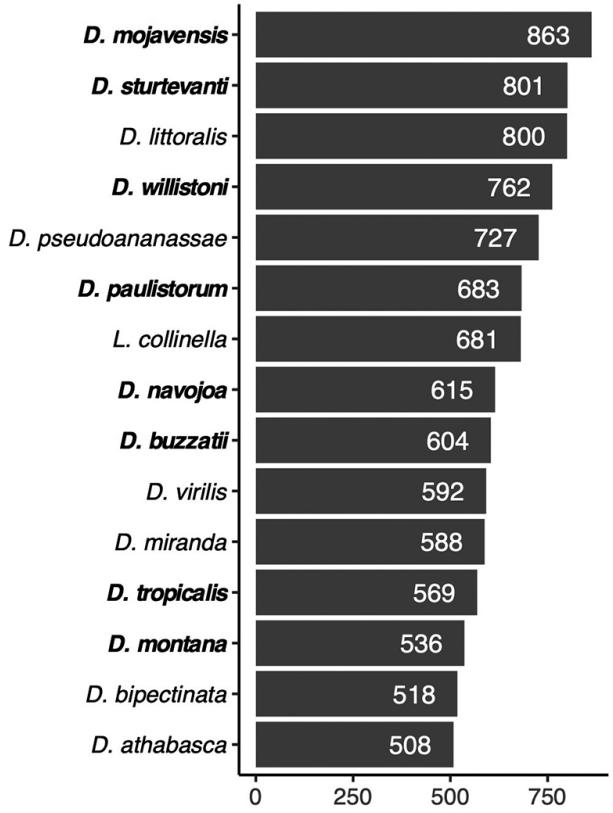

Figure 3 – Number of sequences (X axis) masked as *Galileo* elements by RepeatMasker for the top 15 species (Y axis) with the highest number of sequences. Species highlighted in bold are endemic to the Neotropical region.

Signals of HTT in the *Sophophora* subgenus

The overall congruence between the phylogeny of *Galileo* and that of its host genomes, in terms of clustering species of the same group into the same clade (Figure S1), may be explained by vertical transmission (Acurio, 2015). However, the observed incongruence involving copies found in *Lordiphosa* and species of the *montium* group (Figure 2B) was confirmed as horizontal transfer (HTT) event (Figure 2E). The *Lordiphosa* genus is actually a sister lineage to the *willistoni* group, and its MRCA with the *montium* group diverged around 40 million years ago (Mya) (Suvorov *et al*., 2022). In this case, the oldest HTT event between them (*L. stacklbergi* × *D. punjabiensis*) is estimated to have occurred at around 2.2 Mya; much more recent than their MRCA.

Other cases of HTT involved the *melanogaster* and *montium* groups, whose MRCA diverged around 20 Mya (Suvorov *et al*., 2022); also much older than the oldest HTT detected between them (around 6 Mya for *D. carrolli* × *D. watanabei*). The species involved with HTT events occur in sympatry, mainly in the Palearctic region of Asia (TaxoDros v1.04) – which permits a niche overlap. Addittionaly, *Galileo* exhibits a patchy distribution both in the *Lordiphosa* genus and the *melanogaster* and *montium* groups (Table S1); in this case, the TE is present in some species but absent in another closely related one(s).

Furthermore, a specific THAP binding site for the *Galileo* transposase was identified at the 3' end TIRs (Marzo *et al*., 2013b). The sequences of *Galileo* found in these species involved in HTT cases presented highly conserved and amplified 3' TIRs (Figures 1 and S2-S7), providing further support for the plausibility of such HTT events. Nonetheless, the successful establishment of a TE in new genomes is highly dependent on its transposition rate (Le Rouzic and Capy, 2005), as it must avoid being lost in the population due to genetic drift (Blumenstiel, 2019). While *L. stackelbergi* presented a low number of sequences (49 sequences), *L. collinella* harbors more than 680 sequences (Table S1), similar to *D. buzzatii* (604 sequences), in which *Galileo* was first described. Many other cases of low copy number were also detected (Table S1), and the smallest include *D. ambigua* (10), *D. punjabiensis* (37). and *D. watanabei* (58). The process of stochastic loss of an element may explain both its patchy distribution and low copy number (Blumenstiel, 2019), as observed in *mariner*-like elements in *Drosophila* (Lohe *et al*., 1995) and *Rex* elements in the ray-finned fish *Characidium* (Pucci *et al*., 2018).

