Submitted as an investigation to *Genetics*

Extensive horizontal exchange of transposable elements in the *Drosophila pseudoobscura* group

Tom Hill[1]* & Andrea J. Betancourt[2]

*Corresponding Author

1. 4055 Haworth Hall, The Department of Molecular Biosciences, University of Kansas, 1200 Sunnyside Avenue, Lawrence, KS 66045.

2. Institute of Integrative Biology, University of Liverpool, Liverpool, United Kingdom

Abstract

While the horizontal transfer of a parasitic element can be a potentially catastrophic, it is increasingly recognized as a common occurrence. The horizontal exchange, or lack of exchange, of TE content between species results in different levels of divergence among a species group in the mobile component of their genomes. Here, we examine differences in the TE content of the *Drosophila pseudoobscura* species group. We identify several putative horizontal transfer events, and examine the role that horizontal transfer plays in the spread of TE families to new species and the homogenization of TE content in these species. Despite rampant exchange of TE families between species, we find that both TE content differs hugely across the group, likely due to differing activity of each TE family and differing suppression of TEs due to divergence in Y chromosome size, and its resulting effects of TE regulation. Overall, we show that TE content is highly dynamic in this species group, and that it plays a large role in shaping the differences seen between species.

Data availability

All data used in this study (summarized in table S1) is freely available online through the NCBI short read archive (NCBI SRA: ERR127385, SRR330416, SRR330418, SRR1925723, SRR330426, SRR330420, SRR330423, SRR617430-74). All genomes used are either available through Flybase.org or Popoolation.at.

Introduction

Unlike mammals, which have few active transposable elements (TEs) mostly fixed insertions within species (Hellen and Brookfield 2013a; b), transposable elements (TEs) in *Drosophila* species appear to be highly active, as inferred from a high proportion of polymorphic, and thus presumably recent, insertions (Charlesworth and Langley 1989; Sniegowski and Charlesworth 1994; Charlesworth *et al.* 1997; González *et al.* 2008; Petrov *et al.* 2011). The dynamic nature of TEs in *Drosophila* is reflected in the data from the 12-genomes project (Clark *et al.* 2007). While species in the genus all host LTR, non-LTR retrotransposons and TIR DNA transposons in roughly the same rank order of abundance (Sessegolo *et al.* 2016), the contribution of each appears to differ for different genomes. The proportion of total TE content that is non-LTRs, for example, ranges from ~12% to ~35% (Clark *et al.* 2007; Sessegolo *et al.* 2016).

Under a model of TE evolution where active transposition is followed by suppression and eventually, inactive decayed elements, one might expect that the active families of elements would differ between the species (Kaplan *et al.* 1985; Maruyama and Hartl 1991; Capy *et al.* 1992; Lohe *et al.* 1995; Hartl *et al.* 1997). Instead, the overall content is largely similar (Vieira *et al.* 1999; Lerat *et al.* 2011; Kofler *et al.* 2015b), with many of the same TE insertions found at low frequencies in both species (Kofler *et al.* 2015b). The reason might be that the overall TE content between species be regularly homogenized by horizontal exchange between the species (Bartolomé *et al.* 2009). This process is exemplified by the recent horizontal transfer of the P-element, newly acquired by *D. melanogaster* sometime in the 20th century from a Caribbean species, into *D. simulans* (Kofler *et al.* 2015a; Hill *et al.* 2016).

Here, we investigate these questions in a different *Drosophila* group, the *pseudoobscura* group, using publicly available genome sequences for five species, and an improved genome sequence from *D. pseudoobscura* (Richards *et al.* 2005), and several sequenced third chromosome isolates (Fuller *et al.* 2016). Unlike *D. simulans* and *D. melanogaster*, these species are not cosmopolitan and thus may have had less opportunity to encounter new transposable elements outside their ancestral range. Further, in contrast to most Drosophila, some species in this group were reported to have mostly fixed insertions; we re-examine this claim with new data. We also examine

horizontal exchange between species within the group and from outside the group, and find abundant evidence of recurrent horizontal exchange.

Materials and Methods

Sequence data

All sequence data used is summarized in Table S1. We used publicly available reference genomes for five species: *D. pseudoobscura* (NCBI: PRJNA18793), *D. persimilis* (NCBI: PRJNA29989 genome assembled from Sanger sequence reads, http://popoolation.at/persimilis_genome/ for the genome based on illumina reads), *D. affinis* (NCBI: ERX103526), *D. lowei* (http://popoolation.at/lowei_genome/; Palmieri et al. 2014), *D. miranda* (NCBI: PRJNA77213) and *D. affinis* (http://popoolation.at/affinis_genome/). We also used publicly available paired-end illumina data from inbred lines for four of these species [*D. persimilis* (SRA: SRR330426), *D. miranda* (SRA: SRR1925723), *D. lowei* (SRA: SRR330416 and SRR330418) and *D. affinis* (ENA: ERR127385)]. As we were unable to find publicly available paired-end illumina data for *D. pseudoobscura*, we used a data generated from an individual wild *D. pseudoobscura* made homozygous for the reference third chromosome inversion type (SRA: SRR617430, S. Schaeffer, pers. Comm.; Fuller et al. 2016). As a result, only the third chromosome represents a wild chromosome, the rest of the genome is a mosaic of material from the wild and from the two different balancer stocks used, due to this we limited any population statistical analysis to the third chromosome.

De novo annotation of transposable elements in the D. pseudoobscura group

We annotated TE families in all five species, as well as putative TE sequences in the more diverged species (such as *D. lowei* and *D. affinis*), and compared our *de novo* annotations to the previous annotations for *D. pseudoobscura* and *D. persimilis*. These sequences were identified using *RepeatModeler* and *LTRHarvest* (Ellinghaus *et al.* 2008; Smit and Hubley 2008) and filtered, as outlined in Supplementary Figure 1 to give us a set of ‘high confidence’ TE annotations.

To *de novo* annotate the transposable elements, as shown in Figure S1:

1. We recovered a set of TE candidates for each species using the reference genomes. We used two separate pipelines: *(i) Repeatscout* and *PILER* in the *RepeatModeler* pipeline (default parameters) (Price *et al.* 2005; Smit and Hubley 2008), with all sequences designated as microsatellites and simple repeats removed from the output, and *(ii) LTRHarvest*, which finds LTR retrotransposons (using parameters recommended in the *LTRHarvest* manual*:* -tis -*suf* -lcp -d*es* -*sds* –*dna;* -seed 10*0* -minlenltr 100 -maxlenltr 1000 -mindistltr 1000 -maxdistltr 15000 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3 -similar 90.0 -overlaps best -mintsd 5 -maxtsd 20 -motif tgca -motifmis 0 -vic 60 -longoutput) (Ellinghaus *et al.* 2008). Though this step may bias us to find primarily LTRs, we note that most previously known TEs we find are LTRs, while most (19 of 41) novel elements are DNA transposons (Table S2).
2. Step 1 resulted in a set of 769 candidate TE sequences, ranging from 208bp to 14.5kb. We used BLAST to filter and annotate the candidate TEs (parameters: e-value < 1e-08, -word_size 10, -perc_identity 85) (Altschul *et al.* 1990), by searching a database of all known *Repbase* and *Flybase* transposable element sequences for *Diptera* (including 121 TEs previously found in *D. pseudoobscura, D. persimilis* or *D. miranda*), with sequenced duplicated between the data bases removed using a custom python script.
    a. Sequences that show single BLAST hits (e-value ≤ 1e-08) to this data base were assumed to represent a previously identified TE family. We discarded these sequences and used the Repbase/Flybase TE sequence to represent the family instead. (349 sequences).
    b. From the remaining sequences, those that showed BLAST hits to several TE families, all from one superfamily, were considered to potentially represent a previously unidentified family within that superfamily. (180 sequences).
    c. Of the remaining sequences, those with hits all in a single order, but to multiple superfamilies, were potentially novel TEs within this order. (18 sequences).
    d. For sequences which had no potential TE family assigned in Step 2 (222 sequences), we attempted to find matches by aligning them to the online

NCBI non-redundant database using megablast. Of these, 202 had annotated or predicted genes as the primary BLAST hit; these were discarded. The remaining potentially novel TEs were retained (20 sequences),

To facilitate downstream analysis, we obtained a single representative sequence for the potential novel TEs identified in Steps 2b, c and d, as is already done for those in Step 2a. To do this, we clustered sequences found for all species using *vmatch* (recommended *LTRHarvest* parameters: -dbcluster 95 7 -p -d -seedlength 50 -l 1101 -exdrop 9) (Kurtz 2010). We confirmed these clusters by BLASTing novel TE sequences to themselves and grouping them by similar matches (parameters: e-value < 0.00001, -word_size 10).

4. As these may only represent partial TE sequences, we further assembled the grouped sequences using *Trinity* (default parameters) to collapse similar sequences and get a representative sequence for the cluster, even if only a fragment of the consensus sequence (Haas *et al.* 2013). We checked these assemblies and clusters by aligning sequences from the cluster and with the *Trinity* assembly (if applicable) using *MAFFT* (parameters: --thread 3 --threadit 0 --reorder --leavegappyregion –auto) (Katoh *et al.* 2002), to ensure that the assembly or longest sequence representing the putative novel TE was recovered. From each cluster of similar sequences, we took the longest sequence as the representative fragment of each putatively novel family.
5. Some of the putatively novel families identified in 2b may instead be divergent representatives of known families. To see whether this was the case, we again attempted to identify previously known families among them using the consensus sequences from the five species genomes. We aligned novel TEs pairwise to all *Repbase* TEs using *MAFFT* (parameters: --thread 3 --threadit 0 --reorder --leavegappyregion --auto) and used a custom *python* script to find the number of diverged aligned bases. We defined sequences as belonging to a known family if they were >90% similar to a known family across the sequence, following (Kohany *et al.* 2006). Two families of the novel sequences were found to belong to known families in this way (an I-element and a Jockey element), but were

closely related to insertions in distant relatives of the *obscura* group (*I*-4_DF from *D. funebris* and *Jockey*-8_DRh from *D. rhopaloa,* respectively). We therefore retained these sequences in our data set, as they likely represent diverged copies of these families, or ancient horizontal acquisitions.

