1   Rapid evolutionary diversification of the *flamenco* locus across simulans clade *Drosophila*

2   species

3

4

5

6

7

8   Sarah Signor[1*], Jeffrey Vedanayagam[2], Bernard Y. Kim[3], Filip Wierzbicki[4,5], Robert Kofler[4],

9   and Eric C. Lai[2]

10

11

12

13   *Corresponding author: sarah.signor@ndsu.edu

14

15

16   [1]Biological Sciences, North Dakota State University, Fargo, North Dakota, USA

17   [2]Developmental Biology Program, Sloan-Kettering Institute, 430 East 67th St, ROC-10, New

18   York, NY 10065, USA

19   [3]Department of Biology, Stanford University, Stanford, CA, USA

20   [4]Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria

21   [5]Vienna Graduate School of Population Genetics, Vienna, Austria

22

23

24

Abstract

26   Effective suppression of transposable elements (TEs) is paramount to maintain genomic

27   integrity and organismal fitness. In *D. melanogaster*, *flamenco* is a master suppressor of TEs,

28   preventing their movement from somatic ovarian support cells to the germline. It is transcribed

29   by Pol II as a long (100s of kb), single-stranded, primary transcript, that is metabolized into

30   Piwi-interacting RNAs (piRNAs) that target active TEs via antisense complementarity. *flamenco*

31   is thought to operate as a trap, owing to its high content of recent horizontally transferred TEs

32   that are enriched in antisense orientation. Using newly-generated long read genome data, which

33   is critical for accurate assembly of repetitive sequences, we find that *flamenco* has undergone

34   radical transformations in sequence content and even copy number across *simulans* clade

35   Drosophilid species. *D. simulans flamenco* has duplicated and diverged, and neither copy

36   exhibits synteny with *D. melanogaster* beyond the core promoter. Moreover, *flamenco*

37   organization is highly variable across *D. simulans* individuals. Next, we find that *D. simulans*

38   and *D. mauritiana flamenco* display signatures of a dual-stranded cluster, with ping-pong signals

39   in the testis and/or embryo. This is accompanied by increased copy numbers of germline TEs,

40   consistent with these regions operating as functional dual stranded clusters. Overall, the physical

41   and functional diversity of *flamenco* orthologs is testament to the extremely dynamic

42   consequences of TE arms races on genome organization, not only amongst highly related

43   species, but even amongst individuals.

44

45

46

47   **Introduction**

48   *Drosophila* gonads exemplify two important fronts in the conflict between transposable elements

49   (TEs) and the host – the germline (which directly generates gametes), and somatic support cells

50   (from which TEs can invade the germline) (1, 2). The strategies by which TEs are suppressed in

51   these settings are distinct (3), but share their utilization of piwi-interacting RNAs (piRNAs).

52   These are ~24-32 nt RNAs that are bound by the PIWI subclass of Argonaute proteins, and guide

53   them and associated cofactors to targets for transcriptional and/or post-transcriptional silencing

54   (4–7).

55   Mature piRNAs are processed from non-coding piRNA cluster transcripts, which derive

56   from genomic regions that are densely populated with TE sequences (7–9). However, the

57   mechanisms of piRNA biogenesis differ between gonadal cell types. In the germline, piRNA

58   clusters are transcribed from both DNA strands through non-canonical Pol II activity (6, 10–12),

59   which is initiated by chromatin marks rather than specific core promoter motifs. Moreover, co-

60   transcriptional processes such as splicing and polyadenylation are suppressed within dual strand

61   piRNA clusters (13, 14). On the other hand, in ovarian somatic support cells, piRNA clusters are

62   transcribed from a typical promoter as a single stranded transcript, which can be alternatively

63   spliced as with protein-coding mRNAs (15–18). These rules derive in large part from the study

64   of model piRNA clusters (i.e. the germline *42AB* and somatic *flamenco* piRNA clusters). For

65   both types, their capacity to repress invading TEs is thought to result from random integration of

66   new transposons into the cluster (19). As such, piRNA clusters are adaptive loci that play central

67   roles in the conflict between hosts and TEs.

68   The location and activity of germline piRNA clusters are stochastic and evolutionarily

69   dynamic, as there are many copies of TE families in different locations that may produce

70   piRNAs (9, 20). By contrast, somatic piRNA clusters are not redundant and a single insertion of

71   a TE into a somatic piRNA cluster should be sufficient to prevent that TE from further

72   transposition (1, 18). Thus, *flamenco* should contain only one copy per TE family (18), which is

73   true in the *flamenco* locus of *D. melanogaster* (18). *flamenco* is also the only piRNA cluster

74   which produces a phenotypic effect when altered, as germline clusters can be deleted with no

75   consequences (9).

76   *flamenco* has been a favored model for understanding the piRNA pathway since the

77   discovery of piRNA mediated silencing of transposable elements (6). *flamenco* spans >180 kb of

78   repetitive sequences located in *β*-heterochromatin of the X chromosome (21). Of note, *flamenco*

79   was initially identified, prior to the formal recognition of piRNAs, via transposon insertions that

80   de-repress *gypsy*, *ZAM*, and *Idefix* class elements (21–25). These mutant alleles disrupt the

81   *flamenco* promoter, and consequently abrogate transcription and piRNA production from this

82   locus. By contrast, the recent deletion of multiple model germline piRNA clusters, which

83   eliminate the biogenesis of a bulk of cognate piRNAs, did not de-repress their cognate TEs (9).

84   Thus, the analysis of *flamenco* evolution is presumably more consequential for TE dynamics.

85   Analysis of *flamenco* in various strains of *D. melanogaster* supports that this locus traps

86   horizontally derived TEs to achieve silencing of newly invaded TEs (18). The *flamenco* locus

87   exhibits synteny across the *D. melanogaster* sub-group (26); however, the sequence composition

88   of *flamenco* outside *D. melanogaster* has not been well-characterized (27).

89   In this study, we compare the *flamenco* locus across 10 strains of simulans-clade species,

90   namely *D. simulans*, *D. mauritiana*, and *D. sechellia*. Analysis of piRNAs from ovaries of five

91   genotypes of *D. simulans* found that *flamenco* is duplicated in *D. simulans*. This duplication is

92   old enough that there is no sequence synteny across copies, even though their core promoter

93   regions and the adjacent *dip1* gene duplications are conserved. *flamenco* has also been colonized

94   by abundant (>40) copies of *R1*, a TE that was thought to insert only at ribosomal genes, and to

95   evolve at the same rate as nuclear genes (28). Furthermore, between different genotypes, up to

96   63% of TE insertions are not shared within any given copy of *flamenco*. Despite this, several full

97   length TEs are shared between all genotypes in a similar sequence context. This incredible

98   diversity at the *flamenco* locus, even within a single species, suggests there may be considerable

99   variation in its ability to suppress transposable elements across individuals.

100   Cross-species comparisons further indicate that functions of *flamenco* have diversified.

101   Data from *D. sechellia* and *D. melanogaster* conform with the current understanding of *flamenco*

102   as a uni-strand cluster. However, we find evidence that *D. simulans* and *D. mauritiana flamenco*

103   can act as a dual strand cluster in testis (*D. mauritiana*) and embryos (*D. mauritiana* and *D.*

104   *simulans*), yielding piRNAs from both strands with a ping pong signal. Overall, we infer that the

105   rapid evolution of *flamenco* alleles across individuals and species reflects highly adaptive

106   functions and dynamic biogenesis capacities.

107   **Materials and Methods**

108   *Fly strains*

109   The four *D. simulans* lines *SZ232*, *SZ45*, *SZ244*, and *SZ129* were collected in California from the

110   Zuma Organic Orchard in Los Angeles, CA on two consecutive weekends of February 2012 (29–

111   33). *LNP-15-062* was collected in Zambia at the Luwangwa National Park by D. Matute and

112   provided to us by J. Saltz (J. Saltz pers. comm., (34, 35)). *MD251*, *MD242*, *NS137*, and *NS40*

113   were collected in Madagascar and Kenya (respectively) and are described in (36). The *D.*

114   *simulans* strain *wxD¹* was originally collected by M. Green, likely in California, but its

115   provenance has been lost (pers. comm. Jerry Coyne). *D. mauritiana (w12)* and *D. sechellia*

116   *(Rob3c/Tucson 14021-0248.25)* are described in (37).

117   *Long read DNA sequencing and assembly*

118   *MD242*, four SZ lines and *LNP-15-062* were sequenced on a MinION platform at North Dakota

119   State University (Oxford Nanopore Technologies (ONT), Oxford, GB), with base-calling using

120   guppy (v4.4.2). *MD242*, the four SZ lines, and *LNP-15-062* were assembled with Canu (v2.1)

121   (38) and two rounds of polishing with Racon (v1.4.3) (39). The CA strains were additionally

122   polished with short reads using Pilon (v1.23) (40)(SRR3585779, SRR3585440, SRR3585480,

123   SRR3585391) (29). The first wxD^{1-1} assembly is described here (41). *MD251*, *NS137*, *NS40* and

124   wxD^{1-2} were sequenced on a MinION platform at Stanford University. They were assembled with

125   Flye (42), and polished with a round of Medaka followed by a round of pilon (40). Following

126   this contaminants were removed with blobtools (https://zenodo.org/record/845347, (43)), soft

127   masked with RepeatModeler and Repeatmasker (44, 45), then aligned to the wxD^1 as a reference

128   with Progressive Cactus (46). The assemblies were finished with reference based scaffolding

129   using Ragout (47). *D. mauritiana* and *D. sechellia* were sequenced with PacBio and assembled

130   with FALCON using default parameters

131   (https://github.com/PacificBiosciences/FALCON)(37).The *D. melanogaster* assembly is

132   described here (48). A summary of the assembly statistics is available in Supplementary Table 1.

