OXFORD GENETICS

GENETICS, 2022, 220(2), iyab211
https://doi.org/10.1093/genetics/iyab211
Advance Access Publication Date: 9 December 2021
Investigation
Highlighted Article

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 92697, USA

*Author for correspondence: Department of Ecology and Evolutionary Biology, University of California Irvine, McGaugh Hall 5207, Irvine, CA 92697, USA.
Email: grylee@uci.edu

Abstract

The replicative nature and generally deleterious effects of transposable elements (TEs) raise 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. We investigated this population genetic signal in the likely ancestral *Drosophila melanogaster* population and found evidence supporting the presence of synergistic epistasis among TE insertions, especially TEs expected to exert large fitness impacts. Even though synergistic epistasis of TEs has been predicted to arise through ectopic recombination and TE-mediated epigenetic silencing mechanisms, we only found mixed support for the associated predictions. We observed signals of synergistic epistasis for a large number of TE families, which is consistent with the expectation that such epistatic interaction mainly happens among copies of the same family. Curiously, significant repulsion linkage was also found among TE insertions from different families, suggesting the possibility that synergism of TEs' deleterious fitness effects could arise above the family level and through mechanisms similar to those of simple mutations. Our findings set the stage for investigating the prevalence and importance of epistatic interactions in the evolutionary dynamics of TEs.

Keywords: transposable elements; epistasis; linkage disequilibrium; ectopic recombination; epigenetic effects; Drosophila; synergistic epistasis

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; Deniz *et al.* 2019; Ozata *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 the estimated rate 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.*, <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 toward 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

Received: September 14, 2021. **Accepted:** November 10, 2021

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 Choi and Lee 2020; Kelleher *et al.* 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 perturbating TE evolutionary dynamics.

Synergism among the deleterious 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, Aravin *et al.* 2007; Brennecke *et al.* 2007; Gunawardane *et al.* 2007). With the assumption that the amount of small RNA is in excess, the *probability* of a TE being targeted by small RNA will depend on the TE copy number. Accordingly, the *number* of TEs that will be epigenetically silenced and thus impair host fitness will grow quadratically with increased TE copy number, leading to synergism among the deleterious 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 measure experimentally, but are expected to strongly influence the population dynamics of TEs in nature. Therefore, an orthogonal approach that does not rely on the direct measurement of individual fitness is needed to investigate the predicted synergistic fitness effects of TEs.

To test the presence of 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; Sandler *et al.* 2021) or between species (Callahan *et al.* 2011). In particular, Sohail *et al.* (2017) and Sandler *et al.* (2021) used the correlation between allele frequencies at different sites, or linkage disequilibrium (LD), to demonstrate the presence of synergistic epistasis among nonsense and missense 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 *Drosophila melanogaster* Zambian population (Lack *et al.* 2015). 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. Overall, by examining the LD distribution of TE insertions, we tested for the presence of the predicted synergistic epistasis among TEs and inferred the likely source of such synergism.

Materials and methods

Population genomic data

We used DPGP3 Zambian *D. melanogaster* strains 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), with read length smaller than 100 bp (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 can be found in Supplementary Table S1.

Identification, calling presence/absence, and filtering of TE insertions

To identify TE insertions and determine their presence/absence status in all genomes, we used approaches developed in Lee and Karpen (2017), which combined TE calling pipeline, TIDAL (Rahman *et al.* 2015), and methods of Cridland *et al.* (2013) (also see Supplementary Figure S1 for a flowchart). The reason to employ this two-step approach is that TIDAL only calls the presence of a nonreference TE insertion, but not whether that specific nonreference TE insertion is absent in other genomes. For our analysis, it is important to distinguish, when a TE insertion is not called by TIDAL in another genome, whether this is due to the true absence of that TE insertion or not enough information to resolve (missing data). These two scenarios (true absence *vs* missing data) can be distinguished using methods developed by Cridland *et al.* (2013). It is worth mentioning that both (Cridland *et al.* 2013; Rahman *et al.* 2015) used PCR to validate subsets of the TE calls and confirm the high accuracy of their methods.

Raw reads of DPGP3 genomes were processed by TrimGalore (https://github.com/FelixKrueger/TrimGalore) to remove adaptors and low-quality sequences. We used TIDAL (Rahman *et al.* 2015) to identify nonreference 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 4th chromosome, which is nearly entirely heterochromatic (Riddle and Elgin 2018). This yielded 39,084 potential TE insertion sites.

We used methods developed in Cridland *et al.* (2013) to call the presence/absence of TEs at potential insertion sites, including the genome in which the TE was identified as an insertion by TIDAL to ensure that alleles in all genomes were called the same way. Briefly, we aligned processed reads to Release 6 *D. melanogaster* reference genome using bwa with default parameters (Li and Durbin 2010). Reads that aligned 500 bp 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 sequence database that include canonical TEs and all TEs annotated in the reference genome (retrieved from Flybase (http://flybase.org/)). 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 or aligned to a different TE family from the TIDAL calls, making the family identity of the insertion could not be determined (see Supplementary Figure S1 for an example). 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. Accordingly, we excluded TEs in or near 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) and Huang *et al.* (2014). TE insertions that are within 1 kb 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 4000 missing TE calls, see Supplementary Table S1). It is worth noting that the sequencing coverage of DPGP3 is uniform across genomes, and we did not find associations between average sequencing coverage and the number of identified TE insertions (Supplementary Figure S2). We further filtered 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). In total, 11,527 polymorphic TEs passed these filtering (see Supplementary Data S1 for TE calls). Following (Sohail *et al.* 2017), we further restricted our analysis to rare TEs 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 Liftover (https://genome.ucsc.edu)] to parse out the coding sequence of the longest isoform and then identified synonymous, nonsynonymous, and premature stop codon (loss-of-function or LoF) 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 or synonymous variants), and SNPs with missing data were excluded from the analysis.

Estimation of σ^2/V_a

For both TEs and SNPs, we restricted the analysis to variants with minor allele counts equal to or smaller than five TEs/SNPs because alleles with counts higher than this are less likely 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). σ^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 (MAF) of locus/TE insertion *i*.

Estimation of mean LD

To estimate LD per pair of TEs (mean LD), we used PLINK (Purcell *et al.* 2007) to compute pairwise correlation coefficients (r^2) and used the following equation to back-calculate LD between pairs of TEs: $D_{i,j} = \sqrt{r_{i,j}{}^2 p_i(1-p_i)p_j(1-p_j)}$, 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. This information is retrieved by using the in-phase option of PLINK. By default, PLINK only reports large estimated LD ($r^2 > 0.2$). We set –ld-window-r2 to 0.00001 (*i.e.*, $r^2 >$

0.00001) in order for PLINK to report a more representative distribution of pairwise LD. We also estimated LD among pairs of TEs on different chromosomes using –inter-chr option. We then estimated the mean LD by binning TE pairs according to their physical distance on the same chromosome (<1 kb, 1–10 kb, 10–100 kb, 100 kb–1 Mb) or as TEs on different chromosomes. For each category/bin of TEs, we required at least five pairs of TEs that have PLINK-calculated r^2 to be included in the analysis.

Annotations of TE insertions

Synergistic epistasis, if present, should more likely be observed among TE insertions with large fitness effects. Accordingly, 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 coding sequences, UTRs, and introns, and inferred their distance to the nearest gene. To categorize TEs according to evolutionary constraints of their nearest genes, we estimated *dN/dS* ratios along the *D. melanogaster* linage 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 mean LD 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 100 bp were excluded from the estimation. We also compared TE families for their propensity to be targeted by piRNAs and to exert epigenetic effects. For indexes of TE-mediated epigenetic effects (the proportion of TEs with the effect, the extent and magnitude of the effect), we used estimates from Lee and Karpen (2017). We used small RNA data of two Zimbabwe strains (ZW155 and ZW184) from Luo *et al.* (2020) and followed methods described in the study to identify piRNA reads. The amount of piRNAs corresponding to a TE family (estimated as per million TE reads) and ping-pong fractions were estimated according to Kelleher and Barbash (2013).

Results

We first identified possible TE insertion positions in the Zambian genomes and then determined the presence/absence status of these TEs in individual genomes. In addition to removing TEs with ambiguous family identity or high rates of missing data, we filtered TEs in cosmopolitan inversions or heterochromatin, whose suppressed recombination could bias the distribution of LD among variants even in the absence of epistasis (see *Materials and Methods*). Overall, we identified 11,527 polymorphic TEs in the euchromatic regions of the genome. Consistent with strong selection acting against TE insertions, identified TEs have a frequency spectrum that is highly skewed toward rare variants (Figure 1A). This frequency spectrum for TE insertions is more skewed than that of SNPs in the same genomes, even when compared to SNPs 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 (σ^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 interdependency between the fitness effects of mutations, and purifying selection removing them will result 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 variants. 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. For all euchromatic TE insertions, we found an overdispersed distribution of TE burden ($\sigma^2/V_a = 2.23$, mean LD per pair of TEs $= 1.53 \times 10^{-6}$, Figure 1B). Yet, even for synonymous variants, which are putatively neutral and should have no epistatic interactions, we also found an overdispersed distribution of mutation burden ($\sigma^2/V_a = 7.13$, mean LD per pair of loci $= 2.98 \times 10^{-6}$). This is similar to previously observed overdispersion of synonymous mutational burden using the same population genomes (Sohail *et al.* 2017) and could result from positive LD generated by an unknown demographic history of the population or other yet-to-be-identified sources.

