

## **Synergistic epistasis of the deleterious effects of transposable elements**

Yuh Chwen G. Lee

Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA

Author for correspondence:

Grace Yuh Chwen Lee

Department of Ecology and Evolutionary Biology

University of California, Irvine, Irvine, CA

[grylee@uci.edu](mailto:grylee@uci.edu)

## Abstract

The replicative nature and generally deleterious effects of transposable elements (TEs) give rise to an outstanding question about how TE copy number is stably contained in host populations. Classic theoretical analyses predict that, when the decline in fitness due to each additional TE insertion is greater than linear, or when there is synergistic epistasis, selection against TEs can result in a stable equilibrium of TE copy number. While several mechanisms are predicted to yield synergistic deleterious effects of TEs, we lack empirical investigations of the presence of such epistatic interactions. Purifying selection with synergistic epistasis generates repulsion linkage between deleterious alleles and, accordingly, an underdispersed distribution for the number of deleterious mutations among individuals. We investigated this population genetic signal in an African *Drosophila melanogaster* population and found evidence for synergistic epistasis among TE insertions, especially those expected to have large fitness impacts. Curiously, even though ectopic recombination has long been predicted to generate nonlinear fitness decline with increased TE copy number, TEs predicted to suffer higher rates of ectopic recombination are not more likely to be underdispersed. On the other hand, underdispersed TE families are more likely to show signatures of deleterious epigenetic effects and stronger ping-pong signals of piRNA amplification, a hypothesized source from which synergism of TE-mediated epigenetic effects arises. Our findings set the stage for investigating the importance of epistatic interactions in the evolutionary dynamics of TEs.

## Introduction

Transposable elements (TEs) are genetic elements that copy themselves and move to new genomic locations (Wells and Feschotte 2020). Their replicative nature and generally harmful impacts on host functions (Langley *et al.* 1988; Montgomery *et al.* 1991; Maksakova *et al.* 2006; Hollister and Gaut 2009; Bellen *et al.* 2011; Rebollo *et al.* 2011; Robberecht *et al.* 2013; Lee 2015) make TEs commonly known as “genomic parasites.” To counteract the selfish replication of TEs, a process that depends on the transcription of TE sequences, various hosts have evolved mechanisms to transcriptionally or post-transcriptionally silence TEs (Yang *et al.* 2017 p.; Ozata *et al.* 2019; Deniz *et al.* 2019). In addition, TEs can be excised from the genome during transposition or through ectopic recombination among repeats within or between TE insertions (Devos *et al.* 2002; Lagemaat *et al.* 2005). While transcriptional and post-transcriptional silencing is expected to limit the selfish increase of TEs in host genomes, mutation accumulation experiments still found an appreciable rate of TE replication (transposition rate,  $10^{-5}$ - $10^{-4}$  per copy per genome per generation; (Nuzhdin and Mackay 1995; Maside *et al.* 2000; Pasyukova *et al.* 2004; Díaz-González *et al.* 2011; Adrion *et al.* 2017). Furthermore, this rate of TE increase is at least two orders of magnitude higher than estimated rates of TE excision (Nuzhdin and Mackay 1995; Maside *et al.* 2000; Pasyukova *et al.* 2004; Adrion *et al.* 2017), implying an appreciable net rate of TE increase. At the same time, many eukaryotic genomes only have limited TE abundance (e.g., less than 1% in honeybee (Wells and Feschotte 2020)). Together, these facts pose an outstanding question— how is TE copy number contained in host populations?

Selection against the deleterious fitness effects of TEs has been theoretically proposed as an answer to this puzzle, as it can be a potent evolutionary mechanism counterbalancing the selfish replication of TEs in natural populations (Charlesworth and Charlesworth 1983; Charlesworth and Langley 1989; Lee and Langley 2010; Barrón *et al.* 2014). Empirical investigations have supported the idea that most TE insertions are deleterious and removed from the populations by purifying selection. For example, a dearth of TEs in or near coding sequences is observed across taxa (Kaminker *et al.* 2002; Stuart *et al.* 2016; Laricchia *et al.* 2017). TEs also have frequency spectra that are highly skewed towards rare insertions (Nellåker *et al.* 2012; Cridland *et al.* 2013; Kofler *et al.* 2015; Quadrana *et al.* 2016; Laricchia *et al.* 2017), which is typical for deleterious mutations. Classic theoretical analyses suggest that when natural selection removing TEs cancels out TEs' selfish increase, TE copy number can reach a balance in host populations (Charlesworth and Charlesworth 1983). It was further predicted that, whether TE copy number is *stably* contained in host populations depends on the mode of epistatic interactions among deleterious TE insertions ((Charlesworth and Charlesworth 1983), reviewed in (Kelleher *et al.* 2020; Choi and Lee 2020)). Specifically, when every additional TE exacerbates host fitness with a larger effect, or synergistic epistasis among the deleterious fitness effects of TEs, it is possible to have an equilibrium TE copy number that is stable even with other forces perturbing TE evolutionary dynamics.

Synergism among the fitness effects of TEs has been predicted to arise through two mechanisms. For one, the illegitimate recombination between nonhomologous TE insertions, or ectopic recombination, generates highly deleterious chromosomal rearrangements (Davis *et al.* 1987; Kupiec and Petes 1988; Montgomery *et al.* 1991; Lim

and Simmons 1994; Mieczkowski *et al.* 2006). Empirical evidence suggests that selection against ectopic recombination between TEs is a critical force limiting the selfish increase of TEs in host populations (Langley *et al.* 1988; Montgomery *et al.* 1991; Petrov *et al.* 2003, 2011). Because ectopic recombination happens between two TE insertions, the frequency of the event and the resultant decline in host fitness would naturally depend on the square of TE copy number (Montgomery *et al.* 1987; Langley *et al.* 1988). In other words, each additional TE would incur a higher fitness cost, exhibiting synergistic epistasis. For another, TE-induced changes of local chromatin states are also predicted to give rise to synergistic fitness effects (Lee and Langley 2010; Lee 2015). Small-RNA directed enrichment of repressive epigenetic marks at euchromatic TEs has been identified as a near-universal mechanism to transcriptionally silence TEs in multicellular eukaryotes (Aravin *et al.* 2008; Sienski *et al.* 2012; Le Thomas *et al.* 2013; Marí-Ordóñez *et al.* 2013; McCue *et al.* 2015) (reviewed in (Czech *et al.* 2018; Deniz *et al.* 2019)). Interestingly, these repressive marks could spread to TE-adjacent genic sequences, influencing host functions and, accordingly, fitness (reviewed in (Choi and Lee 2020)). Small RNAs that initiate TE-transcriptional silencing are generated from TE transcripts either directly (e.g., in plants (Xie *et al.* 2004; Kasschau *et al.* 2007)) or indirectly (e.g., via feed-forward "ping-pong cycle" in animals (Gunawardane *et al.* 2007; Brennecke *et al.* 2007; Aravin *et al.* 2007)). Accordingly, the tendency of a TE being targeted and epigenetically silenced by small RNAs, and the consequential deleterious spreading of repressive epigenetic marks, is expected to grow quadratically with increased TE copy number, leading to synergism among the fitness impacts of TEs (Lee and Langley 2010; Choi and Lee 2020). Interestingly, due to the differences in molecular mechanisms, deleterious ectopic recombination and epigenetic

effects of TEs have different predictions about which types of TEs are more likely to exhibit synergistic fitness effects.

Although synergism among the harmful impacts of TEs has been long predicted to be an important theoretical requirement for the stable containment of TE copy number, empirical investigations for its presence and extent in natural populations are still lacking (reviewed in (Kelleher *et al.* 2020)). A direct test for the proposed synergistic fitness effects would come from associations between TE copy number and individual fitness. Even though there is an overall negative association between the copy number of a specific TE family and measurements of fitness components (Mackay 1989; Houle and Nuzhdin 2004; Pasyukova *et al.* 2004), inferring the underlying mode of epistatic interactions from these data is challenging. Fitness is multifaceted, and it is hard to identify *a priori* fitness components impacted by the synergistic effects of TEs. The mode of epistatic interactions may also depend on environmental conditions (Peters and Keightley 2000; Kishony and Leibler 2003; Killick *et al.* 2006), further complicating experimental approaches to infer epistatic fitness effects directly. And importantly, subtle effects on fitness (e.g., 1%) are challenging to experimentally measure, but are expected to strongly influence the population dynamics of TEs in nature. An orthogonal approach that does not rely on the direct measurement of individual fitness is therefore needed to investigate the predicted synergistic fitness effects of TEs.

To test the presence of synergistic epistasis among single-nucleotide variants, several methods that do not rely on direct measurements of fitness have been proposed. These methods infer the mode of epistasis from the nonrandom clustering of variants either within species (Sohail *et al.* 2017) or between species (Callahan *et al.* 2011). In

particular, (Sohail *et al.* 2017) used the correlation between allele frequencies at different sites, or linkage disequilibrium (LD), to demonstrate the presence of synergistic epistasis among loss-of-function single-nucleotide mutations in human and *Drosophila* populations. To test the predicted synergism among TE insertions, we applied this population genetic framework to TE presence/absence polymorphism in a *D. melanogaster* Zambian population (Lack *et al.* 2015), which was also the focal *Drosophila* population in (Sohail *et al.* 2017). This population inhabits the likely ancestral range of the species (Pool *et al.* 2012; Sprengelmeyer *et al.* 2020) and would less likely be influenced by recent demographic history, which could create LD between variants even in the absence of epistatic interactions (Ewens and Spielman 1995; Zavattari *et al.* 2000; Rogers 2014). Importantly, these sequenced *D. melanogaster* strains did not go through intensive inbreeding to establish homozygous lines and were sequenced as haploid embryos (Lack *et al.* 2015). Accordingly, TEs that incur large fitness effects would likely still be represented in the data. By developing a bootstrapping framework, we were able to test for the presence of the predicted synergistic epistasis among TEs and infer from which deleterious mechanisms such synergism likely arises.

## Materials and Methods

### Population genomic data

We used DPGP3 Zambian *D. melanogaster* strains that were sequenced with Illumina paired-end short reads (Lack *et al.* 2015). This dataset includes 197 genomes, and we excluded those that were excluded from (Sohail *et al.* 2017) due to an extreme number of SNPs detected (six genomes), a read length smaller than 100bp (four genomes), or being sequenced in two separate runs (six genomes). An additional eight genomes were removed due to too many missing TE calls (see below). In total, 173 genomes were included in our final analysis. A list of genomes included in the analysis could be found in **Table S1**.

### Identification of TEs

Raw reads of DPGP3 genomes were processed by Trim\_galore (“Babraham Bioinformatics - Trim Galore!”) to remove adaptors and low-quality sequences. We used TIDAL (Rahman *et al.* 2015) to identify TE insertions in these DPGP3 genomes with respect to Release 6 reference genome coordinates. All possible TE calls, irrespective of coverage ratio (an index for the confidence of a TE call in TIDAL) and from all genomes, were combined to generate a list of potential TE insertions. We excluded INE-1, a TE family that experienced an ancient burst of activities and whose copies are mostly fixed in *D. melanogaster* (Kapitonov and Jurka 2003; Singh and Petrov 2004). We also excluded TEs on the 4<sup>th</sup> chromosome, which is nearly entirely heterochromatic (Riddle and Elgin 2018). This yielded 39,084 potential TE insertion sites.