6. From Steps 1-5, we found 567 candidate TE sequences, 349 of which belong to previously described TE families, including all 121 families previously found in the *D. pseudoobscura* group ('known' families), and 445 others (putative 'novel' families). We proceeded to filter sequences from this set which were represented by very few or very short matches to the reference genomes.
    a. First, we used the 567 sequences to repeat mask the reference genome of each species using *RepeatMasker* (parameters: –no_is –norna –no_low –gff –gccalc –u –s –cutoff 200) (Tarailo-Graovac and Chen 2009), following recommendations in (Kofler *et al.* 2012). We required that the families have at least 25 Repeatmasker hits in at least one species (237 sequences retained, 116 known and 121 novel families).
    b. We then estimated the copy number of each TE family for each species from the Illumina short read data from adult females, discarding those estimated to have a median coverage less than 2-fold that of the third chromosome for less than 80% of the length of the sequence. To do this, we mapped short reads to the repeated masked reference genome and the 237 TE sequences retained from the previous step using BWA MEM (parameters: paired end –t 5 -M) (Li and Durbin 2009), and estimated coverage with *bedtools genomecov* (Quinlan and Hall 2010). Due to the poor assembly of the *D. persimilis* genome, we used a reference consisting of the *D. pseudoobscura* genome and the *D. persimilis* TE sequences. (157 sequences retained, from 116 known and 41 families novel to this species group).

We considered these 157 sequences to be a cromulent representation of the TE content in the *pseudoobscura* group, though we recognize that we may have discarded some true TE sequences.

Using this method, we found strong support for 114 of the 121 TE families previously described in *D. pseudoobscura, D. persimilis* or *D. miranda* and 2 TEs previously identified in other species. We also found 41 novel sequences, including two subfamilies of previously known sequences, 30 newly assembled sequences which BLAST exclusively to one super family, and nine potentially new families that BLAST to one TE order. We also found 15 sequences that cannot be assigned an order (either due to BLAST hits to multiple orders, or no BLAST hits). These 15 sequences passed all filters, including being found multiple times in species genomes and did not correspond to genes or other NCBI sequences in a universal BLAST search. To avoid unreliable inferences, we discarded these sequences from downstream analyses, but gave each of the 41 novel sequences an ID (Table S2), and included them in masking and mapping stages. Sequences are available in File S1.

Estimating TE density in the reference genome.

We used *RepeatMasker* v. 4.0.6 to mask each reference genome using the 157 consensus TE sequences and 15 unknown sequences from the *de novo* annotation, (parameters: –no_is –norna –nolow –gff –gccalc –u –s –cutoff 200) (Tarailo-Graovac and Chen 2009). To estimate TE density, we used the density of TE bases per 1MB sliding window (with a step size of 100kb) of the *D. pseudoobscura* reference genome (after removing all N bases [e.g. TE bases / [window size – Ns in chromosome]]), across both assembled scaffolds and unassembled contigs from the reference genome.

Identifying insertions in reference genomes and in sequenced third chromosome lines of D. pseudoobscura

To identify insertion sites in the reference genomes of *D. pseudoobscura* and *D. persimilis*, we used the *PopoolationTE2* pipeline (Kofler *et al.* 2016a). Briefly, we used *RepeatMasker* v. 4.0.6 to mask the *pseudoobscura* genome using the 157 consensus TE sequences and 15 unknown sequences identified above (parameters: –no_is –norna –nolow –gff –gccalc –u –s –cutoff 200) (Tarailo-Graovac and Chen 2009). We chose to use the *D. pseudoobscura* reference, rather than the fragmented *D. persimilis* reference, as it facilitated mapping reads to genomic insertion sites. We expect similar

205 results as these species are closely related (0.018 average synonymous divergence
206 (Noor *et al.* 2007)), and we find that similar numbers of reads map to TEs regardless of
207 whether the *D. pseudoobscura* or *D. persimilis* genome is used (27.63% vs 27.27%).

208 We then mapped available Illumina reads to the repeat masked reference, the
209 consensus TE sequences, and to sequences matching these consensus TEs identified
210 by *RepeatMasker* using BWA-MEM (parameters: paired end –t 5 -M, with secondary
211 alignments reported, but marked) (Li and Durbin 2009). Using masked TE sequences
212 to aids mapping of degenerate TE sequences, as described in (Kofler *et al.* 2016a).

213 Following mapping, we generated a ppileup file summarizing identities and locations of
214 TE insertions for all lines in *PopoolationTE2* (default settings, --map-qual 10) and
215 subsampled to a physical coverage of 25, removing secondary alignments. As these
216 sequences are mostly from inbred lines, we required the estimated frequency to be at
217 least 50% (default parameters, --target-coverage 25, --min-count 5, minimum frequency
218 = 0.5) (Kofler *et al.* 2016a). We then identified the number of insertions per MB window
219 (after adjusting for the number of N bases in the window [e.g. TE number / [window size
220 – Ns in window]]) across the genome of each species.

221 *Expression confirmation of putative TE sequences.*

222 We also used expression data for mRNA (SRA: SRR1956914, taken from (Duff
223 *et al.* 2015)) and small RNAs (SRA: SRR032435, taken from (Leslie *et al.* 2010)) from
224 the *D. pseudoobscura* reference line (MV-25) to examine the expression of novel TEs.

225 Before further analysis, we trimmed all genomic and RNAseq Illumina reads used
226 with *Sickle* to remove low quality sequence data (default parameters for long reads,
227 minimum length = 16 for small RNAs), and removed reads that were unpaired (apart
228 from the small RNA reads) after this step from the sequence data (Joshi and Fass
229 2011).

230 We mapped small RNA sequences from *D. pseudoobscura* to known and novel
231 TEs identified in that species, using publicly available small RNA reads from the
232 reference strain ((Leslie *et al.* 2010), SRA: SRR032435).

233 We first removed non-TE related small RNAs, following (Aravin *et al.* 2007;
234 Rahman *et al.* 2015), by mapping to a database of known *Drosophila* viruses and small
235 RNAs other than those that are TE-related, including miRNAs, viral siRNAs, snoRNA
236 (Rahman *et al.* 2015), using *BWA aln* and allowing for up to 3 mismatches (parameters:
237 -n 3) (Aravin *et al.* 2007; Li and Durbin 2009). We then mapped the remaining reads to
238 the repeat masked *D. pseudoobscura* reference genome and the novel and known TE
239 sequences identified in this study (*BWA aln* parameters: -n 3, maximum 2 alignments).

240 We classified small RNAs by length and orientation using a custom python script
241 and the *Pysam* python library, following (Brennecke *et al.* 2008). Specifically, we
242 considered small RNAs from 21 to 23 to be siRNAs and from 24 to 29 to be piRNAs
243 (Obbard *et al.* 2009). We used *bedtools* (*intersect*, -wa –wb –f 0.3 –r), to check for a
244 10-bp overlap between sense and anti-sense matches and used *sequence logos*
245 (Schneider and Stephens 1990) to check for the 1-T, 10-A bias, both associated with
246 ping-pong amplification, a characteristic feature of piRNAs (Levine and Malik 2011).

247 *Detecting short range horizontal transfer events within the pseudoobscura group*

248 To detect horizontal transfer of TEs within the five species examined, we compared
249 divergence between consensus TE sequences to genomic divergence, following the
250 rationale described in (Bartolomé *et al.* 2009). We limited this analysis to families found
251 in at least 3 species and with an annotation on Repbase.

252 To construct consensus TE sequences for each TE family and each species, we
253 identified the major allele for each species at each variable site using *GATK v3.5-0-*
254 *g36282e4 HaplotypeCaller*, with ploidy levels set to the estimated copy numbers based
255 on coverage of the TE sequence, and using *FastaAlternateReferenceMaker* (default
256 parameters) to generate fasta sequences from the mapped data (DePristo *et al.* 2011).

257 We aligned these consensus sequences from each species using *MAFFT* (parameters:
258 --thread 3 --threadit 0 --reorder --leavegappyregion –auto) (Katoh *et al.* 2002) and
259 generated a phylogeny of each sequence using the *Repbase* annotation and *PhyML*
260 (parameters: -M GTR) (Guindon *et al.* 2010). We obtained a total of 39 annotated
261 alignments that included sequences for *D. affinis* comparisons, and 62 additional
262 sequences for all other species comparisons (noted in Table S2).

263 We estimated synonymous site divergence (d_S) in the TE sequences pairwise
264 between species using *codeml* (with transition–transversion rates estimated from the
265 data, and codon frequencies from the nucleotide frequencies) and the coding regions
266 for these TEs as annotated in *Repbase* (Kohany *et al.* 2006; Yang 2007). We then
267 compared d_S of TEs to that of orthologous genes between species obtained in the same
268 way, taken from Avila et al. (2014). Following Bartolomé et al. (2009), we considered an
269 individual family to show strong evidence of exchange if its d_S value was below the
270 2.5% quantile of the d_S of all nuclear genes, to have potentially transferred if d_S was
271 between the 2.5% and 50% quantiles, and to show no evidence of transferring if above
272 the 50% quantile.

273 We also examined polymorphism within TE families for evidence of horizontal
274 transfer. We estimated Tajima's *D* of each TE using *Popoolation* (Kofler *et al.* 2011),with
275 the TE copy number as the sample size. As negative Tajima's D may reflect recent
276 expansion of a TE family (Bartolomé *et al.* 2009). We compared the levels of
277 polymorphism shared among TEs in each species between potentially transferred TEs
278 (d_S < 2.5% quantile) and TEs that are unlikely to have transferred (d_S > 50% quantile).

279 *Detecting long range horizontal transfer events with other Drosophila species*

280 We attempt to identify long range transfers from other *Drosophila* species. To do this,
281 we separated all known *Drosophila* TEs by their super families, including our set of *D.*
282 *pseudoobscura* group TEs, and aligned the TE sequences within each superfamily
283 using *MAFFT* (Katoh *et al.* 2002) and generated phylogenies for these using *PhyML*
284 (Guindon *et al.* 2010). We then extracted patristic distance matrices for each
285 superfamily using *Patristic* (Fourmant and Gibbs 2006) and compared each distance to
286 the nuclear genome comparison performed previously for these genomes (Chen *et al.*
287 2014).

288

289 **Results and Discussion**

290 *Transposable element annotation of the D. pseudoobscura group genomes*

291 We identified insertions of the 157 well-supported TE families in the reference genome
292 of the five species, and assessed their TE content using four metrics: the proportion of

293 the reference genome masked (using *RepeatMasker* (Tarailo-Graovac and Chen
294 2009)), the proportion of short reads mapping to each TE sequences, the number of
295 insertions in each genome using short read data (using *PopoolationTE2* (Kofler *et al.*
296 2016b), demonstrated across genomes in Supplementary Figure 2) and the estimated
297 copy number of each TE family (Table 1 and Table S2). We also estimated the density
298 of TE content across the genome (in masked bp/Mbp) using the proportion of the
299 reference masked by *Repeatmasker*.