If demographic processes overdisperse the mutational burden of neutral and selected variants similarly, significantly smaller σ^2/V of selected variants than that of bootstrapped neutral variants could support the presence of synergistic epistasis among selected variants. Yet, simulations have found that, even in the absence of epistasis (*i.e.*, multiplicative fitness effects), the build-up of positive LD driven by demographic processes is more severe

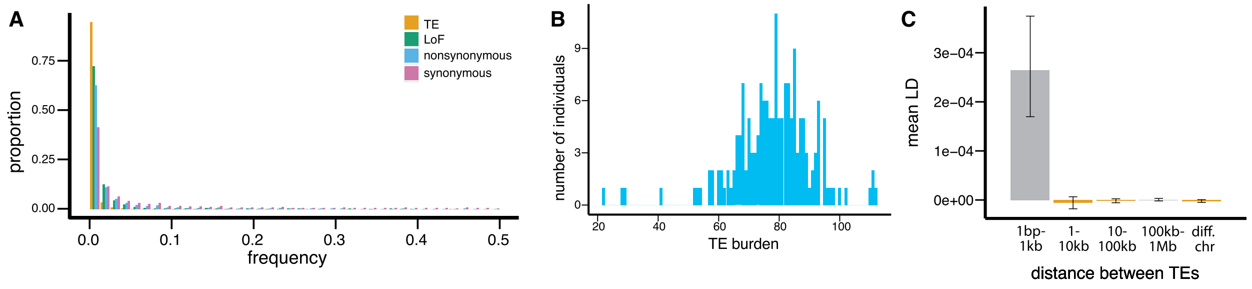

Figure 1 Distributions of allele frequencies and pairwise LD of TEs. (A) Frequency spectra of all the TEs that passed filtering and other SNP variants [loss-of-function (LoF), nonsynonymous, and synonymous]. All the alleles, irrespective of MAF, are included. (B) The distribution of TE burden (TEs with MAF $\leq$ 5) among 173 genomes. (C) Mean LD among pairs of TEs that are of different physical distance (1 bp–1 kb, 1–10 kb, 10–100 kb, 100 kb–1 Mb) or on different chromosomes. Positive mean LD is shown in gray, while negative mean LD is shown in orange. Only TEs with MAF smaller than five are included. Error bars are 95% confidence intervals.

for less selected sites (Sohail *et al.* 2017; Sandler *et al.* 2021). In other words, even in the absence of synergistic epistasis, demographic processes alone could generate less positive LD of selected sites than that of neutral sites.

Alternatively, one could use permutation tests that randomize allele associations to investigate if the observed mean LD of the selected variants is significantly negative (*i.e.*, smaller than zero, Sandler *et al.* 2021). However, as shown by simulations, even with strong selection, demographic processes can still generate appreciable amounts of positive LD among selected sites (Sohail *et al.* 2017; Sandler *et al.* 2021), making this test conservative. In addition, our dataset have a large number of TEs with missing data in at least one genome (99.61%). Missing data are expected to inflate the estimated variance of mutational burden (Sohail *et al.* 2017). In our dataset, the distribution of the number of missing data per genome has a highly skewed L-shape (Supplementary Figure S3), which could have led to the excess of individuals with few TE insertions and overdispersed TE burden (Figure 1B). Indeed, we found a significant negative correlation between TE burden and the amount of missing data per genome ($\rho = -0.32$, $P < 10^{-4}$, Supplementary Figure S4). Accordingly, using permutation to investigate if the mean LD of TEs is significantly negative is not ideal for our dataset either. It is important to note that 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.

Still another possible approach to test for the presence of synergistic epistasis is by investigating the distribution of LD among variants that are of different genetic distances. LD generated by demographic processes is expected to persist over a short genetic distance, given that recombination constantly breaks up associations among alleles (Sandler *et al.* 2021). On the other hand, LD generated by purifying selection with epistasis could persist over a long genetic distance (Ragsdale 2021). In addition, besides 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). In either case, negative LD among variants that are of a large genetic distance would provide stronger support for the presence of synergistic epistasis than that of between nearby variants.

To investigate the mode of epistatic fitness effects of TEs, we calculated mean LD among TE pairs that are on different chromosomes or of different physical distances (1 bp–1 kb, 1–10 kb, 10–100 kb, 100 kb–1 Mb), which serve as proxies for genetic distances. We then estimated the associated 95% confidence interval of mean LD and considered TEs whose confidence interval is entirely negative to show evidence for synergistic epistasis. We focused on rare TEs (present in at most five individuals) to maximize the potential deleterious fitness impacts of TEs. Nevertheless, results based on other TE frequency cutoffs gave consistent results (Supplementary Table S2).

We found very strong positive LD for TE pairs that are within 1 kb, and the associated 95% confidence interval suggests that the observed positive LD is significantly different from zero (Figure 1C). This echoes previously observed strong positive LD among nearby nonsense and missense mutations (Ragsdale 2021; Sandler *et al.* 2021), which could have been generated by demographic processes. On the other hand, pairs of TEs that are 1–10 kb apart, 10–100 kb apart, or on different chromosomes have negative mean LD, although their 95% confidence intervals overlap with zero (Figure 1C, Supplementary Table S2). We further categorized TEs according to their class (DNA or RNA) and type (TIR, non-LTR, or LTR). Again, mean LD for most categories of TEs that are within 1 kb is significantly positive (DNA, RNA, TIR, and LTR, Figure 2A, Supplementary Table S2). On the other hand, for RNA and LTR TEs, pairs of TEs that are on different chromosomes exhibit significant negative mean LD (Figure 2A, Supplementary Table S2). Given that associations between alleles on different chromosomes should quickly break up each generation, this observation suggests strong purifying selection on synergistic fitness effects of RNA and LTR TEs. Overall, we observed that, similar to previous studies on other types of variants, the distribution of LD among pairs of TEs strongly depends on their physical distance. We identified significant negative LD among physically unlinked RNA and LTR TEs, an observation that is consistent with the presence of synergistic epistasis among these TEs.

Physically distant TEs that likely have large fitness effects exhibit significant negative mean LD

In the presence of synergistic epistasis, stronger purifying selection is expected to generate more negative LD among deleterious variants than weaker selection (Sohail *et al.* 2017; Ragsdale 2021). Accordingly, if present, repulsion LD should more likely be identified with TEs that exert large harmful fitness impacts. We thus categorized TEs according to their potential fitness effects and examined their distributions of LD separately. By classifying TEs according to their insertion locations, we found that TEs inside coding sequences have significant negative mean LD for pairs of TEs that are more than 100 kb apart or on different chromosomes (Figure 2B, Supplementary Table S2). This is consistent with the expectation that TEs inserted into coding sequences could