133   The quality of cluster assembly was evaluated using the coverage and soft clip quality as

134   described in (20, 49) (Supplementary File 1).

135   *Short read sequencing and mapping*

136   Short read sequencing was performed by Beijing Genomics Institute on approximately 50 dissected

137   ovaries from adult female flies (*SZ45, SZ129, SZ232, SZ244, LNP-15-062*). Short read libraries from 0-2

138   hour embryos were prepared from *D. melanogaster*, wxD^{1-2}, *D. sechellia*, and *D. mauritiana* (SRAXXX)

139   (50). Small RNA from testis is described in (51, 52). *D. melanogaster* OSC small RNA libraries were

140   downloaded from the SRA (SRR11999160). Libraries were filtered for adapter contamination and short

141   reads between 23-29 bp were retained for mapping with fastp (53). The RNA was then mapped to

142   their respective genomes using bowtie (v1.2.3) and the following parameters (-q -v 1 -p 1 -S -a -

143   m 50 --best --strata) (54, 55). The resulting bam files were processed using samtools (56). To

144   obtain unique reads the bam files were filtered for reads with 1 mapping position. To obtain

145   counts files with weighted mapping the bam files were processed using Rsubreads and the

146   featureCounts function (57).

147   *Defining and annotating piRNA clusters*

148   piRNA clusters were initially defined using proTRAC (58). piRNA clusters were predicted with

149   a minimum cluster size of 1 kb (option "-clsize 1000"), a p-value for minimum read density of

150   0.07 (option "-pdens 0.07"), a minimum fraction of normalized reads that have 1T (1U) or 10A

151   of 0.33 (option "-1Tor10A 0.33") and rejecting loci if the top 1% of reads account for more than

152   90% of the normalized piRNA cluster read counts (option "-distr 1-90"), and a minimal fraction

153   of hits on the main strand of 0.25 (option "-clstrand 0.25"). Note that this ties the piRNA clusters

154   to their function such that participation in the ping pong pathway can be inferred from these

155   patterns. Clusters were annotated using RepeatMasker (v. 4.0.7) and the TE libraries described in

156   Chakraborty et al. (2019) (41, 44). The position of *flamenco* was also evaluated based off of the

157   position of the putative promoter, the *dip1* gene, and the enrichment of *gypsy* elements (15).

158   Fragmented annotations were merged to form TE copies with onecodetofindthemall (59).

159   Fragmented annotations were also manually curated within *flamenco*, particularly because TEs

160   not present in the reference library often have their LTRs and internal sequences classified as

161   different elements.

162   *Aligning the flamenco promoter region*

163   The region around the *flamenco* promotor was extracted from each genotype and species with

164   bedtools getfasta (60). Sequences were aligned with clustal-omega and converted to nexus

165   format (61). Trees were built using a GTR substitution model and gamma distributed rate

166   variation across sites (62). Markov chain monte carlo iterations were run until the standard

167   deviation of split frequencies was below .01, around one million generations. The consensus

168   trees were generated using sumt conformat=simple. The resulting trees were displayed with the

169   R package ape (63).

170   *Detecting ping pong signals in the small RNA data*

171   Ping pong signals were detected using pingpongpro (64) This program detects the presence of

172   RNA molecules that are offset by 10 nt, such that stacks of piRNA overlap by the first 10 nt from

173   the 5' end. These stacks are a hallmark of piRNA mediated transposon silencing. The algorithm

174   also takes into account local coverage and the presence of an adenine at the 10th position. The

175   output includes a z-score between 0 and 1, the higher the z-score the more differentiated the ping

176   pong stacks are from random local stacks.

177   *Annotating shared and unique TE insertions*

178   To align the TE annotations of homologous piRNA clusters, we first extracted the sequences of

179   the clusters and annotated TEs in these sequences using RepeatMasker (open-4.0.7) with a

180   custom TE library and the parameters: -s (sensitive search), -nolow (disable masking of low

181   complexity sequences), and -no_is (skip check for bacterial IS) (37, 65, 66). Finally, we aligned

182   the resulting repeat annotations with Manna using the parameters -gap 0.09 (gap penalty), -mm

183   0.1 (mismatch penalty) -match 0.2 (match score) (20, 37). Manna can be used for aligning the

184   annotations of the transposable elements by relying on synteny to determine insertion homology.

185   Alignments were manually checked for inconsistencies arising from assignment to similar TEs

186   (i.e. *gypsy-3* versus *gypsy-5*) . TEs were considered to be full length if they were present in at

187   least 70% of their reference length and contained internal sequence as well as two LTRs if

188   applicable.

189   **Results**

190   *flamenco in the D. simulans clade*

191   We identified *D. simulans flamenco* from several lines of evidence: piRNA cluster calls from

192   proTRAC, its location adjacent to divergently transcribed *dip1*, the existence of conserved core

193   *flamenco* promoter sequences, and enrichment of *gypsy* elements (Figure 1 & 2); Supplementary

194   Table 2). The *flamenco* locus is at least 376 kb in *D. simulans*. This is an expansion compared

195   with *D. melanogaster*, where *flamenco* is only 156 kb (*Canton-S*). In *D. sechellia flamenco* is

196   363 kb, however in *D. mauritiana* the locus has expanded to at least 840 kb (Supplementary

197   Table 2). This is a large expansion, and it is possible that the entire region does not act as a

198   region controlling somatic TEs. However, evidence that is does include uniquely mapping

199   piRNAs that are found throughout the region and *gypsy* enrichment consistent with a *flamenco-*

200   like locus (Supplementary Figure 1). There are no protein coding genes within the region, and

201   while the neighboring genes on the downstream side of *flamenco* in *D. melanogaster* have

202   moved in *D. mauritiana* (*CG40813- CG41562* at 21.5 MB), the following group of genes

203   beginning with *CG14621* is present and flanks *flamenco* as it is annotated. Thus in *D.*

204   *melanogaster* the borders of *flamenco* are flanked by *dip1* upstream and *CG40813* downstream,

205   while in *D. mauritiana* they are *dip1* upstream and *CG14621* downstream. Between all species

206   the *flamenco* promoter and surrounding region, including the *dip1* gene, are alignable and

207   conserved (Figure 1D).

208   *Structure of the flamenco locus*

209   *D. melanogaster flamenco* bears a characteristic structure, in which the majority of TEs

210   are *gypsy*-class elements in the antisense orientation (79% antisense orientation, 85% of which

211   are *gypsy* elements) (Figure 2C; Supplementary Table 3). In *D. simulans*, *flamenco* has been

212   colonized by large expansions of *R1* transposable element repeats such that on average the

213   percent of antisense TEs is only 50% and the percent of the locus comprised of LTR elements is

214   55%. However, 76% of antisense insertions are LTR insertions, thus the underlying *flamenco*

215   structure is apparent when the *R1* insertions are disregarded (Figure 2C). In *D. mauritiana*

216   *flamenco* is 71% antisense, and of those antisense elements it is 85% LTRs. Likewise in *D.*

217   *sechellia* 78% of elements are antisense, and of those 81% are LTRs. *flamenco* retains the overall

218   structure of a canonical *D. melanogaster*-like *flamenco* locus in all of these species, however in

219   *D. simulans* the nature of the locus is somewhat altered by the abundant *R1* insertions (Figure

220   2C).

221   *flamenco is duplicated in D. simulans*

222   In *D. simulans*, we unexpectedly observed that *flamenco* is duplicated on the X

223   chromosome; the duplication was confirmed with PCR and a restriction digest (Figure 1,

224   Supplementary File 2). These duplications are associated with a conserved copy of the putative

225   *flamenco* enhancer as well as copies of the *dip1* gene located proximal to *flamenco* in *D.*

226   *melanogaster* (Figure 1, 3A). While it is unclear which copy is orthologous to *D. melanogaster*

227   *flamenco*, all *D. simulans* lines bear one copy that aligns across genotypes. We refer to this copy

228   as *D. simulans flamenco*, and the other copies as duplicates. Otherwise, *flamenco* duplicates do

229   not align with one another and lack synteny amongst their resident TEs. Possible evolutionary

230   scenarios are that the *flamenco* duplication occurred early in the *simulans* lineage, that the

231   clusters evolved very rapidly, or that the duplication encompassed only the promoter region and

232   was subsequently colonized by TEs (Figure 1A, 3A).

233   The *flamenco* duplicate is absent in the *D. simulans* reference assembly, w^{501}

234   (GCA_000754195.3), but present in wxD^{1}, suggesting it was polymorphic or absent between the

235   collection of these strains (or was not assembled). The duplicate retains the structure of

236   *flamenco*, with an average of 67% of TEs in the antisense orientation, and 91% of the TEs in the

237   antisense orientation are LTRs. The duplicate of *flamenco* is less impacted by *R1*, with some

238   genotypes having as few as 8 *R1* insertions (Figure 3C).

