

1

2

3 **Title:**

4 The worldwide invasion of *Drosophila suzukii* is accompanied by a large increase of  
5 transposable element load and a small number of putatively adaptive insertions

6 **Authors:**

7 Vincent Mérel<sup>1</sup>, Patricia Gibert<sup>1</sup>, Inessa Buch<sup>1</sup>, Valentina Rodriguez Rada<sup>1</sup>, Arnaud Estoup<sup>2</sup>,  
8 Mathieu Gautier<sup>2</sup>, Marie Fablet<sup>1</sup>, Matthieu Boulesteix<sup>1\*</sup>, Cristina Vieira<sup>1\*</sup>

9 **\* co-corresponding authors:**

10 [matthieu.boulesteix@univ-lyon1.fr](mailto:matthieu.boulesteix@univ-lyon1.fr)

11 [cristina.vieira@univ-lyon1.fr](mailto:cristina.vieira@univ-lyon1.fr)

12 **Affiliations:**

13 <sup>1</sup> Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive  
14 UMR 5558, F-69622 Villeurbanne, France

15 <sup>2</sup> CBGP, Univ Montpellier, CIRAD, INRAE, Institut Agro, IRD, Montpellier, France

## 16 Abstract

17 Transposable Elements (TEs) are ubiquitous and mobile repeated sequences. They are major  
18 determinants of host fitness. Here, we portrayed the TE content of the spotted wing fly  
19 *Drosophila suzukii*. Using a recently improved genome assembly, we reconstructed TE  
20 sequences *de novo*, and found that TEs occupy 47% of the genome and are mostly located in  
21 gene poor regions. The majority of TE insertions segregate at low frequencies, indicating a  
22 recent and probably ongoing TE activity. To explore TE dynamics in the context of biological  
23 invasions, we studied variation of TE abundance in genomic data from 16 invasive and six  
24 native populations (of *D. suzukii*). We found a large increase of the TE load in invasive  
25 populations correlated with a reduced Watterson estimate of genetic diversity  $\hat{\theta}_w$  a proxy of  
26 effective population size. We did not find any correlation between TE contents and bio-climatic  
27 variables, indicating a minor effect of environmentally induced TE activity. A genome-wide  
28 association study revealed that ca. 5,000 genomic regions are associated with TE abundance.  
29 We did not find, however, any evidence in such regions of an enrichment for genes known to  
30 interact with TE activity (e.g. transcription factor encoding genes or genes of the piRNA  
31 pathway). Finally, the study of TE insertion frequencies revealed 15 putatively adaptive TE  
32 insertions, six of them being likely associated with the recent invasion history of the species.

34 **Key words:** *Drosophila suzukii*, Transposable Elements, Biological Invasion, Populations,  
35 Adaptation, PoolSeq.

## 36 Introduction

37 Transposable Elements (TEs) are selfish genetic elements. Despite being mostly neutral  
38 or deleterious, they persist and proliferate in populations by copying and pasting themselves in  
39 genomes (Doolittle and Sapienza 1980; Orgel and Crick 1980; Charlesworth and Charlesworth  
40 1983). The interest for those sequences considerably rose in the 2000's, with the discovery of  
41 some TE insertions having a functional, and potentially adaptive, effect on their host (Mi et al.  
42 2000; Daborn et al. 2002; Niu et al. 2019). The parallel completion of the first sequencing  
43 projects confirmed TE ubiquity and largely contributed to the growing interest for such  
44 sequences (C. elegans Sequencing Consortium 1998; 2000; Lander et al. 2001; 2002;  
45 Schnable et al. 2009).

46 The nature and intensity of TE deleterious effects may vary with their genomic localization  
47 (Mérel et al. 2020). First, TEs close to genes can alter their function. Second, TEs in highly  
48 recombining regions, are more likely to promote ectopic recombination, *i.e.* recombination  
49 between more-or-less identical sequences inserted at different locations in the genome. Third,  
50 recessive deleterious TEs are more likely to impact fitness when located on a chromosome in a  
51 hemizygous state (*e.g.* the X chromosome in males in a XY sex determination system). The  
52 strength of selection acting against TEs hence depends on the genomic region and may result  
53 in a local variation of TE density. In agreement with such expectations, TE density was found to  
54 be negatively correlated with gene density and local recombination rate in several species  
55 (Boissinot et al. 2001; Bartolomé et al. 2002). On the other hand, studies focusing on the *D.*  
56 *melanogaster* genome did not reveal a systematic lower TE content on the X-chromosome,  
57 which is hemizygous in males (Kofler et al. 2012; Cridland et al. 2013).

58 TE insertion frequencies reflect both TE activity and the selection acting upon them. Low  
59 frequency TE insertions are likely to be recent, or strongly selected against, or both. Conversely,  
60 high frequency TE insertions are likely to be old and only weakly subjected to purifying  
61 selection. As mentioned previously, TEs that are in the vicinity of genes and/or located in highly  
62 recombining regions are expected to be selected against. Accordingly, TE insertion frequencies  
63 were found to be negatively correlated with recombination rate and distance to the nearest gene  
64 in *D. melanogaster* (Kofler et al. 2012). In *Drosophila*, the overall distribution of TE frequencies  
65 seems compatible with an active repeatome (Kofler, Nolte, et al. 2015; Hill 2019) For example  
66 80% of the insertions have a frequency lower than 0.2 in *D. melanogaster* and its close relative  
67 *D. simulans* (Kofler, Nolte, et al. 2015).

68 Between population variation of TE content has been reported in various intraspecific studies.  
69 So far, the factors underlying such differences remain unclear. The effective population size ( $N_e$ )  
70 may play a prominent role in modulating TE contents. Considering that TEs are mostly  
71 deleterious, and that small  $N_e$  leads to a less efficient purifying selection, small  $N_e$  should be  
72 associated with high TE content (Lynch and Conery 2003). In support for this hypothesis Lynch  
73 & Connery (2003) found a significant correlation between genome size and estimates of the  
74 scaled mutation rate  $\theta=N_e\mu$  (with  $\mu$  the mutation rate) across populations representative of  
75 various species. At the intraspecific level, if a higher TE content in some populations has  
76 sometimes been suggested to result from a reduction of their  $N_e$  (García Guerreiro et al. 2008;  
77 García Guerreiro and Fontdevila 2011; Talla et al. 2017), to our knowledge the above expected

78 correlation has not been reproduced at this evolutionary scale. Variation in TE content may also  
79 rely on changes in TE activity in relation with the environment (Vieira et al. 1999; Stapley et al.  
80 2015). In *Drosophila*, several laboratory experiments suggest that TE activity may respond to  
81 the environment (García Guerreiro 2012; Horváth et al. 2017), but *in natura* studies considering  
82 the whole repeatome remain rare and a possible confounding effect of the demographic history  
83 cannot be excluded (Lerat et al. 2019). Finally, the host genotype may explain intraspecific  
84 variation of TE abundance. For instance, in *Drosophila*, several studies found different levels of  
85 activity among isogenic lines (Biémont et al. 1987; Pasyukova and Nuzhdin 1993; Díaz-  
86 González et al. 2011).

87 The study of intraspecific variations in TE content and the underlying determining factors is  
88 valuable as TEs may also be important for adaptation (Daborn et al. 2002; Van't Hof et al. 2016;  
89 Niu et al. 2019). Although some TE insertions exhibit a strong signal of positive selection and  
90 have been thoroughly validated experimentally, only few studies aimed at identifying putatively  
91 adaptive insertions at a genome-wide level (González et al. 2008; Li et al. 2018; Rishishwar et  
92 al. 2018; Rech et al. 2019). In addition, most of these studies deal with *D. melanogaster*  
93 (González et al. 2008; González et al. 2010; Blumenstiel et al. 2014; Rech et al. 2019). The most  
94 comprehensive of these studies analyzed genomic data on 60 worldwide natural *D.*  
95 *melanogaster* populations and reported 57 to 300 putatively adaptive insertions (depending on  
96 the degree of evidence considered) among the ~800 polymorphic insertions identified in the  
97 reference genome (Rech et al. 2019). Considering that approximately twice as many non  
98 reference TE insertions as reference insertions may segregate in a single population (Kofler et  
99 al. 2012), quite a high number of TE-induced adaptations is therefore expected. However, it  
100 remains unclear how important TEs are as substrates of adaptation considering the paucity of  
101 studies and their focus on reference genome insertions.

102 Invasive species provide a unique opportunity to study the combined effect of *in natura*  $N_e$   
103 variations and environmental variations both on TE abundance and TE adaptive potential.  
104 Invasive populations often go through demographic bottlenecks allowing to test for an effect of  
105  $N_e$  on TE abundance (Estoup et al. 2016). Individuals from invasive populations also encounter  
106 new environmental conditions, allowing to test for an effect of bio-climatic variables on TE  
107 abundance. Because of the need of colonizing individuals to adapt to new environmental  
108 conditions, biological invasions are often used to study rapid contemporary adaptation  
109 (Lavergne and Molofsky 2007; Rollins et al. 2015). Yet, the particular role of TEs in the rapid

110 adaptation of invasive species remains speculative. In particular, TEs have been proposed to  
111 explain, at least in part, the paradox of invasive species, *i.e.* the successful adaptation to a new  
112 environment despite a reduced genetic diversity caused by small founder population sizes  
113 (Stapley et al. 2015; Estoup et al. 2016; Marin et al. 2020). In response to environmental  
114 changes, TE sequences may be recruited and affect the expression of nearby genes.  
115 Furthermore, if a higher activity of TE is induced in response to environmental changes, the  
116 insertions could thus result in genetic variation, and potentially beneficial alleles.

117 In this paper, we focused on the spotted wing fly *D. suzukii*, a close relative of *D. melanogaster*,  
118 displaying the highest reported TE content among *Drosophila* (Sessegolo et al. 2016). *D.*  
119 *suzukii* is native from Asia and has invaded independently the American and European  
120 continents where it was introduced probably in the late 2000's (Framout et al. 2017). Using the  
121 recently released high-quality genome assembly Dsuz-WT3\_v2.0 based on Long PacBio Reads  
122 (Paris et al. 2020), we constructed a *de novo* TE database and found that TE represented 47 %  
123 of the genome. We further assessed TE insertion frequencies and TE abundance in 22  
124 worldwide populations representative of the native area (n=6) and of the two main invaded  
125 areas in Europe (n=8) and America (n=8). The study of TE frequencies showed that the  
126 repeatome is highly active in *D. suzukii*: 75% of insertion segregated at a frequency < 0.25. We  
127 found that the TE content was significantly higher in invasive populations and was correlated  
128 with a reduction of  $N_e$ . Finally, controlling for population structure, a genome scan conducted on  
129 polymorphic TE insertions identified 15 putatively adaptive TE insertions.

## 130 **Results**

### 131 **A highly repeated reference genome**

132 We found that the high-quality *D. suzukii* assembly Dsuz-WT3\_v2.0 of Paris et al. (2020)  
133 is characterized by a high TE content. Overall, 47.07 % of the reference assembly is annotated  
134 as repeated sequences (fig. 1A). In terms of genomic occupancy, LTR is the predominant TE  
135 order with more than 20% of the sequence assembly corresponding to these elements, then  
136 LINEs (8.77%), DNA elements (6.99%), and RC (6.95%). 4.07% of the assembly is occupied by  
137 unknown repeated sequences. At a lower hierarchical level, the three most represented  
138 superfamilies are *Gypsy*, *Helitron* and *Pao*, corresponding to 13.65%, 6.95% and 6.44% of the  
139 assembly, respectively (supplementary table S1). The average percentage of genomic  
140 occupancy per superfamily is 1.88%. Regarding TE copy numbers, the top three superfamilies

141 are *Helitron*, *Gypsy* and *Pao* (56,493, 39,189 and 15,555 copies, respectively) (fig. 1A). The  
142 average number of copies per superfamily is 4,963.