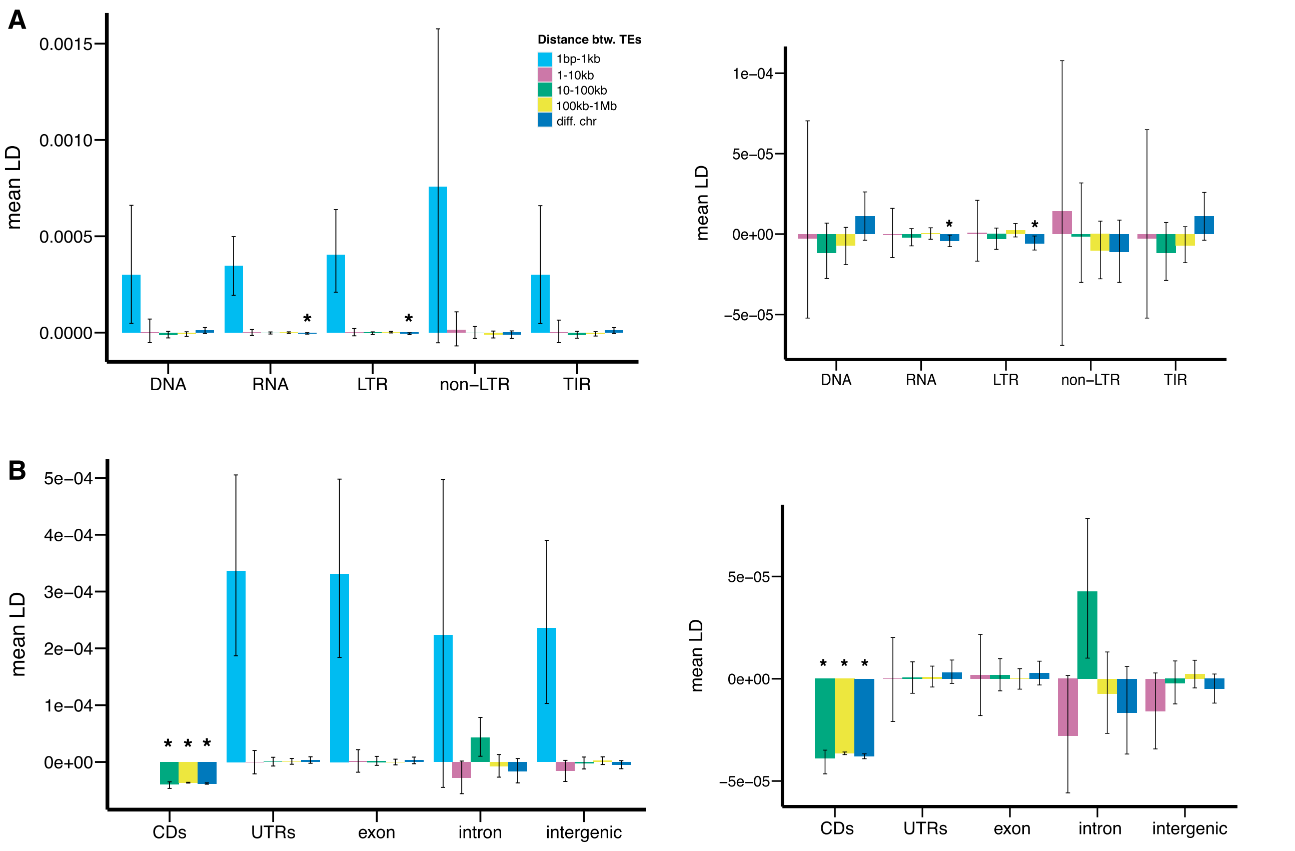

Figure 2 The distributions of LD for different categories of TEs. Mean LD among different classes and types of TEs (A) or TEs inserting at different genomic locations (B). Figures on the right (A,B) excluded nearby TE bins to show the full range of mean LD for physically distant TEs. Error bars are 95% confidence intervals and * denotes 95% confidence intervals that are entirely negative. For TEs in coding sequences (CDs), there are not enough TE pairs for 1 bp–1 kb and 1–10 kb bins (see *Materials and Methods*). Exonic TEs include both TEs in coding sequences and UTRs.

abolish gene function (Bellen *et al.* 2004, 2011) and support the presence of synergistic epistasis among these TEs.

We also categorized TEs according to the evolutionary constraints of their nearest gene. Genes with low ratios of nonsynonymous to synonymous substitution rates, or *dN/dS* ratios, are evolutionarily conserved and generally expected to have essential functions (Larracuente *et al.* 2008; Waterhouse *et al.* 2011). Therefore, TEs in or near these genes could potentially incur large fitness costs. Consistently, we found significant negative mean LD among 1–10 kb TEs whose nearest genes have the second-lowest quartile of *dN/dS* ratio (Figure 3A, Supplementary Table S2). The potentially deleterious effects of TEs could also be inferred from the known mutant phenotypes of their nearest genes [lethal, semi-lethal, inviable (lethal and semi-lethal), and viable (see *Materials and Methods*)]. Similar to analysis based on *dN/dS* ratios, TEs whose nearest genes have known semi-lethal (1–10 kb bin) or inviable (100 kb–1 Mb bin) mutant phenotypes exhibit significant negative mean LD (Figure 3B, Supplementary Table S2). By further restricting to TE insertions within exons, we found significant repulsion LD with more bins of TEs whose nearest genes have known inviable mutant phenotypes (lethal: 100 kb–1 Mb bin; semi-lethal: 1–10 kb, 10–100 kb bins; inviable: 100 kb–1 Mb bin, Figure 3D, Supplementary Table S2). Intriguingly, similar analyses based on *dN/dS* ratio found that mainly exonic TEs whose nearest genes have the largest quartile of *dN/dS* ratio, or being least constrained, have significant negative LD (Figure 3C, Supplementary Table S2, also see Discussions).

In addition to TEs inserting into and disrupting exonic sequences, intergenic TEs could impair host fitness either by disrupting regulatory sequences or through the spreading of repressive epigenetic marks to nearby functional sequences (reviewed in Choi and Lee 2020; Kelleher *et al.* 2020). These effects are similarly expected to result in higher fitness costs when TEs are near essential genes. Consistent with this prediction, physically distant, intergenic TEs show significant negative mean LD if their nearest genes have known inviable mutant phenotypes (for lethal, semi-lethal, and inviable: 1–10 kb and 10–100 kb bins, Figure 3F, Supplementary Table S2). On the other hand, for analyses based on *dN/dS* ratio, we observed negative mean LD among intergenic TEs near genes with all four quartiles of *dN/dS* ratios, instead of mainly among TEs near evolutionarily constrained genes (Figure 3E, Supplementary Table S2, see Discussions). Overall, our analyses that compare the distributions of LD among TEs of different physical distances found significant repulsion LD for TEs expected to have large fitness impacts, especially those inside coding sequences or near genes with known inviable mutant phenotypes.

Due to the low frequencies of TEs included in the analysis, LD among TEs would have a highly skewed distribution, such that most of the observed LD will be negative with occasional positive values. Therefore, when the mean LD is estimated from few pairs of TEs, it is not unlikely to wrongly infer a slightly negative value even in the absence of LD. To evaluate the robustness of our findings concerning this issue, we identified significant negative

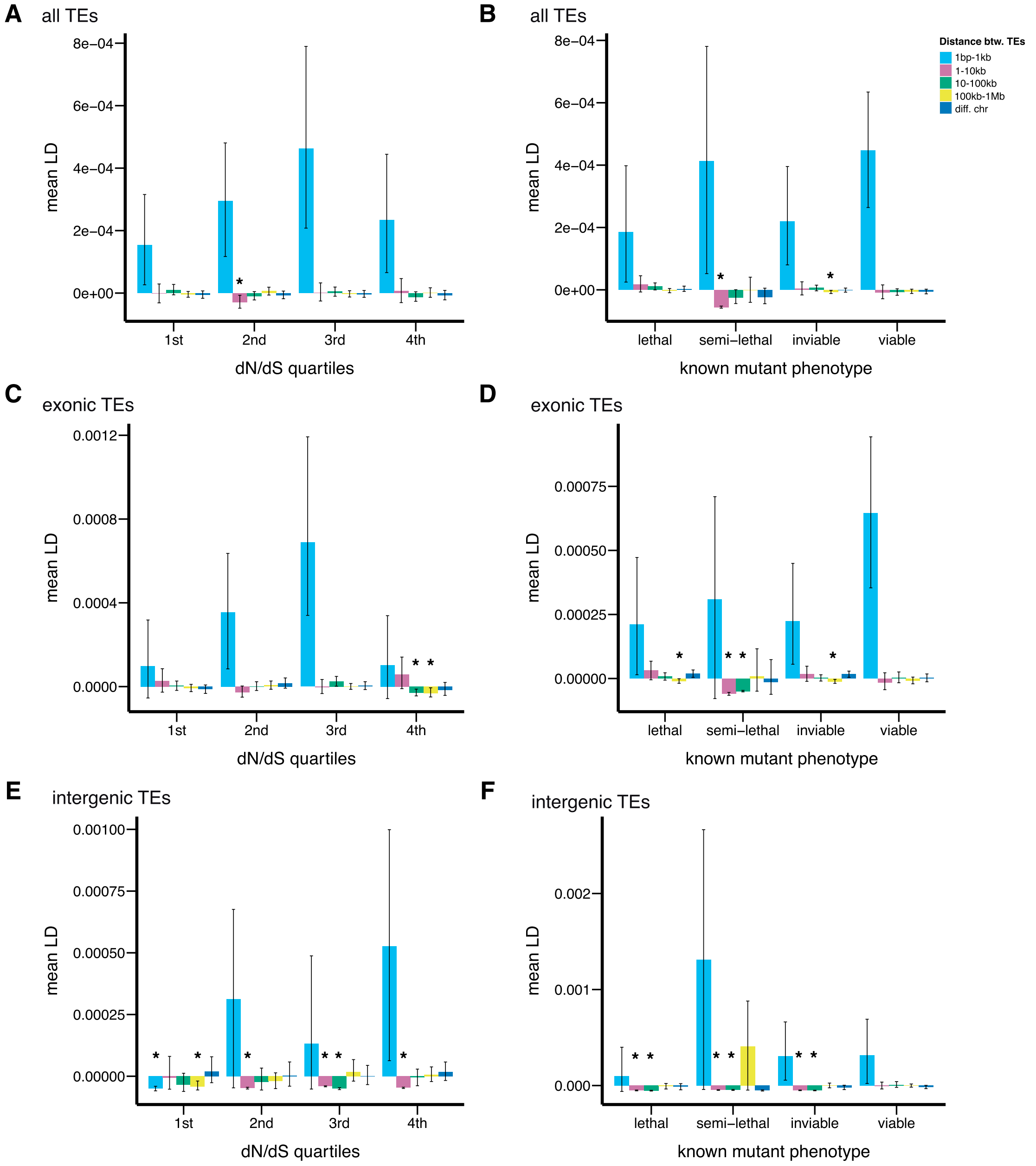

Figure 3 Distributions of LD of TEs near genes with varying evolutionary constraints and essentialities. Mean LD among TEs whose nearest genes have different quartiles of *dN/dS* ratios (A,C,E) or different known mutant phenotypes (B,D,F). Mean LD was estimated for all TEs (A,B), exonic TEs (C,D), and intergenic TEs (E,F). Error bars are 95% confidence intervals and * denotes 95% confidence intervals that are entirely negative.