239   *R1 LINE elements at the flamenco locus*

240   *R1* elements are well-known to insert into rDNA genes, are transmitted vertically, and evolve

241   similarly to the genome background rate (28). They have also been found outside of rDNA

242   genes, but only as fragments. *R1* elements are absent from *flamenco* in the *D. simulans* reference

243   assembly, aside from a single fragmented *R1-1* element (w^{501}). However, as mentioned, *R1*

244   elements are abundant within *flamenco* loci in the *simulans* clade. Outside of *flamenco*, *R1*

245   elements in *D. simulans* are distributed according to expectation, with full length elements

246   occurring only within rDNA (Supplementary File 3). Within *flamenco*, most copies of *R1* occur

247   as tandem duplicates, creating large islands of fragmented *R1* copies (Figure 3A). They are on

248   average 3.7% diverged from the reference R1 from *D. simulans*. Across individual *D. simulans*

249   genomes, ~99 kb of *flamenco* loci consists of *R1* elements, i.e. 26% of their average total length.

250   *SZ45, LNP-15-062, NS40, MD251*, and *MD242* contain 4-7 full length copies of *R1* in the sense

251   orientation, even though all but *SZ45* bear fragmented *R1* copies on the antisense strand. (The

252   *SZ45 flamenco* assembly is incomplete). As the antisense *R1* copies are expected to suppress *R1*

253   transposition, *flamenco* may not suppress these elements effectively. Alternatively, it is possible

254   that *D. simulans flamenco* is still mostly active in the soma, while *R1* is active in the germline,

255   and thus escapes host control by *flamenco*.

256   In *D. mauritiana*, *flamenco* harbors abundant fragments or copies of *R1* (19 on the

257   reverse strand and 20 on the forward strand), and only one large island of *R1* elements. In total,

258   *D. mauritiana* contains 84 kb of *R1* sequence within *flamenco*. In *D. mauritiana* there are 8 full

259   length copies of *R1* at the *flamenco* locus, 7 in antisense, which are not obviously due to a

260   segmental or local duplication. Finally, we find that *D. sechellia flamenco* lacks full length

261   copies of *R1*, and it contains only 18 KB of *R1* sequence (16 fragments on the reverse strand).

262   Yet, all the copies are on the sense strand, which would not produce fragments that can suppress

263   R1 TEs. Essentially the antisense copies of *R1* in *D. mauritiana* should be suppressing the TE,

264   but we see multiple full length antisense insertions, and *D. sechellia* has no antisense copies, but

265   we see no evidence for recent *R1* insertions. From this it would appear that whatever is

266   controlling the transposition of *R1* lies outside of *flamenco*.

267   The presence of long sense-strand *R1* elements within *flamenco* is a departure from

268   expectation (18, 28). There is no evidence of an rDNA gene within the *flamenco* locus that

269   would explain the insertion of *R1* elements there, nor is there precedence for the large expansion

270   of *R1* fragments within the locus. Furthermore, the suppression of *R1* transposition does not

271   appear to be controlled by *flamenco*.

272   *piRNA production from R1*

273   On average *R1* elements within the *flamenco* locus of *D. simulans* produce more piRNA

274   than any other TE within *flamenco* (Supplementary Table 6). *R1* reads mapping to the forward

275   strand constitute an average of 51% of the total piRNAs within the *flamenco* locus from the

276   maternal fraction, ovary, and testis using weighted mapping. The only exception is the ovarian

277   sample from *SZ232* which is a large outlier at only 5%. However reads mapping to the reverse

278   strand account for an average of 84% of the piRNA being produced from the strand in every

279   genotype and tissue – maternal fraction, testis, or ovary. If unique mapping is considered instead

280   of weighted these percentages are reduced by approximately 20%, which is to be expected given

281   that *R1* is present in many repeated copies. Production of piRNA from the reverse strand seems

282   to be correlated with elements inserted in the sense orientation, of which the vast majority are *R1*

283   elements in *D. simulans* (Supplementary Figure 2). The production of large quantities of piRNA

284   cognate to the *R1* element is seemingly pointless – if *R1* only inserts at rDNA genes and are

285   vertically transmitted there is little reason to be producing the majority of piRNA in response to

286   this element.

287   In *D. sechellia* there are very few piRNA produced from *flamenco* in these tissues, and

288   there are no full length copies of *R1*. Likewise overall weighted piRNA production from *R1*

289   elements on either strand is 2.8-5.9% of the total mapping piRNA. In contrast in *D. mauritiana*

290   there are full length *R1* elements and abundant piRNA production in the maternal fraction and

291   testis. In *D. mauritiana* an average of 28% of piRNAs mapping to the forward strand of *flamenco*

292   are arising from *R1*, and 33% from the reverse strand. In *D. mauritiana R1* elements make up a

293   smaller proportion of the total elements in the sense orientation (24%), versus *D. simulans*

294   (55%).

295   *Conservation of flamenco*

296   The *dip1* gene and promoter region adjacent to each copy of *flamenco* are very conserved both

297   within and between copies of *flamenco* (Figure 3A). The phylogenetic tree of the area suggests

298   that we are correct in labeling the two copies as the original *flamenco* locus and the duplicate

299   (Figure 3A). The original *flamenco* locus is more diverged amongst copies while the duplicate

300   clusters closely together with short branch lengths (Figure 3A). The promotor region is also

301   conserved and alignable between *D. melanogaster*, *D. sechellia*, *D. mauritiana*, and *D. simulans*

302   (Figure 1D). However, the same is not true of the *flamenco* locus itself. Approximately 3 kb

303   from the promoter *flamenco* diverges amongst genotypes and species and is no longer alignable

304   by traditional sequence-based algorithms, as the TEs are essentially presence/absence

305   polymorphisms that span multiple kb. There is no conservation of *flamenco* between *D.*

306   *melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* (Figure 4). However, within the

307   *simulans* clade many of the same TEs occupy the locus, suggesting that they are the current

308   genomic invaders in each of these species (Figure 4).

309   In *D. simulans* the majority of full length TEs are singletons – 54% in *flamenco* and 64%

310   in the duplicate. Copies that are full length in one genotype but fragmented in others are counted

311   as shared, not singletons. Almost half of these singletons in the duplicate are due to a single

312   genotype with a unique section of sequence, in this case *MD251*. Singletons are the single largest

313   category of transposable element insertions, followed by fixed insertions. Thus even within a

314   single population there is considerable diversity at the *flamenco* locus, and subsequently

315   diversity in the ability to suppress transposable elements. For example, *gypsy-29* is present in

316   three genotypes either in *flamenco* or the duplicate, which would suggest that these genotypes

317   are able to suppress this transposable element while the other genotypes are not. In contrast

318   *gypsy-3* is present in more than one full length copy in *flamenco* and its duplicate in every

319   genotype but one where it is present in a single copy. There are a number of these conserved full

320   length TEs that are present in all or nearly all genotypes, including *Chimpo*, *gypsy-2*, *Tirant*, and

321   *gypsy-4*. In addition, the *INE1* elements adjacent to the promoter are conserved.

322   It is notable that any full length TEs are shared across all genotypes, given that *wxD¹* was

323   like collected 30-50 years prior to the others, and the collections span continents (Figure 3C).

324   Two facts are relevant to this observation: (1) TEs were shown not to correlate with geography

325   (67) and (2) *D. simulans* is more diverse within populations than between different populations

326   (68–70). Other explanations are also plausible. Selection could be maintaining these full length

327   TEs, *wxD¹* could have had introgression from other lab strains, or a combination of these

328   explanations.

329   *Suppression of TEs by the flamenco locus and the trap model of TE control*

330   In *D. melanogaster*, it was proposed that while germline clusters may have many insertions of a

331   single TE, the somatic 'master regulator' *flamenco* will have a single insertion of each

332   transposon, after which they are silenced and no longer able to transpose (18).

333   Here, we evaluate the following lines of evidence to determine if they support the trap model of

334   transposable element suppression. (1) How many TEs have antisense oriented multicopy

335   elements within *flamenco*? (2) How many TEs have full length and fragmented insertions,

336   suggesting the older fragments did not suppress the newer insertion? (3) How many *de novo*

337   insertions of TEs in the *flamenco* duplicate of *D. simulans* are also present in the original

338   *flamenco* copy?

339   How many TEs have antisense oriented multicopy elements within *flamenco*?

340   Due to the difficulty in classifying degraded elements accurately, for example between multiple

341   classes of *gypsy* element, we will focus here on full length TEs, suggesting recent transposition.

342   In *D. melanogaster* there are 7 full length elements, none of which are present in more than one

343   antisense copy. These elements make up 27% of the *flamenco* locus. Full length copies of five of

344   these elements were also reported previously for other strains of *D. melanogaster* (18)

345   In *D. sechellia* there are 14 full length TEs within the *flamenco* locus, three of which are

346   present in multiple copies. Two of these, *INE1* and *412*, are likely present due to local

347   duplication. In particular the *INE1* elements flanking the promoter, are in the sense orientation,

348   and are conserved between *D. sechellia, D. mauritiana*, and *D. simulans*. The only element

349   present in multiple antisense copies is *GTWIN*. Similar to *D. melanogaster* these 14 elements

350   make up 27% of the *flamenco* locus.

