PLOS GENETICS

RESEARCH ARTICLE

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

Sarah Signor[1]*, Jeffrey Vedanayagam[2,3], Bernard Y. Kim[4], Filip Wierzbicki[5,6], Robert Kofler[5], Eric C. Lai[2]

1 Biological Sciences, North Dakota State University, Fargo, North Dakota, United States of America, **2** Developmental Biology Program, Sloan-Kettering Institute, New York, New York, United States of America, **3** Department of Neuroscience, Developmental and Regenerative Biology, University of Texas at San Antonio, Texas, United States of America, **4** Department of Biology, Stanford University, Stanford, California, United States of America, **5** Institut für Populationsgenetik, Vetmeduni Vienna, Vienna, Austria, **6** Vienna Graduate School of Population Genetics, Vienna, Austria

* sarah.signor@ndsu.edu

Citation: Signor S, Vedanayagam J, Kim BY, Wierzbicki F, Kofler R, Lai EC (2023) Rapid evolutionary diversification of the *flamenco* locus across simulans clade *Drosophila* species. PLoS Genet 19(8): e1010914. https://doi.org/10.1371/journal.pgen.1010914

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: March 7, 2023

Accepted: August 9, 2023

Published: August 29, 2023

Data Availability Statement: All data has been made available in the following repositories: The genomes referenced in this study have been deposited at NCBI Genome under the accession number PRJNA907284. The small RNA data is available at NCBI SRA under the accession number PRJNA913883. These repositories will be made public upon acceptance of the manuscript. The RepeatMasker annotations are available at https://github.com/SignorLab/Flamenco_manuscript.

Abstract

Suppression of transposable elements (TEs) is paramount to maintain genomic integrity and organismal fitness. In *D. melanogaster*, the *flamenco* locus is a master suppressor of TEs, preventing the mobilization of certain endogenous retrovirus-like TEs from somatic ovarian support cells to the germline. It is transcribed by Pol II as a long (100s of kb), single-stranded, primary transcript, and metabolized into ~24–32 nt Piwi-interacting RNAs (piRNAs) that target active TEs via antisense complementarity. *flamenco* is thought to operate as a trap, owing to its high content of recent horizontally transferred TEs that are enriched in antisense orientation. Using newly-generated long read genome data, which is critical for accurate assembly of repetitive sequences, we find that *flamenco* has undergone radical transformations in sequence content and even copy number across *simulans* clade Drosophilid species. *Drosophila simulans flamenco* has duplicated and diverged, and neither copy exhibits synteny with *D. melanogaster* beyond the core promoter. Moreover, *flamenco* organization is highly variable across *D. simulans* individuals. Next, we find that *D. simulans* and *D. mauritiana flamenco* display signatures of a dual-stranded cluster, with ping-pong signals in the testis and/or embryo. This is accompanied by increased copy numbers of germline TEs, consistent with these regions operating as functional dual-stranded clusters. Overall, the physical and functional diversity of *flamenco* orthologs is testament to the extremely dynamic consequences of TE arms races on genome organization, not only amongst highly related species, but even amongst individuals.

Author summary

Transposable element suppression is essential for genomic stability and fertility. To date, many insights have been gained by studying the major suppression loci in *D. melanogaster*, *flamenco* and *42AB*. While *42AB* is an exemplar germline locus, *flamenco* is the

Funding: This work was supported by the National Science Foundation Established Program to Stimulate Competitive Research (NSF-EPSCoR-1826834 and NSF-EPSCoR-2032756, SS), the Austrian Science Fund FWF (https://www.fwf.ac.at/) grant P35093 (RK), a Pathway to Independence award from the National Institute of General Medical Sciences (K99-GM137077, JV), NSF DEB (1209536, JV), NIH-NRSA F32GM135998 (BYK), the National Institute of General Medical Sciences (R01-GM083300, ECL), and National Institutes of Health MSK Core Grant (P30-CA008748, ECL). The funders had no role in study design, data collection and analysis, preparation of, or decision to publish the manuscript.

Competing interests: The authors have declared that no competing interests exist.

master regulator of TEs in the somatic cells of the ovary. Here, we take a closer look at *flamenco* in *simulans*-clade species, to see if what we have learned about *flamenco* in *D. melanogaster* holds true. Certain aspects of *flamenco* are conserved, including enrichment for LTR class TEs arising from horizontal transfer, but other features are diverged. *flamenco* has duplicated in *D. simulans* and may also be serving as a germline suppression cluster. This is also true in *D. mauritiana*, while *D. sechellia flamenco* retains *D. melanogaster*-like features. There is also incredible diversity at *flamenco* within *D. simulans* populations, suggesting important fitness effects at this locus. Overall, our data provide unique insights into the evolutionary dynamics of TE suppression and turnover of piRNA cluster properties.

Introduction

Drosophila gonads exemplify two important fronts in the conflict between transposable elements (TEs) and the host–the germline (which directly generates gametes), and somatic support cells (from which TEs can invade the germline) [1,2]. The strategies by which TEs are suppressed in these settings are distinct [3], but share their utilization of Piwi-interacting RNAs (piRNAs). These are ~24–32 nt RNAs that are bound by the Piwi subclass of Argonaute effector proteins, and guide them and associated cofactors to targets for transcriptional and/or post-transcriptional silencing [4–7].

Mature piRNAs are processed from non-coding piRNA cluster transcripts, which derive from genomic regions that are densely populated with TE sequences [7–9]. However, the mechanisms of piRNA biogenesis differ between gonadal cell types. In the germline, piRNA clusters are transcribed from both DNA strands through non-canonical Pol II activity [6,10–12], which is initiated by chromatin marks rather than specific core promoter motifs. Moreover, co-transcriptional processes such as splicing and polyadenylation are suppressed within dual strand piRNA clusters [12,13]. On the other hand, in ovarian somatic support cells, piRNA clusters are transcribed from a typical promoter as a single stranded transcript, which can be alternatively spliced as with protein-coding mRNAs [14–17]. These rules derive in large part from the study of model piRNA clusters (i.e. the germline *42AB* and somatic *flamenco* piRNA clusters). For both types, their capacity to repress invading TEs is thought to result from random integration of new transposons into the cluster [18]. As such, piRNA clusters are adaptive loci that play central roles in the conflict between hosts and TEs.

The location and activity of germline piRNA clusters are stochastic and evolutionarily dynamic, as there are many copies of TE families in different locations that may produce piRNAs [9,19]. By contrast, somatic piRNA clusters are not redundant and a single insertion of a TE into a somatic piRNA cluster should be sufficient to largely repress that TE from further transposition [1,17]. Thus, *flamenco* should contain only one copy per TE, which is largely true in the *flamenco* locus of *D. melanogaster* [17]. Notably, *flamenco* is also the only piRNA cluster known to produce a phenotypic effect when mutated, since deletions of multiple germline clusters did not activate corresponding TE classes [9].

flamenco has been a favored model for understanding the piRNA pathway since the discovery of piRNA mediated silencing of transposable elements [6]. *flamenco* spans >350 kb of repetitive sequences located in *β*-heterochromatin of the X chromosome [20]. Of note, *flamenco* was initially identified, prior to the formal recognition of piRNAs, via transposon insertions that de-repress *mdg4* (also known as *gypsy)*, *ZAM*, and *Idefix* elements [20–24]. These mutant alleles disrupt the *flamenco* promoter, and consequently abrogate transcription and

piRNA production across the length of this locus. By contrast, the deletion of multiple model germline piRNA clusters, which eliminate the biogenesis of a bulk of cognate piRNAs, surprisingly did not de-repress their cognate TEs [9]. Thus, *flamenco* evolution is potentially more consequential for TE dynamics. Analysis of *flamenco* in various strains of *D. melanogaster* supports that this locus traps horizontally derived TEs to achieve silencing of newly invaded TEs [17]. The *flamenco* locus exhibits synteny across the *D. melanogaster* sub-group [25]; however, the sequence composition of *flamenco* outside *D. melanogaster* has not been well-characterized [3,26].

In this study, we compare the *flamenco* locus across long-read assemblies of the three *simulans*-clade sister species, including 10 strains of *D. simulans*, and one strain each of *D. mauritiana* and *D. sechellia*. Analysis of piRNAs from ovaries of five genotypes of *D. simulans* found that *flamenco* is duplicated in *D. simulans*. There is no sequence synteny across copies, even though their core promoter regions and the adjacent *dip1* gene duplications are conserved. *flamenco* has also been colonized by abundant (>40) copies of *R1*, a TE that was thought to insert only at ribosomal genes, and to evolve at the same rate as nuclear genes [27]. Furthermore, between different genotypes, up to 63% of TE insertions are not shared within any given copy of *flamenco*. Despite this, several full length TEs are shared between all genotypes in a similar sequence context. This incredible diversity at the *flamenco* locus, even within a single species, suggests there may be considerable variation in its ability to suppress transposable elements across individuals.

Cross-species comparisons further support that functions of *flamenco* have diversified. Data from *D. sechellia* and *D. melanogaster* conform with the current understanding of *flamenco* as a uni-strand cluster. However, we find evidence that *D. simulans* and *D. mauritiana flamenco* can act as a dual strand cluster in testis (*D. mauritiana*) and embryos (*D. mauritiana* and *D. simulans*), yielding piRNAs from both strands with a ping-pong signal. Overall, we infer that the rapid evolution of *flamenco* alleles across individuals and species reflects highly adaptive functions and dynamic biogenesis capacities.

Results

flamenco loci across *simulans*-clade Drosophilid species

We identified *D. simulans flamenco* from several lines of evidence: piRNA cluster calls from proTRAC, its location adjacent to divergently transcribed *dip1*, the existence of conserved core *flamenco* promoter sequences, and enrichment of *Ty3/mdg4* elements (Figs 1 and 2 and S1 and S2 Tables). The *flamenco* locus is at least 376 kb in *D. simulans*. This is similar to *D. melanogaster*, where *flamenco* is typically up to 350 kb, though this appears to vary by genotype [28]. In *D. sechellia flamenco* is at least 363 kb, however in *D. mauritiana* the locus has expanded to at least 840 kb (S2 Table). This is a large expansion, and it is possible that the entire region does not act as a region controlling somatic TEs. However, evidence that is does include uniquely mapping piRNAs that are found throughout the region and *Ty3/mdg4* enrichment consistent with a *flamenco*-like locus (S1 Fig). There are no protein coding genes within the 840 kb putative *flamenco* region. The genes that are downstream of *flamenco* in *D. melanogaster* have moved in *D. mauritiana* (*CG40813- CG41562* at 21.5 MB in *D. melanogaster*), and *flamenco* is now flanked by the group of genes beginning with *CG14621* (22.4 MB in *D. melanogaster)*. Thus in *D. melanogaster* the borders of *flamenco* are flanked by *dip1* upstream and *CG40813* downstream, while in *D. mauritiana* they are *dip1* upstream and *CG14621* downstream (but note that *flamenco* does not extend all the way to these genes). Between all species the *flamenco* promoter and surrounding region, including a *dip1* gene, are alignable and conserved (Fig 2D).

