bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 **Title:** Evolutionary history and classification of Micropia retroelements in
2 Drosophilidae species
3
4 **Short title:** Micropia retrotransposable element in Drosophilidae
5
6 Juliana Cordeiro[1#†*] (orcid.org/0000-0003-0902-9638), Tuane L. Carvalho[2†], Vera L. da
7 S. Valente[3], Lizandra J. Robe[2, 4] (orcid.org/0000-0001-8506-9143)
8
9 [1]Departamento de Ecologia, Zoologia e Genética, Instituto de Biologia, Universidade
10 Federal de Pelotas, Pelotas, RS, Brazil
11 [2]Programa de Pós-Graduação em Biodiversidade Animal, Departamento de Biologia
12 Universidade Federal de Santa Maria, Santa Maria, RS, Brazil
13 [3]Departamento de Genética, Instituto de Biociências, Universidade Federal do Rio
14 Grande do Sul, Porto Alegre; Rio Grande do Sul; Brazil
15 [4]Departamento de Biologia, Centro de Ciências Naturais e Exatas, Universidade Federal
16 de Santa Maria, Santa Maria, RS, Brazil
17 [#]Current location: Department of Genetics, University of Wisconsin – Madison, USA.
18
19 * Corresponding author
20 E-mail: jlncdr@gmail.com
21
22 [†] These authors contributed equally to the manuscript
23

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

24 # Abstract

25 Current knowledge indicates TEs have been shaping the evolution of genomes and host
26 species, contributing to the creation of new genes and promoting rearrangements
27 frequently associated with new regulatory networks. Support for these hypothesis
28 frequently result from studies with model species, and *Drosophila* detaches as a great
29 model organism to the study of TEs. Micropia belongs to the Ty3/Gypsy group of LTR
30 retroelements, and comprises one of the least studied *Drosophila* transposable elements.
31 In this study, we assessed the evolutionary history of Micropia within Drosophilidae,
32 while trying to assist in the classification of this TE. At first, we analyzed its presence in
33 the genome of several species from natural populations and then, based on searches
34 within genomic databases, we retrieved Micropia-like sequences from distinct
35 Drosophilidae species genomes. We expanded the knowledge of Micropia distribution
36 within *Drosophila*, and detected an array of divergent sequences, which allowed
37 subdividing this retroelement in 20 subfamilies. Even so, a patchy distribution of
38 Micropia sequences within the Drosophilidae phylogeny could be identified combined
39 with incongruences of the species and the Micropia phylogenies. Comparing dS values
40 between Micropia and host nuclear sequences, we found several cases of unexpected
41 high levels of similarity between Micropia sequences found in divergent species. All
42 these findings propose a hypothesis to the evolution of Micropia within Drosophilidae,
43 including several VTTs and HTTs events, associated to ancestral polymorphisms and
44 recurrent Micropia sequences diversification.

45

46 **Key words:** transposable elements diversification; LTR retrotransposon subfamilies;
47 *cardini* group; *repleta* group; *melanogaster* group; horizontal transposon transfer;
48 vertical transposon transfer

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

49 # Introduction

50 Since Barbara McClintock publication of maize genes moving around the
51 genome, transposable elements (TEs) went from junk to pivotal characters in the control
52 and evolution of genomes. The discovery of unexpected high amounts of TEs in the
53 genome of distinct species has pointed out toward functions of TEs on these genomes
54 [1, 2, 3]. In fact, current knowledge indicates TEs have been shaping the evolution of
55 genomes and host species [4], contributing to the creation of new genes [5, 6] and
56 promoting rearrangements frequently associated with new regulatory networks [7, 8 9].
57 More than this, there is even evidence that TEs may assist in the control of embryonic
58 development [9, 10] and genomic plasticity [11].

59 A large fraction of the genomes of most eukaryotes is composed by TEs known
60 as retroelements [12, 13, 14], some of which belong to the LTR order. Phylogenetic
61 analyses of such retroelements reveal an evolutionary history consisting mainly of
62 vertical transmissions and intraspecific diversification [15]. However, autonomous TEs
63 are able to invade naïve genomes through horizontal transposon transfers (HTT), in
64 which they make copies of themselves and evade host defense systems before becoming
65 fully silenced by genomic anti-TE mechanisms [16, 17]. Although HTTs are still
66 considered rare events, mainly because we can only detect the successful ones, it seems
67 that such events represent an important step in the TEs' life cycle, enabling them to
68 evade the natural progression of their birth-and-death process that can culminate in their
69 extinction [18, 19, 17, 16]. After the HTT event, TEs can have a wide range of positive
70 and/or negative consequences in the host genome [20]; but mainly, they become a new
71 set of sequences were evolution can take place, unveiling their relevance to host genome
72 evolution [21, 22].

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

73 A growing number of studies have identified HTTs using distinct analysis
74 strategies [15, 16, 23, 24, 25]. For instance, a patchy taxonomic distribution among
75 monophyletic species is expected if TEs are moving horizontally rather than being
76 vertically inherited. This patchy distribution associated with incongruences between
77 species and TEs phylogenies and an unexpected high nucleotide identity between TEs
78 found in the genome of divergent species widely strengthens the evidence for HTT [26,
79 17, 25, 27, 28]. According to these criteria, LTR retrotransposons account for
80 approximately 20% of HTT events across the insect's genomes [16]. This value
81 increases when only *Drosophila* genomes are analyzed, e.g. LTR retroelements account
82 for 90% of the HTT events detected across the genomes of *D. melanogaster*, *D.*
83 *simulans* and *D. yakuba* [29].

84 Micropia is a retrotransposon that belongs to the Ty3/Gypsy group of LTR
85 retroelements [30], which is closely related to retroviruses [31, 32]. Micropia was first
86 discovered in the lampbrush loops of the *Drosophila hydei* Y chromosomes. Until
87 recently, there were only four best-characterized Micropia elements, and these were
88 found in the genomes of *D. hydei* (named dhMiF2 and dhMiF8) and *D. melanogaster*
89 (named Dm11 and Dm2) [33, 34, 35]. Recently, complete and probably active Micropia
90 reference sequences were found in the genomes of *D. simulans* and *D. sechellia* [15].
91 Nevertheless, Micropia related sequences are also present in the genomes of several
92 *Drosophila* and *Zaprionus* species, showing an irregular pattern of distribution [36, 37,
93 38, 39, 40, 41]. In some species (like *D. hydei*), Micropia shows an effective
94 transcription based repression mechanism associated with antisense RNAs [37, 41, 42].
95 On the other hand, the genomes of other species (like *D. melanogaster*) seem to be
96 absent from autonomous Micropia sequences [41].

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

97 Here, our goals were to understand the most likely evolutionary history of
98 Micropia retroelement sequences within Drosophilidae, while trying to assist in the
99 classification of this TE. At first, we analyzed its presence in the genome of several
100 species from natural populations and sequenced the detected elements. Then, based on
101 searches within genomic databases, we identified and isolated Micropia-like sequences
102 in the genomes of different species. All these sequences were used to propose a
103 hypothesis to the evolution of Micropia within Drosophilidae, while assessing its
104 subdivision and identifying several cases of HTTs.

105

106 ## Materials and Methods

107 ### Species analyzed

108 For this study, we analyzed the presence/absence of Micropia sequences in the genomes
109 of natural populations of 24 *Drosophila* species using PCR-blot and Dot-blot searches
110 (hereafter "*in vitro searches*") following the methodology described at *In vitro*
111 *searches: DNA manipulation, PCR-blot, Dot-blot and sequencing* (see below) (Table 1).
112 In vitro searches were also previously performed for other three species of the *cardini*
113 group [39], and for *D. melanogaster* [34, 35] and *D. hydei* [33]; the sequences thus
114 obtained were downloaded from GenBank. We also analyzed the presence/absence of
115 Micropia sequences in the genome of 26 species with available genomes at NCBI
116 (blast.ncbi.nlm.nih.gov/Blast.cgi) and Flybase (flybase.org/blast/) websites (hereafter
117 "*in silico* searches") plus two species, *D. suzukii* and *D. buzzatii*, whose genomes are
118 available at personal websites (http://spottedwingflybase.org/ and
119 https://dbuz.uab.cat/welcome.php, respectively), following the criteria described in *In*
120 *silico searches: Genomic analysis* (see below) (Table 1). Thus, *D. buzzatii*, and *D.*
121 *melanogaster* were the only species for which both search strategies were applied. The

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

122 classification scheme adopted for each of these species across this study follows the

123 proposal of [43].