351   *D. mauritiana* contains 22 full length TEs within the *flamenco* locus. Four of these are

352   present in multiple antisense full length copies – *INE1*, *R1*, *Stalker-4*, and *Cr1a*. While some of

353   the five antisense copies of *R1* likely originated from local duplications – they are in the same

354   general region and tend to be flanked by *gypsy-8*, not all of them show these patterns.

355   Furthermore, as aforementioned, there also are full length sense copies of *R1* suggesting *R1* is

356   not being suppressed by *flamenco*. *gypsy-12* and *gypsy-3* have a second antisense copy within

357   *flamenco* that is just below the cutoff to be considered full length – in *gypsy-3* the second copy is

358   10% smaller, for *gypsy-12* it is present but missing an LTR. Full length TEs make up 19% of the

359   *flamenco* locus.

360   In *D. simulans* there are 24 full length TEs present in any of the seven complete *flamenco*

361   assemblies. Six of these are present in multiple antisense copies within a single genome – *INE1*,

362   *Chimpo*, *gypsy-4*, *412*, *Tirant*, and *BEL-unknown*. The two *Tirant* copies are likely a segmental

363   duplication as they flank an *R1* repeat region. In addition, most *INE1* copies are present proximal

364   to the promoter as aforementioned. *Chimpo* is present in three full length copies within *MD242*

365   *flamenco*, with no evidence of local duplication. While there are no full length copies of *R1*

366   inserted in antisense, *R1* is present in full length sense copies despite many genomes containing

367   antisense fragments, suggesting *flamenco* is not suppressing *R1*. On average full length TEs

368   constitute 20% of *flamenco* in *D. simulans*.

369   In the duplicate of *flamenco* in *D. simulans* there are 30 full length TEs present in any

370   one of the five complete *flamenco* duplicate assemblies. However, none of them are multicopy in

371   antisense. However, they are multicopy relative to the original copy of *flamenco*. *gypsy-3*, *BEL-*

372   *unknown*, *Nomad-1*, *Chimpo*, *gypsy-53A*, *R1*, and *INE1* are all multicopy with respect to the

373   original *flamenco* within a given genome. Some of these may have been inherited at the time of

374   duplication, however are full length in both copies suggesting recent transposition. In the

375   duplicate of *flamenco* full length TEs occupy an average of 17% of the locus. *MD251* is an

376   exception which weights the average, with 28% of the locus, while between 10 and 15% is found

377   for the remaining copies. Thus *D. simulans* and *D. mauritiana* overall do not meet the

378   expectation that *flamenco* will contain a single insertion of any given TE.

379   How many TEs have full length and fragmented insertions?

380   Full length elements are younger insertions than fragmented insertions. If a full length element is

381   inserted in *flamenco* and there are fragments in the antisense orientation elsewhere in *flamenco*

382   this indicates that *flamenco* did not successfully suppress the transposition of this element.

383   In *D. melanogaster* two elements have fragments in antisense and a full length TE – *Doc*

384   and *Stalker-2*. *D. sechellia* has 9 elements that are present as a full length TE and a fragment in

385   antisense (including *412*, *GTWIN*, *mdg-1*, and *nomad*) and 6 that are multicopy that are due to a

386   solo LTR (including *blood*, *297*, and *Stalker-4*). *D. mauritiana* has 21 elements that are present

387   in full length and a fragment in antisense (including *blood*, *412*, *gypsy-10-13*, and *R1*), and four

388   elements that are multicopy due to a solo LTR (*mdg-1*, *Idefix*, and *gypsy-7,10).*

389   In *D. simulans*, TEs that fit this criteria in *flamenco* include *gypsy-2*, *gypsy-3*, *gypsy-4*,

390   *gypsy-5*, *Chimpo*, *412*, *INE1*, *R1*, *Tirant*, and *Zam*. *297* and *Nomad-1* are present in full length

391   copies but only multi-copy in the context of solo LTRs. In the duplicate of *flamenco* in *D.*

392   *simulans* this includes *gypsy-2*, *gypsy-3*, *gypsy-5*, *297*, *Stalker-4*, and *R1*. For example in *NS40*

393   there are 7 full length copies of *R1* in the sense orientation that likely duplicated in place, as well

394   as 12 partial copies in the antisense orientation. In the *simulans* clade either fragments of TEs are

395   not sufficient to suppress transposable elements or some elements are able to transpose despite

396   the hosts efforts to suppress them.

397   *Is flamenco a trap for TEs entering through horizontal transfer?*

398   High sequence similarity between TEs in different species suggests horizontal transfer (71).

399   However, because sequence similarity can also exist due to vertical transmission we will use

400   sequence similarity between R1 elements (inserted at rDNA genes) as a baseline for

401   differentiating horizontal versus vertical transfer. There has never been any evidence found for

402   horizontal transfer of *R1* and it is thought to evolve at the same rate as nuclear genes in the

403   *melanogaster* subgroup (18, 28). Of the full length elements present in any genome at *flamenco*

404   62% of them appear to have originated from horizontal transfer. This is similar to previous

405   estimates for *D. melanogaster* in other studies (18). Transfer appears to have occurred primarily

406   between *D. melanogaster*, *D. sechellia*, and *D. willistoni*. This includes some known horizontal

407   transfer events such as *Chimpo* and *Chouto* (72), and others which have not been recorded such

408   as *gypys-29* (*D. willistoni*) and the *Max-element* (*D. sechellia*) (Supplementary File 4). The

409   duplicate of *flamenco* is similar, with 53% of full length TEs originating from horizontal transfer.

410   They are many of the same TEs, with a 46% overlap, thus *flamenco* and its duplicate are trapping

411   many of the same TEs. Both *flamenco* and the duplicate the region appears to serve as a trap for

412   TEs originating from horizontal transfer.

413   In *D. melanogaster* 85% of full length TEs appear to have arisen through horizontal

414   transfer, which is consistent with previous estimates (18). In *D. sechellia* 53% of full length TEs

415   have arisen from horizontal transfer, including some known to have moved by horizontal transfer

416   such as *GTWIN* (*D. melanogaster*/*D. erecta*) (72). *D. mauritiana* has 68% of its full length TEs

417   showing a closer relationship than expected by vertical descent with TEs from *D. sechellia*, *D.*

418   *melanogaster*, and *D. simulans*. The hypothesis that *flamenco* serves as a trap for TEs entering

419   the population through horizontal transfer holds throughout the *simulans* clade.

420   *Flamenco piRNA is expressed in the testis and the maternal fraction*

421   Canonically, *flamenco* piRNA is expressed in the somatic follicular cells of the ovary and

422   not in the germline, and also does not produce a ping pong signal (24). It was not thought to be

423   present in the maternal fraction of piRNAs or other tissues. However, that appears to be variable

424   in different species (Figure 5). We examined single mapping reads in the *flamenco* region from

425   testes and embryos (maternal fraction) in *D. simulans*, *D. mauritiana*, *D. sechellia*, and *D.*

426   *melanogaster*. As a control we also included *D. melanogaster* ovarian somatic cells, where Aub

427   and Ago3 are not expressed and therefore there should be no ping pong signals. In *D. simulans*

428   and *D. mauritiana flamenco* is expressed bidirectionally in the maternal fraction and the testis,

429   including ping pong signals on both strands (Figure 5A & C; Supplementary Figure 1). In *D.*

430   *sechellia*, there is no expression of *flamenco* in either of these tissues. Discarding multimappers

431   in the maternal fraction 63% (*D. mauritiana*) – 36% (*D. simulans*) of the ping pong signatures on

432   the X with a z-score of at least 0.9 are located within *flamenco* (Figure 5C). In the testis the

433   picture is more complicated – in *D. mauritiana* 50% of ping pong signals on the X with a z-score

434   of at least 0.9 are located within *flamenco*, which amounts to a substantial ping pong signature

435   (Supplementary Figure 1). While mapping of piRNA to both strands was observed in *D.*

436   *simulans* testis, there is very little apparent ping pong activity (5 positions in *flamenco* $z > 0.9$;

437   15 potential ping pong signals on the X). In *D. melanogaster*, there is uni-strand expression in

438   the maternal fraction, but it is limited to the region close to the promoter. In *D. melanogaster* no

439   ping pong signals have a z-score above 0.8 in the maternal fraction or the ovarian somatic cells.

440   There are ping pong stacks in *flamenco* in the testis of D. melanogaster (2% of the total on the

441   contig), however they are limited to a single region and are not abundant enough to be strong

442   evidence of ping pong activity.

443   In the duplicate of *flamenco* in the maternal fraction 15% of the ping pong signals with a

444   z-score above 0.9 on the X are within the *flamenco* duplicate. The *flamenco* duplicate does not

445   have a strong signal of the ping pong pathway in the testis. In addition, *flamenco* in these species

446   has been colonized by full length TEs thought to be active in the germline such as *blood*,

447   *burdock, mdg-3*, *Transpac*, and *Bel* (73, 74). *blood* is also present in *D. melanogaster* in a full

448   length copy while there is no evidence of germline activity for *flamenco* in *D. melanogaster*,

449   though no other putative germline TEs are present. The differences in ping pong signals between

450   species and the presence of germline TEs in *D. simulans* and *D. mauritiana* suggests that the role

451   of *flamenco* in these tissues has evolved between species.

