Transposable element dynamics are consistent across the *Drosophila* phylogeny, despite drastically differing content

Tom Hill[1]*

1. 4012 Haworth Hall, The Department of Molecular Biosciences, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS 66045. Email: tom.hill@ku.edu

* Corresponding author

Keywords: Transposable elements, *Drosophila*, fitness.

Abstract

The evolutionary dynamics of transposable elements (TEs) vary across the tree of life and even between closely related species with similar ecologies. In *Drosophila*, most of the focus on TE dynamics has been completed in *Drosophila melanogaster* and the overall pattern indicates that TEs show an excess of low frequency insertions, consistent with their fitness cost in the genome. However, work outside of *D. melanogaster,* in the species *Drosophila algonquin*, suggests that this situation may not be universal, even within *Drosophila*. Here we test whether the pattern observed in *D. melanogaster* is similar across five *Drosophila* species that share a common ancestor more than fifty million years ago. We find that, for most TE families and orders, the patterns are broadly conserved between species, suggesting TEs are primarily costly, and dynamics are conserved in orthologous regions of the host genome. These results suggest that most TEs retain similar activities and fitness costs across the *Drosophila* phylogeny suggesting little evidence of drift in the dynamics of TEs across the phylogeny.

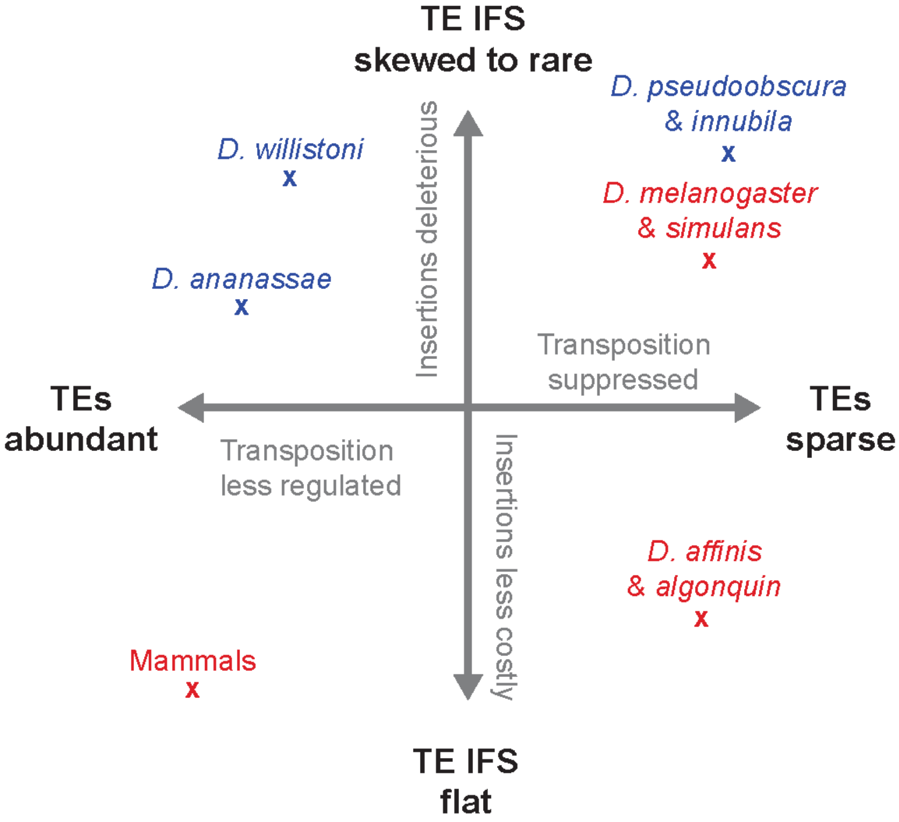
Introduction

Transposable elements are selfish mobile genetic elements found throughout the genomes of a majority of living organisms; these sequences copy and move throughout hosts genomes, mostly to the detriment of the host (McClintock 1953; Orgel and Crick 1980; Charlesworth and Langley 1989; Burt and Trivers 2006; Wicker *et al.* 2007). Mammals, have few active transposable elements (TEs), a large proportion of their genomes are composed of TE insertions fixed within a species population (Hellen and Brookfield 2013a; b). Comparatively, TEs in the fruit fly *Drosophila* appear to be highly active, resulting in polymorphic insertions for most TE families within a species population, with a lower proportion of their genome comprised of TEs (Charlesworth and Langley 1989; Charlesworth *et al.* 1997) .

These differences can be explained with a model described by Lee and Langley (Lee and Langley 2010). TE insertions are primarily deleterious to the host; their insertion can interrupt a gene, cause aberrant expression or differential exon expression (Charlesworth and Langley 1989; Burt and Trivers 2006; Lee and Langley 2010, 2012). Without regulation, TEs are also rampantly expressed and transposing (Lee and Langley 2010; Blumenstiel 2011). To combat this, TE activity is suppressed, in the case of most animals, via the piRNA system (Aravin *et al.* 2007; Brennecke *et al.* 2007, 2008; Lu and Clark 2010). Using small RNAs transcribed from TE sequences, the piRNA system targets and degrades complementary TE mRNAs and cause heterochromatin formation on similar TE insertions (Obbard *et al.* 2009; Blumenstiel 2011; Lee 2015; Senti *et al.* 2015). Within this suppression system, the extent of silencing is then dependent on the expression and copy number of TEs, resulting in the copy number regulation seen in *Drosophila* (Lee and Langley 2010). However, the piRNA system can cause the propagation of heterochromatic silencing marks around TE insertions, resulting in the silencing of nearby genes and position effect variegation (Lee and Langley 2010; Lee 2015). This deleterious side effect, in combination with the deleterious effects of TE insertions suggests TE insertions should be rare in euchromatic regions (Charlesworth and Langley 1989; Charlesworth *et al.* 1997; Lee and Langley 2010).

Within this model, TEs will enter a genome and spread rapidly through a burst of unsuppressed transposition (Kofler *et al.* 2012; Lee and Langley 2012). The TE will be silenced via the piRNA system and regulated so long as piRNAs are produced against the TE (Senti and Brennecke 2010; Blumenstiel 2011). Following this, you should expect larger genomes with fewer active TEs, such as mammals, to have higher TE abundances and TE insertion frequency spectra (IFS) showing no skew towards rare insertions as TE insertions are on average, less costly (Figure 1) (Lee and Langley 2012; Hellen and Brookfield 2013a; Lee 2015). While species with higher effective population sizes, higher coding densities and more active TEs, such as *Drosophila melanogaster*, should have lower abundances of TEs and IFS skewed to rare insertions (Lee and Langley 2010; Petrov *et al.* 2011; Kofler *et al.* 2012, 2015b).

Figure 1: Schematic depicting the model explaining the differences seen between mammals and *Drosophila,* with species analyzed previously in red, species analyzed here in blue. Species have been placed in the schematic based on 1 – the insertion frequency spectrum relative to mammals and *D. melanogaster*, and 2 – TE abundances compared to mammals and *D. melanogaster*.



However, the expectation of lower euchromatic TE abundances, consistent with higher coding densities seen in *Drosophila melanogaster* is not seen in all *Drosophila* species (Clark *et al.* 2007). The dynamic nature of *Drosophila* TEs can be clearly seen in the 12-genomes project, a group of 12 sequenced *Drosophila* species genomes, that span the ~50 million year *Drosophila* genus, with species in both the *Drosophila* and *Sophophora* sister subgenera (Markow and O'Grady 2006; Clark *et al.* 2007). The sequenced species, show striking differences between TE families and orders, and make up differing proportions of the genome, between 5 and 40% across the tree (Sessegolo *et al.* 2016). Additionally, the TE content of two species in the *D. affinis* subgroup, is not comprised of lower copy number families with an excess of low frequency insertions (Hey 1989). Instead they have a few, highly abundant families, with many high frequency insertions, like mammalian genomes, despite their small genome and large effective population sizes (McGaugh *et al.* 2012; Palmieri *et al.* 2014). Though the methods used in this study are not truly comparable to modern techniques of assessing TE abundances, together with the diversity of abundances in the 12 genomes it brings into question the extent to which the previously described model fits outside the *D. melanogaster,* and where within the frame work other species fit (Hey 1989; Clark *et al.* 2007).