124

125 **Table 1. Presence/absence of Micropia sequences in the genomes of Drosophilidae species. Methodology employed, number of sequences**

126 **and GeneBank accession numbers are also shown. * Sequences used as initial BLASTn queries.**

Genus	Subgenus	Group species	Species	Presence/absence	Methodology	GenBank acc. nos.
Drosophila	*Dorsilopha*	*busckii*	*D. busckii*	+	in silico	see Table S2
Drosophila	*Drosophila*	*cardini*	*D. acutilabella*	+	in vitro	FJ748684*, FJ748685*, FJ748686*, FJ748687*, FJ748688*
			D. arawakana	-	in vitro	-
			D. cardini	+	in vitro	FJ748690*, FJ748691*, FJ748692*
			D. cardinoides	+	in vitro	EF090263*, EU149929*, EU149930*
			D. dunni	-	in vitro	-
			D. neocardini	+	in vitro	EF090264*, EU149931*, EU149932*, EU149933*
			D. neomorpha	+	in vitro	FJ748695*, FJ748696*, FJ748697*
			D. nigrodunni	-	in vitro	-
			D. parthenogenetica	+	in vitro	FJ748698*, FJ748699*, GQ339587*, GQ339588*, GQ339589*, GQ339590*
			D. polymorpha	+	in vitro	EF090265*, EF149934*, EF149935*, EF149936*, EF149937*
			D. procardinoides	+	in vitro	FJ748700*, FJ748701*, FJ748702*
			D. similis	-	in vitro	-
		funnebris	*D. funnebris*	-	in vitro	-
		guaramunu	*D. griseolineata*	-	in vitro	-
			D. maculifrons	-	in vitro	-
		guarani	*D. guaru*	-	in vitro	-
			D. ornatifons	-	in vitro	-
		immigrans	*D. albomicans*	+	in silico	see Table S2
			D. immigrans	-	in vitro	-

	tripunctata	*D. bandeirantorum*	-	in vitro	-
		D. mediodiffusa	-	in vitro	-
		D. mediopictoides	-	in vitro	-
		D. mediopunctata	-	in vitro	-
		D. paraguayensis	-	in vitro	-
		D. paramediostriata	-	in vitro	-
		D. tripunctata	-	in vitro	-
Siphlodora	*repleta*	*D. arizonae*	+	in silico	see Table S2
		D. buzzatii	+	in vitro/ in silico	FJ748689*, GQ339579*, GQ339580*, GQ339582*, see Table S2
		D. hydei	+	in vitro	X13304*, X13305*
		D. mercatorum	+	in vitro	FJ748693*, FJ748694*, GQ339583*, GQ339584*, GQ339585* GQ339586*
		D. mojavensis	+	in silico	see Table S2
		D. navojoa	+	in silico	see Table S2
		D. zottii	+	in vitro	FJ748703*, GQ339578*
	virilis	*D. americana*	+	in silico	see Table S2
		D. virilis	+	in silico	see Table S2
Sophophora	*melanogaster*	*D. ananassae*	+	in silico	see Table S2
		D. bipectinata	+	in silico	see Table S2
		D. elegans	+	in silico	see Table S2
		D. erecta	+	in silico	see Table S2
		D. ficusphila	+	in silico	see Table S2
		D. kikkawai	+	in silico	see Table S2
		D. melanogaster	+	in vitro/in silico	X14037*, X14173*, see Table S2
		D. rhopaloa	+	in silico	see Table S2
		D. sechellia	+	in silico	see Table S2
		D. simulans	+	in silico	see Table S2

			D. suzukii	+	in silico	see Table S2
			D. takahashii	+	in silico	see Table S2
			D. yakuba	+	in silico	see Table S2
		obscura	*D. miranda*	-	in silico	-
			D. persimilis	-	in silico	-
			D. subobscura	-	in silico	-
		willistoni	*D. willistoni*	+	in silico	see Table S2
	Haiwaiian Drosophila	-	*D. grimshawi*	-	in silico	-
Phortica	-	*variegata*	*P. variegata*	-	in silico	-
Scaptodrosophila	-	-	*S. lebanonensis*	+	in silico	see Table S2

127

128

129

130

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

131 ***In vitro* searches: DNA manipulation, PCR-blot, Dot-blot and sequencing**

132 Genomic DNA was prepared according to [44]. PCR reactions were performed
133 using Micropia primers to amplify the reverse transcriptase (RT) domain within the pol
134 gene, as described in [39]. The following conditions were used for 25 µl PCR reactions:
135 25 ng of template DNA, 20 pMol of each primer, 0.2 mM of each nucleotide, 1.5 mM
136 MgCl2 and 1 unit Taq DNA polymerase in 1x polymerase buffer (all from Invitrogen).
137 Amplifications parameters were 95°C for 2 min, 35 cycles at 95°C for 30 s, 50-60°C for
138 30 s and 72°C for 1 min, followed by an extension step at 72°C for 10 min. *Drosophila*
139 *hydei* genomic DNA was used as a positive control.

140 In order to confirm the homology of the amplified fragments, a PCR-blot was
141 prepared with the obtained PCR amplicons. The PCR products were separated by
142 electrophoresis using a 1% agarose gel and transferred to nylon membranes (Hybond
143 N+®, GE Healthcare), where hybridization was carried out using an 812 bp fragment of
144 Micropia from *D. hydei* as probe. This fragment ranges from nucleotide 1,777 to 2,589
145 of the *D. hydei* dhMiF2 sequence (GenBank acc. no. X133041), covering part of the RT
146 sequence. The probe label and signal detection were performed using the Gene
147 ImagesTM AlkPhos DirectTM labelling and detection system (GE Healthcare),
148 according to manufacturer's instructions. The membranes were hybridized at 55°C and
149 exposed for 5 min.

150 A Dot-blot procedure was also performed using genomic DNA. Denaturation
151 was performed using 3 µg of genomic DNA in a final volume of 10 µl, which was
152 directly applied onto a nylon membrane (Hybond N+®, GE Healthcare). As positive
153 control, 5 ng (in 10 µl) of the dhMiF2 probe was used. The probe labeling, signal
154 detection, and hybridization temperature were performed as above. Dot-blot revealing
155 film underwent 3 min exposure.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

156 For sequencing, PCR amplicons from each species presenting positive signals
157 for Micropia were separated by 1.5% agarose gel electrophoresis and purified using
158 Illustra GFXTM PCR DNA and Gel Band Purification kit (GE Healthcare) according to
159 the supplier's specifications. The fragments were cloned using pGEM®-T Easy Vector
160 system (Promega). The obtained recombinant plasmids underwent a new PCR reaction
161 using the universal M13 primers at a 55°C annealing temperature. The amplicons were
162 purified using ExoI-SAP (GE Healthcare) and directly sequenced in a
163 MegaBACETM500 (GE Healthcare). Forward and reverse strands were sequenced;
164 ambiguities and compressions were resolved through assemblage in the Staden Package
165 Gap4 program [45]. GenBank accession numbers are indicated in Table 1.
166

167 ***In silico* searches: genomic analysis**

168 BLAST searches were performed in *Drosophila* genomes available at NCBI
169 (blast.ncbi.nlm.nih.gov/Blast.cgi) and Flybase website (flybase.org/blast/), using default
170 parameters. For *D. buzzatii* and *D. suzukii*, searches were performed, respectively, in the
171 '*Drosophila buzzatii* Genome Project' website (dbuz.uab.cat/welcome.php) and in the
172 'Spotted Wing FlyBase' website (spottedwingflybase.org/). The searches were finished
173 in January 2018.
174 The initial BLASTn queries consisted of Micropia reverse transcriptase (RT)
175 nucleotide sequences obtained by [39, 33, 34 and 35] retrieved from GeneBank (Table
176 1). The retrieved sequences obtained during the *in silico* searches showing scores higher
177 than 50 were downloaded, including 2 kb from both sides of each hit. After that, each
178 retrieved sequence was aligned with the set of query sequences using ClustalW, as
179 implemented in MEGA6 software [46]. Sequences that failed to align in this multiple
180 alignment step were further submitted to pairwise or even local alignment against the

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

181 query sequence presenting the highest score in the BLASTn searches (hereafter "best
182 query"). In this case, fragments presenting less than 300 bp of confirmed homology to
183 its best query sequence were withdrawn from the alignment. Furthermore, after
184 compressing the analyzed region, identical nucleotide sequences recorded for the same
185 species were joined in a single sequence.

