Inferred Clades of Transposable Elements in *Drosophila* Suggest Co-evolution with piRNAs

Iskander Said, Michael P. McGurk, Andrew G. Clark, Daniel A. Barbash

Department of Molecular Biology and Genetics, Cornell University, Ithaca NY, 14853, USA

Abstract:

Transposable elements (TEs) are self-replicating "genetic parasites" ubiquitous to eukaryotic genomes. In addition to conflict between TEs and their host genomes, TEs of the same family are in competition with each other. They compete for the same genomic niches while experiencing the same regime of copy-number selection. This suggests that competition among TEs may favor the emergence of new variants that can outcompete their brethren. To investigate the sequence evolution of TEs, we developed a method to infer clades: collections of TEs that share SNP variants and represent distinct TE family lineages. We applied this method to a panel of 85 *Drosophila melanogaster* genomes and found that the genetic variation of several TE families shows significant population structure that arises from the population-specific expansions of single clades. We used population genetic theory to classify these clades into younger versus older clades and found that younger clades are associated with a greater abundance of sense and antisense piRNAs per copy than older ones. Further, we find that the abundance of younger, but not older clades, is positively correlated with antisense piRNA production, suggesting a general pattern where hosts preferentially produce antisense piRNAs from recently active TE variants. Together these findings suggest a co-evolution of TEs and hosts, where new TE variants arise by mutation, then increase in copy number, and the host then responds by producing antisense piRNAs which may be used to silence these emerging variants.

Introduction:

Transposable elements (TEs) are mobile, selfish genetic elements commonly thought of as “genetic parasites". At the start of an invasion TEs begin as a single copy within a host genome, but can transpose and expand rapidly in copy number throughout the population in each successive generation by using the host’s replication machinery (Orgel and Crick 1980; Doolittle and Sapienza 1980). In *Drosophila,* explosive growth in copy number during a TE invasion is thought to be quickly followed by the host acquiring resistance to TE transpositions, commonly through host production of *piwi*-interacting small RNAs, piRNAs which interfere with TE transcripts (Brennecke et al. 2007; Kofler et al. 2018; Le Rouzic and Capy 2005; Aravin, Hannon, and Brennecke 2007). In the germline, piRNA mediated silencing begins with the “primary” piRNA pathway: transcription of long precursor antisense RNAs from TE-rich loci called piRNA clusters that are processed into 21-30 base-pair-long antisense piRNAs. These piRNAs complex with *piwi* clade proteins, bind to nascent sense TE transcripts by recognizing sequence complementarity, and then recruit additional proteins to transcriptionally silence homologous TEs. Additionally, the TE transcript is degraded to form sense piRNAs through a “secondary” piRNA pathway. These sense piRNAs bind to the antisense piRNA precursor to create a positive feedback loop, known as the Ping-Pong cycle, that establishes constitutive silencing (Brennecke et al. 2007; Le Thomas et al. 2014; Aravin, Hannon, and Brennecke 2007; Czech et al. 2018).

Ultimately the TE copy number may reach a steady state, with the rate of transposition dampened by piRNA silencing as well as selection against the deleterious consequences to reproductive fitness of the host organism (Charlesworth and Charlesworth 1983; Lee and Langley 2010; Kelleher, Barbash, and Blumenstiel 2020). However, as TEs expand in copy number, they also acquire polymorphisms in their sequences, which may lead to the formation of new lineages or subfamilies. Multiple lineages of a TE will compete with each other, as long as their polymorphisms are not deactivating (Le Rouzic and Capy 2006; Iwasaki, Kijima, and

Innan 2020). Much in the same way individuals within an ecological system are constrained by a carrying capacity, variants of the same TE may be constrained by the copy-number carrying capacity of the host (Charlesworth and Charlesworth 1983; Le Rouzic and Capy 2006). This dynamic produces an arena of genomic competition of TE variants where selection may drive the propagation of more fit TE lineages, while less fit lineages are purged (Le Rouzic and Capy 2006).

The study of selection on TE population variation has often focused on the fitness of the host organism rather than on the TEs themselves. Much of it centered on the variation of TE insertions within and between populations as well as fitness and phenotypic effects associated with particular insertion loci (Cridland et al. 2013; Blumenstiel et al. 2014; Kofler, Nolte, and Schlötterer 2015). The study of selection on sequence variation of TEs, on the other hand, is much more limited. TEs are typically categorized into classes and subclasses based first on their mechanism of transposition, and then on presence of shared motifs, relative sequence identity, and phylogenetic characteristics (Arkhipova 2017; Makałowski et al. 2019; Wicker et al. 2007). There is extensive systemization of TE families, describing their consensus sequences, open reading frames, and insertion site preferences (Bao, Kojima, and Kohany 2015).

Due to the challenges of quantifying variation within repetitive sequences, however, the empirical study of TE sequence polymorphism is largely limited to analyses of reference genome assemblies. For example, the sequence variation of TE families in the *Drosophila melanogaster* reference genome has been comprehensively described (Lerat, Rizzon, and Biémont 2003; Kaminker et al. 2002). In another example, phylogenetic and evolutionary analyses on retrotransposons within the *Oryza sativa* genome revealed strong purifying selection on protein-coding regions, with occasional bursts of positive selection (Baucom et al. 2009). To our knowledge, studies examining TE sequence variation from population samples are rare, likely because reference genomes are the primary source of full-length TE sequence data.

Ideally, to apply population genetic and molecular evolutionary principles to the genetic variation of TEs, we would study the complete sequences of individual TE insertions across many genomes. This is especially necessary if the aim is to assess competition between TE subfamily lineages, where reconstructing the underlying phylogenies would yield insight into the dynamics of how lineages diversify and potentially compete with each other. However, most current population genomic data comes from short-read sequencing, which does not permit an unambiguous assembly containing all the TEs with their unique variants. The problem is related to haplotype phasing, which can be done with short reads (Excoffier and Slatkin 1995; Delaneau, Coulonges, and Zagury 2008; Clark 1990; Browning and Browning 2007), except here the TE insertions are at nonhomologous positions. Furthermore, the high multiplicity of TEs greatly complicates the task of determining which polymorphisms co-occur in the same insertion, such that with short reads, unambiguous TE haplotypes cannot be recovered as complete sequences of linked SNPs. Although new long-read technologies, like PacBio and Oxford Nanopore, have emerged that greatly reduce phasing problems, their higher cost and relatively high error rates have limited their application to large-scale population genomics studies.

To gain insight into the sequence evolution of actively invading TEs, we sought to resolve some of these challenges by leveraging a straightforward intuition: If a set of SNPs co-occur in the same TE lineage, their copy-number variation should be correlated across genomes, covarying as the copy number of that lineage varies across genomes. To this end, we took advantage of the large sample sizes in a population-genomic dataset to quantify positive correlations in the copy number of SNPs across multiple individuals, and from these we identify groups of SNPs that co-occur within TE lineages. We refer to these groups of SNPs as clades, which are inferred to distinguish lineages of TE subfamilies while sidestepping the task of reconstructing the full phylogenies from short-read data. Applying our method to a set of 85 *D. melanogaster* genomes from the Global Diversity Lines (GDL) (Grenier et al. 2015), we inferred clades in 41 recently active TE families. We then used public PacBio datasets and simulations to validate our inferred clades (Long et al. 2018; Chakraborty et al. 2019). We analyzed the population variation of TE variants and found significant population structure driven by population-specific TE clades, several which are likely active. We additionally analyzed several piRNA libraries from ovaries focused on SNPs that distinguish clades, and found piRNAs are especially enriched for younger TE clades.

New approaches:

Hierarchical clustering of SNPs uncovers TE clades in NGS data:

We leverage large population-genomic datasets to detect TE clades by inferring the co-occurrences of SNPs within a TE lineage. We expect that if two alleles exist within the same lineage they will correlate in copy number, varying together as the TEs of that lineage vary in copy number (Figure 1a). We apply this principle to all pairwise combinations of SNPs within an element to compute a correlation matrix and then use hierarchical clustering to cluster groups of SNPs that are strongly correlated. The result is clusters of SNPs that co-vary in their copy number across samples; because these are inferred to occur within the same TE lineage we refer to these clusters as “clades”. Hierarchical clustering is a particularly appropriate choice for this problem as SNPs within TE lineages are truly related to each other in an underlying tree-like structure that is analogous to a hierarchical clustering dendrogram. The correlations between alleles are unlikely to be a result of co-transposition of multiple TEs because the linkage between TEs is very low and the sampling variation in these data is quite high.

We employed this clade inference method using short-read libraries from the Global Diversity Lines (GDL), 85 *D. melanogaster* lines from populations in Beijing, Ithaca, the Netherlands, Tasmania, and Zimbabwe (Grenier et al. 2015). We aligned the short-read data to the TE consensus sequences of 41 recently active TEs and the *D. melanogaster* reference genome using ConTExt (McGurk and Barbash 2018). Then we calculated allele frequencies of

SNPs from the read pileups of the alignments and calculated copy number from the read depth. In brief, copy number was estimated by dividing the observed read depth at each position on the TE consensus by the expected read depth of single copy sequences inferred from the read depth of the reference genome, with corrections for GC bias (McGurk, Dion-Côté, and Barbash 2020). For each individual we took the allele frequencies at each position and multiplied them by the estimated copy number at each position to generate the copy number of alleles at each position for that individual. We compute pairwise correlations between the copy number of alleles across individuals (Figure 1b), and then employ hierarchical clustering to cluster positively correlated SNPs thus inferring clades (Figure 1c). For each of the 41 TEs analyzed, we report the SNP clusters, as well as the copy number of each inferred clade, calculated by averaging the copy number of the individual alleles (Supplemental File 1).

One important consideration of this method is that we are not identifying the full set of TE insertions within an inferred clade. Rather we are identifying sets of SNPs that distinguish lineages from each other (lineage-informative SNPs), not the complete sequences of the insertions at any particular locus that belong to an inferred clade. Clades are statistical inferences of the set of true lineages that exist for a TE family, but because all TE lineages of a family are related by an underlying phylogeny, there likely does not exist a single correlation cutoff that optimally groups TEs into distinct clades. Rather, any chosen threshold induces some degree of coarse-graining in how it collapses this phylogeny, splitting and merging lineages of TE variants into clades. For example, two closely related lineages may share an ancestral set of SNPs, but have recently diverged and acquired a small number of additional polymorphisms that distinguish them. Depending on the stringency of the clustering cut-off these two lineages may be called as a single clade containing all of the SNPs, or multiple clades each with some subset of SNPs. Thus, a higher stringency in the clustering cut-off will produce many small clusters of tightly correlated SNPs that split lineages, while low stringency cut-offs will produce a few large clusters that merge lineages together. Because of this, clusters of SNPs may be found to occur in the same insertions at some frequency, but whether or not they are merged into one clade depends on the clustering cut-off and how often the two sets of lineage-informative SNPs occur in the same insertion.

We attempt to address these caveats and trade-offs in our analysis by using simulations and PacBio data to validate our inferences. When inferring clades in the GDL short-read data we chose stringent clustering parameters to make the clade calls conservative (splitting distantly related lineages). These sets of parameters were chosen to essentialize TE clades to a minimum number of core SNPs that co-occur with a high positive correlation, while increasing the number of distinct, resolved clades. High stringency cut-offs also minimize the number of false positives that can result from performing many thousand pairwise correlations of allele copy number (Supplemental Figure 1a, 1b). Parameters could be tuned to be less stringent to define clades harboring greater internal SNP variation, but the sets of parameters we chose performed well in our validation.

TE clade inference recovers lineages in simulated and biological datasets.

We first validated a critical assumption of our inference model: that SNPs on the same TE will co-vary in copy number, while SNPs on unrelated TE sequences will not. To do this we asked whether the SNPs in different TE families are positively correlated, expecting that because the SNPs on different TEs are physically unlinked there will be little to no positive correlation between them. We computed the correlation for every pairwise combination of active TEs in our dataset and found, as expected, the correlation of SNPs taken from the same TE family (average *Pearson's r* = 0.23) are generally much greater than from unrelated families (average *Pearson's r* = 0.02) (Supplemental Figure 2a). There is an elevated number of positive correlations between SNPs on different telomeric TEs (average *Pearson's r* = 0.11), likely due to them being linked together in large multimeric arrays exclusively at the ends of chromosomes. There are also very strong positive correlations of SNPs in the *Bari* element

(*Pearson's r* = 0.69*)* and *P-element* (*Pearson's r* = 0.68*)* that are driven by the small number of SNPs segregating in those two families. There are only three SNPs that passed filtering for *Bari* and four for *P-element*, which represent one clade in each TE family.

We next sought to benchmark the performance of our method by simulating TE polymorphism data using a phylogenetic process to generate an inference data set with a known set of lineages to use as validation. Using the simulated data we benchmarked our method by inferring clades under a range of parameters and then reported the clustering quality, as determined by a Silhouette Score. In brief, the Silhouette Score is bound between -1 and 1 and quantifies the cohesiveness and separation of clusters. The optimal score is 1 and implies that clusters of SNPs were often found together in the validation dataset. We find that our method produces high Silhouette Scores (0.75-0.9), correctly inferring clades in the simulated dataset under a wide range of clustering parameters (*Pearson's r* = 0.3-0.9), and only produces errors when clustering parameters are at extremes (Supplemental Figure 2b). The inferred clades also recapitulate the structure of the simulated phylogenetic tree (Supplemental Figure 5b).

We further validated our clade inferences from the GDL short-read data by comparing them to long-read PacBio assemblies of *D. melanogaster* lines that contain molecularly phased haplotypes for the TEs. We looked for sets of SNPs that delineate TE clades in the high-confidence contigs of PacBio assemblies, considering only clades where two or more of the lineage-informative SNPs were detected in the PacBio data, because we were interested in estimating our ability to correctly infer the rate of co-occurrence between detected SNPs in a lineage. Not removing those clades downwardly biased the estimate of our clustering accuracy (Supplemental Figure 3a). Using the filtered set of clades, we found that 70% of the total clades inferred from the GDL data were detected in the PacBio dataset (Figure 1d). We discovered that of the 30% undetected clades in the filtered set, 5% were very likely closely related lineages that had been merged into a single cluster (Supplemental Figure 2b). The remaining 25% of the inferred clades not found in the PacBIo data tend to be large clusters that are missing more lineage-informative SNPs in the PacBio data than the detected clades. This indicates that these undetected clades might be closely related lineages merged into a single cluster as well, but are population specific, or are sufficiently rare that a subset of their SNPs were not sampled in the PacBio data. Therefore the differences in samples between the short-read and PacBio data may be driving poor clustering quality rather than systematic errors, however we cannot completely discount true errors in clustering (Supplemental Figure 4).

Results:

Diversity and variation of TE lineages in the GDL.

To determine whether high sequence diversity of a TE family is due to the evolution of many distinct lineages or a few highly diverged lineages, we assessed the sequence diversity of TE families and the number of clades segregating in the GDL. We calculated the average nucleotide diversity across the GDL of active TE families and found a positive relationship between the number of clades and nucleotide diversity (Figure 2a; *Pearson's r = 0.74; p-value < 0.05*). Both the telomeric TEs and LTR retrotransposons have many families with a high number of clades and high sequence diversity. There are also several LINE-like retrotransposon and DNA transposon families with high sequence diversity, but they tend to have fewer clades than non-LINE TEs with similar sequence diversity. For example, the *R1* family has a high sequence diversity ($\pi \approx 0.026$) and only eight clades, while the non-LINE LTR retrotransposon *Zam* has comparable sequence diversity and 94 clades.

This pattern may be driven by merging SNPs into large clades rather than splitting them into many small ones, so we characterized each clade by the number of lineage-informative SNPs it contains (Figure 2b). In general, the distribution of lineage-informative SNPs is small and tightly distributed, with a median number of two and an interquartile range (IQR) of one. The small cluster size is indicative of a preference for splitting multiple related clades rather than merging them into larger clusters. Clusters of SNPs we discover may therefore not be mutually exclusive and may occur together within a subset of insertions, but with a degree of positive correlation insufficient to pass the clustering threshold. This preference for splitting would upwardly bias the number of clades that we estimate, as large clusters of SNPs with distant genealogical relationships might be broken up into many small clusters of SNPs that are closely related. The exact number of clades segregating in each TE family is affected by parameterization of the clustering, but the relative proportions of clades among different families in our analyses are likely not. Furthermore, the inference of clades using simulated data showed that the number of clades and quality of clusters are robust to clustering parameters (Supplemental Figure 2b, 5a).

Although most TE families have clades with a small number of SNPs, *R1* clades are notable outliers, with a median of 14 SNPs and IQR 19.75 and two clades with 60 and 100 SNPs each. This might be explained by the presence of two independently evolving populations of *R1* elements in *D. melanogaster*, the hundreds of *R1* insertions in the highly repetitive ribosomal DNA array, and a separate lineage of divergent elements that comprise a megabase-sized satellite array (McGurk and Barbash 2018; Roiha et al. 1981; Wellauer and Dawid 1977; Xiong and Eickbush 1988; Luan et al. 1993). Divergence between these lineages likely explains the high sequence diversity. The similarity of sequences within each lineage and dissimilarity between lineages may favor their merging into a handful of large clades during clustering. Additionally, *R1* arrays are physically linked, behaving as a single haplotype that experiences high rates of gene conversion that homogenizes the ribosomal DNA and the *R1* copies (Szostak and Wu 1980; Dvorák, Jue, and Lassner 1987). All of these features could explain why full *R1* clades were not detected in the PacBio data (Figure 1d).

We next addressed whether a single clade dominates a transposable element family in terms of copy number or if instead the clades occur at roughly equal frequency. We determined the proportion of all copies of a TE family in the GDL population belonging to a clade by calculating the “clade population frequency”, dividing the total clade copy number by the average copy number across the lineage-informative positions summed across the GDL (Supplemental File 1). We find that most clades occur in the population at a frequency between ~10-30% (mean 17%; Figure 2c). There is a notable lack of low frequency clades, likely due to the filtering of low copy-number and low population-frequency alleles before calling clades. We found 21 clades with high population frequency (> 40%) occurring in 13 different TE families, including telomeric TEs and several LTR and LINE-like retrotransposons. The *Diver* and *Nomad* LTR retrotransposons had the greatest number of high frequency clades (3-4 per family), but with dramatically different clade population frequency distributions. The *Diver* frequency distribution had many clades spread across the entire range from low to high, while the *Nomad* distribution is clearly split between a handful of low and high frequency clades (Figure 2d). Several other TE families like *Jockey, Doc*, and *DM412* had frequency distributions similar to *Nomad*, while other families such as *Gypsy1*, *Zam*, and *I-element* had a more uniform clade frequency distribution similar to *Diver* (Figure 2d). The *Nomad-like* frequency distributions may reflect a relatively fast copy-number expansion of a handful of clades that outcompeted other lineages, while *Diver-like* distributions may reflect gradual diversification and slow increase in copy number of many clades, possibly driven by stochastic processes.

