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

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

*Corresponding Author: tom.hill@ku.edu

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

145 families, or ancient horizontal acquisitions.

146 6. From Steps 1-5, we found 567 candidate TE sequences, 349 of which belong to
147 previously described TE families, including all 121 families previously found in the
148 *D. pseudoobscura* group ('known' families), and 445 others (putative 'novel'
149 families). We proceeded to filter sequences from this set which were represented
150 by very few or very short matches to the reference genomes.

151 a. First, we used the 567 sequences to repeat mask the reference genome of
152 each species using *RepeatMasker* (parameters: –no_is –norna –no_low –
153 gff –gccalc –u –s –cutoff 200) (Tarailo-Graovac and Chen 2009), following
154 recommendations in (Kofler *et al.* 2012). We required that the families have
155 at least 25 Repeatmasker hits in at least one species (237 sequences
156 retained, 116 known and 121 novel families).

157 b. We then estimated the copy number of each TE family for each species
158 from the Illumina short read data from adult females, discarding those
159 estimated to have a median coverage less than 2-fold that of the third
160 chromosome for less than 80% of the length of the sequence. To do this,
161 we mapped short reads to the repeated masked reference genome and the
162 237 TE sequences retained from the previous step using BWA MEM
163 (parameters: paired end –t 5 -M) (Li and Durbin 2009), and estimated
164 coverage with *bedtools genomecov* (Quinlan and Hall 2010). Due to the
165 poor assembly of the *D. persimilis* genome, we used a reference consisting
166 of the *D. pseudoobscura* genome and the *D. persimilis* TE sequences. (157
167 sequences retained, from 116 known and 41 families novel to this species
168 group).

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

172 Using this method, we found strong support for 114 of the 121 TE families
173 previously described in *D. pseudoobscura, D. persimilis* or *D. miranda* and 2 TEs
174 previously identified in other species. We also found 41 novel sequences, including
175 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 results as these species are closely related (0.018 average synonymous divergence (Noor *et al.* 2007)), and we find that similar numbers of reads map to TEs regardless of whether the *D. pseudoobscura* or *D. persimilis* genome is used (27.63% vs 27.27%).

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

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

218 *Expression confirmation of putative TE sequences.*

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

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

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

229 We first removed non-TE related small RNAs, following (Aravin *et al.* 2007;
230 Rahman *et al.* 2015), by mapping to a database of known *Drosophila* viruses and small
231 RNAs other than those that are TE-related, including miRNAs, viral siRNAs, snoRNA
232 (Rahman *et al.* 2015), using *BWA aln* and allowing for up to 3 mismatches (parameters:
233 -n 3) (Aravin *et al.* 2007; Li and Durbin 2009). We then mapped the remaining reads to

234 the repeat masked *D. pseudoobscura* reference genome and the novel and known TE
235 sequences identified in this study (*BWA aln* parameters: -n 3, maximum 2 alignments).

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

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

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

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

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

259 We estimated synonymous site divergence (d_S) in the TE sequences pairwise
260 between species using *codeml* (with transition–transversion rates estimated from the
261 data, and codon frequencies from the nucleotide frequencies) and the coding regions for
262 these TEs as annotated in *Repbase* (Kohany *et al.* 2006; Yang 2007). We then compared
263 d_S of TEs to that of orthologous genes between species obtained in the same way, taken

264 from Avila et al. (2014). Following Bartolomé et al. (2009), we considered an individual
265 family to show strong evidence of exchange if its d_S value was below the 2.5% quantile
266 of the d_S of all nuclear genes, to have potentially transferred if d_S was between the 2.5%
267 and 50% quantiles, and to show no evidence of transferring if above the 50% quantile.

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

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

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

282

283 **Results and Discussion**

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

285 We identified insertions of the 157 well-supported TE families in the reference genome of
286 the five species, and assessed their TE content using four metrics: the proportion of the
287 reference genome masked (using *RepeatMasker* (Tarailo-Graovac and Chen 2009)), the
288 proportion of short reads mapping to each TE sequences, the number of insertions in
289 each genome using short read data (using *PopoolationTE2* (Kofler *et al.* 2016b),
290 demonstrated across genomes in Supplementary Figure 2) and the estimated copy
291 number of each TE family (Table 1 and Table S2). We also estimated the density of TE
292 content across the genome (in masked bp/Mbp) using the proportion of the reference
293 masked by *Repeatmasker*.

294 Across all species, for all measures of TE content we find a significant linear
295 correlation between measures (Table S2, Spearman's Rank Correlation *p*-value <
296 0.00213), though the strength of the correlation is weak for all species between the
297 proportion of the genome masked at the family level versus the copy number of the TE
298 family, and the insertion count versus the proportion of the genome masked ($\rho < 0.58$),
299 suggesting that the proportion of the genome masked may be an inconsistent measure
300 of TE density. As expected, correlations between measures of TE content in the species
301 with genomes assembled only from short reads are lower (Table 1, Table S2) (Hoen *et al.*
302 2015; Rius *et al.* 2016), and the estimates of TE content for these species are likely more
303 unreliable. We therefore limit our analysis of TE content to the two genomes with available
304 long read data, *D. pseudoobscura* and *D. persimilis*. We also identified 15 sequences that
305 pass all filters, but cannot be assigned to a TE order, we have included these sequences
306 and their statistics, but have not included these sequences in further analyses (e.g. the 2
307 unknown sequences in *D. pseudoobscura*, Table 1, Table S2).