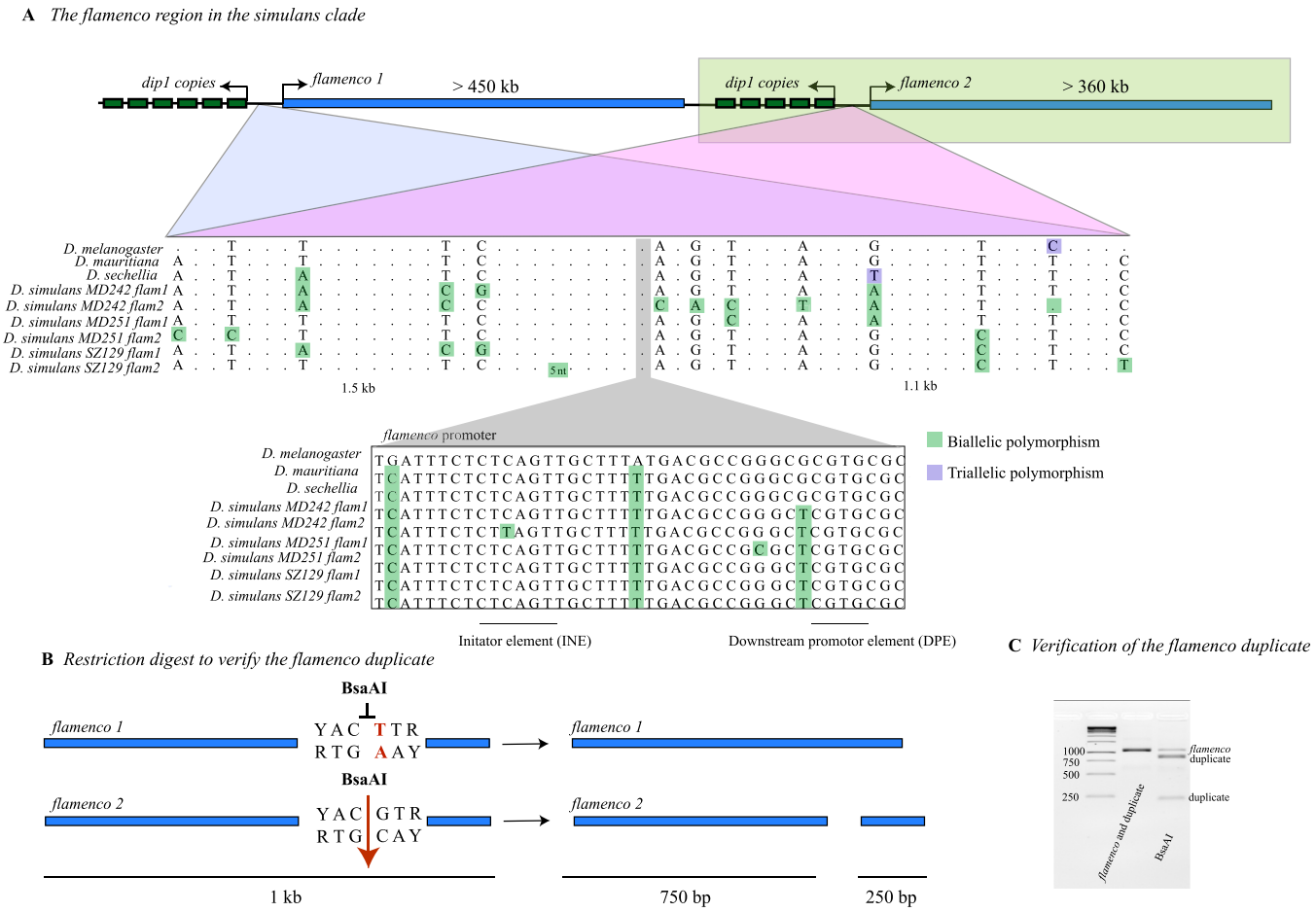

Fig 1. A) The duplication of *flamenco* in the *D. simulans*. Both copies are flanked by copies of the *dip1* gene and copies of the putative *flamenco* promoter. The top portion of the alignment shows ~ 2 kb around the promoter. SNPs are shown if they differentiate copies of *flamenco* within a single 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 duplicated copy of *flamenco* before and after digestion with BsaAI. Both copies of *flamenco* are amplified by the primers, in column two of the gel (Supplemental File 2). In column three of the gel, the original copy of *flamenco* is uncut (band 1), while the duplicate of *flamenco* forms two bands at 750 bp (band 2) and 250 bp (band 3).

https://doi.org/10.1371/journal.pgen.1010914.g001

Structure of the *flamenco* locus

D. melanogaster flamenco bears a characteristic structure, in which the majority of TEs are *Ty3/mdg4* elements in the antisense orientation (79% antisense orientation, 85% of which are *Ty3/mdg4* elements) (Fig 2C and S3 Table). In *D. simulans*, *flamenco* has been colonized by large expansions of *R1* transposable element repeats such that on average the percent of antisense TEs is only 50% and the percent of the locus comprised of LTR elements is 55%. However, 76% of antisense insertions are LTR insertions, thus the underlying *flamenco* structure is apparent when the *R1* insertions are disregarded (Fig 2C). In *D. mauritiana flamenco* is 71% antisense, and of those antisense elements it is 85% LTRs. Likewise in *D. sechellia* 78% of elements are antisense, and of those 81% are LTRs. *flamenco* retains the overall structure of a canonical *D. melanogaster*-like *flamenco* locus in all of these species. That is, *Ty3/mdg4* enrichment, the *flamenco* promoter region, and an enrichment of antisense LTR elements (Fig 2A–2D).

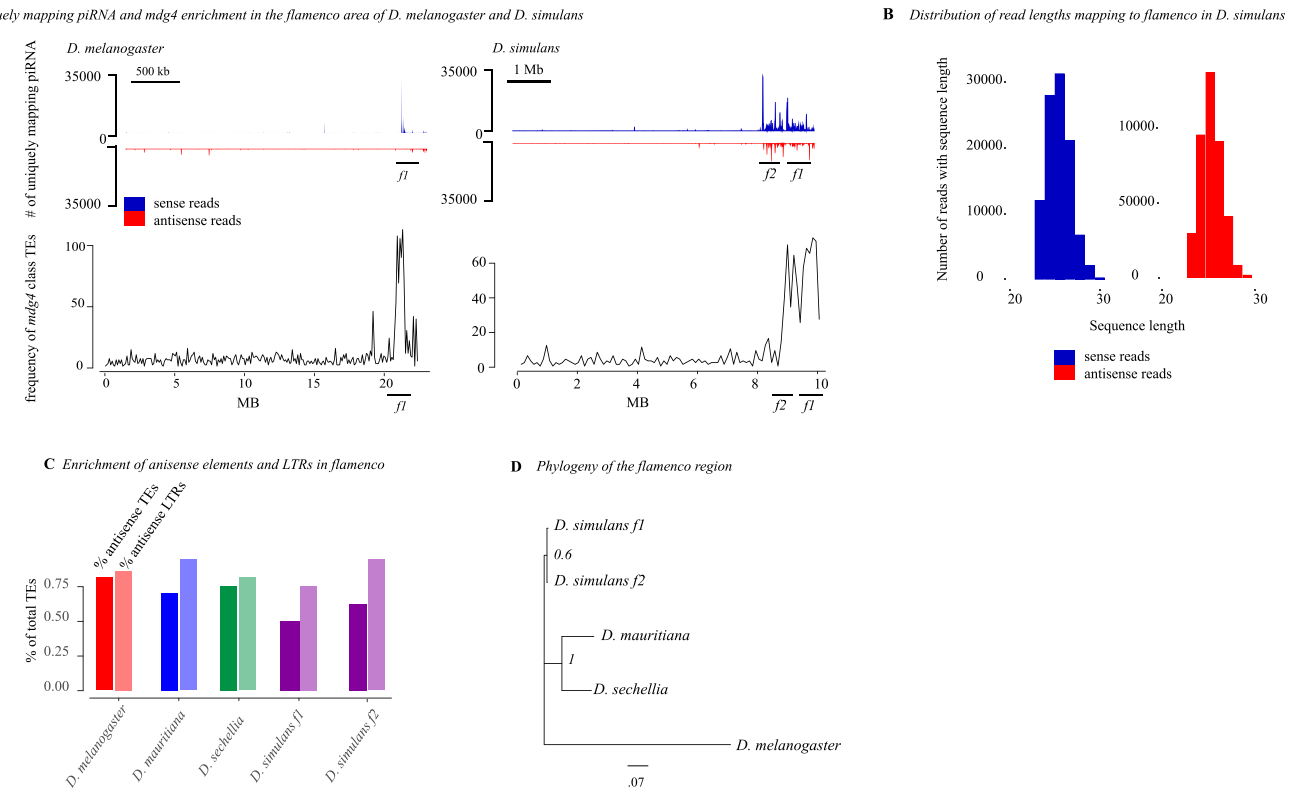

Fig 2. A) Unique piRNA from the ovary and *Ty3/mdg4* enrichment around *flamenco* and its duplicate in *D. simulans* and *D. melanogaster*. piRNA mapping to the entire contig that contains *flamenco* is shown for both species. The top of the panel shows piRNA mapping to *flamenco* and is split by antisense (blue) and sense (red) piRNA. The bottom panel shows the frequency of *Ty3/mdg4* transposon annotations across the contig containing *flamenco*, counted in 100 kb windows. There is a clear enrichment of *mdg4* in the area of *flamenco* and, in *D. simulans*, its duplicate compared to the rest of the contig. B) The distribution of read size for small RNA mapping to *flamenco*. The peak is at approximately 26 bp, within the expected range for piRNA. C) The percent of TEs in *flamenco* in each species which are in the antisense orientation (first bar) and the percent of TEs in the antisense orientation that are also LTR class elements (second bar). D) A phylogenetic tree of the *dip1* and *flamenco* enhancer region for *D. melanogaster* and the *simulans* clade. This region is conserved and alignable between all species. The tree was generated with Mr. Bayes 3.2.7a [74]. Branch lengths are indicated by the scale bar at the bottom, in units of expected changes per site.

