Rapid evolution at the Drosophila telomere: transposable element dynamics at an intrinsically unstable locus.

Michael P McGurk (mpm289@cornell.edu)[*1], Anne-Marie Dion-Côté (anne-marie.dion-cote@umoncton.ca)[*1,2,3], Daniel A Barbash (barbash@cornell.edu)[1]

* *These authors contributed equally to this work*

1. Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY, 14853, United States
2. Department of Evolutionary Biology, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden
3. Present address: Département de Biologie, Université de Moncton, Moncton, NB, E1A 3E9, Canada

RUNNING TITLE

Rapid telomere evolution

KEYWORDS

Telomere, terminal deletions, transposable element, Drosophila, genomic conflict

CORRESPONDING AUTHOR

Daniel A. Barbash

Department of Molecular Biology and Genetics

403 Biotechnology Building

Cornell University

Ithaca, NY 14853

barbash@cornell.edu

ABSTRACT

Drosophila telomeres have been maintained by three families of active transposable elements (TEs), *HeT-A*, *TAHRE* and *TART*, collectively referred to as HTTs, for tens of millions of years, which contrasts with an unusually high degree of HTT interspecific variation. While the impacts of conflict and domestication are often invoked to explain HTT variation, the telomeres are unstable structures such that neutral mutational processes and evolutionary tradeoffs may also drive HTT evolution. We leveraged population genomic data to analyze nearly 10,000 HTT insertions in 85 *D. melanogaster* genomes and compared their variation to other more typical TE families. We observe that occasional large-scale copy number expansions of both HTTs and other TE families occur, highlighting that the HTTs are, like their feral cousins, typically repressed but primed to take over given the opportunity. However, large expansions of HTTs are not caused by the runaway activity of any particular HTT subfamilies or even associated with telomere-specific TE activity, as might be expected if HTTs are in strong genetic conflict with their hosts. Rather than conflict, we suggest instead that distinctive aspects of HTT copy number variation and sequence diversity largely reflect telomere instability, with HTT insertions being lost at much higher rates than other TEs elsewhere in the genome. We extend previous observations that telomere deletions occur at a high rate, and surprisingly discover that more than a third do not appear to have been healed with an HTT insertion. We also report that some HTT families may be preferentially activated by the erosion of whole telomeres, implying the existence of HTT-specific host control mechanisms. We further suggest that the persistent telomere localization of HTTs may reflect a highly successful evolutionary strategy that trades away a stable insertion site in order to have reduced impact on the host genome. We propose that HTT evolution is driven by multiple processes with niche specialization and telomere instability being previously underappreciated and likely predominant.

INTRODUCTION

Transposable elements (TEs) are genomic parasites that can increase their copy number within genomes by a variety of transposition mechanisms. While this provides a replicative advantage to the TE, their mobilization has consequences to the host genome that include DNA double-strand breaks, disruption of open reading frames and regulatory elements, and the perturbation of gene expression (Bourque *et al.* 2018). The presence of dispersed repeats further provides substrates for ectopic recombination, permitting large scale and potentially lethal genome arrangements (Deininger *et al.* 2003). The resulting conflict between TEs and the host genome may progress in several ways.

The simplest models suggest that TE copy number may stabilize at an equilibrium between the fitness advantage that transposition confers to the TEs and the fitness costs imposed on the host genome (Charlesworth and Charlesworth 1983; Charlesworth and Langley 1989). This stabilization is now known to involve complex interactions with the host-defense piRNA system that utilizes small RNAs to repress TE activity (Lee and Langley 2010; Blumenstiel 2011; Kelleher *et al.* 2020). TEs may further alter these dynamics by adopting strategies that mitigate the mutational burden they impose on the genome without sacrificing their replicative success (Cosby *et al.* 2019). Some TEs exhibit insertion site preferences that restrict the set of loci into which they transpose and could in principle limit the potential for deleterious insertions (Sultana *et al.* 2017). Arthropod R-elements are an extreme example, as they only insert in the highly repeated ribosomal DNA (rDNA) genes (Eickbush 2002). Other models that incorporate inactivating mutations suggest saltatory dynamics, where TE families are successively replaced by more active subfamilies until the family is eventually lost from the genome (Le Rouzic *et al.* 2007). TE exaptation or domestication, whereby TE regulatory or coding sequences are co-opted by their host, is another path by which such conflicts may be resolved (Jangam *et al.* 2017). These instances do not typically impact the dynamics of the entire TE family, but rather preserve only a portion of a single TE insertion while the rest of the family independently lives or dies. As such, most examples of domesticated TEs have lost their transposition activity.

The telomeric TEs of Drosophila are a remarkable exception to how this conflict typically proceeds, where the transposition activity of several TE families performs a function essential for

genome integrity. In most eukaryotes, telomeric DNA is comprised of simple repeats synthesized by telomerase that are assembled into a complex nucleoprotein structure. Telomeres serve two major roles: they protect chromosome ends from genetic attrition due to the end-replication problem, and they prevent chromosome ends from being recognized as DNA double-strand breaks, which can lead to chromosome fusions (de Lange 2009). In Diptera, the telomerase gene was lost (Mason *et al.* 2016) and in Drosophila its role in telomere elongation replaced by three non-LTR retrotransposons from the *jockey* clade: *HeT-A, TART* and *TAHRE* (herein collectively referred to as HTTs) (Levis *et al.* 1993; Abad *et al.* 2004). HTTs have several unique features compared to other non-LTR retrotransposons that reflect their specialized function (Mason *et al.* 2008; Pardue and DeBaryshe 2011; Arkhipova 2012). First, they are thought to exclusively insert at the ends of chromosomes, resulting in head-to-tail tandem arrays (Saint-Léandre *et al.* 2019; Pardue and DeBaryshe 2008). Second, they rely upon each other at several steps in their life cycles. *HeT-A* and *TARHE* in *D. melanogaster* carry promoters in their 3' UTR that drive expression of their neighboring element, an innovation only possible because of HTT tandem arrangement (Danilevskaya *et al.* 1997; Maxwell *et al.* 2006). Additionally, *HeT-A* may provide telomere specificity to *TART* and *TAHRE* through its Gag-like protein encoded by ORF1 (Fuller *et al.* 2010). Finally, *HeT-A* is non-autonomous (Biessmann *et al.* 1992a; b; 1994) — it likely depends upon the reverse transcriptase encoded by *TAHRE* and/or *TART* for integration.

Despite being ancestral to Drosophila and having comprised the telomeres for ~60 million years, prior work has uncovered extensive interspecific variation among HTTs, with new families and subfamilies having evolved across the genus and even among closely related species of the *melanogaster* subgroup (Villasante *et al.* 2008; Saint-Léandre *et al.* 2019). In addition, features typical of a family in one species, for example 3'-promoters or terminal repeats, may be absent in other species (Pardue and DeBaryshe 2011). This is a striking degree of variation given that HTTs are required for the fundamental role of protecting chromosome ends. These dramatic changes over the macroevolutionary scale must reflect a complex array of processes ongoing at the population level in ways distinct from more typical TEs. Yet, a comprehensive population genomic study of HTT variation remains to be

102 undertaken, which could provide the power to delineate the forces shaping their contemporary evolution.
103 We consider four potential forces shaping HTT variation at the population scale.

104 First, selection acting at the level of the organism has likely played an important role in both the
105 evolution and dynamics of the HTTs, in ways distinct from more typical TEs. The HTTs comprise the
106 Drosophila telomeres, which is often considered a paradigm of TE domestication, and should favor their
107 long-term persistence and transposition activity. Organismal selection may also favor increased levels of
108 domestication, bringing telomere elongation more tightly under host regulation. In addition, selection
109 might promote new HTT variants that increase organismal fitness in other ways, for example by reducing
110 telomere instability or by having pleiotropic beneficial effects on chromosome structure. Further, unlike
111 other TEs, the genome is under pressure to regulate rather than wholly suppress their activity.
112 Coordinating telomere elongation to prevent either complete erosion or overextension is a feature of
113 more typical telomerase-based systems (Stewart *et al.* 2012; Zhao *et al.* 2014), and mutational studies
114 have shown that Drosophila has recruited the piRNA and heterochromatin maintenance pathways
115 typically involved in TE suppression into this role (Shpiz and Kalmykova 2012). Host factors that interact
116 with the telomeres, such as the capping proteins that assemble onto them and the subtelomeric
117 sequence which directly abut them, may have also evolved properties that regulate HTT activity (Raffa *et*
118 *al.* 2011).

119 Second, selection can act on the HTTs themselves and place them in conflict with each other and
120 with their host, because they are active transposons and thus replicative entities in their own right. Both
121 the continual erosion of the telomeres and the potential fitness impacts of excessively long telomeres
122 imply that space in the telomere is a limiting resource for which all HTT families are competing. HTT
123 variants that are better able to increase in copy number will outcompete those that cannot, for example
124 by escaping regulation by the host genome. Conflict between the HTTs and the host genome (or among
125 the HTTs) is thus an ever present possibility. Indeed, like their feral cousins, studies using mutations in
126 HTT regulators suggest that they are capable of taking over given the opportunity (Shpiz and Kalmykova
127 2012). Similar to the rapid evolution of the Drosophila HTTs (Villasante *et al.* 2007, 2008), some
128 telomere-associated proteins show high evolutionary rates and signatures of positive selection (Raffa *et*

129 *al.* 2011; Lee *et al.* 2017). These evolutionary analyses led Lee et al. to propose that HTTs are in genetic
130 conflict with their Drosophila hosts (Lee *et al.* 2017). Complete resolution of conflict with the genome
131 would likely require somehow separating the reverse transcription machinery that extends the telomere
132 from the HTTs (Arkhipova 2012; Jangam *et al.* 2017; Markova *et al.* 2020).

133 Third, the dynamic and unstable nature of telomeres has the potential to shape HTT evolution.
134 Telomeres continually experience terminal erosion due to the end-replication problem. In addition, their
135 tandemly arrayed structure facilitates the amplification and deletion of sequence through unequal
136 exchange events, and chromosome breaks near the terminus can cause complete loss of a telomere
137 (Begun and Aquadro 1995; Langley *et al.* 2000; Kern and Begun 2008). The expansions, contractions,
138 and deletions of telomeric sequence also have the potential to heighten the rate at which polymorphisms
139 fix within HTT families and to increase the potential for the extinction of lineages, as observed recently in
140 *D. biarmipes* (Saint-Léandre *et al.* 2019). The net effect of these processes is to facilitate rapid sequence
141 change at telomeres in a manner driven by mutational processes rather than selection.

142 Fourth, the trade-off between reducing their impact on the genome versus inhabiting an unstable
143 locus has likely also shaped HTT evolution (Markova et al. 2020). Compared to much of the genome, the
144 telomere is a “safe harbor” that allows the HTTs to accumulate without disrupting essential host genes
145 and also minimizes the potential effects of ectopic recombination (Cosby *et al.* 2019). The tension
146 between this and the inherent instability of the telomeres disposing HTTs to high rates of deletion
147 suggests that the persistence of this strategy reflects the outcome of an evolutionary trade-off. Indeed,
148 we suggest that some of the peculiar features of HTTs, such as 3’-end promoters and head-to-tail
149 arrangement, are likely adaptations to cope with this instability (Pardue and DeBaryshe 2008). Another
150 possible outcome of evolutionary trade-offs (as well as of genetic conflict with other HTTs) would be to
151 escape the telomere. Surprisingly, despite transposing to chromosome ends, *TART* and *TAHRE* each
152 retain a potentially functional endonuclease domain (Casacuberta 2017), which raises the possibility that
153 the HTTs could generate double-strand breaks and insert randomly as do other non-LTR
154 retrotransposons. While limited surveys suggest that HTTs exclusively insert at telomeres in *D.*
155 *melanogaster,* the *melanogaster* subgroup species *D. rhopaloa* has many copies outside the telomere,

156 thus supporting the notion that HTTs can evolve the ability to escape from telomeres (Saint-Léandre *et*
157 *al.* 2019).

158 The extent to which the HTTs vary among species should reflect the compounded effects and
159 interaction among these evolutionary processes playing out over millions of years. However, the
160 timescales involved and the complexity of this array of processes makes it challenging to understand
161 which processes are responsible for which particular aspects of HTT rapid evolution. A clear picture of
162 within-species HTT variation offers the opportunity to understand how these processes play out over
163 shorter timescales where their impacts on HTT evolution can be more clearly distinguished. This requires
164 examining telomere variation along with other active TEs in a large population sample. To date, two
165 studies have presented telomere assembly solely in the reference genome of *D. melanogaster* (Saint-
166 Léandre *et al.* 2019; George *et al.* 2006), while length variation has been assayed only in a few lines, or
167 in a single population for only for a single site in *Het-A* (Siriaco et al. 2002; George et al. 2006; Wei et al.
168 2017).

169 Here, we leverage available population genomic data and the ConTExt pipeline (McGurk and
170 Barbash 2018) to comprehensively analyze HTT sequence and copy number polymorphism in the
171 Drosophila Global Diversity Lines (GDL) (Grenier *et al.* 2015). This dataset encompasses nearly 10,000
172 HTT insertions in 85 strains of *Drosophila melanogaster*, making it the largest study of telomere variation
173 in Drosophila to date. We compare the HTTs to other TE families, including those that are similarly
174 restricted to unstable genomic loci. Within this framework we ask whether the patterns of HTT copy
175 number, organization, and sequence variation require explanations beyond the standard conflict between
176 HTTs and the host genome and, if so, whether this reflects their remarkable symbiosis with the genome
177 or is instead the natural consequence of occupying the unstable telomeres.

178 **RESULTS**

179 ***Analyzing telomere variation in the Global Diversity Lines using ConTExt****:* To explore HTT
180 population variation, we leveraged the paired-end NGS data from the Global Diversity Lines, 85 stocks of
181 *D. melanogaster* collected from five populations: Beijing, Ithaca, Netherlands, Tasmania, and Zimbabwe
182 (Grenier *et al.* 2015). We employed the ConTExt pipeline (McGurk and Barbash 2018) to summarize
183 telomere structure, which is composed of head-to-tail HTT tandem arrays (Figure 1A). In brief, ConTExt
184 aligns repeat-derived reads to repeat-consensus sequences and uses paired-end information to identify
185 the junctions between repeats and neighboring sequence (Figure 1B). In the case of HTT insertions, the
186 neighboring sequence should generally be another HTT insertion or subtelomeric sequence (Figure 1A),
187 and all such junctions should involve the intact 3'-end of a HTT. The range of 5'-truncations resulting
188 from incomplete reverse transcription, which is frequent for all non-LTR retrotransposons, provides
189 additional power to distinguish independent insertions, allowing us to identify insertions and estimate TE
190 copy number within the telomeres.

191 With the data organized in this manner, we first determined sequence coverage relative to each
192 HTT consensus sequence, which provides an estimate of the number of nucleotides that each HTT
193 family contributes to the genome. When we use this estimate to calculate total telomere length, we
194 include all reads mapping to HTTs, but when we make statements about the copy number of particular
195 HTT families or subfamilies, we filter out any ambiguously aligned reads to avoid bias from mismapping
196 between related families/subfamilies (see Materials and Methods for details). While this estimates
197 telomere length, the propensity of HTTs to be 5'-truncated means that it is not a direct measure of copy
198 number: coverage of the 3' end reflects reads derived from all copies, but coverage of the 5' end reflects
199 only full-length insertions (which may, though, have internal deletions). Second, to more directly estimate
200 copy number, we counted the junctions between adjacent HTT insertions identified by ConTExt. This
201 allows us to infer the number of insertions, the degree to which individual insertions are truncated, and
202 which sequences neighbor a given HTT. To account for the possibility that some junctions are present in
203 multiple copies, we employed partial pooling to estimate copy number from the read counts at junctions.
204 The read depth of the junctions may be downwardly biased by several sources, particularly by mapping

205 quality filtering and the presence of additional sequence in the junction, such as poly-A tails, which we
206 account for by calibrating against a set of HTT junctions with known copy number (see Material and
207 Methods). The copy number estimates from modelling read depth are ~30% higher than from simply
208 counting junctions, suggesting some junctions are multicopy, but the estimates are highly correlated with
209 each other (slope = 1.3, r = 0.99, Figure 1C). Similarly, estimating the total telomere length from the
210 inferred copy number of each junction and the estimated extent of truncation of the corresponding HTT
211 concords well with the estimates of telomere length computed from raw coverage (slope = 1.06, r = 0.93,
212 Figure 1D). This suggests that our read count model reasonably accounts for the impacts of filtering out
213 ambiguous alignments. Finally, we examine the sequence of aligned reads themselves to assess for
214 variant alleles.