186 A codon-based alignment was then performed using Muscle [47] as
187 implemented in MEGA6 software. Gaps presented in this matrix were further resolved,
188 in order to leave all sequences in frame, to obtain the aligned amino acid matrix. All
189 these translated sequences were then used as queries to perform exhaustive tBLASTn
190 searches, with the same strategy described above. The final matrix encompassed all
191 sequences obtained under these criteria that presented a minimum overlap of 300bp to
192 the previous nucleotide alignment, after a final codon-based alignment performed in
193 Muscle.

194 After completing the matrix, putative functional RT Micropia sequences were
195 identified by translating each unaligned nucleotide sequence in the different reading
196 frames. Once an Open Reading Frame (ORF) was detected, BLASTn searches further
197 confirmed its identity.

198

199 **Phylogenetic analysis and Micropia subfamilies**

200 Phylogenetic analyses were performed using the amino acid alignment obtained
201 after resolving all gaps and leaving all nucleotide sequences in frame. Fifty amino acid
202 sequences belonging to each of the five main clades recently established by [15] for the
203 Micropia/Sacco group within Ty3/Gypsy were selected from the alignment provided by
204 the authors. These sequences were included as a "taxonomic framework" to guide
205 conclusions related to new Micropia sequences in our phylogenetic analyses, in which a

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

206 Copia-like transposable element sequence from *D. melanogaster* (GenBank access
207 number X01472) was used as outgroup.
208 Bayesian phylogenetic analysis (BA) was performed under a mixed model with
209 gamma correction, as implemented in MrBayes3.1.2 software, through Cipres
210 Computational Resources [48]. This Markov Chain Monte Carlo (MCMC) search was
211 run for 10,000,000 generations, with trees saved every 1,000 after a burn-in of 2,500.
212 The Posterior Probability (PP) of each clade on the 50% majority rule consensus tree
213 was calculated and the resulting tree was visualized in FigTree. The tree so obtained
214 was used to detect intraspecific sequences sharing a most recent common ancestor
215 (MRCA). In these cases, only the sequence with the shortest branch (the most similar to
216 the inferred MRCA sequence) was maintained as representative of that clade in a new
217 round of BA analysis. The final tree was compared to the species tree, as compiled from
218 [49, 50, 51, 52, 53, 54 and 55] data. Subfamilies of the Micropia TE sequences were
219 identified using the criterion established by [30], according to which reciprocally
220 monophyletic sequences with less than 30% of divergence at the amino acid level could
221 be grouped in the same TE subfamily. This analysis was performed in MEGA6, using
222 Poisson amino acid substitution model.
223

224 **dS and divergence time estimates**

225 Pairwise synonymous distance (dS) values were estimated for Micropia in-frame
226 nucleotide sequences and for three host nuclear genes sequences using Nei and Gojobori
227 (1986) method, as implemented in MEGA6. Alcohol-dehydrogenase (*Adh*), alpha–
228 metildopa (*Amd*) and dopa-decarboxylase (*Ddc*) sequences were downloaded from
229 GeneBank or retrieved from the species genomes using BLASTn searches (for
230 GenBank or scaffold accession numbers, see S1 Table). In order to identify if the

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

231 Micropia dS values were significantly lower than those observed for the host nuclear
232 genes, accounting for differences in the number of synonymous sites, a one-tailed
233 Fisher's exact test was performed using R v.3.5.2 [57]. Divergence times were also
234 eventually evaluated using dS estimates and a synonymous substitution rate of 0.016
235 substitutions per site per million years, as calculated for *Drosophila* genes with low
236 codon usage bias [58].
237

238 # Results

239 ## Species analyzed

240 A total of 56 Drosophilidae species were analyzed for the presence/absence of
241 Micropia sequences (Table 1). Thirty species were analyzed by in vitro searches and 28
242 species were analyzed through *in silico* searches.
243

244 ## Patchy distribution of Micropia sequences in the Drosophilidae species genomes

245 The applied methodologies were able to identify the presence of distinct
246 Micropia related sequences in the genome of 34 *Drosophila* species (Table 1 and S2
247 Table). *In vitro* signals of Micropia copies were encountered in *D. melanogaster* and in
248 some species from *cardini* (8 of the 12 species tested) and *repleta* (4 of the 4 species
249 tested) groups, despite the fact that 13 other species were also tested (Table 1, S1 Fig,
250 and data not shown). Conversely, *in silico* searches enabled the isolation of Micropia
251 sequences in the genomes of *D. busckii*, *D. albomicans*, *D. willistoni* and *S.*
252 *lebanonensis*, and in species from the *repleta* (4 of the 4 species tested), *virilis* (2 of the
253 2 species tested) and *melanogaster* (12 of the 12 species tested) groups. None Micropia
254 sequence could be found for *D. grimshawi* (picture wing group), *D. funebris*, *D.*
255 *immigrans* or for any species of the *guaramunu*, *guarani*, *obscura*, and *tripunctata*

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

256 groups. So, interestingly intra-group polymorphisms in the status of presence/absence of
257 Micropia sequences were solely identified for the *cardini* and *immigrans* groups. Fig 1
258 shows the species tree informing the presence and absence of Micropia related
259 sequences in the genomes of each of these species.
260
261 **Fig 1. Phylogenetic reconstruction of species analyzed in this study.** Phylogenetic
262 reconstruction was based on data compiled from [49, 50, 51, 52, 53, 54 and 55]. Species
263 name in black represent presence of Micropia sequences and species name in grey
264 represent absence of such sequences. Distinct branch colours represent distinct
265 subgenera within the *Drosophila* genus, and the classification follows [43]. *Drosophila*
266 genus group species are also indicated to the right. *Scaptodrosophila* and *Phortica* are
267 represented as outgroups of the *Drosophila* genus. Dashed line represents the potential
268 phylogenetic position of *D. zottii*, since there is no molecular phylogeny neither any
269 nuclear or mitochondrial gene available for this species.
270
271 **Phylogenetic analysis, Micropia diversity, and potential coding sequences**
272 As several intraspecific sequences clustered together in the BA phylogenetic tree
273 obtained for the whole set of Micropia sequences (S2 Fig and S1 File), the alignment
274 could be reduced from 298 to 149 sequences (S2 File). The final Micropia phylogenetic
275 tree reinforced reciprocal monophyly of several sets of sequences and confirmed the
276 identity of the retrieved sequences, which were clustered with Micropia sequences
277 obtained by [15] (Fig 2). Further evaluation of the recovered tree topology reveals the
278 presence of four main clusters, which are listed here in ascending order of divergence:
279 the first, presenting the Sacco sequences obtained by [15]; the second, grouping
280 representatives of the Blastopia and MDG3 sequences obtained by [15]; the third,

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

281 presenting the Bicca element recovered by [15]; and the fourth recovering all the
282 Micropia sequences in a major polytomic clade, including sequences obtained by [15].
283
284 **Fig 2. Bayesian phylogenetic tree of the Drosophilidae Micropia sequences**
285 **analyzed in this study.** The phylogenetic tree was based on amino acid sequences
286 following a mixed evolution model with gamma correction. Bargues and Lerat´s
287 sequences [15] were included in the analysis. Numbers from 1 to 20 on the left
288 represent the Micropia subfamilies recovered in our data. Filled circles after Micropia
289 sequence names indicate sequences involved in possible HTT events based on one-
290 tailed Fisher's exact test involving pairwise comparisons of dS values between Micropia
291 and nuclear genes (*Adh* in orange, *Amd* in pink, *Ddc* in purple; see S4 Table). Stars
292 represent the four best-characterized Micropia elements (*D. hydei* dhMiF2 and dhMiF8;
293 and *D. melanogaster* Dm11 and Dm2). The posterior probability of each clade is
294 indicated beside its respective internal branch.
295
296 Following Capy's et al. [30] criteria, we were able to recover 20 potential
297 Micropia subfamilies based on monophyletic sequences (Fig 2) showing amino acid
298 genetic divergence lower than 0.3 (Table 2 and S3 Table). Of these, nine subfamilies are
299 monotypic and represented by a single sequence (subfamilies 2, 6, 8, 13, 16, 17, 18, 19
300 and 20). To the exception of subfamilies 4 and 15 (which were encountered only in
301 species of the *melanogaster* group), all the remaining Micropia subfamilies are
302 composed by species of distinct *Drosophila* species groups and subgenera.