https://doi.org/10.1371/journal.pgen.1010914.g002

flamenco is duplicated in *D. simulans*

In *D. simulans*, we unexpectedly observed that *flamenco* is duplicated on the X chromosome; the duplication was confirmed with PCR and a restriction digest (Figs 1 and S2 and S2 File). While this might in principle represent a second allele of *flamenco* that is very diverged and found in one copy of each genome, the high quality of assemblies of this region makes this unlikely (S1 File). Furthermore, it is found in every assembled *D. simulans* genome and thus is unlikely to be a high frequency balanced polymorphism. These duplications are associated with a conserved copy of the putative *flamenco* enhancer as well as copies of the *dip1* gene located proximal to *flamenco* in *D. melanogaster* (Figs 1 and 3A). While it is unclear which copy is orthologous to *D. melanogaster flamenco*, all *D. simulans* lines bear one copy that aligns across genotypes. We refer to this copy as *D. simulans flamenco*, and the other copies as duplicates. Otherwise, outside of the promoter and *dip1* region, the two copies of *flamenco* do not align with one another and lack synteny amongst their resident TEs. Possible evolutionary scenarios are that the *flamenco* duplication occurred early in the *simulans* lineage, that the clusters evolved very rapidly, or that the duplication encompassed only the promoter region and was subsequently colonized by TEs (Figs 1A and 3A). The duplicate retains the structure of

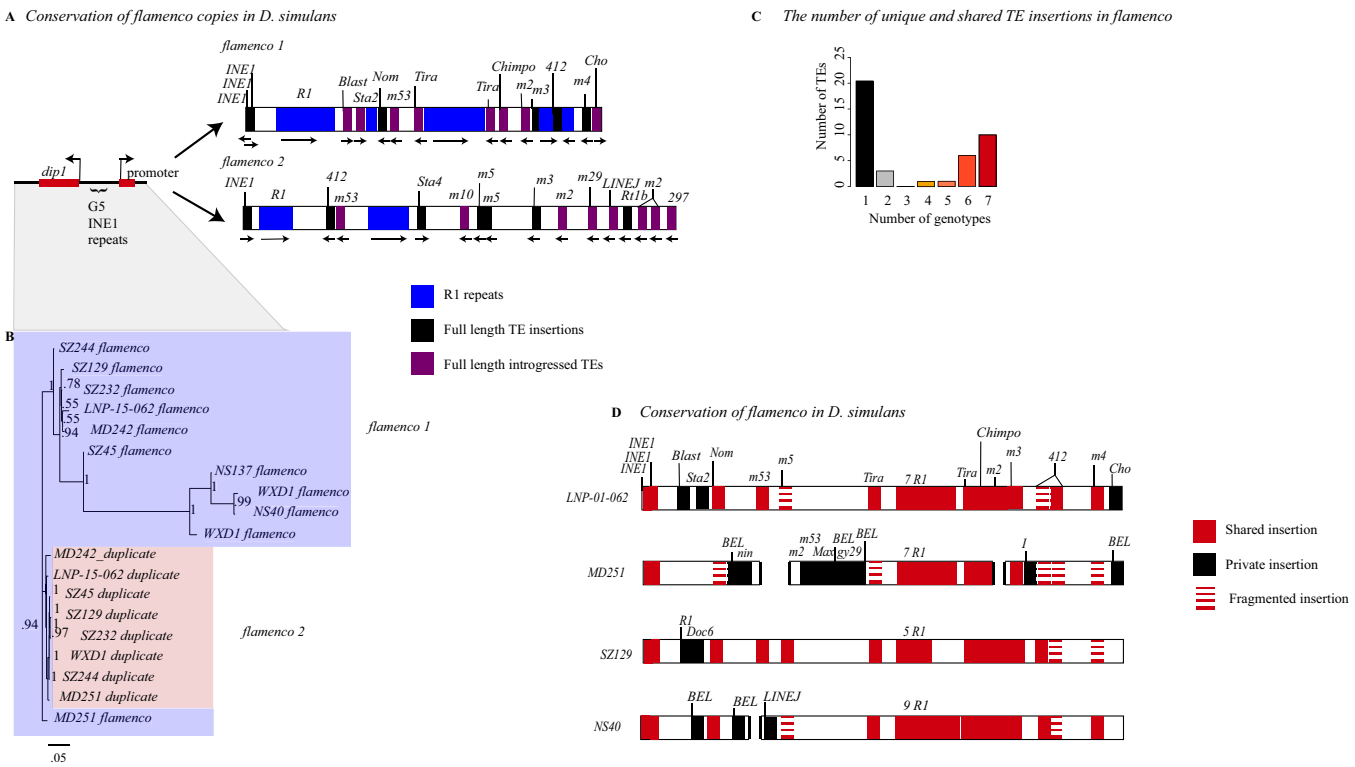

Fig 3. A) A representation of *flamenco* and its duplicate from genotype *LNP-01-062*. R1 repeat regions are shown in blue. Full length transposable elements are labeled. There is no synteny conservation between *flamenco* and its duplicate. Figure is not to scale. B) Divergence between copies of flamenco. This is a phylogenetic tree of *dip1* and the *flamenco* promoter region from each genome. In between *dip1* and the promoter are a series of *G5/INE1* repeats that are found in every genome. Overall this region is fairly conserved, with the duplicate copies all grouping together with short branch lengths (shown in pink). The original copy of *flamenco* is more diverse with some outliers (shown in light blue) but there is good branch support for all the deep branches of the tree. C) The proportion of insertions that are shared by one through seven genotypes (genotypes with complete *flamenco* assemblies). D) Divergence of flamenco within *D. simulans*. Labeled TEs correspond to elements which are present in a full length copy in at least one genome. If they are shared between genomes they are labeled in red, if they are unique they are black. If they are full length in one genome and degraded in other genomes they are represented by stacked dashes. If they are present in the majority of genomes but missing in one, it is represented as a missing that TE, which is agnostic to whether it is a deletion or the element was never present.

https://doi.org/10.1371/journal.pgen.1010914.g003

flamenco, with an average of 67% of TEs in the antisense orientation, and 91% of the TEs in the antisense orientation are LTRs. The duplicate of *flamenco* is less impacted by *R1*, with some genotypes having as few as 8 *R1* insertions (Fig 3C).

The *flamenco* duplicate is absent in the *D. simulans* reference assembly, w^{501} (GCA_000754195.3), but present in wxD^{1}, suggesting it was polymorphic, the duplication had not yet occurred, or the most likely scenario that it was not assembled. A second *flamenco* promoter is present on a 750 bp scaffold in w^{501}, but that is not enough to know if it is a *flamenco* duplicate or an assembly artifact.

flamenco piRNA is expressed in the testis and the maternal fraction

Canonically, *flamenco* piRNA is expressed in the somatic follicular cells of the ovary and not in the germline, and also does not produce a ping-pong signal [23]. It was not thought to be present in the maternal fraction of piRNAs or other tissues. However, that appears to be variable in different species (Fig 4). We examined single mapping reads in the *flamenco* region from testes and embryos (maternal fraction) in *D. simulans*, *D. mauritiana*, *D. sechellia*, and *D.*

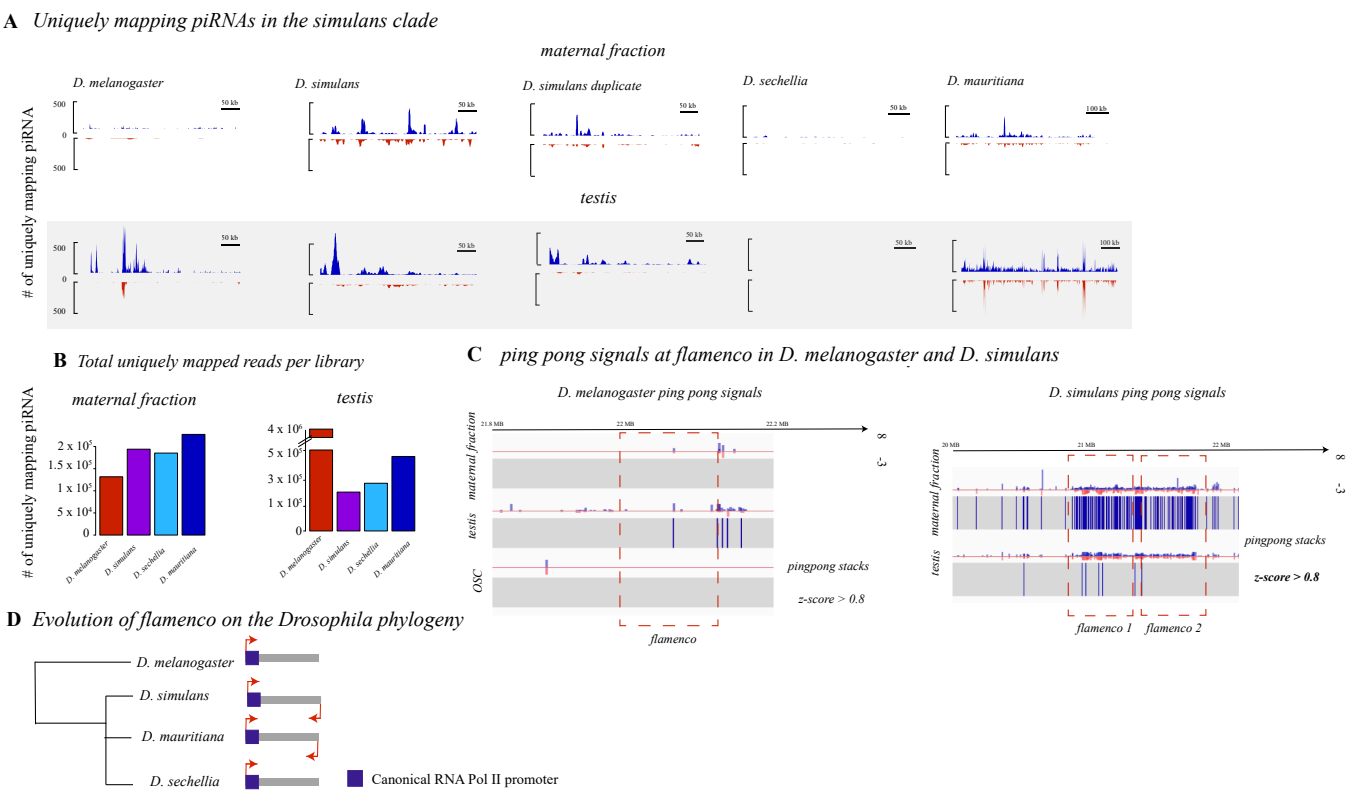

Fig 4. **A**. Expression of single mapping piRNAs in the maternal fraction and testis (gray) of *D. melanogaster* and the *simulans* clade. Sense mapping reads are shown in blue, antisense in red. Libraries are RPM normalized and the axis are the same for each library type i.e. embryo. *D. sechellia* has no expression of *flamenco* in the maternal fraction or the testis. *D. melanogaster* has low expression in the maternal fraction and very little ping-pong activity. *D. simulans* and *D. mauritiana* show dual stranded expression in the testis and maternal fraction. **B**. The total number of uniquely mapping reads for each of the libraries illustrated in A. This is included to demonstrate that a low number of mapping reads does not explain the patterns seen in *D. sechellia* versus *D. mauritiana*. **C.** The height of 10 nt ping-pong stacks at *flamenco* in *D. melanogaster* maternal fraction, testis and ovarian somatic cells is shown on the left. Below each schematic of the height of the stacks is the position of z-scores over 0.8, indicating the likelihood that this is a real ping-pong signal as opposed to an artifact. Scores were produced by pingpongpro [76]. Signals move from red to blue as they approach 1. In the testis, a few ping-pong signals reach this threshold but not enough to indicate ping-pong activity convincingly. On the right are the ping-pong stacks and z-scores for the maternal fraction and testis in *D. simulans*. Only in the maternal fraction are the density of z-scores over 0.8 convincing enough to indicate an active ping-pong cycle in the *flamenco* region. However, the presence of stacks is enriched in testis, thus this may warrant further investigation. *D. mauritiana* also has convincing ping-pong signals in this region (S1 Fig). **D.** A schematic of the evolution of *flamenco* and its mode expression in the *simulans* and *melanogaster* clade.