452   **Discussion**

453   The piRNA pathway is the organisms primary mechanism of transposon suppression.

454   While the piRNA pathway is conserved, the regions of the genome that produce piRNA are

455   labile, particularly in double stranded germline piRNA clusters (9). The necessity of any single

456   cluster for TE suppression in the germline piRNA pathway is unclear, but likely redundant (9).

457   However, *flamenco* is thought to be the master regulator of the somatic support cells of the

458   ovary, preventing *gypsy* elements from hopping into germline cells (18, 21, 23, 24, 75, 76). It is

459   not redundant to other clusters, and insertion of a single element into *flamenco* in *D.*

460   *melanogaster* is sufficient to initiate silencing. Here we show that the function of *flamenco*

461   appears to have diversified in the *D. simulans* clade, acting in at least some tissues as a germline

462   piRNA cluster.

463   *Dual stranded expression of flamenco*

464   In this work, we showed that piRNAs of the *flamenco* locus in *D.simulans* and *D.*

465   *mauritiana* are deposited maternally, align to both strands, and exhibit ping-pong signatures.

466   This is in contrast to *D. melanogaster*, where *flamenco* acts as a uni-strand cluster in the soma

467   (3), our data thus suggest that the *flamenco* locus in *D. simulans* and *D. mauritiana* acts as a

468   dual-strand cluster in the germline. In *D. sechellia* the attributes of *flamenco* uncovered in *D.*

469   *melanogaster* appear to be conserved – no expression in the maternal fraction and the testis and

470   no ping pong signals. Given that *flamenco* is likely a somatic uni-strand cluster in *D. erecta*, we

471   speculate that the conversion into a germline cluster happened in the *simulans* clade (3). Such a

472   conversion of a cluster between the somatic and the germline piRNA pathway is not

473   unprecedented. For example, a single insertion of a reporter transgene triggered the conversion

474   of the uni-stranded cluster *20A* in *D. melanogaster* into a dual-strand cluster (77).

475   The role of *flamenco* in *D. simulans* and *D. mauritiana* as the master regulator of piRNA

476   in somatic support cells may still well be true – the promoter region of the *flamenco* cluster is

477   conserved between species and between copies of *flamenco* within species. This suggests that in

478   at least some contexts (or all) the cluster is still serving as a uni-strand cluster transcribed from a

479   traditional RNA Pol II site (15). However it has acquired additional roles, producing dual strand

480   piRNA and ping pong signals, in these two species, in at least the germline. However, in *D.*

481   *simulans*, the majority of these reverse stranded piRNAs are emerging from the *R1* insertions

482   within *flamenco*. There is no evidence at present that *R1* has undergone an expansion in function

483   in *D. simulans*, thus it is unclear what, if any, functional impact the reverse stranded piRNAs

484   have at the *flamenco* locus.

485   *Duplication of flamenco in D. simulans*

486   In *D. simulans*, *flamenco* is present in 2 genomic copies, and this duplication is present in

487   all sequenced *D. simulans* lines except the reference strain. The *dip1* gene and putative *flamenco*

488   promoter flanking the duplication also has a high similarity in all sequenced lines (Fig. 2B). This

489   raises the possibility that the duplication of *flamenco* in *D. simulans* was positively selected.

490   Such a duplication may be beneficial as it increases the ability of an organism to rapidly silence

491   TEs. Individuals with large piRNA clusters (or duplicated ones) will accumulate fewer

492   deleterious TE insertions than individuals with small clusters (or non-duplicated ones), and

493   duplicated clusters may therefore confer a selective advantage (78).

494   *Rapid evolution of piRNA clusters*

495   A previous work showed that dual- and uni-strand clusters evolve rapidly in *Drosophila*

496   (20). In agreement with this work we also found that the *flamenco*-locus is rapidly evolving

497   between and within species (Fig. 1C, 3B). A major open question remains whether this rapid

498   turnover is driven by selection (positive or negative) or an outcome of neutral processes (eg. high

499   TE activity or insertion bias of TEs). These rapid evolutionary changes at the *flamenco* locus, a

500   piRNA master locus, suggest that there is a constant turnover in patterns of piRNA biogenesis

501   that potentially leads to changes in the level of transposition control between individuals in a

502   population.

503

504

Competing interests
We declare that we have no competing interests.

Authors' contributions
SS conceived the study, performed bioinformatics and drafted portions of the manuscript. FW and RK performed bioinformatics and drafted portions of the manuscript. JV contributed data and bioinformatic analysis. ECL drafted portions of the manuscript and provided data. BYK generated and contributed sequence data.
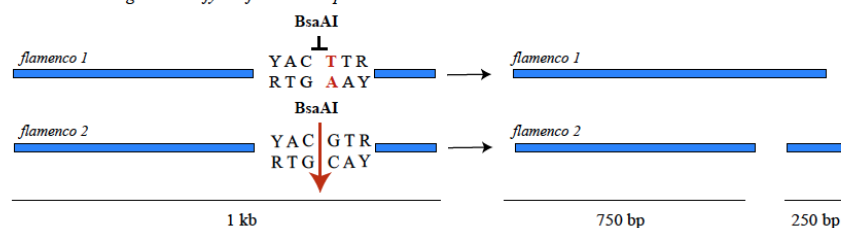
Availability of data and materials
All data has been made available in the following repositories:
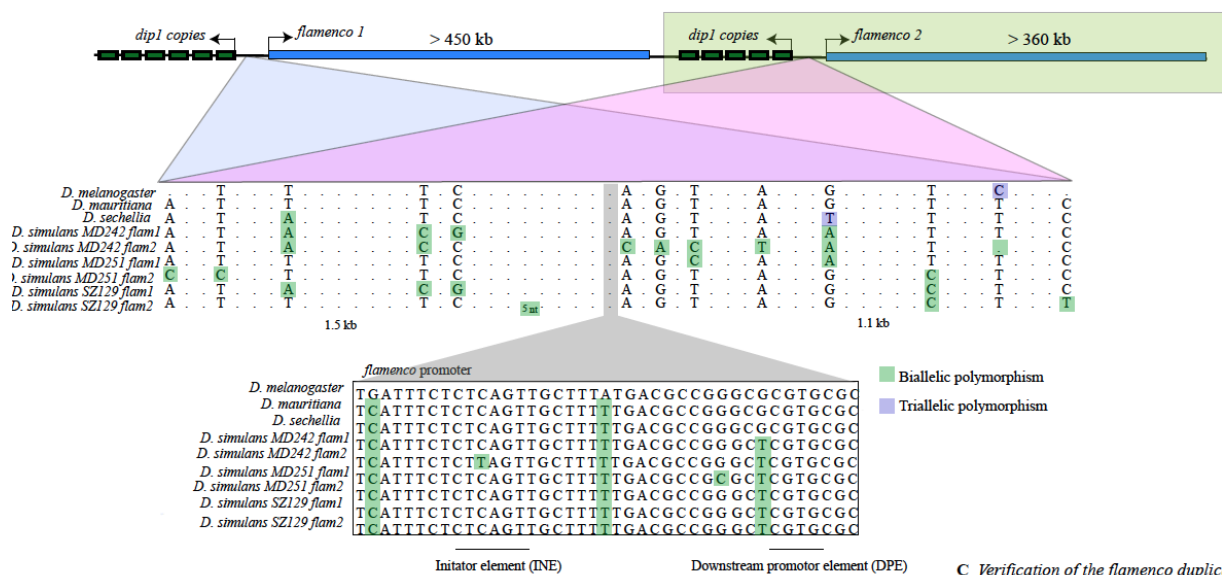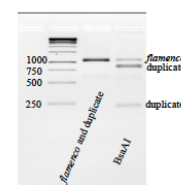
Figure 1. A) The duplication of *flamenco* in the *D. simulans*. Both copies are flanked by copies of the *dip1* gene and copies of the putative *flamenco* promoter. The top portion of the alignment shows ~ 2 kb around the promoter. SNPs are shown if they differentiate copies of *flamenco* within a single genotype of *D. simulans*. Dots do not indicate a single nucleotide, but rather a sequence region where no SNPs differentiate the two copies of *flamenco* within a single genotype. The lower portion illustrates the promoter region with all SNPs illustrated in *D. melanogaster*, *D. sechellia*, *D. mauritiana*, and *D. simulans*. B) A schematic of the restriction digest used to verify the duplicate of *flamenco*. The targeted region is a 1 kb fragment adjacent to the promotor of *flamenco*. Within this region the original *flamenco* copy does not contain a YACGTR site and is not cut by the restriction enzyme BsaAI. The duplicate of *flamenco* is cut into two pieces (750 bp and 250 bp). C) A gel showing the fragments of the original and

545   duplicated copy of *flamenco* before and after digestion with BsaAI. Both copies of *flamenco* are

546   amplified by the primers, in column two of the gel (Supplemental File 2). In column three of the

547   gel, the original copy of *flamenco* is uncut (band 1), while the duplicate of *flamenco* forms two

548   bands at 750 bp (band 2) and 250 bp (band 3).