308 Because the *D. pseudoobscura* and *D. persimilis* genomes were originally
309 assembled from long reads (and the *D. pseudoobscura* genome has also been
310 assembled with the help of PacBio information and Sanger sequence information)
311 (Richards *et al.* 2005; Clark *et al.* 2007), the TE content of these two species is already
312 well-annotated. We found 116 previously identified *pseudoobscura* TE families using our
313 pipeline. We also found two TE families from other species, and 28 additional putative
314 TE families that passed all our filters in these two species. For *D. pseudoobscura*, we
315 were able to use RNAseq data from (Duff *et al.* 2015) to determine whether these
316 elements showed evidence of expression in embryos, using publically available
317 expression data. We estimated RPKM for both novel and known TEs from these data; of
318 the novel TEs, nine had appreciable levels of expression (Figure S3, FPKM > 1), a similar
319 proportion to that of the known TEs (49 of 116). Similarly, we used sequences from
320 embryonic small RNAs to ask if suppressive small RNAs are produced against these TE
321 families. We extracted TEs with at least 20 small RNAs mapping to them, which
322 comprised 114 of the 116 known TEs and all 28 of the novel TE sequences (Table S7 &
323 8). Most of these elements (108 of 140) had piRNAs generated against them (using the
324 24-29nt range generally used to identify piRNAs in other species (Ghildyal and Zamore

325 2009), and 27 elements also had homologous siRNAs (21-23nt small RNAs) (Figure S3,
326 Table S2) (Obbard *et al.* 2009). A subset of the piRNAs, those produced in the germline
327 (Aravin *et al.* 2007), are expected to show signatures of "ping-pong" amplification— small
328 RNAs that match both sense and anti-sense strands of the TE sequence, an enrichment
329 of these that show a 10bp overlap, a uracil in position 1 of sense strand piRNAs, and
330 adenosine in position 10 of the antisense small RNAs) (Aravin *et al.* 2007). We found
331 that 60 elements (53 known families and 7 novel; 36 LTRs, 15 LINEs, 7 DNA transposons
332 & 2 helitrons) showed signatures of ping-pong amplification— from inspecting *Sequence*
333 *Logos* plots (Table S2) (Schneider and Stephens 1990); novel and known elements
334 showed ping-pong small RNAs at similar rates (Figure S3, Mann-Whitney U test W = 24,
335 $p > 0.1676$). As these TE sequences are all multicopy, these measures of expression are
336 mainly useful to show that the putative novel TEs have characteristics like those of the
337 known TE sequences (Mann-Whitney U test W = 37, $p > 0.05$, Figure S3).

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

344 In the remaining species, we find 20 additional families not found in *D.*
345 *pseudoobscura* and *D. persimilis* (Figure 1B). The 57 TE families shared among all five
346 species constitute most of the TE content (73-84% of insertions and 53-78% of each
347 species reference TE content, Table S2), but vary in copy number between species (e.g.
348 HelitronN-1 in *D. miranda* and *D. lowei*, Table S2), possibly due to stochastic expansion
349 and loss of families over time. For example, we find *HelitronN-1_DPe* has 1927 insertions
350 and makes up 1.1% of the genome of *D. miranda*, while it has only 727 insertions in *D.*
351 *lowei,* comprising of 0.14% of the genome (Table 1, Table S2). This is likely due to the
352 collapsing of *Helitron-1* and the closely related ISX sequence that has been co-opted for
353 dosage compensation in *D. miranda* (Ellison and Bachtrog 2013). These differences can
354 be further seen in the distributions of copy numbers in families, which differ between
355 species (Figure S3, S4).

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

372 In *D. persimilis* we found the same TE families as in *D. pseudoobscura*, but
373 estimated 23.59% of the *D. persimilis* reference to be repetitive content versus 12.33% in
374 *D. pseudoobscura*, implying 21.3MB more repetitive content in the *D. persimilis* reference
375 genome compared to *D. pseudoobscura*. Previous annotation from the 12-genomes
376 project found lower TE content as a proportion of the genome than that found here (3%
377 and 8% *vs.* 12.33% and 23.59% here), but a similar ~2-fold enrichment in TEs for *D.*
378 *persimilis* (Clark *et al.* 2007). While it is true that *D. persimilis* has a larger genome than
379 *D. pseudoobscura* ((Gregory 2005), the two species genomes are estimated to differ only
380 by ~2Mb (Bosco *et al.* 2007; Gregory and Johnston 2008).

381 The higher TE content of *D. persimilis* is not due to the presence of additional
382 families, as the same families occur in both species (Table 1, Figure 1). In fact, as these
383 species hybridize occasionally (Noor *et al.* 2007), it would be surprising if their TE families
384 remained very distinct. Estimates of copy number from coverage of short read data
385 (collected from adult females in both species) shows more copies of each TE family in *D.*

386 *persimilis* than *D. pseudoobscura* (46.8 vs. 39.6 on average), but the difference is highly
387 non-significant (Mann-Whitney U, p = 0.669).

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

400 It is possible that an accumulation of TEs in the fixed inversions between
401 *D.pseudoobscura* and *D. persimilis* could explain the large difference in TE content, due
402 to the reduced genomic exchange in these regions (Machado *et al.* 2007), allowing
403 insertions to accumulate in one species but not the other. Consistent with this idea, we
404 find that LTR retrotransposons are at significantly higher densities of TE insertions within
405 these inverted regions in *D. persimilis* when compared to *D. pseudoobscura* and the
406 uninverted regions (File S2, Insertions per MB, using inversion windows defined in (Avila
407 *et al.* 2014); Mann Whitney U test: LTR inside inversions W = 53686, p = 5.674e-05, LTR
408 near inversions W = 16604, p-value = 0.1128 LTR outside inversions W = 290520, p-value
409 = 0.1407). However, we find RC and LINE insertions are at significantly higher densities
410 in *D. persimilis* regardless of genomic location (Insertions per MB, Mann Whitney U test:
411 W > 335780, p-value < 0.0001303 for inside, outside and near inverted regions) and no
412 difference in TIRs (W < 790, p-value > 0.37), suggesting that the fixed inversions are not
413 the only explanation.