https://doi.org/10.1371/journal.pgen.1010914.g004

melanogaster. As a control we also included *D. melanogaster* ovarian somatic cells, where Aub and Ago3 are not expressed and therefore there should be no ping-pong signals. In *D. simulans* and *D. mauritiana flamenco* is expressed bidirectionally in the maternal fraction and the testis, including ping-pong signals on both strands (Figs 4A, 4C and S1). In *D. sechellia*, there is no expression of *flamenco* in either of these tissues. Discarding multimappers in the maternal fraction 63% (*D. mauritiana*)– 36% (*D. simulans*) of the ping-pong signatures on the X with a z-score of at least 0.9 are located within *flamenco* (Fig 4C). In the testis the picture is more complicated–in *D. mauritiana* 50% of ping-pong signals on the X with a z-score of at least 0.9 are located within *flamenco* (S1 Fig). While mapping of piRNA to both strands was observed in *D. simulans* testis, there is very little apparent ping-pong activity (5 positions in *flamenco* $z > 0.9$; 15 potential ping pong signals on the X). In *D. melanogaster*, there is uni-strand expression in the maternal fraction, but it is limited to the region close to the promoter. In *D. melanogaster* no ping-pong signals have a z-score above 0.8 in the maternal fraction or the ovarian somatic cells. There are ping-pong stacks in *flamenco* in the testis of *D. melanogaster*

(2% of the total on the contig); however, they are limited to a single region and are not abundant enough to be strong evidence of ping-pong activity.

In the duplicate of *flamenco* in the maternal fraction 15% of the ping-pong signals with a z-score above 0.9 on the X are within the *flamenco* duplicate. The *flamenco* duplicate does not have a strong signal of the ping-pong pathway in the testis. In addition, *flamenco* in these species has been colonized by full length TEs thought to be active in the germline such as *blood*, *burdock*, *mdg-3*, *Transpac*, and *Bel* [29,30]. The differences in ping-pong signals between species and the presence of germline TEs in *D. simulans* and *D. mauritiana* suggests that the role of *flamenco* in these tissues has evolved between species.

R1 LINE elements at the *flamenco* locus

R1 elements are well-known to insert into rDNA genes, are transmitted vertically, and evolve similarly to the genome background rate [27]. They have also been found outside of rDNA genes, but only as fragments. *R1* elements are abundant within *flamenco* loci in the *simulans* clade. Outside of *flamenco*, *R1* elements in *D. simulans* are distributed according to expectation, with full length elements occurring only within rDNA (S3 File). Within *flamenco*, most copies of *R1* occur as tandem duplicates, creating large islands of fragmented *R1* copies (Fig 3A). They are on average 3.7% diverged from the reference *R1* from *D. simulans*. Across individual *D. simulans* genomes, ~99 kb of *flamenco* loci consists of *R1* elements, i.e. 26% of their average total length. *SZ45*, *LNP-15-062*, *NS40*, *MD251*, and *MD242* contain 4–7 full length copies of *R1* in the sense orientation, even though all but *SZ45* bear fragmented *R1* copies on the antisense strand. (The *SZ45 flamenco* assembly is incomplete, as the scaffold ends before the end of *Ty3/mdg4* enrichment). As the antisense *R1* copies are expected to suppress *R1* transposition, *flamenco* may not suppress these elements effectively. Alternatively, it is possible that *D. simulans flamenco* is still mostly active in the soma, while *R1* is active in the germline, and thus escapes host control by *flamenco*.

In *D. mauritiana*, *flamenco* harbors abundant fragments or copies of *R1* (19 on the reverse strand and 20 on the forward strand), and one large island of *R1* elements. In total, *D. mauritiana* contains 84 kb of *R1* sequence within *flamenco*. In *D. mauritiana* there are 8 full length copies of *R1* at the *flamenco* locus, 7 in antisense, which are not obviously due to a segmental or local duplication. Finally, we find that *D. sechellia flamenco* lacks full length copies of *R1*, and it contains only 18 KB of *R1* sequence (16 fragments on the reverse strand). Yet, all the copies are on the sense strand, which would not produce fragments that can suppress *R1* TEs. Essentially the antisense copies of *R1* in *D. mauritiana* should be suppressing the TE, but we see multiple full length antisense insertions, and *D. sechellia* has no antisense copies, but we see no evidence for recent *R1* insertions. From this it would appear that whatever is controlling the transposition of *R1* lies outside of *flamenco*.

The presence of long sense-strand *R1* elements within *flamenco* is a departure from expectation [17,27]. There is no evidence of an rDNA gene within the *flamenco* locus or the insertion site of *R1* within the 28S rDNA gene that would explain the insertion of *R1* elements there, nor is there precedence for the large expansion of *R1* fragments within the locus. Furthermore, the suppression of *R1* transposition does not appear to be controlled by *flamenco*.

piRNA production from R1

On average *R1* elements within the *flamenco* locus of *D. simulans* produce more piRNA than any other TE within *flamenco*. *R1* reads mapping to the forward strand constitute an average of 51% of the total piRNAs within the *flamenco* locus from the maternal fraction, ovary, and testis using weighted mapping. The maternal fraction constitutes the piRNA deposited by the

mother into the embryo. Weighted mapping refers to mapping where read counts are divided by the number of places they map, i.e. a read that maps to 50 locations is counted as 1/50. The only exception is the ovarian sample from *SZ232* which is a large outlier at only 5%. However *R1* reads mapping to the reverse strand account for an average of 84% of the piRNA being produced from the reverse strand in every genotype and tissue–maternal fraction, testis, or ovary. If unique mapping is considered instead of weighted these percentages are reduced by approximately 20%, which is to be expected given that *R1* is present in many repeated copies. Production of piRNA from the reverse strand seems to be correlated with elements inserted in the sense orientation, of which the vast majority are *R1* elements in *D. simulans* (S3 Fig). The production of large quantities of piRNA cognate to the *R1* element seemingly has no function–if *R1* only inserts at rDNA genes and are vertically transmitted there is little reason to be producing the majority of piRNA in response to this element.

In *D. sechellia* there are very few piRNA produced from *flamenco* in the maternal fraction or testis (which is expected for a cluster that is only active in ovarian somatic tissue), and there are no full length copies of *R1*. Likewise overall weighted piRNA production from *R1* elements on either strand is 2.8–5.9% of the total mapping piRNA. In contrast in *D. mauritiana* there are full length *R1* elements and abundant piRNA production in the maternal fraction and testis. In *D. mauritiana* an average of 28% of piRNAs mapping to the forward strand of *flamenco* are arising from *R1*, and 33% from the reverse strand. In *D. mauritiana R1* elements make up a smaller proportion of the total elements in the sense orientation (24%), versus *D. simulans* (55%).

Conservation of *flamenco*

The *dip1* gene and promoter region adjacent to each copy of *flamenco* are very conserved both within and between copies of *flamenco* (Fig 3A). The phylogenetic tree of the area suggests that we are correct in labeling the two copies as the original *flamenco* locus and the duplicate (Fig 3A). The original *flamenco* locus is more diverged amongst genotypes of *D. simulans* while the duplicate clusters closely together with short branch lengths (Fig 3A). The promotor region is also conserved and alignable between *D. melanogaster*, *D. sechellia*, *D. mauritiana*, and *D. simulans* (Fig 2D). However, the same is not true of the *flamenco* locus itself. Approximately 3 kb from the promoter *flamenco* diverges amongst genotypes and species and is no longer alignable by traditional sequence-based algorithms, as the TEs are essentially presence/absence polymorphisms that span multiple kb. There is no conservation of *flamenco* between *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana* (Fig 5). However, within the *simulans* clade many of the same TEs occupy the locus, suggesting that they are the current genomic invaders in each of these species (Fig 5).

In *D. simulans* the majority of full length TEs are private insertions– 54% in *flamenco* and 64% in the duplicate. Copies that are full length in one genotype but fragmented in others are counted as shared, not private. However, the TE must be full length in at least one genotype to be included in this grouping. Almost half of these private insertions in the duplicate are due to a single genotype with a unique section of sequence, in this case *MD251*. In *flamenco*, private insertions are the single largest category of transposable element insertions, followed by fixed insertions. Thus even within a single population there is considerable diversity at the *flamenco* locus, which could potentially lead to differences in the ability to suppress TEs in the somatic cells of the ovary. For example, full length copies of *297* are present in four genotypes either in *flamenco* or the duplicate, which would suggest that these genotypes are able to suppress this transposable element while the other genotypes are not. Germline suppression is redundant, thus absence of a TE in *flamenco* would not necessarily mean it is not suppressed in the germline. In contrast *mdg4-3* is present in more than one full length copy in *flamenco* and its duplicate in every genotype but one