143 Syntenic relationships with *D. melanogaster* genome have been established for 212 of the 546  
144 contigs of *D. suzukii* assembly. A total of 241 Mb of the 268 Mb assembly have a clearly  
145 identified counterpart in the *D. melanogaster* genome (fig. 1B, supplementary table S2).  
146 Considering the observed bimodal distribution of gene density, we partitioned the *D. suzukii*  
147 assembly into gene-rich regions ( $\geq 7$  genes per 200 kb; 121.8 Mb) and gene-poor regions ( $< 7$   
148 genes per 200 kb; 108 Mb) (fig 1B, supplementary fig. S1). TE fragment density also follows a  
149 bimodal distribution: 127.4 Mb correspond to TE-rich regions ( $\geq 165$  TE fragments per 200 kb)  
150 and 102.4 Mb to TE-poor regions ( $< 165$  TE fragments per 200 kb) (fig. 1B, supplementary fig.  
151 S2). TE-rich regions are enriched in gene-poor regions, and TE-poor regions are enriched in  
152 gene-rich regions ( $\chi^2 = 786.47$ ,  $df = 1$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ). We did not find any difference in  
153 mean TE density between autosomal and X-linked contigs ( $\hat{\mu}_{\text{autosomes}} = 172.00$ ,  
154  $\hat{\mu}_{X\text{-linked}} = 151.93$ ,  $W = 78900$ ,  $p\text{-value} = 0.11$ ). This conclusion holds when comparing  
155 autosomal and X-linked contigs as defined in Paris et al. (2020) using a female-to-male read  
156 mapping coverage ratio ( $\hat{\mu}_{\text{autosomes}} = 176.11$ ,  $\hat{\mu}_{X\text{-linked}} = 150.09$ ,  $W = 79088$ ,  $p\text{-value} = 0.38$ ).  
157 However, when considering only gene-rich regions, the mean TE density was far higher for X-  
158 linked contigs ( $\hat{\mu}_{\text{autosomes}} = 65.31$ ,  $\hat{\mu}_{X\text{-linked}} = 107.54$ ,  $W = 47394$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ). Once  
159 again, this conclusion holds when using autosomal and X-linked contigs as defined by Paris et  
160 al. (2020) ( $\hat{\mu}_{\text{autosomes}} = 65.34$ ,  $\hat{\mu}_{X\text{-linked}} = 107.07$ ,  $W = 47557$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ).

## 161 **An active repeatome in the Watsonville reference population**

162 The female used to establish the WT3 isofemale strain corresponding to the genome  
163 assembly was collected in Watsonville (CA, USA) (Paris et al. 2020). To thoroughly evaluate TE  
164 activity in this reference population, we assessed TE insertion frequencies in a PoolSeq sample  
165 of 50 *D. suzukii* individuals from Watsonville. Because TEs are mostly deleterious, rare TE  
166 insertions are likely to be recent insertions, not yet eliminated by selection, whereas fixed TE  
167 insertions are presumably old insertions weakly submitted to selection. It is worth stressing that,  
168 for the study of TE frequencies and abundances, we first used simulated PoolSeq data to  
169 validate our pipelines and to evaluate their performance and their sensibility to parameters such  
170 as sequencing coverage or number of individuals (see supplementary methods for details).

171 A total of 9,256 insertions were recovered in the reference population. The frequency  
172 distribution is approximately U-shaped (fig. 2A) with a majority of insertions segregating at low  
173 frequency (N = 6934,  $f < 0.25$ ). 1,642 insertions are found at high frequency, in the reference  
174 population ( $f \geq 0.75$ ). Only a minority of insertions are of intermediate frequency (N = 680,  $0.25 \leq$   
175  $f < 0.75$ ). Among the 654 families/pseudofamilies found in the whole dataset, 473 were present  
176 in the reference population. 102 belonged to the DNA order, 98 to the LINE order, 175 to the  
177 LTR order, 46 to the RC and 52 were Unknown. Only 119 TE families/pseudofamilies presented  
178 more than 10 insertions: 25 DNA families/pseudofamilies, 32 LINEs, 32 LTR, 6 RC and 24  
179 Unknown. The vast majority of these families presented a median frequency lower than 0.25 (N  
180 = 80) (fig. 2B). Only four families displayed a median frequency between 0.25 and 0.75. Finally,  
181 35 families had a median frequency superior or equal to 0.75. We did not find evidence that the  
182 number of TE families in these categories differed between TE orders (supplementary table S3;  
183  $\chi^2 = 4.94$ ,  $df = 8$ ,  $p\text{-value} = 0.76$ ). However, the mean frequency was slightly different (  
184  $\hat{\mu}_{DNA} = 0.30$ ,  $\hat{\mu}_{LINEs} = 0.31$ ,  $\hat{\mu}_{LTR} = 0.46$ ,  $\hat{\mu}_{RC} = 0.22$ ,  $\hat{\mu}_{Unknown} = 0.16$ , Kruskal-Wallis  
185  $\chi^2 = 92.35$ ,  $df = 4$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ). TE insertion frequencies were not evenly distributed  
186 along the assembly: mean TE insertion frequency was considerably lower in gene-rich windows.  
187 ( $\hat{\mu}_{rich} = 0.13$ ,  $\hat{\mu}_{poor} = 0.72$ ,  $W = 18863$ ,  $p\text{-value} < 2.2 \times 10^{-16}$ ; supplementary fig. S3).

## 188 **Demography as driver of TE contents in *D. suzukii* populations**

189 Our estimation of TE abundance in the 22 genotyped *D. suzukii* populations (fig. 3A)  
190 indicates substantial variation across populations, with significantly more TEs in invasive than in  
191 native populations and a strong correlation with the Watterson estimate of genetic diversity  
192 obtained from SNPs corresponding to a proxy of population effective size (fig. 3B, C). The mean  
193 number of insertions per Haploid Genome (HG) and per population was 2,793, ranging from  
194 2,113 in the Chinese population CN-Nin to 3,129 in the Hawaiian population (US-Haw). There  
195 was a significant effect of the continent on the mean number of families/pseudofamilies per  
196 population: American and European populations had more families/pseudofamilies than native  
197 populations ( $\hat{\mu}_{America} = 470$ ,  $\hat{\mu}_{Europe} = 468$ ,  $\hat{\mu}_{Asia} = 453$ , Kruskal-Wallis chi-squared = 10.505,  $df = 2$ ,  
198  $p\text{-value} = 0.0052$ ). American and European populations also had more insertions per HG than  
199 native populations ( $\hat{\mu}_{America} = 3008$ ,  $\hat{\mu}_{Europe} = 2928$ ,  $\hat{\mu}_{Asia} = 2326$ , Kruskal-Wallis  $\chi^2 = 14.4$ ,  $df = 2$ ,  $p\text{-}$   
200  $\text{value} = 7.3 \times 10^{-4}$ ). We found a negative linear correlation between the total number of insertions  
201 per HG and per population and the Watterson estimate of genetic diversity obtained from SNPs  
202  $\hat{\theta}_w$ , a proxy of population effective size ( $t = -13.415$ ,  $df = 20$ ,  $p\text{-value} = 1.8 \times 10^{-11}$ , fig. 3C). The

203 variation of  $\hat{\theta}_w$  explains a large proportion of the variance in the total number of insertions per  
204 HG across the populations ( $R^2 = 0.90$ ). The correlation remains significant when considering  
205 only native populations ( $t = -5.22$ ,  $df = 4$ ,  $p\text{-value} = 6.4 \times 10^{-3}$ ), or only invasive populations ( $t = -$   
206  $3.06$ ,  $df = 14$ ,  $p\text{-value} = 8.6 \times 10^{-3}$ ), or only European populations ( $t = -5.46$ ,  $df = 6$ ,  $p\text{-value} =$   
207  $1.6 \times 10^{-3}$ ), but not when considering only American populations ( $t = -1.89$ ,  $df = 6$ ,  $p\text{-value} = 0.11$ ).  
208 The correlation between the number of insertions per HG per population and  $\hat{\theta}_w$  was also  
209 assessed individually for the 83 TE families/pseudofamilies showing an amplitude of variation  
210 superior or equal to 3 copies per HG. After a Benjamini-Hochberg correction for multiple testing,  
211 we found a significant correlation for 63 TE families ( $p\text{-adjusted} < 0.05$ ).

## 212 **Environmental and genotypic effects on TE abundance**

213 Because  $\hat{\theta}_w$  did not explain all the observed variation in TE abundance among the 22  
214 sampled populations, we tested the effect of two other factors: the environmentally induced  
215 changes in TE activity and the genetically determined changes.

216 To test for an effect of environmentally induced changes in TE activity, we used Partial Mantel  
217 tests. We tested the correlation between 19 bioclimatic variables and TE family abundance, for  
218 the 83 TE families showing an amplitude of variation superior or equal to 3 copies per HG,  
219 correcting for population structure. After correction for multiple testing we did not find any  
220 significant correlation (Benjamini-Hochberg correction for multiple testing,  $p\text{-adjusted} < 0.05$ ).

221 To evaluate the effect of genetic variation on TE abundance we performed a genome-wide scan  
222 for association using methods controlling for population structure. To that end we relied on the  
223 13,530,656 bi-allelic variants (mostly SNPs) previously described on the same data set  
224 (Olazcuaga et al. 2020) and searched for association with the population abundance of the 83  
225 TE families/pseudofamilies mentioned above using the BayPass software. Globally, we found  
226 4,856 genomic regions showing evidence of association with population abundance of at least  
227 one TE family. Each region spanned at least 1 kb on the reference assembly and included one  
228 or several significant SNP/InDel separated by less than 1 kb (significance threshold: Bayes  
229 Factor (BF)  $> 20$ ). On average each region was associated with the number of insertions per HG  
230 of 1.37 families (min=1, max=69) and contained 2.40 SNPs/InDels (min=1, max=49). 306  
231 (6.30%) regions overlapped with repeated sequences as annotated in the reference genome,  
232 which is less than expected by drawing SNP/InDel associated regions randomly ( $\hat{\mu} = 9.22\%$ ,



233  $q_{0.025}=7.60\%$ ,  $q_{0.975}=11.16\%$ ; supplementary fig. S4A). Only 14 of these regions contain a TE of  
234 the same family/pseudofamily as the TE abundance they were associated with. Regarding  
235 genes, 2,843 (58.55%) regions were associated with at least one gene, which is less than  
236 expected under random expectations ( $\hat{\mu}=66.97\%$ ,  $\text{quantile}_{0.025}=62.40\%$ ,  $\text{quantile}_{0.975}=70.76\%$ ;  
237 supplementary fig. S4B). Due to their known role in the activity of TEs, we further searched for  
238 enrichment in genes encoding transcription factors and piRNA pathway effectors among the  
239 genes located within our candidate regions. We did not observe any significant enrichment in  
240 genes encoding transcription factors (Observed: 13.63%, Expected:  $\hat{\mu}=15.68\%$ ,  
241  $\text{quantile}_{0.025}=12.00\%$ ,  $\text{quantile}_{0.975}=20.60\%$ ; supplementary fig. S4C) nor in genes involved in the  
242 piRNA pathway (Observed: 0.33%, Expected:  $\hat{\mu}=0.376\%$ ,  $q_{0.025}=0.00\%$ ,  $q_{0.975}=1.18\%$ ;  
243 supplementary fig. S4D). Among the top 10 regions, corresponding to the regions associated  
244 with the highest number of TE families/pseudofamilies, two appeared to be non genic, four  
245 could not be attributed to *D. melanogaster* genome, three were associated with the  
246 mitochondrial genome and one was associated with *blot*, a member of the sodium- and chloride-  
247 dependent neurotransmitter symporter family (<https://flybase.org/reports/FBgn0027660>).

## 248 **A small number of putatively adaptive TE insertions**

249 We investigated the presence of putatively adaptive insertions using a genome scan  
250 combining three methods controlling for population structure implemented in BayPass  
251 (Olazcuaga et al. 2020). First, we assessed overall differentiation (based on the XtX statistics).  
252 Second, we studied allele frequencies differences between two groups of populations (based on  
253 the  $C_2$  statistics ): American invasive vs native populations ( $C_2^{Am}$ ), European invasive vs native  
254 populations ( $C_2^{Eu}$ ), all invasive vs native populations ( $C_2^{WW}$ ). Third, we carried out genome-wide  
255 association with each of the 19 bioclimatic variables (based on the BF).

256 The genome scan was conducted on 7,004 polymorphic TE insertions (MAF > 0.025, 5,944  
257 autosomal insertions and 1,060 X-linked insertions treated separately). We identified a total of  
258 15 putatively adaptive insertions (13 located on autosomal and three on X-linked contigs) (table  
259 1; fig. 4). Nine of these insertions were outliers when considering the global differentiation  
260 statistics XtX. Note that their frequencies were distinct between native Chinese (low  
261 frequencies) and native Japanese populations (high frequencies). One insertion was an outlier  
262 for both the XtX and  $C_2^{Am}$  statistics. Finally, the last five insertions were outliers for the  $C_2^{WW}$   
263 statistics. No significant association was found between TE insertion frequencies and the 19  
264 bioclimatic variables investigated.