mean LD that was inferred from fewer than a hundred pairs of TEs (Supplementary Table S2). While LD for several categories of TEs that are within 1 kb was based on few TE pairs, its value is strongly positive and thus unlikely to be influenced by the issue raised (Figure 2 and Supplementary Table S2). On the other hand, for all distance bins, fewer than a hundred TE pairs were included in the analyses for TEs in coding sequences or intergenic TEs near genes with known semi-lethal phenotype (Supplementary Table S2), which could have contributed to their observed negative mean LD. Nevertheless, for intergenic TEs whose nearest genes have known lethal or inviable (lethal and semi-lethal) phenotypes, their significant negative mean LD was estimated from more than a hundred TE pairs, providing robust support for the presence of synergistic epistasis among intergenic TEs that likely have large fitness effects.

Mean LD of physically distant TEs is significantly negative for many TE families

Both ectopic recombination and epigenetic effects of TEs depend on sequence homology among TE insertions. Mainly copies of the same TE family ectopically recombine, and small RNAs generated from a particular TE family are most 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 investigated whether TEs of individual families exhibit repulsion LD. Specifically, given the predominantly positive LD among nearby TEs (Figures 1C, 2, and 3, Supplementary Table S2), we estimated the mean LD for each TE family, excluding pairs of TEs that are within 1 kb. Following our other analysis, mean LD whose 95% confidence interval is entirely smaller than zero is considered significantly negative.

We found that slightly more than half (55.6%) of TE families (out of 37) have significant negative LD (Supplementary Table S3). This proportion is higher when restricting to TE insertions whose nearest genes have the lowest quartile of *dN/dS* ratio (73%) or have known lethal or inviable mutant phenotypes [80.8% (lethal) and 75.9% (inviable), Supplementary Table S3]. This observation suggests that there may be a mixture of fitness effects among copies of the same TE families, and repulsion LD is more readily detectable among those likely to exert large fitness effects.

Mixed support for the possible source of synergism among the deleterious effects of TEs

To infer the possible source of synergism among the deleterious fitness effects of TEs, we tested several predictions of the ectopic recombination and epigenetic effects models about which categories or families of TEs are more likely to exert synergistic epistasis and thus exhibit repulsion LD. Assuming that the rates of ectopic recombination closely follow that of homologous recombination (Lichten *et al.* 1987), TEs in high recombining regions of the genome should be prone to be involved in ectopic recombination, and if present, synergistic epistasis should more likely be observed among TEs in those genomic regions. Yet, we only observed significant negative mean LD among TEs that are in the lowest quartile of recombination rates (Figure 4A, Supplementary Table S2), which fails to support the prediction of the ectopic recombination model.

On a family level, 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 be more likely to exert synergistic epistasis (Lee and Langley 2010).

To test these predictions, we compared copy numbers, length, and within-family sequence identities between TE families with and without significant negative mean LD. We categorized TE families according to mean LD estimated from all physically distant (>1 kb) TE insertions and physically distant insertions near genes with known inviable mutant phenotypes, which are more likely to exert deleterious fitness effects (see above). 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 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*). For both estimates of TE copy number, TE families with significant negative mean LD have *fewer* TEs than other families, which is the opposite of the predicted direction [Mann–Whitney test, $P = 0.0169$ (euchromatic TE copy number in the reference genome) and 6.9×10^{-10} (TE copy number estimated in our dataset), Table 1, Figure 4B, and Supplementary Figure S5]. When categorizing families according to estimated mean LD among TEs near genes with inviable mutant phenotypes, we observed a similar trend (Figure 4C and Supplementary Figure S5). Yet, it is worth noting that rare TE families are also those that have fewer pairs of TEs included in the LD estimation (Supplementary Table S3). The small number of TE pairs could lead to misleadingly negative LD estimates (see above), resulting in the observed positive associations between TE family abundance and mean LD. While there is

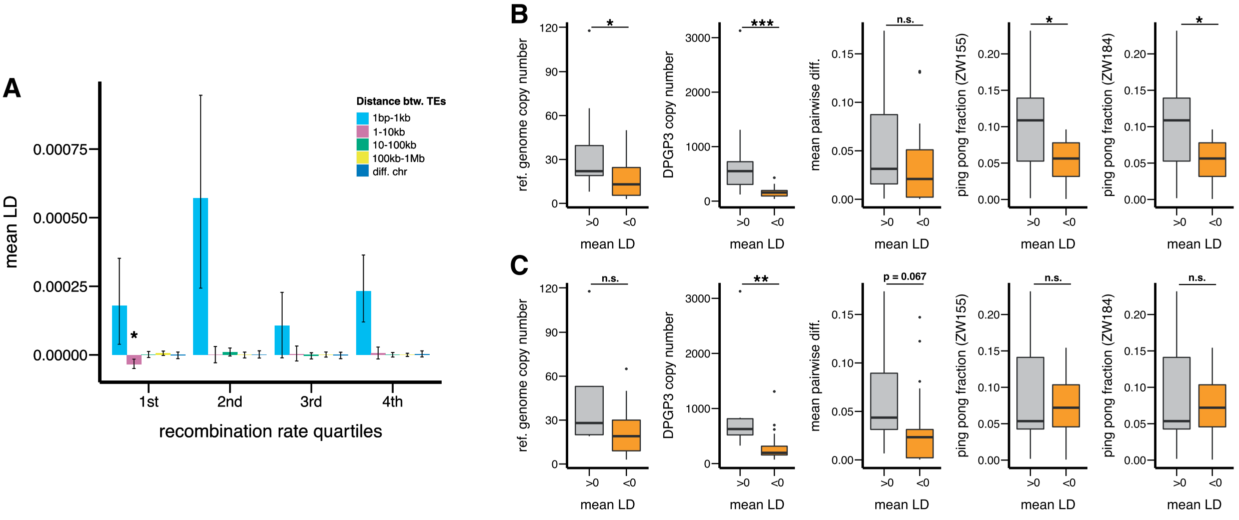

Figure 4 Identification of the possible source of TEs' synergistic deleterious effects. (A) Distribution of LD among TEs in genomic regions with different rates of recombination. Error bars are 95% confidence intervals and * denotes 95% confidence intervals that are entirely negative. (B,C) Comparisons of different attributes of TE families (copy number, mean pairwise difference, and ping-pong fraction) with and without mean negative LD estimated from all TEs (B) or TEs whose nearest genes have known inviable mutant phenotypes (C). Only TE family attributes that have significant differences are shown in this figure, while the distributions of other attributes are shown in Supplementary Figures S5–S7. Mann–Whitney *U*-test, $^{*}P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. n.s. denotes statistically insignificant.

Table 1 Comparisons of biological properties of TE families with and without negative mean LD

	All TEs		Inviable phenotype[a]	
	Mann–Whitney *U*-test, *P*-value	Direction	Mann–Whitney *U*-test, *P*-value	Direction
TE family copy number				
Reference genome	**1.69E−02**	**Negative LD lower**	1.04E−01	
DPGP3 genomes	**6.90E−05**	**Negative LD lower**	2.04E−03	
TE family average length				
Reference genome	8.51E−01		7.31E−01	
TE family mean pairwise difference				
Reference genome	1.56E−01		**6.74E−02**	**Negative LD lower**
TE family epigenetic effects				
Proportion of TEs with epi. effect	2.81E−01		1.56E−01	
Median extent of epi. effect	4.71E−01		8.64E−01	
Median increase of epi. effect	3.23E−01		4.92E−01	
piRNAs correspond to a TE family				
Sense piRNA (strain ZW155)	6.59E−01		2.96E−01	
Sense piRNA (strain ZW184)	7.56E−01		7.56E−01	
Anti-sense piRNA (strain ZW155)	3.33E−01		2.72E−01	
Anti-sense piRNA (strain ZW184)	3.85E−01		3.48E−01	
Ping-pong fraction (strain ZW155)	**3.14E−02**	**Negative LD lower**	8.77E−01	
Ping-pong fraction (strain ZW184)	**1.65E−02**	**Negative LD lower**	9.59E−01	

[a] TEs whose nearest gene have known inviable mutant phenotype. These are p-values and those mentioned in the text are in bold.

no difference between TE families with and without negative mean LD in terms of average TE length (Mann–Whitney *U*-test, $P > 0.05$, Table 1), TE families with significant negative mean LD have lower within-family sequence divergence (for TEs near genes with inviable mutant phenotypes, Mann–Whitney *test*, $P = 0.067$, Table 1, Figure 4C, and Supplementary Figure S5). Overall, by comparing attributes of TE families with and without negative mean LD, we found mixed support for the common predictions of the ectopic recombination and epigenetic effect models about which TE families are more likely to exert synergistic fitness effects.