414 One final possibility, the Y chromosome of these species may also play a role in
415 both the genome size and TE content differences of the species. While considerable
416 variation exists in the size of *D. pseudoobscura* Y chromosome size across types

417 (Dobzhansky 1935, 1937), the *D. persimilis* Y chromosomes are limited to the largest of
418 these types (Types I, II and III). As the *D. pseudoobscura* reference genome was likely
419 generated from a strain containing the smallest Y chromosome type
420 (Standard/Arrowhead, likely type V) (Dobzhansky 1937; Dobzhansky and Sturtevant
421 1937), while the *D. persimilis* genome strain used to generate their genome likely contains
422 the most common *D. persimilis* Y, the largest of the chromosome types (Dobzhansky
423 1937). Previous work has also found Y-linked variation in *D. melanogaster* and *D.*
424 *simulans* to be associated with phenotypic variation in a number of factors including TE
425 regulation, it is possible that a larger Y can cause poorer TE regulation, due to the
426 increased heterochromatin load in the genome (Sackton and Hartl 2013; Francisco and
427 Lemos 2014). This is possibly the case between *D. pseudoobscura* and *D. persimilis*,
428 where the larger Y chromosome may have led to the poorer regulation of TE families,
429 leading to the ~5Mb expansion of TEs in *D. persimilis*.
430
431 *Several transposable element families show evidence of ancient horizontal spread*
432 *between species*
433 As a majority of TEs are likely acquired in a species by horizontal transfer by closely
434 related species (Burt and Trivers 2006; Peccoud *et al.* 2017). We examined our set of
435 TEs for evidence that they had been horizontally acquired from another *Drosophila*
436 species by comparing the patristic distance of all *Repbase* TEs pairwise to the average
437 patristic distance of *pseudoobscura* group TEs (Kohany *et al.* 2006), after building a
438 phylogeny of each superfamily (Supplementary Figure 3).
439 Across 4096 pairwise comparisons, we found 230 where the TE patristic distance
440 was lower than the previously found genic distance (Table S7) (Chen *et al.* 2014). These
441 events were limited to 63 of 157 TEs, with most these TEs having lower patristic distances
442 than entire species groups (such as the *D. rhopaloa/D.elegans/D.ficusphila* species
443 subgroup), consistent with a transfer of the TE between the common ancestor of the
444 species and the *D. pseudoobscura* group, followed by a diversification into multiple
445 elements seen today. 42 of these transferred elements are LTRs, 6 are DNA transposons
446 and 15 are LINEs (Table S7, Figure 2A). While a higher proportion of LTRs are transferred
447 between species, each of these families only appears to have been exchanged with a

448 single species, rather than multiple, likely because of the recent expansion of LTRs in
449 *Drosophila*, compared to the more ancient expansion of most LINEs and DNA
450 transposons (Figure 2B). Among these transferred elements, we find a *piggyBac* element
451 that was acquired from *D. busckii*, several elements from the Asian subgroup of the *D.*
452 *melanogaster* group, (such as *P_226* with *D. elegans, Jockey_185* with *D. rhopaloa* and
453 *I_149* with *D. ficusphila*) and several Jockey elements are closely related to elements
454 found in the *Drosophila* clade species, such as with a *D. virilise* ancestor. We compared
455 the proportion of TEs showing HT events between species to the patristic distance to
456 each species, we find a significant negative linear correlation between the species genic
457 patristic distance and the proportion of TEs (Figure 2C, Binomial GLM logistic regression,
458 z-value = -7.88, p-value< 2e-16), agreeing with previous findings that horizontal
459 acquisition is more likely between closely related species (Peccoud *et al.* 2017).
460
461 *Evidence of recent recurrent horizontal transfer between species*

462 In the *D. melanogaster* group, in addition to occasional bouts of catastrophic
463 invasion, many elements appear to have been transferred commonly between close
464 relatives in the group (Daniels *et al.* 1990; Clark and Kidwell 1997; Bartolomé *et al.* 2009).
465 For *pseudoobscura* group TEs found in at least 3 species which had previously described
466 coding regions (101 TEs, 39 for comparisons to *D. affinis*), we compared the silent site
467 divergence (d_S) of TEs found between species to the d_S of host genes. Overall, we found
468 a significant reduction in synonymous divergence relative to host genes for all
469 comparisons (Mann-Witney U test *p-value* < 0.05), excluding those involving *D. affinis*
470 (Figure 3A). We find 76 TE families below the 97.5% quantile of nuclear gene d_S in at
471 least one comparison suggesting potentially recent transmission between species (51 of
472 62 LTRs, 19 of 30 LINEs and 6 of 8 DNA transposons). Inconsistent with horizontal
473 transfer, there is not a depletion of non-LTR retrotransposon (LINE) elements found here.