Here, we use next generation sequencing data and modern TE content identification methods to assess the TE insertion densities and TE insertion frequency spectra of the euchromatic genome of five *Drosophila* species. We attempt to identify if TEs show patterns consistent with insertions being rare and primarily deleterious, or if they differ between species with differing abundances of TEs. We find that despite differing TE abundances and euchromatic insertion densities between species, most TE insertions have an IFS consistent with families being highly active and deleterious in all species, though some individual families differ in their insertion frequencies between species (Figure 1). This suggests that TEs remain consistently deleterious across the *Drosophila* phylogeny, despite strong phylogenetic differences between species, and large changes in effective population size and TE densities (Sessegolo *et al.* 2016).

Results

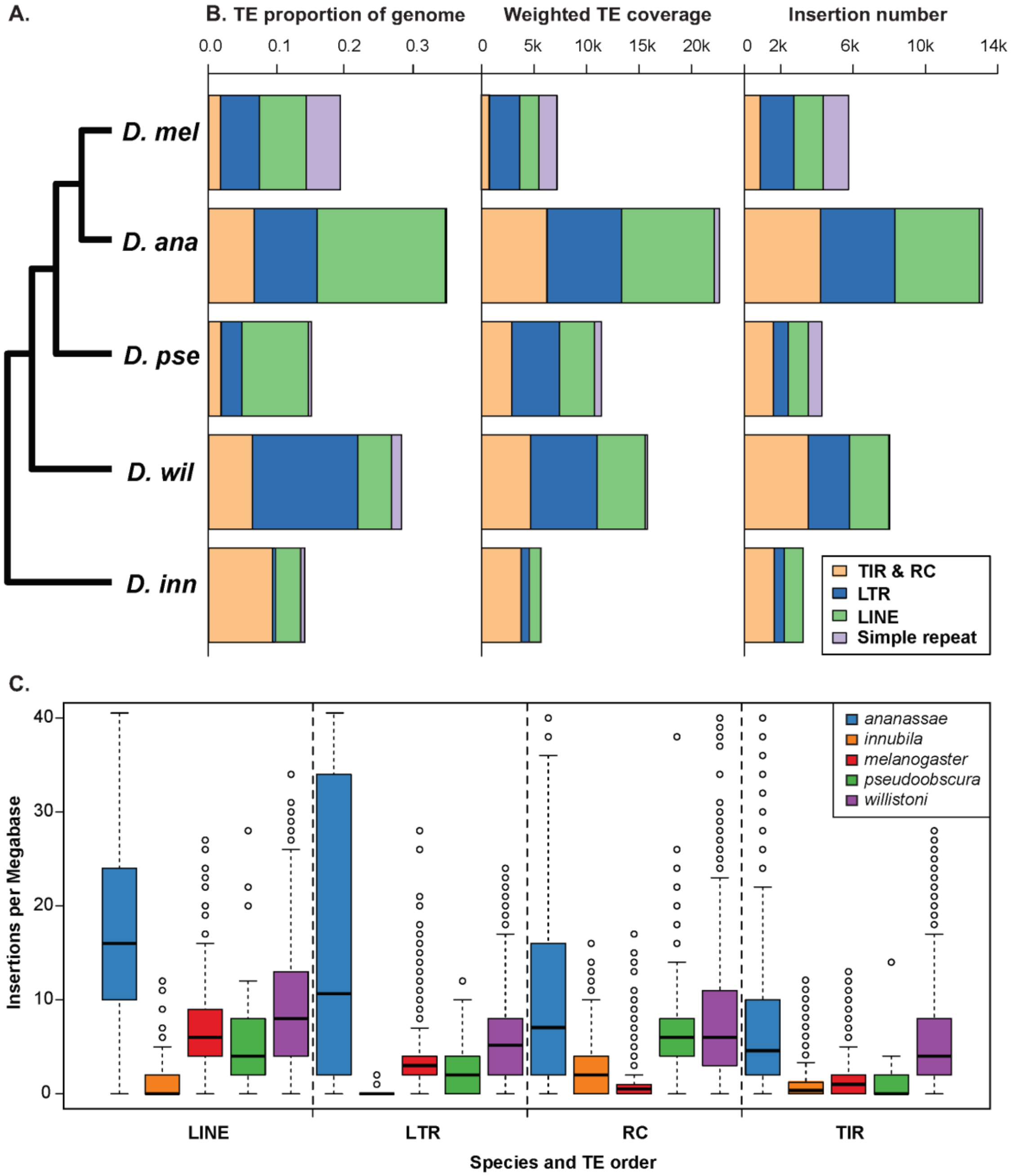
TE content differs drastically across the species examined

To examine the abundance and fitness cost of TE insertions across our *Drosophila* phylogeny of five species (Figure 1, 2A), we generated profiles of the TE content of each species using a combination of *RepeatMasker*, *BEDTools* and *PopoolationTE2* (Tarailo-Graovac and Chen 2009; Quinlan and Hall 2010; Kofler *et al.* 2011b). We estimated the proportion of each genome made up of TE insertions (Tarailo-Graovac and Chen 2009; Quinlan and Hall 2010; Kofler *et al.* 2016), the median copy number of each TE family and the median insertion number of each family in the euchromatic portion of the genome. We grouped families by their orders, either terminal inverted repeat (TIR) and rolling circle (RC) DNA transposons, or long terminal repeat (LTR) and long interspersed nuclear elements (LINE) RNA retrotransposons (Kohany *et al.* 2006; Wicker *et al.* 2007). Within each species, the TE content varies drastically – between 15% and 40% of each genome (Figure 2B), with consistently different numbers of TE copies and euchromatic insertions between species (Figure 2B). As identified elsewhere, there is a significant association between genome size and TE content, as found previously (Supplementary Table 2, *p*-value = 0.00176) (Gregory 2005; Wicker *et al.* 2007; Gregory and Johnston 2008).

The recently assembled and annotated genome of *D. innubila* has considerably lower insertion count numbers, perhaps due to the inferior annotation of TE content compared to other species. Interestingly, the *D. innubila* genome appears to have a lower amount of LTRs than most other studied *Drosophila* species (Hill *et al.* 2019), showing a similar profile to the relatively closely related *D. mojavensis* (Sessegolo *et al.* 2016). Most other species have retrotransposons, such as LTRs and LINEs, making up a large proportion of their repeat content (Figure 2B) (Clark *et al.* 2007). As shown previously, *D. ananassae* and *D. willistoni* have much higher TE content than the other species analyzed here (Clark *et al.* 2007; Sessegolo *et al.* 2016). These species differ in genome size, including an expanded Muller Element F in *D. ananassae* (Clark *et al.* 2007; Leung and Students 2017). In fact, there is an excess of TE content in *D. ananassae* on Muller element F. This element this represents only ~11.6% of the assembled reference genome (based on *D. melanogaster* orthology) but contains ~21.1% of the reference genomes TE content (based on *RepeatMasker* estimates), and so may account for the differences seen here.

To control for this Muller element expansion and other differences in genome size, we measured the TE insertion density per autosomal euchromatic megabase and found a significant excess of TE insertions per MB in *D. ananassae* and *D. willistoni* versus all other species, in all TE orders (Figure 2C, quasi-Poisson GLM, z-value > 19.296, *p*-value < 0.000565). These differences in TE abundances suggest that TE insertions may have differing dynamics between species, even when excluding TE rich regions. Due to the larger genomes and more abundant TE insertions, insertions may be less costly in *D. ananassae* and *D. willistoni* compared to other species and so may be more common in populations, with IFS skewed towards higher frequencies (Aravin *et al.* 2007; Blumenstiel 2011; Levine and Malik 2011).

Figure 2. Transposable Element content (separated by TE order) in populations of five *Drosophila* species. TE content shown as **A.** Cartoon of tree of species assessed here, branches do not accurately represent the distance between species. **B.** Estimated TE profiles including TE proportions of each genome, median TE coverage, weighted by median nuclear coverage, and median TE insertion number. TIR &RCs were combined due to small numbers of either for many species. **C.** TE density per 1 Mb windows across the genome for each species and TE order.



TE insertions are primarily rare across the *Drosophila* phylogeny

Using the TE insertions called with *PopoolationTE2*, we found the insertion frequency spectrum (IFS) across each TE order, across all species, limited to the autosomes (Kofler *et al.* 2016). Like the differing TE insertion numbers and densities across species (Figure 2), the IFS also differ (Supplementary Figure 1, Supplementary Table 2 & 3). Comparing IFSs between TE orders, we find a significant excess of high frequency RC insertions in *D. melanogaster* versus other species (GLM quasi-Binomial *p*-value < 3.5e-5, t-value > 4.151). We also find an excess of rare (low frequency) TIR insertions versus other species in *D. innubila* (*p*-value = 2.37e-5, t = -4.24) and *D. pseudoobscura* (*p*-value = 5.74e-15, t-value = -7.891). Additionally, we find a significant excess of high frequency LTR insertions in *D. ananassae* versus all other species (GLM *p*-value < 2e-16, t-value = 13.243) and an excess of higher frequency LINE insertions in both *D. melanogaster* (GLM *p*-value < 2e-16, t = 12.526) and *D. ananassae* (GLM p-value < 2e-16, t=11.505). While we find IFS differ between species, in all cases TEs are skewed towards rare insertions (Figure 1). The median insertion frequency is below 25% in every TE order across all species and shows no significant differences between species (Supplementary Table 2 & 3, GLM *p*-value > 0.213).