300 Across all species, for all measures of TE content we find a significant linear
301 correlation between measures (Table S2, Spearman's Rank Correlation *p*-value <
302 0.00213), though the strength of the correlation is weak for all species between the
303 proportion of the genome masked at the family level versus the copy number of the TE
304 family, and the insertion count versus the proportion of the genome masked ($\rho < 0.58$),
305 suggesting that the proportion of the genome masked may be an inconsistent measure
306 of TE density. As expected, correlations between measures of TE content in the
307 species with genomes assembled only from short reads are lower (Table 1, Table S2)
308 (Hoen *et al.* 2015; Rius *et al.* 2016), and the estimates of TE content for these species
309 are likely more unreliable. We therefore limit our analysis of TE content to the two
310 genomes with available long read data, *D. pseudoobscura* and *D. persimilis*. We also
311 identified 15 sequences that pass all filters, but cannot be assigned to a TE order, we
312 have included these sequences and their statistics, but have not included these
313 sequences in further analyses (e.g. the 2 unknown sequences in *D. pseudoobscura*,
314 Table 1, Table S2).

315 Because the *D. pseudoobscura* and *D. persimilis* genomes were originally
316 assembled from long reads (and the *D. pseudoobscura* genome has also been
317 assembled with the help of PacBio information and Sanger sequence information)
318 (Richards *et al.* 2005; Clark *et al.* 2007), the TE content of these two species is already
319 well-annotated. We found 116 previously identified *pseudoobscura* TE families using our
320 pipeline. We also found two TE families from other species, and 28 additional putative
321 TE families that passed all our filters in these two species. For *D. pseudoobscura*, we
322 were able to use RNAseq data from (Duff *et al.* 2015) to determine whether these
323 elements showed evidence of expression in embryos, using publically available

324 expression data. We estimated RPKM for both novel and known TEs from these data;
325 of the novel TEs, nine had appreciable levels of expression (Figure S3, FPKM > 1), a
326 similar proportion to that of the known TEs (49 of 116). Similarly, we used sequences
327 from embryonic small RNAs to ask if suppressive small RNAs are produced against
328 these TE families. We extracted TEs with at least 20 small RNAs mapping to them,
329 which comprised 114 of the 116 known TEs and all 28 of the novel TE sequences (Table
330 S7 & 8). Most of these elements (108 of 140) had piRNAs generated against them
331 (using the 24-29nt range generally used to identify piRNAs in other species (Ghildyal
332 and Zamore 2009), and 27 elements also had homologous siRNAs (21-23nt small
333 RNAs) (Figure S3, Table S2) (Obbard *et al.* 2009). A subset of the piRNAs, those
334 produced in the germline (Aravin *et al.* 2007), are expected to show signatures of "ping-
335 pong" amplification— small RNAs that match both sense and anti-sense strands of the
336 TE sequence, an enrichment of these that show a 10bp overlap, a uracil in position 1 of
337 sense strand piRNAs, and adenosine in position 10 of the antisense small RNAs)
338 (Aravin *et al.* 2007). We found that 60 elements (53 known families and 7 novel; 36
339 LTRs, 15 LINEs, 7 DNA transposons & 2 helitrons) showed signatures of ping-pong
340 amplification— from inspecting *Sequence Logos* plots (Table S2) (Schneider and
341 Stephens 1990); novel and known elements showed ping-pong small RNAs at similar
342 rates (Figure S3, Mann-Whitney U test W = 24, p > 0.1676). As these TE sequences
343 are all multicopy, these measures of expression are mainly useful to show that the
344 putative novel TEs have characteristics like those of the known TE sequences (Mann-
345 Whitney U test W = 37, p > 0.05, Figure S3).

346 In all, we found 12.33% and 23.59% make up the reference genomes of these
347 species, (Table 1). In contrast to a previous study, which found similar proportions of
348 LTRs and LINEs in the *D. pseudoobscura* genome (Clark *et al.* 2007), we find over
349 twice as much TE content due to LTR vs. LINE retrotransposons (Table 1); it is worth
350 noting an additional effort was put into finding novel LTRs in the putative TE set using
351 *LTRHarvest* (Ellinghaus *et al.* 2008).

352 In the remaining species, we find 20 additional families not found in *D.*
353 *pseudoobscura* and *D. persimilis* (Figure 1B). The 57 TE families shared among all five
354 species constitute most of the TE content (73-84% of insertions and 53-78% of each

355 species reference TE content, Table S2), but vary in copy number between species (e.g.
356 HelitronN-1 in *D. miranda* and *D. lowei*, Table S2), possibly due to stochastic expansion
357 and loss of families over time. For example, we find *HelitronN-1_DPe* has 1927
358 insertions and makes up 1.1% of the genome of *D. miranda*, while it has only 727
359 insertions in *D. lowei,* comprising of 0.14% of the genome (Table 1, Table S2). This is
360 likely due to the collapsing of *Helitron-1* and the closely related ISX sequence that has
361 been co-opted for dosage compensation in *D. miranda* (Ellison and Bachtrog 2013).
362 These differences can be further seen in the distributions of copy numbers in families,
363 which differ between species (Figure S3, S4).

364 While we annotate the *D. miranda, D. lowei* and *D. affinis* genomes using a
365 pipeline identical to that for other species, we suspect we have underestimated the TE
366 content of these species. There are three main reasons. First, the genome assemblies
367 for these species rely on short reads (Palmieri *et al.* 2014), which can lead to under-
368 representation of the TE content of the genome (Rius *et al.* 2016). Similarly, previous
369 estimates of the TE content of *D. pseudoobscura* and *D. persimilis* were much lower
370 likely due to more fragmented genomes (Clark *et al.* 2007). Second, we may have
371 missed TE families unique to these species, or may have recovered them only as
372 fragments, as it is easier to recover full-length TE sequences if they closely match
373 sequences already in RepBase as is true for the *D. persimilis* and *D. pseudoobscura*
374 sequences. Finally, and likely the most important reason, the material sequenced for the
375 reference genomes for *D. lowei, D. miranda* and *D. affinis* was adult females and not
376 mixed sex embryos as for the others (Richards *et al.* 2005; Clark *et al.* 2007). Thus the
377 other genomes contain the TE-rich Y-chromosome, which appears to be cytologically
378 quite large in these species (Dobzhansky 1935, 1937), and may shows less under-
379 replication of the TE-rich heterochromatin than adult samples.

380 In *D. persimilis* we found the same TE families as in *D. pseudoobscura*, but
381 estimated 23.59% of the *D. persimilis* reference to be repetitive content versus 12.33%
382 in *D. pseudoobscura*, implying 21.3MB more repetitive content in the *D. persimilis*
383 reference genome compared to *D. pseudoobscura*. Previous annotation from the 12-
384 genomes project found lower TE content as a proportion of the genome than that found
385 here (3% and 8% *vs.* 12.33% and 23.59% here), but a similar ~2-fold enrichment in TEs

386 for *D. persimilis* (Clark *et al.* 2007). While it is true that *D. persimilis* has a larger
387 genome than *D. pseudoobscura* ((Gregory 2005), the two species genomes are
388 estimated to differ only by ~2Mb (Bosco *et al.* 2007; Gregory and Johnston 2008).

389 The higher TE content of *D. persimilis* is not due to the presence of additional
390 families, as the same families occur in both species (Table 1, Figure 1). In fact, as
391 these species hybridize occasionally (Noor *et al.* 2007), it would be surprising if their TE
392 families remained very distinct. Estimates of copy number from coverage of short read
393 data (collected from adult females in both species) shows more copies of each TE
394 family in *D. persimilis* than *D. pseudoobscura* (46.8 vs. 39.6 on average), but the
395 difference is highly non-significant (Mann-Whitney U, p = 0.669).

396 Simple coverage differences of TEs, could, in principle, be explained by
397 differences in under-replication of TE between the strains or species. But this is
398 unlikely, as the coverage differences are also consistent with genome-size estimates,
399 insertion number recovered and the proportion of the genome repeat masked
400 (significant associations between all, as stated previously). If the difference is genuine,
401 it could be due to the differences seen in a few families with large numbers of insertions
402 in *D. persimilis,* such as *Gypsy10_Dpse*, *HelitronN-1_Dpe*, *Gypsy17_Dpse*, and
403 *MiniME_DP* (Table S2). Based on coverages of each TE sequence in each species, we
404 estimate that *D. persimilis* has, at most, ~5Mb more TE content than *D. pseudoobscura*,
405 consistent with the minor differences in genome size found between the two species
406 (Bosco *et al.* 2007), suggesting a large amount of genomic content is missing from the
407 *D. pseudoobscura* reference genome.

408 It is possible that an accumulation of TEs in the fixed inversions between
409 *D.pseudoobscura* and *D. persimilis* could explain the large difference in TE content, due
410 to the reduced genomic exchange in these regions (Machado *et al.* 2007), allowing
411 insertions to accumulate in one species but not the other. Consistent with this idea, we
412 find that LTR retrotransposons are at significantly higher densities of TE insertions
413 within these inverted regions in *D. persimilis* when compared to *D. pseudoobscura* and
414 the uninverted regions (File S2, Insertions per MB, using inversion windows defined in
415 (Avila *et al.* 2014); Mann Whitney U test: LTR inside inversions W = 53686, p = 5.674e-
416 05, LTR near inversions W = 16604, p-value = 0.1128 LTR outside inversions W =

290520, p-value = 0.1407). However, we find RC and LINE insertions are at significantly higher densities in *D. persimilis* regardless of genomic location (Insertions per MB, Mann Whitney U test: W > 335780, p-value < 0.0001303 for inside, outside and near inverted regions) and no difference in TIRs (W < 790, p-value > 0.37), suggesting that the fixed inversions are not the only explanation.

One final possibility, the Y chromosome of these species may also play a role in both the genome size and TE content differences of the species. While considerable variation exists in the size of *D. pseudoobscura* Y chromosome size across types (Dobzhansky 1935, 1937), the *D. persimilis* Y chromosomes are limited to the largest of these types (Types I, II and III). As the *D. pseudoobscura* reference genome was likely generated from a strain containing the smallest Y chromosome type (Standard/Arrowhead, likely type V) (Dobzhansky 1937; Dobzhansky and Sturtevant 1937), while the *D. persimilis* genome strain used to generate their genome likely contains the most common *D. persimilis* Y, the largest of the chromosome types (Dobzhansky 1937). Previous work has also found Y-linked variation in *D. melanogaster* and *D. simulans* to be associated with phenotypic variation in a number of factors including TE regulation, it is possible that a larger Y can cause poorer TE regulation, due to the increased heterochromatin load in the genome (Sackton and Hartl 2013; Francisco and Lemos 2014). This is possibly the case between *D. pseudoobscura* and *D. persimilis*, where the larger Y chromosome may have led to the poorer regulation of TE families, leading to the ~5Mb expansion of TEs in *D. persimilis*.