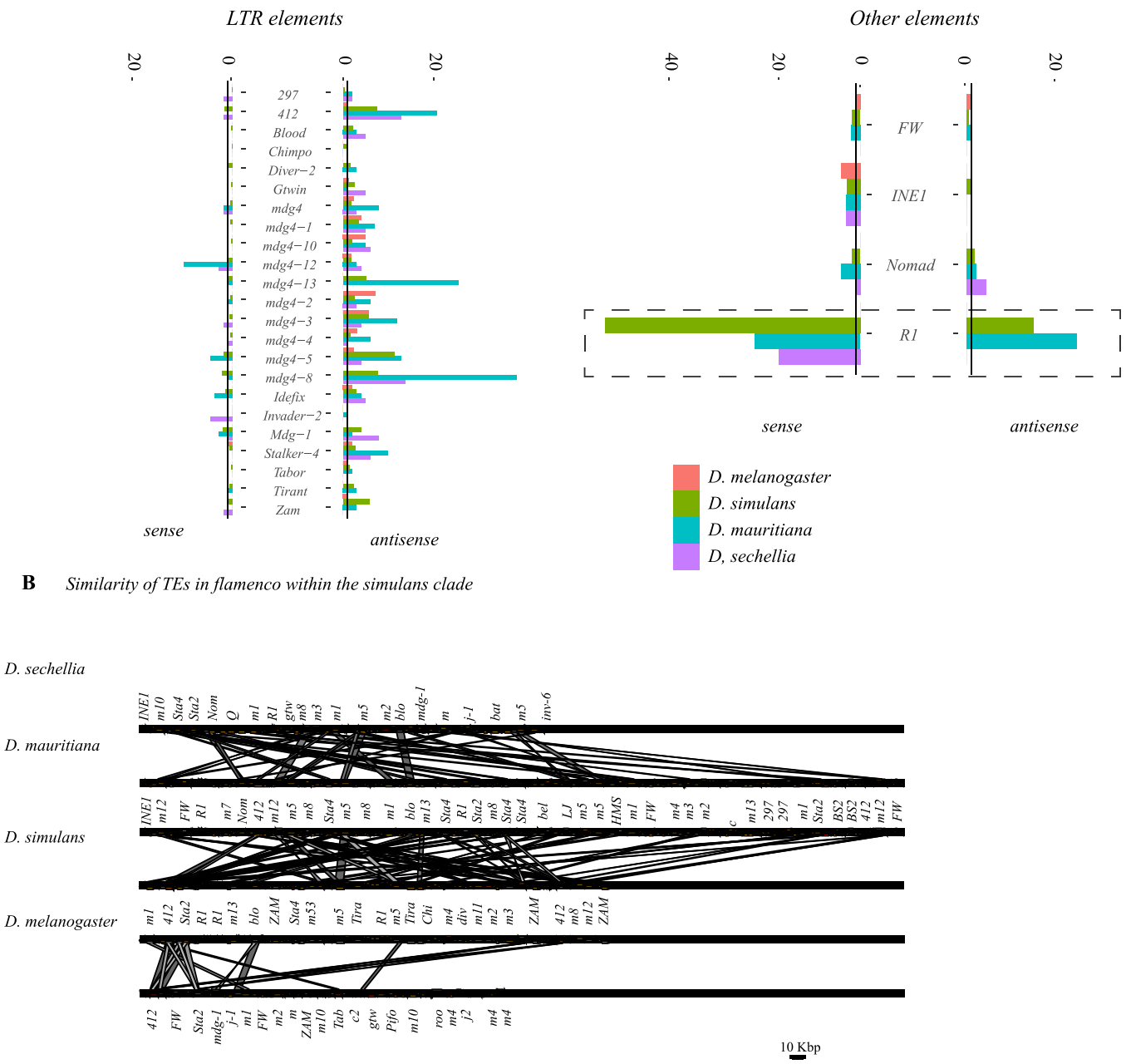

Fig 5. A. Copy number of a subset of transposable elements at *flamenco*. Solo LTRs are indicated by in a lighter shade at the top of the bar. The black line on each bar graph indicates a copy number of one. Values for *D. simulans* are the average for all genotypes with a complete *flamenco* assembly. 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 some of this is likely due to local duplications it is clearly a different pattern than *D. melanogaster*. Enrichment of LTR elements on the antisense strand is clear for all species. **B.** Alignment of *flamenco* in *D. melanogaster*, *D. simulans*, *D. sechellia*, and *D. mauritiana*. There is no conserved synteny between species but there are clearly shared TEs, particularly within the *simulans* clade. The expansion of *D. mauritiana* compared to the other species is apparent.

https://doi.org/10.1371/journal.pgen.1010914.g005

where it is present in a single copy. There are a number of these conserved full length TEs that are present in all or nearly all genotypes, including *Chimpo*, *mdg4-2*, *Tirant*, and *mdg4-4*. In addition, *INE1* elements adjacent to the promoter are conserved.

It is notable that any full length TEs are shared across all genotypes, given that wxD^1 was likely collected 30–50 years prior to the others, and the collections span continents (Jerry Coyne pers. comm.). Two facts are relevant to this observation: (1) TEs were shown not to correlate with geography [31] and (2) *D. simulans* is more diverse within populations than between different populations [32–34]. Other explanations are also plausible. Selection could be maintaining these full length TEs because TE deletions allow for TE reactivation that reduces fitness, wxD^1 could have had introgression from other lab strains, or a combination of these explanations.

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

In *D. melanogaster*, it was proposed that while germline clusters may have many insertions of a single TE, the somatic 'master regulator' *flamenco* will have a single insertion of each transposon, after which they are silenced and no longer able to transpose [17]. While the 'single copy' rule remains a hypothesis, it is largely supported in *D. melanogaster* where the two observed multicopy elements likely arose from segmental duplications. However, this is from an older and partially misassembled *flamenco* (18). In the past, this 'single copy' rule has appeared to apply only to full length insertions, with older degraded copies not effectively suppressing TEs [17]. To evaluate this model we will determine each of the following for full length TEs: (1) How many TEs have antisense oriented multicopy elements within *flamenco*? (2) How many *de novo* insertions of TEs in the *flamenco* duplicate of *D. simulans* are also present in the original *flamenco* copy? (3) How many TEs have full length and fragmented insertions, suggesting the older fragments did not suppress the newer insertion?

First we will evaluate the presence of antisense oriented multicopy elements within *flamenco* in each species. Due to the difficulty in classifying degraded elements accurately, for example between multiple *Ty3/mdg4* elements, we will focus here on full length TEs, suggesting recent transposition. In *D. melanogaster* there are 17 full length TEs (sense and antisense), one of which is present in multiple antisense copies. In *D. sechellia* there are 22 full length TEs within the *flamenco* locus, two of which are multicopy in antisense. *D. mauritiana* contains 41 full length TEs within the *flamenco* locus. Five of these are present in multiple antisense full length copies–*mdg4-5*, *R1*, *Stalker-4*, *jockey-3*, and *Cr1a*.

In *D. simulans* there are 26 full length TEs present in any of the seven complete *flamenco* assemblies. Six of these are present in multiple antisense copies within a single genome–*INE1*, *Chimpo*, *mdg4-4*, *412*, *Tirant*, and *BEL-unknown*. The two *Tirant* copies are likely a segmental duplication as they flank an *R1* repeat region. In the duplicate of *flamenco* in *D. simulans* there are 30 full length TEs, none of which are multicopy in antisense. However, there are TEs that are multicopy in antisense with respect to the original copy of *flamenco*—*mdg4-3*, *BEL-unknown*, *Nomad-1*, *Chimpo*, *mdg4-53A*, *R1*, and *INE1*. The fact that these elements are full length in both copies suggests independent insertions in each cluster rather than inheritance from duplication. Thus *D. simulans* and *D. mauritiana* overall do not meet the expectation that *flamenco* will contain a single insertion of any given TE.

Full length elements are generally younger insertions than fragmented insertions. Although we cannot know the order of insertions and deletions for sure, if a full length element is inserted in *flamenco* and there are fragments in the antisense orientation elsewhere in *flamenco* this suggests that *flamenco* did not successfully suppress the transposition of this element.

In *D. melanogaster* six elements have fragments in antisense that are less than 10% diverged from a full length TE (excluding TEs present in multiple antisense copies). In *D. sechellia* and D. mauritiana this is nine and five elements respectively. In *D. simulans* ten TEs fit this criteria in *flamenco* including *mdg4-2*, *mdg4-3*, *mdg4-5*, *412*, *INE1*, *R1* and *Zam*. In the duplicate of

flamenco in *D. simulans* there are nine TEs that fit this criteria, including *mdg4-2*, *mdg4-3*, *mdg4-5*, *297*, *Stalker-4*, and *R1*. In the *simulans* clade either fragments of TEs are not sufficient to suppress transposable elements or some elements are able to transpose despite the hosts efforts to suppress them.

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

High sequence similarity between TEs in different species could suggests horizontal transfer [35]. However, because sequence similarity can also exist due to vertical transmission we will use sequence similarity between *R1* elements (inserted at rDNA genes) as a baseline for differentiating horizontal versus vertical transfer. There has never been any evidence found for horizontal transfer of *R1* and it is thought to evolve at the same rate as nuclear genes in the *melanogaster* subgroup [17,27]. Similarity of *R1* ranges from 93% (*D. melanogaster*) to 97% (*D. sechellia*) thus TEs with a similarity of $>$ 98% are considered horizontally transferred. While this is not the most robust demonstration of horizontal transfer, it is suggestive. Of the full length elements present in *D. simulans* in any genome at *flamenco* 62% of them appear to have originated from horizontal transfer. This is similar to previous estimates for *D. melanogaster* in other studies [17]. Transfer appears to have occurred primarily between *D. melanogaster*, *D. sechellia*, and *D. willistoni*. This includes some known horizontal transfer events such as *Chimpo* and *Chouto* [36], and others which have not been recorded such as *mdg4-29* (*D. willistoni*) and the *Max-element* (*D. sechellia*) (S4 File). The duplicate of *flamenco* is similar, with 53% of full length TEs originating from horizontal transfer. They are many of the same TEs, with a 46% overlap, thus *flamenco* and its duplicate are trapping many of the same TEs. Both *flamenco* and the duplicate the region appears to serve as a trap for TEs originating from horizontal transfer. Note that we do not know the direction of transfer, thus the originating species could be *D. simulans*.

In *D. melanogaster* 46% of full length TEs appear to correspond to families undergoing recent horizontal transfer [17]. In *D. sechellia* 53% of full length TEs have arisen from horizontal transfer, including some known to have moved by horizontal transfer such as *GTWIN* (*D. melanogaster*/*D. erecta*) [36]. *D. mauritiana* has 68% of its full length TEs showing a closer relationship than expected by vertical descent with TEs from *D. sechellia*, *D. melanogaster*, and *D. simulans*. The hypothesis that *flamenco* serves as a trap for TEs entering the population through horizontal transfer holds throughout the *simulans* clade.

Discussion

The piRNA pathway is the organisms primary mechanism of transposon suppression. While the piRNA pathway is conserved, the regions of the genome that produce piRNA are labile, particularly in double stranded germline piRNA clusters [9]. The necessity of any single cluster for TE suppression in the germline piRNA pathway is unclear, but likely redundant [9]. However, *flamenco* is thought to be the master regulator of the somatic support cells of the ovary, preventing *Ty3*/*mdg4* elements from hopping into germline cells [1,17,20,22,23,37]. It is not redundant to other clusters, and insertion of a single element into *flamenco* in *D. melanogaster* is sufficient to initiate silencing. Here we show that the function of *flamenco* appears to have diversified in the *D. simulans* clade, acting potentially as both a germline and somatic piRNA cluster.