In addition to predictions 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. In addition, 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), the "ping-pong cycle" is thought to be responsible for the majority 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 (Brennecke *et al.* 2007; Gunawardane *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 the ping-pong cycle in the generation of piRNA significantly varies between TE families (Li *et al.* 2009; Malone *et al.* 2009; Kelleher and Barbash 2013). Synergism is thus expected to have a higher tendency to arise for TE families targeted by more piRNAs generated via the ping-pong cycle.

To test these predictions, for each TE family, we used previously reported indexes for the strength of epigenetic effects (Lee and Karpen 2017) and estimated the amount of corresponding piRNAs and ping-pong fractions using ovarian small RNA sequences from two Zimbabwe strains (Luo *et al.* 2020, see *Materials and Methods*). For all three indexes of TEs' epigenetic effects (the proportion of TEs resulting in *cis* spreading of repressive marks, the median extent of this spreading, and the median magnitude of TE-induced increased enrichment of repressive marks), we observed no difference between TE families with and without negative mean LD (Mann–Whitney *U*-test, $P > 0.05$ for all comparisons, Table 1, also see Supplementary Figure S6). Similarly, there are no significant differences in the abundance of sense and anti-sense piRNAs targeting TE families with and without negative mean LD (Mann–Whitney *U*-test, $P > 0.05$ for all comparisons, Table 1, also see Supplementary Figure S7). Intriguingly, contrary to prediction, TE families with negative mean LD have significantly *lower* ping-pong fraction (Mann–Whitney *U*-test, $P = 0.0314$ (ZW155) and 0.0165 (ZW184), Figure 4B, Table 1). Yet, when estimating mean LD from TEs whose nearest genes have inviable mutant phenotypes and are thus more likely to have deleterious fitness effects, we observed the opposite, though statistically insignificant, trend (*i.e.*, TE families with negative mean LD have *larger* ping-pong fractions than other TE families, Mann–Whitney *U*-test, $P > 0.05$ for both strains, Figure 4C, Table 1).

It is worth noting that TE families differ in the average deleteriousness of their insertions (Charlesworth and Langley 1989; Barrón *et al.* 2014), which could be inferred from varying TE population frequencies (*e.g.*, Cridland *et al.* 2013; Kofler *et al.* 2015). The overdispersing effect of demographic processes is expected to influence less selected variants more strongly. If our analysis fails to completely exclude the impacts of recent demographic events, the mean LD of less deleterious TE families would be prone to be inflated and less probable to be observed negative even in the presence of synergistic epistasis. Accordingly, associations between the deleteriousness and various attributes of TE families could also drive patterns observed in Figure 4, B and C. To exclude this possibility, we categorized TE families according to attributes that varied between TE families with and without negative LD and compared their population frequencies. We found no significant differences in population frequencies between TE families with high/low copy numbers in the reference genome or large/small within-family sequence divergence (Mann–Whitney *U*-test *and* Student's *t*-test, $P > 0.05$ for all comparisons, Supplementary Table S4). Also, TE families with low DPGP3 TE copy numbers, which were observed to have negative mean LD, have marginally significantly *higher* TE frequencies (Mann–Whitney *U*-test, $P = 0.046$, Supplementary Table S4). Thus, these TE families should have *lower* deleteriousness of TE insertions, which would not have confounded our inference. On the contrary, TE families with larger ping-pong fractions also have significantly *higher* population frequencies (Mann–Whitney *U*-test and Student's *t*-test, $P < 10^{-8}$ for all comparisons, Supplementary Table S4). Insertions of these TE families are expected to have lower average deleteriousness, and their mean LD is prone to be inflated by demographic processes. We are thus unable to exclude the possibilities that our observed associations between mean negative LD and lower ping pong fractions across TE families are driven by confounding factors unrelated to the mode of TE epistatic interactions. In short, our comparisons of the epigenetic effects and piRNA targeting of TE families with and without negative mean LD do not support the predictions of the epigenetic effects model.

Discussion

Theoretical analysis has predicted that, to stably contain the selfish increase of TEs, each additional TE insertion needs to impose a larger 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; Choi and Lee 2020; Kelleher *et al.* 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 (Charlesworth 1990; Kondrashov 1995; Sohail *et al.* 2017). By leveraging this population genetic signal and examining the distribution of LD among TE insertions with different physical distances, we investigated the predicted synergistic epistasis among potentially deleterious TE insertions in the likely ancestral population of *D. melanogaster*. Pairs of TEs that are within 1 kb tend to have significantly positive mean LD, which excluded the potential role of selective interference (Hill and Robertson 1966; Felsenstein 1974) in shaping the distribution of LD among TEs and could have been generated by recent demographic processes (Sandler *et al.* 2021). On the other hand, we observed significant negative mean LD among physically distant TEs (>1 kb apart), especially those who likely exert large fitness effects. Even more, for some categories of TEs, we observed significant negative mean LD among TE insertions on different chromosomes, where meiosis breaks up linkage among variants every generation. These observations provide strong empirical support for the synergism of the deleterious fitness effects of *D. melanogaster* TEs. It is worth noting that the strong suppression of recombination in heterochromatic regions and chromosomal inversions is expected to generate extensive LD even in the absence of epistatic fitness impacts, and we excluded TEs in these regions from our analysis. Thus, observations made from our studies may not be applicable to TEs in the heterochromatic regions or inside inversions.

The negative LD generated from purifying selection with synergistic epistasis should mainly be observed among TE insertions exerting harmful fitness effects. Consistently, we observed prevalent negative mean LD among TEs inside coding sequences or in/near genes that are evolutionarily constrained or have known inviable mutant phenotypes. Interestingly, negative mean LD among TEs in/near genes with known inviable phenotypes is observed for both genic and intergenic TEs. The latter could be driven by TE deleterious mechanisms that impair host fitness from a distance to genes, such as insertion into regulatory elements or through the spreading of repressive epigenetic marks (reviewed in Choi and Lee 2020; Kelleher *et al.* 2020). On the contrary, when categorizing intergenic TEs according to the *dN/dS* ratio of their nearest genes, we observed negative mean LD among TEs near genes with both low and high *dN/dS* ratios. While this observation seems paradoxical at first glance, fast-evolving genes are recently found to be crucial to organismal survival and play important roles in essential functions (Chen *et al.* 2010; Lee *et al.* 2019; Xia *et al.* 2021).

Supporting the prediction that synergistic epistasis of TEs arises among copies of the same TE families (Montgomery *et al.* 1987; Langley *et al.* 1988; Lee and Langley 2010; Lee 2015), we found that more than half of the TE families have significant negative mean LD. Even more, this proportion increases when restricted to TEs that likely exert large fitness effects. Intriguingly, if purifying selection against synergistic deleterious TEs is key to the stable containment of TE copy numbers, why did we fail to find significant negative LD in some families? Synergistic epistasis should only occur among TEs that impair host fitness. It is plausible that not all the identified TEs are deleterious. Accordingly, including all insertions in the analysis could have weakened our ability to detect negative mean LD among subsets of TEs that are truly deleterious. This conjecture is supported by the observation that some TE families only exhibit significant negative LD when restricted to insertions that likely have large fitness effects (Supplementary Table S2). Also, our analysis could only identify the locations, but not the internal sequences, of TE insertions. Some of the identified TEs could have degenerated and are no longer involved in the population dynamics of its family. Alternatively, the containment of TE copy number could happen above the family levels (see below).

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 prone to be involved in ectopic recombination should be more likely to exhibit synergistic epistasis (Langley *et al.* 1988). While we did not find evidence supporting that TEs in genomic regions with high rates of meiotic recombination are more likely to have significant negative LD, 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 based on 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). Furthermore, we excluded TEs in genomic regions with suppressed recombination, which removes the largest axis of variation in recombination rates and could have limited the power of our analysis.

We also tested other predictions of the ectopic recombination model about which TE families are more likely to exert synergistic epistatic effects, 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). While TE families with significant repulsion LD have higher sequence identity within families (consistent with prediction), they also have fewer copies (opposite to prediction). The observed mixed support for the ectopic recombination model could result from that the estimates of TE family properties used in our study (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 should 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 by which synergistic epistasis was predicted to arise (Lee and Langley 2010; Lee 2015). In *Drosophila*, piRNA targeting initiates the 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 shared with the ectopic recombination model—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 exert synergistic epistasis, which we failed to find support for. We also did not find support for the unique predictions of the epigenetic effects model that TE families with stronger epigenetic effects or targeted by more piRNAs are more likely to exhibit synergistic fitness effects. Again, estimates for the strength of epigenetic effects from a North American population (Lee and Karpen 2017) may not be representative of the focused African population. Similarly, due to the strong population structure (Coughlan *et al.* 2021), piRNA indexes calculated from the Zimbabwe population may differ from those of the Zambia population. Importantly, the complexities of piRNA generation and targeting that are not considered in the current epigenetic effect model could also have led to discrepancies between predictions and observations. For instance, truncated TEs that lost the ability to transcribe would not contribute to piRNA generations through the ping-pong cycle (Sienski *et al.* 2012; Olovnikov *et al.* 2013; Shpiz *et al.* 2014). In addition, targeting of TEs by piRNAs is particularly sensitive to mismatches at specific positions within the piRNA sequences (Wang *et al.* 2014; Mohn *et al.* 2015). Accordingly, simple monotonic relationships could not fully capture how the copy number, length, sequence homology, and piRNA biology of TE families influence the occurrence of piRNA-targeting of TEs and the associated deleterious epigenetic effects.