As these comparisons may be biased by factors such as how the data was generated, the sequencing methods, the quality of the reference genomes and the TE annotation, we limited our analysis to *D. melanogaster*, *D. ananassae* and *D. willistoni*, three species with data generated in similar manners, with similar TE families and high-quality reference genomes. We assessed only insertions in regions of the autosomal genome identified as orthologous using *progressiveMauve* (Darling *et al.* 2004). When comparing the insertions in these orthologous regions, for all comparisons we find the TE dynamics are more consistent between species, with no significant differences in any comparison (Supplementary Table 2, Figure 2, Supplementary Figure 1B: GLM *p*-value > 0.21, t-value < 1.556).

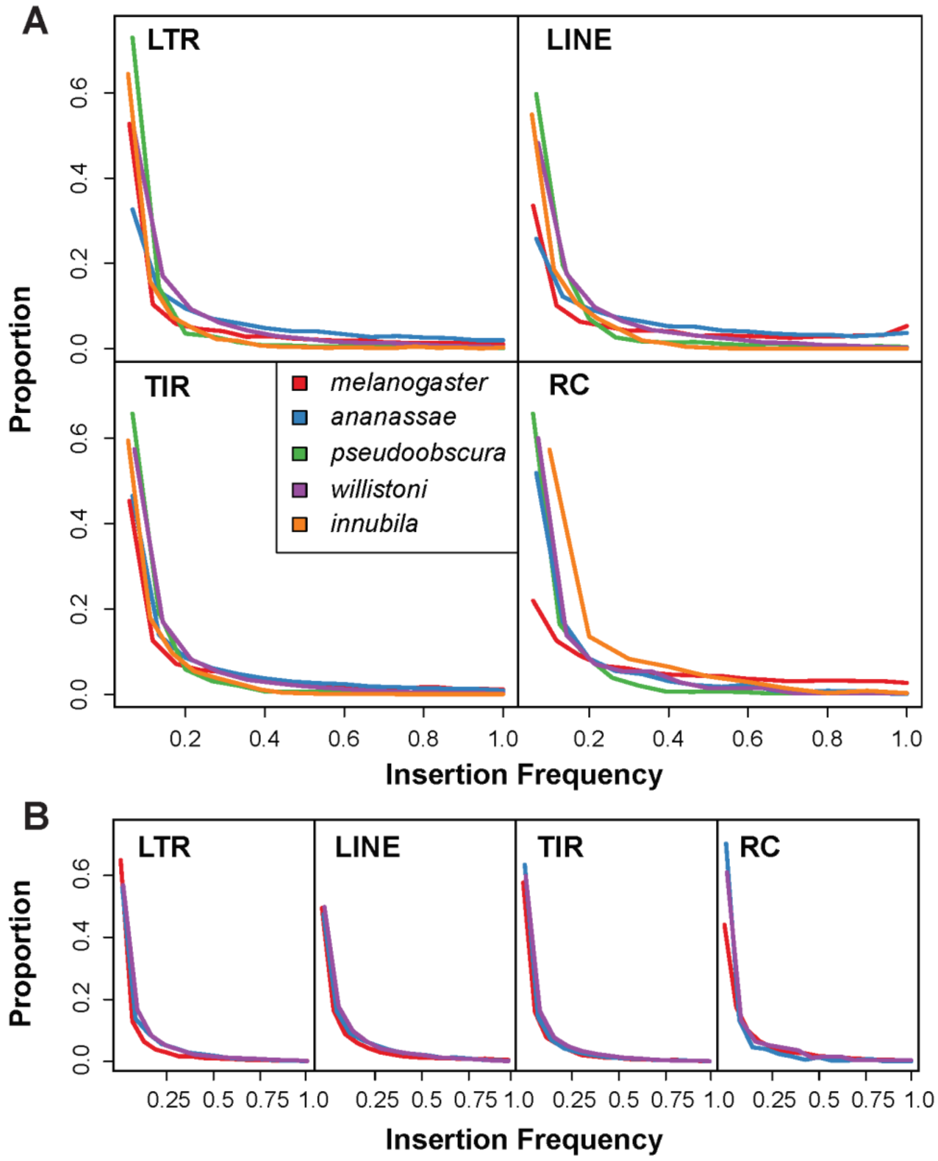
TE site frequency spectra rarely differ when accounting for population structure, insertions are primarily rare

One limitation of the analysis thus far is that all samples except *D. melanogaster* violate our implicit assumption of a single, panmictic population, which may skew the IFS to higher frequencies. This is can be seen in differences in estimated nucleotide site frequency spectrum of each species (limited to Muller element C for *D. pseudoobscura*), specifically finding an excess of high frequency variants in *D. pseudoobscura* when compared to *D. melanogaster* and an excess of low frequency variants in *D. willistoni* and *D. innubila* when compared to *D. melanogaster* (Supplementary Figure 2, GLM quasi-Binomial *p*-value < 0.05). As expected, all SFS show an excess of rare variants consistent with purifying selection, however *D. pseudoobscura* almost fits the neutral expectation, possibly due to the structured populations expected with the segregating inversions found on Muller element C (Dobzhansky and Sturtevant 1937; Dobzhansky and Epling 1948; Fuller *et al.* 2016).

To combat this, we clustered lines based on nuclear polymorphism using a principle component analysis (Supplementary Figure 3). We then took a subset of lines for each species which appears to cluster as a single group in a principle component analysis (Supplementary Figure 3). We also attempted to account for effective population size, on TE content, we find no association between effective population size and total TE content or insertion density, so did not control for this further (LM *p*-value > 0.05, Supplementary Figure 3).

In selected subpopulations, we checked for differences in the nuclear SFS between species and, with no drastic differences seen, we compared TE insertion SFSs between species. We find similar IFS across TE orders, though we do find an excess of high frequency RC insertions in *D. melanogaster* and an excess of high frequency LTR and LINE insertions in *D. ananassae* (Figure 3A, GLM *p*-value = 2e-16). Again, we find no significant differences when comparing orthologous regions (GLM *p*-value > 0.05). As previous, most TE insertions are rare in all species (median frequency < 20%), with *D. ananassae* and *D. melanogaster* having the highest median frequency insertion, we also find no significant difference between median insertion frequency for any species or TE order (GLM *p*-value > 0.352) and no association between TE density or genome size with median insertion frequency (*p*-value > 0.05).

Figure 3: Site frequency spectra for each species, separated by TE order for, A. TE insertions found across total genomes of all species. B. TE insertions called in orthologous regions for *D. melanogaster*, *D. willistoni* and *D. ananassae*.



Only a few, highly active, families differ across species, consistent with differing times of invasion

Our broader comparisons fit with previous work that suggests that most TEs are highly active across a broad species range due to recent acquisition of these TEs (Petrov *et al.* 2011; Kofler *et al.* 2015b), as opposed to other work that suggested TE activity differs between species and families (Hey 1989; Linheiro and Bergman 2012; Rahman *et al.* 2015). As these broad observations may homogenize large differences between TE families, we chose to focus our analysis on specific families, shared between species.

We repeated the previous analysis across 10 TE super families found in all species. While there is a noticeable excess of low frequency insertions in *D. pseudoobscura*, we found no significant difference of insertion frequency between species for TE super family frequency (GLM logistic regression: -1.351 < *t*-value < -0.092, *p*-value > 0.183), however this may be due to few TE insertions in each subgroup or could again be too broad for any real inference (Supplementary Figure 4).

Thus, we attempted to compare the dynamics of specific families shared between these species. We found 55 families shared between *D. melanogaster*, *D. ananassae* and *D. willistoni*, and found insertions within the previously identified orthologous regions. For each TE family we compared the site frequency spectrums for each species. Most these TE families showed no consistent significant differences in TE activity, with only 8 of the 55 TE families showing any significant differences (six after multiple testing correction, Supplementary Table 3-5, Supplementary Figure 5, GLM logistic regression: *p*-value < 0.05). For these elements, one species has an excess of low frequency variants compared to the other two species (Supplementary Figure 5), suggesting this difference may be due to a more recent acquisition than in this species, resulting in higher activity of the family, rather than a consistent difference in activity between species (Bergman and Bensasson 2007; Petrov *et al.* 2011; Kofler *et al.* 2012).