Dual stranded expression of *flamenco*

In this work, we showed that piRNAs of the *flamenco* locus in *D.simulans* and *D. mauritiana* are deposited maternally, align to both strands, and exhibit ping-pong signatures. This is in

contrast to *D. melanogaster*, where *flamenco* acts as a uni-strand cluster in the soma [3], our data thus suggest that the *flamenco* locus in *D. simulans* and *D. mauritiana* acts as a dual-strand cluster in the germline. In *D. sechellia* the attributes of *flamenco* uncovered in *D. melanogaster* appear to be conserved–no expression in the maternal fraction and the testis and no ping-pong signals. Given that *flamenco* is likely a somatic uni-strand cluster in *D. erecta*, we speculate that the conversion into a germline cluster happened in the *simulans* clade [3]. Such a conversion of a cluster between the somatic and the germline piRNA pathway is not unprecedented. For example, a single insertion of a reporter transgene triggered the conversion of the uni-stranded cluster *20A* in *D. melanogaster* into a dual-strand cluster [38].

The role of *flamenco* in *D. simulans* and *D. mauritiana* as the master regulator of piRNA in somatic support cells may still well be true–the promoter region of the *flamenco* cluster is conserved between species and between copies of *flamenco* within species. This suggests that in at least some contexts (or all) the cluster is still serving as a uni-strand cluster transcribed from a traditional RNA Pol II site [14]. However it has acquired additional roles, producing dual strand piRNA and ping-pong signals, in these two species, in at least the germline. However, in *D. simulans*, the majority of these reverse stranded piRNAs are emerging from the *R1* insertions within *flamenco*. There is no evidence at present that *R1* has undergone an expansion in its insertion positions in *D. simulans* (i.e. outside of 28s rDNA), thus it is unclear what, if any, impact the reverse stranded piRNAs have at the *flamenco* locus.

Duplication of *flamenco* in *D. simulans*

In *D. simulans*, *flamenco* is present in 2 genomic copies, and this duplication is present in all sequenced *D. simulans* lines except the reference strain. The *dip1* gene and putative *flamenco* promoter flanking the duplication also has a high similarity in all sequenced lines (Fig 2B). We do not have any direct evidence that *flamenco* is positively selected, but the high similarity between promoter regions across samples from different continents could suggest the possibility that the duplication of *flamenco* in *D. simulans* was positively selected. Such a duplication may be beneficial as it increases the ability of an organism to rapidly silence TEs. Individuals with large piRNA clusters (or duplicated ones) should accumulate fewer deleterious TE insertions than individuals with small clusters (or non-duplicated ones), and duplicated clusters may therefore confer a selective advantage [39].

Rapid evolution of piRNA clusters

A previous work showed that dual- and uni-strand clusters evolve rapidly in *Drosophila* [19]. In agreement with this work we also found that the *flamenco*-locus is rapidly evolving between and within species (Figs 1C and 3B). A major open question remains whether this rapid turnover is driven by selection (positive or negative) or an outcome of neutral processes (eg. high TE activity or insertion bias of TEs). These rapid evolutionary changes at the *flamenco* locus, a piRNA master locus, suggest that there is a constant turnover in patterns of piRNA biogenesis that potentially leads to changes in the level of transposition control between individuals in a population.

Materials and methods

Fly strains

The four *D. simulans* lines *SZ232*, *SZ45*, *SZ244*, and *SZ129* were collected in California from the Zuma Organic Orchard in Los Angeles, CA on two consecutive weekends of February 2012 [40–44]. *LNP-15-062* was collected in Zambia at the Luwangwa National Park by D.

Matute and provided to us by J. Saltz (J. Saltz pers. comm., [45,46]). *MD251*, *MD242*, *NS137*, and *NS40* were collected in Madagascar and Kenya (respectively) and are described in [47]. The *D. simulans* strain wxD^{1} was originally collected by M. Green, likely in California, but its provenance has been lost (pers. comm. Jerry Coyne). *D. mauritiana (w12)* and *D. sechellia (Rob3c/Tucson 14021–0248.25)* are described in [48] In addition, used the *D. melanogaster* dm6 reference assembly, which is strain *iso-1*.

Long read DNA sequencing and assembly

MD242, four SZ lines and *LNP-15-062* were sequenced on a MinION platform at North Dakota State University (Oxford Nanopore Technologies (ONT), Oxford, GB), with base-calling using guppy (v4.4.2). *MD242*, the four SZ lines, and *LNP-15-062* were assembled with Canu (v2.1) [49] and two rounds of polishing with Racon (v1.4.3) [50]. The CA strains were additionally polished with short reads using Pilon (v1.23) [51](SRR3585779, SRR3585440, SRR3585480, SRR3585391) [40]. The first $wxD^{1\text{-}1}$ assembly is described here [52]. *MD251*, *NS137*, *NS40* and $wxD^{1\text{-}2}$ were sequenced on a MinION platform at Stanford University. They were assembled with Flye [53], and polished with a round of Medaka followed by a round of pilon [51]. Following this contaminants were removed with blobtools (https://zenodo.org/record/845347, [54]), soft masked with RepeatModeler and Repeatmasker [55,56], then aligned to the wxD^{1} as a reference with Progressive Cactus [57]. The assemblies were finished with reference based scaffolding using Ragout [58]. *D. mauritiana* and *D. sechellia* were sequenced with PacBio RSII and P6-C4 chemistry and assembled with FALCON using default parameters (https://github.com/PacificBiosciences/FALCON) [48]. A summary of the assembly statistics is available in S1 Table. The quality of cluster assembly was evaluated using the coverage and soft clip quality as described in [19,59] (S1 File). These assemblies were deposited with NCBI under the accession number PRJNA907284.

Short read sequencing and mapping

Short read sequencing was performed by Beijing Genomics Institute on approximately 50 dissected ovaries from adult female flies (*SZ45*, *SZ129*, *SZ232*, *SZ244*, *LNP-15-062*, PRJNA913883). Short read libraries from 0–2 hour embryos were prepared from *D. melanogaster*, $wxD^{1\text{-}2}$, *D. sechellia*, and *D. mauritiana* (PRJNA1003528) [60]. Small RNA from testis is described in [61,62]. *D. melanogaster* OSC and ovarian small RNA libraries were downloaded from the SRA (SRR11999160, SRR11846566)[63]. Libraries were filtered for adapter contamination and short reads between 23–29 bp were retained for mapping with fastp [64]. The RNA was then mapped to their respective genomes (i.e. embryonic piRNA from $wxD^{1\text{-}2}$ was mapped to the wxD^{1} assembly) using bowtie (v1.2.3) and the following parameters (-q -v 1 -p 1 -S -a -m 50—best—strata) [65,66]. The resulting bam files were processed using samtools [67]. To obtain unique reads the bam files were filtered for reads with 1 mapping position. To obtain counts files with weighted mapping the bam files were processed using Rsubreads and the featureCounts function [68].

Defining and annotating piRNA clusters

piRNA clusters were initially defined using proTRAC [69]. piRNA clusters were predicted with a minimum cluster size of 1 kb (option "-clsize 1000"), a p-value for minimum read density of 0.07 (option "-pdens 0.07"), a minimum fraction of normalized reads that have 1T (1U) or 10A of 0.33 (option "-1Tor10A 0.33") and rejecting loci if the top 1% of reads account for more than 90% of the normalized piRNA cluster read counts (option "-distr 1–90"), and a minimal fraction of hits on the main strand of 0.25 (option "-clstrand 0.25"). Clusters were

annotated using RepeatMasker (v. 4.0.7) and the TE libraries described in Chakraborty et al. (2019) [52,55] (available at https://github.com/SignorLab/Flamenco_manuscript). The position of *flamenco* was also evaluated based off of the position of the putative promoter, the *dip1* gene, and the enrichment of *Ty3/mdg4* elements [14]. Enrichment of *Ty3/mdg4* elements was detected by counting the number of annotated *Ty3/mdg4* elements from Repeatmasker present per megabase with *dplyr* [70]. Fragmented annotations were merged to form TE copies with onecodetofindthemall [71]. Fragmented annotations were also manually curated within *flamenco*, particularly because TEs not present in the reference library often have their LTRs and internal sequences classified as different elements.

Aligning the *flamenco* promoter region

1 kb up and downstream of the *flamenco* promotor was extracted from each genotype and species with bedtools [72]. Sequences were aligned with clustal-omega and converted to nexus format [73]. Trees were built using a GTR substitution model and gamma distributed rate variation across sites [74]. Markov chain monte carlo iterations were run until the standard deviation of split frequencies was below .01, around one million generations. The consensus trees were generated using sumt conformat = simple. The resulting trees were displayed with the R package ape [75].

Detecting ping-pong signals in the small RNA data

Ping-pong signals were detected using pingpongpro v1.0 [76] This program detects the presence of RNA molecules that are offset by 10 nt, such that stacks of piRNA overlap by the first 10 nt from the 5' end. These stacks are a hallmark of piRNA mediated transposon silencing. The algorithm also takes into account local coverage and the presence of an adenine at the 10^{th} position. The output includes a z-score between 0 and 1, the higher the z-score the more differentiated the ping-pong stacks are from random local stacks.

Annotating shared and unique TE insertions

To align the TE annotations of homologous piRNA clusters, we first extracted the sequences of the clusters and annotated TEs in these sequences using RepeatMasker (open-4.0.7) with a custom TE library and the parameters: -s (sensitive search), -nolow (disable masking of low complexity sequences), and -no_is (skip check for bacterial IS) [48,77,78]. Finally, we aligned the resulting repeat annotations with Manna using the parameters -gap 0.09 (gap penalty), -mm 0.1 (mismatch penalty) -match 0.2 (match score) [19,48]. Manna can be used for aligning the annotations of the transposable elements by relying on synteny to determine insertion homology. Alignments were manually checked for inconsistencies arising from assignment to similar TEs (i.e. *mdg4-3* versus *mdg4-5*). TEs were considered to be full length if they were present in at least 70% of their reference length and contained internal sequence as well as two LTRs if applicable.

Horizontal transfer

TEs can be transferred vertically or horizontally. To attempt to distinguish between these two scenarios we used sequence similarity at *R1* insertions (at rDNA genes) as a baseline for differentiating the two scenarios. *R1* is only horizontally transferred and it is thought to evolve at the same rate as nuclear genes, therefore *R1* is an example of a vertically transferred TE [17,27].