Several transposable element families show evidence of ancient horizontal spread between species

As a majority of TEs are likely acquired in a species by horizontal transfer by closely related species (Burt and Trivers 2006; Peccoud *et al.* 2017). We examined our set of TEs for evidence that they had been horizontally acquired from another *Drosophila* species by comparing the patristic distance of all *Repbase* TEs pairwise to the average patristic distance of *pseudoobscura* group TEs (Kohany *et al.* 2006), after building a phylogeny of each superfamily (Supplementary Figure 3).

447 Across 4096 pairwise comparisons, we found 230 where the TE patristic distance
448 was lower than the previously found genic distance (Table S7) (Chen *et al.* 2014).
449 These events were limited to 63 of 157 TEs, with most these TEs having lower patristic
450 distances than entire species groups (such as the *D. rhopaloa/D.elegans/D.ficusphila*
451 species subgroup), consistent with a transfer of the TE between the common ancestor
452 of the species and the *D. pseudoobscura* group, followed by a diversification into
453 multiple elements seen today. 42 of these transferred elements are LTRs, 6 are DNA
454 transposons and 15 are LINEs (Table S7, Figure 2A). While a higher proportion of LTRs
455 are transferred between species, each of these families only appears to have been
456 exchanged with a single species, rather than multiple, likely because of the recent
457 expansion of LTRs in *Drosophila*, compared to the more ancient expansion of most
458 LINEs and DNA transposons (Figure 2B). Among these transferred elements, we find a
459 *piggyBac* element that was acquired from *D. busckii*, several elements from the Asian
460 subgroup of the *D. melanogaster* group, (such as *P_226* with *D. elegans, Jockey_185*
461 with *D. rhopaloa* and *I_149* with *D. ficusphila*) and several Jockey elements are closely
462 related to elements found in the *Drosophila* clade species, such as with a *D. virilise*
463 ancestor. We compared the proportion of TEs showing HT events between species to
464 the patristic distance to each species, we find a significant negative linear correlation
465 between the species genic patristic distance and the proportion of TEs (Figure 2C,
466 Binomial GLM logistic regression, z-value = -7.88, p-value< 2e-16), agreeing with
467 previous findings that horizontal acquisition is more likely between closely related
468 species (Peccoud *et al.* 2017).
469
470 *Evidence of recent recurrent horizontal transfer between species*

471 In the *D. melanogaster* group, in addition to occasional bouts of catastrophic
472 invasion, many elements appear to have been transferred commonly between close
473 relatives in the group (Daniels *et al.* 1990; Clark and Kidwell 1997; Bartolomé *et al.*
474 2009). For *pseudoobscura* group TEs found in at least 3 species which had previously
475 described coding regions (101 TEs, 39 for comparisons to *D. affinis*), we compared the
476 silent site divergence (d_S) of TEs found between species to the d_S of host genes.
477 Overall, we found a significant reduction in synonymous divergence relative to host

genes for all comparisons (Mann-Witney U test *p-value* < 0.05), excluding those involving *D. affinis* (Figure 3A). We find 76 TE families below the 97.5% quantile of nuclear gene d_S in at least one comparison suggesting potentially recent transmission between species (51 of 62 LTRs, 19 of 30 LINEs and 6 of 8 DNA transposons). Inconsistent with horizontal transfer, there is not a depletion of non-LTR retrotransposon (LINE) elements found here.

We also compared the phylogenies of the TEs to that of the species, again looking for evidence of horizontal transfer. Of these families, 41 have phylogenies that differ from the species tree and group the two species with little divergence together. It is possible these differences are due to incomplete lineage sorting, or gene tree discordance, it is also possible that horizontal transfer has occurred for this family between these species, and so may support HT for 34 LTRs, 3 DNA transposons and 5 LINEs (All of which are below the 97.5% quantile for genomic d_S, Figure 3B; Table S2). Again, we find no evidence of exchange with *D. affinis*. In the *D. pseudoobscura* subgroup each species can hybridize with others to some degree (though likely not occurring in nature; Machado et al. 2007), therefore, we cannot determine if these apparent transfer events are true horizontal events or vertical transfer via hybridisation. We do see slightly elevated proportions of TIRs & LTRs when comparing phylogenies, consistent with horizontal transfer as suggested previously in (Sánchez-Gracia *et al.* 2005; Bartolomé *et al.* 2009). Conversely, we found d_S between species and *D. affinis* was significantly higher for TEs than host genes, consistent with the allopatric separation limiting HT events seen between species and possibly unconstrained evolution in the TEs (Figure 3A, Mann Whitney U test: $p < 3.5e\text{-}08$, Table S5).

By comparing Tajima's D for each TE in species we can look for strongly negative D, consistent with a copy number expansion following horizontal transmission (Bartolomé *et al.* 2009). All comparisons show equal levels of D in each species, close to 0, implying that each species already share the TE families, resulting in no expansion in copy number (Figure 3C). Consistent with this, we find all the TEs potentially shared between species have shared polymorphism (Figure 3E), which is not expected if acquisition is recent and purely horizontal. However, this result conflicts with our expectation from the nuclear d_S comparison, which we expect to be at similar levels to

509 TE d_S if there is hybridization. This result suggests pervasive transmission between
510 species, resulting in polymorphism being exchanged between species several times,
511 rather than once, resulting in no excess of low frequency polymorphism (Tajima 1989;
512 Bartolomé *et al.* 2009). Alternatively, there is less constraint on polymorphism in
513 transposable elements, allowing polymorphisms to drift to higher frequencies in shorter
514 periods of time following their horizontal acquisition.

515 Interestingly, 10 TE families appear to transfer between species in all
516 comparisons (d_S < 0.25% quantile: 1 TIR, 1 LINE and 8 LTRs), while 21 show no
517 evidence of transfer (d_S > 50% quantile: 1 TIR, 1 helitron, 11 LINEs and 9 LTRs),
518 suggesting that rates of transmission are highly dependent on the TE family and its
519 activity. We also see large differences in copy numbers of each family in each species.
520 We next looked to see if a lack of exchange can lead to changes in copy number of a
521 family and explain the differences between *D. persimilis* and others.

522 We compared changes in copy number over all the species (via the coefficient of
523 variation), for pervasively transferring TEs, non-transferring TEs and all other TEs. We
524 find no difference in the coefficient of variation of copy number for pervasively
525 transferring families and non-transferring families (Figure 3D; Mann Whitney U *p*-
526 value > 0.19 for all comparisons), suggesting that reduced transmission between
527 species isn't altering dynamics of the families compared to pervasively transferring
528 families. This low divergence and no evidence of family expansion has two possible
529 explanations: 1. There may be gene flow to some degree between these species in the
530 wild, while the genes are likely not introgressed due to incompatibilities or lower fitness,
531 their linked TEs will be transpose more readily after hybridisation, becoming unlinked
532 from this gene. This variant will then be maintained in the new host, resulting in reduced
533 divergence for the TE family between the species. 2. Due to the sympatry of the
534 *pseudoobscura* subgroup, there may have been recurrent horizontal transmission
535 between species, for TE families already present in each of the species, resulting in the
536 low d_S, but shared polymorphism and lack of copy number expansion. The lower
537 numbers of LINE families found exchanging between species supports the idea of
538 horizontal exchange, however the supposed numbers of exchanges (up to 61) between
539 species is unprecedented, giving more support to vertical exchange of TEs. Despite this

pervasive TE exchange of some families, TE dynamics may be changing within the species, leading to the differences seen in TE families, densities and family copy numbers.

Like *D. melanogaster*, the *D. pseudoobscura* group shows highly active TEs that appear to be constantly undergoing a cycle of acquisition, expansion and high activity, suppression and finally extinction. Strangely, despite TE exchange between species, the group shows distinct differences in TE content and TE densities. Though some of these differences are due to differences in quality of assembly of each species genome and method used to identify TE insertions, we find a distinct expansion in TE numbers in *D. persimilis*. We find these differences are likely due to stochastic differences in expansion and extinction between shared families, and not due to differing activities in novel and private families compared to these shared families. Overall this suggests that despite frequent gene flow, TE dynamics can evolve rapidly due to stochastic factors across the lifetime of a family.

Due to the history of the first recorded instance of a horizontal transfer of a transposable element, we tend to think that these transfers are rare, likely catastrophic events. However, an expanding body of evidence suggests that these events are likely a common occurrence throughout genomes, becoming more and more common the more closely related two species are. This transfer of elements is possibly even recurrent in some cases and, due to closely related sequences already established within piRNA clusters, do not cause the fitness costs seen with the invasion of P-element into *D. melanogaster.* Our results support the idea that TEs are highly fluid, moving between genomes easily without leading to the expansion of TE content in a species genome, or heavy catastrophic events such as was seen in laboratories with the invasion of P-element.

Acknowledgements

Thanks to L. Endler and D. Gómez Sánchez for discussion and advice concerning the bioinformatic analysis. We are grateful for helpful discussion provided by K. Senti, R. Kofler, B. Charlesworth, A. Clark, R. Unckless and J. Blumenstiel. Thanks for S. W.

570 Schaeffer and M. Noor for providing the data used in this survey and advice concerning
571 how the data should be used. The sequencing the *D. pseudoobscura* line sequencing
572 was funded by a National Institute of Health (NIH) grant R01-GM098478 to S. W.
573 Schaeffer. Funding for TH was provided by the Max Kade foundation.