One important consideration is that due to the way population frequency was calculated, clades with SNPs in commonly deleted portions of a TE may be at a high frequency despite being at a relatively low copy number. This is particularly important for LINE-like elements and DNA transposons that are frequently truncated and internally deleted. Therefore, the clade population frequency does not necessarily reflect the proportion of TE insertions in the clade, but instead the number of TE insertions that have those nucleotide sites in the given clade. High population-frequency clades in *Gypsy*, *Zam*, *HeT-A2*, and *Tart-B1* had very low copy numbers (~1-2 copies on average; Supplemental File 1), likely due to having SNPs in commonly deleted portions of their respective TE sequences. However, we found that many high population-frequency clades were at high copy number (10-40 copies on average). The high frequency clades of *Jockey, DM412*, and *Doc* clades were particularly striking as they reach high copy number and dominate other clades of their respective families. These clades may be at high frequency due to age, having a competitive edge over other variants, or by pure chance.

The majority of clades are young and recently active.

While the limitations of short-read data prevent us from mapping SNPs to specific insertions and assessing their population frequency, the insertion-site frequency spectrum of a TE influences the variance of its copy-number distribution across individuals. Our approach infers TE clades, which are collections of TE insertions that share a subset of SNPs, but the same idea applies -- the insertion-site frequency spectrum, and therefore copy-number distribution, of clades is also expected to be related to their age. We therefore applied population genetic theory to predict the age of clades from their copy-number distributions. Recently active, young TEs will have insertions that are mostly at a low population frequency with little variance. If there is a large number of occupiable sites and no linkage disequilibrium between insertions, the copy-number distribution will follow a Poisson distribution (mean equal to the variance, “dispersed”) (Charlesworth and Charlesworth 1983; McGurk, Dion-Côté, and Barbash 2020). An older lineage will have insertions at variable frequencies -- due to drift increasing the frequency of older insertions -- resulting in the variance being less than the mean (“underdispersion”). A recently active lineage with population structure (*e.g*. a population-specific expansion) or other forms of linkage disequilibrium will have a variance greater than the mean (“overdispersion”) (Charlesworth and Charlesworth 1983). The copy-number distributions of known active and inactive TE families recapitulate these expectations well (McGurk, Dion-Côté, and Barbash 2020).

We analyzed the copy-number distribution of clades by using a two-tailed dispersion test with multiple testing correction to ask whether the distributions are overdispersed, underdispersed, or fit a Poisson (Figure 3a) (Yang, Hardin, and Addy 2009). We found that most high population frequency clades of telomeric TEs and other families such as *Jockey, Copia, Nomad* and *DM412* had copy-number distributions consistent with recent activity (dispersed or overdispersed), while the high frequency clades of *Zam* and *Stalker-4* were classified as older lineages (underdispersed). High population frequency clades in *Doc* and *Diver*, on the other hand, were a mixture of young and old clades. We found that on average old clades were at a slightly higher frequency in the population than young clades (*Dunn's test: p-value < 0.05;* Figure 3b). Overall, 56% of the clades were classified as young and therefore active. The excess of old clades segregating in the GDL is driven by eight families (*Gypsy*, *Gypsy1*, *I-element*, *BS*, *Zam, Bel*, *Diver* and *Burdock*), which accounts for 86% of all the old clades (Figure 3c). The abundance of old clades in these TE families matches their known insertion-site frequency spectra, which is skewed towards older, higher population frequency insertions (Kofler, Nolte, and Schlötterer 2015). However, despite having mostly old and inactive lineages, there are still several young lineages actively transposing in these families.

I-element is a particularly striking example where all but one clade is old (Figure 3c). This is consistent with the known evolutionary history of the *I-element*, as it appears to have invaded *D. melanogaster* populations multiple times, leaving both inactive relics of ancient invasions and younger active copies (Kidwell 1983; Picard et al. 1978; Busseau et al. 1994). Many *D. melanogaster* strains are susceptible to *I-element* invasion, despite having euchromatic insertions, and crosses with strains carrying active *I-elements* result in hybrid dysgenesis (Olovnikov et al. 2013; Ryazansky et al. 2017). Therefore, many of the older clades segregating in *I-element* may be remnants of this ancient invasion.

Curiously, the only *I-element* clade that was predicted to be young shows strong population structure, being at a higher copy number in Beijing than in the other populations (Figure 3d). Strong population structure is expected to inflate the variance in copy number calculated across all populations, so we re-analyzed the *I-element* in the Beijing population. We found the copy-number distribution of the Beijing strains fit a Poisson well, implying that this *I-element* clade is likely young and active (*p-value* = 0.7). This clade is at less than 1% frequency in the other four populations, while at ~11% frequency in the Beijing population. It is quite likely then that this young *I-element* clade invaded the Beijing population after the *D. melanogaster* population migrated to East Asia. Generally, 54% of the likely active clades were overdispersed, which suggests there may be population structure to their geographic distributions as well and potentially indicates ongoing population-specific invasions.

Population structure of TE variation.

The genetic variation of TEs within and between populations is an underexplored facet of TE evolution. Early in the *P*-element and *hobo* invasions, variant lineages emerged and rose to high frequency, entirely replacing the wild-type TE in some populations within a decade (Black et al. 1987; Periquet et al. 1989). These dynamics may reflect selection acting at the level of TEs, with variants outcompeting the ancestral lineage (Le Rouzic and Capy 2006; Iwasaki, Kijima, and Innan 2020; Robillard et al. 2016). The clades we identified provide an opportunity to catch such events in progress. We sought therefore to identify TE lineages that have expanded or contracted in copy number within specific geographic populations, because these might be signatures of selection acting on the TE sequence.

To find clades with population structure we used a Bonferroni-corrected Kruskal-Wallis test to determine which clades rejected the null-hypothesis that their copy number was homogeneously distributed across populations. We found that ~15% of clades were heterogeneously distributed among the five GDL populations, thus indicating population structure (Figure 4a). Some TE families, such as *Burdock* and *Tart-A*, have few or no clades that are enriched for particular populations, while *Jockey, Copia,* and *Tirant* have many clades with population structure.

To quantify the extent of the population structure we compared the average clade copy number across the GDL to the average clade copy number of the subpopulation (Beijing, Ithaca, Netherlands, Tasmania, or Zimbabwe) that was the most differentiated from the entire GDL (Figure 4b). Of the clades that were statistically significant by the Kruskal-Wallis test, the most differentiated populations had a clade copy number that was, on average, ~3.8 copies greater or lesser than the GDL mean. In general, these population structure differences were of modest effect, but *Roo*, *R1,* and *F-element* clades have differences from the GDL mean of ~10-35 copies. These much larger effect sizes might be driven in part by the very high copy number of these three families throughout the genome.

Our analysis of these summary statistics, although informative, does not reveal in which population(s) a clade is enriched. We therefore employed PCA (Principal Component Analysis) on the matrix of SNP frequencies of each TE family in each individual. This allows us to find which SNPs are driving the population variation within a TE family, as well as to visualize which individuals in the GDL carry similar TE variants. We find strong population structure for *Roo* variants with distinct clusters of Beijing and Zimbabwe individuals (Figure 4c). This population structure is driven by population-specific expansions of clades (Figure 4d, 4e). The Beijing- and Zimbabwe-specific clades are at ~45 copies (~33% frequency), and ~50 copies (~40% frequency) in Beijing and Zimbabwe, respectively. The Beijing clade is very rare outside of its respective population, ~1% frequency, which implies that it emerged in East Asia and then expanded in copy number. The Zimbabwe clade, on the other hand, segregates at ~10% frequency in the other populations, implying a more ancestral origin.

We find an analogous pattern in *Tirant* and *Jockey* variation where there is also strong population structure that is driven by population-specific expansions of clades. Much like *Roo,* Ithaca- and Tasmania-specific clade expansions drive the population structure of *Tirant* variation (Supplemental Figure 6a, 6b, 6c). However, in *Jockey* it is the absence of a clade in Zimbabwe that is found in all other populations, coupled with a Zimbabwe-specific expansion of a different clade, that drives that structure (Supplemental Figure 6d, 6e, 6f). In these three families with notable population structure, it is the presence or absence of a single clade that drives the variation rather than multiple variants expanding within the populations. This pattern could be a reflection of selection favoring the expansion of a single lineage in a population, or simply due to genetic drift. In either case it shows that TE lineages are able to expand in copy number and become endemic in a population, dramatically altering the composition of TE variants within those individuals.

Not every population-specific expansion of a clade will be as stark as *Roo*, *Jockey*, or *Tirant*. And as we noted above, ~15% of the several hundred clades discovered show significant heterogeneity in copy number between populations. These clades may not be sufficient to drive variation on a PCA due to their modest effect sizes, but are still significantly different between populations. These small differences may represent stochastic fluctuations in clade copy number between populations, or they may reflect the initial stages of a newly emerging clade in a population rising to high frequency.

Sense and antisense piRNA pools are diverse and reflect the age of variants.

One of the primary mechanisms by which hosts control the proliferation of TEs is through the piRNA pathway. piRNAs are produced in both the sense and antisense direction from two distinct pathways. Antisense piRNAs are generally produced from clusters containing fragments of inactive TEs and target TE transcripts for silencing. Sense piRNAs, in contrast, are generally derived from cleavage of a TE primary transcript, guided by an antisense piRNA. Antisense piRNAs therefore reflect the potential to silence TE expression, while sense piRNAs reflect the cleavage and silencing of TE transcripts. Sense piRNAs further feed back into the production of antisense piRNAs, amplifying the pool of piRNAs targeting that TE sequence (Czech et al. 2018; Aravin, Hannon, and Brennecke 2007; Brennecke et al. 2007). Therefore, the pool of piRNAs a host produces might reflect the genetic variation of active TE families, not just polymorphisms in piRNA clusters.

We asked whether the sequence diversity in the sense and antisense piRNAs ($\boldsymbol{\pi}_{\text{piRNA}}$) tended to be correlated with the sequence diversity of the TEs themselves ($\boldsymbol{\pi}_{\text{TE}}$). Sequence diversity of the piRNAs was quantified by aligning ovarian piRNA libraries from 10 strains from the GDL (two from each population) to TE consensus sequences (Luo et al. 2020). For each TE family we pooled together the piRNA reads from the 10 strains to calculate the average piRNA sequence diversity across its consensus and did a similar procedure with the copy-number data. To reduce technical artifacts we only considered SNPs in the piRNA data whose presence was supported by the corresponding genomic data. We found a strong positive relationship between the sequence diversity of a TE family and the sequence diversity in both sense and antisense piRNAs (+: *Spearman's rho* = 0.90, *p-value* < 0.05; -: *Spearman's rho* = 0.88, *p-value* < 0.05) (Figure 6a).

The ratio ($\boldsymbol{\pi}_{\text{piRNA}}/\boldsymbol{\pi}_{\text{TE}}$, dubbed the “piRNA diversity ratio”) in each TE family estimates how well the piRNA diversity reflects genomic diversity. If the ratio is 1, then the piRNAs are as diverse as the genomic loci that they are derived from. A piRNA diversity ratio less than 1 implies that there is greater unevenness in the proportion of variants found in the piRNA pool than in the genomic sequence, such that some variants may be absent from the piRNAs while others dominate. In contrast, a piRNA diversity ratio greater than 1 implies that variants are present in piRNAs at more equal proportions than in the genomic sequence. Across all families of TEs, the piRNA diversity ratio is approximately 0.6 (Figure 5a).

We found that LTR retrotransposons have the lowest piRNA diversity ratio, 0.58 for either strand, while DNA transposons and LINE-like elements have ratios of ~0.7 (Supplemental Figure 7a). A low sense piRNA diversity ratio may reflect TE families with an abundance of old and inactive TE variants, which would produce few sense transcripts and thus few sense piRNAs. Alternatively, these TE variants may be actively transcribing, but not be effectively silenced by piRNAs and therefore few of their transcripts are processed into sense piRNAs. A lower antisense piRNA diversity is likely due to the fact that antisense piRNAs tend to be generated predominantly from piRNA clusters, which contain only a subset of TE variants that are not necessarily representative of all TE variation. *P-element* has the lowest piRNA diversity ratio of all families, ~0.01, likely reflecting the low genetic diversity of *P-element* ($\pi \approx 0.00076$) and their recent invasion in *D. melanogaster* (Kidwell 1983). Few *P-element* variants are segregating in the population and even fewer presumably have been captured by piRNA clusters.

Although most TE families had a low piRNA diversity ratio, *R2, Blood* and *Hobo* had a piRNA diversity ratio > 1 for one or both strands, and *Roo* was about 1.5x for both strands. These results were not due to an abnormally low average genomic diversity (~0.005 - 0.01). A high sense piRNA diversity ratio implies that many TE variants are transcribed and targeted by piRNAs, while the high diversity of the antisense piRNAs indicates that most variants are present in piRNA clusters or are producing de novo antisense piRNAs.

Given that piRNA content was generally less diverse than the TEs themselves, we wanted to determine which variants were contributing to the diversity of the piRNA pools. Therefore, we calculated the number of mapping sense and antisense piRNAs that contained the lineage-informative SNPs of a clade per copy of that clade and averaged this ratio across the 10 GDL strains. A clade that is both being regulated by piRNAs and being used to produce piRNAs should have high quantities of both sense and antisense piRNAs.

With the exception of the telomeric TEs, antisense piRNA read depth of lineage-informative SNPs was more abundant than sense. This difference was most stark in LTR retrotransposons where over 4x more antisense than sense piRNA reads/copy contained lineage-informative SNPs (Figure 5b, *Wilcoxon signed-rank test: p-value* < 0.05*)*. This suggests that many of these clades have been incorporated into piRNA clusters and are producing antisense piRNAs. The telomeric TEs not only showed similar proportions of sense and antisense piRNAs per copy, but also generally had more piRNAs per clade copy than other TEs, likely reflecting the fact that piRNAs targeting telomeric TEs are generated from the telomeric TEs themselves rather than distinct piRNA cluster loci (Radion et al. 2018). Newly evolved telomeric TE variants do not have to insert by chance into an existing piRNA cluster or become converted into a *de novo* piRNA cluster, but instead can be immediately incorporated into the pools of antisense piRNAs.

We find that many clades produced few or no piRNAs, but there were some, such as in *HeT-A5*, that produced over a hundred piRNAs/copy (Supplemental Figure 7b). The low piRNA diversity ratio we observed for TEs such as *HeT-A5* may therefore reflect the inclusion of only a subset of clades in the primary or secondary piRNA pathway, and this inclusion may not be representative of the copy number of those clades. It is clear that some clades are more likely to be present in the piRNA pool than others. We therefore hypothesized that recently active variants might be more readily targeted by host piRNAs and therefore produce more piRNAs/copy. We used the above described classifications of young and old clades based on the Poisson fit of their copy-number distributions and analyzed the piRNA abundance of the two groups. We found that young clades have significantly higher sense piRNAs/copy, fitting our prediction that these clades are indeed actively transcribed and therefore likely transpositionally active (*Dunn's test: p-value* < 0.05) (Figure 5c). This also held for antisense piRNA read depth, although the difference was less pronounced (*Dunn's test: p-value* < 0.05). It is clear that although young, recently active clades are more readily used as a substrate by the primary piRNA pathway, there are still many older clades that generate antisense piRNAs, perhaps representing old heterochromatic piRNA clusters containing inactive variants. We note that our classifications using the theoretical expectations of the copy-number distribution are imperfect, and some "older" variants may still retain transpositional activity despite their age.

Previous models of evolutionary arms races between TEs and piRNAs predict a positive correlation between the copy number of invading elements and the production of antisense piRNAs, because there is selection on the host genome to silence these elements (Luo et al. 2020). Therefore, for each clade we calculated the Spearman's rank correlation coefficient between its copy number and piRNA read depth for the 10 GDL strains, and compared these values between active and inactive clades. We found that young clades were significantly enriched for positive correlations in both sense and antisense piRNAs (Figure 5d, 5e, 5f; +: *Dunn's test*: *p-value* < 0.05; -: *Dunn's test*: *p-value* < 0.05). We also found that there were 61 and 58 young clades that had a statistically significant correlation between copy number and antisense and sense piRNA read depth, respectively, while only 2 and 4 old clades were statistically significant (*Benjamini-Hochberg: FDR = 10%*). Of the young clades many belonged to telomeric TEs, or recently active LTR and LINE-like retrotransposons, like *Jockey*, and *Tirant*.

Overall, our analysis of piRNA sequence variation shows that host piRNA content changes to respond to the emergence of variant TEs, and that not all variants are represented in the piRNA pool. Young, putatively active TE variants are disproportionately represented in the sense and antisense piRNAs, suggesting that host genomes may be responding to the evolution of new TE lineages.

Discussion:

Clade inference provides a new tool for understanding the evolution of TEs.

We have developed a technique for inferring clades within TE families by leveraging population genomics datasets and heuristic statistical methods. This approach bridges a significant gap in the field of population genomics by obtaining information about TE family substructure from existing short-read datasets. Simulations show that this method reliably identifies clade structures that are consistent with the TE genealogy under a wide parameter space, and we also validated TE clade inferences in *D. melanogaster* by checking them against

PacBio genomes. We then used the clade designations from data on natural *D. melanogaster* populations to infer aspects of TE dynamics and host responses.

The clade calls should be interpreted with some caveats in mind, however. The clusters of lineage-informative SNPs are markers that distinguish clades in the TE genealogy, not the complete set of SNPs within a full-length insertion. Given that two clades may descend from the same ancestral lineage, clusters of SNPs may co-vary within insertions but still be called as distinct TE clades. This behavior reflects a trade-off between merging versus splitting clades and depends on the chosen false-positive rate. This choice will affect the number of clades called for a given TE, but does not likely change the relative proportions of clades among different TEs, which our simulations show are robust to perturbations in clustering parameters. The technique we have developed can be readily applied to any organisms where population-level short-read genomic sequence data and libraries of TE consensus sequences exist.

Extensive population variation of TE lineages.