549
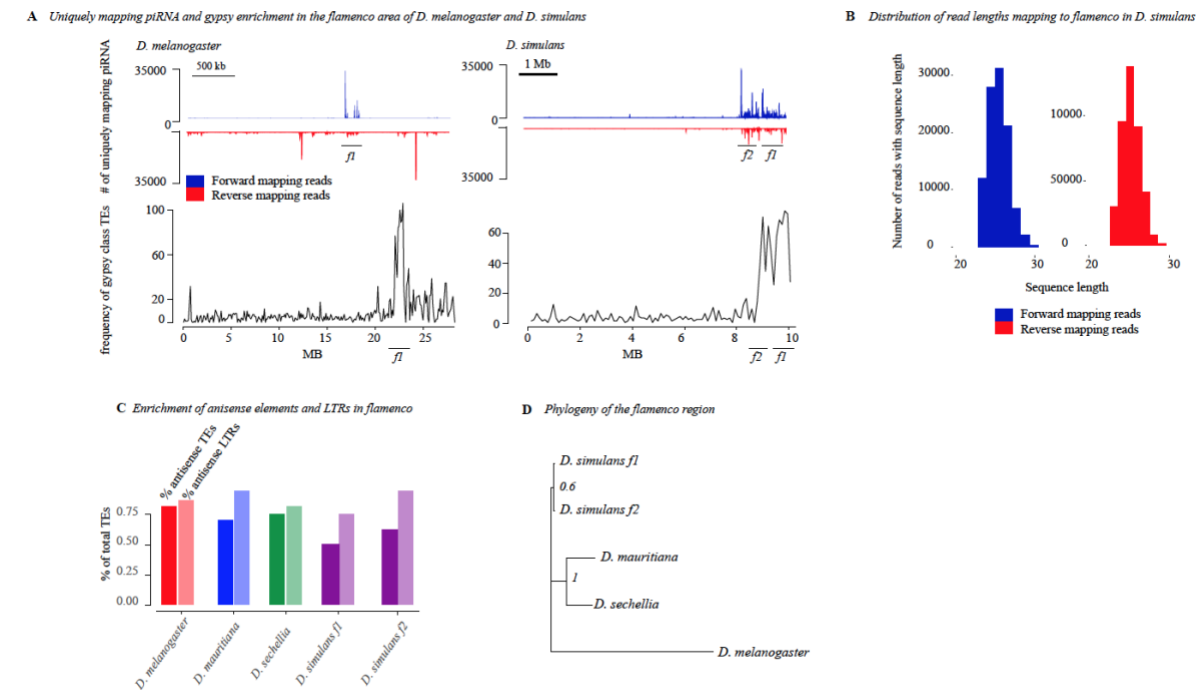
550

551   **Figure 2**. A) Unique piRNA from the ovary and *gypsy* enrichment around *flamenco* and its

552   duplicate in *D. simulans* and *D. melanogaster*. piRNA mapping to the entire contig that contains

553   *flamenco* is shown for both species. The top of the panel shows piRNA mapping to f*lamenco* and

554   is split by antisense (blue) and sense (red) piRNA . The bottom panel shows the frequency of

555   *gypsy*-type transposon annotations across the contig containing *flamenco*, counted in 100 kb

556   windows. There is a clear enrichment of *gypsy* in the area of *flamenco* and, in *D. simulans*, its

557   duplicate compared to the rest of the contig. B) The distribution of read size for small RNA

558   mapping to *flamenco*. The peak is at approximately 26 bp, within the expected range for piRNA.

559   C) The percent of TEs in *flamenco* in each species which are in the antisense orientation (first

560   bar) and the percent of TEs in the antisense orientation that are also LTR class elements (second

561   bar). D) A phylogenetic tree of the *dip1* and *flamenco* enhancer region for *D. melanogaster* and

562   the *simulans* clade. This region is conserved and alignable between all species. The tree was

563   generated with Mr. Bayes (62).

564
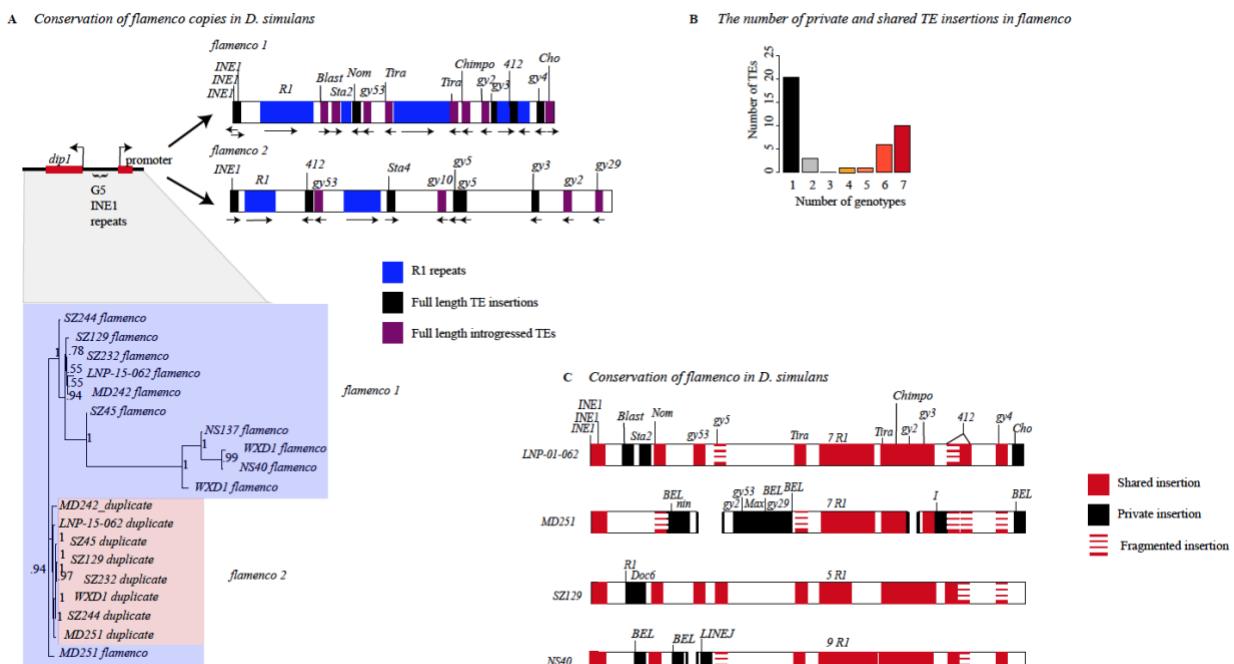
565
566
567
568
569

570
571
572

573

574   Figure 3. A) Divergence between copies of flamenco. Proximal is a phylogenetic tree of *dip1* and

575   the *flamenco* promoter region from each genome. In between *dip1* and the promoter are a series

576   of *G5/INE1* repeats that are found in every genome. Overall this region is fairly conserved, with

577   the duplicate copies all grouping together with short branch lengths (shown in pink). The original

578   copy of *flamenco* is more diverse with some outliers (shown in light blue) but there is good

579   branch support for all the deep branches of the tree. Distal is a representation of *flamenco* and its

580   duplicate. R1 repeat regions are shown in blue. Full length transposable elements are labeled.

581   There is no synteny conservation between *flamenco* and its duplicate. B) The proportion of

582   insertions that are shared by one through seven genotypes (genotypes with complete *flamenco*

583   assemblies). C) Divergence of flamenco within *D. simulans*. Labeled TEs correspond to

584   elements which are present in a full length copy in at least one genome. If they are shared

585   between genomes they are labeled in red, if they are unique they are black. If they are full length

586   in one genome and degraded in other genomes they are represented by stacked dashes. If they are

587   present in the majority of genomes but missing in one, it is represented as a missing that TE,

588   which is agnostic to whether it is a deletion or the element was never present

589

590

591

592

593

Figure 4) A. Copy number of a subset of transposable elements at *flamenco*. Solo LTRs are indicated by in a lighter shade at the top of the bar. The black line on each bar graph indicates a copy number of one. Values for *D. simulans* are the average for all genotypes with a complete *flamenco* assembly. Note that in *D. melanogaster* (green) most TEs have a low copy number. The expansion of *R1* elements in the *simulans* clade is clearly indicated on the right hand panel with a dotted box. Many elements within *flamenco* are multicopy in the *simulans* clade. While

600   some of this is likely due to local duplications it is clearly a different pattern than *D.*

601   *melanogaster*. Enrichment of LTR elements on the antisense strand is clear for all species. **B.**

602   Alignment of *flamenco* in *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana*. There

603   is no conserved synteny between species but there are clearly shared TEs, particularly within the