265 One of the 15 putatively adaptive insertions was close (*i.e.*, 399 bp away) to a SNP/InDel that  
266 had previously been identified in a region potentially associated with *D. suzukii* invasive success  
267 (table 1) (Olazcuaga et al. 2020). For one insertion we did not find any homologous regions in  
268 *D. melanogaster*, four others were in genomic regions without any genes, and the ten remaining  
269 were associated with genes.

270 We further investigated signatures of selection around candidate insertions by estimating local  
271 Tajima's D statistics in the SNP/InDel dataset. Low values of Tajima's D indicate an excess of  
272 rare mutations, one possible signature of a selective sweep due to positive selection. To test if  
273 each of our candidate insertions were associated with selective sweeps, we computed the linear  
274 correlation between its frequency and local Tajima's D values (supplementary fig.S5). Five  
275 statistically significant correlations were found corresponding to the insertions n°4, 9, 10, 12 and  
276 15 (Pearson's product-moment correlation,  $p < 0.05$ ). Only a single insertion was associated  
277 with an extreme local Tajima's D (insertion n°15; Tajima's D < quantile<sub>0.05</sub>), and only for a single  
278 population. The visualization of Tajima's D at a larger scale (*i.e.*, 10 kb upstream - 10 kb  
279 downstream the insertion) confirms the lack of strong effect of the investigated insertions on  
280 Tajima's D (supplementary fig. S6). It is worth noting that, if the effect of our candidate TE  
281 insertion on Tajima's D is globally low, a close investigation of Tajima's D suggests that, at least  
282 in some cases, it is the absence rather than the presence of the insertion that may be adaptive.  
283 As a matter of fact, while the correlation implying an extreme local Tajima's D was negative, the  
284 four other significant correlations between local Tajima's D and insertion frequency were  
285 positive.

## 286 Discussion

287 For most species the repeatome is still a poorly known genomic compartment and much  
288 remains to be understood regarding its variability, dynamics, functional and fitness impacts. This  
289 is all the more important given that TEs appear to be ubiquitous, prompt to invade new  
290 genomes (Kofler, Hill, et al. 2015), and they may drastically impact the host phenotype (Nikitin  
291 and Woodruff 1995; Daborn et al. 2002; Van't Hof et al. 2016). Here we capitalized on a recently  
292 generated long-reads genome assembly and a large set of populational PoolSeq data  
293 (Olazcuaga et al. 2020; Paris et al. 2020) to thoroughly portray the TE content of the non model  
294 invasive species *Drosophila suzukii*.

## 295 **An abundant, unevenly distributed and active repeatome**

296           The observed 47% of TEs in the genome of *D. suzukii* confirmed the outlier position of  
297 this species within the *Drosophila* genus regarding the global amount of TEs. Our estimate is  
298 somewhat higher than those reported in previous studies in *D. suzukii* (Chiu et al. 2013; Ometto  
299 et al. 2013; Sessegolo Camille et al. 2016; Paris et al. 2020). Considering that the assembly of  
300 repeats is often impossible using short paired-end (PE) reads (Rius et al. 2016), it is not  
301 surprising that we recovered more TEs in a long reads genomic assembly than previous studies  
302 investigating TE contents using PE reads assemblies (Chiu et al. 2013; Ometto et al. 2013). In  
303 addition, we here performed a *de novo* reconstruction of TE sequences, which allowed us to  
304 identify more TE families/pseudofamilies, as compared to the previous research work based on  
305 the same assembly (35 %) (Paris et al. 2020). Overall, *de novo* reconstruction of TE sequences  
306 from long read assemblies, such as the 15 *Drosophila* species assemblies recently generated  
307 using nanopore sequencing (Miller et al. 2018), should greatly improve our knowledge of TE  
308 diversity in *Drosophila*.

309 In agreement with the gene disruption hypothesis and observations in a variety of species  
310 (Bartolomé et al. 2002; Medstrand et al. 2002; Wright et al. 2003), we observed a depletion of  
311 TE copies in gene-rich regions of the *D. suzukii* genome. Although it is likely that TEs are  
312 strongly selected against in these regions due to their negative effect on gene function or  
313 expression (Lee and Karpen 2017; Mérel et al. 2020), it is also possible that TE copies are  
314 depleted in these regions because they promote ectopic recombination. In agreement with the  
315 latter hypothesis, gene-rich regions are also known to display high recombination rate in *D.*  
316 *melanogaster* (Adams et al. 2000). The generation of a genomic map of recombination rates in  
317 *D. suzukii* would be needed to disentangle the respective effects of ectopic recombination and  
318 gene disruption.

319 At the chromosomal scale, we did not find a lower density of TEs on the X chromosome  
320 compared to autosomes. This pattern indicates that, if X-linked recessive insertions are more  
321 efficiently selected against than autosomal insertions, the effect on TE abundance is either low  
322 or balanced by another process. When comparing only gene-rich regions, we even found a  
323 higher density of TEs on the X chromosome than on autosomes. Three non-mutually exclusive  
324 explanations can be invoked: (i) there may be a higher insertion rate on the X chromosome,  
325 similar to what was previously found in *D. melanogaster* (Adrion et al. 2017); (ii) the  
326 recombination rate may be lower on the X chromosome, and thus a stronger Muller's ratchet;

327 and (iii) the strength of selection may be reduced by a smaller effective population size for the X  
328 chromosome.

329 Similarly to what has been found in *D. melanogaster* and *D. simulans* (Kofler, Nolte, et al. 2015)  
330 and to what is probably common among *Drosophila* species (Hill 2019), the pattern of TE  
331 insertion frequencies in *D. sukukii* is compatible with an active repeatome. We found differences  
332 in the mean insertion frequency between TE orders, which suggests differences in activity but  
333 could also result from variation in the strength of purifying selection acting against the different  
334 orders (Petrov et al. 2003; Lee and Karpen 2017) Considering the trap model of TE dynamics  
335 (*i.e.* a model in which newly invading TEs are quickly inactivated by host defense (Zanni et al.  
336 2013; Kofler et al. 2018)), an active repeatome suggests a recurrent turnover of TEs, potentially  
337 due to horizontal transfer events. Investigating TE activity in *D. melanogaster* and *D. simulans*,  
338 Kofler and colleagues (Kofler, Nolte, et al. 2015) suggested that such a turnover is influenced by  
339 the colonization history of those species. They propose that the high activity of DNA  
340 transposons in *D. simulans* results from horizontal transfer events from *D. melanogaster* during  
341 *D. simulans* worldwide colonization. In agreement, we detected more families/pseudofamilies in  
342 invasive populations of *D. sukukii* than in the native ones, suggesting that new TE families may  
343 have been acquired during the recent colonization of new areas. However, because the TE  
344 database used here relies on a reference genome obtained from individuals originating from  
345 America (*i.e.* from the Watsonville population), one may expect to find much more  
346 families/pseudofamilies in American than European populations. Yet this is not what we  
347 observed. This could be due to admixture between American and European populations.  
348 However, population genetics studies have shown that gene flow between the two continents is  
349 limited if not absent (Fraitout et al. 2017). It is thus possible that, for technical reasons, we are  
350 simply missing some families that are less abundant in the Asian native range of the species.  
351 The comparison of long read assemblies of genomes generated from individuals originating  
352 from the three continents (Asia, America and Europe) should help shedding light on this issue.

### 353 **Demography, rather than environment or genotype, drives TE content**

354 In agreement with the Lynch and Connery hypothesis (Lynch and Conery 2003), we  
355 found that the TE content in *D. sukukii* is negatively correlated with the Watterson estimate of  
356 genetic diversity  $\hat{\theta}_w$  which may be viewed as a proxy of the population effective size  $N_e$ . The  
357 negative correlation between  $\hat{\theta}_w$  and TE content was significant when considering only

358 European invasive populations, invasive populations as a whole, or only native populations, but  
359 was not significant when considering only American invasive populations. Although a few  
360 studies suggest an increase of TE content following colonization (Nardon et al. 2005; García  
361 Guerreiro et al. 2008; García Guerreiro and Fontdevila 2011; Talla et al. 2017), to our  
362 knowledge it is the first time that a correlation between TE content and  $N_e$  is found at the  
363 intraspecific level. Although several factors may affect  $N_e$ , the variation observed is likely to  
364 result from demographic processes. Indeed, both European and American invasive populations  
365 have encountered bottlenecks (Fraitout et al. 2017). In agreement with this idea, the invasive  
366 population from Hawaii, which experienced the strongest bottleneck (Fraitout et al. 2017),  
367 showed the smallest  $\hat{\theta}_w$  values. It is interesting to note that the negative correlation between  $\hat{\theta}_w$   
368 and TE content remains significant when considering only native populations suggesting that  
369 other demographic event than bottleneck may also be involved (e.g. different stable effective  
370 population sizes and gene flow patterns),

371 Our analysis is controlled for sequencing bias, *i.e.* coverage and insert size, and we are  
372 confident in the biological significance of the correlation observed here. However, it is worth  
373 stressing that our dataset of TE insertions corresponds to a small fraction of the repeatome.  
374 Indeed, the mean number of insertions per HG per population is markedly below the number of  
375 TE copies recovered in the reference genome. We believe that this is due to an impossibility to  
376 properly call TE insertions when TEs are too close or even nested (Vendrell-Mir et al. 2019). It is  
377 thus possible that the negative correlation that we found here exists only for some part of the  
378 genome. Especially it is likely that regions of low TE density, where most of TE insertions are  
379 polymorphic, display the strongest answer to a reduction of selection efficacy. This is simply  
380 because polymorphic insertions can increase in frequency while fixed insertions cannot. One  
381 could also argue that the efficiency of selection is a function of the product between  $N_e$  and  $s$   
382 (with  $s$  the selection coefficient). Therefore, the effects of a reduction of  $N_e$  should be especially  
383 marked in regions where selection against TEs is strong, such as TE-poor / gene-rich regions.

384 We found no significant effect on TE abundance for all the 19 environment variables tested.  
385 This might be surprising at first sight given the large number of studies showing an association  
386 between TE activity and external factors, such as temperature or viral infection (García  
387 Guerreiro 2012; Ryan et al. 2016; Horváth et al. 2017; Roy et al. 2020) Several factors may  
388 explain this discrepancy. First, it is important to notice that, in *Drosophila*, most of these studies  
389 rely on lab experiments, some of them exploring environmental conditions unlikely *in natura*

390 (see (García Guerreiro 2012) for a review). To our knowledge none of these studies established  
391 a link between TE activity and natural environment without any possible confounding effect from  
392 population structure and demographic features. Second, as often in *Drosophila*, most of such  
393 research works were carried out on the same particular species, *D. melanogaster*, so that so far  
394 we do not know much about interspecific variability. Third, although partial Mantel tests allowed  
395 revealing 15 significant correlations between TE abundance and environmental variables in *A.*  
396 *thaliana* populations (Quadrana et al. 2016), we consider our results as conservative, especially  
397 regarding the long discussion about the statistical performance of partial Mantel tests (Diniz-  
398 Filho et al. 2013). More sophisticated statistical methods may be needed to tackle such  
399 relationships into more details.

400 Considering that several studies on *Drosophila* suggest a genotype effect on TE activity  
401 (Biémont et al. 1987; Pasyukova and Nuzhdin 1993; Díaz-González et al. 2011; Adrion et al.  
402 2017), we performed a GWAS on TE abundance to assess this effect in natural populations and  
403 identify the genomic regions involved. Overall, we found ca. 5,000 genomic regions associated  
404 with TE abundance. These regions were not enriched in transcription factor genes nor genes of  
405 the piRNA pathway. As far as we know, no such GWAS study has been carried out in  
406 *Drosophila* populations. Our results are somewhat similar to those found in *A. thaliana*, in which  
407 although a strong causal link between one transcription factor and the abundance of two TE  
408 families was found, no enrichment for any particular function was observed (Quadrana et al.  
409 2016). Comparative genomics between closely related species may help identify a general  
410 pattern. Especially, one could lead the same study using available *D. melanogaster* PoolSeq  
411 data (Kapun et al. 2020), and focus on genes identified in both *D. melanogaster* and *D. sukukii*,  
412 as they might be likely to play a key role in the modulation of TE activity.

### 413 **A potential adaptive role for a limited number of TEs**

414 Similar to studies investigating TE adaptive potential in *D. melanogaster* populations  
415 (González et al. 2008; González et al. 2010; Rech et al. 2019), we found several putatively  
416 adaptive TE insertions in our *D. sukukii* dataset. Overall, we found 15 insertions, six of which  
417 likely to have eased the worldwide invasion of *D. sukukii*. It is important to note that we are  
418 probably missing some insertions, and thus likely underestimating the number of adaptive  
419 insertions sites.