Supporting information

S1 File. Coverage and soft clipping are both good indicators of assembly quality. Because piRNA clusters are so difficulty to assemble, we use an approach here called Cluster Busco. Essentially the rate of soft clipping and coverage are calculated for BUSCO genes. These are then compared to the piRNA clsuters to look for regions with considerably different coverage/soft clipping than BUSCO genes. Here are 99% quantiles for BUSCO genes indicated by the dotted lines for both coverage and soft clipping. Then the rate of soft clipping and coverage are shown as the black line for each assembly. In some cases the assembly was modified based off this information–for example NS40 has a spike in coverage/soft clipping which was an assembly error. The flamenco region in NS40 actually ends at that position.
(PDF)

S2 File. The primer pair used to confirm the flamenco duplicate in D. simulans, as well as the restriction enzyme used to digest the resulting PCR. Note that this will not work in every genotype.
(PDF)

S3 File. The position of R1 insertions and 28S rDNA. This example is from LNP-15-062. The chromosome, start, and end of the 28S rDNA is listed first, followed by the chromosome, start, and end of the R1 insertion. The last column corresponds to the length of the R1 TE. A full length R1 element is about 5429 bp long.
(PDF)

S4 File. TEs are considered horizontally transferred if they have a similarity great than 98%.
(XLSX)

S1 Fig. The height of the ping pong stacks and the distribution of z-scores greater than .8 in *D. mauritiana* maternal fraction and testis. Unlike in *D. simulans* there is a stronger enrichment of ping pong scores in the testis as flamenco, though both show ping pong signals in the maternal fraction.
(PDF)

S2 Fig. A higher resolution map of the transposons and uniquely mapping piRNA for one genotype of *D. simulans*. The original copy of flamenco is on the left, the duplicate on the right. Copy number of the fragmented dip1 gene is variable between strains, LNP-15-062 having more copies than most other genotypes. Dip1 is indicated by the blue boxes, and piRNA is shown as RPM. Note that not all transposons present in flamenco are labeled, given their fragmented nature they are often very dense which would require the labels to be very crowded. Included is enough to give a general map of the area.
(PDF)

S3 Fig. Transcription of piRNA from the reverse strand coordinates with the position of sense oriented R1 elements. Weighted abundance of piRNA mapping is shown for $wxD^{1\text{-}1}$ from the maternal fraction and the testis. piRNA mapping to the forwards strand is shown in blue, the reverse strand in red. The bottom panel shows the weighted abundance of piRNA in the ovary for the strain LNP-15-062. Note that this is a different region as the two genotype's flamenco loci do not show large homologous regions.
(PDF)

S1 Table. Assembly statistics for each of the newly assembled genomes.
(PDF)

S2 Table. The location of *flamenco* and its duplicate in each of the assemblies included in this manuscript.
(PDF)

S3 Table. The percent of annotated TEs that are in the antisense orientation within the flamenco region and the % of those antisense TEs that belong to the LTR class of TEs.
(PDF)

Acknowledgments

Thanks to Colin Meiklejohn for providing some of the fly strains used in this manuscript. We would also like to thank Dimitri Petrov and his lab for providing logistical support to BYK. SS would like to thank Jeff Kittilson for assistance in the laboratory. SS would also like to thank C & F & S Emery for insightful commentary on the manuscript.

Author Contributions

Conceptualization: Sarah Signor, Jeffrey Vedanayagam, Robert Kofler.

Funding acquisition: Sarah Signor, Jeffrey Vedanayagam, Eric C. Lai.

Investigation: Sarah Signor.

Project administration: Eric C. Lai.

Resources: Sarah Signor, Jeffrey Vedanayagam, Bernard Y. Kim, Eric C. Lai.

Software: Bernard Y. Kim.

Supervision: Filip Wierzbicki, Eric C. Lai.

Validation: Filip Wierzbicki.

Visualization: Sarah Signor, Filip Wierzbicki, Robert Kofler.

Writing – original draft: Sarah Signor, Robert Kofler, Eric C. Lai.

Writing – review & editing: Sarah Signor, Robert Kofler, Eric C. Lai.

References

1. Duc C, Yoth M, Jensen S, Mouniée N, Bergman CM, Vaury C, et al. Trapping a somatic endogenous retrovirus into a germline piRNA cluster immunizes the germline against further invasion. Genome Biol. 2019; 20: 127. https://doi.org/10.1186/s13059-019-1736-x PMID: 31227013
2. Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res. 2018; 46: gky761–. https://doi.org/10.1093/nar/gky761 PMID: 30312469
3. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, et al. Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. Cell. 2009; 137: 522–535. https://doi.org/10.1016/j.cell.2009.03.040 PMID: 19395010
4. Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, Nagami T, et al. A Slicer-Mediated Mechanism for Repeat-Associated siRNA 5' End Formation in Drosophila. Science. 2007; 315: 1587–1590. https://doi.org/10.1126/science.1140494 PMID: 17322028
5. Wang SH, Elgin SCR. Drosophila Piwi functions downstream of piRNA production mediating a chromatin-based transposon silencing mechanism in female germ line. Proc National Acad Sci. 2011; 108: 21164–21169. https://doi.org/10.1073/pnas.1107892109 PMID: 22160707
6. Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, Sachidanandam R, et al. Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. Cell. 2007; 128: 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043 PMID: 17346786

7. Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, Snyder B, et al. The Small RNA Profile during Drosophila melanogaster Development. Developmental Cell. 2003; 5: 337–350. https://doi.org/10.1016/s1534-5807(03)00228-4 PMID: 12919683

8. Chirn G, Rahman R, Sytnikova YA, Matts JA, Zeng M, Gerlach D, et al. Conserved piRNA Expression from a Distinct Set of piRNA Cluster Loci in Eutherian Mammals. Plos Genet. 2015; 11: e1005652. https://doi.org/10.1371/journal.pgen.1005652 PMID: 26588211

9. Gebert D, Neubert LK, Lloyd C, Gui J, Lehmann R, Teixeira FK. Large Drosophila germline piRNA clusters are evolutionarily labile and dispensable for transposon regulation. Mol Cell. 2021; 81: 3965–3978. e5. https://doi.org/10.1016/j.molcel.2021.07.011 PMID: 34352205

10. Andersen PR, Tirian L, Vunjak M, Brennecke J. A heterochromatin-dependent transcription machinery drives piRNA expression. Nature. 2017; 549: 54–59. https://doi.org/10.1038/nature23482 PMID: 28847004

11. Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, et al. The Drosophila HP1 Homolog Rhino Is Required for Transposon Silencing and piRNA Production by Dual-Strand Clusters. Cell. 2009; 138: 1137–1149. https://doi.org/10.1016/j.cell.2009.07.014 PMID: 19732946

12. Mohn F, Sienski G, Handler D, Brennecke J. The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. Cell. 2014; 157: 1364–1379. https://doi.org/10.1016/j.cell.2014.04.031 PMID: 24906153

13. Chen Y-CA, Stuwe E, Luo Y, Ninova M, Le Thomas A, Rozhavskaya E, et al. Cutoff Suppresses RNA Polymerase II Termination to Ensure Expression of piRNA Precursors. Mol Cell. 2016; 63: 97–109. https://doi.org/10.1016/j.molcel.2016.05.010 PMID: 27292797

14. Goriaux C, Desset S, Renaud Y, Vaury C, Brasset E. Transcriptional properties and splicing of the flamencopi RNAcluster. EMBO reports. 2014; 15: 411–418. https://doi.org/10.1002/embr.201337898 PMID: 24562610

15. Sienski G, Dönertas D, Brennecke J. Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. Cell. 2012; 151: 964–980. https://doi.org/10.1016/j.cell.2012.10.040 PMID: 23159368

16. Dennis C, Brasset E, Vaury C. flam piRNA precursors channel from the nucleus to the cytoplasm in a temporally regulated manner along Drosophila oogenesis. Mobile DNA. 2019; 10: 203–9. https://doi.org/10.1186/s13100-019-0170-7 PMID: 31312260

17. Zanni V, Eymery A, the MCP of, 2013. Distribution, evolution, and diversity of retrotransposons at the flamenco locus reflect the regulatory properties of piRNA clusters. National Acad Sciences. 110:19842–19847.

18. Bergman CM, Quesneville H, Anxolabéhère D, Ashburner M. Recurrent insertion and duplication generate networks of transposable element sequences in the Drosophila melanogaster genome. Genome Biology. 2006; 7: R112–21. https://doi.org/10.1186/gb-2006-7-11-r112 PMID: 17134480

19. Wierzbicki F, Kofler R, Signor S. Evolutionary dynamics of piRNA clusters in Drosophila. Mol Ecol. 2023; 32:1396–1322. https://doi.org/10.1111/mec.16311 PMID: 34878692

20. Prud'homme N, Gans M, Masson M, Terzian C, Bucheton A. Flamenco, a gene controlling the gypsy retrovirus of Drosophila melanogaster. Genetics. 1995; 139: 697–711. https://doi.org/10.1093/genetics/139.2.697 PMID: 7713426

21. Song SU, Gerasimova T, Kurkulos M, Boeke JD, Corces VG. An env-like protein encoded by a Drosophila retroelement: evidence that gypsy is an infectious retrovirus. Genes & development. 1994; 8: 2046–2057. https://doi.org/10.1101/gad.8.17.2046 PMID: 7958877

22. Mével-Ninio M, Pelisson A, Kinder J, Campos AR, Bucheton A. The flamenco Locus Controls the gypsy and ZAM Retroviruses and Is Required for Drosophila Oogenesis. Genetics. 2007; 175: 1615–1624. https://doi.org/10.1534/genetics.106.068106 PMID: 17277359

23. Pelisson A, Song SU, Prud'homme N, Smith PA, Bucheton A, Corces VG. Gypsy transposition correlates with the production of a retroviral envelope-like protein under the tissue-specific control of the Drosophila flamenco gene. The EMBO Journal. 1995; 13: 4401–4411.