To test this, we calculated Tajima's D for each of the shared 55 TE families. A negative Tajima's D suggests an excess of low frequency variants, consistent with an expansion in copy number following a bottleneck, as would happen with a recent horizontal invasion (Tajima 1989; Bartolomé *et al.* 2009). Among the 55 shared families, we find ten TE families have significant differences in estimations of Tajima's D between species (GLM p-value < 0.05). Only one TE family overlaps with significantly negative Tajima's D and a difference in IFS between species, potentially explained by a more recent invasion of that TE family (Kofler *et al.* 2015a). *P*-element has a significantly different site frequency spectra between species (GLM logistic regression: *p*-value < 0.05), and significantly lower Tajima's D (GLM p-value < 0.05), due to its recent horizontal transfer to *D. melanogaster* from *D. willistoni* (Daniels *et al.* 1990; Khurana *et al.* 2011). Overall these results suggest few TE families differ between species in activity, after accounting for recent acquisitions.

Discussion

Transposable elements, as mobile parasitic elements, are mostly costly to a host organism (Charlesworth and Langley 1989), due to their rampant transposition, leading to the disruption of coding sequences (Charlesworth and Langley 1989; Charlesworth *et al.* 1997; Bachmann and Knust 2008), the misregulation of gene expression (McClintock 1953; Lisch and Bennetzen 2011; Lee 2015) and even because of ectopic recombination and chromosomal breakage between two copies of the same TE family (Charlesworth and Langley 1989; Montgomery *et al.* 1991; Sniegowski and Charlesworth 1994). Deleterious insertions are removed under purifying selection and TE families are rapidly silenced upon their acquisition (Langley *et al.* 1988; Montgomery *et al.* 1991; Lee and Langley 2012), giving an expectation for a site frequency spectrum skewed towards low frequency insertions for more recently active families (Langley *et al.* 1988; Charlesworth and Langley 1989; Montgomery *et al.* 1991; Charlesworth *et al.* 1997; Pasyukova *et al.* 2004). Most of the theoretical and experimental work that led to our understanding of TE dynamics has been completed in *D. melanogaster* (Charlesworth and Langley 1989; Charlesworth *et al.* 1997; Petrov *et al.* 2003), under the assumption that TEs in other *Drosophila* and insects behave in a similar manner, despite some evidence to the contrary (Hey 1989; Kaminker *et al.* 2002a; Bergman and Bensasson 2007). Here we test the validity of this assumption by assessing the TE dynamics in a *D. melanogaster* population and populations of four other increasingly diverged species. We find that, despite the drastic differences in TE content and densities between the species (Figure 2), we observe a pattern of rare insertions across all species, consistent with strong purifying selection against TE insertions in all species (Figure 3, Supplementary Figure 1, Supplementary Table 2 & 4), and the activity of similar families are also mostly conserved between species.

There are several possible explanations for the fact that work predating next generation sequencing technologies suggested differences in TE dynamics among species (Hey 1989). First, these differences may be due to host-specific factors (Supplementary Table 2 - 4, Supplementary Figure 1 & 4), such as how recent the TE family has been established in a species (Hey 1989; Kaminker *et al.* 2002b). Second, high copy number families identified by *In Situ* hybridisation may have be low resolution conflating separate insertions as the same insertion, artificially inflating that insertion's frequency and skewing its frequency higher than in lower copy number samples (Hey 1989). Finally, species genomes may differ in their chromatin states at different parts of genomes, limiting our analyses to well described euchromatic portions could have limited our ability to identify the diversity of TE dynamics in these host species. *D. ananassae*, for example, has an expansive Muller element F, full of transposable elements that was not included in this survey (due to most the chromosome being masked in the reference genome).

Overall, our results support a model where TE families invade of the genome, expand in copy number, are rapidly regulated by the host genome (to differing levels among species), with insertions primarily being deleterious in all species examined, though the selection against insertions appears to differ from species to species to a minor degree.

Materials and Methods

Population genomic data

We used next generation sequencing data from five species collected from three sources, summarized in Supplementary Table 1. For *Drosophila melanogaster*, we downloaded the FastQ files of 100bp paired end reads for a randomly selected set of 17 lines of the DPGP from a population collected from Zambia (SRA accessions: SRR203500-10, SRR204006-12). Similarly, we downloaded the FastQ files of 100bp paired end reads for 45 *Drosophila pseudoobscura* lines (SRA accessions: SRR617430-74). These lines consist of wild flies crossed to balancer stocks for chromosome 3 (Muller element C), this results in an isolated wild third chromosome, but a mosaic of balancer and wild stocks across the remainder of the genome, due to this we restricted our analysis to Muller element C (chromosome 3) in these lines.

We obtained sequencing information for 16 *Drosophila ananassae* isofemale lines and 14 *willistoni* isofemale lines. These lines were sequenced using an illumina HiSeq 2500 to produce 100bp paired end reads for each isofemale line.

Wild *Drosophila innubila* were captured at the Southwest Research Station in the Chiricahua Mountains between September 8th and 15th, 2016. Baits consisted of store-bought white button mushrooms (*Agaricus bisporus*) placed in large piles about 30cm in diameter. A sweep net was used to collect the flies over the baits. Flies were sorted by sex and species at the University of Arizona and males were frozen at -80 degrees C before being shipped on dry ice to Lawrence, KS. All *D. innubila* males were homogenized in 50 microliters of viral buffer (a media meant to preserve viral particles, taken from (Nanda *et al.* 2008)) and half of the homogenate was used to extract DNA using the Qiagen Gentra Puregene Tissue kit (#158689, Germantown, Maryland, USA). We constructed a genomic DNA library using a modified version of the Nextera DNA Library Prep kit (#FC-121-1031, Illumina, Inc., San Diego, CA, USA) meant to conserve reagents (Baym *et al.* 2015). We sequenced the library on two lanes of an Illumina HiSeq 2500 System Rapid-Run to generate paired-end 150 base-pair reads (available at NCBI accession numbers SRR6033015).

We trimmed all data using *Sickle* (minimum length = 50, minimum quality = 20) before mapping, and removed adapter sequences using *Scythe* (Joshi and Fass 2011; Buffalo 2018).

Custom reference genomes

We downloaded the latest *Flybase* reference genome (Flybase.org, as of December 2018) for *D. melanogaster*, *D. ananassae*, *D. pseudoobscura* and *D. willistoni*, and used the *D. innubila* reference genome available on NCBI (NCBI accession: SKCT00000000) (Hill *et al.* 2019).

For the released genomes (*D. melanogaster*, *D. ananassae*, *D. pseudoobscura* and *D. willistoni*), we identified and masked each reference genome using *RepeatMasker* (parameters: -pa 4 –s –gff –gccalc –nolow –norna –no_is) (Tarailo-Graovac and Chen 2009), using a custom repeat library, consisting of *Repbase* TE sequences previously identified in each of the species examined here (Kohany *et al.* 2006).

For. *D. innubila*, we generated a repeat library for the reference genome using *RepeatModeler* (parameters: - engine NCBI) (Smit and Hubley 2008). Then, after identifying each family order by NCBI universal *BLAST* (Altschul *et al.* 1990), used this library as the custom TE library for repeat masking as described above. To validate these *RepeatModeler* consensus sequences for *D. innubila,* we mapped Illumina data to the TE library and kept only TE sequences with at least 1x the genomic coverage across 80% of the sequence (BWA MEM, default parameters (Li and Durbin 2009; Li *et al.* 2009)).

For each species, we then generated a custom reference genome required for the use of *PopoolationTE2* (Kofler *et al.* 2016). For this we merged the masked reference genome, the custom TE library used for masking and the genome TE sequences, extracted using *BEDTools* (Quinlan and Hall 2010). Next, as described in the *PopoolationTE2* manual, we generated a hierarchy for each genome which assigned each TE sequence (all consensus sequences and reference sequences) to a TE family and TE order as described in (Kohany *et al.* 2006; Wicker *et* *al.* 2007), either terminal inverted repeat (TIR) and rolling circle (RC) DNA transposons, or long terminal repeat (LTR) and long interspersed nuclear element (LINE) RNA retrotransposons.