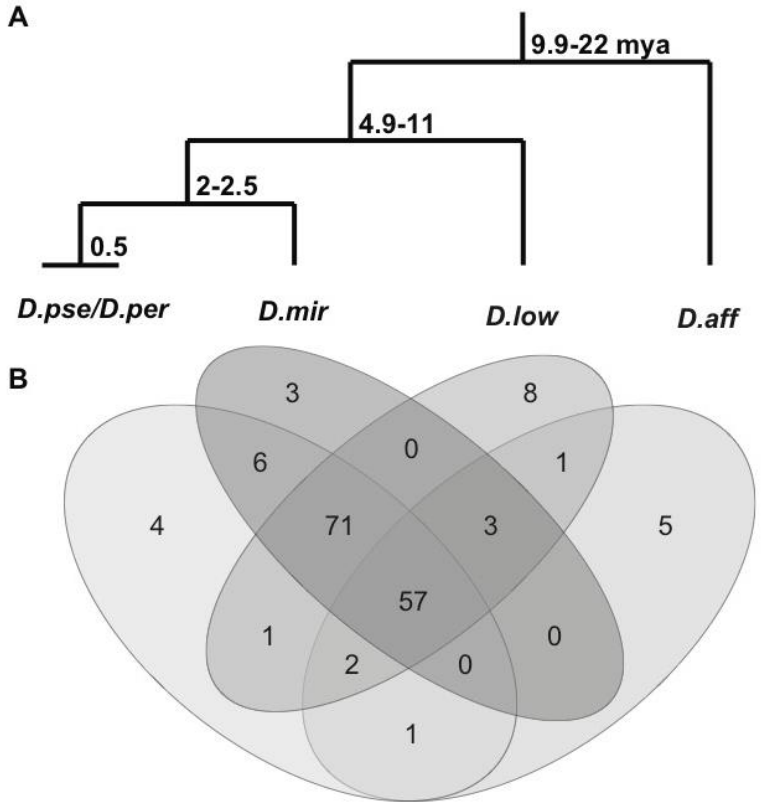
574

575 **Table 1:** Number of TE families and counts by order in species. % of reads mapping to
576 each order in each species, number of copies found based on coverage relative to
577 chromosome 3, % of the reference genome masked by each order for each species and
578 number of insertions found using PopoolationTE2 (Kofler *et al.* 2016a). As LTR
579 elements often exist not as complete insertions, but as solo-LTRs resulting from
580 illegitimate recombination, coverage for the LTR elements was estimated for both solo
581 LTRs and LTR bodies separately, with the average taken across the combined
582 sequences. We tested for extrachromosomal circular DNAs such as from Helitrons and
583 Polintons via comparisons between copy numbers and insertion numbers. We excluded
584 the unknown families from the total insertion counts.

			Reads		**Reference**
		n families	% reads	est. num	% genome
D. pseudoobscura	TIR	31	1.745	414	0.98
	LTR	72	8.875	2230	7.21
	LINE	35	3.633	1121	2.85
	RC	3	1.852	978	1.21
	Polinton	1	0.417	149	0.081
	Unknown	2	0.332	22	0.017
	Total *	142	16.522	4892	12.33
	Total (including Unknown)	144	16.854	4914	12.5
D. persimilis	TIR	31	1.547	413	1.29
	LTR	72	14.273	2260	12.95
	LINE	35	6.956	1301	5.76
	RC	3	4.43	1781	3.41
	Polinton	1	0.034	46	0.18

	Unknown	2	0.543	76	0.025
	Total *	142	27.24	5801	23.59
	Total (including Unknown)	144	27.78	5877	23.615
D. miranda	TIR	31	0.892	262	0.87
	LTR	67	7.19	973	2.21
	LINE	36	5.367	1431	1.25
	RC	5	1.484	1934	1.16
	Polinton	1	0.054	9	0.024
	Unknown	2	0.337	4	0.015
	Total *	140	14.987	4609	5.51
	Total (including Unknown)	142	15.324	4613	5.525
D. loweii	TIR	31	1.396	495	0.382
	LTR	74	6.883	1366	1.55
	LINE	34	3.839	933	0.799
	RC	5	1.245	813	0.363
	Polinton	1	0.054	7	0.013
	Unknown	9	0.641	265	0.087
	Total *	145	13.417	3614	3.1
	Total (including Unknown)	154	14.058	3879	3.187
D. affinis	TIR	9	0.872	278	0.177
	LTR	47	4.328	630	1.427
	LINE	13	5.223	530	0.406
	RC	4	1.351	369	0.245
	Polinton	1	0.068	35	0.041
	Unknown	10	1.192	206	0.098
	Total *	74	11.842	1842	2.29
	Total (including Unknown)	84	13.034	2048	2.394

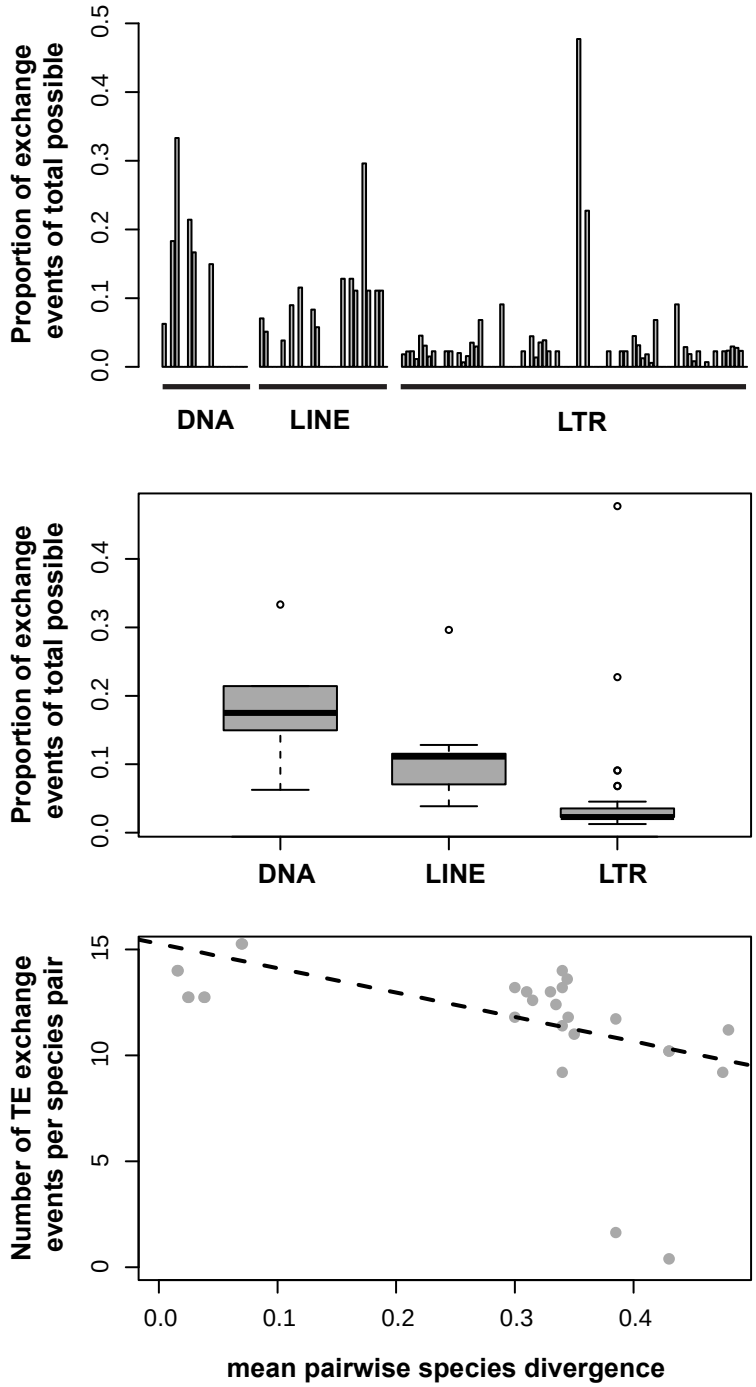
587 **Figure 1: A.** Phylogeny of the *D. pseudoobscura* group and the estimated time of
588 divergence between nodes. **B.** Number of TE families shared between species in the *D.*
589 *pseudoobscura* group, including putative novel families.



590

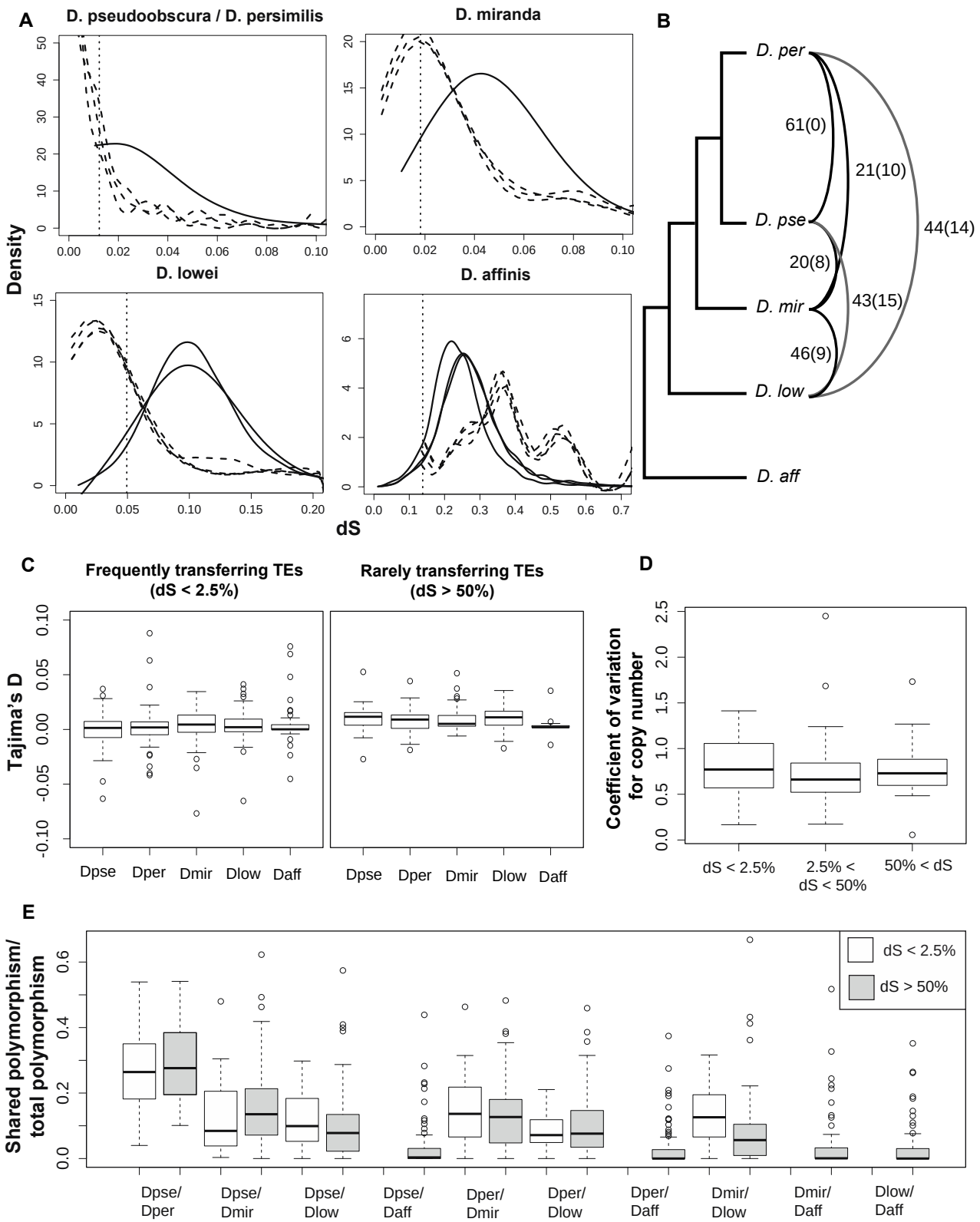
591

592 **Figure 2: A.** Each TE family and the proportion of times they show a lower divergence
593 value than the mean divergence between the host species. **B.** Of the transferring TEs,
594 the proportion of times these TEs are exchanging, grouped by TE order. **C.**
595 Comparisons between the proportion of exchange events and the pairwise divergence
596 between species, for exchanged TEs.