The study of the genetic variation of TEs has previously been largely relegated to reference genome assemblies. By applying population genetic theory to the copy-number distribution of clades, we found that a majority of clades (56%) were young and recently active. This is not wholly unexpected as most of the TE families we assayed have sequence diversity and population insertion-site frequency spectra that reflect recent invasion and activity (Kelleher and Barbash 2013; Kofler et al. 2015). We found that some young clades, such as in *Jockey* and *DM412*, have expanded in copy number dramatically across all populations, accounting for ~40% of all insertions. Other young clades have expanded only within a subset of the populations, sometimes to 3-4x higher copy number than other populations. Interestingly, Tasmanian-specific SNPs for *Tirant* had been previously observed, but our study is the first to put this observation in the context of the emergence and expansion of a TE lineage (Schwarz et al. 2020).

This begs the question: what drives these differences in copy number? One possible cause of local copy-number expansions is the acquisition of adaptive polymorphisms that increase transposition rate. For example, *hobo* elements in *D. melanogaster* with 5 copies of an internal repeat are less active than variants with 3 copies (Souames et al. 2003). There are also segregating polymorphisms in the human *LINE-1* element that account for ~16 fold differences in transposition rate (Lutz et al. 2003; Seleme et al. 2006). Polymorphisms could affect the transposition rate by changing the affinity of a transcription factor to the internal promoter, modifying the conformation of a transposase or reverse transcriptase to increase efficiency, or evading host genome-silencing mechanisms. A more transpositionally efficient variant would eventually displace other variants as it increased in copy number within that population (Le Rouzic and Capy 2006).

It is also possible that differences in clade copy number between populations are caused by neutral processes. Genetic drift and geographic isolation could affect the copy number of variants within a population, thus creating population structure. Population genetic simulations of TEs competing within a population provide a future way to explore these hypotheses.

Antisense piRNA production of variants may be adaptive.

Although the piRNA system can quickly respond to the invasion of TE families into naive populations by producing antisense piRNAs specific to those new invaders (Kofler et al. 2018), its ability to change in response to the emergence of new variants of a TE family has been underexplored. We have shown that the piRNA defense system is surprisingly malleable and seems to often respond to the emergence of new variants by incorporating those variants into antisense piRNAs. The presence of a variant in the antisense piRNAs indicates inclusion of that variant in a piRNA cluster, and may reflect the propensity for the host to silence those variants. We found that, in general, antisense piRNAs had less sequence diversity than genomic TE insertions and that young, recently active clades were overrepresented in the antisense piRNAs.

This is consistent with previous findings that showed a bias for piRNA silencing of active human *LINE-1* elements (Lukic and Chen 2011). In *D. melanogaster*, a positive relationship was found previously between indicators of transposition activity for TE families and their antisense piRNA abundance. However, this relationship seemed to be driven by the removal of inactive TE families from the piRNA pool rather than an increase in the silencing of active elements (Kelleher and Barbash 2013). Our analyses largely concur, as older clades of TEs are significantly less represented in the piRNA pool than younger clades.

Furthermore, the malleability of piRNA content might be beneficial to the host as positive correlations were found between the copy number of young clades and their antisense piRNA read depth. Such positive correlations are predicted under an evolutionary arms race model with strong piRNA silencing (Luo et al. 2020). Recent analyses of TE family copy number in *D. melanogaster* laboratory and natural populations found positive correlations between piRNA read depth and copy number in 6 out 105 families analyzed. These were mostly young and recently expanding TE families, including *P-element* and a handful of telomeric TEs (Luo et al. 2020; Saint-Leandre et al. 2020). By considering the copy-number variation of clades within TE families, our analyses provide much wider evidence of the expected correlation, with antisense piRNA production correlated with copy number for 61 TE clades in 22 out of the 41 likely active TE families considered. These include many telomeric TE clades, for which the positive correlation may have a distinct mechanistic explanation: as nearly all telomeric TEs are found at the telomere ends and do not insert at pericentromeric piRNA clusters, the piRNAs must be generated from the telomeres (Radion et al. 2018). However, we discovered that other active TE families, including *Roo, Jockey, R1,* and *Tirant*, also show this correlation and were not detected previously. The increased power in our analysis would be expected if active lineages preferentially display this correlation between piRNA read depth and TE copy number, with family-level analyses losing statistical power due to the aggregation of young and old clades.

This highlights the importance of integrating sequence polymorphisms into the analysis of TEs and the utility of our clade inference method.

Because the arms-race model predicts significant positive correlations between antisense piRNA abundance and copy number when the strength and efficiency of piRNA silencing is high, it is possible that the host produces antisense piRNAs that are specific to recently active clades to increase silencing efficiency (Luo et al. 2020). Those piRNAs that have perfect sequence complementarity to their targets might have higher specificity in binding and therefore increased silencing efficiency. In *C. elegans* and *D. melanogaster,* deletions or polymorphisms in piRNA binding sites on a transcript can reduce or eliminate silencing (Post et al. 2014; D. Zhang et al. 2018).

Given that piRNA silencing efficiency is affected by sequence complementarity, then two possible models may explain the significant enrichment for recently active TE clades in antisense piRNAs. In the first, natural selection acts to increase the frequency of piRNA clusters segregating in the population that contain active TE variants. piRNA clusters can be highly polymorphic in TE content and rapidly turnover in sequence (S. Zhang, Pointer, and Kelleher 2020; Wierzbicki et al. 2020; Assis and Kondrashov 2009; Zanni et al. 2013). Many distinct piRNA clusters are therefore likely to be segregating in *D. melanogaster* populations*,* each with distinct compositions of TE families and variants. The piRNA clusters that contain newly emerging variants may be selected for if they more efficiently silence novel variants, thus increasing in frequency.

Alternatively, transcriptional activity of piRNA clusters may drive variation in piRNA pools. This epigenetic model is plausible because the transgenerational inheritance of piRNA cluster expression is dependent on maternally deposited piRNAs that trigger the production of primary piRNAs from piRNA clusters (Le Thomas et al. 2014; Brennecke et al. 2008). Maternally deposited piRNAs derived from TE variants may have higher affinity to piRNA clusters that are composed of those same variants, and therefore bias the production of antisense piRNAs towards more active variants. This mechanism could “switch on” the activity of piRNA clusters that contain active TE variants, thus establishing a transgenerational change in piRNA content without altering the frequencies of piRNA clusters in the population. Such transgenerational epigenetic changes may be additionally bolstered by the formation of *de novo* piRNA clusters that can form from individual TE insertions. Euchromatic insertions of TE variants that are young will more likely be transcriptionally active than insertions of older variants, making them a more prominent target of silencing by antisense piRNAs. The piRNA pathway would then silence these actively transcribing TE variants and may convert some into *de novo* piRNA clusters by recruitment of the *Rhino-Deadlock-Cutoff* complex, thus producing variant-containing antisense piRNAs (Olovnikov et al. 2013; Shpiz et al. 2014; Mohn et al. 2014).

In the first model, selection plays a major role in determining the piRNA content in a population and the enrichment of variant-containing antisense piRNAs is strictly adaptive. But in the epigenetic model, the enrichment is not necessarily adaptive. Changes in piRNA content may shift to bias more active variants due to variation in piRNA cluster activity or through the formation of *de novo* clusters, but these variant-containing piRNAs need not be more efficient at silencing TE variants. It is possible that this epigenetic variation is beneficial for the host, or it may be a byproduct of the mechanisms by which piRNA clusters are inherited. These two models are not mutually exclusive, and the underlying observations reveal a fundamental aspect of TE-host coevolution, where a new TE variant emerges in a population, increases in copy number, and then is used as a substrate by the host genome to produce novel antisense piRNAs.

Methods

Aligning short-read data to TE consensus using ConTExt and estimating copy number:

85 short-read libraries from the Global Diversity Lines were aligned to a curated index of RepBase TE consensus sequences and the *D. melanogaster* release 6 reference genome (Bao,

Kojima, and Kohany 2015; Hoskins et al. 2015; Grenier et al. 2015; McGurk, Dion-Côté, and Barbash 2020), using *ConTExt* following the parameters in (McGurk and Barbash 2018). From this output we estimated the copy number of each position for every TE consensus from the read depth, as described in (McGurk, Dion-Côté, and Barbash 2020), and used the read pile-ups to calculate allele frequencies. Copy-number estimates and allele frequencies were generated for each of the 85 short-read libraries for each TE consensus. For Long Terminal Repeats (LTRs), or Perfect Near-Terminal Repeats (PNTRs) sequences the consensus sequence for the repeat unit is too short for copy-number estimation from read depth, so in these cases we used the median copy number from the internal sequence as the estimate of copy number. Additionally, we appended the LTR/PNTR copy number, and allele frequency data to the end of the internal sequence in order to be able to infer SNPs on the LTR/PNTR that co-occur with internal SNPs.

Filtering reads by mapping quality.

When creating the copy number and allele frequency matrices, reads aligned to the TE consensus sequences were filtered for mapping quality as described in (McGurk, Dion-Côté, and Barbash 2020). In brief, rather than using the Bowtie2 mapping quality scores we derived our own metric of filtering ambiguous reads based on the percent identity of the read to the primary (AS) and secondary (XS) alignments. We chose to filter reads in this way because we expect that many reads will be diverged from the consensus if they are derived from polymorphic elements, and we would like to retain that information. We first convert the alignment score of the read alignments to the percent identity to the consensus by assuming all penalties are due to mismatches, and then use these percent identities for AS and XS to compute a score:

$$M = \frac{(AS - XS)^2}{1 - XS}$$

Which reflects the distance between the primary (AS) and secondary (XS) alignment penalized by the divergence of AS to the consensus. If secondary alignments are reported by Bowtie2 we require this score to be greater than 0.05 for the alignment to be included in the analysis. If only a primary alignment is found the alignment must be less than 20% diverged from the consensus to be included.

Calculating sequence diversity of TE families.

From the copy number and allele frequency data derived from the read alignments to the TE consensus we calculated the sequence diversity at each position in the alignment. We multiplied the copy-number matrices by the allele frequency data to generate the estimated number of copies for all alleles across the sequence of the TE consensus, and removed alleles with a copy number < 0.5 as we assumed these low values reflected sequencing errors. We calculated the allele copy number of a TE family for each strain's alignments, as well as pooling allele copy number for all strains belonging to the same population (*e.g.* all Beijing strains), and pooling all strains to obtain global allele copy-number data. We next estimated sequence diversity at each position for each strain, population, and the entire dataset using the allele copy-number data as:

$$\pi = 1 - \sum_{nt \epsilon [A,T,C,G]} (\frac{X_{nt}}{N})^2$$

Where *N* is the total copy number at that position and X_{nt} is the copy number of an allele. When calculating the sequence diversity of the piRNA reads we performed the same procedure, but used a matrix of read counts rather than copy number and did not include any alleles with a copy number < 0.5.

Inferring TE Clades:

We developed a method to infer the co-occurence of SNPs within a TE sequence by finding positive correlations in copy numbers between SNPs across multiple individuals. We performed this inference on a set of 41 recently active TEs (Supplemental File 1). For this approach we only included positions within a TE that had a within-population sequence diversity > 0.1; or had an overall sequence diversity > 0.1. This would be equivalent to filtering out positions where the major allele is present in greater than 95% of copies. After initial diversity filtering, we obtained the copy number estimates of each allele by taking the proportion of reads that mapped to each allele and multiplying it by the estimated copy number at that position. For each position we determined the major allele as being the allele with highest copy number across the entire dataset and then extracted the copy number of the three minor alleles for every strain at every position. The result of this is an *S* x *N* matrix, where *N* is the number of minor alleles that passed our diversity criteria and *S* is the number of strains in the dataset. Each element of this matrix contains the copy number estimates for that allele for each strain. To reduce the rate of false positive correlations caused by low-copy-number alleles, we required that an allele must be present in at least 10 strains to be considered. Additionally, because we are only interested in high frequency alleles, we required alleles to be present in at least 10% frequency either across the GDL, or within a population.

Additionally, we removed strains from the *S* x *N* matrix that were determined to be outliers in copy number as determined from (McGurk, Dion-Côté, and Barbash 2020). These outliers do not represent the natural variation in copy number and instead represent TE copy-number expansions that likely occurred during the inbreeding process of the strain. These massive expansions break assumptions of our method by allowing situations where distinct TE subfamilies may co-expand in copy number and correlate while not existing on the same TE sequence. We perform this data processing for each active TE of interest.

To identify lineage-informative SNPs we perform Hierarchical Clustering with average linkage using a correlation distance on the *S* x *N* matrix (Using the R package *pheatmap*). This clusters together alleles that correlate in copy number. We use this to both seriate a correlation matrix of the alleles, and to directly call clusters by cutting the dendrogram at a correlation distance optimized for each TE, as described in *Choosing a Distance Cut Off for Hierarchical Clustering.* Clusters of minor alleles with more than one allele are lineage-informative SNPs that can be used to distinguish TE clades.

Choosing a Distance Cut Off for Hierarchical Clustering

To justify the correlation cut-off criteria in our Hierarchical Clustering, we generated a null distribution of correlations by permuting the order of the rows of each column of the filtered sets of minor alleles and calculating the pairwise correlation of these permuted SNPs. We performed this operation 1,000 times for each TE. We calculated the pairwise correlations of the un-permuted filtered sets of minor alleles and denoted this as the Test distribution. Due to the large number of pairwise comparisons performed in the clustering (869,042 pairwise correlations) we sought to correct the false positive rate by performing Bonferonni correction on our critical value ($\alpha = 0.05$) by dividing α by the number of pairwise comparisons. This critical value is subtracted from 100 to obtain a percentile that we use to determine cutoffs in the hierarchical clustering, a so-called “Critical Percentile”. We pooled the Test and Null distributions across all TEs of interest and computed the correlation value at the “Critical Percentile” in the Null distribution, r = 0.59 (Supplemental Figure 1a). Although stringent, we found that 6.38% of the SNPs in the Test distribution had a r > 0.59.

We further sought to justify our cut-off by examining the individual Null and Test distributions for each TE and comparing the Null distributions between elements (Supplemental Figure 1b). We observed that there was some degree of variability of the Test and Null distributions for each of the TEs. The “Critical Percentile” of the Null distributions fell between a correlation value of ~0.43 - 0.93 across the samples. Therefore, we optimized the hierarchical clustering distance cutoff for each TE by setting it to the correlation at the “Critical Percentile” for each Null distribution.

PacBio Data and Alignment of TE Consensus

For analysis of PacBio data we used 13 DSPR founders, the OregonR PacBio genome, and five PacBio genomes from the GDL (Chakraborty et al. 2019; Long et al. 2018). RepBase consensus sequences of 41 TEs of interest were processed by substituting ambiguous base calls with a non-ambiguous nucleotide, and then aligned to the PacBio genomes using BLASTn (Bao, Kojima, and Kohany 2015; Camacho et al. 2009; S. F. Altschul et al. 1997; Stephen F. Altschul et al. 1990). For retrotransposons with LTRs or PNTRs we aligned only the internal sequence to simplify the amount of downstream processing of the alignments, and because the majority of SNPs reside in the internal sequence. Alignments were output as XMLs to be analyzed downstream. We also extracted the sequences of each of the alignments as fasta files to be used to construct phylogenies (Supplemental File 2).

Constructing TE Phylogenies from PacBio Data:

We constructed phylogenies of TEs by using TE fasta sequences extracted from PacBio genomes, and then annotated the tips of the phylogeny with inferred clades from the GDL short-read data. We constructed phylogenies for all 41 TEs analyzed (Supplemental File 3). We first extracted fasta sequences of each insertion in the PacBio genomes by taking the sequences from the alignments described above. We excluded TE sequences that were less than 75% full length and then generated a multiple sequence alignment of the remaining sequences using clustalOmega (Sievers et al. 2011). A phylogeny of the sequences was constructed using maximum likelihood and model fitting with the tool iqTree2. A consensus tree was built using 1,000 bootstrap replicates (iqtree2 -s {input} -bb 1000) (Minh et al. 2020; Hoang et al. 2018; Kalyaanamoorthy et al. 2017). We used this consensus tree to generate cladograms, queried the sequences of the tree for inferred clades, and then colored in the tips of the phylogeny if they contained greater than 50% of the SNPs present in the cluster (Supplemental File 3).

Validation of GDL clades using PacBio genomes

We used the alignments of TE consensus sequences to the PacBio genomes to recover our clades inferred from the GDL short-read data. To do this we first needed to extract phased haplotypes from our alignments. We took the alignments and recorded the position on the consensus and the alleles found in each TE insertion of the PacBio genomes. We accounted for gaps created by insertions and deletions by correcting the position, or adding missing values, respectively. The result is a sequence for each alignment that records the position of the nucleotide found in the PacBio genome relative to the position in the RepBase consensus sequence, and the allele found at that position.

We then checked lineage-informative SNPs discovered from the GDL short-read data in the PacBio TE insertions by querying the SNPs against the PacBio insertions for a given TE. For each clade we removed PacBio alignments if they had a deletion at one of the positions of a lineage-informative SNP. On occasions when a SNP was not found in any alignment, that SNP was removed from the analysis. Additionally, we removed clades from the validation if less than two of the SNPs were detected in the PacBio data, because we wanted to be able to assay our ability to infer linkage between detected SNPs. A total of 1,719 out of 4,383 alleles were removed from the PacBio analysis. The proportion varied among TEs, with some elements like *Tart-A* or *P-element* having 95-100% of SNPs missing while other elements like *Doc* had no SNPs missing.

We justify the absence of these alleles as a consequence of the difference of power between the 85 GDL genomes used to make the initial inference, and the 19 PacBio genomes that were used to confirm the method. In addition to power issues, this validation is limited by the differences in population structure between the PacBio genomes and the GDL data.

Although five of the PacBio genomes, B59, I23, N25, T29A, and ZH26, came from GDL populations, the other 14 genomes are from the DSPR. It is possible that alleles that are present at low frequency within the GDL would not be present in the DSPR, because the alleles are population specific. It should also be noted that for retrotransposons only their internal sequence was aligned to the PacBio genome and so SNPs that reside on the LTR/PNTR were not used in this verification.

We recorded the frequency at which each of the complete sets of lineage-informative SNPs was found in its entirety in the PacBio sequences and found that approximately 70% of the clades inferred from the GDL short-reads were found within the PacBio alignments (Fig 2a). We also performed the same analysis on the complete set of clades and found that while the total number of clades detected decreases to 38%, the trends on the percent clades validated for each TE are similar (Supplemental Figure 3a). Missing clades not found in the PacBio genomes may reflect several distinct technical and biological issues. Firstly, some sets of clades do not have perfect linkage between all of their SNPs. This can occur when related lineages that share SNPs are segregating within the population. The clustering algorithm is unable to distinguish these multiple lineages and clusters them together, because they share a significant portion of their SNPs. The other possibility is that the lineage, or a subset of SNPs in the lineage, are rare or specific to a population in the GDL, and were not sampled in the PacBio genomes.