215 The strains in the GDL were sequenced at relatively low coverage, with the expected read count
216 over individual junctions ranging from ~10x to 20x across strains. Variation in read depth means that, in
217 some strains, we may be more likely to miss true HTT junctions, but it explains a relatively small fraction
218 of the variation in HTT copy number estimates (Figure S1A). Therefore, to assess how well ConTExt
219 summarizes telomere composition from low coverage NGS data, we used ART to simulate NGS libraries
220 from five available PacBio GDL genomes (Long *et al.* 2018), matching the read depth and insert size
221 distributions to the corresponding strain's NGS library (Huang *et al.* 2012). We next ran ConTExt on the
222 simulated data and estimated the copy number of the HTT subfamilies using the methods we applied to
223 the true GDL data. We then compared the copy number estimates for each HTT subfamily in the
224 simulated data against the true number evident in the corresponding PacBio assembly. We found these
225 to be strongly concordant, with no evidence of strong upward or downward biases (slope = 1.15, r^2 =
226 0.88, p = 1e-16, Figure 1E). We do note one outlier (Beijing line B59) where ConTExt underestimates
227 *HeT-A* copy number, which likely reflects that, despite performing generally quite well, highly divergent
228 copies of a subfamily are more difficult to capture under our alignment parameters. Finally, the original
229 study of SNPs in the GDL reported some residual heterozygosity, which would both double the number
230 of HTT-HTT junctions and halve their read depth in heterozygous tracts. In the real NGS data, we noted
231 that the read depth over HTT-HTT junctions was about half of what was expected given the coverage of

232 unique sequence, which we corrected for by calibrating against HTT junctions with known copy number,
233 i.e. healed terminal deficiencies (see Materials and Methods, "*Correcting for additional biases in read*
234 *counts over junctions"*). We found that the simulations recapitulate this feature of the data (Figure 1F),
235 suggesting that the alignment process and subsequent filtering of ambiguous reads is likely responsible
236 for reduced coverage over HTT-HTT junctions, rather than residual heterozygosity or
237 underrepresentation of repetitive sequences in the data. We note, however, that our analysis of terminal
238 deficiencies provides evidence of some residual heterozygosity at the telomeres, which we discuss
239 further below. Overall, we find that ConTExt estimates are sufficiently reliable to draw general
240 conclusions about telomere structure in the NGS GDL data.
241 A final aspect of the dataset that must be considered is that the GDL stocks were collected from
242 five populations and inbred for more than 5 years, meaning that lab evolution may be substantial and
243 heterogeneous across the lines. This could intensify background effects on TE activity, as any
244 polymorphic TE insertions and host alleles that affect their activity will frequently become fixed together
245 in a line, whereas in natural populations they would typically be separated by recombination and
246 independent assortment. In most of our analyses, this should not greatly impact our conclusions. For
247 example, the arrangement of HTTs within telomeres should not be highly sensitive to such continued
248 evolution. Lab evolution, furthermore, may present an opportunity to observe more extreme patterns of
249 some biases that do occur in nature through continued mis-regulation. As we describe below, the
250 clearest impact of continued evolution is evident in our analyses as strain-specific copy-number
251 expansions of some TE families, patterns that are unlikely to be observed in recently collected flies.
252
253 ***Telomeres are highly dynamic in Drosophila melanogaster****: Drosophila melanogaster* telomeres are
254 comprised of three non-LTR retrotransposon families *HeT-A*, *TAHRE*, and *TART*, which in our analyses
255 are further subdivided into five *HeT-A* subfamilies (*HeT-A*, *HeT-A1*, *HeT-A-2*, *HeT-A3*, *HeT-A5*) and
256 three *TART* subfamilies (*TART-A*, *TART-B1*, *TART-C*) (see Materials and Methods for details, Table S1
257 for reference insertion coordinates and Supplementary File 1 for consensus sequences). For clarity, we
258 use the terms "*HeT-A* family" and "*TART* family" when referring to the families. Telomere length has

259 previously been reported to vary by two orders of magnitude (Wei *et al.* 2017). However, that survey was
260 based on qPCR estimates of abundance using a primer set that was designed to detect a single
261 consensus *HeT-A*, in a single North American population, so the absolute size range of telomeres in *D.*
262 *melanogaster* remains unknown.

263 To gain a more comprehensive picture of the length and composition of *D. melanogaster*
264 telomeres, we consider how the size and composition of telomeres vary among strains in terms of the
265 sequence abundance and copy number of all HTT families. The median total telomere length in the GDL
266 is 400 kb with a median of 98 HTT insertions per genome, corresponding roughly to 50 kb and 12
267 insertions per each of the eight telomeres (Figure 2A, Table S2). The majority of insertions belong to the
268 *HeT-A* family (64%), with *TAHRE* (17%) and the *TART* family (19%) contributing fewer insertions (Figure
269 2B). While these proportions vary across strains, insertions of each of the three families are identifiable in
270 each of the 85 GDL strains. The relative abundance of *HeT-A* and *TART* family elements is roughly
271 equivalent to previous estimates based on Sanger sequencing (Abad *et al.* 2004; George *et al.* 2006).
272 The approximately equal abundance of *TAHRE* and *TART* family elements contrasts with an analysis of
273 the 4R telomere in the reference strain, which contains 8 *TART* family insertions but no *TAHRE* elements
274 (George *et al.* 2006). This may reflect a peculiar structure of the 4R telomere, because analysis of three
275 other telomeres in the reference strain identified 4 *TAHRE* elements (Abad *et al.* 2004). Other strains
276 also show *TAHRE* on a subset of telomeres (Shpiz *et al.* 2007). The majority of HTT insertions are
277 incomplete, with about 80% of *HeT-A* family and *TAHRE* insertions and 93% of *TART* family insertions
278 being truncated to various degrees (Figure 2C). This is similar to what George et al. (2006) observed on
279 telomere 4R, where 75% of *TART* family insertions (6/8) and 66% of *HeT-A* family insertions (10/15)
280 were incomplete. While the proportion of truncated elements varies considerably among strains, we
281 found that there are no consistent differences among populations.

282 Telomere length displays appreciable variation among individuals, with an interquartile range of
283 193kb and 58 HTT insertions (Table S2). In addition, there is clear potential for extreme gains and losses
284 of telomere sequences, which are observed as significant outliers. Several strains have much shorter
285 total telomere length, with the shortest (ZW140) being 143 kb and having 42 insertions (average of 17 kb

286 or 5 insertions per telomere). Some of the shortest strains are among those that contain unhealed
287 telomere deficiencies (see “Terminal deficiencies are common” below). At the opposite extreme, in I01
288 the telomeres have expanded to 1.7 Mb of sequence, corresponding to 692 insertions (average of 212
289 kb or 86 insertions per telomere). The range of telomere lengths is similar to that found in an earlier study
290 of several strains (including the long-telomere strain *GIII*) that used a different approach (Southern
291 blotting) (George *et al.* 2006). There is some population structure to telomere length, with the estimate of
292 mean telomere length in the Beijing population being 110-140kb longer than the other four populations
293 (Ithaca, Netherlands, Tasmania and Zimbabwe), corresponding to 17-40 more insertions per strain on
294 average. This holds after accounting for how differences in read depth across strains impact the
295 probability of not observing a junction that is truly present (Figure S1A). Increases in the number of *HeT-*
296 *A* family elements drives this heightened copy number in Beijing strains, especially the *HeT-A*1, *HeT-A*2,
297 and *HeT-A*5 subfamilies (Figure S1 B-J).
298
299 ***HTT family composition is similar in long and short telomere strains:*** We described above that the
300 Beijing population has elongated telomeres, and that this is driven by expansions of particular *HeT-A*
301 subfamilies. We thus wondered whether telomere elongation reflects escape from host regulation by
302 specific HTT subfamilies. If a specific HTT escapes host control and drives telomere elongation across
303 lines, this subfamily should be over-represented in long telomere strains compared to short telomere
304 strains. We analyzed telomere composition by comparing HTT family and subfamily copy number and
305 relative proportions between long (90th percentile, > 163 HTT insertions) and short (10th percentile, < 54
306 insertions) telomere strains. At the family level, there is no apparent difference in telomere composition
307 between short and long telomere strains (Figure 2D). As the copy number of *HeT-A* family, *TAHRE* and
308 *TART* family elements increases in long telomere strains, so does their relative proportion, resulting in
309 similar telomere composition. However, some differences arise at the subfamily level (Figure 2E). In long
310 telomere strains, the copy number of *HeTA-1*, *HeTA-5*, *TART-B1* and *TART-C* increases less
311 substantially than other sub-families, resulting in a reduced representation in long telomere strains.
312 Altogether, these observations suggest that while telomere length is under host control, HTT sub-families

313 respond differently to a more or less transposition-permissive state, and that long telomere strains do not
314 result from host regulation escape of specific HTT subfamilies.
315
316 ***Copy number expansions are not specific to HTTs:*** The pronounced copy number expansions of
317 HTTs we observed in some strains raises the question of whether this is specific to the HTTs. We found
318 that large TE expansions are common in GDL strains for both HTTs and non-telomeric TEs (Figure 3A).
319 Interestingly, in line I01, which has the greatest number of HTTs, *Copia* and *MDG3* (both LTR
320 retrotransposons) have also considerably increased in copy number (Figure 3A, arrow heads; Table S2).
321 Total TE copy number is also highest in strain I01, but across all lines there is no strong general
322 tendency for strains with non-HTT copy number expansions to have longer telomeres. Three non-HTT
323 TE families, *Gypsy1, Zam*, and *Gypsy*, that were defined as inactive under our criteria (see Materials and
324 Methods) are present in few copies in most strains, averaging 2, 3, and 4 copies, but expanded to 20, 37,
325 and 78 copies, respectively, in single strains. For the most part, the copy number expansions all occurred
326 in different strains, and we observe no general relationship between the occurrence of TE expansions
327 and HTT copy number, with the exception of the aforementioned strain I01. It is noteworthy that in this
328 strain, all of the *HeT-A* subfamilies and *TAHRE* have expanded considerably in copy number, but not the
329 *TART* family elements, suggesting differential responsiveness among the HTTs to whatever caused the
330 expansion. The correlation between *HeT-A* elements and *TAHRE* is itself consistent with the proposal
331 that *TAHRE* is the autonomous regulator of *HeT-A* (Abad et al. 2004), such that both families are
332 coregulated. We suspect that most or all of the expansions observed occurred during propagation in the
333 laboratory, because in nature the new insertions resulting from such an expansion would be at low
334 population frequency and rapidly become unlinked from each other due to independent assortment and
335 recombination. In contrast, new insertions can more easily reach high population frequency in the
336 laboratory due to either the small population size or the fixation of permissive variants in piRNA clusters
337 or piRNA pathway genes. We conclude that like any other active and selfish TE, HTTs are primed to take
338 over given the opportunity.
339

340 ***HTT copy number variation reflect its highly dynamic environment:*** Given that HTT abundances are
341 highly dynamic across the GDL panel, we wanted to assess whether the extent of this variation was
342 typical of active transposable elements or instead if the HTTs were more (or less) variable in copy
343 number than other TE families. Active elements provide a reasonable baseline because they are in
344 conflict with the genome via transposition. However, the tandem nature of the telomeres additionally
345 permits copy number expansions and contractions through unequal exchange, which may heighten the
346 variability in copy number. Two other TE families, the *R1* and *R2* elements, are also tandemly arrayed
347 within the multi-copy ribosomal RNA genes (rDNA) and thus reside in a dynamic environment. The *R1*
348 and *R2* elements therefore provide a comparison of the HTTs with TEs whose copy number evolves both
349 by transposition and unequal exchange but are not known to provide an essential function.
350 We sought to leverage population genetic theory for the relationship between the mean and
351 variance of TE copy number distributions to interpret the empirical relationship we observed in the GDL.
352 Charlesworth and Charlesworth (1983) highlight that the variance in TE copy number should depend only
353 on the mean and variance of the allele frequency spectrum of TE insertions and the degree of correlation
354 in the presence/absence of distinct insertions (linkage disequilibrium). The intuition is that whatever
355 dynamics and processes underlie the copy number evolution of a TE family, each occupiable locus
356 contains an insertion in some fraction of individuals in a population. The set of loci containing insertions
357 in an individual genome is sampled from this larger population of insertions, each insertion allele being
358 present with probability equal to its population frequency. Assuming that the number of *de novo*
359 insertions per generation is small relative to the number of preexisting insertions, the copy number in an
360 individual genome is then the sum of Bernoulli trials with potentially different success probabilities. If
361 these trials are independent and have identical population frequencies, TE copy number will be Poisson
362 distributed and the variance of copy number across genomes will be equal to the mean genomic copy
363 number. If they are independent but the population frequencies vary, this can be described with a
364 Poisson-binomial distribution, and the variance will be less than the mean. In reality the
365 presence/absence of insertions may not be independent of each other (termed “linkage disequilibrium”)
366 due to a variety of factors, including genetic linkage, selection, and demographic history and population

367 structure. This correlation in the presence/absence of insertions will increase the variance of TE copy
368 number. These factors are synthesized in the mean-variance relation

369 $$Var(Y) = \mu - Var(x_i)T + 4\sum_{i<j} D_{ij}$$

370 where μ is the population mean copy number, *T* is the number of occupiable loci, $Var(x_i)$ is the variance
371 in population frequency of TE insertions and D_{ij} is the linkage disequilibrium of insertions (Charlesworth
372 and Charlesworth 1983). The principal insight is that TE families whose insertions vary over a range of
373 population frequencies should display decreased copy number variation. On the contrary, those families
374 whose insertions are in linkage disequilibrium should display elevated copy number variation. This is
375 what we expect for HTTs, since their tight clustering at eight telomeric tandem arrays may lead to co-
376 transmission or loss of many insertions at once.

377 We examined the mean-variance relationship of TE copy number across families, considering two
378 different estimates of TE copy number (see Materials and Methods) which gave similar results. We found
379 that inactive TE families are generally less variable in copy number than expected relative to a Poisson
380 expectation (Figures 3B, C and S2 A,B). This is consistent with theory, as relatively older (but not
381 ancient) insertions should have more variation in their population frequencies than recent insertions. This
382 is because older insertions have had more time to change in population frequency, leading to a less
383 variable copy number distribution (due to $Var(x_i)T > 0$). In contrast, active TEs not only show greater
384 copy number variation than inactive TEs, but are generally more variable with respect to the Poisson
385 expectation. This greater-than-expected variability may reflect linkage disequilibrium or population
386 structure during the TE family's invasion. We consider continued copy number evolution in the lab stocks
387 (itself a form of population structure) a likely explanation. However, compared to other active elements,
388 HTT copy number is even more variable (Figure 3B, C, B and S2A, B), which could reflect instability of
389 the genomic regions they occupy, coupled with high linkage disequilibrium ($4\sum_{i<j} D_{ij} > 0$).