420 Overall, we did not capture a strong signal of a selective sweep near the candidate adaptive TE  
421 insertions. This may be due to overall large effective population sizes as suggested in  
422 (Olazcuaga et al. 2020), but also to the fact that Tajima's D is unlikely to detect soft selective  
423 sweep, *i.e.* adaptation from standing variation or multiple successive beneficial mutations  
424 (Pennings and Hermisson 2006). An appealing perspective would be to sequence candidate  
425 regions in individual strains and use a haplotype-based analysis. For example, the recently  
426 introduced Comparative Haplotype Identity (xMD) statistics (Lange and Pool 2016; Villanueva-  
427 Cañas et al. 2017) has been shown to perform well for soft sweeps. If the effect of our candidate  
428 TE insertion on Tajima's D is globally low, it highlighted the possibility that the absence rather  
429 than the presence of the insertion may be adaptive, at least for some of our candidate  
430 insertions. More specifically, for four insertions a positive correlation was found between local  
431 Tajima's D and insertion frequency. However, the only extreme local Tajima's D was found in  
432 the population where the putatively adaptive insertion is at its highest frequency, indicating that  
433 it is probably the insertion itself rather than the absence that might be adaptive.

434 One added value to our analysis based on GWAS is that the same type of analysis has been  
435 carried out using SNPs/InDel (Olazcuaga et al. 2020). The authors of this study found 204  
436 markers strongly associated with invasion success distributed over the whole genome. If we  
437 compare this number to our six TE insertions, it seems unlikely that TEs solely may explain the  
438 genetic paradox of invasive species (Stapley et al. 2015). It is worth noting that the level of  
439 variation remains high in invasive *D. suzukii* populations (Framout et al. 2017). Hence, it would  
440 be interesting to carry out similar analyses in invasive species that experienced a more intense  
441 depletion of genetic variation during invasion (Prentis et al. 2009; Zhang et al. 2010; Roux et al.  
442 2011) to assess whether TEs are more likely to be adaptive in invasive populations with low  
443 levels of genetic diversity.

444 At first sight our finding of 15 putatively adaptive polymorphic insertions in worldwide  
445 populations of *D. suzukii* contrasts with the 41 to 300 putatively adaptive polymorphic insertions  
446 found in worldwide populations of *D. melanogaster* (Rech et al. 2019). The difference is even  
447 more blatant considering that we analyzed 7,004 polymorphic insertions, against ~800 in (Rech  
448 et al. 2019). This suggests a largely higher rate of TE induced adaptations during *D.*  
449 *melanogaster* invasion and this despite the much larger, still active and diverse repeatome of *D.*  
450 *suzukii*. This discrepancy could have several non-exclusive explanations. First, it may be due to  
451 historical differences between the two species. *D. melanogaster* experienced a relatively slow

452 and ancient worldwide invasion that started from Africa about ~15,000 ya, whereas *D. suzukii*  
453 came out from its native range in Asia only a few decades ago (Stephan and Li 2007; Fraimout  
454 et al. 2017). Second, the discrepancy may result from intrinsic species differences with respect  
455 to the repeatome contents. For example, *D. melanogaster* TEs could possess more  
456 environment responsive sequences that might be co-opted by the host. Third, it may be due to  
457 differences in the methodology used for the two species. Our analysis relies essentially on the  
458 research of overly differentiated TEs across populations with a correction for population  
459 structure (Gautier 2015; Olazcuaga et al. 2020), whereas in the analysis used for *D.*  
460 *melanogaster* there is no direct methodological control for population structure. In the *D.*  
461 *melanogaster* study (Rech et al. 2019), a TE insertion is considered as putatively adaptive if it is  
462 present at high population frequency (from 10% to 95%), and is located in genomic regions  
463 where recombination rate -and so selection efficacy - is high (ca. 300 putatively adaptive  
464 insertions). Further evidence is collected using a combination of three haplotype-based tests to  
465 detect selective sweeps in the vicinity of candidates, and statistical treatments based on Fst  
466 estimations (with 84 insertions confirmed by at least one test). Applying our statistical  
467 methodologies to the *D. melanogaster* dataset, which also consist in PoolSeq data, would help  
468 to determine if methodology differences can explain the observed discrepancy. Finally, one  
469 could ultimately rely on experimental evolution, applying the same selective pressure to different  
470 *Drosophila* species, to test for an impact of intrinsic species differences on TE adaptive  
471 potential.

472 Our study of TE induced adaptation strongly calls for a validation of candidate insertions. Allele  
473 specific expression assays would allow evaluating if these insertions affect nearby gene  
474 expression (Gonzalez et al. 2009). This would consist in testing a difference of nearby gene  
475 expression between the two alleles of an F1 hybrid between strains with and without the  
476 insertion. While such test should control for genotype effect, as compared to a simple test of  
477 differential expression between strains, it does not preclude for an effect of a SNP/InDel close to  
478 the insertion. Using a CRISPR/Cas9 methodology would also allow (in)validate that the TE(s) of  
479 interest is the causative agent of gene expression change and would allow direct testing for a  
480 phenotypic effect.

## 481 **Conclusion**

482 Our study illustrates the value of an approach combining a long reads based genome  
483 assembly, a *de novo* reconstruction of TE sequences, and PoolSeq population data, to



484 characterize the repeatome of a non model species. Our set of analyses especially highlighted  
485 that the particularly large *D. suzukii* repeatome is probably active and shaped by purifying  
486 selection, similar to that of *D. melanogaster*'s. Additional data, such as local recombination rate,  
487 would also help us shed light on the nature of selection acting on TEs. The analysis of TE  
488 abundance variations in invasive and native populations suggests that a reduction of purifying  
489 selection intensity, in response to demographic processes, can significantly increase TE  
490 content. Our study also indicates that positive selection may act on TE insertions in response to  
491 selective factors that remains to be determined. Experimental validation will allow to (in)validate  
492 a functional impact of our putatively adaptive insertions. Overall, the natural extent of the trends  
493 we uncovered here should be explored into more details, for instance through the application of  
494 similar methods to other (invasive) species that would allow to evaluate the impact of a stronger  
495 bottleneck on both TE content increase and TE adaptive potential.

## 496 **Materials & Methods**

### 497 **Creation of a TE database**

498 A TE database was created by merging previously established consensus of *Drosophila*  
499 TE families and *de novo* reconstructed consensus of *D. suzukii* TE families. The previously  
500 established consensus were obtained by extracting all *Drosophila* consensus annotated as  
501 DNA, LINE, LTR, Other, RC, SINE and Unknown from Dfam and Repbase databases (release  
502 2016-2018 for both) (Hubley et al. 2016; <https://www.girinst.org/repbase/>). Full LTR element  
503 sequences were reconstructed by merging LTRs and their internal parts. *De novo*  
504 reconstruction was performed using an assembly of an American strain from Watsonville,  
505 sequenced using PacBio long reads technology, and the REPET package (v2.5) (Flutre et al.  
506 2011; Paris et al. 2020). Unless otherwise specified, the options were used as in the default  
507 configuration file. Briefly, the genome assembly was cut into batches and aligned to itself using  
508 blastn (ncbi-blast v2.2.6) (Altschul et al. 1990). High-scoring Segment Pairs (HSPs) were  
509 clustered using Recon (v1.08) and Piler (v1.0) (Bao and Eddy 2002; Edgar and Myers 2005). A  
510 structural detection step was performed using LTRHarvest from the GenomeTools package  
511 (v1.5.8) (Ellinghaus et al. 2008; Gremme et al. 2013). LTRHarvest-produced sequences were  
512 clustered using blastclust. Consensus sequences were created for each cluster using MAP  
513 (Huang 1994). Additional consensus sequences were generated using RepeatScout (v1.0.5)  
514 (Price et al. 2005). All consensus, *i.e.* from Recon, Piler, LTRHarvest and RepeatScout, were

515 further submitted to a filtering step. Sequences were retained only if they produced at least 3  
516 hits against the genome assembly with at least 98% query coverage (blastn, blast 2.6.0+).  
517 Structural and coding features were identified and used to classify consensus (see Hoede et al.  
518 (2014) for classification details, the used libraries were  
519 ProfilesBankForREPET\_Pfam27.0\_GypsyDB.hmm, rebase20.05\_aaSeq\_cleaned\_TE.fsa,  
520 rebase20.05\_ntSeq\_cleaned\_TE.fsa). Single satellite repeats, potential host genes and  
521 unclassified sequences were filtered out. Since REPET can easily mis-annotate any pair of  
522 repeats separated by a spacer as TRIM or LARD, those sequences were also removed  
523 (Arkhipova 2017). Remaining sequences were further annotated by homology to previously  
524 established consensus of *Drosophila* TE families. Homology was determined using  
525 RepeatMasker (-cutoff 250, v 1.332) (<http://www.repeatmasker.org/>). We followed the rules  
526 below: 1) if all hits belonged to the same superfamily, the sequence was annotated as  
527 corresponding to that particular superfamily and order; 2) if hits from different superfamilies  
528 were observed the sequence was considered as ambiguous; 3) without any hit, the sequence  
529 was annotated as unknown. Ambiguous sequences were manually curated, sequences which  
530 could be unambiguously attributed to one superfamily according to hits and proteic domains  
531 were kept (proteic domains were investigated using NCBI Conserved Domain Search  
532 (<https://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>). Finally, consensus were clustered in  
533 families using UClust (-id 0.80, -strand both, -maxaccepts 0 -maxrejects 0; v11.0.667) (Edgar  
534 2010). The annotation, superfamily and order, attributed to each cluster, *i.e.* each family, is the  
535 annotation of the longest sequence in the cluster. The generated TE database is accessible at:  
536 <https://github.com/vmerel/Dsu-TE>.

### 537 **Annotation of the reference genome**

538 To recover TE fragments and TE genomic sequence occupancy, the reference genome  
539 assembly was masked using RepeatMasker and the above TE database (-gccalc, -s, -a, -cutoff  
540 200, -no\_is, -nolow, -norna, -u; v 1.332) (<http://www.repeatmasker.org/>). TE density was  
541 evaluated as the number of TE fragments completely within non overlapping genomic windows  
542 of 200 kb. TE copies were reconstructed from TE fragments using OneCodeToFindThemAll  
543 (Bailly-Bechet et al. 2014). Gene density was computed from a run of augustus (-species=fly, -  
544 strand=both, -genemodel=complete; v2.5.5) (Stanke et al. 2008) as the number of genes  
545 completely within non overlapping genomic windows of 200 kb. Promer was used to generate  
546 alignments between *D. melanogaster* and *D. sukuzii* assemblies and establish syntenic  
547 relationships (MUMmer v3.23) (Kurtz et al. 2004). *D. melanogaster* masked assembly was

548 downloaded from UCSC Genome Browser (dm6;  
549 <http://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/>). *D. suzukii* masked assembly was  
550 retrieved from RepeatMasker output (see above). The promoter output was filtered out using the  
551 delta-filter module in order to obtain a one-to-one mapping of reference to query (-q, -r). A file  
552 containing alignment coordinates for alignments of minimum length 100 bp, and in which  
553 overlapping alignments were merged, was generated with the show-coords module (-b, -L 100, -  
554 r). Because the abundance of repeated sequences and the use of masked assemblies may  
555 result in multiple small alignments, alignments separated by less than 20 kb were merged using  
556 a custom script. Note that only alignments implying the 2L, 2R, 3L, 3R, X and 4 chromosomes  
557 of *D. melanogaster* were kept at this step and if a *D. suzukii* contig aligned to several *D.*  
558 *melanogaster* chromosomes only the best pair was conserved (*i.e.* the pair producing the  
559 longest alignment). A graphical visualization of the results was produced using Circos  
560 (Krzywinski et al. 2009).

## 561 **Fly samples and pool sequencing**

562 Pool-sequencing (PoolSeq) data originate from Olazcuaga et al. (2020) where the  
563 detailed associated protocol is described. Briefly, adult wild flies were sampled between 2013  
564 and 2016 from 22 localities of both native and invasive areas (fig. 3A) (Fraimout et al. 2017). Six  
565 samples were collected in the native Asian area, more precisely in four Chinese and two  
566 Japanese localities. The remaining 16 samples were chosen to be representative of two  
567 separate invasion roads: the American invasion road and the European invasion road. The  
568 American invasion road is represented by one Hawaiian sample, one Brazilian sample and six  
569 samples from the United States. The European invasion road corresponds to two German  
570 samples, four French samples (including one from La Réunion Island), one Italian sample and  
571 one Spanish sample. For each population sample, DNA extraction was performed from the  
572 thoraxes of 50 to 100 flies and used to prepare paired-end (PE) libraries (insert size of  
573 ~550 bp). PE sequencing was achieved using a HiSeq 2500 from Illumina to obtain 2×125 bp  
574 reads. Reads were trimmed using the trim-fastq.pl script in the PoPoolation package (-min-  
575 length 75, -quality-threshold 20; v1.2.2) (Kofler et al. 2011).