We used previously developed approaches in (Lee and Karpen 2017), which was based on (Cridland *et al.* 2013), to call the presence/absence of all TEs in the list of



potential insertion sites in DPGP3 genomes, including the genome in which the TE was identified as an insertion by TIDAL. Briefly, following (Lee and Karpen 2017), we aligned processed reads to Release 6 *D. melanogaster* reference genome using bwa with default parameters (Li and Durbin 2010). Sequences that aligned 500bp around identified TE breakpoints were parsed out using samtools (Li 2011) and assembled into contigs using Phrap (Ewing *et al.* 1998) following parameters in (Cridland *et al.* 2013). The assembled contigs were aligned to TE-masked reference genome using blastn (Camacho *et al.* 2009). A TE is identified as absent if a contig is aligned across the TE insertion site. If no contig spanned over the TE insertion site, contigs were blasted to a database of sequences that include canonical TEs and all TEs annotated in the reference genome (retrieved from Flybase). A TE is called present if there were blast hits to TEs and if a contig aligns to the right or left side of the TE insertion site but does not span across the insertion site. All other scenarios were deemed as missing data (i.e., presence/absence status cannot be determined). We excluded TE insertions that are called present, but the contigs aligned to multiple TE families and thus the family identity of the insertion could not be determined. We used this filtering criterion because an important aspect of our analysis relies on TE family identity (see below). In total, this procedure resulted in 25,998 possible polymorphic (presence/absence) TE insertions.

The TE insertion dataset was further filtered with the following criteria. The strong suppression of recombination in pericentromeric regions is by itself expected to generate LD among variants, independently of synergistic epistasis. Accordingly, we excluded TEs in the heterochromatic regions of the genomes (0.5 Mb inward of the epigenetic euchromatin/heterochromatin boundaries identified in (Riddle *et al.* 2011)). Polymorphic

inversions account for a large proportion of population structure (Corbett-Detig and Hartl 2012; Huang *et al.* 2014) and could also create LD among variants. We thus excluded TEs in inversions segregating in the DPGP3 genomes (Lack *et al.* 2015), using inversion breakpoints identified from (Corbett-Detig and Hartl 2012; Huang *et al.* 2014). TE insertions within 1kb to each other, are assigned to the same TE family, and have the same presence/absence calls among all individuals could be two separate TE insertions or one TE insertion that was called twice due to the uncertainty of TE breakpoint identifications. Because we could not distinguish these two possibilities, these 443 TEs were also removed. Following the DPGP3 recommendations, we masked genomic regions with residual heterozygosity, identical by descent, or cosmopolitan admixture (Lack *et al.* 2015). TEs in these regions are considered “missing data.” We then excluded eight genomes whose number of missing TE calls were outliers to other genomes (more than 4,000 missing TE calls, see **Table S1**). We further filter out TE insertions that are called missing data in more than 10% of the genomes or are monomorphic (have the same presence/absence calls among individuals). 11,527 polymorphic TEs passed these filtering. Following (Sohail *et al.* 2017), we further restricted our analysis to rare TEs that are present in equal or fewer than five individuals (11,396 TEs).

### **Identification of SNP variants**

We used genome assemblies of the same 173 strains (see above) from Drosophila Genome Nexus (Lack *et al.* 2015) (in Release 5 reference genome coordinates). We used Flybase annotation 6.07 (converted to Release 5 coordinates by Lifter (<https://genome.ucsc.edu>)) to parse out the coding sequence of the longest isoform and

then identified synonymous, nonsynonymous, and premature stop codon variants. We excluded genes whose annotation in the reference genome contain putative errors (premature stop codon, lacking canonical stop codon, or having a coding sequence length not multiple of three), following (Sohail *et al.* 2017). Multi-allelic variants (a site with more than two alleles), codons with more than two variants (and thus cannot be assigned as either nonsynonymous and synonymous variants), and SNPs with missing data were excluded from the analysis.

### **Estimation and statistical significance of $\sigma^2/V_a$**

For both TEs and SNPs, we restricted the analysis to variants with minor allele counts equal to or smaller than five because TEs/SNPs with allele counts higher than this are unlikely to have deleterious fitness effects. The mutational burden for each individual was estimated as the number of minor alleles of the specific type of variants considered in the genome (Sohail *et al.* 2017).  $\sigma^2$  is estimated as the variance of mutational burden across genomes. Additive genetic variance ( $V_a$ ) was estimated as  $\sum_i 2p_i(1 - p_i)$ , where  $p_i$  is the minor allele frequency of locus/TE insertion  $i$ .

To evaluate the significance of observed  $\sigma^2/V_a$  of TEs, we used bootstrapped synonymous variants to generate a null distribution for  $\sigma^2/V_a$  for the TE dataset. Specifically, we randomly sampled 1,000 sets of synonymous variants that match the TE dataset in three aspects: (1) the number of variants, (2) minor allele counts, and (3) missing data structure, which controls for the number of missing data of a variant with specific minor allele frequency (MAF) and the total number of missing data per genome.

Using the simulated empirical distribution of  $\sigma^2/V_a$ , we estimated the one-sided *p-value* for  $\sigma^2/V_a$  of TEs being smaller than the null expectation.

To estimate LD between per pair of TEs, we used PLINK (Purcell *et al.* 2007) to compute the pairwise correlation coefficient ( $r^2$ ) between all pairs of TEs. For TE insertions  $i$  and  $j$ , LD between them ( $D_{i,j}$ ) is computed as  $\sqrt{r_{i,j}^2 p_i(1-p_i)p_j(1-p_j)}$ , where  $p_i$  and  $p_j$  are the MAF of TE  $i$  and  $j$ . By assuming that TE presence is the derived state, the sign of  $D_{i,j}$  depends on the coupling of TE present alleles, with  $D_{i,j} > 0$  if TE present alleles are on the same haplotype and  $D_{i,j} < 0$  for the opposite situation. We then binned pairs of TEs according to their physical distance on the same chromosome or as TEs on different chromosomes (see below), and calculated the mean LD per pair of TEs.

### **TE insertions and TE family annotations**

To test the predictions that TEs exerting large fitness effects are more likely to show synergistic epistasis, we categorized TEs according to their insertion locations, essentialities of the nearest gene (evolutionary constraints and mutant phenotypes), and local recombination rates. Using Flybase annotation 6.07 and bedtools (Quinlan and Hall 2010), we identified TEs located within exons, UTRs, and introns, and inferred their distance to the nearest gene. To categorize TEs according to evolutionary constraints of nearest genes, we estimated  $dN/dS$  ratios along the *D. melanogaster* lineage using maximum likelihood methods implemented in PAML (v4.9 (Yang 2007)) with alleles from *D. melanogaster*, *D. simulans* (Hu *et al.* 2013) and *D. yakuba* (Clark *et al.* 2007). Genes with fewer than 100 codons or with  $dS < 0.0001$  were treated as missing data. Genes with  $dN/dS$  estimates were binned into four categories according quartiles of  $dN/dS$  estimates: [0,

0.0341), [0.0341, 0.0877), [0.0877, 0.1932) and [0.1932, 15.28). To identify genes with essential functions, we used mutant phenotypes identified by either genetic disruptions or RNAi-mediated expression knockdown (downloaded from Flybase 08/22/2018). We focused on three categories related to survival: "lethal," "semi-lethal," or "viable." For genes that have different reported effects on survival, we chose the most severe phenotype. Local recombination rates around TE insertions were interpolated from the estimates of (Comeron *et al.* 2012). We categorized TEs into four bins according to quartiles of local recombination rates (cM/Mbp): [0, 1.344), [1.344, 2.354), [2.354, 3.64), and [3.64, 14.58).

For our analysis that estimated  $\sigma^2/V_a$  of individual TE families, we compared biological attributes of TE families with and without evidence of synergistic epistasis—specifically, their copy number, length, and sequence similarity. TE family copy numbers were estimated from TEs in the reference genome, excluding those in the heterochromatic regions (see above), and from our TE dataset. The mean length of a TE family was estimated by averaging the length of euchromatic copies of the same TE family in the reference genome. To estimate average pairwise sequence difference, we aligned euchromatic TE insertions of the same TE family in the reference genome using MUSCLE (Edgar 2004), calculated the percentage of pairwise difference (excluding gaps), and averaged that over all pairwise comparisons. TEs shorter than 100bp were excluded from the estimation. We also compared TE families for their propensity to be targeted by piRNAs and to exert epigenetic effects. For piRNA-related indexes, we used the estimated amount of sense and anti-sense piRNAs and ping-pong fraction (the proportion of piRNAs generated by the ping-pong cycle) of two wildtype genotypes (*w1118* and *wK*) from (Kelleher and Barbash 2013). For indexes of TE-mediated epigenetic effects of individual

TE family (proportion of TEs with the effect, the extent and magnitude of the effect), we used estimates from (Lee and Karpen 2017).

## Results

To investigate the mode of epistatic interactions among TEs using the population genetic framework developed by (Sohail *et al.* 2017), we first identified possible TE insertion positions in the Zambian genomes and then determined the presence/absence status of these TEs in individual genomes. After series of filtering steps to remove TEs with ambiguous family identity or presence/absence status, we identified 11,527 polymorphic TEs in the euchromatic regions of the genome (see Materials and Methods). Consistent with strong selection acting against TE insertions, they had a frequency spectrum that is highly skewed towards rare variants (**Figure 1A**). This frequency spectrum for TE insertions is more skewed than that of SNPs in the same genomes, even those resulting in highly deleterious premature stop codons (Lee and Reinhardt 2012) (**Figure 1A**). Thus, despite few cases of adaptive TEs (e.g., (Daborn *et al.* 2002; Schmidt *et al.* 2010; Hof *et al.* 2016), reviewed in (González and Petrov 2009)), the majority of TE insertions in *Drosophila* appear to be deleterious and are strongly selected against (reviewed in (Charlesworth and Langley 1989; Barrón *et al.* 2014)).

### Approach for inferring synergistic epistasis among TE insertions

In the absence of epistatic interactions, each mutation decreases individual fitness to the same extent, and selection acts on each mutation independently. Under this circumstance, the variance of mutational burden ( $\sigma^2$ ), which could be approximated by the number of deleterious mutations in a genome, would equal the sum of genetic variance across all loci ( $V_A$ ) (Sohail *et al.* 2017). In contrast, with epistasis, there is the interdependency of the fitness effects of mutations, and purifying selection removing them results in LD between

alleles (Lewontin and Kojima 1960; Eshel and Feldman 1970; Barton 1995). In particular, selection with synergistic epistasis creates repulsion, or negative, LD among deleterious alleles. The mutational burden will thus have an underdispersed distribution, or smaller variance than would be expected in the absence of epistatic interactions (Charlesworth 1990; Kondrashov 1995).