TE content and copy number differences between genomes

We quantified the amount of TE content for all species in three ways: a) proportion of the reference genome masked with *RepeatMasker*, b) median insertion count of each TE family across all lines in a species and c) median insertion count of each family using *PopoolationTE2*. For b), we found the median coverage for each TE family and the median coverage masked nuclear genome using *BEDTools* (genomeCoverageBed) (Quinlan and Hall 2010), we divided the median TE coverage by the median nuclear coverage (subsampled to 15x coverage) to find the copy number of each family. Then we calculated the median adjusted TE coverage across all lines for each species. For c), we calculated the median TE insertion count for each family in each species, based on TE insertions called using *PopoolationTE2*. To control for differences in genome size across euchromatic regions, we also calculated the insertions per 1 Megabase windows (sliding 250kbp) for each TE order in each line for each species, only for contigs greater than 100kbp with less than 60% of the window masked by *RepeatMasker* (Tarailo-Graovac and Chen 2009).

Calling transposable element insertions across genomes

To identify the TE insertions throughout the genome in each line for each species, we followed the recommended *PopoolationTE2* pipeline for each species (*sourceforge.net/p/popoolation-te2/wiki/Walkthrough/*) (Kofler *et al.* 2016). Though *PopoolationTE2* is designed for use with population pools, we used an adjusted method to call germline insertions in individuals. We subsampled each line to 15x average nuclear coverage and followed the pipeline with appropriate cutoffs to exclude most somatic transpositions (map-qual = 15, min-count = 5, min-distance = -200, max-distance = 500). *PopoolationTE2* gives an estimated frequency of the insertion based on coverage of the TE breakpoint versus the genomic coverage, here we used this as a support score for each TE insertion (Kofler *et al.* 2016). We removed insertions found exclusively in one line with lower than 50% frequency in an individual line, we then merged all remaining insertion files for each species. We also removed all insertions in regions with more than 60% of the Megabase window masked by *RepeatMasker* (Tarailo-Graovac and Chen 2009)*,* we also limited our analysis to scaffolds associated with autosomes in all species.

We used *BEDTools* (Quinlan and Hall 2010) to estimate the frequencies of each family's insertions across each species, combining TE insertions of the same family within 100bp of each other. We used a binomial GLM in R (Team 2013) to assess differences in insertion frequencies between species for each TE order, considering a significant effect of species compared to *D. melanogaster* for a p-value < 0.05 for each set of TE order insertion frequencies. If all species have a significant effect in a consistent direction, we consider this to be a significant effect of *D. melanogaster* on insertion frequency. We also compared the median insertion frequency across species and TE orders and again fit a GLM to compare in R (Team 2013).

For a less bias comparison of insertion frequency spectra, we limited our analyses to genomes with data generated in similar fashions (*D. melanogaster*, *D. ananassae, D. willistoni*), and to orthologous euchromatic regions of the genome. For this we used *progressiveMauve* to identify orthologous regions of each genome (Darling *et al.* 2004), then converted these regions into a bedfile and excluded regions below 100kb, with over 60% of bases masked. We excluded *D. innubila* from this comparison due to its high sequence divergence from all other species and difficulty in finding similar TE families in other species, and *D. pseudoobscura* as it only its Muller element C represented natural variation. We then extracted insertions found in the orthologous regions using *BEDTools* (Quinlan and Hall 2010) to compare insertion frequency spectra in orthologous regions.

Polymorphism and summary statistics across the host genome and TE sequences

We called polymorphism across the host nuclear genome using *GATK HaplotypeCaller* (DePristo *et al.* 2011) for each host and found the nuclear site frequency spectrum for each species using this data, which we confirmed using *ANGSD* (folded spectra, bootstraps = 100, reference sequence given, ancestral sequence not used) (Korneliussen *et al.* 2014). ANGSD was also used to perform a principle component analysis between samples in each species to look for population substructure (Korneliussen *et al.* 2014).

Estimating the effective population size of species

We used the previously generated folded site-frequency spectra from *ANGSD* in *StairwayPlot* for *D. melanogaster*, *D. innubila*, *D. ananassae* and *D. willistoni* (excluding *D. pseudoobscura* due to the method of the data generation) (Korneliussen *et al.* 2014; Liu and Fu 2015). For each estimated effective population size back in time, we found the harmonic mean of the effective size in the past 100,000 years and took that as the average size for the line. We then compared the TE copy number estimations to effective population size.

TE families with dynamics differing between species

We next wanted to identify TE families shared between species to identify differences in activity between species. We aligned families of the same superfamily (defined in the *Repbase* TE database (Kohany *et al.* 2006)) from each species using *MAFFT* and considered families within 95% identity to be the same family in different species (Katoh *et al.* 2002). We then compared the site frequency spectrum of these species using a logistic regression GLM. We also tested for differences in population genetic statistics to assess if differences are due to the recent acquisition of a family in a species. We calculated Watterson's theta, pairwise diversity and Tajima's D using *Popoolation* (Kofler *et al.* 2011a), then compared these statistics across family and species using a generalized linear model, noting significant interactions between species and TE family.

Abbreviations

TE = transposable element, TIR = terminal inverted repeat, LTR = long terminal repeat, LINE = long interspersed nuclear element, RC = rolling circle, GLM – generalized linear model, IFS = insertion frequency spectra.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Funding

This work was supported by a postdoctoral fellowship from the Max Kade foundation (Austria) and a K-INBRE postdoctoral grant (NIH Grant P20 GM103418) to TH.

Competing Interests

The author declares that they have no competing interests.

Authors' contributions

TH performed bioinformatics analysis, statistical analysis, wrote, read and approved the manuscript.

Data availability

D. pseudooscura data available on NCBI SRA: SRR617430- SRR617474. *D. melanogaster* data available on NCBI SRA: SRR203500-10, SRR204006-12. *D. anannasae, D. willistoni* and *D. innubila* data will be made available upon publication. *Drosophila* genomes can be downloaded from Flybase.org or NCBI. All data is available upon request before acceptance.

Acknowledgements

We are extremely grateful for the advice provided by R. Unckless and J. Blumenstiel, for providing the sequencing information, advice on analysis and the production of the manuscript. We are also grateful for helpful discussion provided by J. R. Chapman, A. J. Betancourt, C. Schlotterer, R. Kofler, and B. Charlesworth. Thanks for S. W. Schaeffer for providing the *D. pseudoobscura* data used in this survey and advice concerning how the data should be used.

Supplementary Table 1: Table of *Drosophila* strains used in this study, including information on species, collection location and SRA number.

Supplementary Table 2: Comparison of TE insertion frequencies between species and the fit of GLMs at different levels showing significant differences between species.

Supplementary Table 3: TE insertions across the analysed scaffolds for each of the five species analysed here, with TE family, superfamily, order and TE insertion site occupancy.

Supplementary Table 4: TEs showing significant differences in distributions between species and the median Tajima's D for each species to see if a recent horizontal acquisition was the cause of this difference. NA is given if the TE family is absent from the species in question.

Supplementary Table 5: Table of GLM results for differences in IFS between TE families shared across *D. ananassae*, *melanogaster* and *willistoni* in shared regions of the genome.

Figure S1. A. Insertion frequency spectrum, plots showing the densities of insertions and the proportion of the population these insertions are found in. These spectra are estimated using *PopoolationTE2* for each species*,* separated by TE order. **B.** Insertion frequency spectrum of TE insertions for regions with high similarity, identified using *progressiveMauve*.