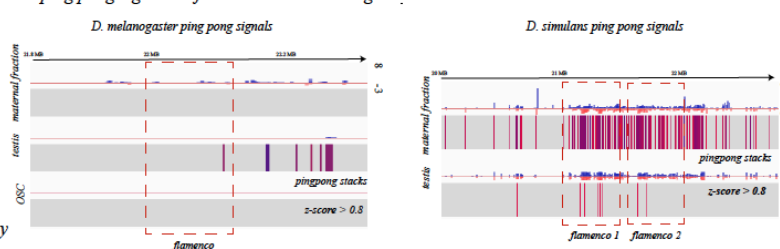604   *simulans* clade. The expansion of *D. mauritiana* compared to the other species is apparent.
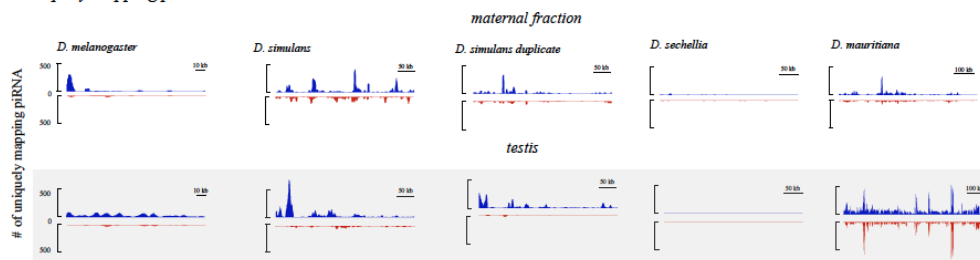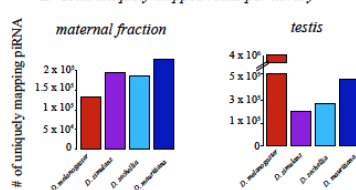
605
606   -
607
608   **Figure 5) A**. Expression of single mapping piRNAs in the maternal fraction and testis (gray) of

609   *D. melanogaster* and the *simulans* clade. Antisense mapping reads are shown in blue, sense in

610   red. Libraries are RPM normalized and scaled across library type. *D. sechellia* has no expression

611   of *flamenco* in the maternal fraction or the testis. *D. melanogaster* has low expression in the

612   maternal fraction and very little ping pong activity. *D. simulans* and *D. mauritiana* show dual

613   stranded expression in the testis and maternal fraction. **B**. The total number of uniquely mapping

614   reads for each of the libraries illustrated in A. This is included to demonstrate that a low number

615   of mapping reads does not explain the patterns seen in *D. sechellia* versus *D. mauritiana*. **C.** The

616   height of 10 nt pingpong stacks at *flamenco* in *D. melanogaster* maternal fraction, testis and

617   ovarian somatic cells is shown on the left. Below each schematic of the height of the stacks is the

618   position of z-scores over 0.8, indicating the likelihood that this is a real ping pong signal as

619   opposed to an artifact. Signals move from red to blue as they approach 1. In the testis a few ping

620   pong signals reach this threshold but not enough to indicate convincingly that there is ping pong

621   activity. On the right are the ping pong stacks and z-scores for the maternal fraction and testis in

622   *D. simulans*. Only in the maternal fraction are the density of z-scores over 0.8 convincing enough

623   to indicate an active ping pong cycle in the *flamenco* region. However, the presence of stacks is

624   enriched in testis, thus this may warrant further investigation. *D. mauritiana* also has convincing

625   ping pong signals in this region (Supplementary Figure 1). **D.** A schematic of the evolution of

626   *flamenco* and its mode expression in the *simulans* and *melanogaster* clade.

627

628
629
630

631   1. C. Duc, *et al.*, Trapping a somatic endogenous retrovirus into a germline piRNA cluster
632   immunizes the germline against further invasion. *Genome Biol* 20, 127 (2019).

633   2. B. Barckmann, *et al.*, The somatic piRNA pathway controls germline transposition over
634   generations. *Nucleic Acids Res* 46, gky761- (2018).

635   3. C. D. Malone, *et al.*, Specialized piRNA Pathways Act in Germline and Somatic Tissues of
636   the Drosophila Ovary. *Cell* 137, 522–535 (2009).

637   4. L. S. Gunawardane, *et al.*, A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5'
638   End Formation in Drosophila. *Science* 315, 1587–1590 (2007).

639   5. S. H. Wang, S. C. R. Elgin, Drosophila Piwi functions downstream of piRNA production
640   mediating a chromatin-based transposon silencing mechanism in female germ line. *Proc*
641   *National Acad Sci* 108, 21164–21169 (2011).

642   6. J. Brennecke, *et al.*, Discrete Small RNA-Generating Loci as Master Regulators of Transposon
643   Activity in Drosophila. *Cell* 128, 1089–1103 (2007).

644   7. A. A. Aravin, *et al.*, The Small RNA Profile during Drosophila melanogaster Development.
645   *Developmental Cell* 5, 337–350 (2003).

646   8. G. Chirn, *et al.*, Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in
647   Eutherian Mammals. *Plos Genet* 11, e1005652 (2015).

648   9. D. Gebert, *et al.*, Large Drosophila germline piRNA clusters are evolutionarily labile and
649   dispensable for transposon regulation. *Mol Cell* 81, 3965-3978.e5 (2021).

650   10. P. R. Andersen, L. Tirian, M. Vunjak, J. Brennecke, A heterochromatin-dependent
651   transcription machinery drives piRNA expression. *Nature* 549, 54–59 (2017).

652   11. C. Klattenhoff, *et al.*, The Drosophila HP1 Homolog Rhino Is Required for Transposon
653   Silencing and piRNA Production by Dual-Strand Clusters. *Cell* 138, 1137–1149 (2009).

654   12. F. Mohn, G. Sienski, D. Handler, J. Brennecke, The Rhino-Deadlock-Cutoff Complex
655   Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell* 157,
656   1364–1379 (2014).

657   13. Y.-C. A. Chen, *et al.*, Cutoff Suppresses RNA Polymerase II Termination to Ensure
658   Expression of piRNA Precursors. *Mol Cell* 63, 97–109 (2016).

659   14. F. Mohn, G. Sienski, D. Handler, J. Brennecke, The Rhino-Deadlock-Cutoff Complex
660   Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell* 157,
661   1364–1379 (2014).

662   15. C. Goriaux, S. Desset, Y. Renaud, C. Vaury, E. Brasset, Transcriptional properties and
663   splicing of the flamencopi RNAcluster. *EMBO reports* 15, 411–418 (2014).

664   16. G. Sienski, D. Dönertas, J. Brennecke, Transcriptional Silencing of Transposons by Piwi and
665   Maelstrom and Its Impact on Chromatin State and Gene Expression. *Cell* 151, 964–980 (2012).

666   17. C. Dennis, E. Brasset, C. Vaury, flam piRNA precursors channel from the nucleus to the
667   cytoplasm in a temporally regulated manner along Drosophila oogenesis. *Mobile DNA* 10, 203–9
668   (2019).

669   18. V. Zanni, A. Eymery, M. C. P. of the, 2013, Distribution, evolution, and diversity of
670   retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters.
671   *National Acad Sciences* https:/doi.org/10.1073/pnas.1313677110/-/dcsupplemental.

672   19. C. M. Bergman, H. Quesneville, D. Anxolabéhère, M. Ashburner, Recurrent insertion and
673   duplication generate networks of transposable element sequences in the Drosophila melanogaster
674   genome. *Genome Biology* 7, R112-21 (2006).

675   20. F. Wierzbicki, R. Kofler, S. Signor, Evolutionary dynamics of piRNA clusters in Drosophila.
676   *Mol Ecol* (2021) https:/doi.org/10.1111/mec.16311.

677   21. N. Prud'homme, M. Gans, M. Masson, C. Terzian, A. Bucheton, Flamenco, a gene
678   controlling the gypsy retrovirus of Drosophila melanogaster. *Genetics* 139, 697–711 (1995).

679   22. S. U. Song, T. Gerasimova, M. Kurkulos, J. D. Boeke, V. G. Corces, An env-like protein
680   encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. *Genes &*
681   *development* 8, 2046–2057 (1994).

682   23. M. Mével-Ninio, A. Pelisson, J. Kinder, A. R. Campos, A. Bucheton, The flamenco Locus
683   Controls the gypsy and ZAM Retroviruses and Is Required for Drosophila Oogenesis. *Genetics*
684   175, 1615–1624 (2007).

685   24. A. Pelisson, *et al.*, Gypsy transposition correlates with the production of a retroviral
686   envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. *The*
687   *EMBO Journal* 13, 4401–4411 (1995).

688   25. A. Bucheton, The relationship between the flamenco gene and gypsy in Drosophila: how to
689   tame a retrovirus. *Trends Genet* 11, 349–353 (1995).

690   26. C. D. Malone, G. J. Hannon, Molecular Evolution of piRNA and Transposon Control
691   Pathways in Drosophila. *Cold Spring Harbor Symposia on Quantitative Biology* 74, 225–234
692   (2010).

693   27. A. G. Clark, *et al.*, Evolution of genes and genomes on the Drosophila phylogeny. *Nature*
694   450, 203–218 (2007).

695   28. D. G. Eickbush, W. C. Lathe, M. P. Francino, T. H. Eickbush, R1 and R2 retrotransposable
696   elements of Drosophila evolve at rates similar to those of nuclear genes. *Genetics* 139, 685–695
697   (1995).

698   29. S. A. Signor, F. N. New, S. Nuzhdin, A Large Panel of Drosophila simulans Reveals an
699   Abundance of Common Variants. *Genome Biology and Evolution* 10, 189–206 (2017).