390 That is, while other TE families are dispersed across the genome, all HTT insertions are clustered
391 at eight tandem arrays, such that many insertions will be transmitted or lost together. Notably, the *R*-
392 elements also display elevated variability comparable to the HTTs (Figure 3B, C and S2 A,B). We

393 conclude that both the HTTs and *R*-elements are more variable in copy number than other active TEs.
394 This could reflect that these elements are more prone to continued CN evolution during stock
395 maintenance, but it is striking that it is those TE families known to be organized in tandem arrays that are
396 the most variable. We suggest therefore that the heightened copy number variation displayed by the
397 HTTs and R-elements reflects both the dynamic nature and the tight physical linkage of the genomic
398 environments caused by the tandem array organization of these families.
399
400 ***The high sequence diversity of HTTs is driven by active variants:*** Sequence evolution of TEs
401 determines their activity and their potential to escape from host suppression. HTTs are highly divergent
402 among species, which has been suggested to reflect continued conflict with their host genome (Lee *et al.*
403 2017; Saint-Léandre *et al.* 2019). We propose that the instability of their genomic niche might also
404 contribute to HTT sequence variation. The heightened copy number variation we observe for the HTTs
405 (and *R*-elements) ought to increase the stochasticity with which lineages go extinct, allowing sequence
406 polymorphism to fix more rapidly in them. We sought, therefore, to understand how the within-species
407 sequence diversity of the HTTs compares to that of other active TEs, especially the rDNA-restricted *R*-
408 elements.
409 For each TE family under consideration, we estimated the copy number of each of the four
410 possible alleles (A, T, G, C) from the number of reads supporting that allele at a given consensus
411 position in a particular strain (Figure 4A; Materials and Methods). We then used these allele copy
412 number estimates to compute the sequence diversity among insertions of a given family in the GDL. The
413 diversity estimates across all insertions in the GDL concord well with those reported from analysis of the
414 ISO-1 reference genome (Kendall's tau = 0.59, p = 6.6e-5) (Figure S2C; Kaminker *et al.* 2002). We find
415 that HTT subfamilies are relatively diverse in their sequence composition, though not obviously more so
416 than other active TE families (Figure 4B). Such sequence diversity could reflect heterogeneity among the
417 transposition-competent elements. Alternatively, the active elements might be relatively homogeneous,
418 but distinct in sequence from older inactive insertions that are relics of prior invasions. This alternative
419 scenario applies to TEs such as *I*-element and *Hobo*, which have invaded *D. melanogaster* multiple

420 times and left both recently active and degenerate copies from past invasions (Bucheton *et al.* 1992;
421 Boussy and Itoh 2004). The shape of a phylogeny from complete sequences of all TEs in the GDL would
422 easily distinguish these scenarios (Khan *et al.* 2006). However, we cannot reconstruct individual
423 elements, much less their phylogeny, with short reads. To circumvent this challenge, we developed an
424 approach to leverage our observation that the copy number variation of active TE families tends to be
425 overdispersed and that of inactive families tends to be underdispersed (Figure 3B, C). We reasoned that
426 high diversity at positions where both the major and the most abundant minor alleles display more
427 variable copy number than expected (see previous section) is suggestive of heterogeneity among the
428 active elements of the family. On the other hand, if the major allele is more variable but the minor allele is
429 less variable than expected, this instead suggests that the active elements are homogeneous at that
430 position, with a distinct population of inactive copies that are divergent from the active lineage (Figure
431 4A). This inference is admittedly indirect, and could be obscured by the difficulties of mapping reads from
432 highly diverse TE families. Our focus on recently active TE families, however, should ameliorate this
433 concern, and the concordance between our sequence diversity estimates for these recently active
434 families and those obtained from prior analysis (Figure S2C) indicates that these are not highly diverse
435 TE families and that our estimates are not strongly biased.
436 Breaking diversity down into the proportion driven by alleles that vary much in copy number
437 (overdispersed) and those that do not (underdispersed), we find that in HTTs, much of the sequence
438 diversity comes from positions with overdispersed minor alleles, while in most other active TE families
439 the diversity largely corresponds instead to positions with underdispersed minor alleles, suggestive of
440 variants in older insertions. (Figure 4B, C). Consistent with a population of heterogeneous active
441 elements, when we consider the mean-variance relations of the major and minor alleles, we find that in
442 the HTTs both generally display the same mean-variance relationship we observed for HTT copy number
443 and tend to be more variable compared to other TEs, as is seen, for example, in *HeT-A* (Figure 3D,
444 Supplemental File 2). In contrast, for most of the other TE families with high sequence diversity including
445 *Hobo* and *I*-element, the variance of the minor allele tends to be shifted down from that of the major allele
446 and is generally less variable, as is seen in *Hobo* (Figure 4E; similar plots for the other TEs in Figure 4B

447 can be found in Supplemental File 2). We suggest that this pattern of sequence diversity of non-HTTs is
448 largely driven by degenerate inactive copies, with only a small fraction of positions in most elements
449 reflecting divergence among active elements. This is consistent with the known invasion history of *Hobo*
450 and *I*-element (Bucheton *et al.* 1992; Boussy and Itoh 2004). *TAHRE* is one exception to the tendency of
451 the HTTs to show mainly highly variable minor alleles, as it appears to have a subset of less variable
452 minor alleles (Figure 4F), suggesting the presence of inactive copies outside of the telomere. We
453 subsequently confirmed the existence of nearly fixed non-telomeric *TAHRE* elements harboring 214/249
454 of these underdispersed alleles in unmapped heterochromatin (see "*HTT transposition is restricted to*
455 *the telomeres*" below), providing additional validation that, despite its indirectness, underdispersion of
456 allele copy number can indeed be indicative of alleles in older insertions. To investigate whether this
457 heightened diversity observed in the HTTs is characteristic of TEs residing within unstable genomic
458 regions, we considered the sequence diversity of the *R*-elements. However, while *R1* also shows high
459 levels of diversity (Figure 4B), the presence of hundreds of tandemly arrayed *R1*-elements per genome
460 evolving independently of the active elements means we cannot distinguish whether this diversity is
461 among the active elements or involves those comprising the tandem array (Roiha and Glover 1981;
462 McGurk and Barbash 2018). A cleaner comparison is with *R2*, which does not form tandem arrays.
463 Unlike the HTTs, *R2* displays considerable variation consistent with inactive and degenerate copies,
464 which likely reflects accretion from the edge of the rDNA array. This accretion of degenerate and
465 fragmented copies of repeat units is commonly observed at the edge of satellite arrays (McAllister and
466 Werren 1999). The absence of accreted HTT relics may simply reflect the action of telomere erosion,
467 which readily deletes telomeric sequence, including sequence accreted into the subtelomere. *R2*
468 displays lower levels of sequence diversity among active copies (Figure 4B, Supplementary File 3),
469 which may indicate that even for TEs within unstable niches, the HTTs are evolving rapidly, but this is a
470 limited comparison because we are comparing to only one element. However, we can more confidently
471 conclude that the high sequence diversity of HTTs is driven by active variants whose appearance and
472 maintenance is favored by the unstable genomic niche they occupy.
473

474 ***HTT transposition is restricted to the telomeres:*** While HTT transposition has been apparently
475 restricted to the telomeres for millions of years, the retention of intact endonuclease ORFs in *TART* and
476 *TAHRE* suggests that HTTs could potentially transpose elsewhere in the genome (Casacuberta 2017).
477 Transposition of HTTs outside the telomere might reflect selfish HTT behavior that is in conflict with the
478 host (Saint-Léandre *et al.* 2019). Non-telomeric *HeT-A* family and *TART* sequences are found on the
479 heterochromatic Y chromosome of *D. melanogaster* and other Drosophila (Danilevskaya *et al.* 1993;
480 Losada *et al.* 1997; Agudo *et al.* 1999; Berloco *et al.* 2005). However, Berloco et al. (2005) noted that
481 "these non-telomeric tandem repeats differ from the telomeric elements in at least two respects: they are
482 interspersed with other repetitive sequences and they are not always oriented in the same direction" and
483 postulated that chromosomal rearrangements rather than non-telomeric insertions were responsible.
484 These structural differences suggest that non-telomeric HTT copies may result from gene conversion,
485 capture by other TEs, or rearrangements, rather than by non-telomeric insertion (Danilevskaya *et al.*
486 1993; Agudo *et al.* 1999).
487 Outside of the Y chromosome, there are two non-telomeric regions of the mapped *D.*
488 *melanogaster* reference genome that contain sequence with >80% homology to the HTT consensus
489 sequences, both on 3L: a ~700 bp stretch between 16,576,994-16,577,693 with ~95% identity to a
490 fragment of the *TART-A* terminal repeat and a 15 kb stretch between 25,216,958-25,230,409 containing
491 fragments with lower identity (<85%) to *HeT-A*, *TAHRE*, and *TART-A* sequence. Their short lengths
492 (several hundred base pairs or smaller) and adjacency with other HTT sequence suggest that they either
493 arose from recombination-mediated transfer of telomeric sequence (Agudo et al 1999) or represent
494 ancient insertions that have largely degraded. Non-telomeric insertions were recently reported as
495 common, however, in the *melanogaster* subgroup species *D. rhopaloa* (Saint-Léandre *et al.* 2019). We
496 therefore used ConTExt to search the *D. melanogaster* GDL for HTT insertions in assembled regions as
497 well as in other repeats. We identified 34 junctions between HTT 3'-ends and subtelomeric unique
498 sequence, these reflect healed terminal deficiencies and are discussed separately below. Beyond the
499 subtelomere, we only found junctions supporting seven loci in the mapped chromosomes potentially
500 harboring non-telomeric HTT sequence. Both 3' and 5' junctions of the aforementioned *TART-A* terminal

501 repeat fragment at ~16,576,000 on 2L were identifiable, with at least one of the junctions detectable in
502 82/85 strains. The only other non-telomeric HTT with both 3' and 5' junctions detectable was a short
503 *HeT-A* fragment spanning 5,276 - 5,574 of the consensus at position ~7,537,400 on 3L in strain I29. The
504 remaining five loci were supported by only a single junction with a *HeT-A* subfamily, and the only one
505 present in more than one strain (3/85) falls at the edge of the previously mentioned stretch of 3L
506 harboring degenerate *HeT-A* fragments (25,229,995). Furthermore, none had both 5' and intact 3'
507 junctions; it is therefore challenging to distinguish these loci from sequencing artifacts. Assuming that
508 they are genuine, it is likely that these fragments arrived outside the telomere due to a recombination-
509 based mechanism, though two of the seven non-telomeric HTTs may have intact 3'-ends and thus could
510 represent rare non-telomeric insertions. We conclude that non-telomeric insertions of HTT elements are
511 very rare or negligible in *D. melanogaster.*

512 For perspective, we searched for *R*-element insertions outside the rDNA. While domestication of
513 the HTTs could keep them constrained to the telomeres, there are unlikely to be similar constraints on
514 the *R*-elements because they serve no obvious function. Rather, their restricted localization to the rDNA
515 is probably a property purely intrinsic to the *R*-element transposition machinery. Thus, the extent to which
516 we detect *R*-element insertions outside of their niche provides both a positive control validating our
517 approach and insight into the extent to which niche specificity can be maintained by TEs alone without
518 external control by the host. We identified no instances of *R2* elements inserted outside of rDNA units in
519 the GDL data. By contrast, we found 33 insertions of *R1* outside of rDNA units where both the intact 3'-
520 end and (often truncated) 5' junctions were identifiable. Most such junctions were restricted to single
521 strains, though one insertion in the 3L pericentric heterochromatin was present in 70/85 strains,
522 suggesting it occurred more distantly in the past. Overall, while this corresponds to only 0.34% of all *R1*
523 elements in the GDL, it nonetheless reflects more insertion site promiscuity outside of its typical niche
524 than we observe for the HTTs. On the whole, though, the tight restriction of the *R*-elements and HTTs to
525 their respective niches suggests it is largely a property of the elements themselves and that host control
526 may not be required to maintain highly specific insertion sites.

527 However, while we did not identify clear evidence of non-telomeric HTT insertions in well-mapped
528 regions of chromosomes, several of our observations suggest there may be copies of *TAHRE* outside
529 the telomere but within unmapped heterochromatic regions. First, a subset of *TAHRE* SNPs vary less in
530 copy number than expected of alleles found in active elements, suggesting the possibility of inactive
531 copies outside the telomeres (Figure 4F). Second, we find that few strains lack full-length *TAHRE*
532 insertions in contrast to the other HTTs (Figure 2C), suggesting there may be tandem *TAHRE* copies that
533 are rarely lost. Finally, we observe an *R1* inserted into *TAHRE* sequence which has reached moderate
534 population frequency (12/15) among Beijing lines. Given the instability of the telomeres, we think it
535 unlikely that such structures could be present in so many genomes if they involved telomeric *TAHRE*.
536 We therefore searched for non-telomeric *TAHRE* elements in unmapped heterochromatic
537 sequences missing from the reference genome. As the unmapped contigs in the reference genome are
538 too short to determine whether *TAHRE*-containing contigs are surrounded by telomeric repeats or
539 instead embedded in non-telomeric repeats, we searched for heterochromatic *TAHRE* insertions in the
540 PacBio long-read assembly (GCA_000778455.1) of the reference strain ISO1 (Berlin *et al.* 2015). We
541 identified one 140kb scaffold (JSAE01000744) which contains three tandem *TAHRE* insertions
542 embedded among many other TEs. These are highly similar to each other, though 5% divergent from the
543 *TAHRE* consensus. Both the 3'- and 5'-ends of two insertions are evident and they form tandem
544 junctions. These *TAHREs* likely correspond to the full-length *TAHRE* that we found above in most
545 strains. A majority (214/249; 86%) of the underdispersed *TAHRE* minor alleles (which led us to search
546 for non-telomeric *TAHRE*s) correspond to polymorphisms present in these heterochromatic elements
547 (Figure 4F), indicating that these are also the cause of the underdispersed alleles and providing
548 validation for our inference that underdispersion reflects older insertions. Some polymorphisms within
549 these non-telomeric copies are present in more than three copies and display overdispersed variation
550 more consistent with active elements, suggesting that some extant active *TAHRE*s are closely related to
551 the lineage that gave rise to these tandem *TAHRE*. Both the *gag* and *Pol* ORFs are interrupted by
552 several nested TE insertions that are present in most (82/85) strains. Tandem TEs indeed can form by
553 multiple, independent insertions at the same locus, but generally constitute a minority of total insertions

554 and often form at strong consensus insertion sites (McGurk and Barbash 2018). If these tandem *TAHRE*
555 elements reflect true non-telomeric insertions, at least two independent *TAHRE* insertions must have
556 occurred at this locus. Given that we observed no single insertions outside of the telomeres, we consider
557 double insertion outside of the telomere to be unlikely (although it is possible that non-telomeric *TAHRE*
558 insertions occurred more frequently in the past). We instead propose that these *TAHRE* elements
559 originated as insertions within heterochromatin-adjacent telomeres, XR or 4L being likely candidates, and
560 were subsequently pushed into the adjacent heterochromatin by accretion from the edge of the array.
561
562 ***Terminal deficiencies are common:*** The high degree of copy number variability displayed by the HTTs
563 appears to be shaped by the dynamic nature of the telomere. We sought, therefore, to further
564 characterize the stability of the telomeres. Beyond the action of unequal exchange between telomeres to
565 amplify and delete sequence, there are other possible mechanisms by which telomeric sequence may
566 turn over. First, the end replication problem directly causes telomere erosion, which in some cases may
567 reduce the chromosome ends to the subtelomeric TAS arrays or even into centromere-proximal unique
568 sequence. Second, DNA double-strand breaks in the subtelomere may result in sudden terminal
569 deficiencies.
570 We consider three lines of evidence to identify deficiencies. First, to identify the subset of terminal
571 deficiencies that extend through the HTT array and subtelomeric repeats into unique sequence, we
572 searched for a sharp loss of (homozygous) or reduction in (heterozygous) read depth over telomere-
573 proximal unique sequence. Second, deficiencies which have been healed by new HTT insertions can be
574 detected as junctions between the 3'-ends of HTT elements and subtelomeric unique sequence
575 (Biessmann *et al.* 1990b). Such insertions onto the end of a broken chromosome can be readily
576 distinguished from chromosome-internal insertions, which would produce junctions with unique sequence
577 involving both the 3' and 5' ends of the HTT. Considering any subtelomere that met either of these
578 criteria, we found evidence for 56 terminal deficiencies in 48/85 strains (Figure 5A, Supplemental File 3).
579 As a third and independent test, we looked for concomitant deletion of the subtelomeric satellite. The left
580 arms of chromosomes 2 and 3 harbor arrays of the same subtelomeric satellite, TAS-L. If we have