474 We also compared the phylogenies of the TEs to that of the species, again looking
475 for evidence of horizontal transfer. Of these families, 41 have phylogenies that differ from
476 the species tree and group the two species with little divergence together. It is possible
477 these differences are due to incomplete lineage sorting, or gene tree discordance, it is
478 also possible that horizontal transfer has occurred for this family between these species,

479 and so may support HT for 34 LTRs, 3 DNA transposons and 5 LINEs (All of which are
480 below the 97.5% quantile for genomic d_S, Figure 3B; Table S2). Again, we find no
481 evidence of exchange with *D. affinis*. In the *D. pseudoobscura* subgroup each species
482 can hybridize with others to some degree (though likely not occurring in nature; Machado
483 et al. 2007), therefore, we cannot determine if these apparent transfer events are true
484 horizontal events or vertical transfer via hybridisation. We do see slightly elevated
485 proportions of TIRs & LTRs when comparing phylogenies, consistent with horizontal
486 transfer as suggested previously in (Sánchez-Gracia *et al.* 2005; Bartolomé *et al.* 2009).
487 Conversely, we found d_S between species and *D. affinis* was significantly higher for TEs
488 than host genes, consistent with the allopatric separation limiting HT events seen
489 between species and possibly unconstrained evolution in the TEs (Figure 3A, Mann
490 Whitney U test: $p < 3.5e\text{-}08$, Table S5).

491 By comparing Tajima's D for each TE in species we can look for strongly negative
492 D, consistent with a copy number expansion following horizontal transmission (Bartolomé
493 *et al.* 2009). All comparisons show equal levels of D in each species, close to 0, implying
494 that each species already share the TE families, resulting in no expansion in copy number
495 (Figure 3C). Consistent with this, we find all the TEs potentially shared between species
496 have shared polymorphism (Figure 3E), which is not expected if acquisition is recent and
497 purely horizontal. However, this result conflicts with our expectation from the nuclear d_S
498 comparison, which we expect to be at similar levels to TE d_S if there is hybridization. This
499 result suggests pervasive transmission between species, resulting in polymorphism being
500 exchanged between species several times, rather than once, resulting in no excess of
501 low frequency polymorphism (Tajima 1989; Bartolomé *et al.* 2009). Alternatively, there is
502 less constraint on polymorphism in transposable elements, allowing polymorphisms to
503 drift to higher frequencies in shorter periods of time following their horizontal acquisition.

504 Interestingly, 10 TE families appear to transfer between species in all comparisons
505 (d_S < 0.25% quantile: 1 TIR, 1 LINE and 8 LTRs), while 21 show no evidence of transfer
506 (d_S > 50% quantile: 1 TIR, 1 helitron, 11 LINEs and 9 LTRs), suggesting that rates of
507 transmission are highly dependent on the TE family and its activity. We also see large
508 differences in copy numbers of each family in each species. We next looked to see if a

509 lack of exchange can lead to changes in copy number of a family and explain the
510 differences between *D. persimilis* and others.

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

531

532 Like *D. melanogaster*, the *D. pseudoobscura* group shows highly active TEs that
533 appear to be constantly undergoing a cycle of acquisition, expansion and high activity,
534 suppression and finally extinction. Strangely, despite TE exchange between species, the
535 group shows distinct differences in TE content and TE densities. Though some of these
536 differences are due to differences in quality of assembly of each species genome and
537 method used to identify TE insertions, we find a distinct expansion in TE numbers in *D.*
538 *persimilis*. We find these differences are likely due to stochastic differences in expansion

539 and extinction between shared families, and not due to differing activities in novel and
540 private families compared to these shared families. Overall this suggests that despite
541 frequent gene flow, TE dynamics can evolve rapidly due to stochastic factors across the
542 lifetime of a family.

543 Due to the history of the first recorded instance of a horizontal transfer of a
544 transposable element, we tend to think that these transfers are rare, likely catastrophic
545 events. However, an expanding body of evidence suggests that these events are likely a
546 common occurrence throughout genomes, becoming more and more common the more
547 closely related two species are. This transfer of elements is possibly even recurrent in
548 some cases and, due to the presence of closely related sequences within piRNA clusters,
549 reducing the fitness costs of a transfer event, such as the invasion of P-element into *D.*
550 *melanogaster.* Our results support the idea that TEs are highly fluid, moving between
551 genomes easily without leading to the expansion of TE content in a species genome, or
552 heavy catastrophic events such as was seen in laboratories with the invasion of P-
553 element.

554 **Acknowledgements**

555 Thanks to L. Endler and D. Gómez Sánchez for discussion and advice concerning the
556 bioinformatic analysis. We are grateful for helpful discussion provided by K. Senti, R.
557 Kofler, B. Charlesworth, A. Clark, R. Unckless and J. Blumenstiel. Thanks for S. W.
558 Schaeffer and M. Noor for providing the data used in this survey and advice concerning
559 how the data should be used. The *D. pseudoobscura* line sequencing was funded by a
560 National Institute of Health (NIH) grant R01-GM098478 to S. W. Schaeffer. Funding for
561 TH was provided by the Max Kade foundation.