To more finely describe the co-occurence of SNPs in clades, we computed the pairwise Jaccard Score of SNPs within and between inferred clades. The Jaccard Score is computed as the number of times two SNPs occur together in a TE insertion in the PacBio genomes divided by the total number of times that either one or both SNPS are present. We converted these scores into a Jaccard Distance by simply calculating 1 - Jaccard Score, such that two SNPs that always co-occur will have a distance of 0 while two that never co-occur have a distance of 1.

We used these Jaccard Distance matrices to quantify the cohesiveness and separation of lineage-informative SNPs by computing Silhouette Scores. Silhouette Scores are a common metric for evaluation clustering performance and are calculated as being the mean distance between an individual in a cluster and its other cluster members subtracted by the mean distance between this individual and the members of the closest neighboring cluster. These values are then normalized such that they are bounded between -1 and 1, where positive scores imply that SNPs within a cluster, or clade, co-occur with each other more often than they co-occur with SNPs from a neighboring clade, and negative scores imply the opposite.

By comparing the Silhouette Score of clades to their frequency in the PacBio data, we can classify clades in four ways (Supplemental Figure 3b). "Full clade": the Silhouette Score is a positive value, and the frequency is greater than 0 *i.e.* clustering quality is good, and this arrangement of SNPs is found in the validation data. "Multiple derived lineages": the Silhouette Score is positive, but the frequency is 0. In this case we reason that the sets of alleles co-segregate in multiple lineages that share SNPs with each other, and the algorithm merges these multiple lineages into one cluster. "Incomplete clade", the Silhouette score is negative, but the frequency is greater than 0. These arrangements of SNPs are found to exist in the validation data, but may be a result of splitting lineages, or can reflect a high degree of relatedness with other clusters. "Errors", the Silhouette Score is negative and the frequency is 0. In these cases the SNPs are found in the validation set, but do not seem to co-exist on the same TE sequence.

Using these classifications we find that 41.4% of inferred clades were "Full clades", 5.5% were "Multiple derived lineages", 28.4% were "Incomplete clades", and 24.7% were "Errors". "Errors" may be a result of erroneous clustering, but likely reflect a mismatch between the SNPs found in the PacBio dataset and the short-read dataset, as nearly 39% of SNPs found in the GDL population were not found in the PacBio dataset. We also find that "Error" clades are larger (composed of more SNPs) than "Full clades" and "Incomplete clades", but not "Multiple derived lineages" (*Wilcoxon rank sum test)*. "Error" clades also tend to have more SNPs missing in the

PacBio dataset than “Incomplete clades” (*Wilcoxon rank sum test: p-value < 0.05)*, and although have a higher mean number of SNPs missing than “Full clades” (2.3 SNPs vs 0.5 SNPs) this difference was not significant.

As a qualitative assessment of the clustering we generated graphs of the SNPs for TEs where each node is a SNP, and the edge weights are the Jaccard Distance between SNPs. The SNPs are then colored by which clade they belong to (Supplemental File 4). Using *Jockey* as an example, short edge lengths connect the SNPs of “Cluster_3”, giving us strong confidence that this is a “Full clade” as expected, while the SNPs in “Cluster_12” are disconnected and fall into the “Error” category (Supplemental Figure 3c). Interestingly, we see that “Cluster_7” is centrally located in the graph with strong connections to many other SNPs. This high degree of connectivity with other clusters implies that it is an ancestral variant that has given rise to the other variants seen in the graph. In support of this hypothesis we find that “Cluster_7” is at ~40% frequency in the GDL, and is present in nearly all Jockey insertions.

As an additional qualitative assessment we examined the annotated phylogenies of TE sequences from the PacBio alignments (Supplemental File 3). We expect that clades will cluster together along branches, and that ancestral variants will be uniformly distributed across the branches -- meaning the older variants will be more widely shared. Generally, we find that our expectations hold true: older clades are widely distributed across the tree and are at higher frequency, while TEs that share newer, lower frequency clades cluster together. The tree for Jockey is shown as an illustrative pattern of the expected structure (Supplemental Figure 2d). We find that “Cluster_3”, “Cluster_8”, and “Cluster_9” form cohesive groups within the phylogeny, and that “Cluster_7” seems to be an ancestral variant that is shared among all the TE sequences in this tree.

Simulating Artificial Clades from Phylogenies

To generate the sequences of artificial clades we first used a birth-death process to generate a topology of the evolutionary history of a TE (R package *treeSim) (Love, Huber, and Anders 2014; Stadler 2011)*. We reason that transposition events in a TEs evolutionary history can be considered births, while a deactivating mutation or excision, would be equivalent to the extinction of a lineage. We used a birth rate of 1×10^{-4} and a death rate of 1×10^{-5} (Le Rouzic, Payen, and Hua-Van 2013). We simulated 20,560 generations of this process which generated a tree with 2,500 extant tips and 248 extinct tips -- a large but manageable number of sequences. We retained the extinct tips in the topology as they would represent TEs that are no longer active but still segregate in the population. We used this topology to generate sequences evolving neutrally by generating a random ancestral sequence of length 3,000 bp and dropping mutations via a Poisson process along the branch lengths with a mutation rate of 1×10^{-7} and no recombination, thus generating sequences for each tip (R package *simSeq*) (Schliep 2011). For a population of 85 individuals (the same number of individuals as in the GDL sample) we generated a copy-number distribution by drawing each individual's copy number from a Poisson distribution with a mean copy number of 25. We then used this distribution to randomly sample from all extant and extinct lineages with replacement (Supplemental File 5).

Simulations of Short-Reads from Artificial Clades

We aimed to simulate data that would be obtained from short-read libraries generated genomes harboring TE insertions that were aligned to a consensus sequence using ConTExt. We used arrays of known sequences that “reside” in each simulated individual in our population to generate TE copy number, and an allele proportion matrix.

TE copy number is simply the number of copies of an artificial TE that an individual has within their “genome”. The allele proportion matrix contains the proportion of artificial TE sequences that contain an A, T, C, or G at a given position for each strain plus pseudocounts added to represent a 0.1% Illumina sequencing error.

We then used these two reference files to simulate allele copy number pileups that replicate the inputs we used for our analysis of the GDL short-read data. For each strain we generated the coverage of our simulated library by drawing from a Poisson distribution with a target coverage as our lambda parameter:

$$E(R) \sim Pois(\lambda = Coverage)$$

We call the values drawn from this distribution our Expected Reads, E(R). We then use E(R) to generate the observed number of reads, O(R), that map to a given position of our TE. We reason that the number of reads observed at a position would be the E(R) multiplied by the TE copy number, CN. Therefore, we draw O(R) from another Poisson distribution where lambda is E(R) times the CN:

$$O(R) \sim Pois(\lambda = E(R) \times CN)$$

With the O(R) obtained for all positions of a TE for a given simulated library we now will use this to estimate the observed copy number, O(CN). We do this by adapting methods of copy number estimation, but instead of estimating E(R) with library specific parameters, we use our known E(R) from our simulated library (McGurk, Dion-Côté, and Barbash 2020). In short we divide O(R) by E(R) to obtain our O(CN):

$$O(CN) = \frac{O(R)}{E(R)}$$

We now randomly sample O(R) number of reads from a multinomial distribution parameterized by the allele proportion matrix, thereby generating read counts that map to A, T, C or G. We use the proportion of reads that map to each nucleotide to generate a mapped allele proportion matrix that we multiply to O(CN) to obtain the number of copies observed for each allele. This was output as our final simulated allele copy-number matrix where we have recorded the copy number of each allele of a TE for 85 simulated libraries (Supplemental File 5). These simulated data are identical in structure to the data structure that was used to infer clades from the GDL short-read data. We used our clade inference pipeline described in the above sections to infer clades using the same population specific parameters as the GDL ($\pi >$ 0.1, Population Frequency > 10%). We examined the effect of correlation clustering cut offs at 0.1 intervals from -1 to 1.

Simulations of sequence evolution and copy-number data to benchmark clade inference:

We benchmarked the performance of our clade inference method using the aforementioned simulations of sequence evolution to create artificial TE sequences segregating in a simulated population. We used the artificial sequences as a validation set for the clade inferences by calculating the frequency of inferred clades in the validation set, and Silhouette Scores for each clade. To examine the effect that clustering correlation cut-offs had on inference quality, we inferred clades from the same simulated data set using correlation cut-offs ranging from -1 to 1 in 0.1 intervals and computed an average Silhouette Score for the run. We found that relaxed correlation cut-offs, *i.e.* $r < 0$, produced a singular cluster that encompassed the entire set of SNPs. We assigned the lowest possible score, -1, to these results as they are uninformative. Conversely, a very stringent cut-off, *i.e.* $r = 1$, produced no clusters and also was assigned a -1 score. The intermediate scores between 0 and 0.9 were the most informative. We found that Silhouette Scores steadily increase as the cut-off stringency increases until reaching a peak at $r = 0.8$ (Figure 2b). This result shows that increasing clustering stringency produces tighter, more cohesive clusters, as would be expected, until a limit is reached and an under-clustering behavior emerges.

To finely describe the quality of individual clades we used both frequency in the validation set and Silhouette Scores to classify them (Supplemental Figure 4a). The relaxed clustering cutoffs, *i.e. r* = (0, 0.1, 0.2), have more “Multiple Derived lineages” as would be expected, while if the cut-off is more stringent, *r* = 0.9, we see an “Incomplete Clade”. However, in general we see many high frequency “Full Clades” called at a wide range of clustering parameters, but the optimum seems to be, *r* = 0.7-0.8 (Supplemental Figure 4a).

Our final assessment of clustering quality was to annotate the underlying phylogeny that was used to generate the sequences with the clades that were inferred from the simulated data. We took the results from one of the optimal clustering cut-offs, *r* = 0.7, and colored the tips of the phylogeny if that sequence contained more than 50% of the SNPs that form a cluster. We found that the tips that contain the same clades form groupings in the phylogeny as expected. Additionally, we can see that “Cluster_3” is an older clade that is more widely distributed along the phylogeny, as was observed for old clades in the real TE sequences (Supplemental Figure 4b).

These simulations allowed us to generate a realistic dataset to validate our clade inferences. We find in general that clustering behaves better under more stringent clustering cut-off parameters (Supplemental Figure 2b, Supplemental Figure 4a), and our inferred clades are correctly identifying phylogenetic relationships (Supplemental Figure 4b).

Processing and aligning small RNA data

Public piRNA libraries were all created from female *D. melanogaster* ovaries, and are available through the SRA (see SRA acessions). We obtained libraries from 10 GDL strains (two from each population) (Luo et al. 2020). piRNA reads were trimmed using *Trimmomatic* and aligned to an index of curated RepBase repeat consensus sequences using *Bowtie2* using the parameters: *-N 1 -L 10 -i S,1,0.5 -p 8 --score-min L,0,-1.2 -D 100 -R 5* (Langmead et al. 2019; Langmead and Salzberg 2012; Bolger, Lohse, and Usadel 2014; Bao, Kojima, and

Kohany 2015; McGurk, Dion-Côté, and Barbash 2020). After alignment, reads were filtered by base quality (Q > 30), by size (21-30 base pairs) and by mapping quality as described for the genomic data in the above sections. From the remaining reads we generate SNP read pileups with the python module *pysam* using the *pileup* function (https://github.com/pysam-developers/pysam), akin to *samtools mpileup* (Li et al. 2009). We separated reads out by sense and antisense to get SNP pileups derived from the secondary piRNA pathway, and the primary piRNA pathway, respectively. The result is a matrix containing the number of sense and antisense reads that map to each position and that read's nucleotide at that position. After generating the matrices, we used a Size Factor Normalization approach to normalize the total read depth of all repeats that reads were aligned to. We generated a table of read counts for each TE from the SNP pileups, and then followed the protocols described by DESeq2, but used a custom script to handle our unique data structure (Love, Huber, and Anders 2014). The normalized read depth SNP pileups were used as the primary data for all piRNA analyses in this study. To calculate piRNA read depth of each clade we averaged the sense and antisense piRNA read depth across all alleles of each clade across the strains. We then added pseudocounts of one to the clade piRNA read depth and to the clade copy number of the strains before computing piRNAs/copy. This was done to regularize data for log-transformation. We used these values to calculate the average sense and antisense piRNA read depth per clade copy across the 10 GDL strains.

Data and code availability:

Short-read, PacBio and piRNA data used in this study were previously published and are publicly available: GDL NGS libraries are available under SRA accession SRP050151 (Grenier et al. 2015). GDL PacBio genomes are available under SRA accession SRP142531 (Long et al. 2018), and DSPR PacBio genomes are available under BioProject accession PRJNA418342 (Chakraborty et al. 2019). GDL piRNA data is available under SRA accession SRP068882 (Luo et al. 2020). Code and processed data used to infer clades are publicly available through Github (https://github.com/is-the-biologist/TE_CladeInference).

Acknowledgements

We thank Clayton Hughes for help with the piRNA alignments and Satyam Srivastav for invaluable discussion on the epigenetic inheritance of piRNA clusters.

Figure 1. Outline and examples of the TE clade inference method. (a) Cartoon depicting the method. The genomes of three individuals (orange rectangles) contain copies of an ancestral TE (blue rectangles) and a derived TE with SNPs A and B (blue rectangles with a red and purple stripe, respectively). As the copy number of the derived TE varies in copy number across individuals 1-3, so does the copy number of SNPs A and B. This relationship in copy number is depicted as a cartoon scatterplot, where each red dot represents the copy number of SNPs A and B in one of nine individuals sampled in the population. The copy number of the SNPs is positively correlated because the SNPs are physically linked. (b) Scatterplot depicting the correlation in copy number across GDL individuals for two SNPs in the *Jockey* element. Each dot represents the copy number of the SNPs C at position 2238 (C_2238) and C at position 2402 (C_4204) for each individual, colored by their population of origin. The degree of correlation of these two SNPs is high (*Pearson's r = 0.82*), suggesting that they are physically linked and represent a clade. Black dashed line is a linear fit of the data drawn for emphasis. (c) Heatmap showing correlation of the copy number of all SNPs from the *Jockey* element. Cells in the heatmap are seriated via hierarchical clustering to create clusters of tightly correlated SNPs, which are inferred to be *Jockey* clades segregating in the population. The cells are shaded by the pairwise Pearson's correlation between SNP copy number. The SNPs from (b) are outlined in a block box. (d) The percent of clades inferred from GDL data that were then detected in a set of PacBio genomes (includes only clades where at least two SNPs were detected at any

frequency in the PacBio data). The results are separated by TE family, with total clades shown on the far right. Fraction of clades validated over the total number of clades found are placed above each bar.

Figure 2. Summary statistics of TE clades inferred from GDL short-reads. (a) Average nucleotide diversity for each TE family vs. the number of clades inferred for that family, colored by TE class. (b) The number of phase-informative SNPs that compose each clade inferred for every TE family. Each point represents a clade and is colored by TE class. (c) Histogram of the population frequency of all clades from 41 recently active TE families. (d) Boxplots of the clade population frequency of all clades separated by TE family. Each point represents a clade and is colored by TE class.

Figure 3. Age of clades are inferred by their copy-number distribution. (a) Mean-variance relationship of the clade copy-number distributions for clades from all families. The copy-number distributions for each clade were tested for goodness of fit to a Poisson distribution, and then colored based on acceptance or rejection of this test: “overdispersed” (rejected, red), “dispersed” (fail to reject, yellow), or “underdispersed” (rejected, purple). (b) Population frequency of young (red: “dispersed" and “overdispersed”) or old (purple: “underdispersed”) clades across all TE families. (c) Number of discovered clades per TE family that are young (red: “dispersed" and “overdispersed”) or old (purple: “underdispersed”). (d) Boxplot of the copy-number distribution of the sole putatively active *I-element* clade from (c) for each GDL population. There is a significant elevation in the copy number of this clade in Beijing.

Figure 4. Population structure and variation of TE clades is common. (a) Boxplot showing the result of Kruskal-Wallis tests on the clade copy number between GDL populations for each TE family. Each dot represents the negative log base-2 transformed p-value for a single clade. Red dashed line is the Bonferonni corrected critical value. 15% of the clades had a p-value less

than the critical value, and showed heterogeneity in copy number between populations (*Bonferonni correction:* =0.05919). (b) Each dot represents the average copy number of a clade across GDL and the average copy number of the population that is most differentiated from the GDL average. Most differentiated is defined as the greatest absolute difference between the population mean and the GDL mean. Clades are colored by whether they are statistically significant by Kruskal-Wallis test (sig., red), or not (n.s., grey). (c) PCA on the minor allele frequency of *Roo* element SNPs in the GDL. Each dot represents the principal components derived from the minor allele frequencies of an individual. Beijing (red), and Zimbabwe (purple) clusters can be seen. (d) Boxplot of copy number of a *Roo* clade enriched for Beijing (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe). (e) Boxplot of copy number of a different *Roo* clade enriched for Zimbabwe. (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe).

Figure 5. piRNA diversity and the average sense and antisense piRNAs/copy of TE clades for 10 GDL strains. (a) Scatterplot where each point represents the sense (+, red) or antisense (-, blue) piRNA sequence diversity ($\boldsymbol{\pi}_{\text{piRNA}}$) for a TE family plotted against the genomic TE sequence diversity ($\boldsymbol{\pi}_{\text{TE}}$) of the TE family. The grey dashed line represents the 1:1 expectation of piRNA diversity:genomic diversity and the black dashed line represents a linear fit between the piRNA diversity and genomic diversity. (b) Average clade piRNAs/copy for sense (+, red), and antisense (-, blue) separated by TE class. Significant differences between sense and antisense piRNAs/copy were found in clades for LTR and LINE-like elements (sig.), but not telomeric or DNA transposons (n.s. ; *Wilcoxon signed-rank test*) . (c) piRNAs/copy of putatively young clades (likely active) and old clades (likely inactive). Young clades had greater piRNAs/copy than older clades (+, *Dunn's test p-value* < 0.05; -, *Dunn's test p-value* < 0.05). (d) Spearman's correlation calculated for copy number and piRNA read depth for putatively young and old clades. Young clades had a greater Spearman's correlation than inactive clades for sense and antisense piRNA read depth (+, *Dunn's test p-value* < 0.05; -, *Dunn's test p-value* < 0.05). (e) Copy number vs. sense (+, red) and antisense (-, blue) piRNA read depth for a young, recently active *Jockey* clade, and (f) for an old, putatively inactive *I-element* clade.