Still another possibility for why we did not observe unequivocal support for the predictions about which TE families are more likely to show synergism is—such epistatic fitness effects may arise above the family level. Consistently, significant negative mean LD was observed among insertions of different TE families (Figures 1–3, Supplementary Table S2). Furthermore, several deleterious effects of TEs, such as disruptions of coding sequences (Bellen *et al.* 2004, 2011) or altering gene expression through inserting into regulatory sequences (Chuong *et al.* 2017), could be functionally equivalent to that of deleterious SNPs. Recently, synergistic fitness effects have been inferred for loss-of-function (Sohail *et al.* 2017) and missense (Sandler *et al.* 2021) SNPs. Though the underlying molecular details for these identified epistatic interactions among SNPs are still unclear, TEs may exert synergistic epistasis through similar mechanisms irrespective of their family identity. Indeed, negative mean LD among TEs is of a similar order of magnitude to that of loss-of-function SNPs [the most negative mean LD of TEs in Supplementary Table S2—mean: -5.96×10^{-5}, 95% CI: $(-6.54 \times 10^{-5}, -5.14 \times 10^{-5})$ *vs* 1-10kb bin of LoF has significant negative mean LD—mean: -1.08×10^{-4}, 95% CI: $(-1.27 \times 10^{-4}, -9.16 \times 10^{-5})$] and radical missense SNPs in biological networks (Sandler *et al.* 2021). Such observation suggests a similar extent of synergistic fitness effects between TEs and deleterious SNPs. The containment of TEs could thus depend on synergistic epistasis both within and between TE families, or even between TEs and other types of deleterious variants. Studies of the population dynamics of TEs may need to go beyond the usually presumed "within-family" regulation and jointly consider other TE families as well as broader genomic contexts.

It is worth noting that the statistical power for some of our current analyses may be limited due to challenges associated with studying TEs. For example, TEs have a frequency spectrum that is highly skewed toward rare variants (Supplementary 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 restricted with short-read sequencing data. Some of these limitations may be alleviated in the near future with the growing number of genomes sequenced by 3rd-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. Our mixed support for the predictions of ectopic recombination and epigenetic effect models suggests a need to incorporate additional biological details to refine the models for how synergistic fitness effects of TEs may arise within and, perhaps, between TE families. 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.

Data availability

The presence and absence status of 11,396 TEs included in the analysis could be found in Supplementary Data S1.

Supplementary material is available at *GENETICS* online.

Acknowledgments

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 Thoronton provided helpful comments on the earlier version of the manuscript. I also appreciate Andrea Betancourt's helpful editing of the manuscript. Three reviewers provided helpful suggestions that significantly improved the study.

Funding

This work was supported by NIH R00 GM121868 and NIH R35GM142494.

Conflicts of interest

The authors declare that there is no conflict of interest.

Literature cited

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S. 2017. Genome-wide estimates of transposable element insertion and deletion rates in *Drosophila melanogaster*. Genome Biol Evol. 9: 1329–1340. doi:10.1093/gbe/evx050.

Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, *et al.* 2008. A piRNA pathway primed by individual transposons is linked to *de novo* DNA methylation in mice. Mol Cell. 31:785–799. doi:10.1016/j.molcel.2008.09.003.

Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ. 2007. Developmentally regulated piRNA clusters implicate MILI in transposon control. Science. 316:744–747. doi:10.1126/science.1142612.

Barrón MG, Fiston-Lavier A-S, Petrov DA, González J. 2014. Population genomics of transposable elements in Drosophila. Annu Rev Genet. 48:561–581. doi:10.1146/annurev-genet-120213-092359.

Barton NH. 1995. A general model for the evolution of recombination. Genet Res. 65:123–144. doi:10.1017/S0016672300033140.

Bellen HJ, Levis RW, He Y, Carlson JW, Evans-Holm M, *et al.* 2011. The Drosophila gene disruption project: progress using transposons with distinctive site specificities. Genetics. 188:731–743. doi: 10.1534/genetics.111.126995.

Bellen HJ, Levis RW, Liao G, He Y, Carlson JW, *et al.* 2004. The BDGP gene disruption project: single transposon insertions associated with 40% of Drosophila genes. Genetics. 167:761–781. doi: 10.1534/genetics.104.026427.

Brennecke J, Aravin AA, Stark A, Dus M, Kellis M, *et al.* 2007. Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. Cell. 128:1089–1103. doi:10.1016/j.cell.2007.01.043.

Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI. 2011. Correlated evolution of nearby residues in Drosophilid proteins. PLoS Genet. 7:e1001315.doi:10.1371/journal.pgen.1001315.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, *et al.* 2009. BLAST+: architecture and applications. BMC Bioinformatics. 10: 421.doi:10.1186/1471-2105-10-421.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD. 2019. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. Nat Commun. 10:1–11. doi:10.1038/s41467-019-12884-1.

Charlesworth B. 1990. Mutation-selection balance and the evolutionary advantage of sex and recombination. Genet Res. 55:199–221. doi:10.1017/s0016672300025532.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genet Res. 42:1–27. doi:10.1017/S0016672300021455.

Charlesworth B, Langley CH. 1989. The population genetics of Drosophila transposable elements. Annu Rev Genet. 23:251–287. doi:10.1146/annurev.ge.23.120189.001343.

Charlesworth B, Lapid A. 1989. A study of ten families of transposable elements on X chromosomes from a population of *Drosophila melanogaster*. Genet Res. 54:113–125.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. Science. 330:1682–1685. doi:10.1126/science.1196380.

Choi JY, Lee YCG. 2020. Double-edged sword: the evolutionary consequences of the epigenetic silencing of transposable elements. PLoS Genet. 16:e1008872.doi:10.1371/journal.pgen.1008872.

Chuong EB, Elde NC, Feschotte C. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nat Rev Genet. 18: 71–86. doi:10.1038/nrg.2016.139.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, *et al.* 2007. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 450:203–218. doi:10.1038/nature06341.

Comeron JM, Ratnappan R, Bailin S. 2012. The many landscapes of recombination in *Drosophila melanogaster*. PLoS Genet. 8: e1002905.doi:10.1371/journal.pgen.1002905.

Corbett-Detig RB, Hartl DL. 2012. Population genomics of inversion polymorphisms in *Drosophila melanogaster*. PLoS Genet. 8: e1003056.doi:10.1371/journal.pgen.1003056

Coughlan JM, Dagilis AJ, Serrato-Capuchina A, Elias H, Peede D, *et al.* 2021. Patterns of and processes shaping population structure and introgression among recently differentiated *Drosophila melanogaster* populations. doi.org/10.1101/2021.06.25.449842.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. Mol Biol Evol. 30:2311–2327. doi:10.1093/molbev/mst129.

Czech B, Hannon GJ. 2016. One loop to rule them all: the ping-pong cycle and piRNA-guided silencing. Trends Biochem Sci. 41: 324–337. doi:10.1016/j.tibs.2015.12.008.

Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, *et al.* 2018. piRNA-guided genome defense: from biogenesis to silencing. Annu Rev Genet. 52:131–157. doi:10.1146/annurev-genet-120417-031441.

Daborn PJ, Yen JL, Bogwitz MR, Goff GL, Feil E, *et al.* 2002. A single P450 allele associated with insecticide resistance in Drosophila. Science. 297:2253–2256. doi:10.1126/science.1074170.

Davis PS, Shen MW, Judd BH. 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 USA. 84:174–178. doi:10.1073/pnas.84.1.174.

Debladis E, Llauro C, Carpentier M-C, Mirouze M, Panaud O. 2017. Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. BMC Genomics. 18:537.doi:10.1186/s12864-017–3753-z.

Deniz Ö, Frost JM, Branco MR. 2019. Regulation of transposable elements by DNA modifications. Nat Rev Genet. 20:417–431. doi:10.1038/s41576-019-0106-6.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 12:1075–1079. doi:10.1101/gr.132102.

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

Dumont BL, Broman KW, Payseur BA. 2009. Variation in genomic recombination rates among heterogeneous stock mice. Genetics. 182:1345–1349. doi:10.1534/genetics.109.105114.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. doi:10.1093/nar/gkh340.

Ellison CE, Cao W. 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 Res. 48:290–303. doi:10.1093/nar/gkz1080.