Lineage sorting explains the similarity between the *saltans*, *virilis* and *willistoni* groups

Marzo *et al*. (2008) described a high similarity between the copies found in the genomes of *D. virilis* and *D. willistoni*. Interestingly, the first belongs to the *Drosophila* subgenus, while the latter belongs to the *Sophophora* subgenus – their MRCA diverged around 49.9 Mya (Suvorov *et al*., 2022). Acurio (2015) later confirmed this close relationship, identifying it along with the *guarani* and *tripunctata* groups (*Drosophila* subgenus). Our results further corroborate both studies by expanding the sample size to include *D. littoralis* and *D. novamexicana* (*virilis* group).

Interestingly, *Galileo* sequences found in each of these two groups clustered into sister clades that corresponded to their host species, with the addition of sequences from the *saltans* group in the latter. This clade (*virilis* + *saltans* + *willistoni*) was the first to split in the evolution of *Galileo* – also congruent with Marzo *et al*. (2008). These authors also proposed two explanations for the incongruence between the phylogenies of *Galileo* and its host genomes: lineage sorting with ancestral HTT (Acurio, 2015); or horizontal transfer itself. As no signal of HTT was detected between or within these three species groups (Figure 2A), lineage sorting is a plausible explanation (Cummings, 1994). In this case, the transposon is vertically transmitted, but its copies coalesce prior to the split between the host species (Tenaillon *et al*., 2010) or are differentially lost along the branches of the species tree (Marzo *et al*., 2008).

Conclusions

The evolutionary history of *Galileo* in Drosophilidae is marked mostly by vertical and possibly ancient horizontal transmissions, as identified by Acurio (2015), with stochastic loss through genetic drift occurring while species diverged. In addition, its high fragmentation level is compatible with the characteristics of MITEs, although a thorough characterization is still needed to confirm this. *Galileo* found favorable conditions for its amplification in the heterogeneous Neotropical region, with an astounding copy number detected in Drosophilidae species inhabiting this area. Finally, considering the potential of *Galileo* to induce chromosomal rearrangements and their

evolutionary implications, the HTT described between *Lordiphosa* and the *montium* group, and between the latter and the *melanogaster* group, these results raise an intriguing question (Alfredo Ruiz, personal communication): could evolution be infectious?

Acknowledgements

The authors thank Dr. Arnaud Le Rouzic and Dr. Gabriel L. Wallau for technical assistance with the *vhica* analysis and Dr. Alfredo Ruiz for inspiring comments in early stages of this study. This study was supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), under scholarship No. 141319/2020-8, and research productivity grant No. 312781/2018-0.

Conflict of Interest

The authors declare that there is no conflict of interest that could be perceived as prejudicial to the impartiality of the reported research.

Author Contributions

HRMA conceptualization, data curation, formal analysis, investigation, methodology, visualization, writing-original draft, writing-review and editing; SP data curation, formal analysis, methodology, software, writing-review and editing; MD conceptualization, formal analysis, methodology, supervision, writing-review and editing; VLSV conceptualization, project administration, resources, supervision, writing-review and editing.

References

Acurio AE (2015) Coevolutionary analysis of the transposon *Galileo* in the genus *Drosophila*. M. Sc. Thesis, Autonomous University of Barcelona, 219 p.

Antoniolli HRM, Deprá M and Valente VLS (2023) Patterns of genome size evolution versus fraction of repetitive elements in *statu nascendi* species: the case of the *willistoni* subgroup of *Drosophila* (Diptera, Drosophilidae). Genome 66:193–201.

Bailly-Bechet M, Haudry A and Lerat E (2014) "One code to find them all": A perl tool to conveniently parse RepeatMasker output files. Mob DNA 5:13.

Bertocchi NA, Oliveira TD, Deprá M, Goñi B and Valente VLS (2022) Interpopulation variation of transposable elements of the *hAT* superfamily in *Drosophila willistoni* (Diptera: Drosophilidae): *in-situ* approach. Genet Mol Biol 45:e20210287.

Blumenstiel JP (2019) Birth, school, work, death, and resurrection: The life stages and dynamics of transposable element proliferation. Genes 10:336.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL and Feschotte C (2018) Ten things you should know about transposable elements. Genome Biol 19:199.

Cáceres M, Ranz JM, Barbadilla A, Long M and Ruiz A (1999) Generation of a widespread *Drosophila* inversion by a transposable element. Science 285:415–418.

Canapa A, Barucca M, Biscotti MA, Forconi M and Olmo E (2015) Transposons, genome size, and evolutionary insights in animals. Cytogenet Genome Res 147:217–239.

Capella-Gutiérrez S, Silla-Martínez JM and Gabaldón T (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25:1972–1973.