Table 2. Pairwise amino acid genetic distances between Micropia subfamilies.

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19
Subf. 02	0.357																		
Subf. 03	0.313	0.345																	
Subf. 04	0.327	0.343	0.297																
Subf. 05	0.385	0.407	0.338	0.335															
Subf. 06	0.436	0.397	0.412	0.423	0.315														
Subf. 07	0.410	0.442	0.324	0.341	0.376	0.376													
Subf. 08	0.690	0.739	0.533	0.620	0.625	0.496	0.302												
Subf. 09	0.407	0.521	0.328	0.360	0.368	0.351	0.181	0.433											
Subf. 10	0.421	0.407	0.341	0.345	0.381	0.400	0.217	0.451	0.237										
Subf. 11	0.468	0.461	0.431	0.405	0.454	0.422	0.274	0.512	0.284	0.271									
Subf. 12	0.435	0.425	0.360	0.368	0.480	0.441	0.366	0.635	0.384	0.361	0.426								
Subf. 13	0.664	0.401	0.580	0.574	0.549	0.751	0.518	0.586	0.584	0.535	0.573	0.437							
Subf. 14	0.413	0.441	0.358	0.382	0.393	0.374	0.332	0.577	0.334	0.317	0.385	0.354	0.539						
Subf. 15	0.381	0.512	0.346	0.363	0.398	0.406	0.323	0.586	0.318	0.324	0.385	0.302	0.440	0.304					
Subf. 16	1.499	0.857	1.345	1.425	1.338	1.460	1.437	1.368	1.401	1.359	1.302	1.484	1.408	1.391	1.507				
Subf. 17	0.406	0.511	0.349	0.375	0.418	0.426	0.297	0.586	0.313	0.327	0.398	0.292	0.431	0.336	0.140	1.365			
Subf. 18	0.401	0.449	0.330	0.375	0.418	0.390	0.339	0.611	0.312	0.318	0.391	0.288	0.430	0.310	0.149	1.533	0.137		
Subf. 19	0.376	0.505	0.334	0.351	0.371	0.390	0.266	0.598	0.285	0.287	0.337	0.301	0.465	0.309	0.119	1.375	0.157	0.160	
Subf. 20	0.366	0.460	0.330	0.356	0.309	0.432	0.233	0.496	0.238	0.285	0.367	0.247	0.651	0.280	0.145	1.509	0.138	0.123	0.113

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

307 As a result, there are clear cases of incongruence between the species and TE's
308 phylogenies (Figs 1 and 2, respectively), in which Micropia sequences found in the
309 genomes of distantly related species are clustered in the same subfamily in the Micropia
310 phylogeny, and copies within a unique genome do not share a unique and exclusive
311 common ancestor. For example, subfamily 7 (Fig 2) comprises sequences within the
312 genome of *cardini* and *repleta* group species, belonging to the *Drosophila* and
313 *Siphlodora* subgenera, respectively, together with sequences encountered within the
314 genome of *D. willistoni*, which belongs to the *Sophophora* subgenus. As concerns the
315 presence of divergent copies within the same genome, the cases of *D. buzzatii* (*repleta*
316 group), *D. americana* (*virilis* group) and *D. willistoni* (*willistoni* group) should be
317 highlighted, since these species present Micropia sequences widely spread over the tree,
318 nested in five, six and nine of the subfamilies, respectively.

319 The analysis of potential coding sequences for the 98 sequences of Micropia
320 presented in the final tree (sequences of [15] were not included in this analysis, as well
321 as the outgroup Copia sequence) shows that approximately 49% of them (48 from 98)
322 putatively encode for reverse transcriptase enzyme (S4 Table). In fact, from the total set
323 of 34 species with Micropia sequences evaluated here, only *D. erecta, D. kikkaway, D.*
324 *mojavensis* and *D. polymorpha* do not possess potentially encoding sequences.

325

326 **dS estimates and identification of horizontal transposon transfer (HTT) events**

327 The use of *Adh*, *Amd* and *Ddc* nuclear gene sequences held a total of 4,367,
328 4,370 and 4,558 pairwise dS comparisons, respectively (S5 Table). Micropia dS values
329 were lower than those found for the host nuclear genes in 277 cases (significance in the
330 Fisher exact test - with p value < 0.05 - were obtained for 96, 266 and 207 comparisons
331 involving *Adh*, *Amd* and *Ddc*, respectively), revealing patterns incompatible with

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

332 vertical transposon transmission (VTT). Thus, signals of HTTs account for 2.2%, 6.1%
333 and 4.5% of the comparisons performed with *Adh*, *Amd* and *Ddc*, respectively. Fig 2
334 highlights all species involved in at least one case of significantly lower Micropia dS
335 value. Indeed, only 19 of 97 sequences of Micropia for which the Fisher Exact Test
336 could be performed do not present any signal of involvement in HTTs events
337 (sequences of [15] were not included in this analysis, as well as that from the outgroup
338 and from *D. zotti*, for which none of the three nuclear genes have been previously
339 characterized).
340

341 # Discussion

342 ## Micropia classification

343 By comparing our data with those of Bargues and Lerat's [15], it is possible to
344 show that our non-stringent methodology retrieved sequences belonging to Micropia
345 within the Micropia/Sacco group of the *Ty3/gypsy* retrotransposable elements. Within
346 this group, Micropia is recovered as a monophyletic lineage that is sister to the Bica
347 group of LTR retroelements. The Bayesian phylogeny of these sequences highlights the
348 existence of a high array of divergent sequences, which are compatible with the
349 subdivision of Micropia in specific groups. Nevertheless, the taxonomic status
350 represented by these remains a matter of debate.

351 In fact, except for the very well accepted criteria used to classify TEs in classes
352 and subclasses proposed by [59], in general, there is no consensus over the criteria
353 adopted to achieve TEs families and subfamilies [60]. Several authors used different
354 strategies to identify new TE families and subfamilies, whether based on nucleotide
355 and/or amino acid sequence similarities [30, 61, 62, 63, 64, 65, 66]. Given the
356 abundance and diversity of TEs, [64] proposed a classification for eukaryotic TEs based

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

357 uniquely on nucleotide similarities. Nevertheless, given the absence of evolutionary
358 criteria based on reciprocal monophyly, this system is yet widely controversial. So, we
359 adopted here more conservative criteria, according to which different subfamilies are
360 established based on reciprocal monophyly and divergence values higher than 0.3 at the
361 amino acid level [30].

362 Adopting these criteria, our data shows the existence of at least 20 potential
363 Micropia subfamilies that form the reciprocally monophyletic groups or monotypic
364 lineages shown in Fig 2. Several of these subfamilies are spread over distinct
365 *Drosophila* subgenera and groups, although only subfamilies 7 and 12 could be sampled
366 across species of *Sophophora*, *Drosophila*, and *Siphlodora*. In this sense, most
367 sequences within the *Drosophila* subgenus species are clustered in subfamily 7, whereas
368 sequences of *Siphlodora* are highly intermingled in the topology, but are predominantly
369 nested in subfamilies 3, 7, 10 and 12. The other Micropia subfamilies are mostly
370 comprised by sequences within species of the *Sophophora* subgenus, especially by
371 sequences within the *melanogaster* group. Interestingly, sequences of Micropia used by
372 [15] are distributed across nine of the 20 subfamilies here established, showing the wide
373 diversity of Micropia sequences in Drosophilidae species genomes.