Supplemental Figure 1. Null and test distributions used to choose optimal correlation cut offs for hierarchical clustering. Null distributions were constructed by permuting the allele copy number matrices for each TE family 1,000 times and recording all pairwise correlations between permuted alleles. The test distributions were computed by performing the pairwise comparisons of the allele copy number matrices for each TE family. (a) Aggregations of all test distributions (orange) and null distributions (teal) for all TE families. (b) Boxplots of individual test distributions (orange) and null distributions (teal) for each TE family.

Supplemental Figure 2. Clade inference method validated using PacBio genomes and simulations. (a) Average pairwise correlation of allele copy number between TE families and within TE families from GDL short-reads. (b) Clustering quality (Silhouette Score) of inferred clades from simulated polymorphism data. Clustering accuracy was high when the correlation cutoff for clustering was ≥ 0.3 and ≤ 0.9.

Supplemental Figure 3. Additional validation of GDL clades using PacBio data. (a) The percent of all clades inferred from GDL data that were then detected in a set of PacBio genomes (including those with no SNPs detected in the PacBio data). The results are separated by TE family, with total clades shown on the far right (colors are for visual separation and are not quantitative). (b) Each clade is classified by a Silhouette Score and insertion frequency in the PacBio genomes: Full clade (Silhouette Score > 0, Frequency > 0; green), Incomplete clade (Silhouette Score =< 0, Frequency > 0; orange), Multiple derived lineages (Silhouette Score > 0, Frequency = 0; blue), and Errors (Silhouette Score =< 0, Frequency = 0; red). Results are separated by TE family. (c) Network graph showing Jaccard distances between SNPs and their corresponding clades, for the *Jockey* element. Each node is a SNP and is colored by the clade it belongs to. Edge weights are the Jaccard distance between SNPs calculated from insertions in the PacBio genome. (d) Cladogram of *Jockey* insertions that are >=75% full-length from all PacBio genomes. Tips of the cladogram are colored by the clade(s) they contain (an insertion is considered to correspond to a clade if it has > 50% of the SNPs that define the clade). Insertions with similar composition of clades cluster together within the cladogram.

Supplemental Figure 4. Analysis of missing lineage-informative SNPs in validation. Number of missing lineage-informative-SNPs (a) and total number of lineage-informative SNPs (b) for each inferred clade. Each point represents a clade and is classified by a Silhouette Score and insertion frequency in the PacBio genomes: Full clade (Silhouette Score > 0, Frequency > 0), Incomplete clade (Silhouette Score =< 0, Frequency > 0), Multiple derived lineages (Silhouette Score > 0, Frequency = 0), and Errors (Silhouette Score =< 0, Frequency = 0).

Supplemental Figure 5. Benchmark of clade inference method using polymorphism data from a simulated phylogeny. (a) Clades are classified by a Silhouette Score and insertion frequency in the simulated reference sequences: Full clade (Silhouette Score > 0, Frequency > 0; green), Incomplete clade (Silhouette Score =< 0, Frequency > 0; orange), Multiple derived clades (Silhouette Score > 0, Frequency = 0; blue), and Errors (Silhouette Score =< 0, Frequency = 0; none shown). Classification was done on correlation cutoffs from -1 to 1, but only presenting informative runs from -1 - 0.9. (b) Cladogram of a simulated sequence phylogeny used to generate polymorphism data. Tips of the cladogram are colored by the clade(s) they contain (a sequence is considered to contain a clade if it has > 50% of the SNPs that define the clade).

Supplemental Figure 6. PCAs of minor allele frequencies from TEs that show strong population structure. (a) PCA on the minor allele frequency of *Tirant* elements reveals an

Ithaca cluster (teal) and a Tasmania cluster (orange). (b) Boxplot of clade copy number of the *Tirant* clade enriched for Tasmania (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe). (c) Boxplot of clade copy number of a *Tirant* clade enriched strongly for Ithaca, and moderately for Netherlands (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe). (d) PCA on the minor allele frequency of *Jockey* elements reveals a cluster for Zimbabwe (purple). (e) Boxplot of clade copy number of a *Jockey* element enriched for Zimbabwe (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe). (f) Boxplot of clade copy number of a *Jockey* element depleted in Zimbabwe (B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe).

Supplemental Figure 7. Additional piRNA diversity ratio and piRNA abundance data for recently active TEs. (a) Sense (red) and antisense (blue) piRNA diversity ratio ($\boldsymbol{\pi}_{\text{piRNA}}/\boldsymbol{\pi}_{\text{TE}}$) for all TE families from 10 GDL strains separated by class. (b) Average clade piRNAs/copy for sense (+, red), and antisense (-, blue) separated by TE family.

Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. “Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs.” *Nucleic Acids Research* 25 (17): 3389–3402.

Altschul, Stephen F., Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Aravin, Alexei A., Gregory J. Hannon, and Julius Brennecke. 2007. “The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race.” *Science* 318 (5851): 761–64.

Arkhipova, Irina R. 2017. “Using Bioinformatic and Phylogenetic Approaches to Classify Transposable Elements and Understand Their Complex Evolutionary Histories.” *Mobile DNA* 8 (December): 19.

Assis, Raquel, and Alexey S. Kondrashov. 2009. “Rapid Repetitive Element-Mediated Expansion of piRNA Clusters in Mammalian Evolution.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (17): 7079–82.

Bao, Weidong, Kenji K. Kojima, and Oleksiy Kohany. 2015. “Repbase Update, a Database of Repetitive Elements in Eukaryotic Genomes.” *Mobile DNA* 6 (June): 11.

Baucom, Regina S., James C. Estill, Jim Leebens-Mack, and Jeffrey L. Bennetzen. 2009. “Natural Selection on Gene Function Drives the Evolution of LTR Retrotransposon Families in the Rice Genome.” *Genome Research* 19 (2): 243–54.

Black, D. M., M. S. Jackson, M. G. Kidwell, and G. A. Dover. 1987. “KP Elements Repress P-Induced Hybrid Dysgenesis in Drosophila Melanogaster.” *The EMBO Journal* 6 (13): 4125–

35.

Blumenstiel, Justin P., Xi Chen, Miaomiao He, and Casey M. Bergman. 2014. “An Age-of-Allele Test of Neutrality for Transposable Element Insertions.” *Genetics* 196 (2): 523–38.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Brennecke, Julius, Alexei A. Aravin, Alexander Stark, Monica Dus, Manolis Kellis, Ravi Sachidanandam, and Gregory J. Hannon. 2007. “Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila.” *Cell* 128 (6): 1089–1103.

Brennecke, Julius, Colin D. Malone, Alexei A. Aravin, Ravi Sachidanandam, Alexander Stark, and Gregory J. Hannon. 2008. “An Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing.” *Science* 322 (5906): 1387–92.

Browning, Sharon R., and Brian L. Browning. 2007. “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies by Use of Localized Haplotype Clustering.” *American Journal of Human Genetics* 81 (5): 1084–97.

Busseau, I., M. C. Chaboissier, A. Pélisson, and A. Bucheton. 1994. “I Factors in Drosophila Melanogaster: Transposition under Control.” *Genetica* 93 (1-3): 101–16.

Camacho, Christiam, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L. Madden. 2009. “BLAST+: Architecture and Applications.” *BMC Bioinformatics* 10 (December): 421.

Chakraborty, Mahul, J. J. Emerson, Stuart J. Macdonald, and Anthony D. Long. 2019. “Structural Variants Exhibit Widespread Allelic Heterogeneity and Shape Variation in Complex Traits.” *Nature Communications* 10 (1): 4872.

Charlesworth, Brian, and Deborah Charlesworth. 1983. “The Population Dynamics of Transposable Elements.” *Genetical Research* 42 (1): 1–27.

Clark, A. G. 1990. “Inference of Haplotypes from PCR-Amplified Samples of Diploid Populations.” *Molecular Biology and Evolution* 7 (2): 111–22.

Cridland, Julie M., Stuart J. Macdonald, Anthony D. Long, and Kevin R. Thornton. 2013. “Abundance and Distribution of Transposable Elements in Two Drosophila QTL Mapping Resources.” *Molecular Biology and Evolution* 30 (10): 2311–27.

Czech, Benjamin, Marzia Munafò, Filippo Ciabrelli, Evelyn L. Eastwood, Martin H. Fabry, Emma Kneuss, and Gregory J. Hannon. 2018. “piRNA-Guided Genome Defense: From Biogenesis to Silencing.” *Annual Review of Genetics* 52 (November): 131–57.

Delaneau, Olivier, Cédric Coulonges, and Jean-François Zagury. 2008. “Shape-IT: New Rapid and Accurate Algorithm for Haplotype Inference.” *BMC Bioinformatics* 9 (1): 1–14.

Doolittle, W. F., and C. Sapienza. 1980. “Selfish Genes, the Phenotype Paradigm and Genome Evolution.” *Nature* 284 (5757): 601–3.

Dvorák, J., D. Jue, and M. Lassner. 1987. “Homogenization of Tandemly Repeated Nucleotide Sequences by Distance-Dependent Nucleotide Sequence Conversion.” *Genetics* 116 (3): 487–98.

Excoffier, L., and M. Slatkin. 1995. “Maximum-Likelihood Estimation of Molecular Haplotype Frequencies in a Diploid Population.” *Molecular Biology and Evolution* 12 (5): 921–27.

Grenier, Jennifer K., J. Roman Arguello, Margarida Cardoso Moreira, Srikanth Gottipati, Jaaved Mohammed, Sean R. Hackett, Rachel Boughton, Anthony J. Greenberg, and Andrew G. Clark. 2015. “Global Diversity Lines - a Five-Continent Reference Panel of Sequenced Drosophila Melanogaster Strains.” *G3* 5 (4): 593–603.

Hoang, Diep Thi, Olga Chernomor, Arndt von Haeseler, Bui Quang Minh, and Le Sy Vinh. 2018. “UFBoot2: Improving the Ultrafast Bootstrap Approximation.” *Molecular Biology and Evolution* 35 (2): 518–22.

Hoskins, Roger A., Joseph W. Carlson, Kenneth H. Wan, Soo Park, Ivonne Mendez, Samuel E. Galle, Benjamin W. Booth, et al. 2015. “The Release 6 Reference Sequence of the Drosophila Melanogaster Genome.” *Genome Research* 25 (3): 445–58.

Iwasaki, Watal M., T. E. Kijima, and Hideki Innan. 2020. "Population Genetics and Molecular Evolution of DNA Sequences in Transposable Elements. II. Accumulation of Variation and Evolution of a New Subfamily." *Molecular Biology and Evolution* 37 (2): 355–64.

Kalyaanamoorthy, Subha, Bui Quang Minh, Thomas K. F. Wong, Arndt von Haeseler, and Lars S. Jermiin. 2017. "ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates." *Nature Methods* 14 (6): 587–89.

Kaminker, Joshua S., Casey M. Bergman, Brent Kronmiller, Joseph Carlson, Robert Svirskas, Sandeep Patel, Erwin Frise, et al. 2002. "The Transposable Elements of the Drosophila Melanogaster Euchromatin: A Genomics Perspective." *Genome Biology* 3 (12): RESEARCH0084.

Kelleher, Erin S., and Daniel A. Barbash. 2013. "Analysis of piRNA-Mediated Silencing of Active TEs in Drosophila Melanogaster Suggests Limits on the Evolution of Host Genome Defense." *Molecular Biology and Evolution* 30 (8): 1816–29.

Kelleher, Erin S., Daniel A. Barbash, and Justin P. Blumenstiel. 2020. "Taming the Turmoil Within: New Insights on the Containment of Transposable Elements." *Trends in Genetics: TIG* 36 (7): 474–89.

Kidwell, M. G. 1983. "Evolution of Hybrid Dysgenesis Determinants in Drosophila Melanogaster." *Proceedings of the National Academy of Sciences of the United States of America* 80 (6): 1655–59.

Kofler, Robert, Viola Nolte, and Christian Schlötterer. 2015. "Tempo and Mode of Transposable Element Activity in Drosophila." *PLoS Genetics* 11 (7): e1005406.

Kofler, Robert, Kirsten-André Senti, Viola Nolte, Ray Tobler, and Christian Schlötterer. 2018. "Molecular Dissection of a Natural Transposable Element Invasion." *Genome Research* 28 (6): 824–35.

Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.

Langmead, Ben, Christopher Wilks, Valentin Antonescu, and Rone Charles. 2019. "Scaling Read Aligners to Hundreds of Threads on General-Purpose Processors." *Bioinformatics* 35 (3): 421–32.

Lee, Yuh Chwen G., and Charles H. Langley. 2010. "Transposable Elements in Natural Populations of Drosophila Melanogaster." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 365 (1544): 1219–28.

Lerat, Emmanuelle, Carène Rizzon, and Christian Biémont. 2003. "Sequence Divergence within Transposable Element Families in the Drosophila Melanogaster Genome." *Genome Research* 13 (8): 1889–96.

Le Rouzic, Arnaud, and Pierre Capy. 2005. "The First Steps of Transposable Elements Invasion: Parasitic Strategy vs. Genetic Drift." *Genetics* 169 (2): 1033–43.

———. 2006. "Population Genetics Models of Competition between Transposable Element Subfamilies." *Genetics* 174 (2): 785–93.

Le Rouzic, Arnaud, Thibaut Payen, and Aurélie Hua-Van. 2013. "Reconstructing the Evolutionary History of Transposable Elements." *Genome Biology and Evolution* 5 (1): 77–86.

Le Thomas, Adrien, Evelyn Stuwe, Sisi Li, Jiamu Du, Georgi Marinov, Nikolay Rozhkov, Yung-Chia Ariel Chen, et al. 2014. "Transgenerationally Inherited piRNAs Trigger piRNA Biogenesis by Changing the Chromatin of piRNA Clusters and Inducing Precursor Processing." *Genes & Development* 28 (15): 1667–80.

Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.

Long, Evan, Carrie Evans, John Chaston, and Joshua A. Udall. 2018. "Genomic Structural

Variations Within Five Continental Populations of.” *G3* 8 (10): 3247–53.

Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. “Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2.” *Genome Biology* 15 (12): 550.

Luan, D. D., M. H. Korman, J. L. Jakubczak, and T. H. Eickbush. 1993. “Reverse Transcription of R2Bm RNA Is Primed by a Nick at the Chromosomal Target Site: A Mechanism for Non-LTR Retrotransposition.” *Cell* 72 (4): 595–605.

Lukic, Sergio, and Kevin Chen. 2011. “Human piRNAs Are under Selection in Africans and Repress Transposable Elements.” *Molecular Biology and Evolution* 28 (11): 3061–67.

Luo, Shiqi, Hong Zhang, Yuange Duan, Xinmin Yao, Andrew G. Clark, and Jian Lu. 2020. “The Evolutionary Arms Race between Transposable Elements and piRNAs in Drosophila Melanogaster.” *BMC Evolutionary Biology* 20 (1): 14.

Lutz, Sheila M., Bethaney J. Vincent, Haig H. Kazazian Jr, Mark A. Batzer, and John V. Moran. 2003. “Allelic Heterogeneity in LINE-1 Retrotransposition Activity.” *American Journal of Human Genetics* 73 (6): 1431–37.

Makałowski, Wojciech, Valer Gotea, Amit Pande, and Izabela Makałowska. 2019. “Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics.” *Methods in Molecular Biology* 1910: 177–207.

McGurk, Michael P., and Daniel A. Barbash. 2018. “Double Insertion of Transposable Elements Provides a Substrate for the Evolution of Satellite DNA.” *Genome Research* 28 (5): 714–25.

McGurk, Michael P., Anne-Marie Dion-Côté, and Daniel A. Barbash. 2020. “Rapid Evolution at the Drosophila Telomere: Transposable Element Dynamics at an Intrinsically Unstable Locus.” *Genetics*, December. https://doi.org/10.1093/genetics/iyaa027.

Minh, Bui Quang, Heiko A. Schmidt, Olga Chernomor, Dominik Schrempf, Michael D. Woodhams, Arndt von Haeseler, and Robert Lanfear. 2020. “IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.” *Molecular Biology and Evolution* 37 (5): 1530–34.

Mohn, Fabio, Grzegorz Sienski, Dominik Handler, and Julius Brennecke. 2014. “The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila.” *Cell* 157 (6): 1364–79.

Olovnikov, Ivan, Sergei Ryazansky, Sergey Shpiz, Sergey Lavrov, Yuri Abramov, Chantal Vaury, Silke Jensen, and Alla Kalmykova. 2013. “De Novo piRNA Cluster Formation in the Drosophila Germ Line Triggered by Transgenes Containing a Transcribed Transposon Fragment.” *Nucleic Acids Research* 41 (11): 5757–68.

Orgel, L. E., and F. H. Crick. 1980. “Selfish DNA: The Ultimate Parasite.” *Nature* 284 (5757): 604–7.

Periquet, G., M. H. Hamelin, Y. Bigot, and Kai Hu. 1989. “Presence of the Deleted Hobo Element Th in Eurasian Populations of Drosophila Melanogaster.” *Genetics, Selection, Evolution: GSE* 21 (1): 107.

Picard, G., J. C. Bregliano, A. Bucheton, J. M. Lavige, A. Pelisson, and M. G. Kidwell. 1978. “Non-Mendelian Female Sterility and Hybrid Dysgenesis in Drosophila Melanogaster.” *Genetical Research* 32 (3): 275–87.

Post, Christina, Josef P. Clark, Yuliya A. Sytnikova, Gung-Wei Chirn, and Nelson C. Lau. 2014. “The Capacity of Target Silencing by Drosophila PIWI and piRNAs.” *RNA* 20 (12): 1977–86.

Radion, Elizaveta, Valeriya Morgunova, Sergei Ryazansky, Natalia Akulenko, Sergey Lavrov, Yuri Abramov, Pavel A. Komarov, Sergey I. Glukhov, Ivan Olovnikov, and Alla Kalmykova. 2018. “Key Role of piRNAs in Telomeric Chromatin Maintenance and Telomere Nuclear Positioning in Drosophila Germline.” *Epigenetics & Chromatin* 11 (1): 40.

Robillard, Émilie, Arnaud Le Rouzic, Zheng Zhang, Pierre Capy, and Aurélie Hua-Van. 2016. “Experimental Evolution Reveals Hyperparasitic Interactions among Transposable Elements.” *Proceedings of the National Academy of Sciences of the United States of*

America 113 (51): 14763–68.

Roiha, H., J. R. Miller, L. C. Woods, and D. M. Glover. 1981. “Arrangements and Rearrangements of Sequences Flanking the Two Types of rDNA Insertion in D. Melanogaster.” *Nature* 290 (5809): 749–53.