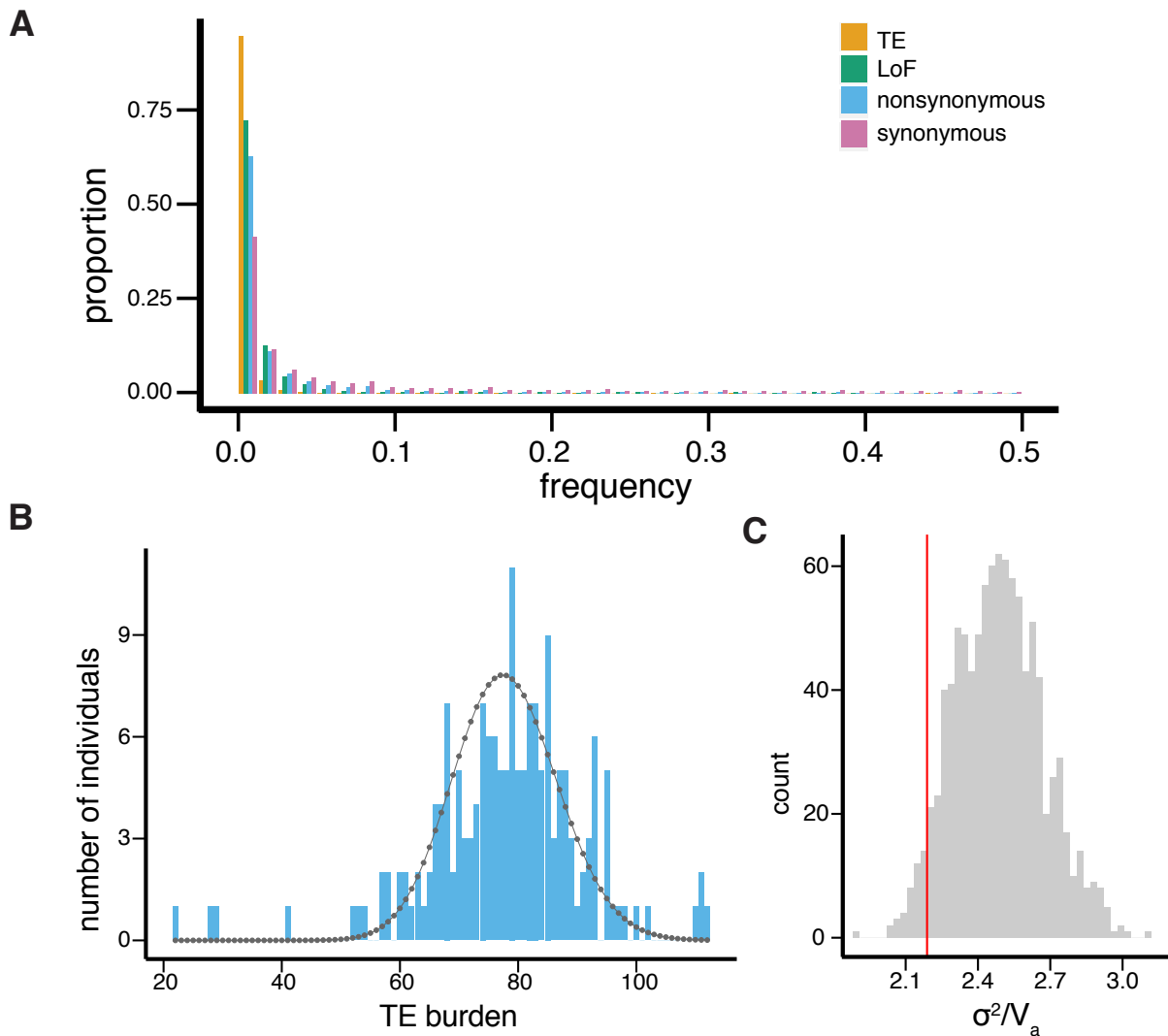
We estimated “TE burden” as the number of rare TEs in the individual genome (see Materials and Methods). In the absence of other factors that impact the distribution of mutational burden, a reduced variance of TE burden when compared to additive genetic variance ( $\sigma^2/V_a < 1$ ) would support synergistic fitness effects of TEs. Yet, even for synonymous variants, which are putatively neutral and should show no epistatic interactions, we found an overdispersed distribution ( $\sigma^2/V_a = 7.13$ ). This is similar to previously observed overdispersion of synonymous mutational burden using the same population genomes (Sohail *et al.* 2017), and could result from LD generated by an unknown demographic history of the population or other yet to be identified sources. Further, in our dataset, a large number of TEs are annotated as missing data in at least one genome (99.61%, see Materials and Methods). Missing data is expected to inflate the variance of estimated mutational burden (Sohail *et al.* 2017) and lead to a spurious inference of overdispersion (or  $\sigma^2/V_a > 1$ ). Indeed, by randomly masking alleles of synonymous variants, we found a significantly elevated ratio of  $\sigma^2/V_a$  (**Figure S1**). However, excluding TEs with any missing data, which was implemented in previous studies focusing on SNPs (Sohail *et al.* 2017), would reduce the number of polymorphic TEs to only 44 insertions.



Because of these additional factors that could affect the estimated distribution of TE burden, we compared the distribution of TE burden to that of putatively neutral sets of control loci, or synonymous variants. We modified the previously proposed bootstrapping approach (Sohail *et al.* 2017) and randomly sampled sets of synonymous variants that have a matching number of variants, MAF, and missing data structure as the TE dataset. We used these random sets of synonymous variants to generate an empirical null distribution of  $\sigma^2/V_a$  and estimated the *p-value* for the observed  $\sigma^2/V_a$  of TEs being smaller than that of synonymous variants, an approach that assumes that the same factors influence both synonymous variants and TEs.

**Figure 1. TEs have a skewed frequency spectrum and underdispersed distribution.**

(A) Frequency spectra of all the TEs that passed filtering and other SNP variants [loss-of-function (LoF), nonsynonymous, and synonymous]. (B) The distribution of TE burden (of TEs with  $MAF \leq 5$ ) among 173 genomes. Poisson distribution with identical mean is shown as black dots and line. (C) Resampling distribution of  $\sigma^2/V_a$  for synonymous variants that match number, MAF, and missing data structure as the TE data set is shown in gray bars while the red line shows the observed  $\sigma^2/V_a$  of TEs with  $MAF \leq 5$ .



## **TE burden for insertions that likely have large fitness effects has underdispersed distributions**

For all euchromatic TE insertions, we found an overall overdispersed distribution ( $\sigma^2/V_a = 2.23$ ), and there is an excess of individuals with too large or too small TE burden when compared to Poisson distribution with the same mean of TE burden (**Figure 1B**). Yet, there is also a larger than expected number of individuals with intermediate numbers of TEs (**Figure 1B**). The especially heavy left tail of individuals with too small TE burden could be a result of the abundant missing data for TE insertion (see above). However, despite this overdispersion, TE insertions have significantly lower  $\sigma^2/V_a$  than synonymous controls ( $p$ -value = 0.043, **Figure 1C**). This observation suggests that TE burden may be underdispersed in the absence of population structure, missing data, and/or other unknown factors inflating the variance. We further partitioned TEs according to TE class and type and estimated  $\sigma^2/V_a$  for each class (Table 1). Among them, the mutational burden of non-LTR (or LINE), retrotransposons that lack long-terminal repeats, has an underdispersed distribution that is significant compared to randomly sampled synonymous variants ( $\sigma^2/V_a = 0.98$ ,  $p$ -value = 0.048; Table 1, **Figure S2**).

Strong purifying selection is expected to weaken the overdispersion of mutational burden generated by population structure (Sohail *et al.* 2017). Signals of underdispersion, if present, should more likely be identified with TEs that exert strong deleterious fitness impacts. We thus categorized TEs according to their potential fitness impacts and examined their distribution separately. By classifying TEs according to their insertion locations, we indeed found that TEs inside coding sequences ( $\sigma^2/V_a = 0.97$ ,  $p$ -value = 0.49), UTRs ( $\sigma^2/V_a = 1.46$ ,  $p$ -value = 0.043) or exons ( $\sigma^2/V_a = 1.47$ ,  $p$ -value = 0.013) likely have

underdispersed distributions (**Table 1** and **Figure S3**). This is consistent with the expectation that TEs inserting into coding sequences or UTRs could abolish gene function (Bellen *et al.* 2004, 2011). We also categorized TEs according to the evolutionary constraints of their nearest gene. Genes that have low ratios between nonsynonymous to synonymous substitution rates, or  $dN/dS$  ratio, are highly conserved and generally expected to have essential functions (Larracunte *et al.* 2008; Waterhouse *et al.* 2011). TEs in or near these genes could potentially result in higher fitness costs. Consistently, we found that mutational burden of TEs whose nearest genes have the lowest quartile of  $dN/dS$  ratio underdisperses ( $\sigma^2/V_a = 0.97$ ,  $p$ -value = 0.098; **Table 1** and **Figure S4**). Similarly, TEs whose nearest genes have known lethal mutant phenotypes have significantly smaller  $\sigma^2/V_a$  than randomly sampled synonymous variants ( $\sigma^2/V_a = 1.31$ ,  $p$ -value = 0.045; **Table 1** and **Figure S5**). If restricting to insertions within exons, there is a significantly underdispersed distribution for TEs near genes that have the second-lowest quartile of  $dN/dS$  ratio ( $\sigma^2/V_a = 0.89$ ,  $p$ -value = 0.028; **Table 1** and **Figure S6**). TE burden of exonic TEs in genes with semi-lethal phenotypes also underdisperses ( $\sigma^2/V_a = 0.99$ ,  $p$ -value = 0.385; **Table 1** and **Figure S6**).

In addition to TEs inserting into and disrupting exonic sequences, intergenic TEs could impair host fitness through two mechanisms that were predicted to result in synergistic fitness effects of TEs. The illegitimate recombination between nonhomologous TEs could generate highly deleterious chromosomal rearrangements, irrespective of whether TEs insert inside or outside genes (Davis *et al.* 1987; Kupiec and Petes 1988; Montgomery *et al.* 1991; Lim and Simmons 1994; Mieczkowski *et al.* 2006). Assuming that the rates of ectopic recombination closely follow that of homologous recombination

(Lichten *et al.* 1987), TEs in high recombining regions of the genomes should be prone to be involved in ectopic recombination. Yet, we did not find TEs in high recombining regions of the genome having an underdispersed distribution (**Table 1**), which fails to support the prediction of the ectopic recombination model. On the other hand, epigenetically silenced intergenic TEs could result in the spreading of repressive epigenetic marks to adjacent sequences, which disrupts gene functions ((Rebollo *et al.* 2011; Lee 2015), reviewed in (Choi and Lee 2020)). Such an effect is likely to be restricted to intergenic TEs that are near essential genes. Supporting this prediction, there is a significantly underdispersed distribution of intergenic TEs near genes with known lethal mutant phenotypes ( $\sigma^2/V_a = 0.87$ ,  $p$ -value = 0.034, **Table 1** and **Figure S6**). It is worth noting that the deleterious epigenetic effects of intergenic TEs could not be entirely disentangled from TE-mediated genetic disruption of regulatory sequences (Kelleher *et al.* 2020; Choi and Lee 2020). Overall, we found that the distribution of mutational burden for TEs that are expected to have large fitness impacts have underdispersed distributions, especially for TEs inside exons or near essential genes.

### **TE burden of many families is underdispersed**

Both ectopic recombination and epigenetic effects of TEs depend on sequence homology among TE insertions. Mainly copies of the same TE family recombine, and small RNAs generated from a particular TE family are mostly effective on insertions of the very same TE family. Accordingly, both models predict that the synergistic epistasis would arise among insertions of *the same TE family* (Montgomery *et al.* 1987; Langley *et al.* 1988; Lee and Langley 2010; Lee 2015). We thus estimated  $\sigma^2/V_a$  for individual TE families,

restricting to those that have at least 20 insertions (46 of 86, or 53.49%, annotated TE families, **Table S1**). Twenty-two out of 46 TE families (47.83%) have  $\sigma^2/V_a < 1$  (**Table S2**), and that of *Jockey*, an abundant LINE TE family, is significantly reduced when compared to randomly sampled synonymous variants ( $\sigma^2/V_a = 0.81$ ,  $p$ -value = 0.003, **Table S2** and **Figure S6**). Restricting to TEs that are in or near essential genes found an even higher percentage of TE families with  $\sigma^2/V_a < 1$  (lowest quartile of  $dN/dS$  ratio: 64.29% of 28 families; with known lethal phenotype: 58.82% of 34 families; with inviable phenotype: 54.29% of 35 families). Among these families, *Roo*, the most abundant TE family in *D. melanogaster* (Kaminker *et al.* 2002; Kofler *et al.* 2015), and *Jockey* have significantly underdispersed distributions when compared to synonymous controls (**Table S2** and **Figure S6**).