374

375 **Micropia evolutionary history**

376 In addition to this pattern of high diversity, our data also show that the
377 evolutionary history of Micropia retroelement in *Drosophila* is characterized by several
378 VTTs and HTTs events. Although VTTs may comprise the predominant form of
379 transmission (94-98% of the events), HTTs is clearly an important way these genomic
380 parasites have to evade genomic extinction [17, 18]. In our data, the evidence for HTT
381 in Micropia evolution came from three main sources: the patchy distribution within

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

382 Drosophilidae phylogeny, the incongruence between Micropia and species phylogenies,
383 and the significantly lower dS values presented by some Micropia sequences in
384 comparison to nuclear host genes [17, 26]. In the first line of evidence, PCR and Dot-
385 Blot analyses provided some interesting results, especially when they were evaluated
386 considering the results obtained through genomic data, aiming to get inferences about
387 presence/absence patterns along the Drosophilidae phylogeny. Sequence analysis was
388 further performed using amino acid data to reconstruct the Micropia phylogenetic
389 relationships and using codon-aligned nucleotide data in order to measure synonymous
390 distances. This whole set of results enabled to envision a hypothesis about the evolution
391 of Micropia sequences within Drosophilidae.

392 The *cardini* group species was the best-represented *Drosophila* group in our
393 analysis, and 80% of its species had their genome analyzed (12 from the 15 described
394 species; [67]). Of these, eight species presented Micropia sequences. Conversely, the
395 *melanogaster* and the *repleta* groups, for which several species have sequenced
396 genomes, presented the higher percentage of species containing Micropia copies on the
397 genomes here analyzed (100%). The number of isolated sequences is generally higher
398 for species of the last groups, for which whole genome sequences are frequently
399 available. Nevertheless, the use of *in vitro* methodologies to investigate the presence of
400 TEs in non-model group species revealed here an important strategy to establish a
401 robust evolutionary hypothesis for the element. For example, using such methodologies
402 we were able to identify the absence of Micropia copies in the genome of several
403 species belonging to distinct groups (*funnebris*, *guaramunu*, *guarani*, *immigrans* and
404 *tripunctata*), confirming, therefore, the patchy distribution of Micropia in the
405 *Drosophila* subgenus.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

406 The *cardini* group species showed an interesting Micropia distribution pattern.
407 Micropia sequences are present only in the genome of species occurring in the
408 mainland, from south North America to southern South America [68]. The other four
409 species, *D. arawakana*, *D. dunni*, *D. nigrodunni* and *D. similis*, which seem to be
410 devoid of Micropia (S1 Fig), are endemic to the Caribbean islands [68]. The clustering
411 of the Micropia sequences presented by the mainland *cardini* species and their
412 straightforward similarity in amino acid sequences suggest the element has invaded the
413 genome of these species around 1.5Mya, which is quite earlier than the divergence
414 times estimated for the target species (4 - 35 Mya, as estimated by [51]). Considering
415 this, it is interesting to note that 73% (8 of 11) of the Micropia RT sequences analyzed
416 for the *cardini* group species seem to be potentially capable to code for reverse
417 transcriptase enzyme, which is also an evidence in favour of a recent invasion. This
418 invasion apparently occurred through multiple HTTs, as can be inferred through the
419 comparison of pairwise Micropia dS values and orthologous nuclear genes dS values.
420 This methodology is able to detect HTTs between closely-related species [29]. In fact,
421 all the 51 comparisons involving only species of the *cardini* group showed significantly
422 lower dS values for Micropia than for any of the three evaluated nuclear genes.
423 Nevertheless, although several HTTs events seem to have occurred between species of
424 the *cardini* group, it is quite probable that the ancestor sequence of this group came
425 from a species belonging to the *repleta* group (or another related group not analyzed
426 here), for which at least some sequences of subfamily 7 seem to have evolved through
427 VTTs. This can be seen, for example, by the absence of rejection of the null hypothesis
428 of VTT in the comparison of dS values between the sequences Dhydei_X13304 and
429 Dbuzzatti_04_2 and those of the host nuclear genes. This pattern is also corroborated by
430 [39].

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

431 Several other HTTs might also have occurred within the *melanogaster* group
432 (53.3% of potentially coding sequences) and evidence for these can be found within
433 subfamilies 1, 4, 10, 11 and 14. In subfamily 10, for example, the Micropia copies in *D.*
434 *melanogaster*, *D. simulans* and *D. sechellia* genomes are identical, suggesting recent
435 events of HTTs. Conversely, in subfamily 1, there are clear incongruences between
436 Micropia and species phylogeny, and a sequence encountered in *D. suzukii* may have
437 been recently transferred to *D. rhopaloa*, given the earlier branching of the Micropia
438 sequences from *D. suzukii* genome, and this event occurred around 5 Mya. In fact, these
439 species are included in different subgroups of the *melanogaster* group, for which
440 divergence times at the same divergence level are older than 10Mya [46].
441 Interestingly, signals of HTTs are less straightforward among species of the
442 *repleta* group, and despite the presence of sequences nested in different Micropia
443 subfamilies, only subfamily 7 presents some evidence of HTT involving *D. hydei*, *D.*
444 *buzzatii* and *D. mercatorum*. Such events were dated to approximately 1.25 Mya, which
445 is quite more recent than the divergence times estimated for these species (4-16 Mya –
446 51]). Interestingly, there are two common features between these events and those
447 presented above for the *cardini* group: also here multiple HTTs can be inferred, and
448 these lie in the same time confidence interval as those discussed above. Moreover, all
449 the evaluated species of both, the *cardini* and the *repleta* groups occur in the Neotropics
450 [67], which faced severe climatic oscillations during this period [69]. Since it was
451 already shown that these events possibly changed the distribution of several species of
452 *Drosophila* [70. 71], they may have led to several secondary contacts which created the
453 necessary conditions for HTT.
454 All the HTTs discussed so far occurred between closely related species,
455 comprising the same species group. According to [16], it is expected that the more

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

456 species sampled within a group, the more HTT events will be discovered, since
457 retrotransposons show low HTT rates between distantly related lineages. Nevertheless,
458 considering the dS comparisons performed within each of the Micropia subfamilies, in
459 association to the incongruences between species and Micropia phylogenies, we were
460 also able to hypothesize the occurrence at least seven other HTTs involving species
461 from distinct *Drosophila* groups or even distinct subgenera, as follow:
462 - Subfamily 3: since this subfamily is widely spread in the genome of species from the
463 subgenus *Siphlodora*, there must have occurred one HTT from one species of the
464 *Siphlodora* subgenus to *D. suzukii*, the only species of the *melanogaster* group with
465 sequences belonging to this Micropia subfamily;
466 - Subfamily 7: the sequences Dhydei_X13304 and X13305 do not present signals of
467 HTT with Dbuzzatti_04_2, so these sequences might be the presumably ancestral copies
468 within this subfamily. In this way, besides the HTTs within the *cardini* and *repleta*
469 groups discussed above, and that from one species of the *repleta* group (possibly *D.*
470 *hydei*) to another species of the *cardini* group, there might have occurred at least one
471 HTT from *D. buzzatii* to *D. willistoni*.
472 - Subfamily 11: as Damericana_121 does not show signals of HTT comparing with
473 Dbusckii_03, they might represent ancestral sequences. In this way, it might have
474 occurred at least one HTT to species of the *melanogaster* group.
475 - Subfamily 12: given the absence of HTTs signals among several species of the
476 *melanogaster* group, as well as among species of the *Siphlodora* subgenus, most of
477 these copies possibly evolved through VTT since the most recent common ancestor
478 (MRCA) of both lineages. Nevertheless, there is evidence of one HTT presumably from
479 *D. sechellia* to *D. willistoni*, one from *D. ananassae* to *D. albomicans*, and one
480 involving the MRCA of the *melanogaster* and *Siphlodora* lineages.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

481 - Subfamily 14: this Micropia subfamily is widespread in the *melanogaster* group, from
482 which a HTT presumably occurred to *D. americana*.

483 In conclusion, the Micropia evolutionary history is based on VTTs and HTTs
484 events with a high diversification of sequences leading to the distinct subfamilies here
485 detected, with some sequences still capable to encode RT enzyme. Moreover, species
486 from the *repleta* and *melanogaster* group seem to have played an important role in most
487 HTT events inferred here within *Drosophila*. The wide distribution range occupied by
488 some species of these groups possibly contributed to these phenomena, by providing
489 more chances to HTT due to the overlapped distribution with other species [16].