Ryazansky, Sergei, Elizaveta Radion, Anastasia Mironova, Natalia Akulenko, Yuri Abramov, Valeriya Morgunova, Maria Y. Kordyukova, Ivan Olovnikov, and Alla Kalmykova. 2017. “Natural Variation of piRNA Expression Affects Immunity to Transposable Elements.” *PLoS Genetics* 13 (4): e1006731.

Saint-Leandre, Bastien, Pierre Capy, Aurelie Hua-Van, and Jonathan Filée. 2020. “piRNA and Transposon Dynamics in Drosophila: A Female Story.” *Genome Biology and Evolution* 12 (6): 931–47.

Schliep, Klaus Peter. 2011. “Phangorn: Phylogenetic Analysis in R.” *Bioinformatics* 27 (4): 592–93.

Schwarz, Florian, Filip Wierzbicki, Kirsten-André Senti, and Robert Kofler. 2020. “Tirant Stealthily Invaded Natural Drosophila Melanogaster Populations during the Last Century.” *Evolutionary Biology*. bioRxiv.

Seleme, Maria del Carmen, Melissa R. Vetter, Richard Cordaux, Laurel Bastone, Mark A. Batzer, and Haig H. Kazazian Jr. 2006. “Extensive Individual Variation in L1 Retrotransposition Capability Contributes to Human Genetic Diversity.” *Proceedings of the National Academy of Sciences of the United States of America* 103 (17): 6611–16.

Shpiz, Sergey, Sergei Ryazansky, Ivan Olovnikov, Yuri Abramov, and Alla Kalmykova. 2014. “Euchromatic Transposon Insertions Trigger Production of Novel Pi- and Endo-siRNAs at the Target Sites in the Drosophila Germline.” *PLoS Genetics* 10 (2): e1004138.

Sievers, Fabian, Andreas Wilm, David Dineen, Toby J. Gibson, Kevin Karplus, Weizhong Li, Rodrigo Lopez, et al. 2011. “Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega.” *Molecular Systems Biology* 7 (October): 539.

Souames, Sémi, Claude Bazin, Eric Bonnivard, and Dominique Higuet. 2003. “Behavior of the Hobo Transposable Element with Regard to TPE Repeats in Transgenic Lines of Drosophila Melanogaster.” *Molecular Biology and Evolution* 20 (12): 2055–66.

Stadler, Tanja. 2011. “Simulating Trees with a Fixed Number of Extant Species.” *Systematic Biology* 60 (5): 676–84.

Szostak, J. W., and R. Wu. 1980. “Unequal Crossing over in the Ribosomal DNA of Saccharomyces Cerevisiae.” *Nature* 284 (5755): 426–30.

Wellauer, P. K., and I. B. Dawid. 1977. “The Structural Organization of Ribosomal DNA in Drosophila Melanogaster.” *Cell* 10 (2): 193–212.

Wicker, Thomas, François Sabot, Aurélie Hua-Van, Jeffrey L. Bennetzen, Pierre Capy, Boulos Chalhoub, Andrew Flavell, et al. 2007. “A Unified Classification System for Eukaryotic Transposable Elements.” *Nature Reviews. Genetics* 8 (12): 973–82.

Wierzbicki, Filip, Florian Schwarz, Odontsetseg Cannalonga, and Robert Kofler. 2020. “Generating High Quality Assemblies for Genomic Analysis of Transposable Elements.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2020.03.27.011312.

Xiong, Y. E., and T. H. Eickbush. 1988. “Functional Expression of a Sequence-Specific Endonuclease Encoded by the Retrotransposon R2Bm.” *Cell* 55 (2): 235–46.

Yang, Zhao, James W. Hardin, and Cheryl L. Addy. 2009. “A Score Test for Overdispersion in Poisson Regression Based on the Generalized Poisson-2 Model.” *Journal of Statistical Planning and Inference* 139 (4): 1514–21.

Zanni, Vanessa, Angéline Eymery, Michael Coiffet, Matthias Zytnicki, Isabelle Luyten, Hadi Quesneville, Chantal Vaury, and Silke Jensen. 2013. “Distribution, Evolution, and Diversity of Retrotransposons at the Flamenco Locus Reflect the Regulatory Properties of piRNA Clusters.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (49): 19842–47.

Zhang, Donglei, Shikui Tu, Michael Stubna, Wei-Sheng Wu, Wei-Che Huang, Zhiping Weng, and Heng-Chi Lee. 2018. “The piRNA Targeting Rules and the Resistance to piRNA Silencing in Endogenous Genes.” *Science* 359 (6375): 587–92.

Zhang, Shuo, Beverly Pointer, and Erin S. Kelleher. 2020. “Rapid Evolution of piRNA-Mediated Silencing of an Invading Transposable Element Was Driven by Abundant de Novo Mutations.” *Genome Research* 30 (4): 566–75.

Figure 1

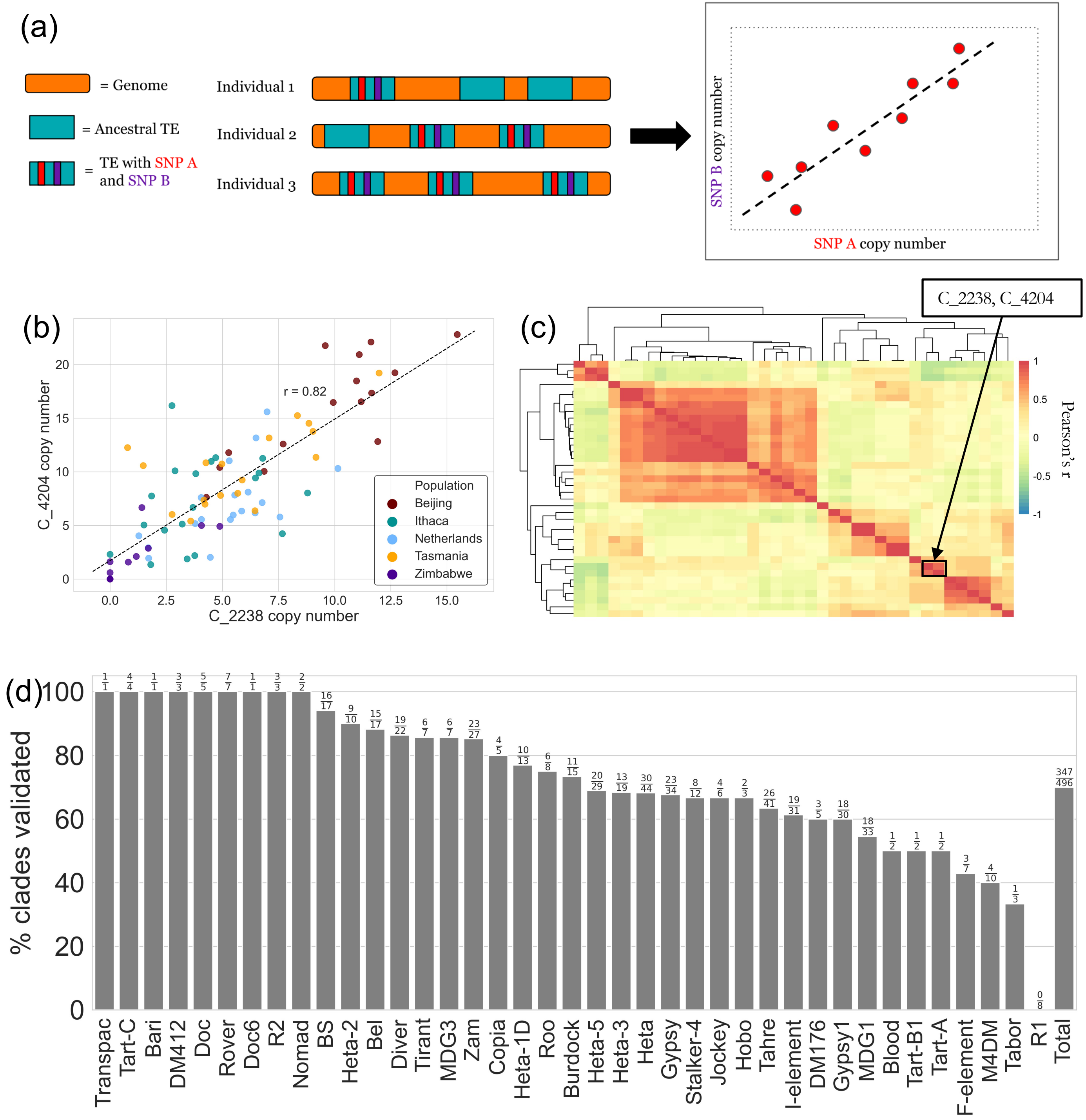

(a)
= Genome
= Ancestral TE
= TE with SNP A and SNP B
Individual 1
Individual 2
Individual 3
SNP B copy number
SNP A copy number
(b)
C_4204 copy number
C_2238 copy number
r = 0.82
Population
Beijing
Ithaca
Netherlands
Tasmania
Zimbabwe
0
5
10
15
20
0.0
2.5
5.0
7.5
10.0
12.5
15.0
(c)
C_2238, C_4204
Pearson's r
1
0.5
0
-0.5
-1
(d)
% clades validated
0
20
40
60
80
100
Transpac 1/1
Tart-C 4/4
Bari 1/1
DM412 3/3
Doc 5/5
Rover 7/7
Doc6 1/1
R2 3/3
Nomad 2/2
BS 16/17
Heta-2 9/10
Bel 15/17
Diver 19/22
Tirant 6/7
MDG3 6/7
Zam 23/27
Copia 4/5
Heta-1D 10/13
Roo 6/8
Burdock 11/15
Heta-5 20/29
Heta-3 13/19
Heta 30/44
Gypsy 23/34
Stalker-4 8/12
Jockey 4/6
Hobo 2/3
Tahre 26/41
I-element 19/31
DM176 3/5
Gypsy1 18/30
MDG1 18/33
Blood 1/2
Tart-B1 1/2
Tart-A 1/2
F-element 3/7
M4DM 4/10
Tabor 1/3
R1 0/8
Total 347/496

Figure 2

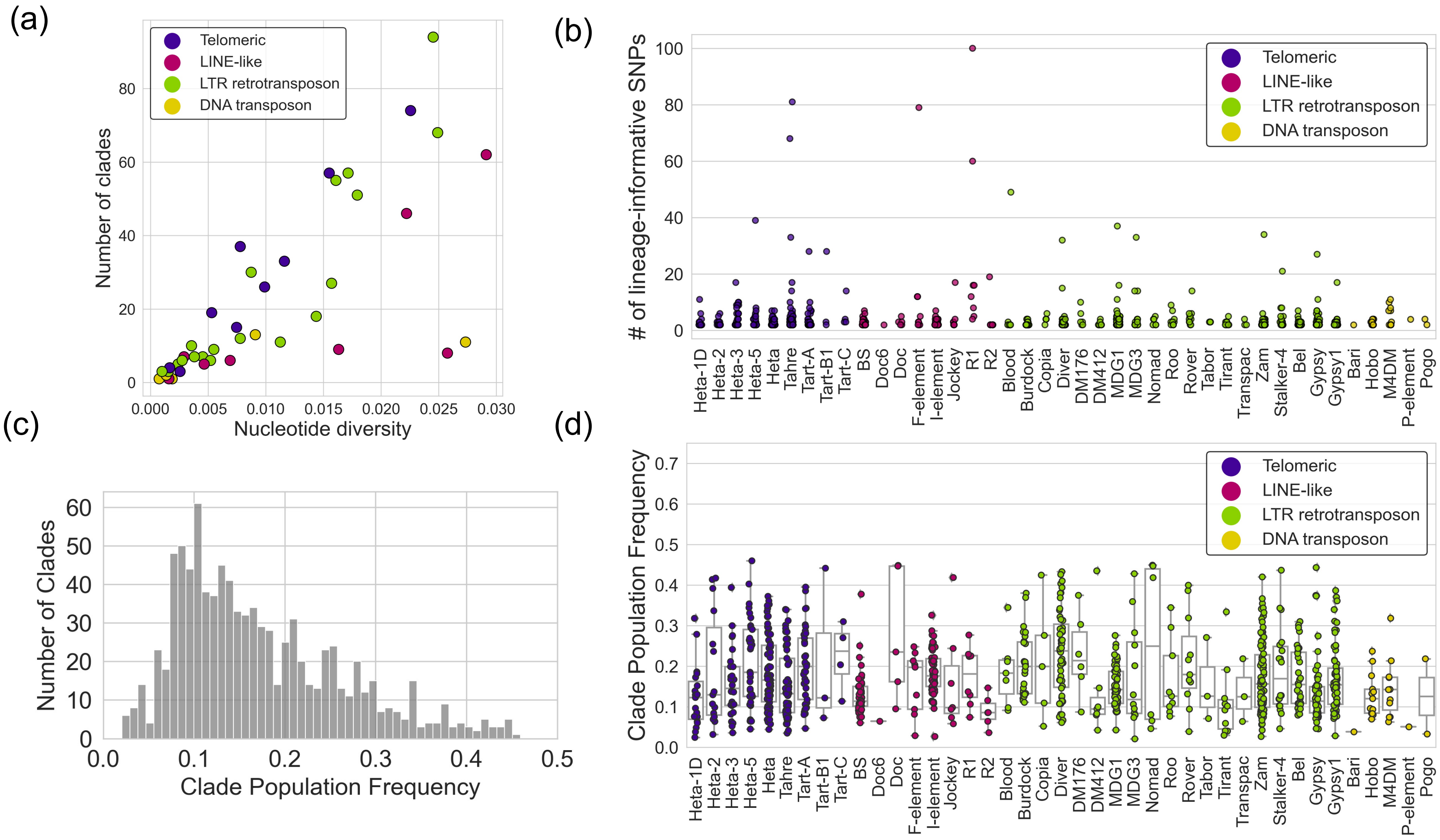

(a)
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
Number of clades
0
20
40
60
80
0.000
0.005
0.010
0.015
0.020
0.025
0.030
Nucleotide diversity
(b)
# of lineage-informative SNPs
0
20
40
60
80
100
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
Heta-1D
Heta-2
Heta-3
Heta-5
Heta
Tahre
Tart-A
Tart-B1
Tart-C
BS
Doc6
Doc
F-element
I-element
Jockey
R1
R2
Blood
Burdock
Copia
Diver
DM176
DM412
MDG1
MDG3
Nomad
Roo
Rover
Tabor
Tirant
Transpac
Zam
Stalker-4
Bel
Gypsy
Gypsy1
Bari
Hobo
M4DM
P-element
Pogo
(c)
Number of Clades
0
10
20
30
40
50
60
0.0
0.1
0.2
0.3
0.4
0.5
Clade Population Frequency
(d)
Clade Population Frequency
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
Heta-1D
Heta-2
Heta-3
Heta-5
Heta
Tahre
Tart-A
Tart-B1
Tart-C
BS
Doc6
Doc
F-element
I-element
Jockey
R1
R2
Blood
Burdock
Copia
Diver
DM176
DM412
MDG1
MDG3
Nomad
Roo
Rover
Tabor
Tirant
Transpac
Zam
Stalker-4
Bel
Gypsy
Gypsy1
Bari
Hobo
M4DM
P-element
Pogo

Figure 3

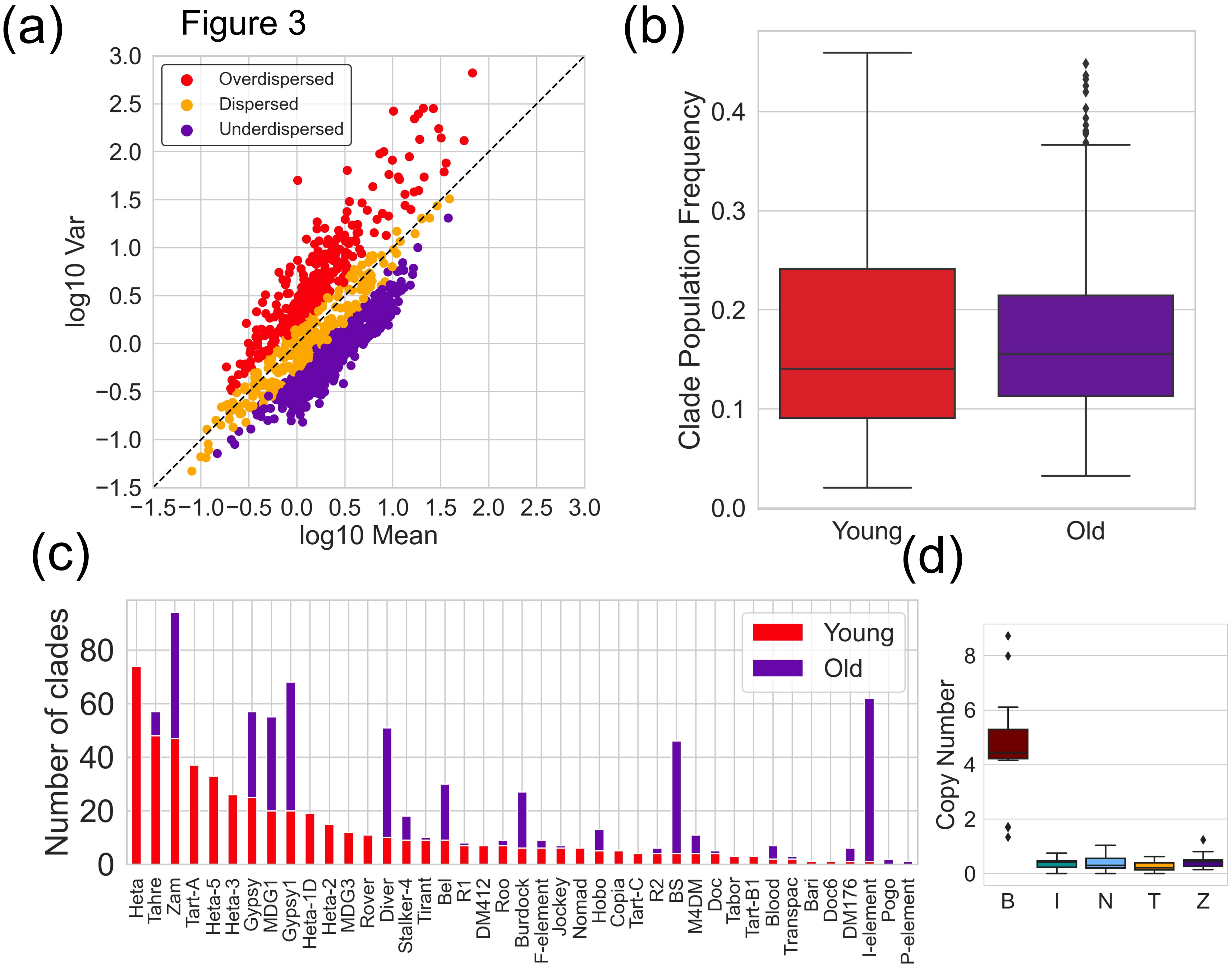

(a)
Overdispersed
Dispersed
Underdispersed
log10 Var
log10 Mean
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5 3.0
(b)
Clade Population Frequency
0.4
0.3
0.2
0.1
0.0
Young
Old
(c)
Number of clades
80
60
40
20
0
Young
Old
Heta
Tahre
Zam
Tart-A
Heta-5
Heta-3
Gypsy
MDG1
Gypsy1
Heta-1D
Heta-2
MDG3
Rover
Diver
Stalker-4
Tirant
Bel
R1
DM412
Roo
Burdock
F-element
Jockey
Nomad
Hobo
Copia
Tart-C
R2
BS
M4DM
Doc
Tabor
Tart-B1
Blood
Transpac
Bari
Doc6
DM176
I-element
Pogo
P-element
(d)
Copy Number
8
6
4
2
0
B
I
N
T
Z