562

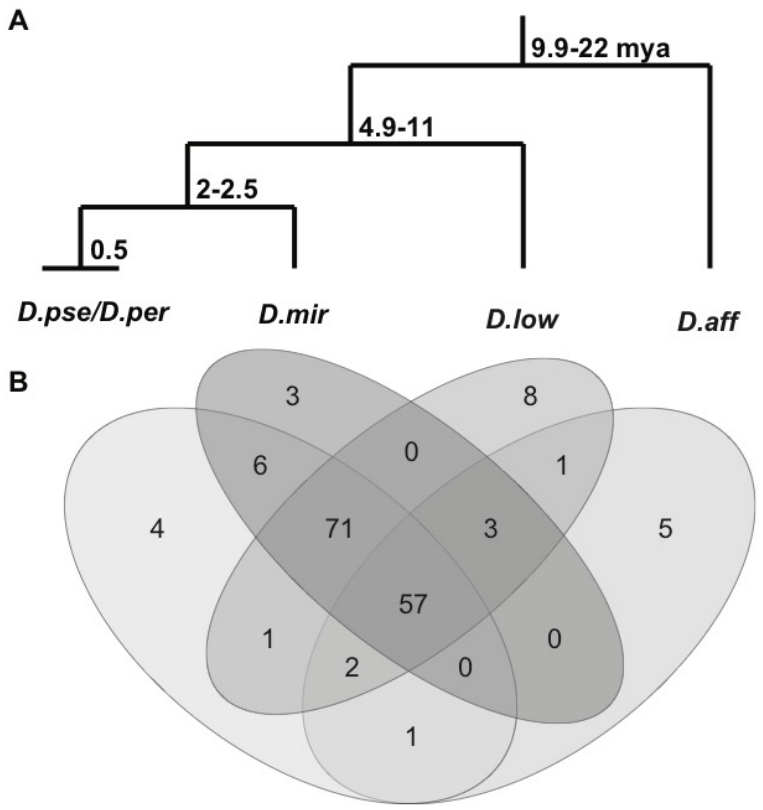
563 **Table 1:** Number of TE families and counts by order in species. % of reads mapping to
564 each order in each species, number of copies found based on coverage relative to
565 chromosome 3, % of the reference genome masked by each order for each species and
566 number of insertions found using PopoolationTE2 (Kofler *et al.* 2016a). As LTR
567 elements often exist not as complete insertions, but as solo-LTRs resulting from
568 illegitimate recombination, coverage for the LTR elements was estimated for both solo

569 LTRs and LTR bodies separately, with the average taken across the combined
570 sequences. We tested for extrachromosomal circular DNAs such as from Helitrons and
571 Polintons via comparisons between copy numbers and insertion numbers. We excluded
572 the unknown families from the total insertion counts.

			Reads		**Reference**	**PPTE2**
Species	TE Order	No. families	% reads	est. num	% genome	est. num
D. pseudoobscura	TIR	31	1.745	414	0.98	292
	LTR	72	8.875	2230	7.21	1846
	LINE	35	3.633	1121	2.85	927
	RC	3	1.852	978	1.21	978
	Polinton	1	0.417	149	0.081	29
	Unknown	2	0.332	22	0.017	6
	Total *	142	16.522	4892	12.33	4072
	Total (including Unknown)	144	16.854	4914	12.5	4078
D. persimilis	TIR	31	1.547	413	1.29	392
	LTR	72	14.273	2260	12.95	1919
	LINE	35	6.956	1301	5.76	958
	RC	3	4.43	1781	3.41	1755
	Polinton	1	0.034	46	0.18	46
	Unknown	2	0.543	76	0.025	7
	Total *	142	27.24	5801	23.59	5070
	Total (including Unknown)	144	27.78	5877	23.615	5077
D. miranda	TIR	31	0.892	262	0.87	258
	LTR	67	7.19	973	2.21	925
	LINE	36	5.367	1431	1.25	1059
	RC	5	1.484	1934	1.16	1934
	Polinton	1	0.054	9	0.024	9
	Unknown	2	0.337	4	0.015	4
	Total *	140	14.987	4609	5.51	4185

	Total (including Unknown)	142	15.324	4613	5.525	4189
D. loweii	TIR	31	1.396	495	0.382	381
	LTR	74	6.883	1366	1.55	740
	LINE	34	3.839	933	0.799	449
	RC	5	1.245	813	0.363	523
	Polinton	1	0.054	7	0.013	7
	Unknown	9	0.641	265	0.087	241
	Total *	145	13.417	3614	3.1	2100
	Total (including Unknown)	154	14.058	3879	3.187	2341
D. affinis	TIR	9	0.872	278	0.177	230
	LTR	47	4.328	630	1.427	832
	LINE	13	5.223	530	0.406	339
	RC	4	1.351	369	0.245	369
	Polinton	1	0.068	35	0.041	35
	Unknown	10	1.192	206	0.098	206
	Total *	74	11.842	1842	2.29	1805
	Total (including Unknown)	84	13.034	2048	2.394	2011