581 identified true terminal deficiencies of 3L, the TAS array should be lost from those chromosomes and
582 TAS-L copy number reduced. Consistent with true terminal deficiencies on 3L, the TAS-L copy number is
583 roughly half that of strains without detectable deficiencies in strains with HTT-3L junctions (Figure 5B).
584 Of the 56 terminal deficiencies, 14 display evidence of heterozygosity, suggesting these are not
585 fixed in the fly stocks and perhaps arose after collection. The remaining 42/56 deficiencies are all
586 consistent with homozygous deficiencies, as expected in inbred strains. Of the 56 deficiencies, 34
587 showed evidence of an HTT-subtelomere junction, indicating that the lost telomere has been replaced by
588 a new HTT array. All but five of those 34 deficiencies extend sufficiently far into unique sequence that we
589 could confirm a clear drop in coverage after the HTT-subtelomere junction. For 22/56 deficiencies,
590 however, we observe only a loss of subtelomeric unique sequence based on read depth but no evidence
591 of repair in the form of an HTT junction. Some of these may reflect our failure to detect a junction that is
592 truly present, for example due to variation in read counts or repeat-masked sequence (note that five
593 putatively unhealed deficiencies sit at the edge of masked sequence, Figure 5A middle panel). However,
594 total telomere length is generally lower in these strains with putatively unhealed deficiencies compared to
595 strains where a HTT junction is identifiable (Figure 5C), consistent with truly missing a HTT-based
596 telomere. This suggests that a remarkably high proportion of telomere deficiencies are not healed by
597 HTTs.
598 At least 39 of the deficiency breakpoints are independently derived, involving different sequence
599 coordinates (Figure S2D). Most (44/56) are found on 3L and often delete the most distal gene, *mthl8,*
600 with a few also deleting the next annotated gene *CG43149*; a similar high rate of 3L terminal deficiencies
601 was discovered by Kern and Begun (2008). We note that the density of deficiencies is twice as high
602 distal to *mthl8* than it is between *mthl8* and *CG43149,* suggesting that deletions of *mthl8* nonetheless
603 have some deleterious effect even if not as serious as loss of *CG43149* (Figure 5A, middle panel). Only
604 9 and 3 deficiencies are present on 2R and 3R, respectively, but the distribution of these are also limited
605 by genes within several kilobases of the telomere, with few extending beyond *CG30429* on 2R and
606 beyond *Map205* on 3R. We observed no deficiencies on 2L, likely because there is only 400 bp of non-
607 repetitive sequence between the gene *lethal (2) giant larvae* and the TAS array. Mutations in *l(2)gl* are

608 homozygous lethal; therefore terminal deletions removing *l(2)gl* would not be expected in the GDL lines
609 we sampled because they have been inbred to homozygosity, but they are common on heterozygous
610 chromosomes in natural populations and in heterozygous lab stocks (Golubovsky 1978; Green and
611 Shepherd 1979; Mechler *et al.* 1985; Roegiers *et al.* 2009). Taken together, our analysis reveals that
612 terminal deletions occur at high frequency on all chromosomes, and also highlights the fitness impacts of
613 terminal erosion and deletion.
614
615 ***Patterns of HTT insertions suggest different families are active under different conditions*:** The
616 presence of distinct families and divergent subfamilies of TEs is a common feature of Drosophila
617 telomeres and opens up the potential for specialization of elements and differential regulation by the host
618 genome. We investigated two possible cases. First, we hypothesized that HTT specialization could affect
619 patterns of HTT interspersion. For example, competition among elements that is resolved by limiting their
620 expression to different cell types or development stages would result in finding adjacent copies less
621 frequently than expected. Conversely, cooperation among elements that requires their co-expression for
622 successful transposition would result in an increased frequency of interspersed insertions. We found that
623 elements of the same subfamily tend to neighbor each other, suggesting that they tend to be reverse
624 transcribed multiple times at a given chromosome end (Figure 6A, B). We also observed that elements of
625 the *TAHRE* and *HeT-A* families are generally interspersed as would be expected under random ordering.
626 *TART* family elements, however, are less commonly found adjacent to *HeT-A* family and *TAHRE*
627 insertions than expected. Further, the tendency of *TART* family elements to be found adjacent to their
628 own subfamily is much stronger than that of *HeT-A* subfamilies and *TAHRE*. Clustering of *TART* family
629 elements has been previously detected (George *et al.* 2006). The interspersion of *HeT-A* family and
630 *TAHRE* elements contrasted with the low frequency of junctions of either of them with *TART* family
631 elements may be a consequence of the *HeT-A* family relying upon the expression of *TAHRE*'s
632 polymerase for reverse transcription, implying that the *HeT-A* family and *TAHRE* mobilize together (Shpiz
633 *et al.* 2007). It further suggests that *TAHRE* and the *TART* family may not be active at the same time. To
634 confirm these observations, we examined the assembled telomeres in the release 6 reference genome

635 (Hoskins *et al.* 2015), and observed several instances where multiple full-length copies of the same
636 element are in tandem (Figure 6C). We conclude that the pattern of HTT arrays are not random, and that
637 this likely reflects the dependence of the *HeT-A* family on *TAHRE's* retrotransposition activity as well as
638 possible specialization of the *TART* family versus *TAHRE*. The three *TART-A* elements in tandem on the
639 X chromosome telomere, however, share terminal repeats, which is consistent with recombination
640 between the 3' and 5' terminal repeats (Ke and Voytas 1997) and the principal mechanism by which TEs
641 with terminal repeats generate tandems in Drosophila (McGurk and Barbash 2018). Thus the higher rate
642 of self-adjacency observed for *TART* family elements may reflect recombination between termini as
643 previously suggested by George et al. (2006).

644 We also considered the possibility that certain families are more involved in the recovery of lost
645 telomeres than others, which would result in an enrichment near the base of the telomere. On the
646 autosomes, tandemly repeated telomere associated sequences (TAS-L and TAS-R) sit at the base of
647 each telomere. An HTT insertion adjacent to a TAS-L or TAS-R repeat therefore represents the first
648 insertion after the HTT array eroded completely. Similarly, if an HTT has healed a terminal deficiency
649 extending beyond the TAS array, it will form a junction with unique subtelomeric sequence (as described
650 above; see also Kern and Begun (2008)). In both cases, if an HTT family is particularly activated upon
651 the complete erosion of a telomere, it will be enriched next to these subtelomeric sequences relative to
652 the other HTT families. We find that *HeT-A2* and *TART* family elements are underrepresented and
653 enriched, respectively, at the telomere base. The most striking result though is that *TAHRE* is 4-fold
654 underrepresented at the base (Figure 6D). Since *HeT-A* and *TAHRE* likely share the same transposition
655 machinery, one would expect them to show the same pattern. Our results therefore suggest the
656 possibility that *HeT-A* elements are more proficient at using *TAHRE's* transpositional machinery to heal
657 telomeres than *TAHRE* itself. Regardless, our data suggest that the HTTs differ in their propensity to
658 heal completely eroded telomeres, and some (particularly the *TART* family) may be especially activated
659 by telomere loss.

660 **DISCUSSION**

661 Linear chromosomes require a specialized mechanism to replicate and maintain their ends, and
662 most eukaryotes use the highly conserved telomerase enzyme to do so. Yet Drosophila and a few other
663 groups have dispensed with telomerase and instead use the HTT retrotransposons to perform this
664 essential function, highlighting the remarkable flexibility of evolution (Pardue and DeBaryshe 2008).
665 Extensive efforts have provided full assemblies of several telomeres in *Drosophila melanogaster* and
666 other Drosophila species (Saint-Léandre *et al.* 2019; George *et al.* 2006). Other studies have hinted that
667 there may be extensive variation in telomere length and structure within populations (Kern and Begun
668 2008; Wei *et al.* 2017), but in the absence of a large-scale population genomics analysis, it remains
669 difficult to distinguish among very different evolutionary models that could govern telomere dynamics.

670 While long-read sequencing technologies are beginning to provide glimpses into structural variation
671 of complex repetitive regions (Chang *et al.* 2019; Weissensteiner *et al.* 2017), the vast majority of
672 population genomics data comes from short-read Illumina sequencing. Here, we leveraged the ConTExt
673 pipeline and available population genomic data to investigate HTT variation in *D. melanogaster* at an
674 unprecedented scale, examining 85 lines derived from five world-wide populations and encompassing
675 close to 10,000 HTT insertions.

676

677 ***Caveats and technical considerations***

678 There are some limitations resulting from the data we used that are derived from pools of ~50
679 inbred whole females of varying ages. First, the GDL lines have continued to evolve under inbreeding
680 and laboratory evolution, which could result in increased variance of TE copy number, and fixation or
681 loss of certain TE variants. However, this also would apply to most other studies of Drosophila telomeres
682 in which strains have not been freshly collected from the wild before sequencing, including reference
683 genome assembly (Saint-Léandre *et al.* 2019; George *et al.* 2006). Although not necessarily reflecting
684 natural variation, inbreeding and continued lab evolution may provide increased power to detect TEs
685 escaping repression or differences in their susceptibility to repression. Second, telomere length might
686 differ across tissues and over an individual's lifespan. However, this is unlikely to affect our estimates as

687 1) these differences should be specific to individual flies and 2) germline erosion is reported to be
688 approximately 70 bp per generation in Drosophila (Levis 1989; Biessmann *et al.* 1990a), much lower
689 than the variance we report here. Some of our inferences about rates and patterns of telomere variation
690 could in principle be extended using a mutation accumulation framework, although the relatively low per-
691 generation rates would require very large sample sizes. Finally, a powerful, albeit extremely costly, way
692 to further probe variation in telomere structure would be to directly assemble full telomeres from wild
693 populations, as Saint-Léandre *et al*. (2019) have reported from single strains of multiple Drosophila
694 species using long-read sequencing.

695 The ConTExt method infers the presence and arrangement of TEs from paired-end Illumina data.
696 We validated ConTExt here by examining five genomes for which both Illumina and PacBio assembly
697 data are available (Long *et al.* 2018), and simulating Illumina short reads from the PacBio assembly. We
698 found a high concordance for HTT copy number estimates between the simulated and actual ConTExt
699 analyses (Figure 1C).

700 We have further confidence in our analyses based on the concordance of our results with a wide
701 range of previous studies. For example, the large fraction of truncated HTTs we observed is consistent
702 with the assembled telomeres from the *D. melanogaster* reference strain (George *et al.* 2006). In
703 addition, the high prevalence of tandem HTTs is similar between the two studies. For sequence diversity,
704 our estimates for TE families across the 85 GDL concord with estimates reported from analyses of the
705 ISO-1 reference strain (Kaminker *et al.* 2002). We also identified a large number of terminal deficiencies,
706 consistent with a more limited previous study (Kern and Begun 2008). When examining copy number, we
707 found a strong correlation between the number of *HeT-A* family and *TAHRE* elements in outlier strains,
708 which supports prior suggestions that *TAHRE* is the autonomous regulator of HTT transposition (Abad *et*
709 *al.* 2004). Finally, our identification of *R*-elements inserted outside of the rDNA agrees with previous
710 studies (Stage and Eickbush 2009).
711

712 ***Frameworks for considering HTT evolution***

713 HTTs are often considered to be a canonical example of transposon domestication, perhaps
714 reflecting an adaptive response to ameliorate a reduction in telomerase activity (Mason *et al.* 2008;
715 Pardue and DeBaryshe 2011; Arkhipova 2012; Shpiz and Kalmykov 2012; Servant and Deininger 2015).
716 Others though have noted the rapid evolution of HTTs and telomere-associated proteins and proposed
717 that HTTs remain in genetic conflict with their hosts (Saint-Léandre *et al.* 2019; Lee *et al.* 2017; Cosby *et*
718 *al.* 2019; Markova *et al.* 2020). Markova et al. (2020) have also proposed recently that telomeric
719 localization of HTTs might reflect a strategy of site-specific integration to minimize deleterious effects on
720 the host, rather than domestication. They also noted that the HTTs are likely to maintain the ability to
721 cause genetic conflict by evolving through new mutations or gene conversion. We suggest that a fourth
722 factor, niche instability, is an additional and under recognized factor that influences HTT evolution. We
723 outline and examine our results below in the context of these four distinct processes that may affect
724 telomere dynamics in Drosophila.

725

726 ***Evidence for domestication or host selection***

727 The pioneering essays on selfish and junk DNA suggested that repetitive DNAs might occasionally
728 be co-opted by their hosts to serve essential functions (Doolittle and Sapienza 1980; Orgel and Crick
729 1980). Active TEs will experience continued selection for their own replicative capacity and so complete
730 domestication likely requires ablating transpositional activity (Jangam *et al.* 2017). Consistent with this,
731 most described examples of domesticated TEs have lost the ability to transpose. In contrast, the HTTs
732 have not. The extent of HTT domestication is a challenging question to address with population genomic
733 data, as it likely relates to the ways in which their activity is regulated by the host genome, how this
734 regulation is distinct from that of other TEs, and how dependent they have become upon telomere-
735 specific host factors. The elaborate system of HTT control also speaks to host domestication. Like other
736 TEs, the HTTs are controlled by the piRNA pathway, though unlike most TEs it is the active elements
737 themselves that act as the piRNA clusters (Shpiz and Kalmykova 2012). This potentially makes it
738 challenging for the host to exert stable control of telomeres since the piRNA source is at an inherently

739 unstable location. HTTs outside of the telomere could thus indicate evidence of host control. We suggest
740 that the heterochromatic, likely inactive, *TAHRE* tandem array we discovered in 82/85 strains could be a
741 locus of host control, functioning as either a source of *TAHRE* piRNAs or producing dominant-negative
742 *TAHRE* proteins.

743 Full domestication of HTTs would imply that the HTT transposition rate does not increase beyond a
744 point where telomere length becomes excessive for host fitness. Although there is clearly extensive
745 variation in telomere length, there are only limited data on whether and how telomere length may affect
746 fitness (Walter *et al.* 2007). While our analysis thus does not provide direct evidence for domestication,
747 the lack of full domestication would cause genetic conflict, which we discuss next.

748

749 ***Evidence for genetics conflict***

750 Typical TEs exist in conflict with the host genome, as they gain a replicative advantage through
751 transposition but also deleteriously impact host fitness by disrupting functional sequence, facilitating
752 ectopic recombination, and creating intrachromosomal breaks (Kelleher *et al.* 2020). HTTs insert
753 specifically at the ends of chromosomes and so avoid causing these harmful impacts. If long telomeres
754 have no negative fitness impacts, then adoption of this insertional strategy may have largely resolved the
755 typical TE-host conflict. On the other hand, if excessive elongation causes decreased host fitness,
756 conflict of two sorts arises: 1) conflict with the host genome and 2) conflict among the HTTs competing
757 for limited insertion sites.

758 First, while there should exist an optimal rate of telomere elongation for chromosome integrity,
759 higher rates of transposition may be optimal for the HTTs as they are for TEs in general, with faster
760 replicating subfamilies replacing slower ones (Charlesworth and Langley 1986). This conflict with the
761 host should skew telomeres toward longer-than-optimal lengths and require the evolution of host
762 suppression to shift the rate of elongation closer to the host's optimum. The optimal average telomere
763 length for *D. melanogaster* is unknown, so this potential skew cannot be detected from population data.
764 However, we observed several strains with exceptionally long telomeres, the longest of which also
765 exhibited extreme CN expansions of several non-telomeric TEs, suggesting heterochromatin

766 maintenance or piRNA pathway defects. The extent of telomere elongation in these strains highlights the
767 degree to which host suppression limits HTT replication, and the extent to which HTTs expand in its
768 absence, suggestive of conflict. As noted above though, these relatively rare strains with significant TE
769 expansions might reflect laboratory evolution of the strains rather than true natural variation.

770 Second, there is potential conflict among HTTs, as they are all restricted to the same genomic
771 region. This should favor the emergence of variants with higher rates of transposition, displacing less
772 active variants. While we found some differences in the proportion of HTT subfamilies between long and
773 short telomere strains, the three major families are equally represented, suggesting that longer telomeres
774 result from loss of host control rather than runaway expansion of one family at the expense of others. We
775 therefore do not have evidence of strong genetic conflict among HTTs.

776 Conflict with either the host or other HTTs should lead to the fixation of new sequence variants
777 within the HTTs, and ongoing conflict may result in substantial diversification. It is tempting, therefore, to
778 infer conflict from the number of distinct HTT subfamilies and the sequence diversity within them.
779 However, our observation that the copy number variation of most HTT SNPs is consistent with recently
780 active elements, in contrast to typical TEs where most SNPs are consistent with older insertions,
781 suggests that the rapid turnover of the telomeres has a major influence on the patterns of HTT
782 diversification. This distinctly high rate of sequence loss must be accounted for when trying to infer
783 evolutionary forces influencing HTT evolution (see below).