Figure 4

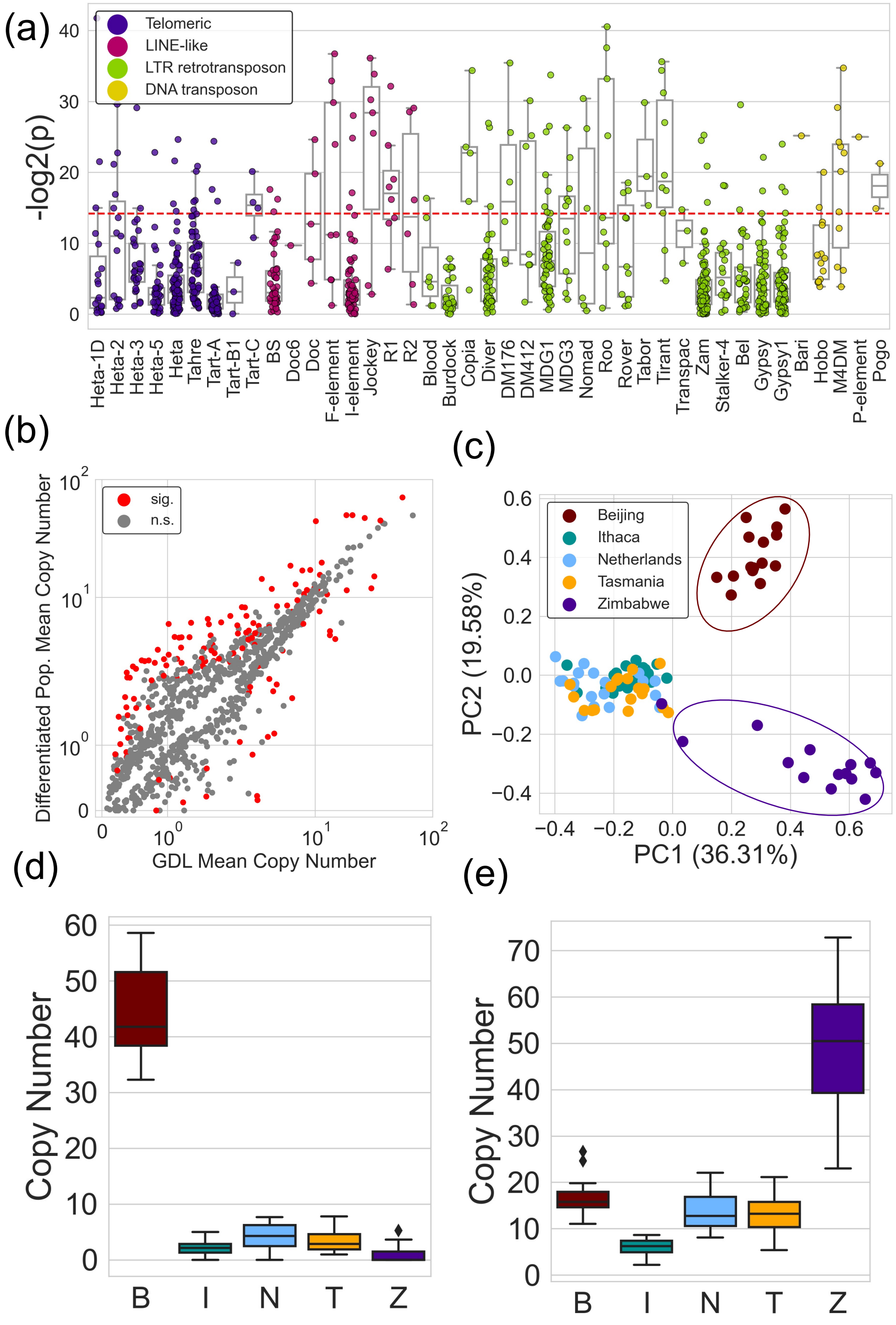
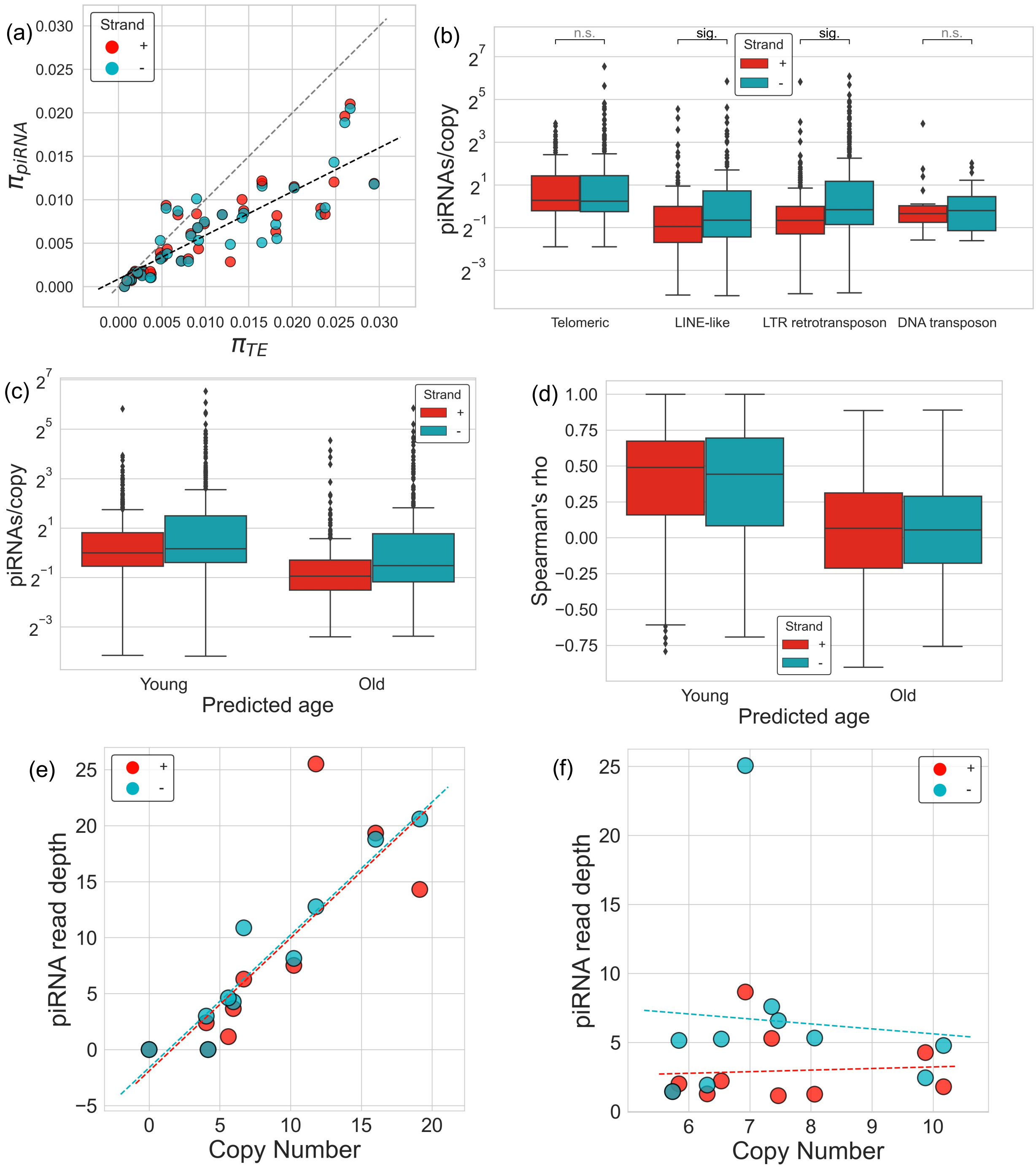
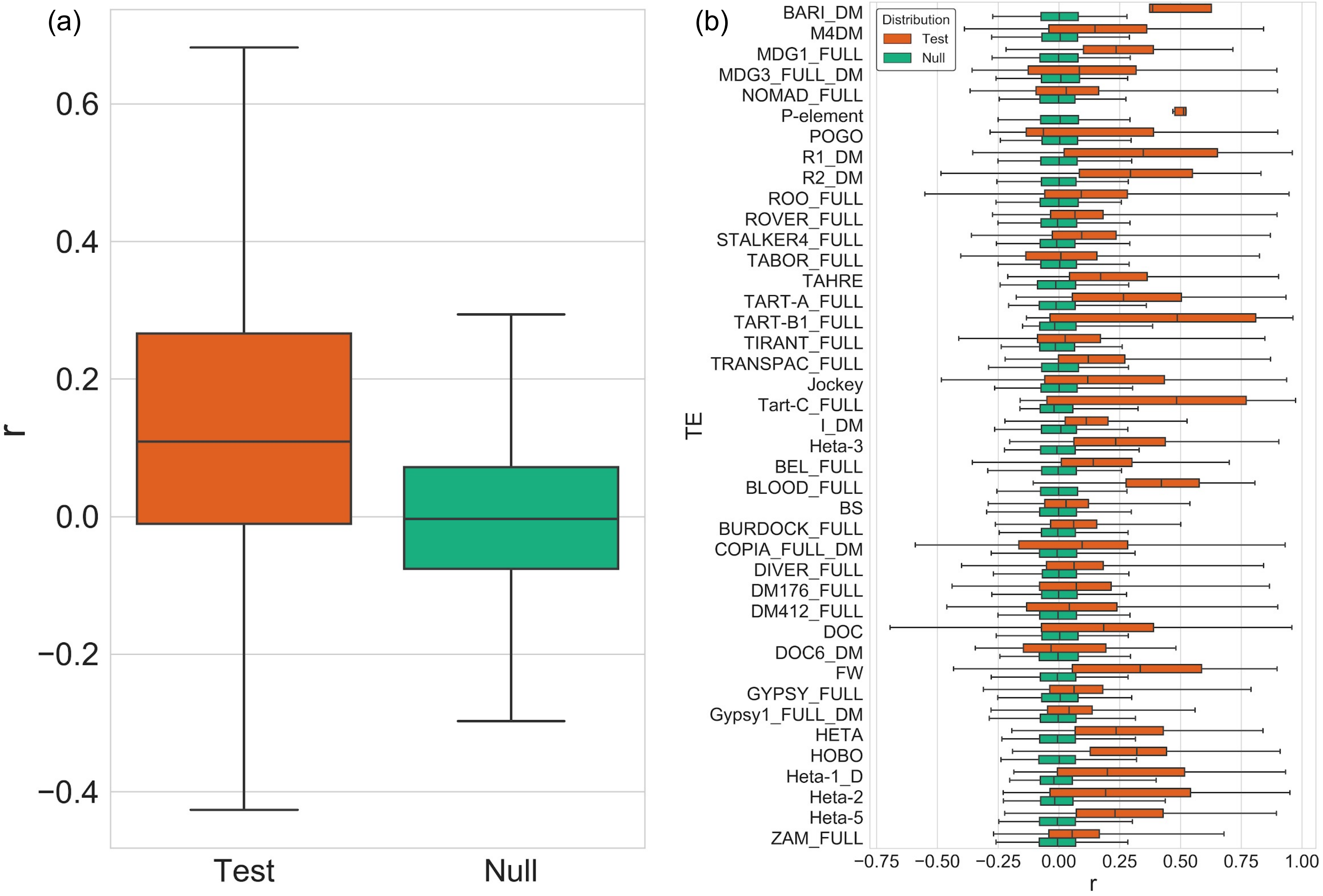

(a)
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
-log2(p)
0
10
20
30
40
Heta-1D
Heta-2
Heta-3
Heta-5
Heta
Tahre
Tart-A
Tart-B1
Tart-C
BS
Doc6
Doc
F-element
I-element
Jockey
R1
R2
Blood
Burdock
Copia
Diver
DM176
DM412
MDG1
MDG3
Nomad
Roo
Rover
Tabor
Tirant
Transpac
Zam
Stalker-4
Bel
Gypsy
Gypsy1
Bari
Hobo
M4DM
P-element
Pogo
(b)
sig.
n.s.
Differentiated Pop. Mean Copy Number
GDL Mean Copy Number
0
10^0
10^1
10^2
(c)
Beijing
Ithaca
Netherlands
Tasmania
Zimbabwe
PC2 (19.58%)
PC1 (36.31%)
−0.4
−0.2
0.0
0.2
0.4
0.6
(d)
Copy Number
0
10
20
30
40
50
60
B
I
N
T
Z
(e)
Copy Number
0
10
20
30
40
50
60
70
B
I
N
T
Z

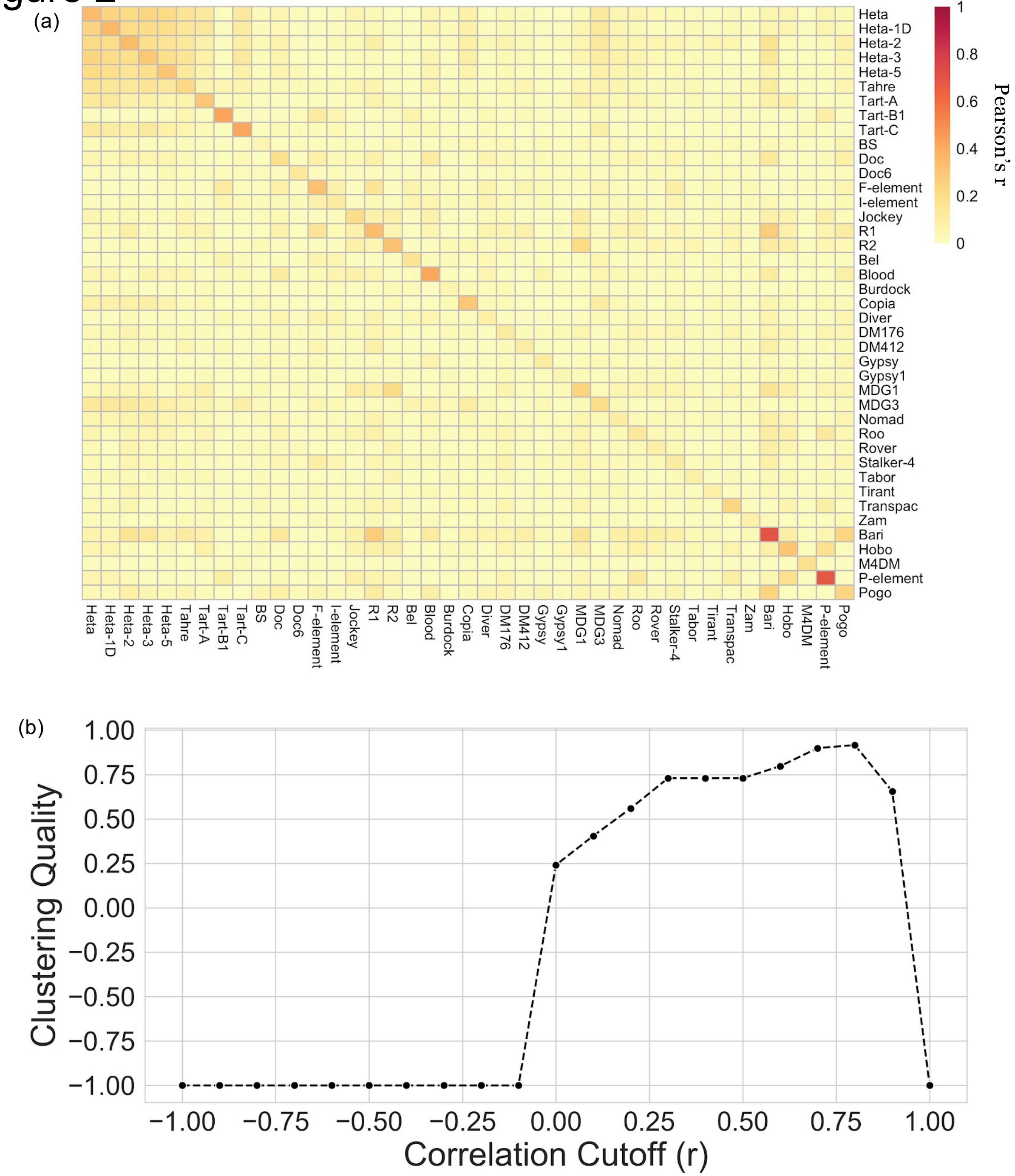
Figure 5


(a)
Strand
+
-
0.030
0.025
0.020
0.015
0.010
0.005
0.000
π_{piRNA}
0.000 0.005 0.010 0.015 0.020 0.025 0.030
π_{TE}
(b)
n.s.
sig.
sig.
n.s.
Strand
+
-
2^7
2^5
2^3
2^1
2^{-1}
2^{-3}
piRNAs/copy
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
(c)
Strand
+
-
2^7
2^5
2^3
2^1
2^{-1}
2^{-3}
piRNAs/copy
Young
Old
Predicted age
(d)
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
Spearman's rho
Strand
+
-
Young
Old
Predicted age
(e)
+
-
25
20
15
10
5
0
−5
piRNA read depth
0 5 10 15 20
Copy Number
(f)
+
-
25
20
15
10
5
0
piRNA read depth
6 7 8 9 10
Copy Number