Eshel I, Feldman MW. 1970. On the evolutionary effect of recombination. Theor Popul Biol. 1:88–100. doi:10.1016/0040–5809(70)90043-2.

Ewens WJ, Spielman RS. 1995. The transmission/disequilibrium test: history, subdivision, and admixture. Am J Hum Genet. 57: 455–464.

Ewing B, Hillier L, Wendl MC, Green P. 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.

Garcia JA, Lohmueller KE. 2020. Negative linkage disequilibrium between amino acid changing variants reveals interference among deleterious mutations in the human genome. Evol Biol.

González J, Petrov DA. 2009. The adaptive role of transposable elements in the Drosophila genome. Gene. 448:124–133. doi:10.1016/j.gene.2009.06.008.

Gunawardane LS, Saito K, Nishida KM, Miyoshi K, Kawamura Y, *et al.* 2007. A slicer-mediated mechanism for repeat-associated siRNA 5’ end formation in Drosophila. Science. 315:1587–1590. doi:10.1126/science.1140494.

Hill WG, Robertson A. 1966. The effect of linkage on limits to artificial selection. Genet Res. 8:269–294.

Hof AEV, Campagne P, Rigden DJ, Yung CJ, Lingley J, *et al.* 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature. 534:102–105. doi:10.1038/nature17951.

Hollister JD, Gaut BS. 2009. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res. 19: 1419–1428. doi:10.1101/gr.091678.109.

Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, *et al.* 2015. The Release 6 reference sequence of the *Drosophila melanogaster* genome. Genome Res. 25:445–458. doi:10.1101/gr.185579.114.

Houle D, Nuzhdin SV. 2004. Mutation accumulation and the effect of copia insertions in *Drosophila melanogaster*. Genet Res. 83:7–18. doi:10.1017/s0016672303006505.

Hu TT, Eisen MB, Thornton KR, Andolfatto P. 2013. A second-generation assembly of the Drosophila simulans genome provides new insights into patterns of lineage-specific divergence. Genome Res. 23:89–98. doi:10.1101/gr.141689.112.

Huang W, Massouras A, Inoue Y, Peiffer J, Ràmia M, *et al.* 2014. Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. Genome Res. 24: 1193–1208. doi:10.1101/gr.171546.113.

Hunter CM, Huang W, Mackay TFC, Singh ND. 2016. The genetic architecture of natural variation in recombination rate in *Drosophila melanogaster.* PLoS Genet. 12:e1005951.doi:10.1371/journal.pgen.1005951.

Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, *et al.* 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3:research0084.

Kapitonov VV, Jurka J. 2003. Molecular paleontology of transposable elements in the *Drosophila melanogaster* genome. Proc Natl Acad Sci USA. 100:6569–6574. doi:10.1073/pnas.0732024100.

Kasschau KD, Fahlgren N, Chapman EJ, Sullivan CM, Cumbie JS, *et al.* 2007. Genome-wide profiling and analysis of Arabidopsis siRNAs. PLoS Biol. 5:e57.doi:10.1371/journal.pbio.0050057.

Kelleher ES, Barbash DA. 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. doi:10.1093/molbev/mst081.

Kelleher ES, Barbash DA, Blumenstiel JP. 2020. Taming the turmoil within: new insights on the containment of transposable elements. Trends Genet. 36:474–489. doi:10.1016/j.tig.2020.04.007.

Killick SC, Carlsson AM, West SA, Little TJ. 2006. Testing the pluralist approach to sex: the influence of environment on synergistic interactions between mutation load and parasitism in Daphnia magna. J Evol Biol. 19:1603–1611. doi:10.1111/j.1420–9101.2006.01123.x.

Kishony R, Leibler S. 2003. Environmental stresses can alleviate the average deleterious effect of mutations. J Biol. 2:14.

Kofler R, Nolte V, Schlötterer C. 2015. Tempo and mode of transposable element activity in Drosophila. PLoS Genet. 11:e1005406.doi:10.1371/journal.pgen.1005406.

Kondrashov AS. 1995. Dynamics of unconditionally deleterious mutations: Gaussian approximation and soft selection. Genet Res. 65:113–121. doi:10.1017/S0016672300033139.

Kupiec M, Petes TD. 1988. Allelic and ectopic recombination between Ty elements in yeast. Genetics. 119:549–559.

Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, *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. 199: 1229–1241. doi:10.1534/genetics.115.174664.

Lagemaat L. N V D, Gagnier L, Medstrand P, Mager DL. 2005. Genomic deletions and precise removal of transposable elements mediated by short identical DNA segments in primates. Genome Res. 15:1243–1249. doi:10.1101/gr.3910705.

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

Laricchia KM, Zdraljevic S, Cook DE, Andersen EC. 2017. Natural variation in the distribution and abundance of transposable elements across the *Caenorhabditis elegans* species. Mol Biol Evol. 34: 2187–2202. doi:10.1093/molbev/msx155.

Larracuente AM, Sackton TB, Greenberg AJ, Wong A, Singh ND, *et al.* 2008. Evolution of protein-coding genes in Drosophila. Trends Genet. 24:114–123. doi:10.1016/j.tig.2007.12.001.

Lee YCG. 2015. The role of piRNA-mediated epigenetic silencing in the population dynamics of transposable elements in *Drosophila melanogaster.* PLoS Genet. 11:e1005269.doi:10.1371/journal.pgen.1005269.

Lee YCG, Karpen GH. 2017. Pervasive epigenetic effects of Drosophila euchromatic transposable elements impact their evolution. eLife. 6:e25762. doi:10.7554/eLife.25762.

Lee YCG, Langley CH. 2010. Transposable elements in natural populations of *Drosophila melanogaster*. Philos Trans R Soc Lond B Biol Sci. 365:1219–1228. doi:10.1098/rstb.2009.0318.

Lee YCG, Reinhardt JA. 2012. Widespread polymorphism in the positions of stop codons in *Drosophila melanogaster*. Genome Biol Evol. 4:533–549. doi:10.1093/gbe/evr113.

Lee YCG, Ventura IM, Rice GR, Chen D-Y, Colmenares SU, *et al.* 2019. Rapid evolution of gained essential developmental functions of a young gene via interactions with other essential genes. Mol Biol Evol. 36:2212–2226. doi:10.1093/molbev/msz137.

Le Thomas A, Rogers AK, Webster A, Marinov GK, Liao SE, *et al.* 2013. Piwi induces piRNA-guided transcriptional silencing and establishment of a repressive chromatin state. Genes Dev. 27:390–399. doi:10.1101/gad.209841.112.

Lewontin RC, Kojima K. 1960. The evolutionary dynamics of complex polymorphisms. Evolution. 14:458–472. doi:10.1111/j.1558–5646.1960.tb03113.x.

Li C, Vagin VV, Lee S, Xu J, Ma S, *et al.* 2009. Collapse of germline piRNAs in the absence of Argonaute3 reveals somatic piRNAs in flies. Cell. 137:509–521. doi:10.1016/j.cell.2009.04.027.

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. doi:10.1093/bioinformatics/btr509.

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

Lichten M, Borts RH, Haber JE. 1987. Meiotic gene conversion and crossing over between dispersed homologous sequences occurs frequently in *Saccharomyces cerevisiae*. Genetics. 115:233–246.

Lim JK, Simmons MJ. 1994. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays. 16:269–275. doi:10.1002/bies.950160410.

Luo S, Zhang H, Duan Y, Yao X, Clark AG, *et al.* 2020. The evolutionary arms race between transposable elements and piRNAs in *Drosophila melanogaster*. BMC Evol Biol. 20:14.doi:10.1186/s12862-020–1580-3.

Mackay TF. 1989. Transposable elements and fitness in *Drosophila melanogaster*. Genome. 31:284–295.

Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, *et al.* 2006. Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line. PLoS Genet. 2:e2.doi:10.1371/journal.pgen.0020002.

Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, *et al.* 2009. Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell. 137:522–535. doi:10.1016/j.cell.2009.03.040.

Marí-Ordóñez A, Marchais A, Etcheverry M, Martin A, Colot V, *et al.* 2013. Reconstructing de novo silencing of an active plant retrotransposon. Nat Genet. 45:1029–1039. doi:10.1038/ng.2703.

Maside X, Assimacopoulos S, Charlesworth B. 2000. Rates of movement of transposable elements on the second chromosome of *Drosophila melanogaster*. Genet Res. 75:275–284.

McCue AD, Panda K, Nuthikattu S, Choudury SG, Thomas EN, *et al.* 2015. ARGONAUTE 6 bridges transposable element mRNA-derived siRNAs to the establishment of DNA methylation. EMBO J. 34:20–35. doi:10.15252/embj.201489499.

Mieczkowski PA, Lemoine FJ, Petes TD. 2006. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast *Saccharomyces cerevisiae*. DNA Repair (Amst). 5:1010–1020. doi:10.1016/j.dnarep.2006.05.027.