24. Bucheton A. The relationship between the flamenco gene and gypsy in Drosophila: how to tame a retrovirus. Trends Genet. 1995; 11: 349–353. https://doi.org/10.1016/s0168-9525(00)89105-2 PMID: 7482786

25. Malone CD, Hannon GJ. Molecular Evolution of piRNA and Transposon Control Pathways in Drosophila. Cold Spring Harbor Symposia on Quantitative Biology. 2010; 74: 225–234. https://doi.org/10.1101/sqb.2009.74.052 PMID: 20453205

26. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007; 450: 203–218. https://doi.org/10.1038/nature06341 PMID: 17994087

27. Eickbush DG, Lathe WC, Francino MP, Eickbush TH. R1 and R2 retrotransposable elements of Drosophila evolve at rates similar to those of nuclear genes. Genetics. 1995; 139: 685–695. https://doi.org/10.1093/genetics/139.2.685 PMID: 7713425

28. Lopik J van Trapotsi M-A, Hannon GJ, Bornelöv S, Nicholson BC. Unistrand piRNA clusters are an evolutionary conserved mechanism to suppress endogenous retroviruses across the Drosophila genus. Biorxiv. 2023; 2023.02.27.530199. https://doi.org/10.1101/2023.02.27.530199

29. Durdevic Z, Pillai RS, Ephrussi A. Transposon silencing in the Drosophila female germline is essential for genome stability in progeny embryos. Life Sci Alliance. 2018; 1: e201800179. https://doi.org/10.26508/lsa.201800179 PMID: 30456388

30. Czech B, Preall JB, McGinn J, Hannon GJ. A Transcriptome-wide RNAi Screen in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway. Mol Cell. 2013; 50: 749–761. https://doi.org/10.1016/j.molcel.2013.04.007 PMID: 23665227

31. Lerat E, Goubert C, Guirao-Rico S, Merenciano M, Dufour A-B, Vieira C, et al. Population specific dynamics and selection patterns of transposable element insertions in European natural populations. Molecular Ecology. 2018; 1–42. https://doi.org/10.1111/mec.14963 PMID: 30506554

32. Singh RS. Population genetics and evolution of species related to Drosophila melanogaster. Annual Review of Genetics. 1989; 23: 425–453. https://doi.org/10.1146/annurev.ge.23.120189.002233 PMID: 2515792

33. Machado HE, Bergland AO, O'Brien KR, Behrman EL, Schmidt PS, Petrov DA. Comparative population genomics of latitudinal variation in Drosophila simulans and Drosophila melanogaster. Molecular Ecology. 2016; 25: 723–740. https://doi.org/10.1111/mec.13446 PMID: 26523848

34. Sedghifar A, Saelao P, Begun DJ. Genomic patterns of geographic differentiation in Drosophila simulans. Genetics. 2016. https://doi.org/10.1534/genetics.115.185496 PMID: 26801179

35. Loreto ELS, Carareto CMA, Capy P. Revisiting horizontal transfer of transposable elements in Drosophila. Heredity. 2008; 100: 545–554. https://doi.org/10.1038/sj.hdy.6801094 PMID: 18431403

36. Bargues N, Lerat E. Evolutionary history of LTR-retrotransposons among 20 Drosophila species. Mobile Dna-uk. 2017; 8: 7. https://doi.org/10.1186/s13100-017-0090-3 PMID: 28465726

37. Pélisson A and T. About the origin of retroviruses and the co-evolution of the gypsy retrovirus with the Drosophila flamenco host gene. ["Capy, Pierre"], editors. 1997; 29–37. https://doi.org/10.1007/978-94-011-4898-6_3

38. Luo Y, He P, Kanrar N, Toth KF, Aravin A. Maternally inherited siRNAs initiate piRNA cluster formation. bioRxiv 2022.02.08.479612. https://doi.org/10.1101/2022.02.08.479612

39. Kofler R. piRNA Clusters Need a Minimum Size to Control Transposable Element Invasions. Schaack S, editor. Genome Biology and Evolution. 2020; 12: 736–749. https://doi.org/10.1093/gbe/evaa064 PMID: 32219390

40. Signor SA, New FN, Nuzhdin S. A Large Panel of Drosophila simulans Reveals an Abundance of Common Variants. Genome Biology and Evolution. 2017; 10: 189–206. https://doi.org/10.1093/gbe/evx262 PMID: 29228179

41. Signor S, Nuzhdin S. Dynamic changes in gene expression and alternative splicing mediate the response to acute alcohol exposure in Drosophila melanogaster. Heredity. 2018; 121:342–360. https://doi.org/10.1038/s41437-018-0136-4 PMID: 30143789

42. Signor S. Population genomics of Wolbachia and mtDNA in Drosophila simulans from California. Scientific Reports. 2017; 1–11. https://doi.org/10.1038/s41598-017-13901-3 PMID: 29042606

43. Signor SA, Abbasi M, Marjoram P, Nuzhdin SV. Social effects for locomotion vary between environments in Drosophila melanogaster females. Evolution. 2017; 71: 1765–1775. https://doi.org/10.1111/evo.13266 PMID: 28489252

44. Signor S. Transposable elements in individual genotypes of Drosophila simulans. Ecology and Evolution. 2020; 130: 499–11. https://doi.org/10.1002/ece3.6134 PMID: 32273997

45. Matute DR, Gavin-Smyth J, Liu G. Variable post-zygotic isolation in Drosophila melanogaster/D. simulanshybrids. Journal of Evolutionary Biology. 2014; 27: 1691–1705. https://doi.org/10.1111/jeb.12422 PMID: 24920013

46. Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of Drosophila simulans and D. sechellia. Payseur BA, editor. PLoS Genetics. 2018; 14: e1007341–29. https://doi.org/10.1371/journal.pgen.1007341 PMID: 29684059

47. Rogers RL, Cridland JM, Shao L, Hu TT, Andolfatto P, Thornton KR. Landscape of Standing Variation for Tandem Duplications in Drosophila yakuba and Drosophila simulans. Molecular Biology and Evolution. 2014; 31: 1750–1766. https://doi.org/10.1093/molbev/msu124 PMID: 24710518

48. Chakraborty M, Chang C-H, Khost DE, Vedanayagam J, Adrion JR, Liao Y, et al. Evolution of genome structure in the Drosophila simulans species complex. 2020; 139: 1067–63. https://doi.org/10.1101/2020.02.27.968743

49. Koren S., Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017; 27:722–736. https://doi.org/10.1101/gr.215087.116 PMID: 28298431

50. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017; 27: 737–746. https://doi.org/10.1101/gr.214270.116 PMID: 28100585

51. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. Plos One. 2014; 9: e112963. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509

52. Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nature Communications. 2019; 1–11. https://doi.org/10.1038/s41467-019-12884-1 PMID: 31653862

53. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019; 37: 540–546. https://doi.org/10.1038/s41587-019-0072-8 PMID: 30936562

54. Laetsch DR, Blaxter ML. BlobTools: Interrogation of genome assemblies. F1000research. 2017; 6: 1287. https://doi.org/10.12688/f1000research.12232.1

55. Tarailo-Graovac M, Chen N. Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Current Protocols in Bioinformatics. 2009; 1–14. https://doi.org/10.1002/0471250953.bi0410s25 PMID: 19274634

56. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc National Acad Sci. 2020; 117: 9451–9457. https://doi.org/10.1073/pnas.1921046117 PMID: 32300014

57. Armstrong J, Hickey G, Diekhans M, Fiddes IT, Novak AM, Deran A, et al. Progressive Cactus is a multiple-genome aligner for the thousand-genome era. Nature. 2020; 587: 246–251. https://doi.org/10.1038/s41586-020-2871-y PMID: 33177663

58. Kolmogorov M, Armstrong J, Raney BJ, Streeter I, Dunn M, Yang F, et al. Chromosome assembly of large and complex genomes using multiple references. Genome Res. 2018; 28: 1720–1732. https://doi.org/10.1101/gr.236273.118 PMID: 30341161

59. Wierzbicki F, Schwarz F, Cannalonga O, Kofler R. Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. Mol Ecol Resour. 2022; 22: 102–121. https://doi.org/10.1111/1755-0998.13455 PMID: 34181811

60. Vedanayagam Jeffrey. Evolutionary Genomics of piRNA Mediated Transposon Silencing in Drosophila. University of Rochester. 2016.

61. Vedanayagam J, Lin C-J, Papareddy R, Nodine M, Flynt AS, Wen J, et al. Endogenous RNAi silences a burgeoning sex chromosome arms race. Biorxiv. 2022; 2022.08.22.504821. https://doi.org/10.1101/2022.08.22.504821

62. Vedanayagam J, Lin C-J, Lai EC. Rapid evolutionary dynamics of an expanding family of meiotic drive factors and their hpRNA suppressors. Nat Ecol Evol. 2021; 5: 1613–1623. https://doi.org/10.1038/s41559-021-01592-z PMID: 34862477

63. Schwarz F, Wierzbicki F, Senti K-A, Kofler R. Tirant stealthily invaded natural Drosophila melanogaster populations during the last century. Mol Biology Evol. 2020; 38: msaa308–. https://doi.org/10.1093/molbev/msaa308 PMID: 33247725

64. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Biorxiv. 2018; 274100. https://doi.org/10.1093/bioinformatics/bty560 PMID: 30423086

65. Axtell MJ. ShortStack: Comprehensive annotation and quantification of small RNA genes. RNA. 2013; 19: 740–751. https://doi.org/10.1261/rna.035279.112 PMID: 23610128

66. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009; 10: R25. https://doi.org/10.1186/gb-2009-10-3-r25 PMID: 19261174

67. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943

68. Liao Y, Smyth GK, Shi W. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 2019; 47: gkz114–. https://doi.org/10.1093/nar/gkz114 PMID: 30783653

69. Rosenkranz D, Zischler H. proTRAC—a software for probabilistic piRNA cluster detection, visualization and analysis. Bmc Bioinformatics. 2012; 13: 5. https://doi.org/10.1186/1471-2105-13-5 PMID: 22233380

70. Wickham H, Francois R. dplyr: A Grammar of Data Manipulation. 2015.

71. Bailly-Bechet M, Haudry A, Lerat E. "One code to find them all": a perl tool to conveniently parse RepeatMasker output files. Mobile Dna-uk. 2014; 5: 13. https://doi.org/10.1186/1759-8753-5-13

72. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

73. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018; 27: 135–145. https://doi.org/10.1002/pro.3290 PMID: 28884485

74. Ronquist F, Teslenko M, Mark P van der, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. Systematic Biology. 2012; 61: 539–542. https://doi.org/10.1093/sysbio/sys029 PMID: 22357727

75. Paradis E, Claude J, Strimmer K. APE: Analyses of Phylogenetics and Evolution in R language. Bioinformatics. 2004; 20: 289–290. https://doi.org/10.1093/bioinformatics/btg412 PMID: 14734327

76. Uhrig S, Klein H. PingPongPro: a tool for the detection of piRNA-mediated transposon-silencing in small RNA-Seq data. Bioinformatics. 2018; 35: 335–336. https://doi.org/10.1093/bioinformatics/bty578 PMID: 29985981

77. Smit, Hubley A and, Green R and, P. RepeatMasker Open-4.0. 2013–2015. 2005.

78. Quesneville H, Bergman CM, Andrieu O, Autard D, Nouaud D, Ashburner M, et al. Combined Evidence Annotation of Transposable Elements in Genome Sequences. Plos Comput Biol. 2005; 1: e22. https://doi.org/10.1371/journal.pcbi.0010022 PMID: 16110336