490

491 **Acknowlegements**

492 We thank Dr. Emmanuele Lerat for kindly provide the sequences belonging to
493 Micropia/Sacco group of Ty3/Gypsy retroelements; Christopher McAllester for
494 suggestions on early version of the manuscript; and all researchers from the Laboratório
495 de Drosophilidae at UFRGS. This study was funded by the Brazilian institutions
496 FAPERGS, CNPq and CAPES.

497

498 **References**

499 1. Pardue M-L, DeBaryshe PG. Retrotransposons that maintain chromosome ends.
500 Proceedings of the National Academy of Science. 2011; 108:20317–24.
501 2. Jangam D, Feschotte C, Betran E. Transposable element domestication as an
502 adaptation to evolutionary conflicts. Trends in Genetics. 2017; 33:817-831
503 3. Joly-Lopez Z, Bureau TE. Exaptation of transposable element coding sequences.
504 Current Opinion in Genetics & Development. 2018; 49:34–42.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

505 4. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al.
506 Ten things you should know about transposable elements. Genome Biology. 2018;
507 19(1), 199.
508 5. Volff JN. Turning junk into gold: domestication of transposable elements and the
509 creation of new genes in eukaryotes. Bioessays. 2006; 28(9), 913-922.
510 6. Sinzelle L, Izsvak Z, Ivics Z. Molecular domestication of transposable elements:
511 from detrimental parasites to useful host genes. Cellular and Molecular Life
512 Sciences. 2009; 66(6), 1073-1093.
513 7. Feschotte C. Transposable elements and the evolution of regulatory networks. Nature
514 Reviews Genetics. 2008; 9(5), 397.
515 8. Lee HE, Ayarpadikannan S, Kim HS. Role of transposable elements in genomic
516 rearrangement, evolution, gene regulation and epigenetics in primates. Genes &
517 Genetic Systems. 2016; 15-00016.
518 9. Loreto ELS, Deprá M, Diesel JF, Panzera Y, Valente VLS. *Drosophila* relics hobo
519 and hobo-MITEs transposons as raw material for new regulatory
520 networks. Genetics and Molecular Biology. 2018; 41(1), 198-205.
521 10. Muotri AR, Marchetto MC, Coufal NG, Gage FH. The necessary junk: new
522 functions for transposable elements. Human Molecular Genetics. 2007; 16(R2),
523 R159-R167.
524 11. Tsushima A, Gan P, Kumakura N, Narusaka M, Takano Y, Narusaka Y, Shirasu K.
525 Genomic plasticity mediated by transposable elements in the plant pathogenic
526 fungus *Colletotrichum higginsianum*. Genome Biology and Evolution. 2019; 11(5):
527 1487-1500
528 12. Kidwell MG. Transposable elements and the evolution of genome size in
529 eukaryotes. Genetica. 2002; 115(1), 49-63.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

530 13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial
531 sequencing and analysis of the human genome. Nature. 2001; 409, 860–921
532 14. Schnable PS et al. The B73 maize genome: complexity, diversity, and dynamics.
533 Science. 2009; 326, 1112–1115
534 15. Bargues N, Lerat E. Evolutionary history of LTR retrotransposons among 20
535 *Drosophila* species Mobile DNA. 2017; 8:7.
536 16. Peccoud J, Loiseau V, Cordaux R, Gilbert C. Massive horizontal transfer of
537 transposable elements in insects. Proceedings of the National Academy of
538 Sciences. 2017; 114(18), 4721-4726.
539 17. Schaack S, Gilbert C, Feschotte C. Promiscuous DNA: horizontal transfer of
540 transposable elements and why it matters for eukaryotic evolution. Trends in
541 Ecology & Evolution. 2010; 25(9), 537-546
542 18. Gilbert C, Feschotte C. Horizontal acquisition of transposable elements and viral
543 sequences: patterns and consequences. Current Opinion in Genetics &
544 Development. 2018; 49:15–24
545 19. Maumus F, Fiston-Lavier A-S, Quesneville H. Impact of transposable elements on
546 insect genomes and biology. Current Opinion in Insect Science. 2015; 7, 1-7.
547 20. McDonald JF. Evolution and consequences of transposable elements. Current
548 Opinion in Genetics and Development. 1993; 3: 855-864.
549 21. Capy P, Gasperi G, Biémont C, Bazin C. Stress and transposable elements: co-
550 evolution or useful parasites? Heredity. 2000; 85, 101–106.
551 22. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements:
552 from conflicts to benefits. Nature Reviews Genetics. 2017; 18, 71–86

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

553 23. El Baidouri M, Carpentier M-C, Cooke R, Gao D, Lasserre E, Llauro C, Mirouse M,
554 Picault N, Jackson SA, Panaud O. Widespread and frequent horizontal transfers of
555 transposable elements in plants. Genome Res. 2014; 24:831–838
556 24. Ivancevic AM, Walsh AM, Kortschak RD, Adelson DL. Jumping the fine LINE
557 between species: Horizontal transfer of transposable elements in animals catalyses
558 genome evolution. BioEssays. 2013; 35:1071–1082
559 25. Wallau GL, Capy P, Loreto E, Le Rouzic A, Hua-Van A. VHICA, a new method to
560 discriminate between vertical and horizontal transposon transfer: Application to the
561 mariner family within Drosophila. Mol Biol Evol. 2016; 33:1094–1109.
562 26. Loreto ELS, Carareto CM, Capy P. Revisiting horizontal transfer of transposable
563 elements in *Drosophila*. Heredity. 2008; 100, 545–554
564 27. Silva JC et al. Factors that affect the horizontal transfer of transposable elements.
565 Curr. Issues Mol. Biol. 2004; 6, 57–71
566 28. Wallau GL, Ortiz MF, Loreto ELS. Horizontal transposon transfer in eukarya:
567 detection, bias, and perspectives. Genome Biology and Evolution. 2012; 4(8), 801-
568 811.
569 29. Bartolomé C, Bello X, Maside X. Widespread evidence for horizontal transfer of
570 transposable elements across *Drosophila* genomes. Genome Biol. 2009; 10:R22
571 30. Capy P, Bazin C, Higuet D, Langin T. 1st ed. Dynamic and Evolution of
572 Transposable Elements. RG Landes Company, Austin. 1997.
573 31. Kim A, Terzian C, Santamaria P, Pélisson A, Purd'homme N, Bucheton A.
574 Retroviruses in invertebrates: the Gypsy retrotransposon is apparently an infectious
575 retrovirus of *Drosophila melanogaster*. Proc Natl Acad Sci. 1994; 91:1285–9.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

576 32. Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG. An env-like protein
577 encoded by a *Drosophila* retroelement: evidence that Gypsy is an infectious
578 retrovirus. Genes Dev. 1994; 8:2046–57
579 33. Huijser P, Kirchhoff C, Lankenau D-H, Hennig W. Retrotransposon-like sequences
580 are expressed in Y chromosomal lampbrush loops of *Drosophila hydei*. J Mol Biol.
581 1988; 203:689–697
582 34. Lankenau D-H, Huijser P, Jansen E, Miedema K, Hennig W. Micropia: a
583 retrotransposon of *Drosophila* combining structural features of DNA viruses,
584 retroviruses and non-viral transposable elements. J Mol Biol. 1988; 2:233–246
585 35. Lankenau D-H, Huijser P, Jansen E, Miedema K, Hennig W. DNA sequence
586 comparison of Micropia transposable elements from *Drosophila hydei* and
587 *Drosophila melanogaster*. Chromosoma. 1990; 99:111–117
588 36. Almeida LM, Castro JP, Carareto CMA. Micropia transposable element occurrence
589 in *Drosophila* species of the *saltans* group. DIS. 2001; 84:114–117
590 37. Almeida LM, Carareto CMA. Identification of two subfamilies of Micropia
591 transposable element in species of the *repleta* group of *Drosophila*. Genetica. 2004;
592 121:155–164
593 38. Almeida LM, Carareto CMA. Sequence heterogeneity and phylogenetic
594 relationships between the Copia retrotransposon in *Drosophila* species of the
595 *repleta* and *melanogaster* groups. Genet Sel Evol. 2006; 38:535–550
596 39. Cordeiro J, Robe LJ, Loreto EL, Valente VL. The LTR retrotransposon Micropia in
597 the *cardini* group of *Drosophila* (Diptera: Drosophilidae): a possible case of
598 horizontal transfer. Genetica. 2008; 134(3): 335--344.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