784

785 ***Evidence for niche specialization of HTTs and tradeoffs***

786 One route by which TEs might attenuate conflict with the genome is by limiting their deleterious
787 impacts, for example by inserting solely at gene-poor loci. Both the HTTs and R-elements appear to have
788 adopted such strategies, inserting at gene-poor but unstable loci, thereby trading long-term insertion
789 sites for reduced host fitness costs (Markova *et al.* 2020). TE subfamilies could conceivably reacquire a
790 more general insertion pattern to escape both the instability and competition with the other TEs targeting
791 the same limited set of occupiable sites. We observed no clear evidence of HTT insertions outside of the
792 telomere, but identified 33 strain-specific *R1*-element insertions outside of the rDNA. Multispecies

793 surveys of *R*-elements and HTTs have provided similar observations, with low to moderate fractions of
794 *R*-element being inserted outside of the rDNA in most species (Stage and Eickbush 2007) but few, if any,
795 non-telomeric HTT insertions (with the possible exception of *D. rhopaloa*, Saint-Léandre *et al.* 2019).
796 This could reflect domestication locking the HTTs more tightly into their insertion site than *R*-elements,
797 perhaps through reliance on telomere proteins. But a neutral explanation based on TE
798 mobilization/transposition mechanism is also conceivable. *R*-elements have a functional endonuclease
799 which is required to insert in the rDNA. *R*-elements therefore only need to relax the sequence specificity
800 of their endonuclease to escape the rDNA. In contrast, it is not clear if the HTTs depend on
801 endonuclease to insert at the chromosome ends, even if *TART* and *TAHRE* elements retain a potentially
802 functional endonuclease domain (Morrish *et al.* 2007; Arkhipova 2012). If HTTs do not utilize an
803 endonuclease as part of their normal lifecycle, then escape from the telomeres likely requires
804 reacquisition of a functional endonuclease, thus locking telomeric TEs more tightly into their insertion site
805 than the *R*-elements.
806 We observed that HTTs have a propensity to form tandem insertions of the same subfamily, which
807 might reflect another aspect of niche specialization that mitigates some of the fitness cost of being
808 exposed to terminal erosion. *Het-A* insertions are known to include short tags of sequence derived from
809 the upstream element whose 3'-end-promoter initiated transcription, which George *et al*. (2010) suggest
810 provides a buffer to protect the element from erosion. We note that the tandems we describe are 1)
811 generally between elements of the same *HeT-A* subfamily and 2) the tandems present in the reference
812 genome include arrays of full-length elements (Fig 6C), rather than a single full-length element preceded
813 by short fragments. We suggest, therefore, that these tandems reflect sequential insertion of the same
814 subfamily to the chromosome end. The most terminal HTT on a telomere is at imminent risk of loss by
815 erosion. An HTT that generates two or three tandem copies at the same chromosome end exposes the
816 distal-most insertion to immediate erosion but buffers the more proximal elements for potentially
817 hundreds of generations. By contrast, creating two independent copies on two different telomeres
818 generates the same number of new insertions, but both are subject to immediate erosion. For the *HeT-A*
819 family and *TAHRE*, this is potentially an evolved strategy of niche specialization, because they do not

820 have an inherent mechanism such as internal repeats to catalyze tandem formation. *TART* family
821 elements do, however, so their occurrence in tandem is more likely to reflect a mutational process, as we
822 discuss below.
823
824 ***Evidence for mutational processes and genome instability driving HTT turnover***
825 The erosion and replacement of sequence is a fundamental characteristic of telomeres, and
826 occupying such an unstable locus thus distinguishes the HTTs from most other TE families. Our
827 observations both clarify the extent of this instability and highlight its impact on HTT evolution. First, the
828 presence of terminal repeats in *TART* family elements predisposes them to tandem expansions by
829 ectopic recombination, similar to LTR retrotransposons. The high rate of tandem *TART* elements with
830 shared terminal repeats we observed is evidence of recombination-mediated changes in telomere length.
831 This extends upon prior observations of *TART* family tandems in the ISO-1 reference genome (George *et*
832 *al.* 2006), and highlights that not at all gains and loss of sequence result from transposition and erosion.
833 Second, we discovered a high prevalence of terminal deletions in the GDL, extending previous
834 discoveries (Levis 1989; Walter *et al.* 1995; Kern and Begun 2008). Surprisingly, we also discovered that
835 about one third of terminal deletions are unhealed by an HTT. It is possible that unhealed deletions are
836 more common in inbred lab stocks; for example, if inbreeding allows alleles to fix that reduce HTT
837 activity. Regardless though, most telomeres are presumably eventually healed over evolutionary time,
838 and this creates opportunities for new HTT subfamilies to expand.
839 Third and most generally, we find that for positions within HTT sequences that display multiple
840 alleles, both alleles tend to exhibit high copy number variation consistent with being recently active
841 elements. This contrasts sharply with other active TEs for which only one allele generally appears to be
842 recently active while the other typically shows low copy number variation, reflective of being older and
843 likely inactive. We suggest this results from HTT insertions being lost much more rapidly in the telomeres
844 than other TEs elsewhere in the genome, where the relics of past transposition may be preserved for
845 millions of years. In the telomeres, the window into the past is much more narrow, due to continuous
846 erosion. Therefore, the diversity of HTT families and subfamilies does not necessarily demonstrate rapid

847 evolution for HTTs, but may instead simply reflect that each species is recreating its entire population of
848 HTT insertions all the time as erosion and recombination erase older insertions. This will be reflected in
849 their phylogenetic pattern, as the HTTs show common ancestors with long terminal branches (Saint-
850 Léandre *et al.* 2019; Villasante *et al.* 2007), rather than a birth/death pattern with very short terminal
851 branches such as for active primate LINE-1 elements (Khan *et al.* 2006).
852
853 ***Conclusion***
854 HTTs are often described as a clear-cut case of domestication. They clearly serve an essential host
855 function, and non-telomeric insertions appear to be negligible. But niche specialization is a plausible
856 alternative, whereby telomere localization reflects an evolutionary strategy of the HTTs to reduce their
857 fitness impact on the genome (Markova *et al.* 2020). In this interpretation, HTTs may have become the
858 dominant mechanism for forming telomeres as a by-product of evolving preferential localization to
859 telomeres. Species such as silkworm might represent a transitional stage, as they have telomere-
860 associated retrotransposons while maintaining telomerase (Okazaki *et al.* 1993, 1995; Pardue and
861 DeBaryshe 2008).
862 On the other hand, genetic conflict has been suggested to be a major driver of the observed rapid
863 evolution of HTT families and subfamilies, telomeric proteins, and subtelomeric sequences (Saint-
864 Léandre *et al.* 2019; Lee *et al.* 2017; Cosby *et al.* 2019; Saint-Léandre and Levine 2020). But direct
865 evidence for conflict is elusive, largely due to limited data on fitness consequences of telomere length
866 variation. We further suggest that the high instability of chromosome ends is an alternative explanation of
867 HTT variability. However, our results are also consistent with prior observations that host sequences
868 regulate the HTTs and further suggest that their susceptibility to regulation varies among the HTTs and is
869 actively evolving, consistent with ongoing conflict between these genomic endosymbionts and their host
870 genome. We conclude that multiple evolutionary forces and mechanistic processes interact to explain the
871 patterns of HTT variation, with telomere instability being an under-recognized contributor.

872 **MATERIALS AND METHODS**

873 ***The repeat index and nomenclature***

874 We used the manually curated repeat index described in McGurk and Barbash (2018), which
875 contains consensus sequences for the known *D. melanogaster* TE families as well as satellite repeats,
876 including the left and right telomere associated sequences (TAS-L and TAS-R). The telomeric TE
877 sequences in this index are somewhat distinct from the Repbase entries (Jurka *et al.* 2005). First, the
878 long perfect near-terminal repeats (PNTR) found in the *TART* family elements are given entries separate
879 from the internal sequences, similar to how the terminal repeats of LTR retrotransposons are given their
880 own entries separate from the internal sequence in Repbase. Second, the index includes four *HeT-A*
881 family entries in addition to the Repbase consensus (Table S1), as many of the *HeT-A* family insertions
882 in the reference genome are quite divergent from the consensus sequence (< 85% identity, Figure S4A).
883 To ensure reads derived from *HeT-A* family insertions would align to the repeat index, we extracted all
884 insertions annotated as *HeT-A* in the UCSC genome browser Repeat Masker track for the release 6 *D.*
885 *melanogaster* genome. We then performed all pairwise alignments using BLASTn and constructed a
886 graph by construing insertions as nodes and connected all pairs of sequence that shared >90% identity
887 with edges (McGurk and Barbash 2018). We partitioned this graph into communities of homologous
888 insertions using the Louvain algorithm (Blondel *et al.* 2008). This identified four *HeT-A* family
889 communities, one of which appeared to contain subcommunities. We split the large community into two
890 subcommunities by reclustering with the alignment cutoff set at 95% identity. From each community we
891 manually chose a full-length element as representative and added these to the repeat index, labelling
892 them *HeT-A1* to *HeT-A5*. However, we removed *HeT-A4* due to high sequence similarity with *Het-A3*.
893 We refer to the original *HeT-A* sequence from RepBase as "*HeT-A*" (Bao *et al.* 2015). The final set of
894 *Het-A* family sequences all have less than 85% identity to each other, with most pairs having roughly
895 80% (Figure S3A, Supplemental File 1). For TEs that have internal repeats, we removed them by
896 scanning along each consensus, and for each 70-mer in the consensus masking all subsequent 70-mers
897 within 5 mismatches.

898

899 ***Categorizing TE families as active***

900 We categorize TE families as active or inactive based on their percent identity and population
901 frequency as summarized in Kelleher and Barbash (2013). We consider a family as putatively active (or
902 recently so) if the mean pairwise identity between individual insertions in the reference genome is greater
903 than 95% (Kaminker *et al.* 2002; Kelleher and Barbash 2013) and the average population frequency of
904 insertions is less than 0.4 (Kofler *et al.* 2012). We additionally consider the telomeric TEs and the two *R*-
905 elements as being active, as well as including *P*-element, *Hobo*, and *I*-element. While this approach is
906 likely to misclassify some truly active elements as inactive, the two categories should be fairly
907 representative of the active and inactive families in *D. melanogaster*.

908

909 ***Sequence data and read mapping***

910 Sequencing data come from a previously published study in which pools of ~50 females from 85
911 wild-derived *Drosophila melanogaster* lines covering five continents were sequenced with Illumina 100 nt
912 paired-end reads to an average depth of 12.5X (Grenier *et al.* 2015, NCBI BioProject PRJNA268111).

913 We employed the ConTExt pipeline to organize these data and discover structural and sequence
914 variation within the repetitive portions of these genomes (McGurk and Barbash 2018). Briefly, ConTExt
915 aligns repeat-derived reads to the corresponding repeat consensus sequences (using Bowtie2 v. 2.1.0,
916 Langmead and Salzberg 2012). Mixture modeling is used to infer the set of underlying structures that
917 generated the set of discordantly aligned read pairs in each sample, and ConTExt subsequently clusters
918 these junctions to determine which structures are present in multiple samples. We modified the pipeline
919 here to infer sequence variation within repeat families from the aligned reads (described below).
920 Additionally, rather than align reads first to individual insertions and then collapse these alignments onto
921 consensus sequences (McGurk and Barbash 2018), here we aligned reads directly to the consensus
922 sequences using permissive alignment parameters (`--score-min L,0,-2.5 -L 11 -N 1 -i`
923 `S,1,.5 -D 100 -R 5`).

924

925 ***Filtering ambiguous alignments***

926 While Bowtie2's mapping quality summarizes the ambiguity of alignments, it heavily penalizes
927 divergence from the reference sequences and is defined in a rather opaque manner involving nested
928 conditionals. Because we expect some reads to derive from TE insertions diverged from the consensus,
929 it is undesirable to penalize reads with no secondary alignments but which are nonetheless diverged
930 from the consensus. We instead filter ambiguous alignments by directly considering the primary and
931 secondary alignment scores, which we convert to percent identities assuming all penalties are due to
932 mismatches. We use these to define a score *M* reflecting the distance between the primary (*AS*) and
933 secondary (*XS*) hits:

934 $$M = \frac{AS - XS}{1 - XS}(AS - XS) = \frac{(AS - XS)^2}{1 - XS}$$

935 This score summarizes the distance between the primary and secondary alignments and is
936 penalized by the divergence of best alignment from the consensus, but it does so in a predictable
937 fashion. If a secondary alignment is reported by Bowtie2, we require this score to be greater than 0.05 for
938 the alignment to be considered unambiguous and included in the analysis. If the primary hit perfectly
939 matches the consensus, the secondary alignment must be more than 5% diverged from the consensus
940 for the read to be included in the analysis. If the primary alignment is 10% divergent, the secondary
941 alignment must be more than 20% diverged from the consensus for the read to be included in the
942 analysis (Figure S3B). Finally, we exclude any read whose primary alignment is more than 20% diverged
943 from the consensus.

944

945 ***Overview of statistical analyses***

946 We describe the details of each model employed in the subsequent sections, but outline the
947 general logic of our analyses here. We approached the analyses of the data from the perspective of
948 Bayesian parameter estimation, seeking to define probability distributions over all possible values of our
949 statistics of interest to guide our interpretations. To accomplish model-checking, we employ posterior
950 predictive simulations to evaluate the ability of our models to account for features of the observed data

951 as a means of model checking. When the failure of a model to account for features of the data is an
952 observation of interest, we report the posterior predictive p-value (Gelman 2013), which unlike the
953 frequentist p-value is conditioned upon the set of models most consistent with the data rather than a
954 predefined null-hypothesis.

955 When relating read counts to copy number, we employ the negative binomial distribution as it can
956 model overdispersed counts and the sum of its random variables are also negative binomially distributed,
957 allowing us to describe the read count distributions under different copy numbers. That is, if the read
958 depth of a single copy structure is negatively binomially distributed with mean μ and dispersion α, we
959 can model the read depth *x* of a structure present in *n* copies as $x \sim NB(n\mu, n\alpha)$. We further employ
960 partial pooling, which jointly estimates the copy number of individual structures and the underlying copy
961 number distribution from which junctions arise, using the complete set of observed read counts to
962 constrain the individual copy number estimates and prevent outlier read counts from being unduly
963 interpreted as multicopy structures. When making inferences about copy number distributions, we focus
964 on identifying the mean and variance parameters of the distributions. However, TE copy number, and
965 consequently telomere length, are prone to outliers relative to a normal distribution, even in log scale,
966 which likely reflect atypical events that resulted in a copy number expansion. Therefore, we model the
967 copy number distribution as a mixture of two distributions, one reflecting typical copy number variation
968 and the other reflecting the distribution after extreme copy number expansion, and incorporate latent
969 variables that assign each observed copy number as an inlier or an outlier.

970 When estimating relative quantities such as proportions or frequencies we employ beta-binomial
971 and Dirichlet-multinomial models, and whenever we believe that strain-specific factors, such as telomere
972 composition, are relevant, we incorporate this hierarchical structure into the model.