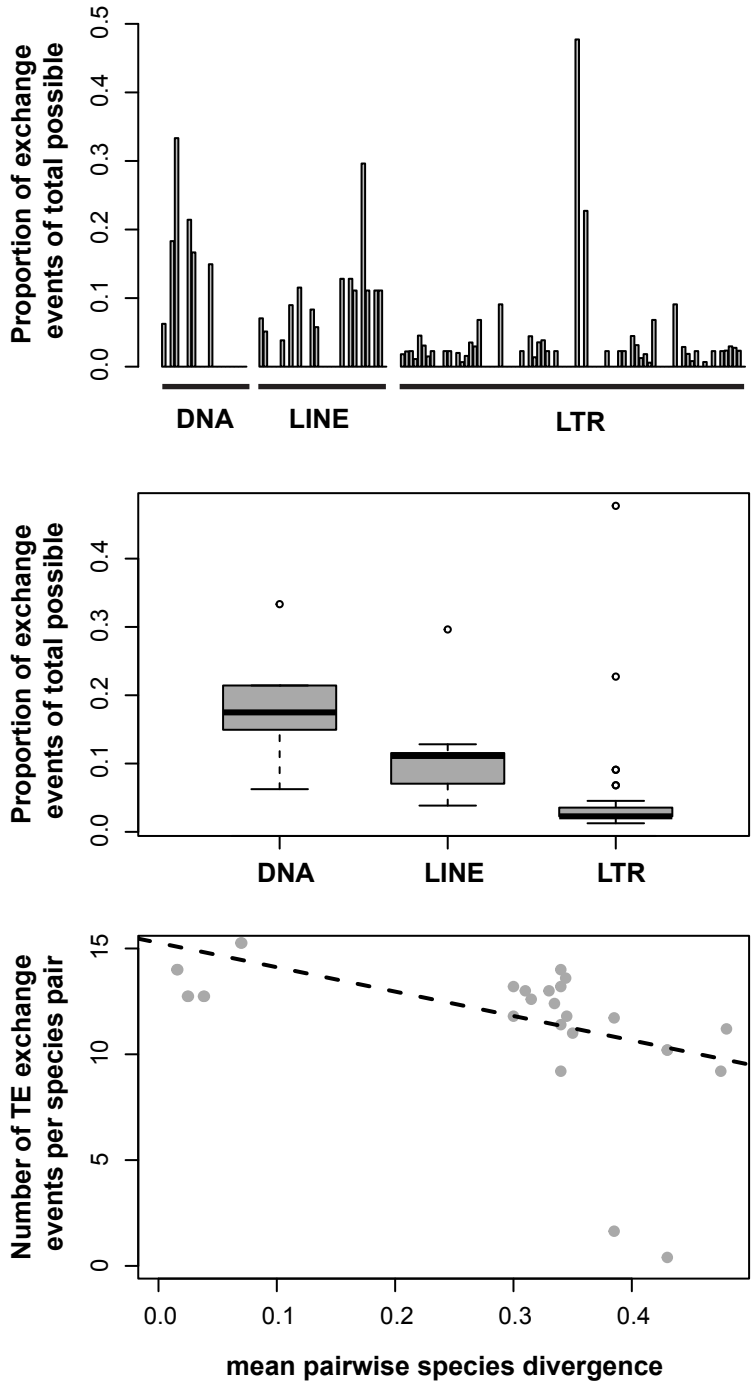
Figure 1: A. Phylogeny of the *D. pseudoobscura* group and the estimated time of divergence between nodes. **B.** Number of TE families shared between species in the *D. pseudoobscura* group, including putative novel families.


A
9.9-22 mya
4.9-11
2-2.5
0.5
D.pse/D.per
D.mir
D.low
D.aff
B
3
8
6
0
1
4
71
3
5
57
1
0
2
0
1