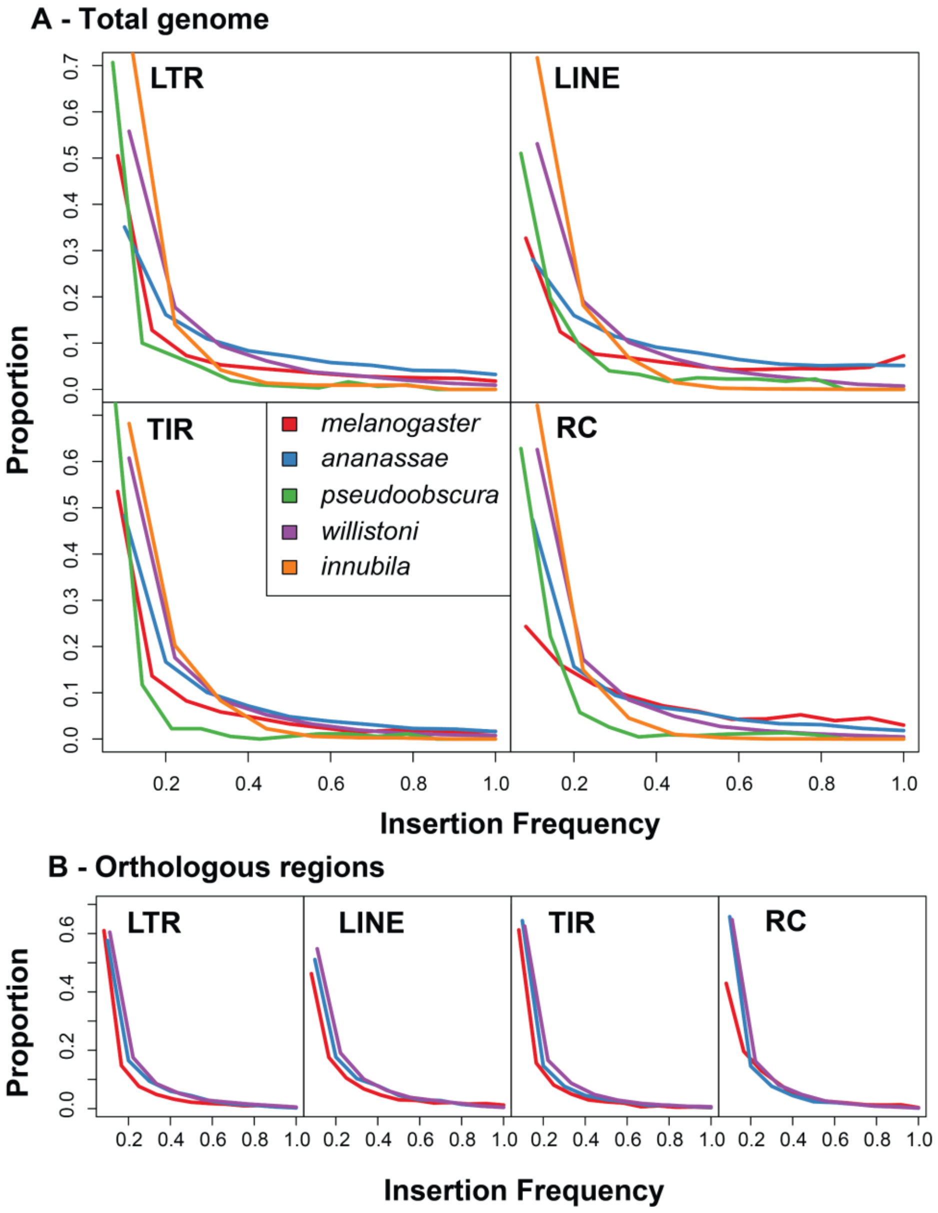


Figure S2: Site frequency spectra the nuclear genome of species analyzed here, calculated using ANGSD. The theoretical neutral site frequency spectrum is layered on top in red.

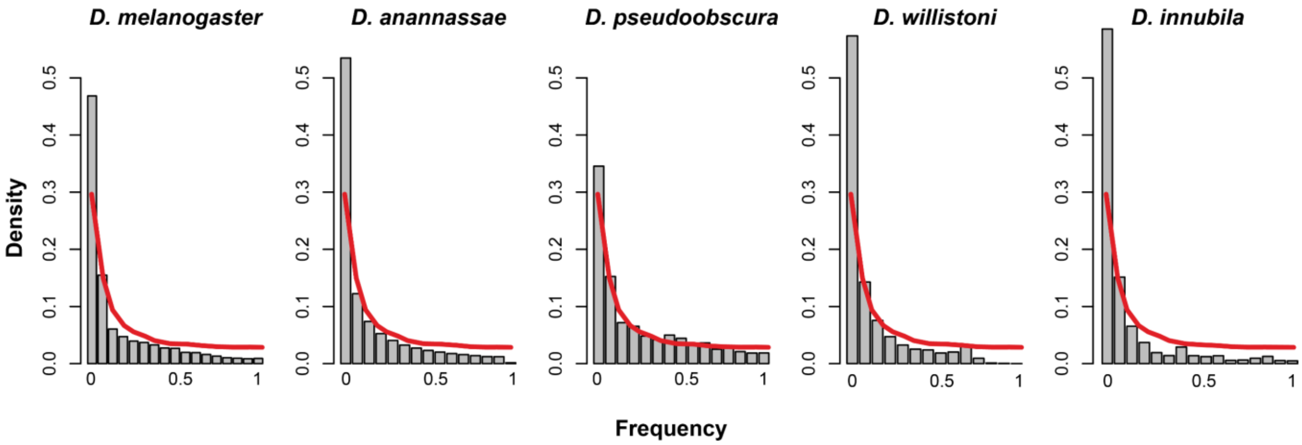


Figure S3: Principle component analysis for nuclear polymorphism for each species. Subpopulations are colored differently when known. E.G. Muller C inversion karyotype for *D. pseudoobscura* and Arizona sky island place of collection for *D. innubila* (both colored arbitrarily). Circled clusters are the lines used in the subset analysis, chosen arbitrarily based on the clustering seen in the PCAs. TE copy number for each species (+- 2 * standard deviations) is also compared to estimated effective population size from *StairwayPlot*.