599 40. Setta N, Van-Sluys MA, Capy P, Carareto CM. Multiple invasions of Gypsy
600 andMicropia retroelements in genus *Zaprionus* and *melanogaster* subgroup of the
601 genus *Drosophila*. BMC Evolutionary Biology. 2009; 9(1), 279.
602 41. Lankenau D-H. The retrotransposon family Micropia in *Drosophila* species. In:
603 McDonald J (eds) Transposable elements and evolution. 1993. Kluwer Publishers,
604 Amsterdam pp 232–241
605 42. Lankenau S, Corces GV, Lankenau D-H. The *Drosophila* Micropia retrotransposon
606 encodes a testis-specific antisense RNA complementary to reverse transcriptase.
607 Mol Biol Evol. 1994. 17:1542–1557
608 43. Yassin, A. Phylogenetic classification of the Drosophilidae Rondani (Diptera): the
609 role of morphology in the postgenomic era. Sys. Entomol. 2013; 38: 349–364.
610 44. Sassi AK, Herédia FO, Loreto ELS, Valente VLS, Rohde C. Transposable elements
611 P and gypsy in natural populations of *Drosophila willistoni*. Genet Mol Biol. 2005;
612 28:734–739
613 45. Staden R. The Staden sequence analysis package. Mol. Biotechnol. 1996; 5:233–
614 241
615 46. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular
616 Evolutionary Genetics Analysis Version 6.0. Molecular Biology and Evolution.
617 2013; 30: 2725-2729.
618 47. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high
619 throughput, Nucleic Acids Research. 2004; 32(5), 1792-1797.
620 48. Miller MA, Pfeiffer W, Schwartz T. (2010) Creating the CIPRES Science Gateway
621 for inference of large phylogenetic trees. Proceedings of the Gateway Computing
622 Environments Workshop (GCE), 14 Nov. 2010, New Orleans, LA pp 1 – 8.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

623 49. Brisson JA, Wilder J, Hollocher H. Phylogenetic analysis of the *cardini* group of
624 *Drosophila* with respect to changes in pigmentation. Evolution. 2006; 60:1228–
625 1241
626 50. Gao J, Watabe H, Aotsuka T, Pang J, Zang Y. Molecular phylogeny of the
627 *Drosophila obscura* species group, with emphasis on the Old World species BMC
628 Evolutionary Biology. 2007, 7:87
629 51. Robe LJ, Loreto EL, Valente VL. Radiation of the Drosophila subgenus
630 (Drosophilidae, Diptera) in the Neotropics. Journal of Zoological Systematics and
631 Evolutionary Research. 2010; 48(4), 310-321.
632 52. Robe LJ, ValenteVL, Loreto EL. Phylogenetic relationships and macro-evolutionary
633 patterns within the *Drosophila tripunctata* “radiation” (Diptera: Drosophilidae).
634 Genetica. 2010; 138(7), 725-735.
635 53. Yang Y, Hou Z-C, Qian Y-H, Kang H, Zeng Q-T. Increasing the data size to
636 accurately reconstruct the phylogenetic relationships between nine subgroups of the
637 Drosophila melanogaster species group (Drosophilidae, Diptera). Molecular
638 Phylogenetics and Evolution. 2012; 62: 214-223
639 54. Seetharam AS, Stuart GW. Whole genome phylogeny for 21 *Drosophila* species
640 using predicted 2b-RAD fragments. PeerJ. 2013; 1:e226
641 55. O’Grady PM, DeSalle R. Phylogeny of the genus *Drosophila*. Genetics. 2018;
642 209(1), 1-25.
643 56. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and
644 nonsynonymous nucleotide substitutions. Mol Biol Evol. 1986; 3:418–426
645 57. R Core Team. R: A Language and Environment for Statistical Computing. 2018.
646 58. Sharp PM, Li WH. On the rate of DNA sequence evolution in *Drosophila*. Journal
647 of Molecular Evolution. 1989; 28(5), 398-402.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

648 59. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends
649 Genet. 1989; 5, 103–107.
650 60. Piégu B, Bire S, Arensburger P, Bigot Y. A survey of transposable element
651 classification systems–a call for a fundamental update to meet the challenge of their
652 diversity and complexity. Molecular Phylogenetics and Evolution. 2015; 86, 90-
653 109.
654 61. Lohe AR, Moriyama EN, Lidholm DA, Hartl DL. Horizontal transmission, vertical
655 inactivation, and stochastic loss of mariner-like transposable elements. Molecular
656 Biology and Evolution. 1995; 12(1), 62-72.
657 62. Clark JB, Kidwell MG. A phylogenetic perspective on P transposable element
658 evolution in *Drosophila*. Proc. Natl. Acad. Sci. U.S.A. 1997; 94(21): 11428-11433.
659 63. Herédia FO, Loreto ELS, Valente VLS. Complex evolution of gypsy in drosophilid
660 species. Mol Biol Evol. 2004; 21:1–12
661 64. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified
662 classification system for eukaryotic transposable elements. Nature Reviews
663 Genetics. 2007; 8(12), 973.
664 65. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements
665 in eukaryotic genomes. Mobile DNA. 2015; 6: 11
666 66. Arkhipova IR. Using bioinformatic and phylogenetic approaches to classify
667 transposable elements and understand their complex evolutionary histories. Mobile
668 DNA. 2017; 8(1), 19.
669 67. Bächli G. Taxodros database: The database on taxonomy of Drosophilidae. URL:
670 http://www.taxodros.unizh.ch/. Last accessed in April 2019

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

671 68. Heed WB, Russel JS. Phylogeny and population structure in island and continental
672 species of the cardini group of *Drosophila* studied by inversion analysis. Univ
673 Texas Publs. 1971; 7103:91–130
674 69. Souza CRG. Quaternário do Brasil. 1st ed. Holos Editora. 2005.
675 70. Franco FF, Manfrin MH. Recent demographic history of cactophilic Drosophila
676 species can be related to Quaternary palaeoclimatic changes in South
677 America. Journal of Biogeography. 2013; 40(1), 142-154.
678 71. Cenzi de Ré F, Gustani EC, Oliveira APF, Machado LP, Mateus RP, Loreto ELS,
679 & Robe LJ. Brazilian populations of Drosophila maculifrons (Diptera:
680 Drosophilidae): low diversity levels and signals of a population expansion after the
681 Last Glacial Maximum. Biological Journal of the Linnean Society. 2014. 112(1),
682 55-66.
683

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Supporting information

S1 Fig. In vitro searches for Micropia within genomes. A: PCR-blot results of species from the *cardini* and *repleta* groups. B: Dot-blot on genomic DNA confirming the pattern seen on the PCR-blot. In both cases, the probe used was an 812bp PCR fragment from *D. hydei* dhMiF2 sequence. Control: 5µl (in 10 µl) of Micropia probe.

S2 Fig. Bayesian phylogenetic tree of all Micropia sequences recovered by our searches within the Drosophilidae species analyzed in this study. The phylogenetic tree was based on amino acid sequences following a mixed evolution model with gamma correction. Bargues and Lerats´s sequences [15] were included in the analysis. The posterior probability of each clade is indicated beside its respective internal branch.

S1 Table. GenBank accession numbers of nuclear genes used in the dS analysis. - : data not available.

S2 Table. Micropia retroelement related sequences retrieved through *in silico* searches. Species scaffold: represents the scaffold in the species genome where Micropia sequence was found. First nt: first nucleotide in the scaffold where Micropia RT sequence homologous to our query was detected. Last nt: last nucleotide in the scaffold where Micropia RT sequence homologous to our query was detected. Methodology: database and *in silico* search methodology used to find the Micropia best match query.

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

S3 Table. Amino acid genetic distances between sequences belonging to the same Micropia subfamily. Data for each subfamily are in distinct sheet in the Excel file.