To infer the possible source of synergism among TEs, we compared different attributes of TE families with  $\sigma^2/V_a$  smaller or greater than 1. TE families with  $\sigma^2/V_a < 1$  should be enriched with those that have synergistic epistasis among copies and thus underdispersed distributions. Factors that may confound  $\sigma^2/V_a$  estimates are unlikely to have differential influence across TE families, making this comparison between groups of TE families less likely biased.

Ectopic recombination and epigenetic effect models share some predictions about which TE families are more likely to exhibit synergistic fitness effects. Both models predict that abundant TE families would elicit higher fitness costs per TE copy (Langley *et al.* 1988; Lee and Langley 2010). Because both mechanisms depend on sequence homology, TEs that are long or have high sequence identity with other copies would represent larger targets for both ectopic recombination and small-RNA mediated epigenetic silencing. Accordingly,

it is predicted that TE families that are longer in length or have higher sequence identities among copies should more likely to exert synergistic epistasis (Lee and Langley 2010).

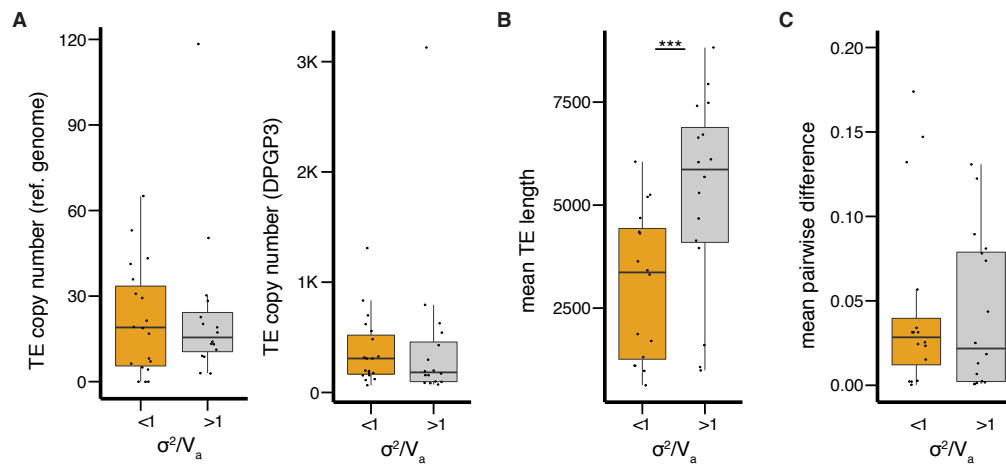
We investigated whether TE families with  $\sigma^2/V_a < 1$  have larger copy numbers, longer lengths, and higher within-family sequence identities. We used two estimates of euchromatic TE copy numbers: from our data, which is representative of natural populations, and from the reference genome annotation (Kaminker *et al.* 2002; Hoskins *et al.* 2015), which is comprehensive. Because we were unable to assemble internal sequences of TEs with the short-read Illumina data of the focused population, we used annotated euchromatic TEs in the reference genome to estimate the average insertion length and sequence divergence of TEs (see Materials and Methods). We found no significant associations between both estimates of TE copy number and whether a TE family has  $\sigma^2/V_a$  smaller than 1 (*Mann-Whitney U test*,  $p$ -value > 0.05 for all comparisons, **Table 2** and **Figure 2A**). Similarly, the sequence divergence between TE families with  $\sigma^2/V_a < 1$  or > 1 are not significantly different, irrespective of whether we restricted the analysis to insertions in or near essential genes (*Mann-Whitney U test*,  $p$ -value > 0.05 for all comparisons, **Table 2** and **Figure 2B**). On the other hand, the length of TE families with  $\sigma^2/V_a < 1$  are significantly *shorter* than other TE families when restricting the analysis to TE insertions in or near essential genes (*Mann-Whitney U test*,  $p$ -value = 0.0095 (lethal phenotypes) and 0.0075 (inviable phenotypes), **Table 2** and **Figure 2C**). The same unexpected pattern held when the analysis used canonical, instead of average, TE length (*Mann-Whitney U test*,  $p$ -value = 0.021 (lethal phenotypes), 0.0056 (inviable phenotypes), **Table 2** and **Figure 2C**). Overall, we did not find that TE families with  $\sigma^2/V_a < 1$  are more

abundant, longer in length, or with higher sequence identifies than other TE families,  
defying shared predictions of the ectopic recombination and epigenetic effect models.



## Figure 2. Comparing biological attributes between TE families with $\sigma^2/V_a < 1$ or $> 1$

Boxplots for (A) TE copy number (in the reference genome or in our current dataset), (B) average and canonical TE length, and (C) mean pairwise difference for TE families with and without an underdispersed distribution ( $\sigma^2/V_a < 1$ ) when restricting the analysis to TEs near/in genes with known inviable mutant phenotypes. Also see Table 2. \*\*\**Mann-Whitney test p-value* < 0.001.



## **TE families with stronger epigenetic effects are more likely to have underdispersed distribution**

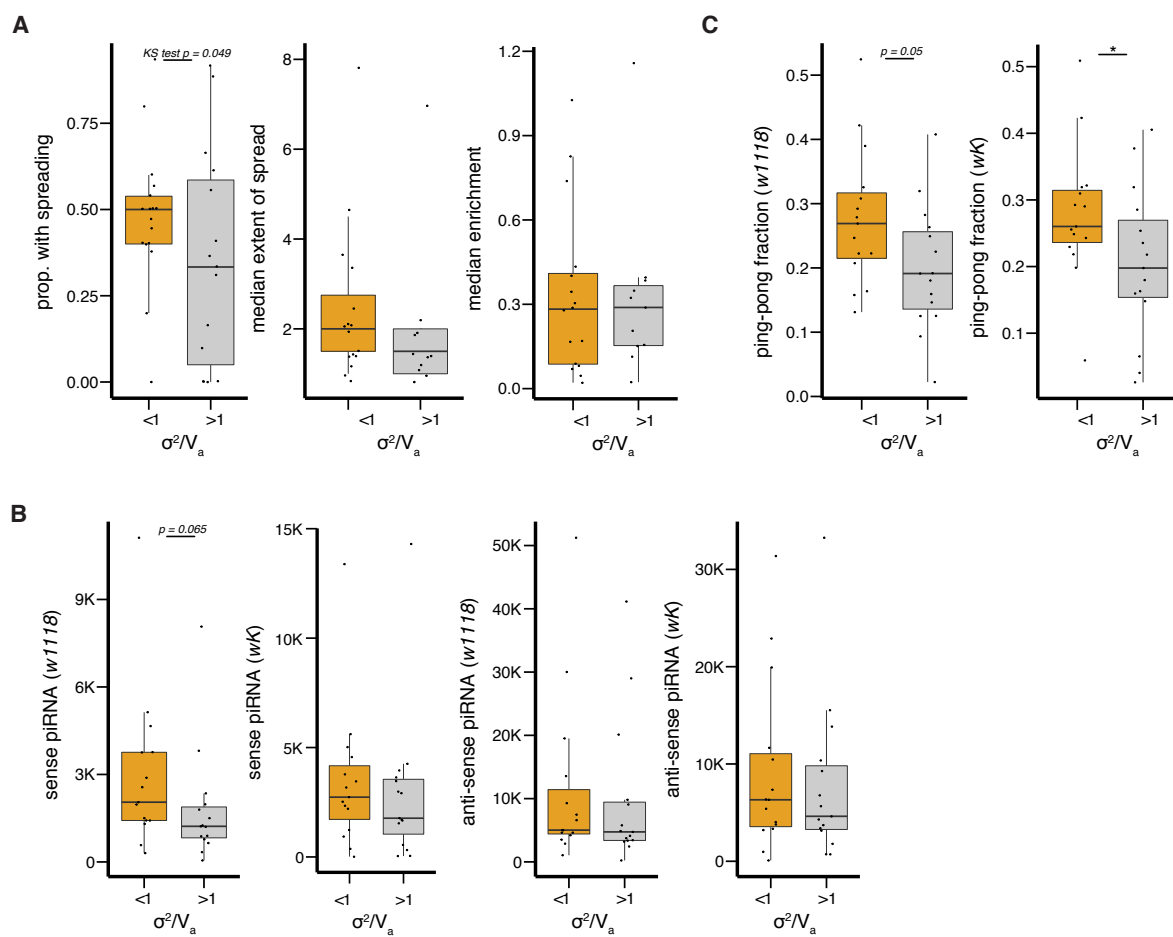
In addition to predictions that are shared with the ectopic recombination model, the epigenetic effect model has several unique predictions about which TE families are prone to exhibit synergistic fitness effects. The propensity to elicit epigenetic effects varies significantly among TE families (reviewed in (Choi and Lee 2020)), and intuitively, TE families that exert stronger such effects are more likely to interact synergistically. The synergism among the deleterious epigenetic effects of TEs in *Drosophila* was predicted to arise through the molecular details for piRNA production (Lee and Langley 2010). While other mechanisms also generate piRNAs ((Malone *et al.* 2009), reviewed in (Czech *et al.* 2018)), “ping-pong cycle” is thought to be responsible for the majorities of the piRNA amplification in flies. In this feed-forward cycle, TE transcripts, which are a source of sense piRNA precursors, and anti-sense piRNA precursors are reciprocally cleaved to generate mature sense and anti-sense piRNAs ((Gunawardane *et al.* 2007; Brennecke *et al.* 2007), reviewed in (Czech and Hannon 2016; Czech *et al.* 2018)). The amount of piRNAs, and accordingly the number of epigenetically silenced TEs and their associated deleterious effects, should grow quadratically or even exponentially with TE copy number (Lee and Langley 2010; Lee 2015). Interestingly, the involvement of ping-pong cycle in the generation of piRNA significantly vary between TE families (Li *et al.* 2009; Malone *et al.* 2009; Kelleher and Barbash 2013). Synergism is expected to have a higher tendency to arise for TE families that are targeted by more piRNAs generated via the ping-pong cycle.

To test these predictions, we used previously estimated indexes for the strength of epigenetic effects for individual TE family: the proportion of TEs resulting in *cis* spreading

of repressive marks, median extent of this spreading, and median magnitude of TE-induced increased enrichment of repressive marks (Lee and Karpen 2017). While there is an overall trend that TE families with  $\sigma^2/V_a < 1$  are more likely to exert epigenetic effects and result in a greater extent of the spreading of silencing marks (**Figure 3A**), only the estimated proportion of TEs with epigenetic effects have a statistically significant shifted distribution (shifted towards large value for families with  $\sigma^2/V_a < 1$ , *Kolmogorov-Smirnov test*,  $p$ -value = 0.049 when restricting to TEs in/near genes with known inviable mutant phenotypes, **Figure 3A**). We also compared the extent of piRNA targeting between TE families with  $\sigma^2/V_a$  smaller or greater than one (see Materials and Methods). We found that TE families with  $\sigma^2/V_a < 1$  are generally targeted by more piRNAs (**Figure 3B**). However, none of the comparisons are statistically significant, except for a marginally insignificant higher amount of sense piRNAs for TE families with  $\sigma^2/V_a < 1$  (for TEs in or near genes with known inviable mutant phenotypes, *Mann-Whitney U test*,  $p$ -value = 0.065, **Table 2 and Figure 3B**). Interestingly, we found that ping-pong fraction, which estimates the involvement of ping-pong cycle in piRNA generation, is significantly higher for TE families with  $\sigma^2/V_a < 1$  when focusing on TE insertions near or in genes with known inviable mutant phenotypes (*Mann-Whitney U test*,  $p$ -value = 0.050 (*w1118*) and 0.041 (*wK*), **Table 2 and Figure 3C**). Our observations are consistent with the predictions that synergistic epistasis may arise through piRNA amplification from the ping-pong cycle and the resultant deleterious epigenetic effects.