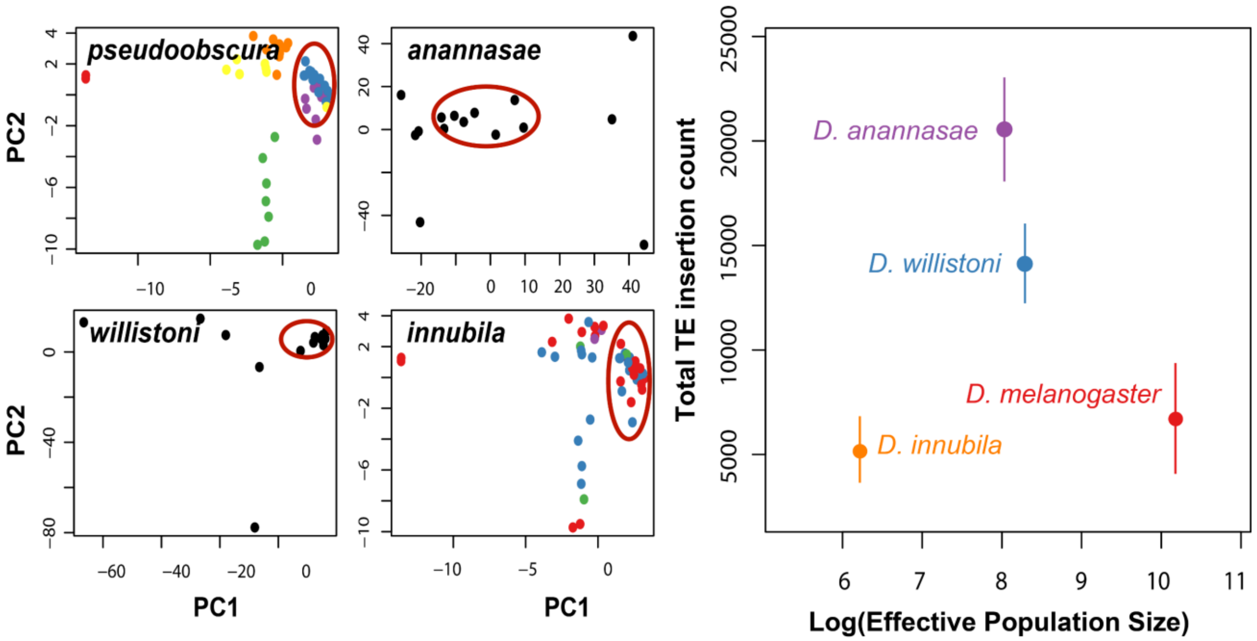
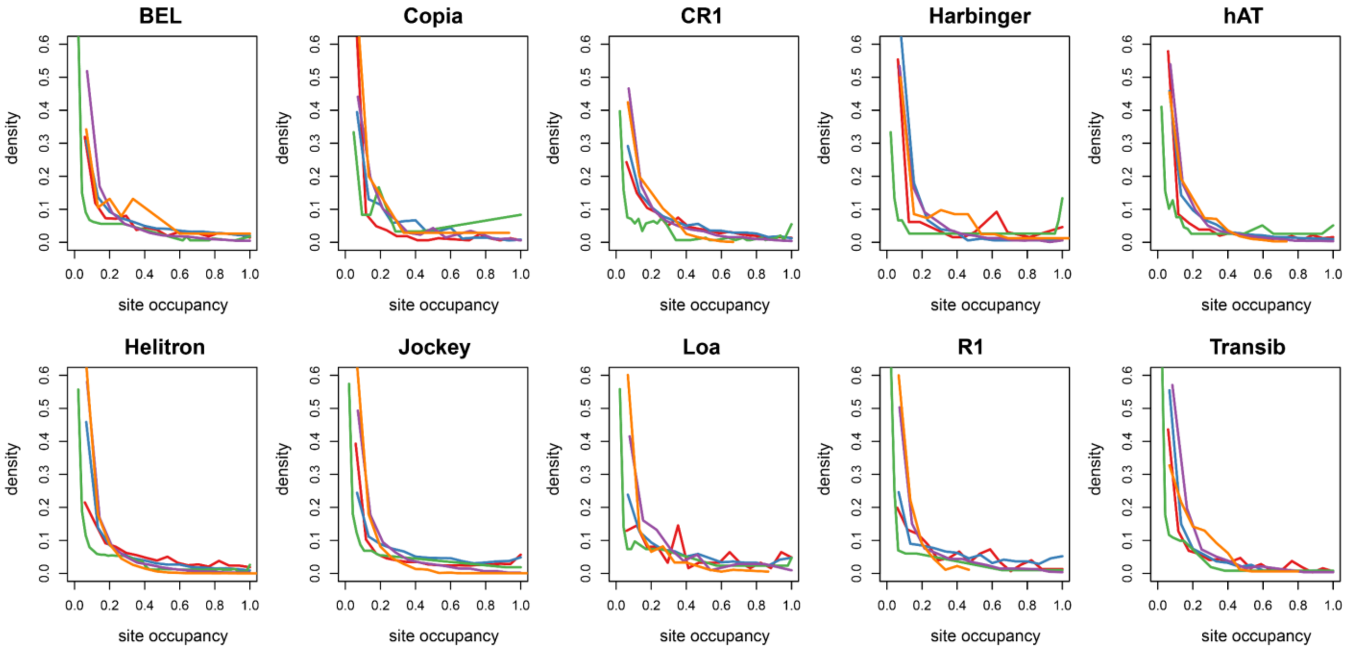
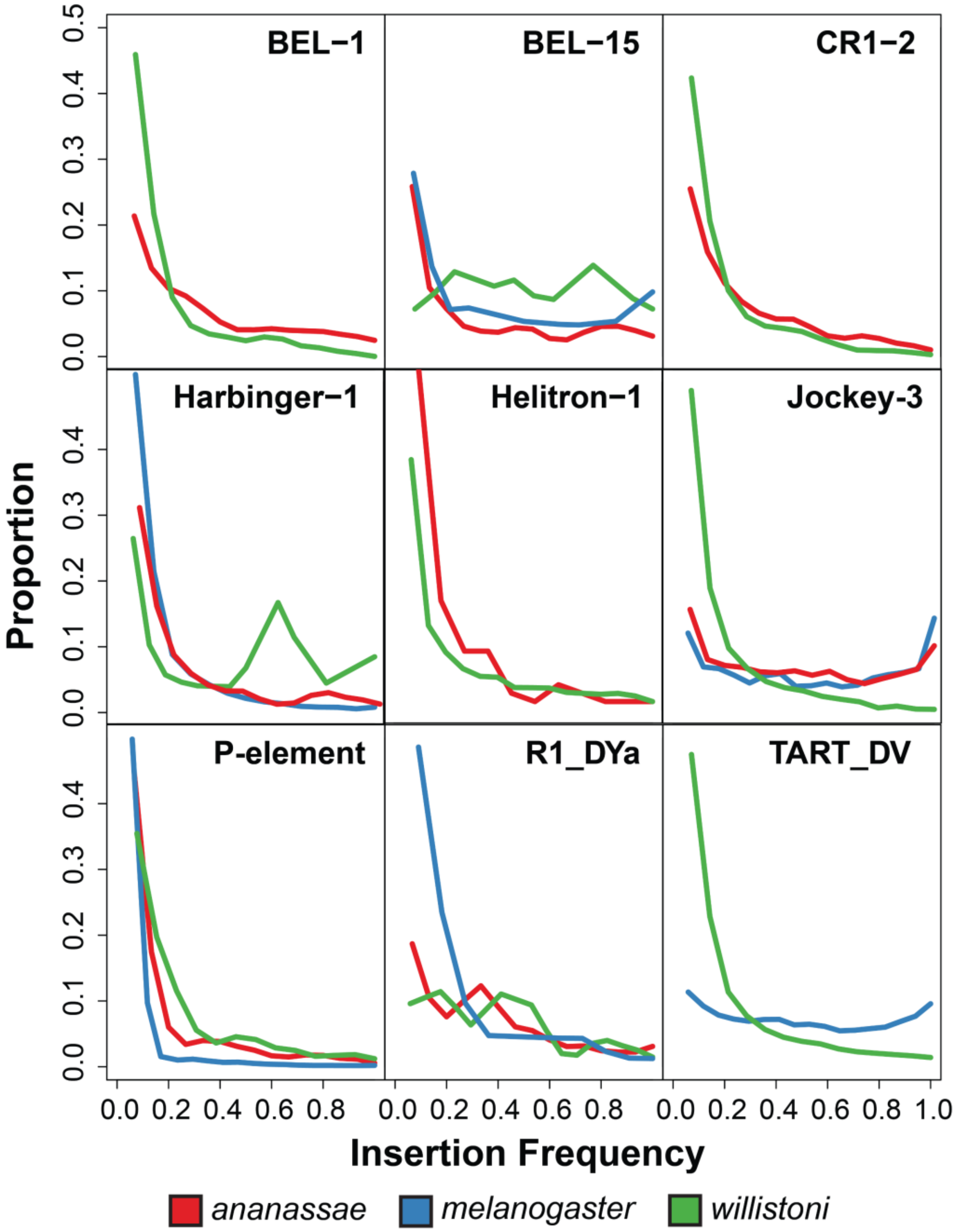


Figure S4: Insertion frequency per species for shared TE superfamilies.



464 **Figure S5:** Site frequency spectrum of TEs shared between species that are significantly different
465 in at least one comparison. Spectra are weighted by copy number. These are the 9 of 55
466 comparisons to show significant differences in distribution between species. The peak at ~60% in
467 Harbinger-1 in *D. willistoni* is caused by a small number of insertions at 60% frequency and low
468 insertion numbers found in the *D. willistoni*.



469

470

Bibliography

Altschul S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.

Aravin A. A., G. J. Hannon, and J. Brennecke, 2007 The piwi-piRNA pathway provides an adaptive defense in the transposon arms race. Science (80-.). 318: 761–764.

Bachmann A., and E. Knust, 2008 The use of P-element transposons to generate transgenic flies. Methods Mol. Biol. 420: 61–77.

Bartolomé C., X. Bello, and X. Maside, 2009 Widespread evidence for horizontal transfer of transposable elements across Drosophila genomes. Genome Biol. 10: R22.

Baym M., S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, *et al.*, 2015 Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One 1–15.

Bergman C. M., and D. Bensasson, 2007 Recent LTR retrotransposon insertion contrasts with waves of non-LTR insertion since speciation in Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. 104: 11340–11345.

Blumenstiel J. P., 2011 Evolutionary dynamics of transposable elements in a small RNA world. Trends Genet. 27: 23–31.

Brennecke J., A. A. Aravin, A. Stark, M. Dus, M. Kellis, *et al.*, 2007 Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila*. Cell 128: 1089–1103.

Brennecke J., C. D. Malone, A. A. Aravin, R. Sachidanandam, A. Stark, *et al.*, 2008 An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322:

Buffalo V., 2018 Scythe

Burt A., and R. Trivers, 2006 *Genes in Conflict*.

Capy P., T. Langin, D. Higuet, P. Maurer, and C. Bazin, 1997 Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100: 63–72.

Charlesworth B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. 23: 251–87.

Charlesworth B., C. H. Langley, and P. D. Sniegowski, 1997 Transposable element distributions in *Drosophila*. Genetics 147: 1993–5.

Clark A. G., M. B. Eisen, D. R. Smith, C. M. Bergman, B. Oliver, *et al.*, 2007 Evolution of genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.

Daniels S. B., K. R. Peterson, L. D. Strausbaugh, M. G. Kidwell, and A. Chovnick, 1990 Evidence for horizontal transmission of the P transposable element between *Drosophila* species. Genetics 124: 339–355.

Darling A. C. E., B. Mau, F. R. Blattner, and N. T. Perna, 2004 Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements 1394–1403.

DePristo M. A., E. Banks, R. Poplin, K. V Garimella, J. R. Maguire, *et al.*, 2011 A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43: 491–8.

Dobzhansky T., and A. H. Sturtevant, 1937 Inversions In Chromosomes of <i>Drosophila pseudoobscura. Genetics 23: 28–64.

Dobzhansky T., and C. Epling, 1948 The suppression of crossing over in inversion heterozygotes of Drosophila pseudoobscura. Proc. Natl. Acad. Sci. U. S. A. 34: 137–41.