## 576 **TE frequency pipeline**

577 To obtain TE insertion frequencies in PoolSeq samples a calling of TEs was done using  
578 PoPoolationTE2 (Kofler et al. 2016), the reference genome and the newly constructed  
579 database. To make sure that no reads from TE sequences could map on the masked assembly,

580 TE reads were simulated, mapped on the masked assembly and aligned positions were also  
581 masked. Reads simulation was performed using the script `create-reads-for-te-sequences.py`  
582 (Kofler et al. 2016): reads of 125 bp reads, coverage of 1024 X per TE sequence in the  
583 database. Because we do not expect a split read based TE calling tool such as PoPoolationTE2  
584 to accurately call for insertions shorter than the insert size, TE sequences shorter than 500 bp  
585 were removed before calling. Moreover, as PoPoolationTE2 filters out insertions with reads  
586 mapping on more than one family, families with cross-mapping were grouped in pseudofamilies.  
587 Two families were brought together if at least 1% of reads from one sequence of the first family  
588 were mapped on a sequence of the second family (read simulation: 125 bp reads, coverage of  
589 100 X per consensus). Concerning the TE calling, reads were mapped using `bwa bwasmw`  
590 (v0.7.17) (Li and Durbin 2010) and paired-end information restored using the `se2pe` script  
591 provided with the PoPoolationTE2 package (v1.10.04) (Kofler et al. 2016). One unique `ppileup`  
592 file was generated with all samples specifying a minimum mapping quality of 15. The remaining  
593 modules of PoPoolationTE2 were used as follow: `identifySignatures: -mode joint, -signature-`  
594 `window minimumSampleMedian, -min-valley minimumSampleMedian, -min-count 2;`  
595 `updatestrand: -map-qual 15, -max-disagreement 0.5; frequency; filterSignatures -min-`  
596 `coverage 10, -max-otherte-count 2, -max-structvar-count 2; pairupSignatures -min-distance -`  
597 `200, -max-distance 300.` The final output contained frequencies in the 22 populations for each  
598 called TE insertion. See supplementary methods for the validation work on simulated data.

## 599 **TE abundance pipeline**

600 TE abundances, as the numbers of insertions per HG per population, were estimated in  
601 PoolSeq samples by summing insertion frequencies in each sample. Since this pipeline also  
602 relies on the estimation of TE frequencies in PoolSeq samples, it is very similar to the TE  
603 frequency pipeline. However, the last steps were modified to account for differences in coverage  
604 and insert sizes between samples and to allow an unbiased comparison of TE abundance  
605 across samples. After the `ppileup` step the following analyses were performed:  
606 `subsamplePpileup: -target-coverage 30; identifySignatures -mode separate, -signature-`  
607 `window minimumSampleMedian, -min-valley minimumSampleMedian, -min-count 2;`  
608 `updatestrand: -map-qual 15, -max-disagreement 0.5; frequency; filterSignatures: -min-`  
609 `coverage 10; -max-otherte-count 2; -max-structvar-count 2; pairupSignatures: -min-distance -`  
610 `200; -max-distance 300.` See supplementary methods for the validation work on simulated data.

## 611 **Evaluation of population genetics statistics**

612 We estimated Watterson's theta ( $\hat{\theta}_w$ ) and Tajima's D statistics in non-overlapping  
613 1000 bp windows using PoPoolation (v1.2.2) (Kofler et al. 2011). Forward and Reverse trimmed  
614 reads were mapped separately using bwa aln (-o 2 -d 12 -e 12 -n 0.01; v0.7.17) (Li and Durbin  
615 2010). A paired-end alignment file was generated using bwa sampe. Reads were filtered for a  
616 minimum mapping quality of 20 and a pileup file generated with samtools (v1.7) (Li et al. 2009).  
617 Each pileup file was split into two files: one corresponding to autosomal contigs and another  
618 corresponding to X-linked contigs (autosomal and X-linked contigs as determined in Olazcuaga  
619 et al. (2020)). PoPoolation was used as follows: -min-count 2 -min-coverage 8 -max-coverage  
620 250 -min-qual 20. The pool-size argument was modified accordingly between autosomal and X-  
621 linked pileup.

## 622 **Genome Wide Association Study with TE family abundance**

623 All genome scans were performed using BayPass (v2.2) (Gautier 2015; Olazcuaga et al.  
624 2020), a package aiming at identifying markers evolving under selection and/or associated to  
625 population-specific covariates, taking into account the shared history of the populations. For  
626 each SNP/InDel previously called in these PoolSeq samples (Olazcuaga et al. 2020), we  
627 estimated 83 Bayes Factors (BF), reflecting their association with the number of insertions per  
628 HG of 83 families/pseudofamilies (based on a linear regression model). The 83 chosen TE  
629 families/pseudofamilies were those displaying an amplitude of variation of at least three  
630 insertions per HG across the complete dataset. To improve computing time BayPass was run  
631 on data subsets. Data concerning TE abundance was split into three subsets of 28, 28 and 27  
632 families, respectively. For SNPs/InDel, we used the data subsets of Olazcuaga et al. (2020), for  
633 which the 11,564,472 autosomal variants are divided into 154 subsets and the 1,966,184 X-  
634 linked variants into 26 subsets. Since we used the importance sampling algorithm implemented  
635 in Baypass to assess BFs, and single run estimations may be unstable, a total of three runs  
636 were performed for each combination of TE subsets-SNP/InDel subsets and the median of BFs  
637 computed (Gautier et al. 2018). Note that different pool size files were used for autosomal and  
638 X-linked variants to take into account differences in the number of autosomes and X  
639 chromosomes in each PoolSeq sample. In accordance to Jeffrey's rule, a SNP/InDel was  
640 considered as associated with a TE family/pseudofamily abundance for a BF superior to 20  
641 deciban (dB) (Jeffreys 1961).

642 SNP/InDel locations were used to define genomic regions associated with TE abundance.  
643 Variants were gathered if separated by less than 1 kb. If the spanned genomic interval was less  
644 than 1 kb or if a variant could not be found, the region was obtained by adding 500 bp on both  
645 sides. For each region we looked for overlapping TEs using the RepeatMasker annotation (gff  
646 file, see *Annotation of the reference genome*). We also investigated gene content. First, we  
647 retrieved homologous regions in the *D. melanogaster* genome using BLAT against the *D.*  
648 *melanogaster* masked assembly downloaded from UCSC Genome Browser  
649 (<http://hgdownload.soe.ucsc.edu/goldenPath/dm6/bigZips/>; BLAT v.36x4, -t=dnax -q=dnax). We  
650 then checked for genes overlapping the best hit subject sequence using the UCSC Genome  
651 Browser gff annotation file. Note that if the best hit score was lower than 100 we considered that  
652 no homologous region was retrieved. The number of transcription factor genes among the  
653 genes retrieved was obtained by comparing their IDs to those of the gene group Transcription  
654 factor on flybase (<https://flybase.org/reports/FBgg0000745.html>). Similarly, the number of genes  
655 involved in the piRNA pathway was obtained by comparing gene IDs to those listed in Ozata et  
656 al. (2019). In order to test if the candidate regions were enriched in TEs we generated random  
657 expectations by applying the above to 1000 randomly selected SNPs 250 times. For computing  
658 time reasons, for genes, transcription factor genes, or genes involved in the piRNA pathway, we  
659 used 500 randomly selected SNPs 125 times.

## 660 **Correlation between climatic variables and TE family abundance**

661 Partial Mantel tests were used to test the correlation between bioclimatic variables and  
662 TE family abundance correcting for population structure (as in Quadrana et al. (2016)). 19  
663 bioclimatic variables from the worldclim dataset (Fick and Hijmans 2017) were considered:  
664 annual mean temperature, mean diurnal range, isothermality, temperature seasonality, max  
665 temperature of warmest month, minimum temperature of coldest month, temperature annual  
666 range, mean temperature of wettest quarter, mean temperature of driest quarter, mean  
667 temperature of warmest quarter, mean temperature of coldest quarter, annual precipitation,  
668 precipitation of wettest month, precipitation of driest month, precipitation seasonality,  
669 precipitation of wettest quarter, precipitation of driest quarter, precipitation of warmest quarter,  
670 precipitation of coldest quarter. The 83 families with an amplitude of variation of at least three  
671 insertions per HG between populations were considered. The population structuring of genetic  
672 diversity is summarized by the scaled covariance matrix of population allele frequencies ( $\Omega$ )  
673 estimated with Baypass, one autosomal subset randomly chosen was used (the correlation of  
674 the posterior means of the estimated  $\Omega$  elements across SNP subsamples had previously been

675 verified (Olazcuaga et al. 2020)). Partial Mantel tests were conducted using the R package  
676 `ecodist` (Goslee and Urban 2007). P-values were further adjusted to account for multiple testing  
677 applying the Benjamini-Hochberg correction (Benjamini and Hochberg 1995).

## 678 **Screening for putatively adaptive TE insertions**

679 A genome scan for putatively adaptive TE insertions was performed using BayPass with  
680 the output of the TE frequency pipeline (v2.2) (Gautier 2015; Olazcuaga et al. 2020). Insertions  
681 with Minor Allelic Frequency (MAF) inferior to 0.025 were removed before the analysis.  
682 Autosomal and X-linked contigs were analyzed separately. Three statistics were computed to  
683 detect putatively adaptive TE insertions:  $XtX$ ,  $C_2$  and the Bayes Factor (BF) for Environmental  
684 Association Analysis. Briefly,  $XtX$  corresponds to a global differentiation statistics,  $C_2$  contrasts  
685 allelic frequencies between user-defined groups of populations, and BF measures the support of  
686 the association between a marker and a covariate (usually an environmental variable). Because  
687 Bayes Factor was computed using the importance sampling algorithm, and single run  
688 estimations may be unstable, BF were estimated as the median over five estimates obtained  
689 from independent runs of Baypass (Gautier et al. 2018). In accordance to Jeffrey's rule, a BF  
690 superior to 20 deciban (dB) was considered as decisive evidence supporting an association  
691 (Jeffreys 1961).  $XtX$  and  $C_2$  estimates came from one single run and simulation was used to  
692 determine a significance threshold. The R function `simulate.baypass()` provided within the  
693 BayPass package was used to simulate read count data (`nsnp=10000`, `pi.maf=0`). We used the  
694 physical coverage estimated from the `ppileup` file using the module `stat-coverage` of  
695 `PoPoolationTE2` (Kofler et al. 2016). BayPass was run on this simulated dataset to estimate the  
696 null distribution of the  $XtX$  and the  $C_2$  statistics. An insertion was considered as overly  
697 differentiated (for  $XtX$ ) or associated to the tested contrast (for  $C_2$ ) if the corresponding statistics  
698 exceeded the 99.9% quantile of the estimated null distribution. The populations whose  
699 frequencies were contrasted using the  $C_2$  were: populations of the invasive American road and  
700 the native ones ( $C_2^{Am}$ ), populations of the invasive European road and the native ones ( $C_2^{Eu}$ ),  
701 invasive populations and the native ones ( $C_2^{WW}$ ). This choice was made according to the  
702 invasion roads inferred using microsatellite markers (Fraimout et al. 2017), the populations  
703 structure assessed with SNP/InDel markers called in these samples (Olazcuaga et al. 2020)  
704 and the population structure assessed here with TE markers (supplementary fig. S5). For each  
705 putatively adaptive insertion, gene vicinity in a 1 kb region centered on the insertion was  
706 investigated as described in the paragraph "Genome Wide Association Study with TE family  
707 abundance". The presence of the insertion in a region of selective sweep was assessed using

708 Tajima's D. For the 22 populations, we investigated if the Tajima's D estimated in the 1 kb  
709 window containing this insertion was inferior to the quantile 0.05 of Tajima's D distribution in this  
710 population. More precisely, to prevent for a difference between autosome and X chromosome,  
711 autosomal insertions were compared to the autosomal Tajima's D distribution and X-linked  
712 insertions to the X chromosome Tajima's D distribution (with autosomal and X-linked contigs as  
713 defined in Paris et al. (2020)). We also checked if the insertion was close to SNPs/InDels  
714 previously identified as potentially adaptive during *D. sukikii* invasion (considering a maximum  
715 distance of 5 kb) (Olazcuaga et al. 2020).