**Figure 3. Comparing piRNA-targeting and epigenetic effects between TE families with  $\sigma^2/V_a < 1$  or  $> 1$**

(A) Boxplots for the proportion of TEs showing spreading of repressive marks (left), the median extent of TE-mediated spreading of repressive marks (middle), and TE-mediated increase in the enrichment of repressive marks (right) for TE families with and without underdispersed distributions ( $\sigma^2/V_a < 1$ ). The distributions of proportion of TEs with spreading effects are significantly different, as indicated by *Kolmogorov-Smirnov test*. (B and C) Boxplots for sense and anti-sense piRNAs (B) and ping-pong fraction (C) corresponding to TE families with  $\sigma^2/V_a < 1$  or  $> 1$  when restricting to TEs that are in/near genes with known inviable mutant phenotypes. The amount of piRNA was estimated as the number of piRNA reads corresponding to a TE family per 1 million TE-corresponding piRNAs in (Kelleher and Barbash 2013). Results of two genotypes (w1118 and wK), both of which are wildtypes for the piRNA pathway genes, are shown. Also, see Table 2. \**Mann-Whitney test*,  $p < 0.05$ .



## Discussion

Theoretical analysis has predicted that, to stably contain the selfish increase of TEs, each additional TE insertion needs to impose a higher fitness cost than the last one, leading to purifying selection accelerating the removal of TEs with increased TE copy number (Charlesworth and Charlesworth 1983). This theoretical requirement has been extensively discussed in the context of TE evolutionary dynamics (Charlesworth and Langley 1989; Lee and Langley 2010; Kelleher *et al.* 2020; Choi and Lee 2020), and is predicted to be biologically plausible under several deleterious mechanisms of TEs, including TE-mediated ectopic recombination (Montgomery *et al.* 1987; Langley *et al.* 1988) and the spreading of silencing marks (Lee and Langley 2010; Lee 2015). However, the presence and prevalence of synergistic fitness effects among TEs are yet to be empirically tested.

Purifying selection with synergistic epistasis generates repulsion linkage among variants and, accordingly, an underdispersion of mutational burden when compared to selection without epistatic interactions (Charlesworth 1990; Kondrashov 1995; Sohail *et al.* 2017). Using approaches that leverage this population genetic signal (Sohail *et al.* 2017), we investigated the predicted synergistic epistasis among potentially deleterious TE insertions in the likely ancestral population of *D. melanogaster*. To account for the impacts of known (demography and missing data) and unknown factors that overdisperse TE burden, we compared the distribution of TE burden to that of bootstrapped synonymous variants to identify TEs that are likely underdispersed, an approach that has been implemented in previous studies investigating LD among variants (Sohail *et al.* 2017; Garcia and Lohmueller 2020) (but see (Sandler *et al.* 2020) for potential caveats). An important assumption of this approach is that overdispersing factors would influence the

mutational burden of TEs and synonymous variants similarly. Factors that only overdispense TE burden (e.g., variation in *trans* factors that influences TE silencing between genomes (Lee and Karpen 2017)) could not be addressed with this approach.

We found that TE burden likely has an underdispersed distribution, especially for TEs that are inserted into exons of highly conserved genes and expected to exert large fitness impacts. On the other hand, intergenic TEs near essential genes also have an underdispersed distribution. This observation could be driven by deleterious mechanisms of TEs that could impair host fitness from a distance to genes, such as insertion into regulatory elements or through the spreading of repressive epigenetic marks (reviewed in (Kelleher *et al.* 2020; Choi and Lee 2020)). By comparing various attributes of TE families with and without underdispersed distributions, we found that underdispersing TE families show stronger epigenetic effects and are targeted by more piRNAs generated from the ping-pong cycle. These observations are consistent with the hypothesis that synergism may arise through the deleterious epigenetic effects of TEs. Overall, our discoveries empirically supported the theoretically predicted synergistic fitness effects of TE insertions.

In addition to purifying selection with synergistic epistasis, repulsion LD could arise through selective interference among variants that are separated by small genetic distances (Hill and Robertson 1966; Felsenstein 1974; Garcia and Lohmueller 2020). To address this possibility, we estimated LD between pairs of TEs that are of different physical distance on the same chromosome or on different chromosomes (see Materials and Methods). Contrary to the prediction of selective interference, we observed negative LD mainly among TE pairs that are at least 1kb apart (**Figure S7**). In fact, for categories of TEs that we observed to have underdispersed distributions (nonLTR TEs, TEs in exons, and TEs

near essential genes), repulsion LD mainly arises among pairs of TEs that are physically separated by at least 0.1Mb (**Figure S7**). Because our analysis excluded TEs that are in or near pericentromeric heterochromatin, where recombination is strongly suppressed, observed negative linkage among TEs that are physically distant are less likely driven by selective interference than by synergistic fitness effects of TEs.

Different from the proposed source of synergistic epistasis of simple mutations (de Visser *et al.* 2011), synergistic fitness effects of TEs have been predicted to arise through unique mechanisms by which TEs impair host fitness. The illegitimate recombination between nonallelic TEs is predicted to lead to an accelerated removal of TEs with increased TE copy number, or synergistic epistasis (Montgomery *et al.* 1987; Langley *et al.* 1988). Under this model, TEs that are prone involved in ectopic recombination should be more likely to exhibit synergistic epistasis (Langley *et al.* 1988; Dolgin and Charlesworth 2008). While we did not find evidence supporting that TEs in genomic regions with high rates of meiotic recombination are more likely to underdisperse, several assumptions of our analysis could have confounded the results. Recombination landscapes vary between individuals (Dumont *et al.* 2009; Comeron *et al.* 2012; Hunter *et al.* 2016) and populations (Samuk *et al.* 2020), and could have been different between our studied Zambian population and the cosmopolitan population from which the recombination rate was estimated (Comeron *et al.* 2012). We also assumed that the rate of ectopic recombination closely mirrors that of homologous recombination (Lichten *et al.* 1987). This assumption has been questioned by the observed lack of TEs at the tip of the *D. melanogaster* X chromosome, where homologous recombination is strongly suppressed and TEs are expected to accumulate (Langley *et al.* 1988; Charlesworth and Lapid 1989).



We also found that TE families with underdispersed distribution do not follow the predictions of the ectopic recombination model, including being larger in abundance (Montgomery *et al.* 1987; Langley *et al.* 1988), longer in length (Petrov *et al.* 2003) or having higher sequence homology within families (Lee and Langley 2010; Petrov *et al.* 2011). These observations could result from our estimated properties of TE families (from the reference genome) are not representative of the studied population (an African population). On the other hand, our results may suggest that predictions of the ectopic recombination model need to be revised by incorporating additional biological details. It is recently proposed that the dependency of ectopic recombination on TE copy number should plateau when the number of TE insertions in the genome is large, and the process is unlikely limited by the number of potential recombining targets (Kelleher *et al.* 2020). According to this revised model, synergistic epistasis would only arise when TE copy number is below a certain threshold. In addition, the efficiency of recombination is observed to jointly depend on the length and sequence similarities of (reviewed in (Radman and Wagner 1993; Waldman 2008)), as well as the spatial distance between and orientation of, recombining partners (reviewed in (Renkawitz *et al.* 2014)). A model that incorporates these biological details may provide better predictions for the conditions by which synergistic epistasis may arise via ectopic recombination.

TE-mediated spreading of silencing marks is another mechanism from which synergistic epistasis was predicted to arise (Lee and Langley 2010; Lee 2015). In *Drosophila*, this process is initiated by piRNA-directed epigenetic silencing of TEs (Sienski *et al.* 2012; Le Thomas *et al.* 2013) (reviewed in (Czech *et al.* 2018)). Accordingly, many predictions of the model depend on how piRNAs are generated and target TE sequences.

These include predictions that are shared with the ectopic recombination model but not supported by our observations—TE families that are abundant (Lee and Langley 2010; Lee 2015; Lee and Karpen 2017), long (Lee 2015), and homogenous in sequences (Lee and Langley 2010) are more likely to exhibit synergistic epistasis. Again, the complexities of piRNA generation and targeting that are not considered in the current epigenetic effect model could have led to these discrepancies between predictions and observations. For instance, truncated TEs that lost the ability to transcribe would not contribute to piRNA generations through ping-pong cycle (Sienski *et al.* 2012; Olovnikov *et al.* 2013; Shpiz *et al.* 2014). Also, the targeting of TEs by piRNAs is abolished when the sequence divergence between the two is too large (Post *et al.* 2014; Kotov *et al.* 2019) and is particularly sensitive to mismatches at specific positions within the piRNA sequences (Wang *et al.* 2014; Mohn *et al.* 2015). Simple monotonic relationships could not fully capture how the abundance, length, and sequence homology of TE families influence the occurrence piRNA-targeting of TEs and the associated deleterious epigenetic effects.

On the other hand, the epigenetic effect model uniquely predicts that TE families resulting in more extensive spreading of repressive epigenetic marks, which is initiated by the piRNA-mediated silencing of TEs, should more likely to exhibit synergistic fitness effects. Supporting this prediction, we found that TE families with underdispersed distributions have a higher tendency to elicit epigenetic effects. Interestingly, we also found that piRNAs that target TE families with underdispersed distributions are more likely generated through the ping-pong cycle, which is a predicted source from which synergistic fitness effects of TEs arise (Lee and Langley 2010; Choi and Lee 2020). These observations suggest that TE-mediated epigenetic effects likely contribute to the synergistic epistatic

interactions among TE insertions, which could drive the stable containment of TE copy number.

It is worth noting that the statistical power for some of our current analyses may be limited due to challenges associated with studying TEs. TEs have a frequency spectrum that is highly skewed towards rare variants (Figure S1, also see (Stewart *et al.* 2011; Nellåker *et al.* 2012; Cridland *et al.* 2013; Kofler *et al.* 2015; Quadrana *et al.* 2016; Laricchia *et al.* 2017)). This low allele frequency would limit the range of possible LD estimates (Sved and Hill 2018), potentially restricting our ability to detect repulsion LD even if synergistic epistasis among deleterious TEs is present. Also, our ability to identify TEs and infer their biological properties (e.g., length and sequence identity) is limited with short-read sequencing data. Some of these limitations may be alleviated in the near future with the growing number of genomes that are sequenced by 3<sup>rd</sup>-generation long reads, which could significantly improve the identification of TEs and the assembly of their sequences (e.g., Debladis *et al.* 2017; Chakraborty *et al.* 2019; Ellison and Cao 2020).

By leveraging population genetic signals to circumvent direct measurements of individual fitness, we provided empirical evidence for the presence of synergistic epistasis among potentially deleterious TE insertions. While we identified that TE-mediated epigenetic effects may result in synergistic epistasis, our observations also suggest a need to incorporate additional biological details to refine predictions for how synergistic fitness effects of TEs may arise. With revised models and the expanding capacity of TE identifications with long-read sequencing, our analysis framework could provide a path forward to investigate the mode, prevalence, and importance of epistatic interactions in the evolutionary dynamics of TEs.