700   30. S. Signor, S. Nuzhdin, Dynamic changes in gene expression and alternative splicing mediate
701   the response to acute alcohol exposure in Drosophila melanogaster. *Heredity* (2018).

702   31. S. Signor, Population genomics of Wolbachia and mtDNA in Drosophila simulans from
703   California. *Scientific Reports*, 1–11 (2017).

704   32. S. A. Signor, M. Abbasi, P. Marjoram, S. V. Nuzhdin, Social effects for locomotion vary
705   between environments in Drosophila melanogaster females. *Evolution* 71, 1765–1775 (2017).

706   33. S. Signor, Transposable elements in individual genotypes of Drosophila simulans. *Ecology*
707   *and Evolution* 130, 499–11 (2020).

708   34. D. R. Matute, J. Gavin-Smyth, G. Liu, Variable post-zygotic isolation in Drosophila
709   melanogaster/D. simulanshybrids. *Journal of Evolutionary Biology* 27, 1691–1705 (2014).

710   35. D. R. Schrider, J. Ayroles, D. R. Matute, A. D. Kern, Supervised machine learning reveals
711   introgressed loci in the genomes of Drosophila simulans and D. sechellia. *PLoS Genetics* 14,
712   e1007341-29 (2018).

713   36. R. L. Rogers, *et al.*, Landscape of Standing Variation for Tandem Duplications in Drosophila
714   yakuba and Drosophila simulans. *Molecular Biology and Evolution* 31, 1750–1766 (2014).

715   37. M. Chakraborty, *et al.*, Evolution of genome structure in the Drosophila simulansspecies
716   complex. 139, 1067–63 (2020).

717   38. , Genome Res.-2017-Koren-gr.215087.116.

718   39. R. Vaser, I. Sović, N. Nagarajan, M. Šikić, Fast and accurate de novo genome assembly from
719   long uncorrected reads. *Genome Res* 27, 737–746 (2017).

720   40. B. J. Walker, *et al.*, Pilon: An Integrated Tool for Comprehensive Microbial Variant
721   Detection and Genome Assembly Improvement. *Plos One* 9, e112963 (2014).

722   41. M. Chakraborty, J. J. Emerson, S. J. Macdonald, A. D. Long, Structural variants exhibit
723   widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications*,
724   1–11 (2019).

725   42. M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using
726   repeat graphs. *Nat Biotechnol* 37, 540–546 (2019).

727   43. D. R. Laetsch, M. L. Blaxter, BlobTools: Interrogation of genome assemblies.
728   *F1000research* 6, 1287 (2017).

729   44. M. Tarailo-Graovac, N. Chen, Using RepeatMasker to Identify Repetitive Elements in
730   Genomic Sequences. *Current Protocols in Bioinformatics*, 1–14 (2009).

731   45. J. M. Flynn, *et al.*, RepeatModeler2 for automated genomic discovery of transposable
732   element families. *Proc National Acad Sci* 117, 9451–9457 (2020).

733   46. J. Armstrong, *et al.*, Progressive Cactus is a multiple-genome aligner for the thousand-
734   genome era. *Nature* 587, 246–251 (2020).

735   47. M. Kolmogorov, *et al.*, Chromosome assembly of large and complex genomes using multiple
736   references. *Genome Res* 28, 1720–1732 (2018).

737   48. F. Wierzbicki, F. Schwarz, O. Cannalonga, R. Kofler, Generating high quality assemblies for
738   genomic analysis of transposable elements. *Biorxiv*, 2020.03.27.011312 (2020).

739   49. F. Wierzbicki, F. Schwarz, O. Cannalonga, R. Kofler, Novel quality metrics allow
740   identifying and generating high-quality assemblies of piRNA clusters. *Mol Ecol Resour* 22, 102–
741   121 (2022).

742   50. Vedanayagam, Jeffrey, "Evolutionary Genomics of piRNA Mediated Transposon Silencing
743   in Drosophila," University of Rochester. (2016).

744   51. J. Vedanayagam, *et al.*, Endogenous RNAi silences a burgeoning sex chromosome arms race.
745   *Biorxiv*, 2022.08.22.504821 (2022).

746   52. J. Vedanayagam, C.-J. Lin, E. C. Lai, Rapid evolutionary dynamics of an expanding family
747   of meiotic drive factors and their hpRNA suppressors. *Nat Ecol Evol* 5, 1613–1623 (2021).

748   53. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: an ultra-fast all-in-one FASTQ preprocessor.
749   *Biorxiv*, 274100 (2018).

750   54. M. J. Axtell, ShortStack: Comprehensive annotation and quantification of small RNA genes.
751   *RNA* 19, 740–751 (2013).

752   55. B. Langmead, C. Trapnell, M. Pop, S. L. Salzberg, Ultrafast and memory-efficient alignment
753   of short DNA sequences to the human genome. *Genome Biol* 10, R25 (2009).

754   56. H. Li, *et al.*, The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–
755   2079 (2009).

756   57. Y. Liao, G. K. Smyth, W. Shi, The R package Rsubread is easier, faster, cheaper and better
757   for alignment and quantification of RNA sequencing reads. *Nucleic Acids Res* 47, gkz114-
758   (2019).

759   58. D. Rosenkranz, H. Zischler, proTRAC - a software for probabilistic piRNA cluster detection,
760   visualization and analysis. *Bmc Bioinformatics* 13, 5 (2012).

761   59. M. Bailly-Bechet, A. Haudry, E. Lerat, "One code to find them all": a perl tool to
762   conveniently parse RepeatMasker output files. *Mobile Dna-uk* 5, 13 (2014).

763   60. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic
764   features. *Bioinformatics* 26, 841–842 (2010).

765   61. F. Sievers, D. G. Higgins, Clustal Omega for making accurate alignments of many protein
766   sequences. *Protein Sci* 27, 135–145 (2018).

767   62. F. Ronquist, *et al.*, MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model
768   Choice Across a Large Model Space. *Systematic Biology* 61, 539–542 (2012).

769   63. E. Paradis, J. Claude, K. Strimmer, APE: Analyses of Phylogenetics and Evolution in R
770   language. *Bioinformatics* 20, 289–290 (2004).

771   64. S. Uhrig, H. Klein, PingPongPro: a tool for the detection of piRNA-mediated transposon-
772   silencing in small RNA-Seq data. *Bioinformatics* 35, 335–336 (2018).

773   65. Smit, A. and Hubley, R. and Green, P, *RepeatMasker Open-4.0. 2013--2015* (2005).

774   66. H. Quesneville, *et al.*, Combined Evidence Annotation of Transposable Elements in Genome
775   Sequences. *Plos Comput Biol* 1, e22 (2005).

776   67. E. Lerat, *et al.*, Population specific dynamics and selection patterns of transposable element
777   insertions in European natural populations. *Molecular Ecology*, 1–42 (2018).

778   68. R. S. Singh, Population genetics and evolution of species related to Drosophila melanogaster.
779   *Annual Review of Genetics* 23, 425–453 (1989).

780   69. H. E. Machado, *et al.*, Comparative population genomics of latitudinal variation in
781   Drosophila simulans and Drosophila melanogaster. *Molecular Ecology* 25, 723–740 (2016).

782   70. A. Sedghifar, P. Saelao, D. J. Begun, Genomic patterns of geographic differentiation in
783   Drosophila simulans. *Genetics* (2016) https:/doi.org/10.1534/genetics.115.185496.

784   71. E. L. S. Loreto, C. M. A. Carareto, P. Capy, Revisiting horizontal transfer of transposable
785   elements in Drosophila. *Heredity* 100, 545–554 (2008).

786   72. N. Bargues, E. Lerat, Evolutionary history of LTR-retrotransposons among 20 Drosophila
787   species. *Mobile Dna-uk* 8, 7 (2017).

788   73. Z. Durdevic, R. S. Pillai, A. Ephrussi, Transposon silencing in the Drosophila female
789   germline is essential for genome stability in progeny embryos. *Life Sci Alliance* 1, e201800179
790   (2018).

791   74. B. Czech, J. B. Preall, J. McGinn, G. J. Hannon, A Transcriptome-wide RNAi Screen in the
792   Drosophila Ovary Reveals Factors of the Germline piRNA Pathway. *Mol Cell* 50, 749–761
793   (2013).

794   75. A. and T. Pélisson, About the origin of retroviruses and the co-evolution of the gypsy
795   retrovirus with the Drosophila flamenco host gene. 29–37 (1997).

796   76. C. Duc, *et al.*, Trapping a somatic endogenous retrovirus into a germline piRNA cluster
797   immunizes the germline against further invasion. *Genome Biol* 20, 127 (2019).

798   77. Y. Luo, P. He, N. Kanrar, K. F. Toth, A. Aravin, Maternally inherited siRNAs initiate piRNA
799   cluster formation https:/doi.org/10.1101/2022.02.08.479612.

800   78. R. Kofler, piRNA Clusters Need a Minimum Size to Control Transposable Element
801   Invasions. *Genome Biology and Evolution* 12, 736–749 (2020).

802   79. A. M. Waterhouse, J. B. Procter, D. M. A. Martin, M. Clamp, G. J. Barton, Jalview Version
803   2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* 25, 1189–1191
804   (2009).

805