Fuller Z. L., G. D. Haynes, S. Richards, and S. W. Schaeffer, 2016 Genomics of Natural Populations: How Differentially Expressed Genes Shape the Evolution of Chromosomal Inversions in. Genetics.

Gregory T. R., 2005 Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. 6: 699–708.

Gregory T. R., and J. S. Johnston, 2008 Genome size diversity in the family Drosophilidae. Heredity (Edinb). 101: 228–38.

Hellen E. H. B., and J. F. Y. Brookfield, 2013a The diversity of class II transposable elements in mammalian genomes has arisen from ancestral phylogenetic splits during ancient waves of proliferation through the genome. Mol. Biol. Evol. 30: 100–108.

Hellen E. H. B., and J. F. Y. Brookfield, 2013b Transposable element invasions. Mob. Genet. Elements 3: e23920.

Hey J., 1989 The transposable portion of the genome of Drosophila algonquin is very different from that in Drosophila melanogaster. Mol. Biol. Evol. 6: 66–79.

Hill T., B. Koseva, and R. L. Unckless, 2019 The genome of Drosophila innubila reveals lineage-specific patterns of selection in immune genes. Mol. Biol. Evol. in press: 1–29.

Joshi N., and J. Fass, 2011 Sickle: A sliding window, adaptive, quality-based trimming tool for fastQ files. 1.33.

Kaminker J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, *et al.*, 2002a The Transposable Elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 3: 0084.

Kaminker J. S., C. M. Bergman, B. Kronmiller, R. Svirskas, S. Patel, *et al.*, 2002b The transposable elements of the Drosophila melanogaster euchromatin : a genomics perspective. Genome Biol. 3: 1–20.

Katoh K., K. Misawa, K. Kuma, and T. Miyata, 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059–66.

Khurana J. S., J. Wang, J. Xu, B. S. Koppetsch, T. C. Thomson, *et al.*, 2011 Adaptation to P-element transposon invasion in *Drosophila melanogaster*. Cell 147: 1551–1563.

Kofler R., P. Orozco-terWengel, N. de Maio, R. V. Pandey, V. Nolte, *et al.*, 2011a Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6.

Kofler R., R. V. Pandey, and C. Schlötterer, 2011b PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27: 3435–3436.

Kofler R., A. J. Betancourt, and C. Schlötterer, 2012 Sequencing of pooled DNA Samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PloS Genet. 8: 1–16.

Kofler R., T. Hill, V. Nolte, A. J. Betancourt, and C. Schlötterer, 2015a The recent invasion of natural Drosophila simulans populations by the P-element. Proc. Natl. Acad. Sci. U. S. A. 112.

Kofler R., V. Nolte, and C. Schlötterer, 2015b Tempo and mode of transposable element activity in *Drosophila*. PLoS Genet 11: e1005406.

Kofler R., G. Daniel, and C. Schlötterer, 2016 PoPoolationTE2 : comparative population genomics of transposable elements using Pool-Seq. Mol. Biol. Evol. 1–12.

Kohany O., A. J. Gentles, L. Hankus, and J. Jurka, 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.

Korneliussen T. S., A. Albrechtsen, and R. Nielsen, 2014 ANGSD: Analysis of Next Generation Sequencing Data. BMC Bioinformatics 15: 356.

Langley C. H., E. Montgomery, R. Hudson, N. Kaplan, and B. Charlesworth, 1988 On the role of unequal exchange in the containment of transposable element copy number. Genet. Res. 52: 223–235.

Lee Y. C. G., and C. H. Langley, 2010 Transposable elements in natural populations of Drosophila melanogaster. Philos. Trans. R. Soc. B Biol. Sci. 365: 1219–1228.

Lee Y. C. G., and C. H. Langley, 2012 Long-term and short-term evolutionary impacts of transposable elements on *Drosophila*. Genetics 192: 1411–1432.

Lee Y. C. G., 2015 The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in Drosophila melanogaster. PLOS Genet. 11: 1–24.

Leung W., and P. Students, 2017 Retrotransposons Are the Major Contributors to the Expansion of the Drosophila ananassae Muller. G3 7: 2439–2460.

Levine M. T., and H. S. Malik, 2011 Learning to protect your genome on the fly. Cell 147: 1440–1441.

Levis R. W., R. Ganesan, K. Houtchens, L. A. Tolar, and F. Sheen, 1993 Transposons in place of telomeric repeats at a *Drosophila* telomere. Cell 75: 1083–1093.

Li H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–60.

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–9.

Linheiro R. S., and C. M. Bergman, 2012 Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. PLoS One 7: e30008.

Lisch D., and J. L. Bennetzen, 2011 Transposable element origins of epigenetic gene regulation. Curr. Opin. Plant Biol. 14: 156–161.

Liu X., and Y.-X. Fu, 2015 Exploring population size changes using SNP frequency spectra. Nat. Genet. 47: 555–559.

Lu J., and A. G. Clark, 2010 Population dynamics of PIWI-interacting RNAs (piRNAs) and their targets in *Drosophila*. Genome Res. 20: 212–227.

Markow T. A., and P. O'Grady, 2006 Drosophila*: a guide to species identification*.

McClintock B., 1953 Induction of instability at selected loci in Maize. Genetics 38: 579–599.

McGaugh S. E., C. S. S. Heil, B. Manzano-Winkler, L. Loewe, S. Goldstein, *et al.*, 2012 Recombination modulates how selection affects linked sites in *Drosophila*. PLoS Biol. 10: 1–17.

Montgomery E. A., S. Huang, C. H. Langley, and B. H. Judd, 1991 Chromosome rearrangement by ectopic recombination in Drosophila melanogaster: genome structure and evolution. Genetics 129: 1085–1098.

Nanda S., G. Jayan, F. Voulgaropoulou, A. M. Sierra-Honigmann, C. Uhlenhaut, *et al.*, 2008 Universal virus detection by degenerate-oligonucleotide primed polymerase chain reaction of purified viral nucleic acids. J. Virol. Methods 152: 18–24.

Obbard D. J., K. H. J. Gordon, A. H. Buck, and F. M. Jiggins, 2009 The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 364: 99–115.

Orgel L. E., and F. H. C. Crick, 1980 Selfish DNA: the ultimate parasite. Nature 284: 604–607.

Palmieri N., C. Kosiol, and C. Schlötterer, 2014 The life cycle of Drosophila orphan genes. Elife 3: 1–21.

Pardue M.-L., and P.G. DeBaryshe, 2003 Retrotransposons provide an evolutionarily robust non-telomerase mechanism to maintain telomeres. Annu. Rev. Genet. 37: 485–511.

Pasyukova E. G., S. V Nuzhdin, T. V Morozova, and T. F. C. Mackay, 2004 Accumulation of transposable elements in the genome of Drosophila melanogaster is associated with a decrease in fitness. J. Hered. 95: 284–90.

Petrov D. A., Y. T. Aminetzach, J. C. Davis, D. Bensasson, and A. E. Hirsh, 2003 Size matters: Non-LTR retrotransposable elements and ectopic recombination in *Drosophila*. Mol. Biol. Evol. 20: 880–892.

Petrov D. a, A.-S. Fiston-Lavier, M. Lipatov, K. Lenkov, and J. González, 2011 Population genomics of transposable elements in Drosophila melanogaster. Mol. Biol. Evol. 28: 1633–1644.

Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–2.

Rahman R., G.-W. Chirn, A. Kanodia, Y. A. Sytnikova, B. Brembs, *et al.*, 2015 Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. Nucleic Acids Res. 43: 10655–72.

Senti K. A., and J. Brennecke, 2010 The piRNA pathway: A fly's perspective on the guardian of the genome. Trends Genet. 26: 499–509.

Senti K. A., D. Jurczak, R. Sachidanandam, and J. Brennecke, 2015 piRNA-guided slicing of transposon transcripts enforces their transcriptional silencing via specifying the nuclear piRNA repertoire. Genes Dev. 29: 1747–1762.

Sessegolo C., N. Burlet, A. Haudry, C. Biémont, C. Vieira, *et al.*, 2016 Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol. Lett. 12: 521–524.

Smit A. F. A., and R. Hubley, 2008 RepeatModeler Open-1.0

Sniegowski P. D., and B. Charlesworth, 1994 Transposable element numbers in cosmopolitan inversions from a natural population of Drosophila melanogaster. Genetics 137: 815–827.

Tajima F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tarailo-Graovac M., and N. Chen, 2009 Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma.

Team R. C., 2013 R: A Language and Environment for Statistical Computing

Wicker T., F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. 8: 973–82.