578

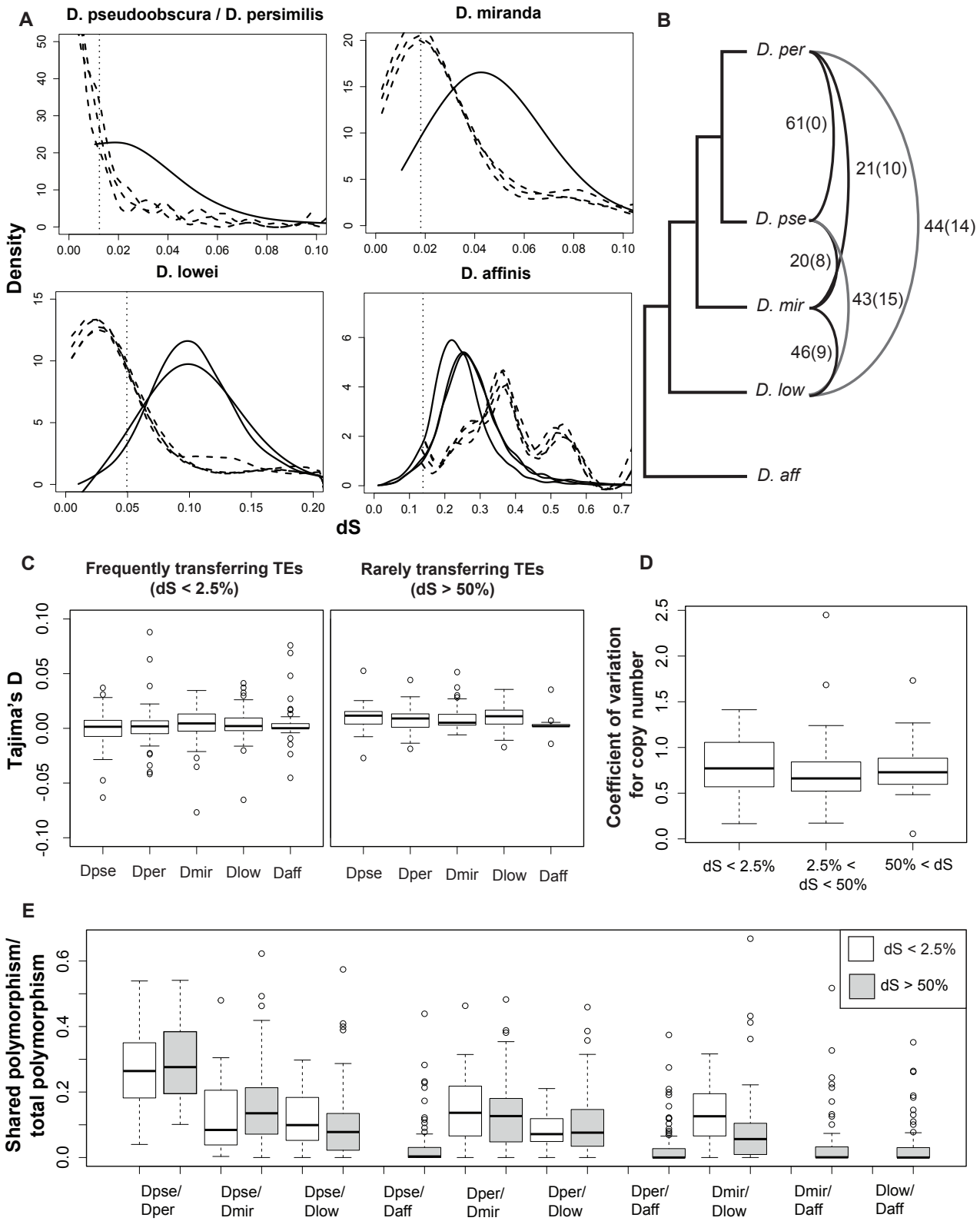
579

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



585

586 **Figure 3: A.** Pairwise comparison of silent site diversity (d_S) for nuclear genes (solid line)
587 and shared TEs (dashed lines) between *D. pseudoobscura, D. pseudoobscura bogotana,*
588 *D. persimilis* and other species. The lower 2.5% quartile for nuclear d_S is shown as the
589 dotted vertical line **B.** The number of transfer events for transposable elements based on
590 d_S, the number in brackets shows events that can be seen in the assembled phylogenies.
591 Note that many events could be occurring between species vertically as well as
592 horizontally. **C.** Comparison of Tajimas D across species for frequently exchanged TEs
593 and rarely exchanged TEs shows no difference, suggesting no population expansion. **D.**
594 No more variation in copy number of rarely exchanged TEs than with frequently
595 exchanged TEs. **E.** Proportion of shared nucleotide polymorphism sites between TE
596 sequences in species, out of total nucleotide polymorphism sites, divided by TE families
597 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

598

599

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.

632 Bosco G., Campbell P., Leiva-Neto J. T., Markow T. A., 2007 Analysis of Drosophila
633 species genome size and satellite DNA content reveals significant differences
634 among strains as well as between species. Genetics 177: 1277–1290.
635 Brennecke J., Malone C. D., Aravin A. A., Sachidanandam R., Stark A., *et al.*, 2008 An
636 epigenetic role for maternally inherited piRNAs in transposon silencing. Science
637 322: 1387–1392.
638 Burt A., Trivers R., 2006 *Genes in Conflict*.
639 Capy P., Koga A., David J. R., Hartl D. L., 1992 Sequence analysis of active Mariner
640 elements in natural populations of Drosophila simulans. Genetics 130: 499–506.
641 Charlesworth B., Langley C. H., 1989 The population genetics of *Drosophila*
642 transposable elements. Annu. Rev. Genet. 23: 251–87.
643 Charlesworth B., Langley C. H., Sniegowski P. D., 1997 Transposable element
644 distributions in *Drosophila*. Genetics 147: 1993–5.
645 Chen Z., Sturgill D., Qu J., Jiang H., 2014 Comparative validation of the D.
646 melanogaster modENCODE transcriptome annotation. Genome …: 1209–1223.
647 Clark J. B., Kidwell M. G., 1997 A phylogenetic perspective on P transposable element
648 evolution in *Drosophila*. Proc Nat Acad Sci USA 94: 11428–11433.
649 Clark A. G., Eisen M. B., Smith D. R., Bergman C. M., Oliver B., *et al.*, 2007 Evolution of
650 genes and genomes on the Drosophila phylogeny. Nature 450: 203–218.
651 Daniels S. B., Peterson K. R., Strausbaugh L. D., Kidwell M. G., Chovnick A., 1990
652 Evidence for horizontal transmission of the P transposable element between
653 *Drosophila* species. Genetics 124: 339–355.
654 DePristo M. A., Banks E., Poplin R., Garimella K. V, Maguire J. R., *et al.*, 2011 A
655 framework for variation discovery and genotyping using next-generation DNA
656 sequencing data. Nat. Genet. 43: 491–8.
657 Dobzhansky T., 1935 The Y Chromosome of Drosophila Pseudoobscura. Genetics 20:
658 366–76.
659 Dobzhansky T., 1937 Further data on the variation of the Y chromosome in *Drosophila*
660 *pseudoobscura*. : 340–346.
661 Dobzhansky T., Sturtevant A. H., 1937 Inversions In Chromosomes of <i>Drosophila
662 pseudoobscura. Genetics 23: 28–64.
663 Duff M. O., Olson S., Wei X., Garrett S. C., Osman A., *et al.*, 2015 Genome-wide
664 identification of zero nucleotide recursive splicing in Drosophila. Nature 521: 376–
665 379.
666 Ellinghaus D., Kurtz S., Willhoeft U., 2008 LTRharvest, an efficient and flexible software
667 for de novo detection of LTR retrotransposons. BMC Bioinformatics 9: 18.
668 Ellison C. E., Bachtrog D., 2013 Dosage Compensation via Transposable Element
669 Mediated Rewiring of a Regulatory Network. Science (80-.). 342: 846–850.
670 Fourmant M., Gibbs M. J., 2006 Patristic: a program for calculating patristic distances
671 and graphically comparing the components of genetic change. BMC Evol. Biol. 6:
672 1–5.
673 Francisco F. O., Lemos B., 2014 How Do Y-Chromosomes Modulate Genome-Wide
674 Epigenetic States : Genome Folding , Chromatin Sinks , and Gene Expression. J.
675 Genomics 2: 94–103.
676 Fuller Z. L., Haynes G. D., Richards S., Schaeffer S. W., 2016 Genomics of Natural
677 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.