## Acknowledgment

I would like to thank Charles Langley for inspiring this study and Sohail Mashaal for helpful discussions. Andrea Betancourt, Jae Choi, Brandon Gaut, Yuheng Huang, and Kevin Thornton provided helpful comments on the manuscript. I also appreciate Andrea Betancourt's careful editing of the manuscript. This work was supported by NIH R00 GM121868.

## References

- Adrion J. R., M. J. Song, D. R. Schrider, M. W. Hahn, and S. Schaack, 2017 Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in *Drosophila Melanogaster*. *Genome Biol Evol* 9: 1329–1340. <https://doi.org/10.1093/gbe/evx050>
- Aravin A. A., R. Sachidanandam, A. Girard, K. Fejes-Toth, and G. J. Hannon, 2007 Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316: 744–747. <https://doi.org/10.1126/science.1142612>
- Aravin A. A., R. Sachidanandam, D. Bourc'his, C. Schaefer, D. Pezic, *et al.*, 2008 A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. *Mol. Cell* 31: 785–799. <https://doi.org/10.1016/j.molcel.2008.09.003>
- Babraham Bioinformatics - Trim Galore!
- Barrón M. G., A.-S. Fiston-Lavier, D. A. Petrov, and J. González, 2014 Population Genomics of Transposable Elements in *Drosophila*. *Annual Review of Genetics* 48: 561–581. <https://doi.org/10.1146/annurev-genet-120213-092359>
- Barton N. H., 1995 A general model for the evolution of recombination. *Genet. Res.* 65: 123–144. <https://doi.org/10.1017/S0016672300033140>
- Bellen H. J., R. W. Levis, G. Liao, Y. He, J. W. Carlson, *et al.*, 2004 The BDGP Gene Disruption Project: Single Transposon Insertions Associated With 40% of *Drosophila* Genes. *Genetics* 167: 761–781. <https://doi.org/10.1534/genetics.104.026427>
- Bellen H. J., R. W. Levis, Y. He, J. W. Carlson, M. Evans-Holm, *et al.*, 2011 The *Drosophila* Gene Disruption Project: Progress Using Transposons With Distinctive Site Specificities. *Genetics* 188: 731–743. <https://doi.org/10.1534/genetics.111.126995>
- 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. <https://doi.org/10.1016/j.cell.2007.01.043>
- Callahan B., R. A. Neher, D. Bachtrog, P. Andolfatto, and B. I. Shraiman, 2011 Correlated Evolution of Nearby Residues in *Drosophilid* Proteins. *PLoS Genet* 7: e1001315. <https://doi.org/10.1371/journal.pgen.1001315>

- Camacho C., G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, *et al.*, 2009 BLAST+ architecture and applications. *BMC Bioinformatics* 10: 421.  
<https://doi.org/10.1186/1471-2105-10-421>
- Chakraborty M., J. J. Emerson, S. J. Macdonald, and A. D. Long, 2019 Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications* 10: 1–11. <https://doi.org/10.1038/s41467-019-12884-1>
- Charlesworth B., and D. Charlesworth, 1983 The Population Dynamics of Transposable Elements. *Genetics Research* 42: 1–27.  
<https://doi.org/10.1017/S0016672300021455>
- Charlesworth B., and C. H. Langley, 1989 The population genetics of *Drosophila* transposable elements. *Annu. Rev. Genet.* 23: 251–287.  
<https://doi.org/10.1146/annurev.ge.23.120189.001343>
- Charlesworth B., and A. Lapid, 1989 A study of ten families of transposable elements on X chromosomes from a population of *Drosophila melanogaster*. *Genet. Res.* 54: 113–125.
- Charlesworth B., 1990 Mutation-selection balance and the evolutionary advantage of sex and recombination. *Genet Res* 55: 199–221.  
<https://doi.org/10.1017/s0016672300025532>
- Choi J. Y., and Y. C. G. Lee, 2020 Double-edged sword: The evolutionary consequences of the epigenetic silencing of transposable elements. *PLoS Genetics* 16: e1008872.  
<https://doi.org/10.1371/journal.pgen.1008872>
- 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.  
<https://doi.org/10.1038/nature06341>
- Comeron J. M., R. Ratnappan, and S. Bailin, 2012 The Many Landscapes of Recombination in *Drosophila melanogaster*. *PLoS Genet* 8: e1002905.  
<https://doi.org/10.1371/journal.pgen.1002905>
- Corbett-Detig R. B., and D. L. Hartl, 2012 Population Genomics of Inversion Polymorphisms in *Drosophila melanogaster*. *PLoS Genet* 8: e1003056.  
<https://doi.org/10.1371/journal.pgen.1003056>
- Cridland J. M., S. J. Macdonald, A. D. Long, and K. R. Thornton, 2013 Abundance and Distribution of Transposable Elements in Two *Drosophila* QTL Mapping Resources. *Mol Biol Evol* 30: 2311–2327. <https://doi.org/10.1093/molbev/mst129>
- Czech B., and G. J. Hannon, 2016 One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends in Biochemical Sciences* 41: 324–337.  
<https://doi.org/10.1016/j.tibs.2015.12.008>
- Czech B., M. Munafò, F. Ciabrelli, E. L. Eastwood, M. H. Fabry, *et al.*, 2018 piRNA-Guided Genome Defense: From Biogenesis to Silencing. *Annual Review of Genetics* 52: 131–157. <https://doi.org/10.1146/annurev-genet-120417-031441>
- Daborn P. J., J. L. Yen, M. R. Bogwitz, G. Le Goff, E. Feil, *et al.*, 2002 A Single P450 Allele Associated with Insecticide Resistance in *Drosophila*. *Science* 297: 2253–2256.  
<https://doi.org/10.1126/science.1074170>
- Davis P. S., M. W. Shen, and B. H. Judd, 1987 Asymmetrical pairings of transposons in and proximal to the white locus of *Drosophila* account for four classes of regularly occurring exchange products. *Proc Natl Acad Sci U S A* 84: 174–178.  
<https://doi.org/10.1073/pnas.84.1.174>

- Debladis E., C. Llauro, M.-C. Carpentier, M. Mirouze, and O. Panaud, 2017 Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics* 18: 537. <https://doi.org/10.1186/s12864-017-3753-z>
- Deniz Ö., J. M. Frost, and M. R. Branco, 2019 Regulation of transposable elements by DNA modifications. *Nature Reviews Genetics* 20: 417–431. <https://doi.org/10.1038/s41576-019-0106-6>
- Devos K. M., J. K. M. Brown, and J. L. Bennetzen, 2002 Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in *Arabidopsis*. *Genome Res.* 12: 1075–1079. <https://doi.org/10.1101/gr.132102>
- Díaz-González J., J. F. Vázquez, J. Albornoz, and A. Domínguez, 2011 Long-term evolution of the roo transposable element copy number in mutation accumulation lines of *Drosophila melanogaster*. *Genet Res (Camb)* 1–7. <https://doi.org/10.1017/S0016672311000103>
- Dolgin E. S., and B. Charlesworth, 2008 The effects of recombination rate on the distribution and abundance of transposable elements. *Genetics* 178: 2169–2177. <https://doi.org/10.1534/genetics.107.082743>
- Dumont B. L., K. W. Broman, and B. A. Payseur, 2009 Variation in Genomic Recombination Rates Among Heterogeneous Stock Mice. *Genetics* 182: 1345–1349. <https://doi.org/10.1534/genetics.109.105114>
- Edgar R. C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32: 1792–1797. <https://doi.org/10.1093/nar/gkh340>
- Ellison C. E., and W. Cao, 2020 Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of *Drosophila melanogaster*. *Nucleic Acids Research* 48: 290–303. <https://doi.org/10.1093/nar/gkz1080>
- Eshel I., and M. W. Feldman, 1970 On the evolutionary effect of recombination. *Theor Popul Biol* 1: 88–100. [https://doi.org/10.1016/0040-5809\(70\)90043-2](https://doi.org/10.1016/0040-5809(70)90043-2)
- Ewens W. J., and R. S. Spielman, 1995 The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 57: 455–464.
- Ewing B., L. Hillier, M. C. Wendl, and P. Green, 1998 Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* 8: 175–185.
- Felsenstein J., 1974 The Evolutionary Advantage of Recombination. *Genetics* 78: 737–756.
- García J. A., and K. E. Lohmueller, 2020 *Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome*. *Evolutionary Biology*.
- González J., and D. A. Petrov, 2009 The adaptive role of transposable elements in the *Drosophila* genome. *Gene* 448: 124–133. <https://doi.org/10.1016/j.gene.2009.06.008>
- Gunawardane L. S., K. Saito, K. M. Nishida, K. Miyoshi, Y. Kawamura, *et al.*, 2007 A slicer-mediated mechanism for repeat-associated siRNA 5' end formation in *Drosophila*. *Science* 315: 1587–1590. <https://doi.org/10.1126/science.1140494>
- Hill W. G., and A. Robertson, 1966 The effect of linkage on limits to artificial selection. *Genet. Res* 8: 269–294.

- Hof A. E. van't, P. Campagne, D. J. Rigden, C. J. Yung, J. Lingley, *et al.*, 2016 The industrial melanism mutation in British peppered moths is a transposable element. *Nature* 534: 102–105. <https://doi.org/10.1038/nature17951>
- Hollister J. D., and B. S. Gaut, 2009 Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res.* 19: 1419–1428. <https://doi.org/10.1101/gr.091678.109>
- Hoskins R. A., J. W. Carlson, K. H. Wan, S. Park, I. Mendez, *et al.*, 2015 The Release 6 reference sequence of the *Drosophila melanogaster* genome. *Genome Res.* gr.185579.114. <https://doi.org/10.1101/gr.185579.114>
- Houle D., and S. V. Nuzhdin, 2004 Mutation accumulation and the effect of copia insertions in *Drosophila melanogaster*. *Genet Res* 83: 7–18. <https://doi.org/10.1017/s0016672303006505>
- Hu T. T., M. B. Eisen, K. R. Thornton, and P. Andolfatto, 2013 A second-generation assembly of the *Drosophila simulans* genome provides new insights into patterns of lineage-specific divergence. *Genome Res.* 23: 89–98. <https://doi.org/10.1101/gr.141689.112>
- Huang W., A. Massouras, Y. Inoue, J. Peiffer, M. Ràmia, *et al.*, 2014 Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.* 24: 1193–1208. <https://doi.org/10.1101/gr.171546.113>
- Hunter C. M., W. Huang, T. F. C. Mackay, and N. D. Singh, 2016 The Genetic Architecture of Natural Variation in Recombination Rate in *Drosophila melanogaster*. *PLOS Genetics* 12: e1005951. <https://doi.org/10.1371/journal.pgen.1005951>
- Kaminker J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, *et al.*, 2002 The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. *Genome Biol.* 3: RESEARCH0084.
- Kapitonov V. V., and J. Jurka, 2003 Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. *PNAS* 100: 6569–6574. <https://doi.org/10.1073/pnas.0732024100>
- Kasschau K. D., N. Fahlgren, E. J. Chapman, C. M. Sullivan, J. S. Cumbie, *et al.*, 2007 Genome-Wide Profiling and Analysis of *Arabidopsis* siRNAs. *PLOS Biology* 5: e57. <https://doi.org/10.1371/journal.pbio.0050057>
- Kelleher E. S., and D. A. Barbash, 2013 Analysis of piRNA-Mediated Silencing of Active TEs in *Drosophila melanogaster* Suggests Limits on the Evolution of Host Genome Defense. *Mol Biol Evol* 30: 1816–1829. <https://doi.org/10.1093/molbev/mst081>
- Kelleher E. S., D. A. Barbash, and J. P. Blumenstiel, 2020 Taming the Turmoil Within: New Insights on the Containment of Transposable Elements. *Trends in Genetics* 36: 474–489. <https://doi.org/10.1016/j.tig.2020.04.007>
- Killick S. C., A. M. Carlsson, S. A. West, and T. J. Little, 2006 Testing the pluralist approach to sex: the influence of environment on synergistic interactions between mutation load and parasitism in *Daphnia magna*. *Journal of Evolutionary Biology* 19: 1603–1611. <https://doi.org/10.1111/j.1420-9101.2006.01123.x>
- Kishony R., and S. Leibler, 2003 Environmental stresses can alleviate the average deleterious effect of mutations. *J Biol* 2: 14.
- Kofler R., V. Nolte, and C. Schlötterer, 2015 Tempo and Mode of Transposable Element Activity in *Drosophila*. *PLoS Genet* 11: e1005406. <https://doi.org/10.1371/journal.pgen.1005406>