973 For Bayesian modelling we employ PyMC3 (v. 3.9). For each model, we draw samples from the
974 posterior distribution in two chains to assess convergence, sampling continuous random variables with
975 the No-U-Turn Sampler initialized with jitter+adapt_diag as implemented in PyMC3. The No-U-Turn
976 Sampler (NUTS) is the current state of the art in Monte Carlo sampling, capable of efficiently exploring
977 high-dimensional parameter spaces and drawing uncorrelated samples from the posterior. While superior

978 to the Metropolis algorithm in most regards, it is unable to sample discrete random variables due to its
979 reliance on gradient information. Consequently, we represent copy number in our analyses with positive
980 continuous random variables rather than with discrete variables, preferring greater confidence in the
981 posterior sample afforded by NUTS. We note that treating copy number as continuous is not dissimilar
982 from the commonly used approach of estimating copy number by dividing the observed read counts by
983 the expected read depth. Further, as the genomic DNA for each GDL strain was obtained from a pool of
984 individuals, the average copy number per individual in the stock is better described by a continuous
985 rather than discrete variable. Similarly, when we incorporate outlier detection in our models we
986 marginalize out the binary labels representing the outlier status of individual observations to avoid the
987 need for Gibbs sampling. We retain the interpretability provided by these labels by using the posterior
988 distribution to reconstruct the probability that each observation reflects an outlier.
989

990 ***Modelling read depth***

991 To infer the copy number of multicopy sequence, we need to know the read depth distribution of
992 single-copy sequences. We estimate this from coverage of the two major autosomes in the reference
993 genome, using the same filtering steps described in McGurk and Barbash (2018). As in McGurk and
994 Barbash (2018), we consider the coverage of sequence not by the reads themselves, but rather by the
995 interval between read pairs. However, here we model the read depth of single-copy sequence with
996 negative binomial distributions, allowing the mean and overdispersion to vary with %GC. To better
997 handle junctions with extreme %GC, we model the relationship between the mean and overdispersion of
998 %GC in each strain *j* with library-specific functions:

999 $$\mu_j(g) = 2^{\beta_2 g^2 + \beta_1 g + \beta_0}$$

1000 $$\alpha_j(g) = 2^{\beta_2 g^2 + \beta_1 g + \beta_0}$$

1001 where *g* is the expected %GC of read pairs spanning a position and the βs are the coefficients of each
1002 quadratic function (Figure S3D, D).

1003 Additionally, there is a clear excess of positions with zero coverage in the data that is likely due to
1004 filtering, but homozygous deletions may also contribute. We account for this by fitting a zero-inflated
1005 negative binomial distribution when inferring the mean and overdispersion functions, to ensure the
1006 excess of zeroes does not downwardly bias the expected read depth. We allow the amount of zero-
1007 inflation to vary by %GC, relating it with a logistic function as the expected proportion of zeros must be
1008 restricted between 0 and 1:

1009 $$\varphi_j(g) = \frac{1}{1+e^{-(\beta_2 g^2+\beta_1 g+\beta_0)}}$$

1010 For each sample we obtain maximum-likelihood estimates of these functions using PyMC3.

1011

1012 ***Estimating copy number from read depth***

1013 For each position of a consensus sequence we count the number of concordant read pairs which
1014 span it in each strain. We estimate the number of copies containing a given position by dividing the read
1015 count at each position by the expected read depth given the model of GC-bias we inferred for that strain
1016 (Figure S3C, D). From this we estimate sequence abundance by summing the copy number of each
1017 position. The edges of the sequence have reduced mappability as most read pairs span a junction
1018 between the repeat and some other sequence. To account for this, we exclude the first and last 500 bp
1019 from our estimate of sequence abundance and divide the resulting sum by $\frac{L-1000}{L}$, where L is the length
1020 of the consensus sequence. We do not estimate sequence abundance for consensus sequences shorter
1021 than 1,000 bp.

1022 When estimating the total amount of HTT sequence comprising the telomeres, we do not filter
1023 reads for ambiguous alignments, as most ambiguity in reads aligning to the HTTs is caused by homology
1024 among related subfamilies, so while it may be unclear which HTT gave rise to such a read it is almost
1025 certainly derived from the telomere. However, when estimating the copy number of specific HTT families
1026 or subfamilies, we do filter out ambiguously aligned reads.

1027

1028 ***Estimating the breakpoints of terminal deficiencies from read depth***

1029 We identified terminal deficiencies by visually examining the coverage of subtelomeric sequence
1030 for obvious loss of read depth (Supplemental File S3). To estimate the breakpoint of a terminal deficiency
1031 based on read depth, we sought to identify a sequence coordinate beyond which the copy number,
1032 estimated as the ratio between the observed and expected coverage, drops below 1. For each of the
1033 autosome arms, we considered coverage of the most telomere-proximal 100 kb sequence and thinned it
1034 by retaining only every hundredth position, to account for autocorrelation in the coverage signal. We then
1035 removed any position that had more than five repeat-masked nucleotides within 100 bp to reduce the
1036 impact of repeat masking on coverage. At each position, we divided the observed read count by the
1037 expected read count given the local GC-content. We then fit a step function to these copy-number
1038 estimates, *x*, where the expected copy number differs from 1 on the telomere-proximal side of a
1039 breakpoint, *c*. We estimate this breakpoint by minimizing the sum of squared residuals (SSR), defined for
1040 positions on the centromere proximal side of the breakpoint as

1041 $$SSR_{cent} = \sum_i (x_i - 1\,)^2$$

1042 and on the telomere-proximal side as

1043 $$SRR_{tel} = \sum_i (x_i - \mu_{tel})^2$$

1044 where μ_{tel} is the mean copy number estimate on the telomere-proximal side of the breakpoint.
1045 We combine these as

1046 $$SSR = SSR_{cent} + SSR_{tel}$$

1047 We noted a tendency for the last 10 kb of the 2R subtelomere to have reduced coverage in most
1048 strains, which may result from decreased mappability. We were therefore conservative when calling
1049 deficiencies in this region, requiring a complete loss of read depth to call a deficiency in the absence of
1050 an HTT-subtelomere junction, and only calling heterozygous deficiencies if a HTT-subtelomere junction
1051 accompanied the drop in coverage.

1052 If we had evidence of an HTT-subtelomere junction at a breakpoint, we used the sequence
1053 coordinates of the junction as the deficiency breakpoint. In estimating the number of independently
1054 derived deficiencies, we considered any pair of deficiencies within 200 nt of each other as potentially the
1055 same allele and collapsed them.
1056
1057 ***Analysis of truncated HTTs***
1058 HTTs, like all non-LTR retrotransposons, are frequently 5'-truncated due to incomplete reverse
1059 transcription, in addition to being susceptible to telomere erosion. While ConTExt does not permit the
1060 reconstruction of HTT insertions, we can determine the extent to which each insertion is 5'-truncated. A
1061 full-length insertion must have the intact 5'-end, though the presence of the 5'-end does not guarantee
1062 that there are not deletions within the insertion. We note, however, that we observe few junctions
1063 consistent with internal deletions within HTTs. Given the high rate of truncation shown in Figure 2C,
1064 strains containing only truncated insertions of a particular family might be common. But because such
1065 elements are unlikely to encode for functional proteins, and only the autonomous *TAHRE* and *TART*
1066 family elements encode reverse transcriptase, some full-length elements are likely maintained by
1067 selection. We found that 33% of strains lack any full-length *TART* family elements, but only 2 out of 85
1068 strains lack full-length *TAHRE*. To assess the likelihood of this observation, we model the number of full-
1069 length elements and then used posterior predictive simulations to ask whether we observe fewer strains
1070 without any full-length *TAHRE* elements than would be expected under the fitted model.
1071 We estimate the length of insertions by considering the coordinates of an element's 5'-end in an
1072 HTT-HTT junction, and define an element as being full-length if its estimated end is within 300 bp of the
1073 expected 5'-end of its consensus sequence. As elements of the *TART* family are flanked by terminal
1074 repeats, we require the junction to involve the 5'-end of the PNTR rather than the internal sequence to be
1075 considered full-length, though some fraction of such junctions may involve the 3'-PNTR of *TART* instead
1076 and reflect truncated elements. As the counts are overdispersed, for each family we model the number of
1077 full-length insertions, k_i out of all N_i insertions in strain *i* with a beta-binomial distribution. For
1078 interpretability, we estimate mean, q, and concentration, ν, parameters that respectively describe the

1079 expected binomial rate of full-length elements and the variability around this which drives the
1080 overdispersion. We define the priors in this model as

1081 $$q \sim Beta(\alpha = 1, \beta = 1)$$

1082 $$\nu \sim Exponential\ (\lambda = .1)$$

1083 placing a uniform prior on the rate of full-length elements. We model the likelihood as

1084 $$p \sim Beta(\alpha = q\nu, \beta = (1-q)\nu)$$

1085 $$k_i \sim Binomial(p, N_i)$$

1086 To assess the extent to which the model accounts for the fraction of strains lacking any full-length
1087 elements, we employ posterior predictive simulations to obtain a posterior p-value that describes the
1088 probability of observing as few or fewer strains without full-length elements present in the GDL given the
1089 posterior distribution. In 99% of the simulations, the fitted model predicted more strains entirely missing
1090 full-length *TAHRE* than we observed. One interpretation is that selection acts on the genome to maintain
1091 the reverse transcriptase encoded by *TAHRE*. Alternatively, in most genomes there may be a full-length
1092 *TAHRE* insertion in tandem with another HTT that is rarely deleted, perhaps because it is located outside
1093 of the telomere. We present in the Results evidence in favor of this alternative explanation, but stress
1094 that these *TAHRE* elements are old and likely non-functional.
1095

1096 ***Interpreting sequence variation within repeats***

1097 In our analyses of sequence variation, we estimate the copy numbers of the different alleles
1098 found at a given position. We first count the number of reads supporting each allele at that position,
1099 excluding nucleotides whose PHRED base quality score is less than 30 as well as those in the first and
1100 last five bases of the read. This filtering score reduces the contribution of sequencing and alignment
1101 errors, but decreases the read count. Therefore, allele copy number estimates based on these counts
1102 are downwardly biased. To infer the copy number of these alleles, we therefore also compute the
1103 proportion of reads that support each allele and multiply by the estimated copy number at that position. If
1104 the copy number of an allele is estimated to be a small fraction (< 0.2), we assume it reflects sequencing
1105 errors and treat it as zero.

1106 To estimate the sequence diversity at a position, we pool these allele copy numbers across all
1107 strains in the GDL, treating all copies of the family as members of the same population of TEs. We then
1108 estimate diversity as

1109 $$diversity = 1 - \sum_{nt \in \{A,T,C,G\}} \left(\frac{X_{nt}}{N}\right)^2$$

1110 where *N* is the total copy number at that position and X_{nt} is the copy number of a given allele. We
1111 define the major allele as the allele with the greatest copy number in the GDL and the minor allele as that
1112 with the second greatest copy number. We estimate the mean and variance of these alleles from their
1113 estimated copy number in each strain.
1114

1115 ***Correcting for additional biases in read counts over junctions***

1116 In our analyses of the read counts over HTT-HTT junctions we often observe fewer reads than
1117 expected given a single-copy junction, which could reflect residual heterozygosity, the presence of
1118 additional sequence within the junctions such as 3'-tags or polyA-tails, or reads having been filtered out
1119 due to alignment ambiguity. We could determine the zygosity status of the junctions between HTTs and
1120 healed terminal deficiencies independently of their read counts by examining the coverage of unique
1121 sequence on the telomere-proximal side of the junction. Despite that most such deficiencies are truly
1122 homozygous, we noted the same deflation of read counts we observed for HTT-HTT junctions, leading
1123 us to conclude this reflects a downward bias rather than heterozygosity. To assess the degree of this
1124 bias, we modelled the read counts over these HTT-deficiency junctions as

1125 $$x_{ij} \sim NB\left(\lambda n_{ij}\mu_j(g_{ij}), \qquad \lambda n_{ij}\alpha_j(g_{ij})\right)$$

1126 where n_{ij} is the true copy number of the junctions (1 if homozygous and 0.5 if heterozygous), $\mu_j(g_{ij})$ is
1127 the expected read count given the %GC (Figure S3C, D) and $\alpha_j(g_{ij})$ is the degree of overdispersion.
1128 The degree of bias is λ over which we place a prior of

1129 $$log\ \lambda \sim Normal\ (0, .5)$$

1130 Fitting this model, we find that the read count over these junctions is 43% (95% credible interval 35% -
1131 49%) of what we expected given the read depth distributions observed for single copy sequence. We
1132 incorporate this bias (and our uncertainty about its exact degree) into our copy number estimates of
1133 junctions between HTTs and other repeats, calibrating their read counts against this subset of HTT
1134 junctions with known copy number. When we use these corrected copy number estimates to infer the
1135 total amount of telomeric sequence in each strain, we find a strong concordance with our estimates
1136 based on consensus coverage without MapQ filtering (Figure 1D).
1137
1138 ***Identifying TE copies from their junctions***
1139 We identify and count the numbers of euchromatic TE insertions as described in McGurk and
1140 Barbash (2018), by identifying the 3' and/or 5' junctions the TE forms with its insertion site. For *R*-
1141 elements, we estimate the copy number from the read count of the junction between their 3'-end and the
1142 rDNA. As all such *R*-element junctions all have the same sequence coordinates and are present in
1143 multiple copies rather than being a set of single and low-copy junctions like the HTT-HTT junctions, we
1144 estimate their copy as the ratio between the observed and expected read count (assuming a single copy
1145 structure with the same %GC) without needing to employ partial pooling to avoid overestimating their
1146 copy number. As described in the next section, we employ partial pooling to estimate the copy number of
1147 HTT insertions, as we expect there might be both single and multi-copy HTT-HTT junctions.
1148 However, as we estimated the copy number of TE insertions (by counting junctions) in unique
1149 sequence differently from that of the HTTs and the *R*-elements (by also considering read counts), we
1150 worried that some differences among these categories could reflect differences in the behavior of these
1151 estimators rather than true biological differences. To rule this out, we also estimated copy number from
1152 the read depth near the 3'-ends of each family's consensus, estimating copy number as the average
1153 sequence abundance over a 1 kb interval ending 500 bp from the element's 3'-end. Considering only the
1154 3'-end helps limit the impact of 5'-truncation on estimated copy number. While this ensures the copy
1155 numbers of all TE families are estimated in the same way, it does not distinguish TE-derived satellites,
1156 such as the large *R1*- and *Bari1*-derived satellite arrays, from dispersed TEs. We are, however, able to

1157 distinguish these in the copy number estimates based on junctions rather than read depth (i.e. tandem
1158 *Bari1* and *R1*-elements form junctions with themselves, rather than with unique sequence or the *R*-
1159 element's rDNA insertion site).
1160
1161 ***Modelling copy number of HTT-HTT junctions***
1162 While many of the junctions we identify may reflect single-copy structures, it is likely that an
1163 appreciable fraction of junctions are present in multiple copies. We therefore estimated the copy
1164 numbers of HTT junctions based on their read counts. For a junction *i* in sample *j*, we model its read
1165 depth x_{ij} as arising from a negative binomial distribution truncated at 2, as we only analyzed junctions
1166 supported by at least two reads

1167 $$x_{ij} \sim NB_{\geq 2}\left(n_{ij}\mu_j\left(g_{ij}\right)\lambda, \qquad n_{ij}\alpha_j\left(g_{ij}\right)\lambda\right)$$

1168 Here μ_j and α_j are the library-specific functions that describe how the read depth and overdispersion
1169 vary with the expected %GC of read pairs spanning the junction, and g_{ij} and λ are the aforementioned
1170 bias. Finally, n_{ij} is the quantity of interest: the underlying copy number of the junction.
1171 If we estimated this copy number for each junction independently with weakly informative priors,
1172 we would upwardly bias our copy-number estimates. This is because *a priori* we should expect that most
1173 HTTs junctions are truly single-copy and that higher than expected read counts often reflect single-copy
1174 structures that generated more reads than expected by chance. But, as we do not know ahead of time
1175 what fraction are single-copy, we sought to infer the underlying copy number distribution at the same
1176 time we estimate the copy numbers of the individual junctions. This accomplishes the parameter
1177 estimation analog of a multiple-test correction, by shrinking copy-number estimates away from extreme
1178 values unless the data justify believing the copy number greater than 1. We assume that the mode of this
1179 copy-number distribution is 1, and represent how much junction copy number varies around this mode
1180 with a variance parameters (Figure S2 A, B)
1181 Thus we model the overall copy-number distribution of HTT junctions as

1182 $$s \sim HalfCauchy\ (\beta = 1)$$

$$log_{10} n_{ij} \sim Normal\ (0, s)$$

where s is the standard deviation of the copy number distribution. We model the reduced read counts over HTT junctions as

$$log\ \lambda \sim Normal\ (-0.38, 0.07)$$

We bring these together to model the read counts over individual HTT junctions with a negative binomial distribution, truncated at 2 to reflect the minimum junction read count for inclusion in this analysis:

$$x_{ij} \sim NB_{\geq 2}\big(n_{ij}\mu_j(g_{ij})\lambda, \qquad n_{ij}\alpha_j(g_{ij})\lambda\big)$$

Estimating the proportion of missed junctions

The GDL strains were sequenced at relatively low depth, ranging from 10-20 reads expected per junction. This, coupled with the filtering out of HTT reads that could not be assigned unambiguously, means that some fraction of junctions truly present in a strain must have been missed by our analysis. We therefore sought to quantify the extent of this. To this end, for each of our 2,000 posterior samples of the read count model's parameters we simulated new read counts from a non-truncated negative binomial distribution

$$x'_{ij} \sim NB\big(n_{ij}\mu_j(g_{ij})\lambda, \qquad n_{ij}\alpha_j(g_{ij})\lambda\big)$$

This provided a large set of read counts generated from junctions whose copy number, GC-content, and read depth are well-matched to those observed in the GDL. For each strain in each of the 2,000 replicates of this simulation, we computed that fraction of junctions with fewer than two reads. For each strain, we take the average of this value across all 2,000 as an estimate of the rate at which truly present junctions would be missed.