724 Kofler R., Daniel G., Schl C., 2016a PoPoolationTE2 : comparative population genomics
725 of transposable elements using Pool-Seq. : 1–7.
726 Kofler R., Daniel G., Schlötterer C., 2016b PoPoolationTE2 : comparative population
727 genomics of transposable elements using Pool-Seq. Mol. Biol. Evol.: 1–12.
728 Kohany O., Gentles A. J., Hankus L., Jurka J., 2006 Annotation, submission and
729 screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC
730 Bioinformatics 7: 474.
731 Kurtz S., 2010 The Vmatch large scale sequence analysis software Features of Vmatch.
732 : 1–12.
733 Lerat E., Burlet N., Biémont C., Vieira C., 2011 Comparative analysis of transposable
734 elements in the *melanogaster* subgroup sequenced genomes. Gene 473: 100–9.
735 Leslie L. K., Cohen J. T., Newburger J. W., Alexander M. E., Wong J. B., *et al.*, 2010
736 Evolutionary flux of canonical microRNAs and mirtrons in Drosophila. Nat. Genet.
737 42: 6–10.
738 Levine M. T., Malik H. S., 2011 Learning to protect your genome on the fly. Cell 147:
739 1440–1441.
740 Li H., Durbin R., 2009 Fast and accurate short read alignment with Burrows-Wheeler
741 transform. Bioinformatics 25: 1754–60.
742 Lohe A. R., Moriyama E. N., Lidholm D. A., Hartl D. L., 1995 Horizontal transmission,
743 vertical inactivation, and stochastic loss of mariner-like transposable elements. Mol.
744 Biol. Evol. 12: 62–72.
745 Machado C. a, Haselkorn T. S., Noor M. a F., 2007 Evaluation of the genomic extent of
746 effects of fixed inversion differences on intraspecific variation and interspecific gene
747 flow in Drosophila pseudoobscura and Drosophila persimilis. Genetics 175: 1289–
748 306.
749 Maruyama K., Hartl D. L., 1991 Evolution of the transposable element Mariner in
750 *Drosophila* Species. Genet. Soc. Am. 128: 319–329.
751 Noor M. a F., Garfield D. a, Schaeffer S. W., Machado C. a, 2007 Divergence between
752 the Drosophila pseudoobscura and D. persimilis genome sequences in relation to
753 chromosomal inversions. Genetics 177: 1417–28.
754 Obbard D. J., Gordon K. H. J., Buck A. H., Jiggins F. M., 2009 The evolution of RNAi as
755 a defence against viruses and transposable elements. Philos. Trans. R. Soc. Lond.
756 B. Biol. Sci. 364: 99–115.
757 Palmieri N., Kosiol C., Schlötterer C., 2014 The life cycle of *Drosophila* orphan genes.
758 Elife 3: 1–21.
759 Peccoud J., Loiseau V., Cordaux R., Gilbert C., 2017 Massive horizontal transfer of
760 transposable elements in insects. Proc. Natl. Acad. Sci. 114: 4721–4726.
761 Petrov D. a, Fiston-Lavier A.-S., Lipatov M., Lenkov K., González J., 2011 Population
762 genomics of transposable elements in *Drosophila melanogaster*. Mol. Biol. Evol.
763 28: 1633–1644.
764 Price A. L., Jones N. C., Pevzner P. a., 2005 De novo identification of repeat families in
765 large genomes. Bioinformatics 21: 351–358.
766 Quinlan A. R., Hall I. M., 2010 BEDTools: a flexible suite of utilities for comparing
767 genomic features. Bioinformatics 26: 841–2.
768 Rahman R., Chirn G.-W., Kanodia A., Sytnikova Y. A., Brembs B., *et al.*, 2015 Unique
769 transposon landscapes are pervasive across Drosophila melanogaster genomes.

770 Nucleic Acids Res. 43: 10655–72.
771 Richards S., Liu Y., Bettencourt B. R., Hradecky P., Letovsky S., *et al.*, 2005
772 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal,
773 gene, and cis-element evolution. Genome Res. 15: 1–18.
774 Rius N., Guillén Y., Delprat A., Kapusta A., Feschotte C., *et al.*, 2016 Exploration of the
775 Drosophila buzzatii transposable element content suggests underestimation of
776 repeats in Drosophila genomes. BMC Genomics 17: 344.
777 Sackton T. B., Hartl D. L., 2013 Meta-Analysis Reveals that Genes Regulated by the
778 Preferentially Localized to Repressive Chromatin. Genome Biol. Evol. 5: 255–266.
779 Sánchez-Gracia A., Maside X., Charlesworth B., 2005 High rate of horizontal transfer of
780 transposable elements in Drosophila. Trends Genet. 21: 200–203.
781 Schneider T. D., Stephens R. M., 1990 Sequence logos: a new way to display
782 consensus sequences. Nucleic Acids Res. 18: 6097–6100.
783 Sessegolo C., Burlet N., Haudry A., Biémont C., Vieira C., *et al.*, 2016 Strong
784 phylogenetic inertia on genome size and transposable element content among 26
785 species of flies. Biol. Lett. 12: 521–524.
786 Smit A. F. A., Hubley R., 2008 RepeatModeler Open-1.0.
787 Sniegowski P. D., Charlesworth B., 1994 Transposable element numbers in
788 cosmopolitan inversions from a natural population of Drosophila melanogaster.
789 Genetics 137: 815–827.
790 Tajima F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA
791 polymorphism. Genetics 123: 585–595.
792 Tarailo-Graovac M., Chen N., 2009 Using RepeatMasker to identify repetitive elements
793 in genomic sequences. Curr. Protoc. Bioinforma.
794 Vieira C., Lepetit D., Dumont S., Biémont C., 1999 Wake up of transposable elements
795 following *Drosophila simulans* worldwide colonization. Mol. Biol. Evol. 16: 1251–5.
796 Yang Z., 2007 PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol.
797 24: 1586–1591.
798