- Kondrashov A. S., 1995 Dynamics of unconditionally deleterious mutations: Gaussian approximation and soft selection. *Genetics Research* 65: 113–121. <https://doi.org/10.1017/S0016672300033139>
- Kotov A. A., V. E. Adashev, B. K. Godneeva, M. Ninova, A. S. Shatskikh, *et al.*, 2019 piRNA silencing contributes to interspecies hybrid sterility and reproductive isolation in *Drosophila melanogaster*. *Nucleic Acids Res* 47: 4255–4271. <https://doi.org/10.1093/nar/gkz130>
- Kupiec M., and T. D. Petes, 1988 Allelic and ectopic recombination between Ty elements in yeast. *Genetics* 119: 549–559.
- Lack J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, *et al.*, 2015 The *Drosophila* Genome Nexus: A Population Genomic Resource of 623 *Drosophila melanogaster* Genomes, Including 197 from a Single Ancestral Range Population. *Genetics* genetics.115.174664. <https://doi.org/10.1534/genetics.115.174664>
- Lagemaat L. N. van de, L. Gagnier, P. Medstrand, and D. L. Mager, 2005 Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. *Genome Res.* 15: 1243–1249. <https://doi.org/10.1101/gr.3910705>
- 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.
- Laricchia K. M., S. Zdraljevic, D. E. Cook, and E. C. Andersen, 2017 Natural Variation in the Distribution and Abundance of Transposable Elements Across the *Caenorhabditis elegans* Species. *Mol Biol Evol* 34: 2187–2202. <https://doi.org/10.1093/molbev/msx155>
- Larracunte A. M., T. B. Sackton, A. J. Greenberg, A. Wong, N. D. Singh, *et al.*, 2008 Evolution of protein-coding genes in *Drosophila*. *Trends in Genetics* 24: 114–123. <https://doi.org/10.1016/j.tig.2007.12.001>
- Le Thomas A., A. K. Rogers, A. Webster, G. K. Marinov, S. E. Liao, *et al.*, 2013 Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. *Genes Dev* 27: 390–399. <https://doi.org/10.1101/gad.209841.112>
- Lee Y. C. G., and C. H. Langley, 2010 Transposable elements in natural populations of *Drosophila melanogaster*. *Philos Trans R Soc Lond B Biol Sci* 365: 1219–1228. <https://doi.org/10.1098/rstb.2009.0318>
- Lee Y. C. G., and J. A. Reinhardt, 2012 Widespread Polymorphism in the Positions of Stop Codons in *Drosophila melanogaster*. *Genome Biol Evol* 4: 533–549. <https://doi.org/10.1093/gbe/evr113>
- 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: e1005269. <https://doi.org/10.1371/journal.pgen.1005269>
- Lee Y. C. G., and G. H. Karpen, 2017 Pervasive epigenetic effects of *Drosophila* euchromatic transposable elements impact their evolution. *eLife* 6. <https://doi.org/10.7554/eLife.25762>
- Lewontin R. C., and K. Kojima, 1960 The Evolutionary Dynamics of Complex Polymorphisms. *Evolution* 14: 458–472. <https://doi.org/10.1111/j.1558-5646.1960.tb03113.x>



- Li C., V. V. Vagin, S. Lee, J. Xu, S. Ma, *et al.*, 2009 Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. *Cell* 137: 509–521.  
<https://doi.org/10.1016/j.cell.2009.04.027>
- Li H., and R. Durbin, 2010 Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* 26: 589–595.  
<https://doi.org/10.1093/bioinformatics/btp698>
- Li H., 2011 A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987–2993. <https://doi.org/10.1093/bioinformatics/btr509>
- Lichten M., R. H. Borts, and J. E. Haber, 1987 Meiotic Gene Conversion and Crossing Over Between Dispersed Homologous Sequences Occurs Frequently in *Saccharomyces cerevisiae*. *Genetics* 115: 233–246.
- Lim J. K., and M. J. Simmons, 1994 Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. *BioEssays* 16: 269–275.  
<https://doi.org/10.1002/bies.950160410>
- Mackay T. F., 1989 Transposable elements and fitness in *Drosophila melanogaster*. *Genome* 31: 284–295.
- Maksakova I. A., M. T. Romanish, L. Gagnier, C. A. Dunn, L. N. van de Lagemaat, *et al.*, 2006 Retroviral Elements and Their Hosts: Insertional Mutagenesis in the Mouse Germ Line. *PLoS Genet* 2: e2. <https://doi.org/10.1371/journal.pgen.0020002>
- Malone C. D., J. Brennecke, M. Dus, A. Stark, W. R. McCombie, *et al.*, 2009 Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* 137: 522–535. <https://doi.org/10.1016/j.cell.2009.03.040>
- Marí-Ordóñez A., A. Marchais, M. Etcheverry, A. Martin, V. Colot, *et al.*, 2013 Reconstructing de novo silencing of an active plant retrotransposon. *Nature Genetics* 45: 1029–1039. <https://doi.org/10.1038/ng.2703>
- Maside X., S. Assimakopoulos, and B. Charlesworth, 2000 Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. *Genet. Res.* 75: 275–284.
- McCue A. D., K. Panda, S. Nuthikattu, S. G. Choudury, E. N. Thomas, *et al.*, 2015 ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. *The EMBO Journal* 34: 20–35.  
<https://doi.org/10.15252/embj.201489499>
- Mieczkowski P. A., F. J. Lemoine, and T. D. Petes, 2006 Recombination between retrotransposons as a source of chromosome rearrangements in the yeast *Saccharomyces cerevisiae*. *DNA Repair (Amst.)* 5: 1010–1020.  
<https://doi.org/10.1016/j.dnarep.2006.05.027>
- Mohn F., D. Handler, and J. Brennecke, 2015 piRNA-guided slicing specifies transcripts for Zucchini dependent, phased piRNA biogenesis. *Science* 348: 812–817.  
<https://doi.org/10.1126/science.aaa1039>
- Montgomery E., B. Charlesworth, and C. H. Langley, 1987 A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. *Genet. Res.* 49: 31–41.
- Montgomery E. A., S. M. 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.

- Nellåker C., T. M. Keane, B. Yalcin, K. Wong, A. Agam, *et al.*, 2012 The genomic landscape shaped by selection on transposable elements across 18 mouse strains. *Genome Biology* 13: R45. <https://doi.org/10.1186/gb-2012-13-6-r45>
- Nuzhdin S. V., and T. F. Mackay, 1995 The genomic rate of transposable element movement in *Drosophila melanogaster*. *Mol. Biol. Evol.* 12: 180–181.
- Olovnikov I., S. Ryazansky, S. Shpiz, S. Lavrov, Y. Abramov, *et al.*, 2013 De novo piRNA cluster formation in the *Drosophila* germ line triggered by transgenes containing a transcribed transposon fragment. *Nucl. Acids Res.* 41: 5757–5768. <https://doi.org/10.1093/nar/gkt310>
- Ozata D. M., I. Gainetdinov, A. Zoch, D. O'Carroll, and P. D. Zamore, 2019 PIWI-interacting RNAs: small RNAs with big functions. *Nat Rev Genet* 20: 89–108. <https://doi.org/10.1038/s41576-018-0073-3>
- 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: 284–290. <https://doi.org/10.1093/jhered/esh050>
- Peters A. D., and P. D. Keightley, 2000 A Test for Epistasis Among Induced Mutations in *Caenorhabditis elegans*. *Genetics* 156: 1635–1647.
- 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. <https://doi.org/10.1093/molbev/msg102>
- 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. <https://doi.org/10.1093/molbev/msq337>
- Pool J. E., R. B. Corbett-Detig, R. P. Sugino, K. A. Stevens, C. M. Cardeno, *et al.*, 2012 Population Genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genet* 8: e1003080. <https://doi.org/10.1371/journal.pgen.1003080>
- Post C., J. P. Clark, Y. A. Sytnikova, G.-W. Chirn, and N. C. Lau, 2014 The capacity of target silencing by *Drosophila* PIWI and piRNAs. *RNA* 20: 1977–1986. <https://doi.org/10.1261/rna.046300.114>
- Purcell S., B. Neale, K. Todd-Brown, L. Thomas, M. A. R. Ferreira, *et al.*, 2007 PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81: 559–575. <https://doi.org/10.1086/519795>
- Quadrana L., A. B. Silveira, G. F. Mayhew, C. LeBlanc, R. A. Martienssen, *et al.*, 2016 The *Arabidopsis thaliana* mobilome and its impact at the species level. *eLife* 5: e15716. <https://doi.org/10.7554/eLife.15716>
- Quinlan A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26: 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Radman M., and R. Wagner, 1993 Mismatch recognition in chromosomal interactions and speciation. *Chromosoma* 102: 369–373. <https://doi.org/10.1007/BF00360400>
- Rahman R., G. Chirn, A. Kanodia, Y. A. Sytnikova, B. Brembs, *et al.*, 2015 Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. *Nucl. Acids Res.* 43: 10655–10672. <https://doi.org/10.1093/nar/gkv1193>