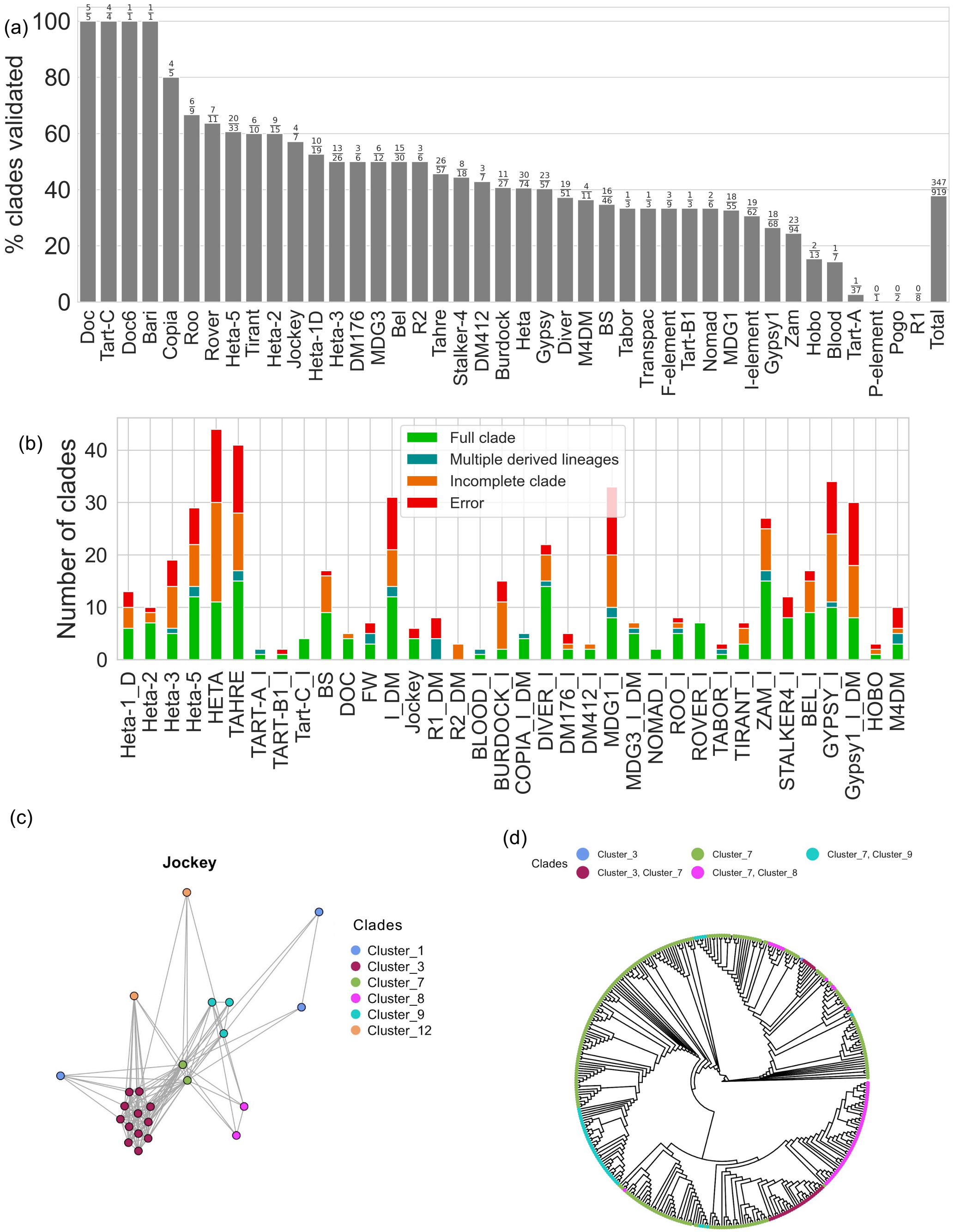
Supplemental Figure 1



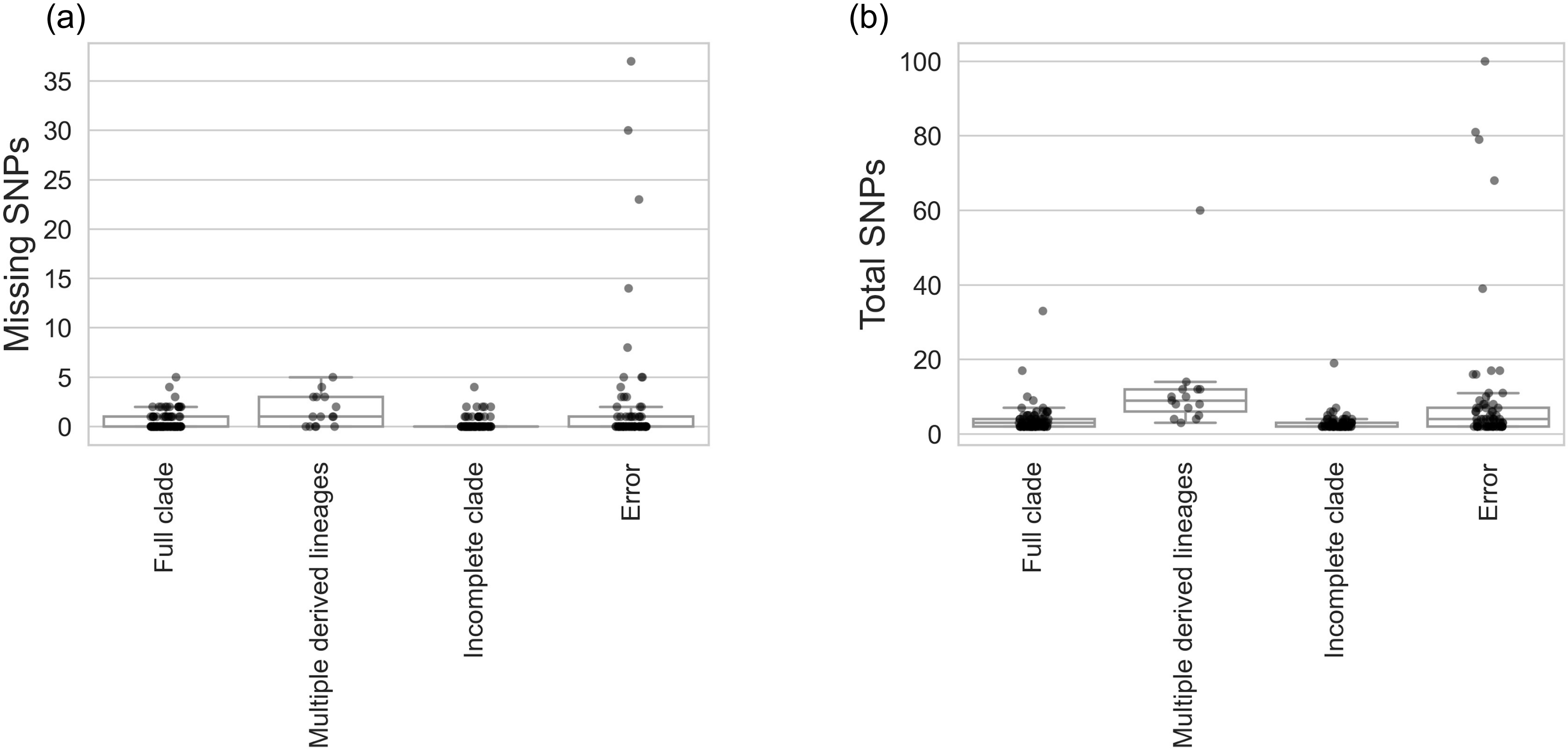
Supplemental Figure 2



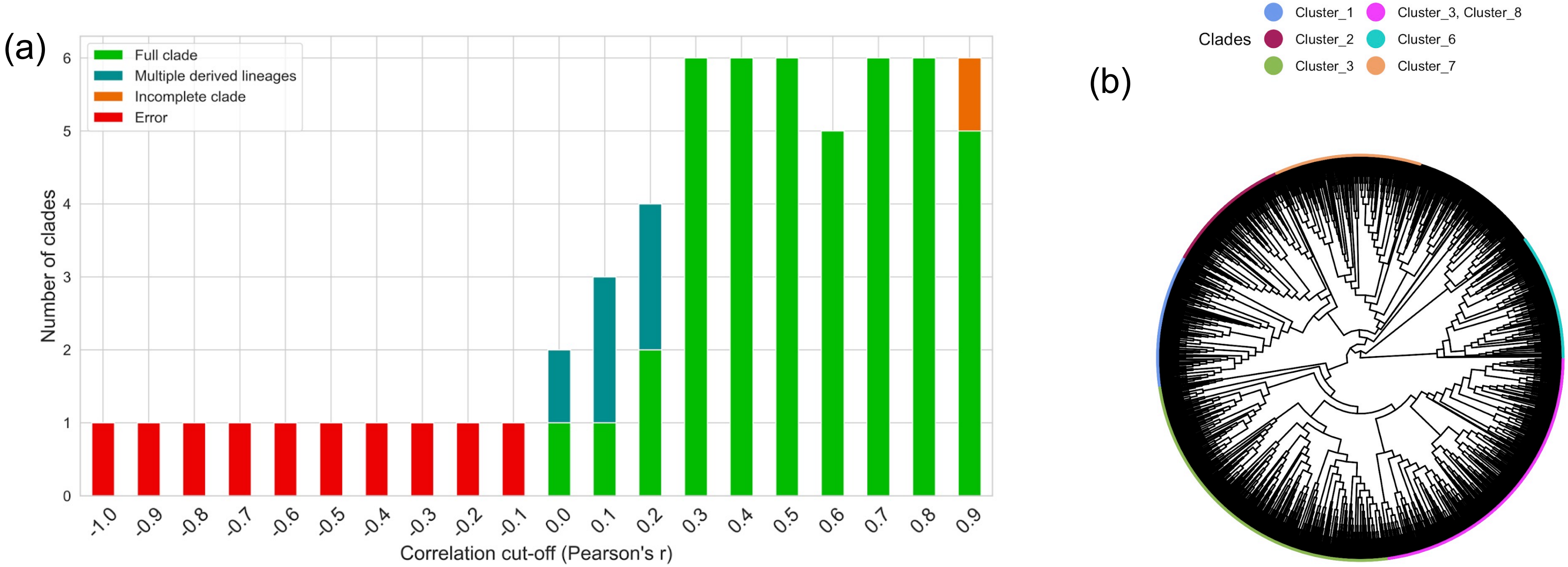
Supplemental Figure 3



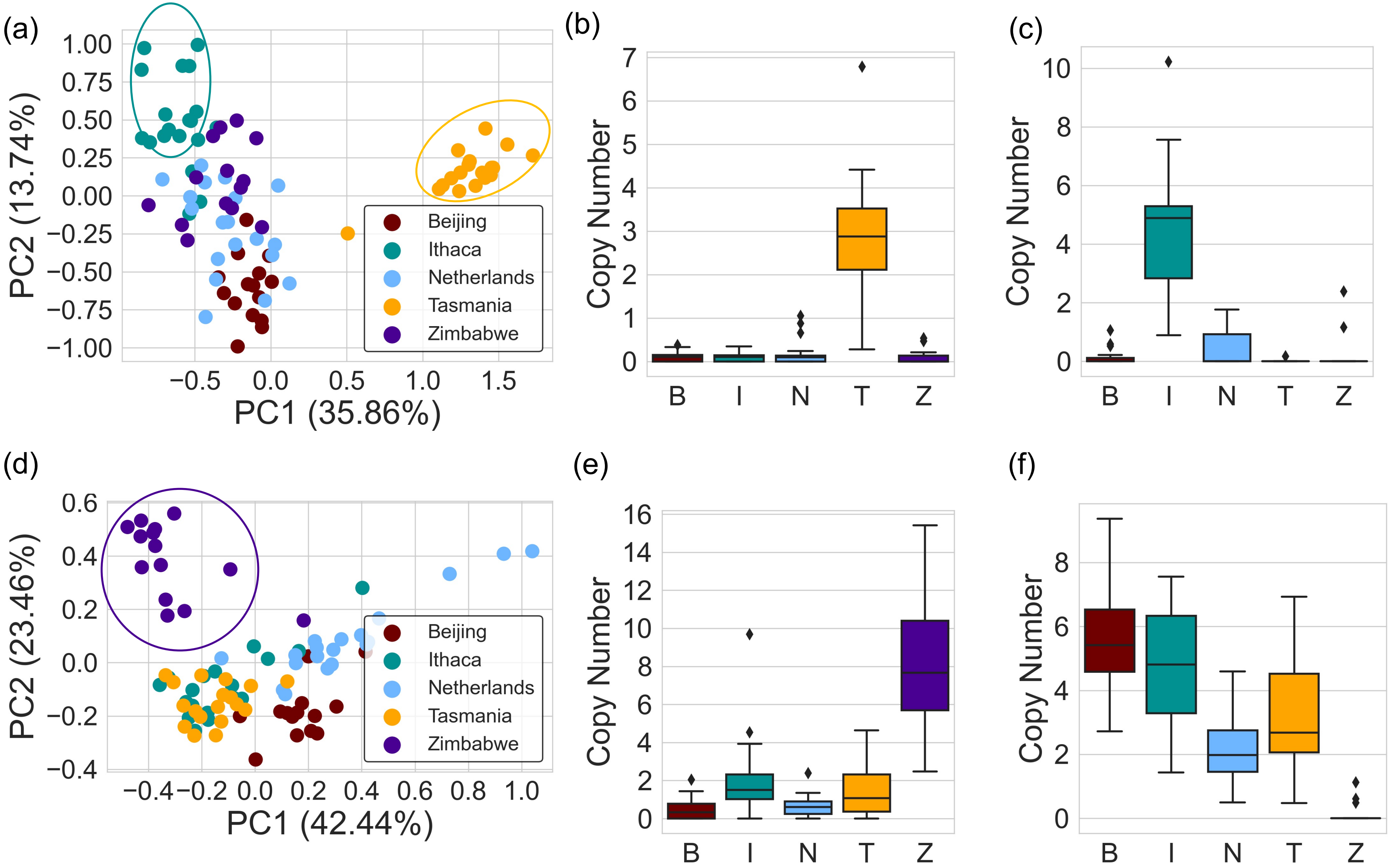
Supplemental Figure 4



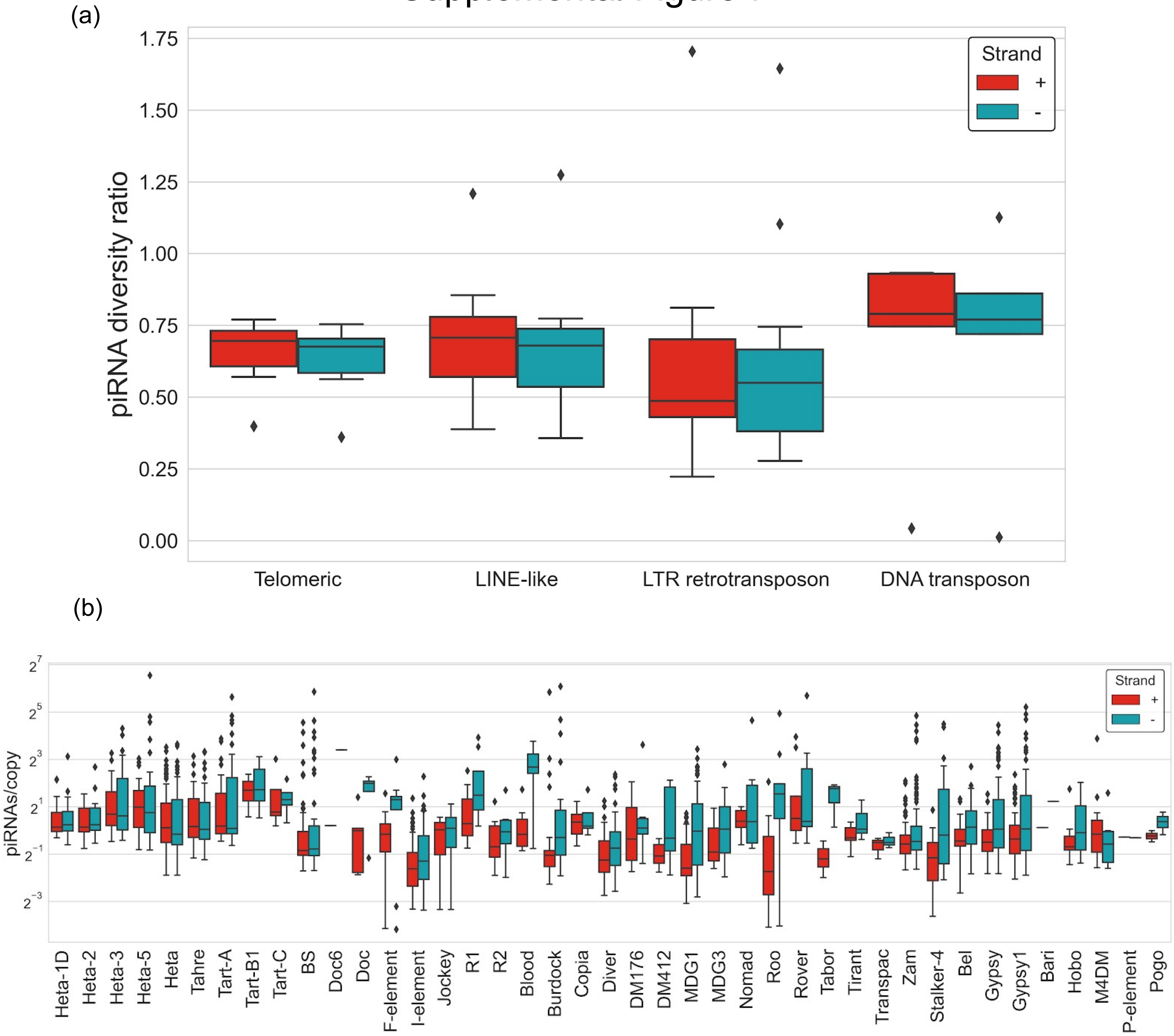
Supplemental Figure 5



Supplemental Figure 6



Supplemental Figure 7


(a)
piRNA diversity ratio
Telomeric
LINE-like
LTR retrotransposon
DNA transposon
Strand
(b)
piRNAs/copy
2^7
2^5
2^3
2^1
2^{-1}
2^{-3}
Strand
Heta-1D
Heta-2
Heta-3
Heta-5
Heta
Tahre
Tart-A
Tart-B1
Tart-C
BS
Doc6
Doc
F-element
I-element
Jockey
R1
R2
Blood
Burdock
Copia
Diver
DM176
DM412
MDG1
MDG3
Nomad
Roo
Rover
Tabor
Tirant
Transpac
Zam
Stalker-4
Bel
Gypsy
Gypsy1
Bari
Hobo
M4DM
P-element
Pogo