597

598 **Figure 3: A.** Pairwise comparison of silent site diversity (d_S) for nuclear genes (solid
599 line) and shared TEs (dashed lines) between *D. pseudoobscura, D. pseudoobscura*
600 *bogotana, D. persimilis* and other species. The lower 2.5% quartile for nuclear d_S is
601 shown as the dotted vertical line **B.** The number of transfer events for transposable
602 elements based on d_S, the number in brackets shows events that can be seen in the
603 assembled phylogenies. Note that many events could be occurring between species
604 vertically as well as horizontally. **C.** Comparison of Tajimas D across species for
605 frequently exchanged TEs and rarely exchanged TEs shows no difference, suggesting
606 no population expansion. **D.** No more variation in copy number of rarely exchanged TEs
607 than with frequently exchanged TEs. **E.** Proportion of shared nucleotide polymorphism
608 sites between TE sequences in species, out of total nucleotide polymorphism sites,
609 divided by TE families with low Ks relative to nuclear genes and TEs with higher d_S.


A
D. pseudoobscura / D. persimilis
D. miranda
D. lowei
D. affinis
Density
dS
B
D. per
D. pse
D. mir
D. low
D. aff
61(0)
21(10)
44(14)
20(8)
43(15)
46(9)
C
Frequently transferring TEs
(dS < 2.5%)
Rarely transferring TEs
(dS > 50%)
Tajima's D
Dpse
Dper
Dmir
Dlow
Daff
D
Coefficient of variation
for copy number
dS < 2.5%
2.5% <
dS < 50%
50% < dS
E
Shared polymorphism/
total polymorphism
dS < 2.5%
dS > 50%
Dpse/
Dper
Dpse/
Dmir
Dpse/
Dlow
Dpse/
Daff
Dper/
Dmir
Dper/
Dlow
Dper/
Daff
Dmir/
Dlow
Dmir/
Daff
Dlow/
Daff

610

611

Figure S1: Pipeline for TE annotation.

Figure S2: TE density across the genomes of each species, found using *PopoolationTE2,* sorted by TE order.

Figure S3: Comparison between putatively novel and known TE sequences for (A) length, (B) expression, (C) small RNA silencing expression and (D-F) copy number.

Figure S4: Distribution of TE copy numbers per species.

Table S1: *D. pseudoobscura* lines used in this study

Table S2: TEs found in *D. obscura* group. Sorted by if they are previously discovered or novel, then by Order and super family. Transmission states if the TE family is found to transfer between species

Table S3: Diagonal shows the total number of families found in each species for comparison.

Table S4: GLMs for three recombination maps versus TE accumulation, divided by order and super family. Done for both TE count (quasipoisson GLM) and TE density (binomial GLM). Significant values ($p < 0.05$) are shown in bold.

Table S5: For instances where no dS for nuclear comparisons are available, we used the dS between D. pseudoobscura and the species of interest.

Table S6: Number of unique and shared polymorphic sites for each species comparison, for each TE family.

Table S7: *D. pseudoobscura* TEs and the patristic distance from other TEs in their superfamily group, compared to the patristic distance between the TEs fly species and *D.pseudoobscura*

Bibliography

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

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

Avila V., Marion de Procé S., Campos J. L., Borthwick H., Charlesworth B., *et al.*, 2014 Faster-X effects in two Drosophila lineages. Genome Biol. Evol. 6: 2968–82.

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

Bosco G., Campbell P., Leiva-Neto J. T., Markow T. A., 2007 Analysis of Drosophila species genome size and satellite DNA content reveals significant differences among strains as well as between species. Genetics 177: 1277–1290.

Brennecke J., Malone C. D., Aravin A. A., Sachidanandam R., Stark A., *et al.*, 2008 An epigenetic role for maternally inherited piRNAs in transposon silencing. Science 322: 1387–1392.

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

Capy P., Koga A., David J. R., Hartl D. L., 1992 Sequence analysis of active Mariner elements in natural populations of Drosophila simulans. Genetics 130: 499–506.

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

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

Chen Z., Sturgill D., Qu J., Jiang H., 2014 Comparative validation of the D. melanogaster modENCODE transcriptome annotation. Genome …: 1209–1223.

Clark J. B., Kidwell M. G., 1997 A phylogenetic perspective on P transposable element evolution in *Drosophila*. Proc Nat Acad Sci USA 94: 11428–11433.

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

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

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

Dobzhansky T., 1935 The Y Chromosome of Drosophila Pseudoobscura. Genetics 20: 366–76.

Dobzhansky T., 1937 Further data on the variation of the Y chromosome in *Drosophila pseudoobscura*. : 340–346.

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

Duff M. O., Olson S., Wei X., Garrett S. C., Osman A., *et al.*, 2015 Genome-wide identification of zero nucleotide recursive splicing in Drosophila. Nature 521: 376–379.

Ellinghaus D., Kurtz S., Willhoeft U., 2008 LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18.

Ellison C. E., Bachtrog D., 2013 Dosage Compensation via Transposable Element Mediated Rewiring of a Regulatory Network. Science (80-.). 342: 846–850.

Fourmant M., Gibbs M. J., 2006 Patristic: a program for calculating patristic distances and graphically comparing the components of genetic change. BMC Evol. Biol. 6: 1–5.

Francisco F. O., Lemos B., 2014 How Do Y-Chromosomes Modulate Genome-Wide Epigenetic States : Genome Folding , Chromatin Sinks , and Gene Expression. J. Genomics 2: 94–103.

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

Chromosomal Inversions in. Genetics.
Ghildyal M., Zamore P. D., 2009 Small silencing RNAs: an expanding universe. Nat Rev Genet. 10: 94–108.
González J., Lenkov K., Lipatov M., Macpherson J. M., Petrov D. a, 2008 High rate of recent transposable element-induced adaptation in *Drosophila melanogaster*. PLoS Biol. 6: e251.
Gregory T. R., 2005 Synergy between sequence and size in large-scale genomics. Nat. Rev. Genet. 6: 699–708.
Gregory T. R., Johnston J. S., 2008 Genome size diversity in the family Drosophilidae. Heredity (Edinb). 101: 228–38.
Guindon S., Dufayard J.-F., Lefort V., Anisimova M., Hordijk W., *et al.*, 2010 New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst. Biol. 59: 307–21.
Haas B. J., Papanicolaou A., Yassour M., Grabherr M., Blood P. D., *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8: 1494–512.
Hartl D. L., Lozovskaya E. R., Nurminsky D. I., Lohe A. R., 1997 What restricts the activity of mariner-like transposable elements? Trends Genet. 13: 197–201.
Hellen E. H. B., Brookfield J. F. Y., 2013a The diversity of class II transposable elements in mammalian genomes has arisen from ancestral phylogenetic splits during ancient waves of proliferation through the genome. Mol. Biol. Evol. 30: 100–108.
Hellen E. H. B., Brookfield J. F. Y., 2013b Transposable element invasions. Mob. Genet. Elements 3: e23920.
Hill T., Schlötterer C., Betancourt A. J., 2016 Hybrid Dysgenesis in Drosophila simulans Associated with a Rapid Invasion of the P-Element. PLoS Genet. 12.
Hoen D. R., Hickey G., Bourque G., Casacuberta J., Cordaux R., *et al.*, 2015 A call for benchmarking transposable element annotation methods. Mob. DNA 6: 13.
Joshi N., Fass J., 2011 Sickle: A sliding window, adaptive, quality-based trimming tool for fastQ files. 1.33.
Kaplan N., Darden T., Langley C. H., 1985 Evolution and extinction of transposable elements in Mendelian populations. Genetics 109: 459–480.
Katoh K., Misawa K., Kuma K., Miyata T., 2002 MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 30: 3059–66.
Kofler R., Orozco-terWengel P., Maio N. de, Pandey R. V., Nolte V., *et al.*, 2011 Popoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6.
Kofler R., Betancourt A. J., Schlötterer C., 2012 Sequencing of pooled DNA Samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PloS Genet. 8: 1–16.
Kofler R., Hill T., Nolte V., Betancourt A. J., Schlötterer C., 2015a The recent invasion of natural Drosophila simulans populations by the P-element. Proc. Natl. Acad. Sci. U. S. A. 112.
Kofler R., Nolte V., Schlötterer C., 2015b Tempo and mode of transposable element activity in *Drosophila*. PLoS Genet 11: e1005406.

Kofler R., Daniel G., Schl C., 2016a PoPoolationTE2 : comparative population genomics of transposable elements using Pool-Seq. : 1–7.
Kofler R., Daniel G., Schlötterer C., 2016b PoPoolationTE2 : comparative population genomics of transposable elements using Pool-Seq. Mol. Biol. Evol.: 1–12.
Kohany O., Gentles A. J., Hankus L., Jurka J., 2006 Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474.
Kurtz S., 2010 The Vmatch large scale sequence analysis software Features of Vmatch. : 1–12.
Lerat E., Burlet N., Biémont C., Vieira C., 2011 Comparative analysis of transposable elements in the *melanogaster* subgroup sequenced genomes. Gene 473: 100–9.
Leslie L. K., Cohen J. T., Newburger J. W., Alexander M. E., Wong J. B., *et al.*, 2010 Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. Nat. Genet. 42: 6–10.
Levine M. T., Malik H. S., 2011 Learning to protect your genome on the fly. Cell 147: 1440–1441.
Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–60.
Lohe A. R., Moriyama E. N., Lidholm D. A., Hartl D. L., 1995 Horizontal transmission, vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol. Biol. Evol. 12: 62–72.
Machado C. a, Haselkorn T. S., Noor M. a F., 2007 Evaluation of the genomic extent of effects of fixed inversion differences on intraspecific variation and interspecific gene flow in Drosophila pseudoobscura and Drosophila persimilis. Genetics 175: 1289–306.
Maruyama K., Hartl D. L., 1991 Evolution of the transposable element Mariner in *Drosophila* Species. Genet. Soc. Am. 128: 319–329.
Noor M. a F., Garfield D. a, Schaeffer S. W., Machado C. a, 2007 Divergence between the Drosophila pseudoobscura and D. persimilis genome sequences in relation to chromosomal inversions. Genetics 177: 1417–28.
Obbard D. J., Gordon K. H. J., Buck A. H., Jiggins F. M., 2009 The evolution of RNAi as a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 364: 99–115.
Palmieri N., Kosiol C., Schlötterer C., 2014 The life cycle of *Drosophila* orphan genes. Elife 3: 1–21.
Peccoud J., Loiseau V., Cordaux R., Gilbert C., 2017 Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. 114: 4721–4726.
Petrov D. a, Fiston-Lavier A.-S., Lipatov M., Lenkov K., González J., 2011 Population genomics of transposable elements in *Drosophila melanogaster*. Mol. Biol. Evol. 28: 1633–1644.
Price A. L., Jones N. C., Pevzner P. a., 2005 De novo identification of repeat families in large genomes. Bioinformatics 21: 351–358.
Quinlan A. R., Hall I. M., 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–2.
Rahman R., Chirn G.-W., Kanodia A., Sytnikova Y. A., Brembs B., *et al.*, 2015 Unique transposon landscapes are pervasive across Drosophila melanogaster genomes.