## 716 **Acknowledgements**

717 This work was supported by the French National Research Agency (ANR-16-CE02-  
718 0015-01 – SWING) and performed using the computing facilities of the CC LBBE/PRABI. We  
719 sincerely thank C. Mermet-Bouvier for technical help. We are also grateful to B. Prud'homme  
720 and F. Sabot for constructive discussion about this article.

## 721 **Figure caption and tables**

722 ***Table 1: Description of the 15 putatively adaptive TE insertions.***

723 *Each insertion is an outlier when considering one or a combination of the global differentiation*  
724 *statistics ( $XtX$ ) and statistics contrasting allelic frequencies between native populations and*  
725 *populations of the invasive American road ( $C_2^{Am}$ ) or populations of the invasive European road*  
726 *( $C_2^{Eu}$ ) or all invasive populations ( $C_2^{WW}$ ).*



Insertion	Statistics	Gene vicinity	Outlier SNP nearby	A/X	TE Order
1	$C_2^{Am} - XtX$	ASPP	F	A	Unknown
2	$C_2^{WW}$	dia	F	A	Unknown
3	$C_2^{WW}$	-	T	A	DNA
4	$C_2^{WW}$	NA	F	X	Unknown
5	$C_2^{WW}$	inaE	F	X	Unknown
6	$C_2^{WW}$	-	F	X	DNA
7	$XtX$	Mical	F	A	DNA
8	$XtX$	CG30015	F	A	Unknown
9	$XtX$	-	F	A	Unknown
10	$XtX$	CR31386	F	A	Unknown
11	$XtX$	-	F	A	Unknown
12	$XtX$	Dop1R2	F	A	Unknown
13	$XtX$	jing	F	A	Unknown
14	$XtX$	CG14282	F	A	Unknown
15	$XtX$	GATAe	F	A	Unknown

727 Note.—The fourth column indicates whether a SNP potentially evolving under positive selection  
728 had been detected less than 5 kb away in Olazcuaga et al. (2020) (F=False, T=True). The fifth  
729 column indicates whether the insertion is located on an autosomal (A) or X-linked contig (X).

730 **Figure 1: Main features of the TE content in the *D. suzukii* reference genome.**

731 A. TE copy numbers and TE genomic occupancy. Barplot representing TE copy numbers for the  
732 20 TE superfamilies displaying the highest copy numbers Piechart illustrating genomic  
733 sequence occupancy of each TE order (in percentages of the assembly). Class I TEs are shown  
734 in green (light green for LINES and darker green for LTR Elements). Class II TEs are shown in  
735 blue (light blue for DNA and darker blue for Rolling Circles (RC)). Non repeated sequences are  
736 shown in gray. B. Distribution of TEs and genes. TE density (pink outer graph) and gene density  
737 (yellow inner graph) are shown for windows of 200 kb. The maximum value of gene density is  
738 54. The maximum number of TE fragments is 713. Syntenic relationships with *D. melanogaster*  
739 assembly are shown inside using light links for regions of low gene density (< 7 genes per 200  
740 kb) and dark links for regions of high gene density (>= 7 genes per 200 kb). Contigs are  
741 surrounded by black strokes. Ticks on *D. melanogaster* assembly are separated by one Mb.

742 **Figure 2: TE activity in the *D. sukuzii* reference population from Watsonville (USA).**

743 A. Frequency distributions of TE insertions. B. Population frequencies for each TE family (in  
744 black) or pseudofamily (in gray). Only families/pseudofamilies with more than 10 insertions in  
745 the reference population are shown. DNA and Rolling Circles (RC) have been grouped for  
746 graphical reasons.

747 **Figure 3: TE dynamics in native and invasive *D. sukuzii* populations.**

748 A. Geographic location and historical status of the 22 *D. sukuzii* population samples genotyped  
749 using a pool-sequencing methodology. Population samples from the native range are in green  
750 and those from the invaded range are in orange (American invasion route) or blue (European  
751 invasion route) (Frainout et al. 2017). B. TE content in *D. sukuzii* populations, as the numbers  
752 of insertions per haploid genome (HG). C. Correlation between TE content and Watterson's  
753 theta in *D. sukuzii* population samples.

754 **Figure 4: Frequencies of each of the 15 putatively adaptive insertions in the 22 *D. sukuzii***  
755 **populations.**

756 Insertion number is indicated on the left together with the associated BayPass statistics. XtX  
757 corresponds to a global differentiation statistic,  $C_2$  to a statistic contrasting allelic frequencies  
758 between native populations and populations of the invasive American road ( $C_2^{Am}$ ) or populations  
759 of the invasive European road ( $C_2^{Eu}$ ) or all invasive populations ( $C_2^{WW}$ ).

760 **Supplementary data:**

761 Supplementary data are available at Molecular Biology and Evolution online.

## 762 Figures

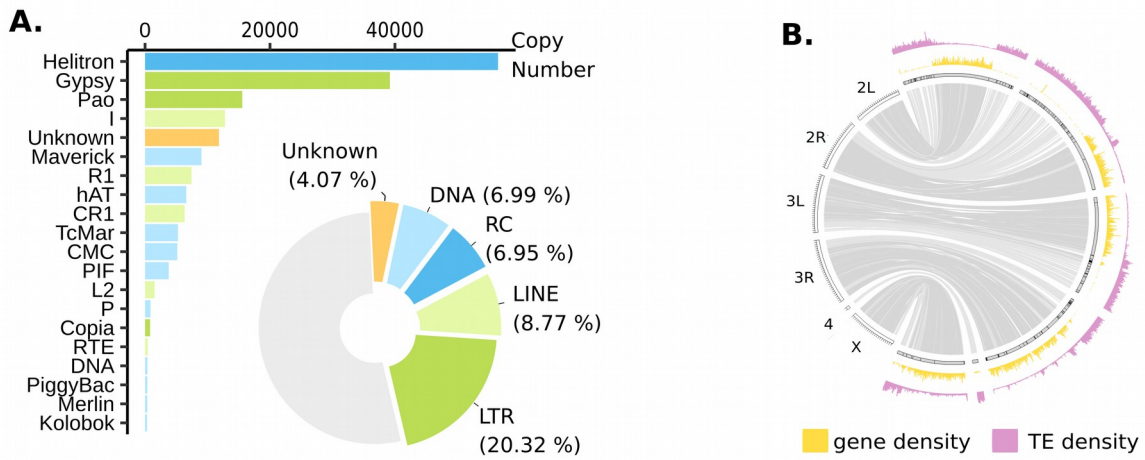


Figure 1

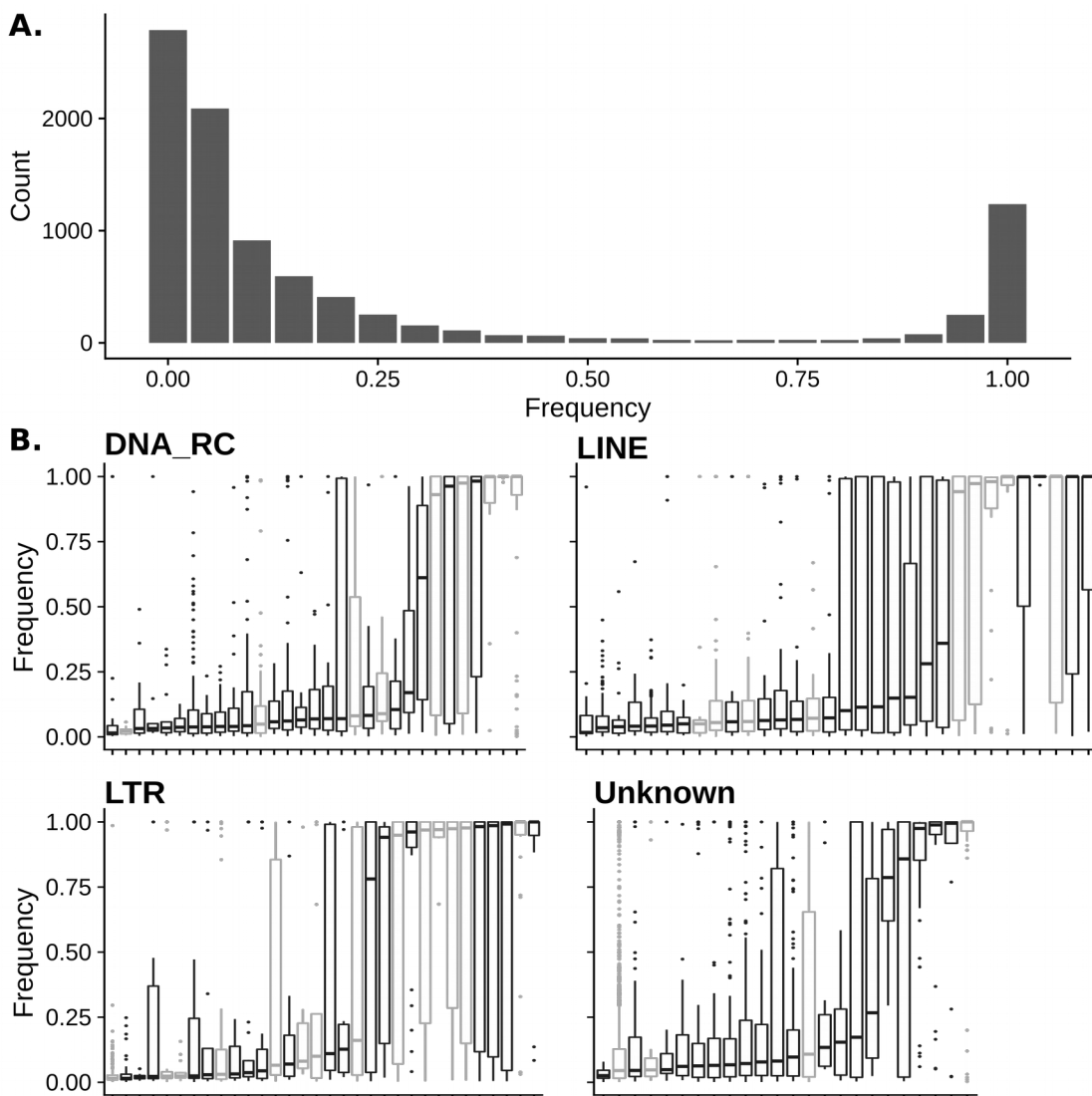
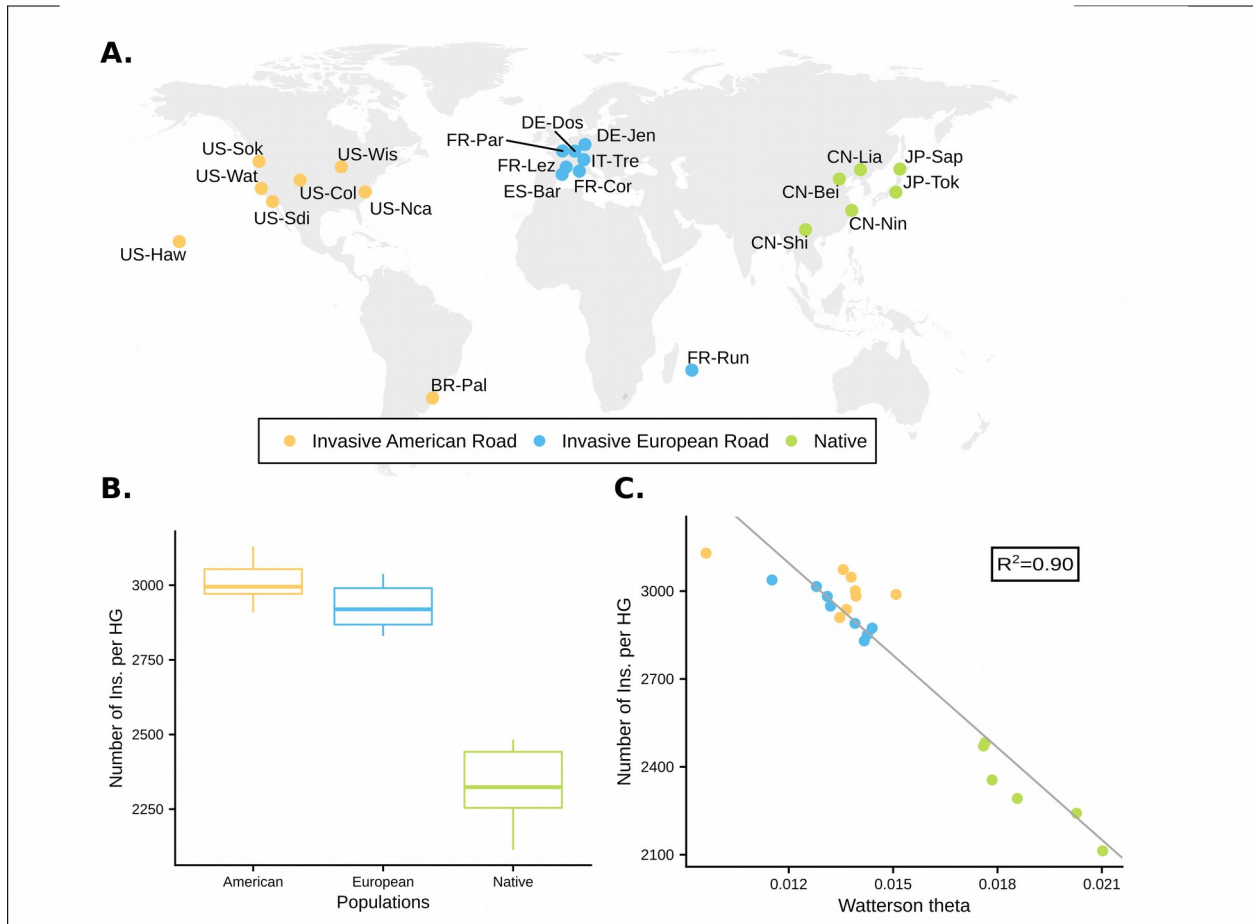
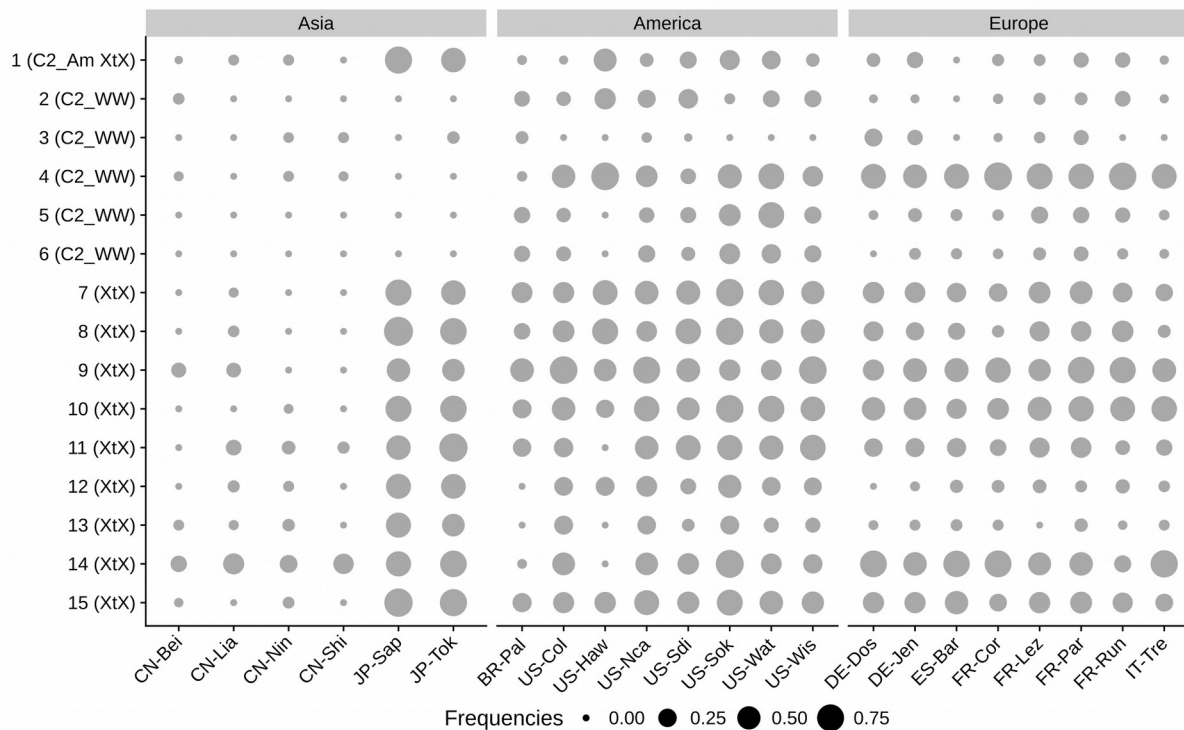