- Rebollo R., M. M. Karimi, M. Bilenky, L. Gagnier, K. Miceli-Royer, *et al.*, 2011 Retrotransposon-Induced Heterochromatin Spreading in the Mouse Revealed by Insertional Polymorphisms. *PLoS Genet* 7: e1002301. <https://doi.org/10.1371/journal.pgen.1002301>
- Renkawitz J., C. A. Lademann, and S. Jentsch, 2014 Mechanisms and principles of homology search during recombination. *Nature Reviews Molecular Cell Biology* 15: 369–383. <https://doi.org/10.1038/nrm3805>
- Riddle N. C., A. Minoda, P. V. Kharchenko, A. A. Alekseyenko, Y. B. Schwartz, *et al.*, 2011 Plasticity in patterns of histone modifications and chromosomal proteins in *Drosophila* heterochromatin. *Genome Res.* 21: 147–163. <https://doi.org/10.1101/gr.110098.110>
- Riddle N. C., and S. C. R. Elgin, 2018 The *Drosophila* Dot Chromosome: Where Genes Flourish Amidst Repeats. *Genetics* 210: 757–772. <https://doi.org/10.1534/genetics.118.301146>
- Robberecht C., T. Voet, M. Z. Esteki, B. A. Nowakowska, and J. R. Vermeesch, 2013 Nonallelic homologous recombination between retrotransposable elements is a driver of de novo unbalanced translocations. *Genome Res.* 23: 411–418. <https://doi.org/10.1101/gr.145631.112>
- Rogers A. R., 2014 How Population Growth Affects Linkage Disequilibrium. *Genetics* 197: 1329–1341. <https://doi.org/10.1534/genetics.114.166454>
- Samuk K., B. Manzano-Winkler, K. R. Ritz, and M. A. F. Noor, 2020 Natural Selection Shapes Variation in Genome-wide Recombination Rate in *Drosophila pseudoobscura*. *Current Biology* 30: 1517–1528.e6. <https://doi.org/10.1016/j.cub.2020.03.053>
- Sandler G., S. I. Wright, and A. F. Agrawal, 2020 Using Patterns of Signed Linkage Disequilibria to Test for Epistasis in Flies and Plants. *bioRxiv* 2020.11.25.399030. <https://doi.org/10.1101/2020.11.25.399030>
- Schmidt J. M., R. T. Good, B. Appleton, J. Sherrard, G. C. Raymant, *et al.*, 2010 Copy number variation and transposable elements feature in recent, ongoing adaptation at the *Cyp6g1* locus. *PLoS Genet.* 6: e1000998. <https://doi.org/10.1371/journal.pgen.1000998>
- Shpiz S., S. Ryazansky, I. Olovnikov, Y. Abramov, and A. Kalmykova, 2014 Euchromatic Transposon Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the *Drosophila* Germline. *PLoS Genet* 10: e1004138. <https://doi.org/10.1371/journal.pgen.1004138>
- Sienski G., D. Dönertas, and J. Brennecke, 2012 Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* 151: 964–980. <https://doi.org/10.1016/j.cell.2012.10.040>
- Singh N. D., and D. A. Petrov, 2004 Rapid Sequence Turnover at an Intergenic Locus in *Drosophila*. *Mol Biol Evol* 21: 670–680. <https://doi.org/10.1093/molbev/msh060>
- Sohail M., O. A. Vakhrusheva, J. H. Sul, S. L. Pulit, L. C. Francioli, *et al.*, 2017 Negative selection in humans and fruit flies involves synergistic epistasis. *Science* 356: 539–542. <https://doi.org/10.1126/science.aah5238>
- Sprengelmeyer Q. D., S. Mansourian, J. D. Lange, D. R. Matute, B. S. Cooper, *et al.*, 2020 Recurrent Collection of *Drosophila melanogaster* from Wild African Environments and Genomic Insights into Species History. *Molecular Biology and Evolution* 37: 627–638. <https://doi.org/10.1093/molbev/msz271>

- Stewart C., D. Kural, M. P. Strömberg, J. A. Walker, M. K. Konkel, *et al.*, 2011 A Comprehensive Map of Mobile Element Insertion Polymorphisms in Humans. *PLoS Genet* 7: e1002236. <https://doi.org/10.1371/journal.pgen.1002236>
- Stuart T., S. R. Eichten, J. Cahn, Y. V. Karpievitch, J. O. Borevitz, *et al.*, 2016 Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. *eLife*.
- Sved J. A., and W. G. Hill, 2018 One Hundred Years of Linkage Disequilibrium. *Genetics* 209: 629–636. <https://doi.org/10.1534/genetics.118.300642>
- Visser J. A. G. M. de, T. F. Cooper, and S. F. Elena, 2011 The causes of epistasis. *Proc. R. Soc. B*. 278: 3617–3624. <https://doi.org/10.1098/rspb.2011.1537>
- Waldman A. S., 2008 Ensuring the fidelity of recombination in mammalian chromosomes. *BioEssays* 30: 1163–1171. <https://doi.org/10.1002/bies.20845>
- Wang W., M. Yoshikawa, B. W. Han, N. Izumi, Y. Tomari, *et al.*, 2014 The Initial Uridine of Primary piRNAs Does not Create the Tenth Adenine That is the Hallmark of Secondary piRNAs. *Mol Cell* 56: 708–716. <https://doi.org/10.1016/j.molcel.2014.10.016>
- Waterhouse R. M., E. M. Zdobnov, and E. V. Kriventseva, 2011 Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. *Genome Biol Evol* 3: 75–86. <https://doi.org/10.1093/gbe/evq083>
- Wells J. N., and C. Feschotte, 2020 A Field Guide to Eukaryotic Transposable Elements. *Annual Review of Genetics* 54: 539–561. <https://doi.org/10.1146/annurev-genet-040620-022145>
- Xie Z., L. K. Johansen, A. M. Gustafson, K. D. Kasschau, A. D. Lellis, *et al.*, 2004 Genetic and Functional Diversification of Small RNA Pathways in Plants. *PLOS Biology* 2: e104. <https://doi.org/10.1371/journal.pbio.0020104>
- Yang Z., 2007 PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586–1591. <https://doi.org/10.1093/molbev/msm088>
- Yang P., Y. Wang, and T. S. Macfarlan, 2017 The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. *Trends in Genetics* 33: 871–881. <https://doi.org/10.1016/j.tig.2017.08.006>
- Zavattari P., E. Deidda, M. Whalen, R. Lampis, A. Mulargia, *et al.*, 2000 Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. *Human Molecular Genetics* 9: 2947–2957. <https://doi.org/10.1093/hmg/9.20.2947>

**Table 1. Mutational burden and additive genetic variance of different categories of TEs**

TE type	no. TEs	mean TE burden	$\sigma^2$	$V_a$	$\sigma^2/V_a$	one-sided <i>p</i> -value
All TEs	11396	77.92	180.43	80.80	2.23	<b>0.043<sup>a</sup></b>
<b>TE class</b>						
DNA	2522	17.58	25.01	18.23	1.37	0.691
RNA	8354	56.58	122.52	58.69	2.09	0.494
<b>TE types</b>						
TIR	2522	17.58	25.01	18.23	1.37	0.685
LTR	6697	44.16	88.67	45.84	1.93	0.639
nonLTR	1657	12.42	12.61	12.84	<b>0.98<sup>b</sup></b>	<b>0.048</b>
<b>TE insertion locations</b>						
coding exons	120	0.73	0.74	0.76	<b>0.97</b>	0.493
UTRs	5609	38.26	57.86	39.65	1.46	<b>0.043</b>
exons	5729	38.99	66.19	44.90	1.47	<b>0.013</b>
intron	1498	10.49	12.24	10.85	1.13	0.381
intergenic	4169	28.44	40.97	29.54	1.39	0.172
<b>dN/dS ratio of nearest gene</b>						
1st quartile: [0, 0.0341)	2538	17.50	11.82	12.16	<b>0.97</b>	0.098
2nd quartile: [0.0341, 0.0877)	2702	18.48	13.84	12.78	1.08	0.080
3rd quartile: [0.0877, 0.1932)	2840	19.35	15.86	13.35	1.19	0.058
4th quartile: [0.1932, )	1686	11.40	8.75	7.99	1.09	0.364
<b>known pheno.<sup>c</sup> of nearest gene</b>						
lethal	4351	9.75	40.25	30.84	1.31	<b>0.045</b>
semi-lethal	827	5.64	6.23	5.83	1.07	0.487
non-viable (lethal and semi-lethal)	5178	35.38	53.88	36.66	1.47	0.116
viable	4116	28.09	43.89	29.12	1.51	0.449

<sup>a,b</sup>  $\sigma^2/V_a < 1$  (<sup>b</sup>) or bootstrapping *p*-value < 0.05 (<sup>a</sup>) are in bold

<sup>c</sup> phenotype

**(cont.) Table 1. Mutational burden and additive genetic variance of different categories of TEs**

TE type	no. TEs	mean TE burden	$\sigma^2$	$V_a$	$\sigma^2/V_a$	one-sided p-value
<b>TEs in exons, dN/dS ratio of nearest gene</b>						
dNdS [0, 0.0341)	1500	10.21	11.62	10.59	1.10	0.298
dNdS [0.0341, 0.0877)	1666	11.35	10.74	11.75	<b>0.91</b>	<b>0.028</b>
dNdS [0.0877, 0.1932)	1762	12.10	13.46	12.52	1.08	0.138
dNdS [0.1932, )	780	5.32	6.58	5.54	1.19	0.942
<b>TEs in exons, known pheno. of nearest gene</b>						
lethal	2855	19.36	26.35	20.08	1.31	0.412
semi-lethal	438	2.99	3.08	3.10	<b>0.99</b>	0.385
non-viable	3293	22.35	30.30	23.17	1.31	0.228
viable	2033	13.91	20.89	14.40	1.45	0.953
<b>Intergenic TEs, dN/dS ratio of nearest gene</b>						
1st quartile: [0, 0.0341)	651	4.55	4.81	4.73	1.02	0.408
2nd quartile: [0.0341, 0.0877)	736	5.05	5.54	5.24	1.06	0.524
3rd quartile: [0.0877, 0.1932)	718	4.79	4.47	4.97	<b>0.90</b>	0.075
4th quartile: [0.1932, )	802	5.38	5.58	5.58	1.00	0.248
<b>Intergenic TEs, known pheno. of nearest gene</b>						
lethal	800	5.60	5.06	5.80	<b>0.87</b>	<b>0.034</b>
semi-lethal	226	1.43	1.55	1.49	1.04	0.676
non-viable (lethal and semi-lethal)	1026	7.03	7.46	7.29	1.02	0.257
viable	1671	11.38	13.71	11.81	1.16	0.388
<b>Local recombination rate (cM/Mbp)</b>						
1st quartile: [0, 1.344)	2838	19.65	28.75	20.33	1.41	0.659
2nd quartile: [1.344, 2.354)	2852	19.73	23.49	20.40	1.15	0.073
3rd quartile: [2.354, 3.64)	2849	19.09	24.13	19.79	1.22	0.174
4th quartile: [3.64, )	2857	19.45	30.10	20.28	1.48	0.775

**Table 2. Comparisons of biological properties of TE families with  $\sigma^2/Va < 1$  or  $> 1$**

	all TEs	in or near essential genes		
		1st quartile dN/dS	lethal phenotype	inviable phenotype
<b>TE family copy number</b>				
reference genome	0.367	0.221	0.473	0.895
DPGP3 genomes	0.912	0.944	0.506	0.354
<b>TE family average length</b>				
reference genome	0.674	0.745	<b>0.010</b>	<b>0.007</b>
<b>TE family mean pairwise difference</b>				
reference genome	0.400	0.669	0.332	0.696
<b>TE family epigenetic effects</b>				
proportion of TEs with epi. effect <sup>a</sup>	0.391	0.559	0.214	<b>0.172<sup>d</sup></b>
median extent of epi. effect <sup>b</sup>	0.565	0.777	0.571	0.214
median increase of epi. effect <sup>c</sup>	0.208	0.742	0.334	0.134
<b>piRNAs correspond to a TE family</b>				
sense piRNA (strain <i>w1118</i> )	0.966	0.868	0.106	<b>0.065</b>
sense piRNA (strain <i>wK</i> )	0.878	0.224	0.811	0.678
anti-sense piRNA (strain <i>w1118</i> )	0.686	0.868	0.106	0.461
anti-sense piRNA (strain <i>wK</i> )	0.512	0.224	0.811	0.653
ping-pong fraction (strain <i>w1118</i> )	0.321	0.673	0.325	<b>0.050</b>
ping-pong fraction (strain <i>wK</i> )	0.234	1.000	0.325	<b>0.041</b>

<sup>a</sup>proportion of TEs in a family that result in spreading of repressive epigenetic marks in *Lee and Karpen (2017)*

<sup>b</sup>median extent of spreading of repressive epigenetic marks of a TE family in *Lee and Karpen (2017)*

<sup>c</sup>median TE-induced increase in enrichment of repressive epigenetic marks of a TE family in *Lee and Karpen (2017)*

<sup>d</sup>Kolmogorov-Smirnov test, p-value = 0.049