Nucleic Acids Res. 43: 10655–72.
Richards S., Liu Y., Bettencourt B. R., Hradecky P., Letovsky S., *et al.*, 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res. 15: 1–18.
Rius N., Guillén Y., Delprat A., Kapusta A., Feschotte C., *et al.*, 2016 Exploration of the Drosophila buzzatii transposable element content suggests underestimation of repeats in Drosophila genomes. BMC Genomics 17: 344.
Sackton T. B., Hartl D. L., 2013 Meta-Analysis Reveals that Genes Regulated by the Preferentially Localized to Repressive Chromatin. Genome Biol. Evol. 5: 255–266.
Sánchez-Gracia A., Maside X., Charlesworth B., 2005 High rate of horizontal transfer of transposable elements in Drosophila. Trends Genet. 21: 200–203.
Schneider T. D., Stephens R. M., 1990 Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18: 6097–6100.
Sessegolo C., Burlet N., Haudry A., Biémont C., Vieira C., *et al.*, 2016 Strong phylogenetic inertia on genome size and transposable element content among 26 species of flies. Biol. Lett. 12: 521–524.
Smit A. F. A., Hubley R., 2008 RepeatModeler Open-1.0.
Sniegowski P. D., Charlesworth B., 1994 Transposable element numbers in cosmopolitan inversions from a natural population of Drosophila melanogaster. Genetics 137: 815–827.
Tajima F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
Tarailo-Graovac M., Chen N., 2009 Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinforma.
Vieira C., Lepetit D., Dumont S., Biémont C., 1999 Wake up of transposable elements following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. 16: 1251–5.
Yang Z., 2007 PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24: 1586–1591.

Figure S1: Pipeline for TE annotation.

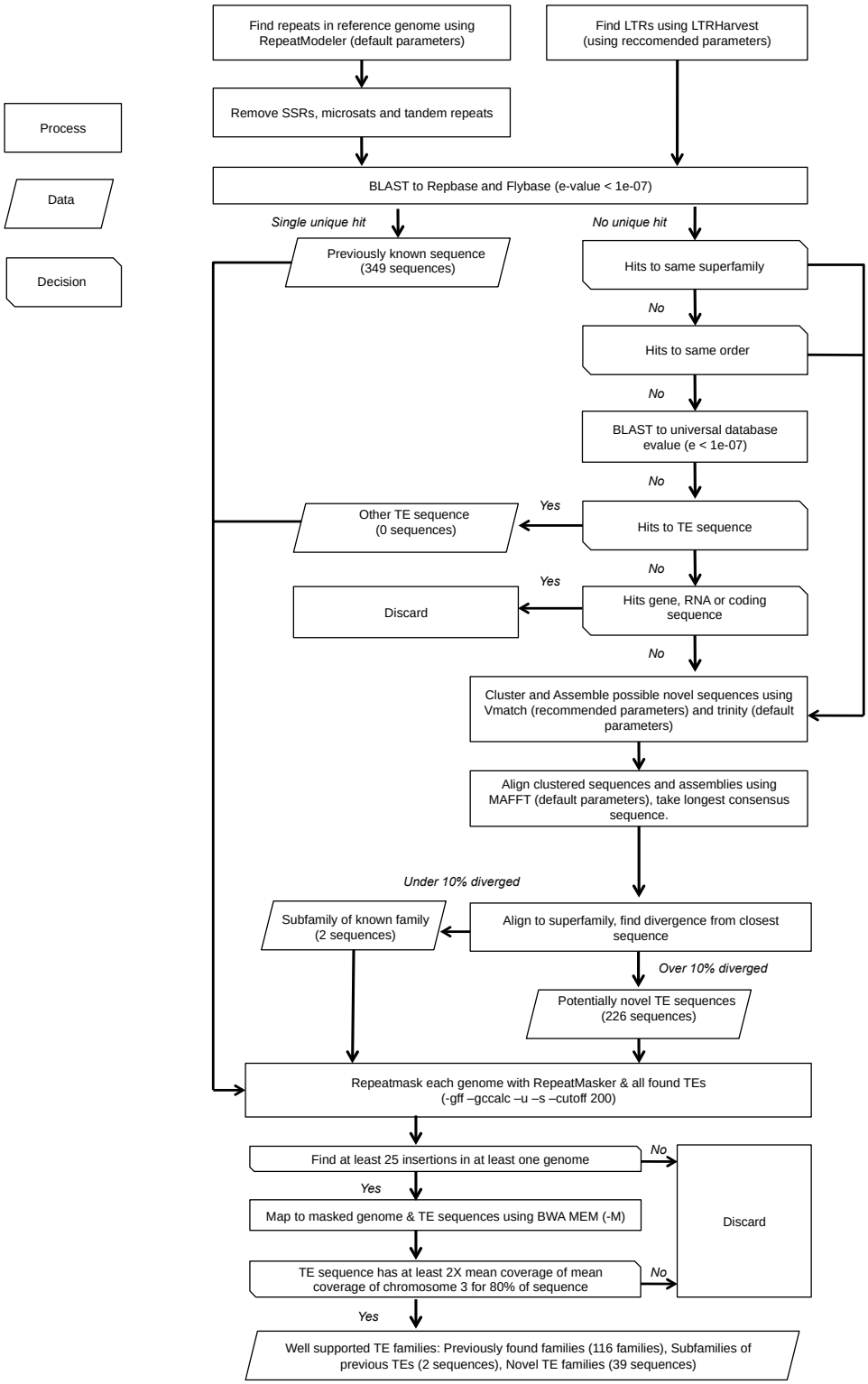


Figure S2: TE density across the genomes of each species, found using *PopoolationTE2,* sorted by TE order.