Figure 2



**Figure 3**



**Figure 4**

## 765 References

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The Genome Sequence of *Drosophila melanogaster*. *Science*. 287(5461):2185–2195. doi:10.1126/science.287.5461.2185.

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in *Drosophila Melanogaster*. *Genome Biol Evol*. 9(5):1329–1340. doi:10.1093/gbe/evx050.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *Journal of Molecular Biology*. 215(3):403–410. doi:10.1016/S0022-2836(05)80360-2.

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. 2000. *Nature*. 408(6814):796.

Arkhipova IR. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob DNA*. 8. doi:10.1186/s13100-017-0103-2.

Bailly-Bechet M, Haudry A, Lerat E. 2014. “One code to find them all”: a perl tool to conveniently parse RepeatMasker output files. *Mobile DNA*. 5(1):13. doi:10.1186/1759-8753-5-13.

Bao Z, Eddy SR. 2002. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. *Genome Res*. 12(8):1269–1276. doi:10.1101/gr.88502.

Bartolomé C, Maside X, Charlesworth B. 2002. On the Abundance and Distribution of Transposable Elements in the Genome of *Drosophila melanogaster*. *Mol Biol Evol*. 19(6):926–937. doi:10.1093/oxfordjournals.molbev.a004150.

Benjamini Y, Hochberg Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*. 57(1):289–300.

Biémont C, Aouar A, Arnault C. 1987. Genome reshuffling of the copia element in an inbred line of *Drosophila melanogaster*. *Nature*. 329(6141):742–744. doi:10.1038/329742a0.

Blumenstiel JP, Chen X, He M, Bergman CM. 2014. An Age of Allele Test of Neutrality for Transposable Element Insertions. *Genetics*. 196(2):523–538. doi:10.1534/genetics.113.158147.

Boissinot S, Entezam A, Furano AV. 2001. Selection against deleterious LINE-1-containing loci in the human lineage. *Mol Biol Evol*. 18(6):926–935. doi:10.1093/oxfordjournals.molbev.a003893.

C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*. 282(5396):2012–2018. doi:10.1126/science.282.5396.2012.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. *Genetics Research*. 42(1):1–27. doi:10.1017/S0016672300021455.

Chiu JC, Jiang X, Zhao L, Hamm CA, Cridland JM, Saelao P, Hamby KA, Lee EK, Kwok RS, Zhang G, et al. 2013. Genome of *Drosophila suzukii*, the Spotted Wing *Drosophila*. *G3 (Bethesda)*. 3(12):2257–2271. doi:10.1534/g3.113.008185.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and Distribution of Transposable Elements in Two *Drosophila* QTL Mapping Resources. *Mol Biol Evol*. 30(10):2311–2327. doi:10.1093/molbev/mst129.

Daborn PJ, Yen JL, Bogwitz MR, Goff GL, Feil E, Jeffers S, Tijet N, Perry T, Heckel D, Batterham P, et al. 2002. A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. *Science*. 297(5590):2253–2256. doi:10.1126/science.1074170.

Díaz-González J, Vázquez JF, Albornoz J, Domínguez A. 2011. Long-term evolution of the roo transposable element copy number in mutation accumulation lines of *Drosophila melanogaster*. *Genetics Research*. 93(3):181–187. doi:10.1017/S0016672311000103.

Diniz-Filho JAF, Soares TN, Lima JS, Dobrovolski R, Landeiro VL, de Campos Telles MP, Rangel TF, Bini LM. 2013. Mantel test in population genetics. *Genet Mol Biol.* 36(4):475–485. doi:10.1590/S1415-47572013000400002.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 284(5757):601–603. doi:10.1038/284601a0.

Edgar R, Myers E. 2005. PILER: Identification and classification of genomic repeats. *Bioinformatics (Oxford, England).* 21 Suppl 1:i152-8. doi:10.1093/bioinformatics/bti1003.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics.* 26(19):2460–2461. doi:10.1093/bioinformatics/btq461.

Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics.* 9(1):18. doi:10.1186/1471-2105-9-18.

Estoup A, Ravigné V, Hufbauer R, Vitalis R, Gautier M, Facon B. 2016. Is There a Genetic Paradox of Biological Invasion? *Annual Review of Ecology, Evolution, and Systematics.* 47(1):51–72. doi:10.1146/annurev-ecolsys-121415-032116.

Fick SE, Hijmans RJ. 2017. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology.* 37(12):4302–4315. doi:10.1002/joc.5086.

Flutre T, Duprat E, Feuillet C, Quesneville H. 2011. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE.* 6(1):e16526. doi:10.1371/journal.pone.0016526.

Fraimout A, Debat V, Fellous S, Hufbauer RA, Foucaud J, Pudlo P, Marin J-M, Price DK, Cattel J, Chen X, et al. 2017. Deciphering the Routes of invasion of *Drosophila suzukii* by Means of ABC Random Forest. *Mol Biol Evol.* 34(4):980–996. doi:10.1093/molbev/msx050.

García Guerreiro MP, Chávez-Sandoval BE, Balanyà J, Serra L, Fontdevila A. 2008. Distribution of the transposable elements bilbo and gypsy in original and colonizing populations of *Drosophila subobscura*. *BMC Evolutionary Biology*. 8(1):234. doi:10.1186/1471-2148-8-234.

García Guerreiro MP, Fontdevila A. 2011. Osvaldo and Isis retrotransposons as markers of the *Drosophila buzzatii* colonisation in Australia. *BMC Evolutionary Biology*. 11(1). doi:10.1186/1471-2148-11-111.

García Guerreiro MPG. 2012. What makes transposable elements move in the *Drosophila* genome? *Heredity*. 108(5):461–468. doi:10.1038/hdy.2011.89.

Gautier M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*. 201(4):1555–1579. doi:10.1534/genetics.115.181453.

Gautier M, Yamaguchi J, Foucaud J, Loiseau A, Ausset A, Facon B, Gschloessl B, Lagnel J, Loire E, Parrinello H, et al. 2018. The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Current Biology*. 28(20):3296-3302.e7. doi:10.1016/j.cub.2018.08.023.

González J, Karasov TL, Messer PW, Petrov DA. 2010. Genome-Wide Patterns of Adaptation to Temperate Environments Associated with Transposable Elements in *Drosophila*. *PLOS Genetics*. 6(4):e1000905. doi:10.1371/journal.pgen.1000905.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. 2008. High Rate of Recent Transposable Element-Induced Adaptation in *Drosophila melanogaster*. *PLOS Biology*. 6(10):e251. doi:10.1371/journal.pbio.0060251.



Gonzalez J, Macpherson JM, Petrov DA. 2009. A Recent Adaptive Transposable Element Insertion Near Highly Conserved Developmental Loci in *Drosophila melanogaster*. *Molecular Biology and Evolution*. 26(9):1949–1961. doi:10.1093/molbev/msp107.

Goslee SC, Urban DL. 2007. The ecodist Package for Dissimilarity-based Analysis of Ecological Data. *Journal of Statistical Software*. 22(1):1–19. doi:10.18637/jss.v022.i07.

Gremme G, Steinbiss S, Kurtz S. 2013. GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Trans Comput Biol Bioinformatics*. 10(3):645–656. doi:10.1109/TCBB.2013.68.

Hill T, unpublished data, <https://www.biorxiv.org/content/10.1101/651059v2.full>, last accessed April 5, 2019

Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. 2014. PASTEC: an automatic transposable element classification tool. *PLoS ONE*. 9(5):e91929. doi:10.1371/journal.pone.0091929.

Horváth V, Merenciano M, González J. 2017. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. *Trends in Genetics*. 33(11):832–841. doi:10.1016/j.tig.2017.08.007.

Huang X. 1994. On global sequence alignment. *Comput Appl Biosci*. 10(3):227–235. doi:10.1093/bioinformatics/10.3.227.

Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, Smit AFA, Wheeler TJ. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res*. 44(D1):D81–D89. doi:10.1093/nar/gkv1272.

Initial sequencing and comparative analysis of the mouse genome. 2002. *Nature*. 420(6915):520.

Jeffreys H. 1961. *Theory of Probability*, Ed. 3 Oxford University Press. Oxford.