Carvalho TL, Cordeiro J, Vizentin-Bugoni J, Fonseca PM, Loreto ELS and Robe LJ (2023) Horizontal transposon transfer and their ecological drivers: The case of flower-breeding Drosophila. Genome Biol Evol 15:evad068.

Casals F, Cáceres M and Ruiz A (2003) The foldback-like transposon *Galileo* is involved in the generation of two different natural chromosomal inversions of *Drosophila buzzatii*. Mol Biol Evol 20:674–685.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.

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

Cummings MP (1994) Transmission patterns of eukaryotic transposable elements: Arguments for and against horizontal transfer. Trends Ecol Evol 9:141–145.

Colonna Romano N and Fanti L (2022) Transposable elements: Major players in shaping genomic and evolutionary patterns. Cells 11:1048.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B and Flouri T (2020) ModelTest-NG: A new and scalable tool for the selection of DNA and protein evolutionary models. Mol Biol Evol 37:291–294.

Delprat A, Negre B, Puig M and Ruiz A (2009) The transposon *Galileo* generates natural chromosomal inversions in *Drosophila* by ectopic recombination. PLoS One 4:e7883.

Deprá M, Ludwig A, Valente VL and Loreto EL (2012) *Mar*, a MITE family of *hAT* transposons in *Drosophila*. Mob DNA 3:13.

Dotto BR, Carvalho EL, Silva AF, Duarte Silva LF, Pinto PM, Ortiz MF and Wallau GL (2015) HTT-DB: Horizontally transferred transposable elements database. Bioinformatics 31:2915–2917.

Edgar RC (2004) MUSCLE: Multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797.

Fattash I, Rooke R, Wong A, Hui C, Luu T, Bhardwaj P and Yang G (2013) Miniature inverted-repeat transposable elements: Discovery, distribution, and activity. Genome 56:475–486.

Finnegan DJ (1989) Eukaryotic transposable elements and genome evolution. Trends Genet 5:103–107.

Gilbert C and Feschotte C (2018) Horizontal acquisition of transposable elements and viral sequences: Patterns and consequences. Curr Opin Genet Dev 49:15–24.

Gonçalves JW, Valiati VH, Delprat A, Valente VL and Ruiz A (2014) Structural and sequence diversity of the transposon *Galileo* in the *Drosophila willistoni* genome. BMC Genomics 15:792.

Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev Camb Philos Soc 76:65–101.

Kidwell MG and Lisch D (1997) Transposable elements as sources of variation in animals and plants. Proc Natl Acad Sci U S A 94:7704–7711.

Le Rouzic A and Capy P (2005) The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. Genetics 169:1033–1043.

Lim JK and Simmons MJ (1994) Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays 16:269–275.

Lohe AR, Moriyama EN, Lidholm DA, Hartl DL (1995) Horizontal transmission, vertical inactivation, and stochastic loss of *mariner*-like transposable elements. Mol Biol Evol 12:62–72.

Loreto ELS, Carareto CMA and Capy P (2008) Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity 100:545–554.

Marzo M, Puig M and Ruiz A (2008) The *Foldback*-like element *Galileo* belongs to the *P* superfamily of DNA transposons and is widespread within the *Drosophila* genus. Proc Natl Acad Sci U S A 105:2957–2962.

Marzo M, Bello X, Puig M, Maside X and Ruiz A (2013a) Striking structural dynamism and nucleotide sequence variation of the transposon *Galileo* in the genome of *Drosophila mojavensis*. Mob DNA 4:6.

Marzo M, Liu D, Ruiz A and Chalmers R (2013b) Identification of multiple binding sites for the THAP domain of the *Galileo* transposase in the long terminal inverted-repeats. Gene 525:84-91.

Manni M, Berkeley MR, Seppey M, Simão FA and Zdobnov EM (2021) BUSCO Update: Novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 38(10):4647–4654.

Melo ES and Wallau GL (2020) Mosquito genomes are frequently invaded by transposable elements through horizontal transfer. PLOS Genetics 16:e1008946.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A and Lanfear R (2020) IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol 37:1530–1534.

Miranda FR, Machado AF, Clozato CL and Silva SM (2022) Nine biomes and nine challenges for the conservation genetics of Neotropical species, the case of the vulnerable giant anteater (*Myrmecophaga tridactyla*). Biodivers Conserv 31:2515–2541.

Negm S, Greenberg A, Larracuente AM and Sproul JS (2021) RepeatProfiler: A pipeline for visualization and comparative analysis of repetitive DNA profiles. Mol Ecol Resour 21:969–981.