We note that there may be some downward bias to this estimate, as the set of observed junctions will be more enriched for multicopy junctions than the set of missed junctions. So these simulations assume a higher proportion of multicopy junctions than may have been truly present in the GDL.

1209

1210 ***Modelling telomere length and copy number distributions***

1211 Telomere length and HTT copy number distributions share similar features, so we model both
1212 under the same framework. First, as both are restricted to positive values, we model their log-
1213 transformed values. Second, we observe values that are outliers relative to this log-normal distribution.
1214 These outliers, for reasons discussed in the Results, almost certainly reflect events that occurred while
1215 the strains were maintained as lab stocks. Consequently, we model the data as a mixture of two
1216 distributions, one largely reflecting natural variation and the other reflecting anomalous copy number
1217 expansions.

1218 To infer the telomere-length distribution of each population in a way that accounts for the
1219 observed outliers, we modeled the $\log_{10}$-transformed total telomere length x_i of a typical individual *i* in
1220 population j as arising from a Normal distribution, with population-specific means, μ_i, and standard
1221 deviation, σ_i.

1222 $$log_{10} x_i \sim Normal(\mu = \mu_j, \sigma = \sigma_j)$$

1223 If the observation reflects an atypical copy number expansion we model it as arising instead from

1224 $$log_{10} x_i \sim Normal\left(\mu = \mu_j + 2\sigma_j + \mu_{out}, \sigma = \frac{1}{2}\mu_{out}\right)$$

1225 where μ_{out} reflects how much greater than two standard deviations above the inlier mean the outliers
1226 tend to be and over which we place a prior which considers an order of magnitude copy number increase
1227 a plausible outlier

1228 $$\mu_{out} \sim HalfNormal(\sigma = .5)$$

1229 The outlier status of each individual could be modeled with a variable $z_i$which equals 0 for inliers
1230 and 1 for outliers so that

1231 $$log_{10} x_i \sim Normal(\mu = \mu_j + z_i(2\sigma_j + \mu_{out}), \sigma = \sigma_j(1 - z_i) + \frac{1}{2}\mu_{out} z_i)$$

1232 and we would model z_i as being Bernoulli distributed with a probability of being an outlier *p* arising from a
1233 uniform distribution on the interval $\left(0, \frac{1}{2}\right)$

$$z_i \sim Bernoulli(p)$$

$$p \sim Uniform(0, \frac{1}{2})$$

However, as binary variables are challenging to sample we marginalize these labels out of the model (Hogg *et al.* 2010),

$$P(\, log_{10}x_i \mid \theta\,) = P(\, log_{10}x_i \mid \theta, z_i = 0\,)P(\, z_i = 0\,) + P(\, log_{10}x_i \mid \theta, z_i = 1\,)P(\, z_i = 1\,)$$

where θ represents all model parameters ($\mu_j, \mu_{out}, \sigma_j$) other than the outlier labels, and $P(z_i = 1) = p$ and $P(z_i = 0) = 1 - p$. To compute the posterior probability that an observation is an outlier we use

$$P(\, z_i = 1 \mid log_{10}x_i, \theta\,) = \frac{P(\, log_{10}x_i \mid z_i = 1, \theta\,)P(\, z_i = 1)}{P(\, log_{10}x_i \mid z_i = 0, \theta\,)P(\, z_i = 0\,) + P(\, log_{10}x_i \mid z_i = 1, \theta\,)P(\, z_i = 1\,)}$$

and average over all values of θ sampled from the posterior. Those strains that have a probability of being outliers are depicted in Figure 2A and Figure S1B-J as black dots.

When modelling telomere length we describe our prior beliefs as

$$\mu_i \sim Normal\ (\mu = 5, \sigma = 1.5)$$

$$\sigma_i \sim HalfNormal\ (\sigma = 2)$$

These priors imply that average total telomere lengths may be as small as 1 kb or as large 100 Mb. Similarly, the prior on σ is open to the standard deviation of telomere length encompassing four orders of magnitude. Both priors are somewhat overly permissive allowing the data to drive the parameter estimates with little constraint from our prior beliefs.

Modeling the propensity of different HTTs to heal telomere erosion

To determine whether some families have a higher propensity to heal telomere erosion, we compared the frequency with which an element is found next to subtelomeric sequence against the proportion of HTT copies it comprises in each strain. As junctions with subtelomeric sequence involve the 3'-ends of the HTTs, we estimate the copy number of each family from the number of junctions involving the element's 3'-end, ensuring any biases resulting from mappability or structural variation at the 3'-end of the elements affect both estimates. We employ a Dirichlet-Multinomial model where the number of

junctions x_{ij} between an HTT family i and the known chromosome 2 and 3 telomere associated sequence (TAS) repeats across all strains *j* arises as

$$x_j \sim Multinomial(N = n_j, p = p_j)$$

where n_j is the total number HTT-TAS and HTT-deficiency junctions we detect in strain *j*, p_j is a vector describing the probability of finding each HTT family at the base of the telomere in strain *j*,

$$p_{ij} = \frac{\lambda_i q_{ij}}{\Sigma_i \lambda_i q_{ij}}$$

λ_i is the relative enrichment of family i at the base of the telomere, and q_{ij} is the number of HTT elements belonging to family i in the strain *j*, incorporating the differences in telomere composition among strains into the model. We model our prior beliefs about the relative enrichment as being wholly agnostic using a uniform Dirichlet prior

$$\lambda_{\cdot} \sim Dirichlet(\alpha_{\cdot} = 1)$$

We fit this model considering the HTT families as well as subfamilies.

Modelling interspersion among HTTs

To assess the tendency of particular HTTs to neighbor each other we also employ a Dirichlet-Multinomial model. Here we consider an interspersion matrix *X* where each cell X_{ijk} counts the number of times the 3'-end of element *i* forms a junction with 5'-sequence of element *j* in strain *k*. We model the vector of counts in strain *k* as arising from

$$X_k \sim Multinomial(N = N_k, p = p_k)$$

where p_k is a matrix describing the probability that two elements *i* and *j* neighbor each other. This relates to both a tendency, λ_{ij}, of two elements *i* and *j* to neighbor each other and the proportion of HTTs elements *i* and *j* found in strain *k*

$$p_{ijk} = \frac{\lambda_{ij} q_{\cdot jk} q_{i \cdot k}}{\Sigma_i \Sigma_j \lambda_{ij} q_{ijk}}$$

where $q_{\cdot jk}$ and $q_{i \cdot k}$ are the marginal counts

$$q_{i \cdot k} = \Sigma_j q_{ijk}$$

1284 $$q_{\cdot jk} = \Sigma_i q_{ijk}$$

1285 We model prior beliefs about the parameters of interest, the vectorized matrix of associations, with a
1286 uniform Dirichlet distribution

1287 $$\lambda \sim Dirichlet(\alpha_0 .. \alpha_m = 1)$$

1288 We note that this does not model the interspersion under reorderings of existing telomeres but rather
1289 under the generation of new telomeres drawn from the observed proportions.

1290

1291 ***Assessing performance on simulated data***

1292 We simulated Illumina Hiseq 2000 paired-end data from five available GDL PacBio genome
1293 assemblies (B59, I23, N25, T29A, and ZH26; Long *et al.* 2018) using ART (version: MountRanier). To
1294 match the characteristics of the GDL NGS data, we set the genome-wide read depth and fragment size
1295 distributions to match those reported for the particular GDL strains. We ran these simulated datasets
1296 through the ConTExt pipeline using the same parameters as for the real GDL data.

1297 To assess our ability to correctly estimate the copy number of HTT insertions, we compared the
1298 copy-number estimates from the simulated data to the copy number evident in the assemblies from
1299 which the data were simulated. To determine the ground truth for HTT copy numbers in each assembly,
1300 we aligned the HTT consensus sequences to the assemblies using BLAST and filtered out hits with less
1301 than 90% identity to a consensus. We then counted the number of hits corresponding to the intact 3'-end
1302 of each HTT family in the assembly.

1303 **Data availability**

1304 Sequencing data come from a previously published study (NCBI BioProject PRJNA268111, Grenier *et al.*
1305 2015). ConTExt is located at https://github.com/LaptopBiologist/ConTExt. Release 6 of the reference
1306 genome is described in Hoskins *et al.* (2015); the 5 PacBio assemblies of GDL strains are described in
1307 Long *et al.* (2018).

1308

1309 **Competing interests**

1310 The authors declare that they have no competing interests.

1311

1312 **Funding**

1313 This work was supported by NIH grants R01-GM074737 and R01-GM119125 to D.A.B. A.M.D.C. was
1314 supported by postdoctoral fellowships from the FRQ-S (33616), NSERC (PDF-51651-2018) and the
1315 Lawski foundation as well as an NSERC Discovery grant (RGPIN-2019-0544). The funding bodies had
1316 no role in the design of the study and collection, analysis, and interpretation of data and in writing the
1317 manuscript.

1318

1319 **Author's contributions**

1320 Conceptualization: M.P.M, A.M.D.C. and D.A.B., Method development: M.P.M., Analyses: M.P.M and
1321 A.M.D.C., Interpretation: M.P.M, A.M.D.C. and D.A.B. All authors wrote, read and approved the final
1322 manuscript.

1323

1324 **Acknowledgements**

1325 We are grateful to Alla Kalmykova, Mia Levine, Homa Papoli and Jesper Boman who commented on an
1326 earlier version of this manuscript. We also thank Andrew Clark, Cédric Feschotte, Alexander Suh and the
1327 members of their labs for stimulating discussions.

1328

1329 **FIGURE LEGENDS**

1330 **Figure 1. Methodological overview.** (A) Graphical representation of a Drosophila telomere and the
1331 junctions we use to query telomere variation: i) junctions between adjacent HTTs of the same subfamily,
1332 ii) a junction between adjacent insertions of two different HTT subfamilies, iii) a junction between the HTT
1333 at the base of a telomere and its neighboring TAS repeat, and iv) a junction between a healed terminal
1334 deficiency and unique sequence. The bottom diagram v) depicts the structure of a chromosome with an
1335 unhealed deficiency. (B) An example using *HeT-A5* to depict the correspondence between the junctions
1336 of adjacent insertions of the same subfamily (i in A) and Illumina read pairs from strain N16. Each dot
1337 corresponds to a read pair where both ends map unambiguously to the *HeT-A5* consensus sequence.
1338 The Y-axis corresponds to the *HeT-A5* plus strand and the X-axis to the minus strand. The diagonal of
1339 dots colored black corresponds to concordant reads that align as would be expected given the
1340 consensus element. Non-concordant reads aligning off this diagonal reflect junctions between tandem
1341 elements (above the diagonal) or internal deletions (below the diagonal). Small indels may shift reads
1342 just above or below the diagonal, for example the grey cluster near the diagonal. Junctions between
1343 different families of repeats are detected by considering plots where the X- and Y-axes correspond to
1344 different consensus sequences. ConTExt identifies junctions by clustering the non-concordant reads,
1345 cluster assignments are reflected by the color of the dots. The five clusters forming a horizontal line
1346 across the top of the plot correspond to five distinct tandem junctions between the 3' end of a *HeT-A5*
1347 element and the generally truncated 5'-end of an adjacent *HeT-A5*. Two of the tandem arrangements are
1348 illustrated above the plot. (C-F) Comparisons of different approaches for estimating telomere length and
1349 HTT copy number, and comparisons against simulated data. The blue lines indicate OLS linear
1350 regressions and the red dotted lines indicate a one-to-one relationship for comparison. (C) The
1351 relationship between the total number of distinct HTT-HTT junctions identified by ConTExt in each strain
1352 (X-axis) and the total HTT copy number inferred from the read depth over these junctions (Y-axis). D) A
1353 comparison of total telomere length in each strain estimated from the mapping-quality-filtered read count
1354 of junctions (Y-axis) and from coverage of HTT consensus sequences without mapping quality filtering
1355 (X-axis). The junction-based telomere-length estimates are obtained by multiplying the inferred copy

1356 number of each identified HTT-HTT junction by the length of the distal element inferred from its degree of
1357 5' truncation. (E) Correlation in CN estimates from simulated Illumina datasets with true copy number.
1358 Each dot represents the copy number of an HTT family in one of the five PacBio genomes. The X-axis
1359 indicates its copy number estimated from the number of 3'-ends detectable in BLAST alignments
1360 between the PacBio assembly and the HTT consensus sequences. The Y-axis is the copy number
1361 estimated by ConTExt from data simulated from the PacBio genomes using ART. (F) The downward bias
1362 in copy number we observe in the true GDL data is recapitulated in the simulations. The Y-axis is the
1363 observed read count divided by the read count expected given the junction's local GC content. The
1364 boxplots depict the distribution of this ratio across all identified junctions in the true GDL data and the
1365 simulated data.
1366
1367 **Figure 2. Telomeres are highly dynamic in *Drosophila melanogaster*.** (A) Telomere length
1368 distribution (in kb) as estimated from HTT sequence abundance for each strain, grouped by population.
1369 Filled circles represent outlier strains. B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe.
1370 (B) A ternary plot depicting the proportion of each HTT family in each GDL strain. The angle of the tick on
1371 each axis indicates the corresponding gridline for that HTT family. (C) Proportion of full-length elements
1372 per subfamily per strain. (D) Telomere composition depicted by proportion of total telomere length per
1373 HTT subfamily as estimated from copy number. White corresponds to short telomere strains (bottom
1374 10%), grey to long telomere strains (top 10%). (E) Telomere composition depicted by proportion of total
1375 telomere length per HTT family as estimated from copy number. White and grey are the same as in (D).
1376
1377 **Figure 3. Comparing HTT variability to other TEs within *Drosophila melanogaster*.** (A) Copy
1378 number of selected TEs per strain as estimated from junctions. Each dot indicates the copy number of a
1379 TE family in a single strain, and red dots indicate strains with extreme copy-number expansions (four
1380 standard deviations greater than the mean). TEs are grouped as: Left, TEs with copy number outliers;
1381 middle, HTTs; right, all other active TEs. Outliers occurring in strain I01 are indicated with arrows. (B)
1382 Scatter plot showing the relationship between mean copy number as estimated from junction data of TE