Kapun M, Barrón MG, Staubach F, Obbard DJ, Wiberg RAW, Vieira J, Goubert C, Rota-Stabelli O, Kankare M, Bogaerts-Márquez M, et al. 2020. Genomic Analysis of European *Drosophila melanogaster* Populations Reveals Longitudinal Structure, Continent-Wide Selection, and Previously Unknown DNA Viruses. *Mol Biol Evol.* 37(9):2661–2678. doi:10.1093/molbev/msaa120.

Kofler R, Betancourt AJ, Schlötterer C. 2012. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet.* 8(1):e1002487. doi:10.1371/journal.pgen.1002487.

Kofler R, Gómez-Sánchez D, Schlötterer C. 2016. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Molecular biology and evolution*:msw137.

Kofler R, Hill T, Nolte V, Betancourt AJ, Schlötterer C. 2015. The recent invasion of natural *Drosophila simulans* populations by the P-element. *PNAS.* 112(21):6659–6663. doi:10.1073/pnas.1500758112.

Kofler R, Nolte V, Schlötterer C. 2015. Tempo and Mode of Transposable Element Activity in *Drosophila*. *PLOS Genetics.* 11(7):e1005406. doi:10.1371/journal.pgen.1005406.

Kofler R, Orozco-terWengel P, Maio ND, Pandey RV, Nolte V, Futschik A, Kosiol C, Schlötterer C. 2011. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLOS ONE.* 6(1):e15925. doi:10.1371/journal.pone.0015925.

Kofler R, Senti K-A, Nolte V, Tobler R, Schlötterer C. 2018. Molecular dissection of a natural transposable element invasion. *Genome Res.* 28(6):824–835. doi:10.1101/gr.228627.117.

Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Res.* 19(9):1639–1645. doi:10.1101/gr.092759.109.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5(2):R12. doi:10.1186/gb-2004-5-2-r12.

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. *Nature.* 409(6822):860–921. doi:10.1038/35057062.

Lange JD, Pool JE. 2016. A haplotype method detects diverse scenarios of local adaptation from genomic sequence variation. *Mol Ecol.* 25(13):3081–3100. doi:10.1111/mec.13671.

Lavergne S, Molofsky J. 2007. Increased genetic variation and evolutionary potential drive the success of an invasive grass. *Proceedings of the National Academy of Sciences.* 104(10):3883–3888. doi:10.1073/pnas.0607324104.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of *Drosophila* euchromatic transposable elements impact their evolution. *Elife.* 6. doi:10.7554/eLife.25762.

Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, González J. 2019. Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Molecular Ecology.* 28(6):1506–1522. doi:10.1111/mec.14963.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics.* 26(5):589–595. doi:10.1093/bioinformatics/btp698.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 25(16):2078–2079. doi:10.1093/bioinformatics/btp352.

Li Z-W, Hou X-H, Chen J-F, Xu Y-C, Wu Q, González J, Guo Y-L. 2018. Transposable Elements Contribute to the Adaptation of *Arabidopsis thaliana*. Van De Peer Y, editor. *Genome Biology and Evolution*. 10(8):2140–2150. doi:10.1093/gbe/evy171.

Lynch M, Conery JS. 2003. The Origins of Genome Complexity. *Science*. 302(5649):1401–1404. doi:10.1126/science.1089370.

Marin P, Genitoni J, Barloy D, Maury S, Gibert P, Ghalambor CK, Vieira C. 2020. Biological invasion: The influence of the hidden side of the (epi)genome. *Functional Ecology*. 34(2):385–400. doi:10.1111/1365-2435.13317.

Medstrand P, van de Lagemaat LN, Mager DL. 2002. Retroelement distributions in the human genome: variations associated with age and proximity to genes. *Genome Res*. 12(10):1483–1495. doi:10.1101/gr.388902.

Mérel V, Boulesteix M, Fablet M, Vieira C. 2020. Transposable elements in *Drosophila*. *Mobile DNA*. 11(1):23. doi:10.1186/s13100-020-00213-z.

Mi S, Lee X, Li X, Veldman GM, Finnerty H, Racie L, LaVallie E, Tang XY, Edouard P, Howes S, et al. 2000. Syncytin is a captive retroviral envelope protein involved in human placental morphogenesis. *Nature*. 403(6771):785–789. doi:10.1038/35001608.

Miller DE, Staber C, Zeitlinger J, Hawley RS. 2018. Highly Contiguous Genome Assemblies of 15 *Drosophila* Species Generated Using Nanopore Sequencing. *G3 (Bethesda)*. 8(10):3131–3141. doi:10.1534/g3.118.200160.

Nardon C, Deceliere G, Løevenbruck C, Weiss M, Vieira C, Biémont C. 2005. Is genome size influenced by colonization of new environments in dipteran species? *Molecular Ecology*. 14(3):869–878. doi:10.1111/j.1365-294X.2005.02457.x.

Nikitin AG, Woodruff RC. 1995. Somatic movement of the mariner transposable element and lifespan of *Drosophila* species. *Mutation Research/DNAging*. 338(1):43–49. doi:10.1016/0921-8734(95)00010-4.

Niu X-M, Xu Y-C, Li Z-W, Bian Y-T, Hou X-H, Chen J-F, Zou Y-P, Jiang J, Wu Q, Ge S, et al. 2019. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. *PNAS*. 116(14):6908–6913. doi:10.1073/pnas.1811498116.

Olazcuaga L, Loiseau A, Parrinello H, Paris M, Fraimout A, Guedot C, Diepenbrock LM, Kenis M, Zhang J, Chen X, et al. 2020. A Whole-Genome Scan for Association with Invasion Success in the Fruit Fly *Drosophila suzukii* Using Contrasts of Allele Frequencies Corrected for Population Structure. *Mol Biol Evol*. 37(8):2369–2385. doi:10.1093/molbev/msaa098.

Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D, et al. 2013. Linking Genomics and Ecology to Investigate the Complex Evolution of an Invasive *Drosophila* Pest. *Genome Biol Evol*. 5(4):745–757. doi:10.1093/gbe/evt034.

Orgel LE, Crick FHC. 1980. Selfish DNA: the ultimate parasite. *Nature*. 284(5757):604–607. doi:10.1038/284604a0.

Ozata DM, Gainetdinov I, Zoch A, OCarroll D, Zamore PD. 2019. PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet*. 20(2):89–108. doi:10.1038/s41576-018-0073-3.

Paris M, Boyer R, Jaenichen R, Wolf J, Karageorgi M, Green J, Cagnon M, Parrinello H, Estoup A, Gautier M, et al. 2020. Near-chromosome level genome assembly of the fruit pest *Drosophila suzukii* using long-read sequencing. *Scientific Reports*. 10(1):11227. doi:10.1038/s41598-020-67373-z.

Pasyukova EG, Nuzhdin SV. 1993. Doc and copia instability in an isogenic *Drosophila melanogaster* stock. *Mol Gen Genet*. 240(2):302–306. doi:10.1007/bf00277071.

Pennings PS, Hermisson J. 2006. Soft Sweeps III: The Signature of Positive Selection from Recurrent Mutation. *PLoS Genet*. 2(12). doi:10.1371/journal.pgen.0020186.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size Matters: Non-LTR Retrotransposable Elements and Ectopic Recombination in *Drosophila*. *Mol Biol Evol.* 20(6):880–892. doi:10.1093/molbev/msg102.

Prentis P, Sigg D, Raghu S, Dhileepan K, Pavasovic A, Lowe A. 2009. Understanding invasion history: Genetic structure and diversity of two globally invasive plants and implications for their management. *Diversity and Distributions.* 15. doi:10.1111/j.1472-4642.2009.00592.x.

Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. *Bioinformatics.* 21 Suppl 1:i351-358. doi:10.1093/bioinformatics/bti1018.

Quadrana L, Bortolini Silveira A, Mayhew GF, LeBlanc C, Martienssen RA, Jeddloh JA, Colot V. 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. Zilberman D, editor. *eLife.* 5:e15716. doi:10.7554/eLife.15716.

Rech GE, Bogaerts-Márquez M, Barrón MG, Merenciano M, Villanueva-Cañas JL, Horváth V, Fiston-Lavier A-S, Luyten I, Venkataram S, Quesneville H, et al. 2019. Stress response, behavior, and development are shaped by transposable element-induced mutations in *Drosophila*. *PLoS Genet.* 15(2):e1007900. doi:10.1371/journal.pgen.1007900.

Rishishwar L, Wang L, Wang J, Yi SV, Lachance J, Jordan IK. 2018. Evidence for positive selection on recent human transposable element insertions. *Gene.* 675:69–79. doi:10.1016/j.gene.2018.06.077.

Rius N, Guillén Y, Delprat A, Kapusta A, Feschotte C, Ruiz A. 2016. Exploration of the *Drosophila buzzatii* transposable element content suggests underestimation of repeats in *Drosophila* genomes. *BMC Genomics.* 17. doi:10.1186/s12864-016-2648-8.

Rollins LA, Richardson MF, Shine R. 2015. A genetic perspective on rapid evolution in cane toads (*Rhinella marina*). *Mol Ecol.* 24(9):2264–2276. doi:10.1111/mec.13184.

Roux JJL, Brown GK, Byrne M, Ndlovu J, Richardson DM, Thompson GD, Wilson JRJ. 2011. Phylogeographic consequences of different introduction histories of invasive Australian *Acacia* species and *Paraserianthes lophantha* (Fabaceae) in South Africa. *Diversity and Distributions*. 17(5):861–871. doi:10.1111/j.1472-4642.2011.00784.x.

Roy M, Viginier B, Saint-Michel É, Arnaud F, Ratinier M, Fablet M. 2020. Viral infection impacts transposable element transcript amounts in *Drosophila*. *PNAS*. 117(22):12249–12257. doi:10.1073/pnas.2006106117.

Ryan CP, Brownlie JC, Whyard S. 2016. Hsp90 and Physiological Stress Are Linked to Autonomous Transposon Mobility and Heritable Genetic Change in Nematodes. *Genome Biol Evol*. 8(12):3794–3805. doi:10.1093/gbe/evw284.

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 326(5956):1112–1115. doi:10.1126/science.1178534.

Sessegolo C, Burlet N, Haudry A. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters*. 12(8):20160407. doi:10.1098/rsbl.2016.0407.

Sessegolo Camille, Burlet Nelly, Haudry Annabelle. 2016. Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. *Biology Letters*. 12(8):20160407. doi:10.1098/rsbl.2016.0407.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*. 24(5):637–644. doi:10.1093/bioinformatics/btn013.

Stapley J, Santure AW, Dennis SR. 2015. Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Molecular Ecology*. 24(9):2241–2252. doi:10.1111/mec.13089.

Stephan W, Li H. 2007. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity*. 98(2):65–68. doi:10.1038/sj.hdy.6800901.

Talla V, Suh A, Kalsoom F, Dincă V, Vila R, Friberg M, Wiklund C, Backström N. 2017. Rapid Increase in Genome Size as a Consequence of Transposable Element Hyperactivity in Wood-White (Leptidea) Butterflies. *Genome Biol Evol.* 9(10):2491–2505. doi:10.1093/gbe/evx163.

Van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. *Nature.* 534(7605):102–105. doi:10.1038/nature17951.

Vendrell-Mir P, Barteri F, Merenciano M, González J, Casacuberta JM, Castanera R. 2019. A benchmark of transposon insertion detection tools using real data. *Mobile DNA.* 10(1):53. doi:10.1186/s13100-019-0197-9.

Vieira C, Lepetit D, Dumont S, Biémont C. 1999. Wake up of transposable elements following *Drosophila simulans* worldwide colonization. *Mol Biol Evol.* 16(9):1251–1255.

Villanueva-Cañas JL, Rech GE, de Cara MAR, González J. 2017. Beyond SNPs: how to detect selection on transposable element insertions. *Methods in Ecology and Evolution.* 8(6):728–737. doi:10.1111/2041-210X.12781.

Wright SI, Agrawal N, Bureau TE. 2003. Effects of Recombination Rate and Gene Density on Transposable Element Distributions in *Arabidopsis thaliana*. *Genome Res.* 13(8):1897–1903. doi:10.1101/gr.1281503.

Zanni V, Eymery A, Coiffet M, Zytnicki M, Luyten I, Quesneville H, Vaury C, Jensen S. 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. *Proc Natl Acad Sci USA.* 110(49):19842–19847. doi:10.1073/pnas.1313677110.

Zhang Y-Y, Zhang D-Y, Barrett S. 2010. Genetic uniformity characterizes the invasive spread of water hyacinth (*Eichhornia crassipes*), a clonal aquatic plant. *Molecular ecology.* 19:1774–86. doi:10.1111/j.1365-294X.2010.04609.x.