Mohn F, Handler D, Brennecke J. 2015. piRNA-guided slicing specifies transcripts for Zucchini dependent, phased piRNA biogenesis. Science. 348:812–817. doi:10.1126/science.aaa1039

Montgomery E, Charlesworth B, Langley CH. 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 EA, Huang SM, Langley CH, Judd BH. 1991. Chromosome rearrangement by ectopic recombination in *Drosophila melanogaster*: genome structure and evolution. Genetics. 129:1085–1098.

Nellåker C, Keane TM, Yalcin B, Wong K, Agam A, *et al.* 2012. The genomic landscape shaped by selection on transposable elements across 18 mouse strains. Genome Biol. 13:R45.doi:10.1186/gb-2012-13-6-r45.

Nuzhdin SV, Mackay TF. 1995. The genomic rate of transposable element movement in *Drosophila melanogaster*. Mol Biol Evol. 12:180–181.

Olovnikov I, Ryazansky S, Shpiz S, Lavrov S, Abramov Y, *et al.* 2013. *De novo* piRNA cluster formation in the Drosophila germ line triggered by transgenes containing a transcribed transposon fragment. Nucleic Acids Res. 41:5757–5768. doi:10.1093/nar/gkt310.

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

Pasyukova EG, Nuzhdin SV, Morozova TV, Mackay TFC. 2004. Accumulation of transposable elements in the genome of *Drosophila melanogaster* is associated with a decrease in fitness. J Hered. 95:284–290. doi:10.1093/jhered/esh050.

Peters AD, Keightley PD. 2000. A test for epistasis among induced mutations in *Caenorhabditis elegans*. Genetics. 156:1635–1647.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: non-LTR retrotransposable elements and ectopic recombination in Drosophila. Mol Biol Evol. 20:880–892. doi:10.1093/molbev/msg102.

Petrov DA, Fiston-Lavier A-S, Lipatov M, Lenkov K, González J. 2011. Population genomics of transposable elements in *Drosophila melanogaster*. Mol Biol Evol. 28:1633–1644. doi:10.1093/molbev/msq337.

Pool JE, Corbett-Detig RB, Sugino RP, Stevens KA, Cardeno CM, *et al.* 2012. Population genomics of Sub-Saharan *Drosophila melanogaster*: African diversity and Non-African admixture. PLoS Genet. 8:e1003080.doi:10.1371/journal.pgen.1003080.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, *et al.* 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 81:559–575. doi:10.1086/519795.

Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, *et al.* 2016. The *Arabidopsis thaliana* mobilome and its impact at the species level. eLife. 5:e15716.doi:10.7554/eLife.15716.

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

Radman M, Wagner R. 1993. Mismatch recognition in chromosomal interactions and speciation. Chromosoma. 102:369–373. doi:10.1007/BF00360400.

Ragsdale AP. 2021. Can we distinguish modes of selective interactions using linkage disequilibrium? bioRxiv 2021.03.25.437004. doi:10.1101/2021.03.25.437004.

Rahman R, Chirn G, Kanodia A, Sytnikova YA, Brembs B, *et al.* 2015. Unique transposon landscapes are pervasive across *Drosophila melanogaster* genomes. Nucleic Acids Res. 43:10655–10672. doi:10.1093/nar/gkv1193.

Rebollo R, Karimi MM, Bilenky M, Gagnier L, Miceli-Royer K, *et al.* 2011. Retrotransposon-induced heterochromatin spreading in the mouse revealed by insertional polymorphisms. PLoS Genet. 7:e1002301.doi:10.1371/journal.pgen.1002301.

Renkawitz J, Lademann CA, Jentsch S. 2014. Mechanisms and principles of homology search during recombination. Nat Rev Mol Cell Biol. 15:369–383. doi:10.1038/nrm3805.

Riddle NC, Elgin SCR. 2018. The Drosophila dot chromosome: where genes flourish amidst repeats. Genetics. 210:757–772. doi:10.1534/genetics.118.301146.

Riddle NC, Minoda A, Kharchenko PV, Alekseyenko AA, Schwartz YB, *et al.* 2011. Plasticity in patterns of histone modifications and chromosomal proteins in Drosophila heterochromatin. Genome Res. 21:147–163. doi:10.1101/gr.110098.110.

Robberecht C, Voet T, Esteki MZ, Nowakowska BA, Vermeesch JR. 2013. Nonallelic homologous recombination between retrotransposable elements is a driver of *de novo* unbalanced translocations. Genome Res. 23:411–418. doi:10.1101/gr.145631.112.

Rogers AR. 2014. How population growth affects linkage disequilibrium. Genetics. 197:1329–1341. doi:10.1534/genetics.114.166454.

Samuk K, Manzano-Winkler B, Ritz KR, Noor MAF. 2020. Natural selection shapes variation in genome-wide recombination rate in Drosophila pseudoobscura. Curr Biol. 30:1517–1528.e6. doi:10.1016/j.cub.2020.03.053.

Sandler G, Wright SI, Agrawal AF. 2021. Patterns and causes of signed linkage disequilibria in flies and plants. Mol Biol Evol. 38:4310–4321. doi:10.1093/molbev/msab169.

Schmidt JM, Good RT, Appleton B, Sherrard J, Raymant GC, *et al.* 2010. Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. PLoS Genet. 6:e1000998.doi:10.1371/journal.pgen.1000998.

Shpiz S, Ryazansky S, Olovnikov I, Abramov Y, Kalmykova A. 2014. Euchromatic transposon insertions trigger production of novel pi- and endo-siRNAs at the target sites in the Drosophila germline. PLoS Genet. 10:e1004138.doi:10.1371/journal.pgen.1004138.

Sienski G, Dönertas D, Brennecke J. 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. Cell. 151:964–980. doi:10.1016/j.cell.2012.10.040.

Singh ND, Petrov DA. 2004. Rapid sequence turnover at an intergenic locus in Drosophila. Mol Biol Evol. 21:670–680. doi:10.1093/molbev/msh060.

Sohail M, Vakhrusheva OA, Sul JH, Pulit SL, Francioli LC, *et al.*; Alzheimer's Disease Neuroimaging Initiative. 2017. Negative selection in humans and fruit flies involves synergistic epistasis. Science. 356:539–542. doi:10.1126/science.aah5238.

Sprengelmeyer QD, Mansourian S, Lange JD, Matute DR, Cooper BS, *et al.* 2020. Recurrent collection of *Drosophila melanogaster* from wild African environments and genomic insights into species history. Mol Biol Evol. 37:627–638. doi:10.1093/molbev/msz271.

Stewart C, Kural D, Strömberg MP, Walker JA, Konkel MK, *et al.*; 1000 Genomes Project. 2011. A Comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genet. 7:e1002236.doi:10.1371/journal.pgen.1002236.

Stuart T, Eichten SR, Cahn J, Karpievitch YV, Borevitz JO, *et al.* 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife. 5:e20777.

Sved JA, Hill WG. 2018. One hundred years of linkage disequilibrium. Genetics. 209:629–636. doi:10.1534/genetics.118.300642.

Visser J. A G M D, Cooper TF, Elena SF. 2011. The causes of epistasis. Proc Biol Sci. 278:3617–3624. doi:10.1098/rspb.2011.1537.

Waldman AS. 2008. Ensuring the fidelity of recombination in mammalian chromosomes. Bioessays. 30:1163–1171. doi:10.1002/bies.20845.

Wang W, Yoshikawa M, Han BW, Izumi N, Tomari Y, *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. doi:10.1016/j.molcel.2014.10.016

Waterhouse RM, Zdobnov EM, Kriventseva EV. 2011. Correlating traits of gene retention, sequence divergence, duplicability and essentiality in vertebrates, arthropods, and fungi. Genome Biol Evol. 3:75–86. doi:10.1093/gbe/evq083.

Wells JN, Feschotte C. 2020. A field guide to eukaryotic transposable elements. Annu Rev Genet. 54:539–561. doi:10.1146/annurev-genet-040620-022145.

Xia S, VanKuren NW, Chen C, Zhang L, Kemkemer C, *et al.* 2021. Genomic analyses of new genes and their phenotypic effects reveal rapid evolution of essential functions in Drosophila development. PLoS Genet. 17:e1009654.doi:10.1371/journal.pgen.1009654.

Xie Z, Johansen LK, Gustafson AM, Kasschau KD, Lellis AD, *et al.* 2004. Genetic and functional diversification of small RNA pathways in plants. PLoS Biol. 2:e104.doi:10.1371/journal.pbio.0020104.

Yang P, Wang Y, Macfarlan TS. 2017. The role of KRAB-ZFPs in transposable element repression and mammalian evolution. Trends Genet. 33:871–881. doi:10.1016/j.tig.2017.08.006.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586–1591. doi:10.1093/molbev/msm088.

Zavattari P, Deidda E, Whalen M, Lampis R, Mulargia A, *et al.* 2000. Major factors influencing linkage disequilibrium by analysis of different chromosome regions in distinct populations: demography, chromosome recombination frequency and selection. Hum Mol Genet. 9:2947–2957. doi:10.1093/hmg/9.20.2947.

Communicating editor: T. Slotte