Pace JK, Gilbert C, Clark MS and Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc Natl Acad Sci U S A 105:17023–17028.

Peccoud J, Loiseau V, Cordaux R and Gilbert C (2017) Massive horizontal transfer of transposable elements in insects. Proc Natl Acad Sci U S A 114:4721–4726.

Panaud O (2016) Horizontal transfers of transposable elements in eukaryotes: The flying genes. C R Biol 7–8: 296–299.

Pucci MB, Nogaroto V, Moreira-Filho O and Vicari R (2018) Dispersion of transposable elements and multigene families: Microstructural variation in *Characidium* (Characiformes: Crenuchidae) genomes. Genet Mol Biol 41:585–592.

Ranwez V, Douzery EJP, Cambon C, Chantret N and Delsuc F (2018) MACSE v2: Toolkit for the alignment of coding sequences accounting for frameshifts and stop codons. Mol Biol Evol 35:2582–2584.

Rebollo R, Horard B, Hubert B and Vieira C (2010) Jumping genes and epigenetics: Towards new species. Gene 454:1–7.

Ren L, Huang W, Cannon EKS, Bertioli DJ and Cannon SB (2018) A mechanism for genome size reduction following genomic rearrangements. Front Genet 9:454.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA and Huelsenbeck JP (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539-542.

Schaack S, Gilbert C and Feschotte C (2010) Promiscuous DNA: Horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol Evol 25:537–546.

Sharp PM and Li WH (1989) On the rate of DNA sequence evolution in *Drosophila*. J Mol Evol 28: 398–402.

Suvorov A, Kim BY, Wang J, Armstrong EE, Peede D, D'Agostino ERR, Price DK, Waddell PJ, Lang M, Courtier-Orgogozo V *et al*. (2022) Widespread introgression across a phylogeny of 155 *Drosophila* genomes. Curr Biol 32:111–123.

Tamura K, Stecher G and Kumar S (2021) MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol 38:3022–3027.

Tenaillon MI, Hollister JD and Gaut BS (2010) A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471–478.

Thompson JD, Higgins DG and Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680.

Wallau GL, Capy P, Loreto E, Le-Rouzic A and Hua-Van A (2015) VHICA, a new method to discriminate between vertical and horizontal transposon transfer: Application to the *Mariner* Family within *Drosophila*. Mol Biol Evol 33:1094–1109.

Wells JN and Feschotte C (2020) A field guide to eukaryotic transposable elements. Annu Rev Genet 54:539–561.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P and Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982.

Internet Resources

TaxoDros. The database on Taxonomy of Drosophilidae v1.04, https://www.taxodros.uzh.ch/ (accessed 10 April 2023)

Blin K (2021) NCBI Genome Downloading Scripts, https://github.com/kblin/ncbi-genome-download/ (accessed 10 November 2022)

McGowan J (2020) jamiemcg/BUSCO_phylogenomics: BUSCO v4, https://zenodo.org/records/7334954 (accessed 8 April 2023)

Rambaut A (2018) FigTree v1.4.4, https://github.com/rambaut/figtree/ (accessed 9 April 2023)

Smit A, Hubley R and Green P (2023) RepeatMasker 4.0, http://www.repeatmasker.org/ (accessed 10 November 2022)

Supplementary material

The following online material is available for this article:

Figure S1 – Phylogenetic relationships between sequences of *Galileo* found across genomes of Drosophilidae, reconstructed through Bayesian Inference.

Figure S2 – Normalized coverage graphs for Dana*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S3 – Normalized coverage graphs for Dbuz*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S4 – Normalized coverage graphs for Dmoj*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S5 – Normalized coverage graphs for Dper*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S6 – Normalized coverage graphs for Dvir*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Figure S7 – Normalized coverage graphs for Dwil*Galileo* used for searching transposase sequences across genomes of Drosophilidae.

Table S1 – List and taxonomy of Drosophilidae species included in this study, including positive results from BLASTn searches of the *Galileo* transposase sequences and accession numbers to the genome assembly and short-read sequencing data on NCBI.

Table S2 – List of genes used to normalize the results of profile and abundance of *Galileo* across the analyzed genomes in this study.

Table S3 – List of genes used for Codon Usage Bias (CUB) comparisons in the analysis of horizontal transposon transfer (HTT) in vhica R package.

Table S4 – Statistics of assembly completeness for each analyzed genome.

Table S5 – Statistically significant results of pairwise comparisons of horizontal transposon transfer performed with vhica at HTT-DB.

Associate Editor: Loreta Brandão de Freitas