S4 Table. Potentially coding sequences and their respective coding frame. Sequences presenting stop codons are represented by a dash (-). The involvement in HTT was identified by the Fisher exact test (see Table S5)

S5 Table. Parwise comparative analysis of dS values between Micropia and *Adh*, *Amd* and *Ddc* nuclear gene sequences. Comparisons suggesting horizontal transposon transfer events were statistically tested by one-sided tail Fisher's exact test (Ost). Colors represent the p values lower than 0.05 (see Fig 2) to: $Ost_{Micropia\text{-}Adh}$ (orange), $Ost_{Micropia\text{-}Amd}$ (pink) and $Ost_{Micropia\text{-}Ddc}$ (purple).

bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

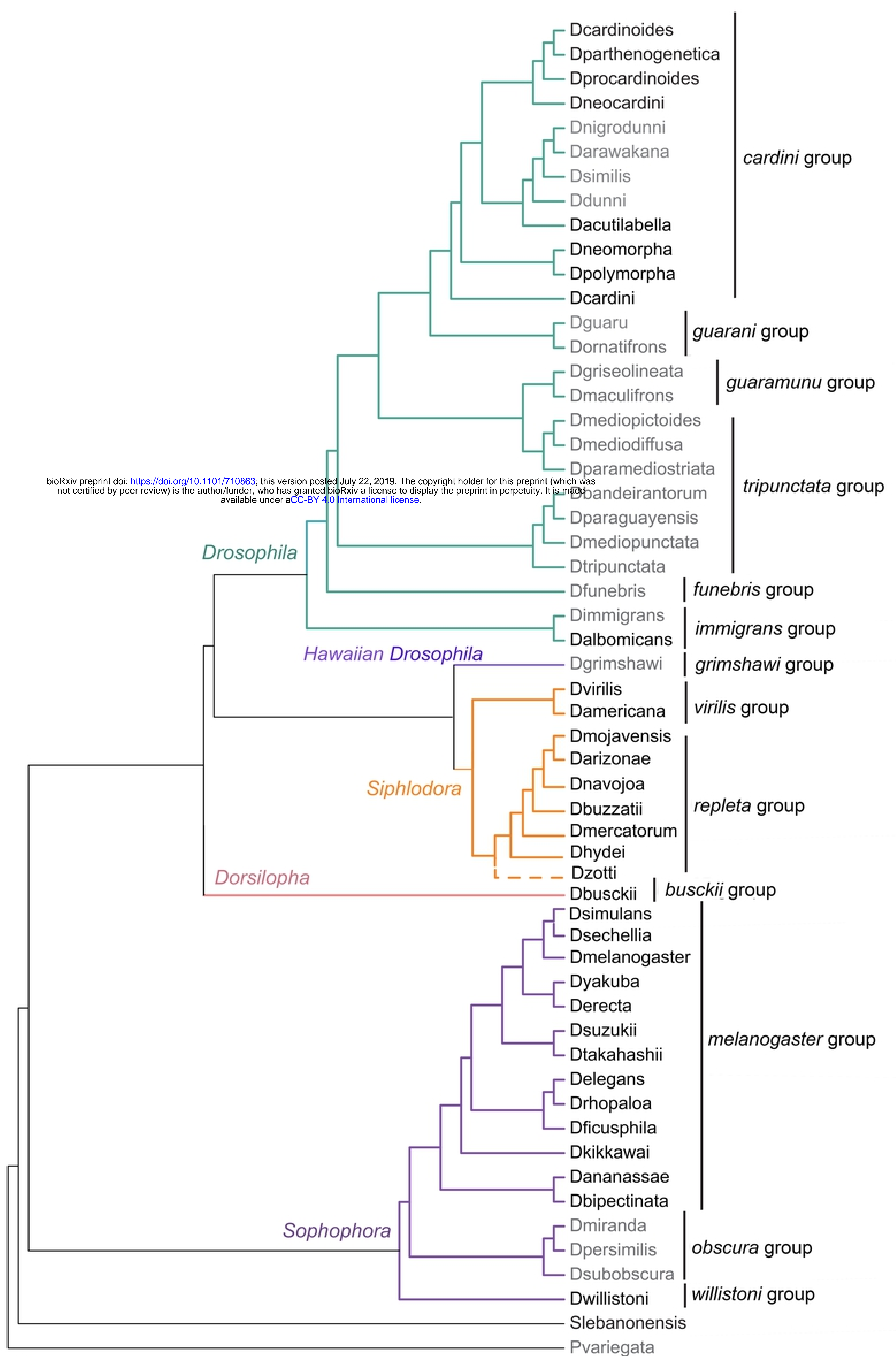
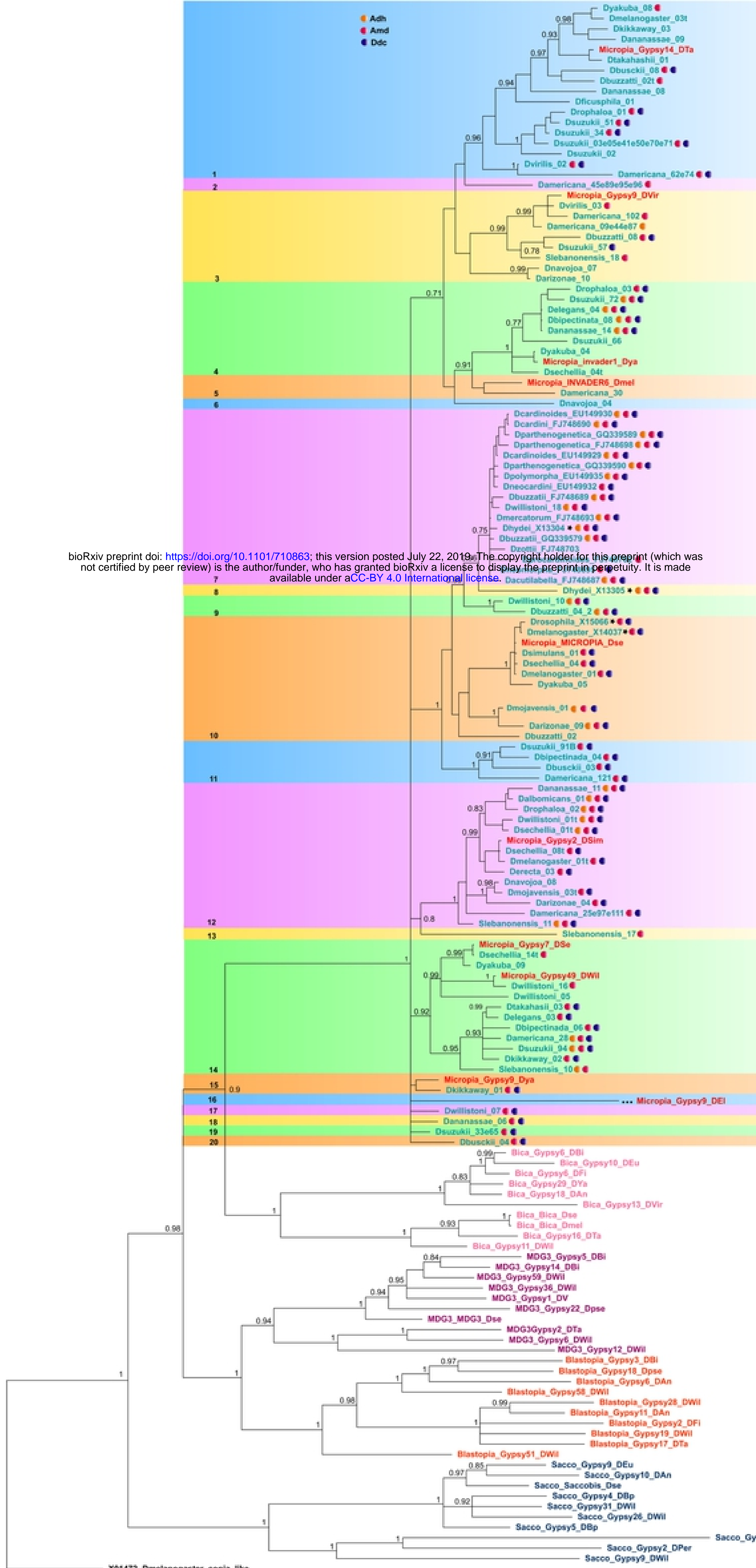


Fig 1



bioRxiv preprint doi: https://doi.org/10.1101/710863; this version posted July 22, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Fig 2