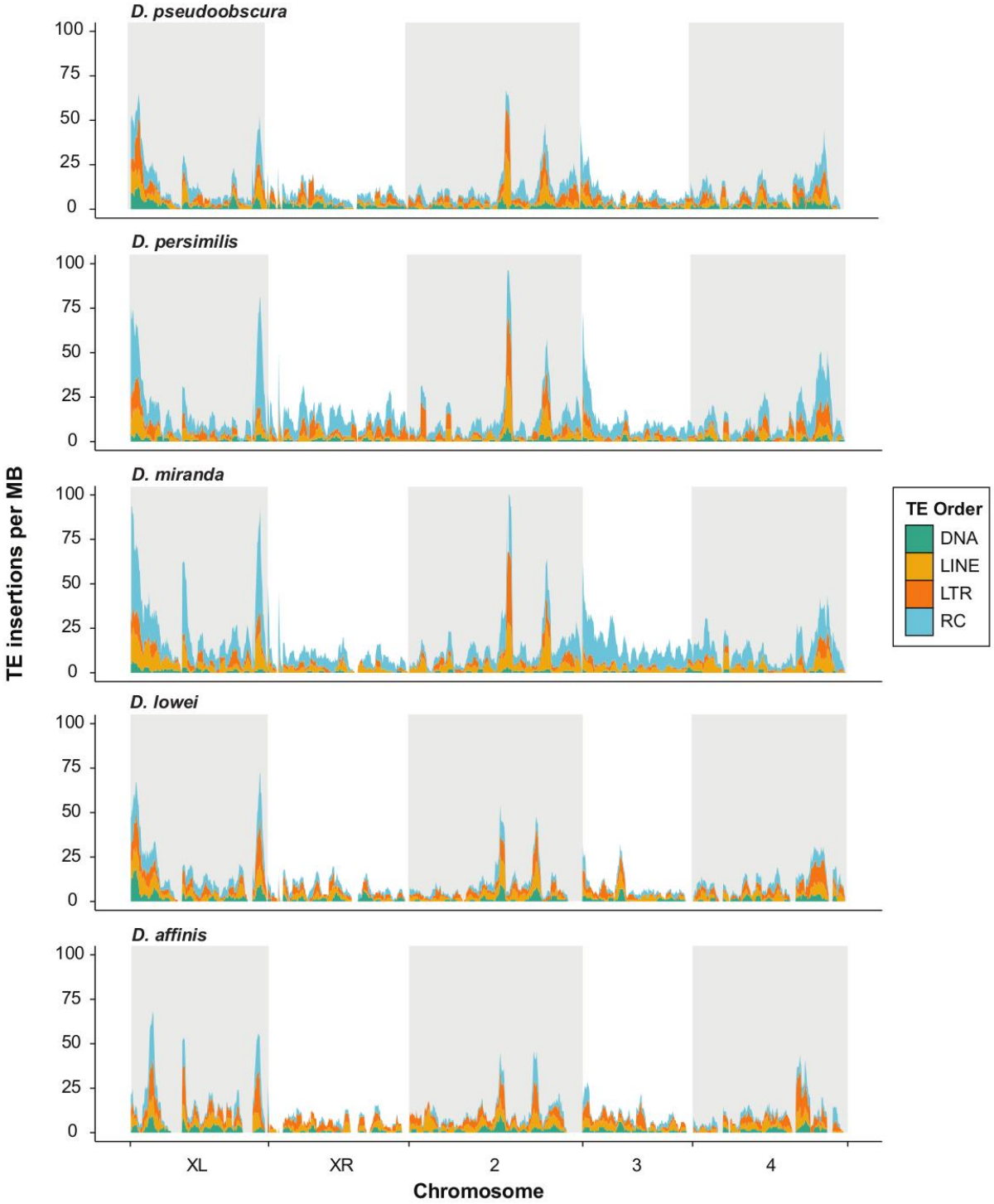
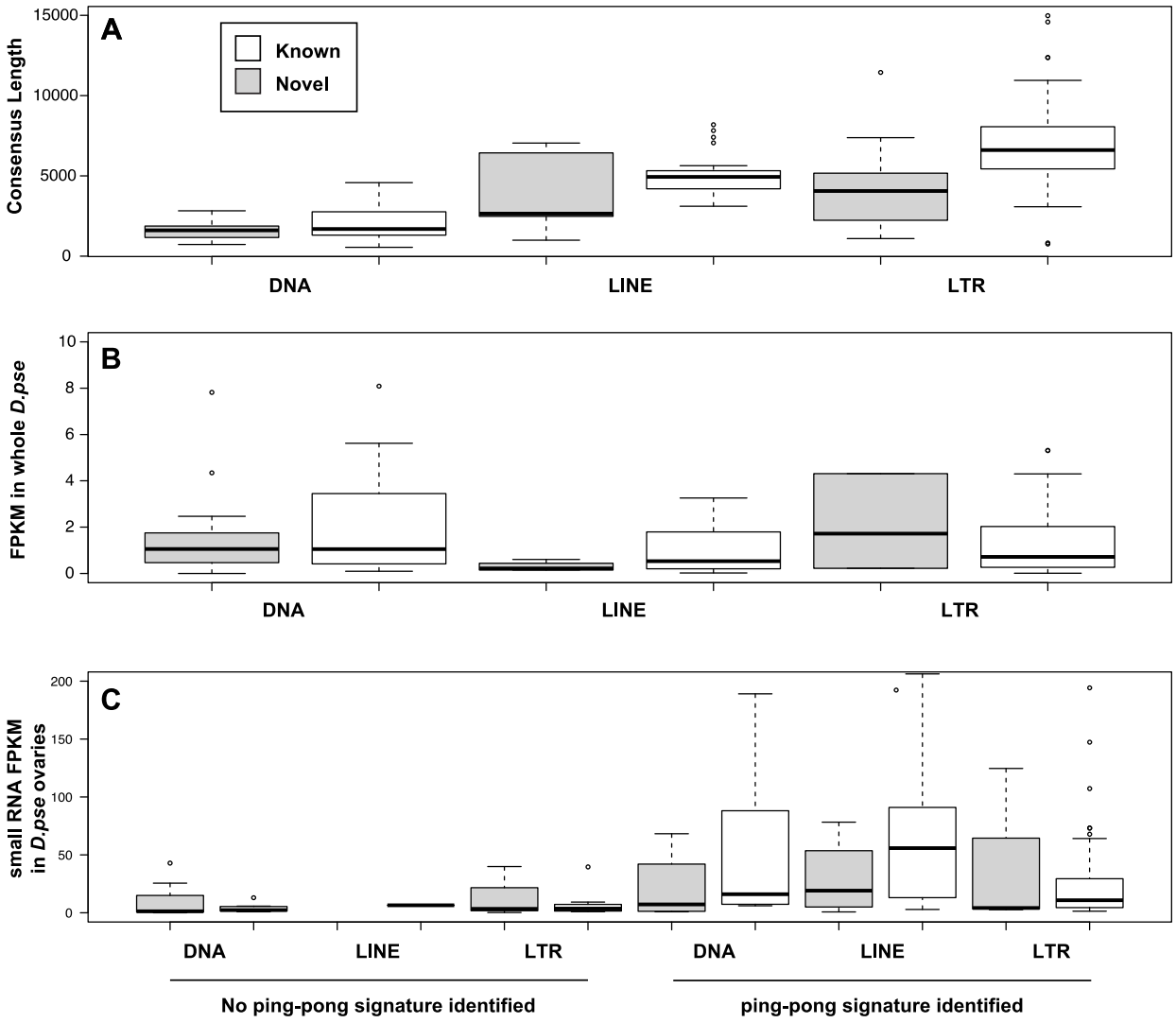
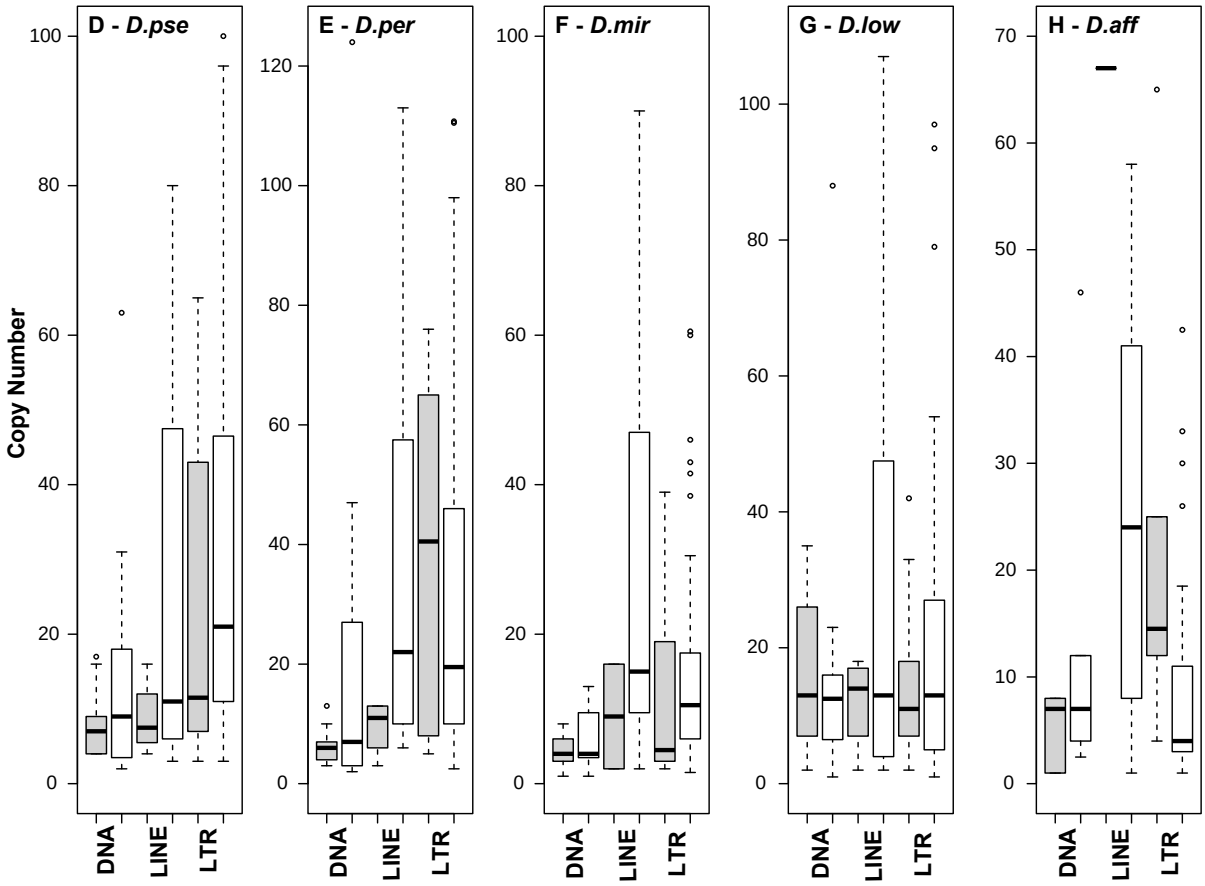


Figure S3: Comparison between putatively novel (grey) and known TE sequences (white) for (A) length, (B) expression, (C) small RNA silencing expression and (D-F) copy number.




D - D.pse
E - D.per
F - D.mir
G - D.low
H - D.aff
Copy Number
DNA
LINE
LTR

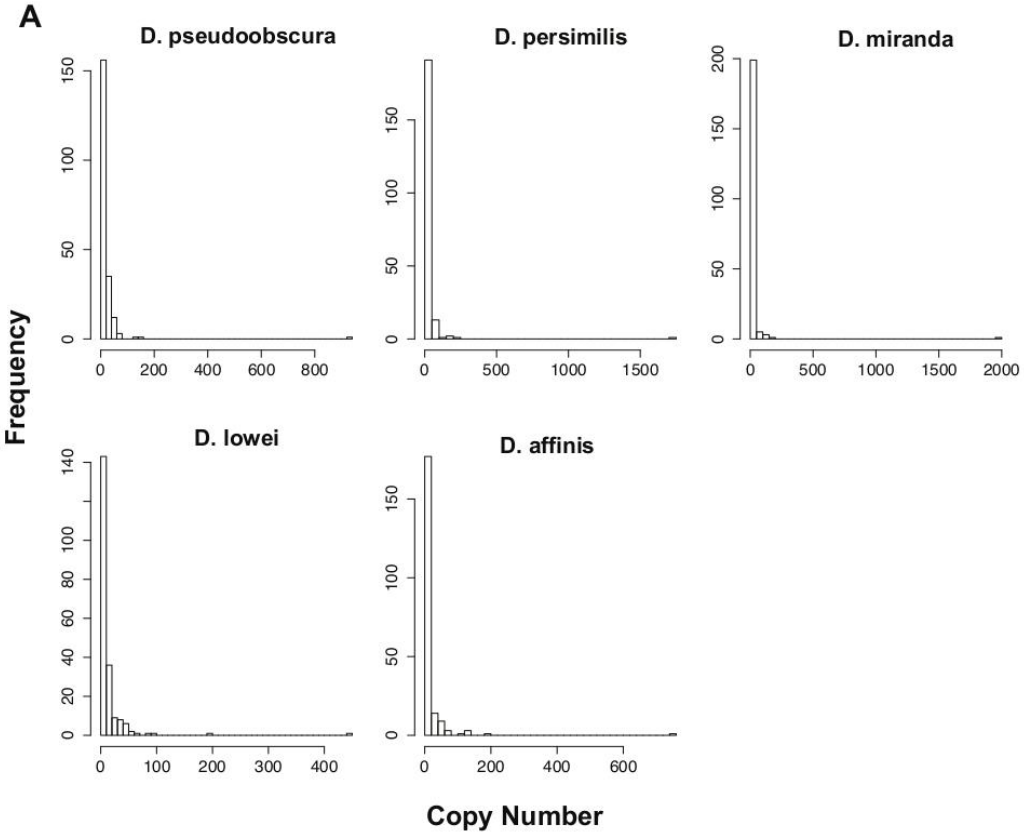



Figure S4: Distribution of TE copy numbers per species.