1383 families (log scale) and their variance (log scale) across the GDL. Designations of active and inactive
1384 TEs are from prior estimates of sequence divergence and population frequency as described in the
1385 Materials and Methods. Solid line represents the expected relationship under the assumption of little
1386 variation in population frequency and low linkage disequilibrium among insertions. Shaded regions
1387 summarize the distributions of mean and variance for inactive (grey), active (yellow), and the HTT and *R*-
1388 elements (purple) TE families, covering two standard deviations of bivariate Gaussians matched to the
1389 moments of the data they are approximating. (C) Boxplots depicting the distribution of variance-to-mean
1390 ratios ($\log_{10}$) in each of the four categories of the TEs.
1391
1392 **Figure 4. HTT copy number and sequence variation in the context of other TEs.** (A) A schematic of
1393 how allele copy number is analyzed. Top: Hypothetical copy number of a TE family in a single strain,
1394 including two positions where polymorphisms exist such that the copy number can be partitioned into the
1395 major (black) and minor alleles (orange). Bottom: A schematic depicting hypothetical copy number
1396 variation of the most common minor allele across six strains. If the variance in copy number is greater or
1397 less than the mean, the copy number variation is called overdispersed (light red) or underdispersed
1398 (blue), respectively. (B) Sequence diversity across all copies of each active TE family in the GDL,
1399 estimated from the depth of reads supporting each possible allele. The contribution of positions where
1400 the minor allele displays overdispersed copy number variation, suggesting a variant in an active element,
1401 is indicated in red. The contribution of positions where the minor allele displays underdispersed copy
1402 number variation, suggesting a variant in an inactive element, is indicated in blue. (C) Boxplots depicting
1403 the fraction of diversity contributed by positions with overdispersed minor alleles for both HTTs and non-
1404 HTT active TEs. The difference in medians is 49% ($p = 3e-4$, permutation test, 100,000 permutations).
1405 (D-E) The mean-variance relationships of *HeT-A* (D) and *hobo* (E) broken down by the copy number of
1406 the major and minor alleles. Each dot reflects the observed mean and variance of the copy number of the
1407 major (blue) and minor (gold) alleles of positions with >0.1 sequence diversity. For reference, the shaded
1408 regions are re-plotted from Figure 3B. (F) The mean-variance relationship of *TAHRE*'s minor alleles,

1409 colored by whether the minor allele is found in the heterochromatic *TAHRE* insertions (blue) or is likely
1410 telomeric (red). Shaded regions are re-plotted from Figure 3B.
1411
1412 **Figure 5. Terminal deficiencies are frequent and tend to be healed by *TART* family elements.** (A)
1413 Location of all identified terminal deficiencies (triangles represent individual terminal deletions) and
1414 elements involved (see colour legend; "none" indicates a terminal deficiency lacking any HTT-
1415 subtelomere junction) across all lines. Thin and thick black bars represent UTRs and exons respectively,
1416 thin lines represent introns. Top, Chromosome 2R; middle, Chromosome 3L; bottom, Chromosome 3R.
1417 X-axis scales are in kilobases; note that the telomeres are to the right for 2R and 3R, and to the left for
1418 3L. (B) TAS-L copy-number boxplot in strains with (+) or without (–) deficiencies on 3L. (C) Boxplots
1419 comparing the total telomere length of strains with homozygous deficiencies and heterozygous
1420 deficiencies without observed HTT-subtelomere junctions (unhealed) to the telomere length of strains
1421 with healed deficiencies and without deficiencies. Strains with healed deficiencies (whether homozygous
1422 or heterozygous) will have eight telomeres whereas those with homozygous unhealed deficiencies will
1423 have only seven. Note that the Y-axis is in $\log_{10}$-scale.
1424
1425 **Figure 6. HTT insertions tend to be enriched adjacent to themselves.** (A) The observed frequency
1426 with which two HTTs neighbor each other relative to the expected frequency ($\log_2$). (B) Boxplots of the
1427 posterior distributions describing the degree to which elements tend to neighbor themselves ($\log_2$). The
1428 whiskers reflect the 95% credible intervals. (C) A visualization of the HTT subfamilies, depicted as thick
1429 bars, in the X chromosome telomere of the Release 6 reference. We depict alignments with at least 90%
1430 identity to the consensus; if a region is homologous to two elements, we assign it to the element with the
1431 greatest homology, which was only an issue due to some homology between *TART-A* and *TART-C*. The
1432 upward ticks indicate the 3'-end of an element and the downward ticks the 5'-end; a full length insertion
1433 has both ticks. TART-A_PNTR is the TART-A Perfect Near-Terminal Repeat. (D) Top: The observed
1434 proportion with which each subfamily is found anywhere in the telomere (white) or is the first HTT found
1435 at the base of a telomere (grey). The error bars are 95% credible intervals computed analytically for

1436 Dirichlet-Multinomial models with uniform priors. Bottom: Boxplots summarizing posterior samples of the
1437 relative enrichment ($\log_2$) of each subfamily at the base of the telomere, accounting for telomere
1438 composition differences across strains. The whiskers span the 95% credible interval, determined as
1439 quantiles of the posterior sample.

1440 **SUPPLEMENTARY TABLES**

1441 **Table S1: HTT copy numbers in Pacbio assemblies and corresponding ConTExt estimates.** This
1442 table presents for each HTT family, in each of the five Pacbio assemblies, the copy number estimated
1443 from BLAST alignments of consensus sequences against the assembly and the copy number that was
1444 estimated from applying ConTExt to Illumina data simulated from each assembly.

1445

1446 **Table S2: TE copy number per strain.** The estimated copy number of each TE family in the 85 GDL
1447 strains.

1448

1449 **SUPPLEMENTARY FIGURES**

1450 **Figure S1. The copy number of the HTT subfamilies in each strain, broken down by population**
1451 **and corrected by the estimated fraction of missing junctions.** A) The percent of junctions for all HTT
1452 subfamilies inferred from posterior predictive simulations to have been missed in each strain due to read
1453 depth variation, broken down by population. To ensure interpretations of population structure are not
1454 driven by these missing junctions for the rest of this figure only, when assessing the presence of
1455 population structure we roughly corrected our copy number estimates by dividing by one minus the
1456 inferred fraction of missing junctions for each HTT subfamily. Inferred missing junctions for each HTT
1457 subfamily: B) *HeT-A*, C) *HeT-A1*, D) *HeT-A2*, E) *HeT-A3*, F) *HeT-A5*, G) *TAHRE*, H) *TART-A*, I) *TART-B*,
1458 J) *TART-C.*

1459

1460 **Figure S2. Alternative and supporting visualizations of data .** (A) Scatter plot showing the
1461 relationship between mean copy number (log scale) and variance (log scale) of TE families across the
1462 GDL as estimated from the read depth at the 3'-end of TE consensus sequences rather than junctions.
1463 Designations of active and inactive TEs are from prior estimates of sequence divergence and population
1464 frequency as described in the Materials and Methods. Solid line represents the expected relationship
1465 under the assumption of little variation in population frequency and low linkage disequilibrium among
1466 insertions. (B) Boxplots depicting distributions of variance-to-mean ratios ($\log_{10}$) for each of the four

1467 categories of the TE families based on these copy number estimates. (C) A scatter plot comparing our
1468 estimates of sequence diversity for the TE families in Figure 4B with those reported in Kaminker et al
1469 (2003). This is restricted to TE families included in both analyses. (D) A swarm plot depicting the
1470 distribution of distances between terminal deficiency breakpoints, where for each breakpoint the distance
1471 is plotted relative to the nearest other breakpoint we discovered in a different sample.
1472
1473 **Figure S3. Details on interpreting alignments to telomeric transposable elements** (A) Percent
1474 identities of all >300 bp insertions annotated as *HeT-A* in the UCSC genome browser repeat-masker
1475 track of the Release 6 reference-genome telomeres. Percent identities are based on a ClustalOmega
1476 multiple sequence alignment with default parameters. The Repbase consensus sequence and the
1477 representatives we chose for each *HeT-A* subfamily in our repeat index are indicated as such. The white
1478 cells indicate no alignment between a pair of sequences, all involving alignments with the two shortest
1479 *HeT-A* fragments. (B) A depiction of how we define ambiguous alignments. The black line indicates how
1480 diverged the second best alignment must be for a primary alignment with a given percent identity to be
1481 considered unambiguous. (C) The expected relationship between read depth and %GC for each of the
1482 GDL strains. The variability of this relationship across strains is what necessitated controlling for %GC
1483 when estimating copy number. (D) A heatmap of the expected relationship between read depth and
1484 %GC, sorted by population.

1485 **SUPPLEMENTARY FILES**

1486 **Supplementary File 1. FASTA file of HTT consensus sequences for five *HeT-A* subfamilies (*HeT-***
1487 ***A*, *HeT-A1*, *HeT-A-2*, *HeT-A3*, *HeT-A5*), *TAHRE,* and three *TART* subfamilies (*TART-A*, *TART-B1*,**
1488 ***TART-C*).** The positions of the insertions in the Release 6 *Drosophila melanogaster* reference genome
1489 used as representative sequences for the *HeT-A* subfamilies are indicated.

1490

1491 **Supplementary File 2. Plots of mean-variance relationships between allele copy number for each**
1492 **active TE family.** As in Figures 4D-E, each dot reflects the observed mean and variance of the copy
1493 number of the major (blue) and minor (gold) alleles of positions with >0.1 sequence diversity. For
1494 reference, the shaded regions are re-plotted from Figure 3B**.**

1495

1496 **Supplementary File 3. Visualization of read coverage on chromosomes with evidence for terminal**
1497 **deficiencies.** Each blue dot is the estimated copy number at a given position, determined as the ratio
1498 between the observed and expected read depth. The black lines below the plots indicate which regions
1499 of the reference are repeat-masked (and should show little or no read coverage). The orange trendlines
1500 are a 10 kb rolling window average of estimated copy number and the red lines indicate the average
1501 copy number estimate to the left and right of the inferred deficiency breakpoint. The vertical red line in
1502 some plots indicates the location of the HTT-subtelomere junction. Below each plot we note the
1503 breakpoint coordinate we estimated based on read depth as well as the percent of the coverage prior to
1504 the breakpoint remaining in the deleted regions, from which we infer the heterozygosity or homozygosity
1505 of the deficiency.

1506 **References**

1507 Abad J. P., B. De Pablos, K. Osoegawa, P. J. De Jong, A. Martín-Gallardo, *et al.*, 2004 *TAHRE*, a novel
1508 telomeric retrotransposon from Drosophila melanogaster, reveals the origin of Drosophila telomeres.
1509 Mol. Biol. Evol. 21: 1620–1624.

1510 Agudo M., A. Losada, J. P. Abad, S. Pimpinelli, P. Ripoll, *et al.*, 1999 Centromeres from telomeres? The
1511 centromeric region of the *Y* chromosome of *Drosophila melanogaster* contains a tandem array of
1512 telomeric HeT-A- and TART-related sequences. Nucleic Acids Research 27: 3318–3324.

1513 Arkhipova I. R., 2012 Telomerase, Retrotransposons, and Evolution, pp. 265–299 in *Telomerases*, edited
1514 by Lue N. F., Autexier C. John Wiley & Sons, Inc., Hoboken, NJ, USA.

1515 Bao W., K. K. Kojima, and O. Kohany, 2015 Repbase Update, a database of repetitive elements in
1516 eukaryotic genomes. Mob. DNA 6: 11.

1517 Begun D. J., and C. F. Aquadro, 1995 Evolution at the tip and base of the X chromosome in an African
1518 population of *Drosophila melanogaster*. Mol. Biol. Evol. 12: 382–390.

1519 Berloco M., L. Fanti, F. Sheen, R. W. Levis, and S. Pimpinelli, 2005 Heterochromatic distribution of *HeT-*
1520 A- and *TART*-like sequences in several *Drosophila* species. Cytogenet. Genome Res. 110: 124–
1521 133.

1522 Biessmann H., S. B. Carter, and J. M. Mason, 1990a Chromosome ends in Drosophila without telomeric
1523 DNA sequences. Proc. Natl. Acad. Sci. U. S. A. 87: 1758–1761.

1524 Biessmann H., J. M. Mason, K. Ferry, M. d'Hulst, K. Valgeirsdottir, *et al.*, 1990b Addition of telomere-
1525 associated HeT DNA sequences "heals" broken chromosome ends in Drosophila. Cell 61: 663–673.

1526 Biessmann H., L. E. Champion, M. O'Hair, K. Ikenaga, B. Kasravi, *et al.*, 1992a Frequent transpositions
1527 of *Drosophila melanogaster* HeT-A transposable elements to receding chromosome ends. EMBO J.
1528 11: 4459–4469.

1529 Biessmann H., K. Valgeirsdottir, A. Lofsky, C. Chin, B. Ginther, *et al.*, 1992b HeT-A, a transposable
1530 element specifically involved in “healing” broken chromosome ends in *Drosophila melanogaster*.
1531 Mol. Cell. Biol. 12: 3910–3918.

1532 Biessmann H., B. Kasravi, T. Bui, G. Fujiwara, L. E. Champion, *et al.*, 1994 Comparison of two active
1533 HeT-A retroposons of *Drosophila melanogaster*. Chromosoma 103: 90–98.

1534 Blondel V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, 2008 Fast unfolding of communities in
1535 large networks. Journal of Statistical Mechanics: Theory and Experiment 2008: P10008.

1536 Blumenstiel J. P., 2011 Evolutionary dynamics of transposable elements in a small RNA world. Trends
1537 Genet. 27: 23–31.

1538 Bourque G., K. H. Burns, M. Gehring, V. Gorbunova, A. Seluanov, *et al.*, 2018 Ten things you should
1539 know about transposable elements. Genome Biol. 19: 199.

1540 Boussy I. A., and M. Itoh, 2004 Wanderings of hobo: a transposon in *Drosophila melanogaster* and its
1541 close relatives. Genetica 120: 125–136.

1542 Bucheton A., C. Vaury, M. C. Chaboissier, P. Abad, A. Pélisson, *et al.*, 1992 I elements and the
1543 Drosophila genome. Genetica 86: 175–190.

1544 Casacuberta E., 2017 Drosophila: retrotransposons making up telomeres. Viruses **2017** 9: 192.

1545 Chang C.-H., A. Chavan, J. Palladino, X. Wei, N. M. C. Martins, *et al.*, 2019 Islands of retroelements are
1546 the major components of Drosophila centromeres. PLoS Biol 17(5): e3000241.

1547 Charlesworth B., and D. Charlesworth, 1983 The population dynamics of transposable elements. Genet.
1548 Res. 42: 1–27.

1549 Charlesworth B., and C. H. Langley, 1986 The evolution of self-regulated transposition of transposable
1550 elements. Genetics 112: 359–383.

1551 Charlesworth B., and C. H. Langley, 1989 The population genetics of Drosophila transposable elements.
1552 Annu. Rev. Genet. 23: 251–287.

1553 Cosby R. L., N. C. Chang, and C. Feschotte, 2019 Host–transposon interactions: conflict, cooperation,
1554 and cooption. Genes Dev. 33: 1098–1116.

1555 Danilevskaya O., A. Lofsky, E. V. Kurenova, and M. L. Pardue, 1993 The Y chromosome of *Drosophila*
1556 *melanogaster* contains a distinctive subclass of Het-A-related repeats. Genetics 134: 531–543.

1557 Danilevskaya O. N., I. R. Arkhipova, M. L. Pardue, and K. L. Traverse, 1997 Promoting in Tandem: The
1558 Promoter for Telomere Transposon *HeT-A* and Implications for the Evolution of Retroviral LTRs. Cell
1559 88: 647–655.

1560 Deininger P. L., J. V. Moran, M. A. Batzer, and H. H. Kazazian Jr, 2003 Mobile elements and mammalian
1561 genome evolution. Curr. Opin. Genet. Dev. 13: 651–658.

1562 Doolittle W. F., and C. Sapienza, 1980 Selfish genes, the phenotype paradigm and genome evolution.
1563 Nature 284: 601–603.

1564 Eickbush T. H., 2002 R2 and related site-specific non-long terminal repeat retrotransposons, pp. 813–
1565 835 in *Mobile DNA II*, American Society of Microbiology.

1566 Fuller A. M., E. G. Cook, K. J. Kelley, and M.-L. Pardue, 2010 Gag proteins of Drosophila telomeric
1567 retrotransposons: collaborative targeting to chromosome ends. Genetics 184: 629–636.

1568 Gelman A., 2013 Two simple examples for understanding posterior p-values whose distributions are far
1569 from uniform. Electron. J. Stat. 7: 2595–2602.

1570 George J. A., P. G. DeBaryshe, K. L. Traverse, S. E. Celniker, and M.-L. Pardue, 2006 Genomic
1571 organization of the Drosophila telomere retrotransposable elements. Genome Res. 16: 1231–1240.

1572 George J. A., K. L. Traverse, P. G. DeBaryshe, K. J. Kelley, and M.-L. Pardue, 2010 Evolution of diverse

1573 mechanisms for protecting chromosome ends by Drosophila *TART* telomere retrotransposons. Proc.
1574 Natl. Acad. Sci. U. S. A. 107: 21052–21057.

1575 Golubovsky M. D., 1978 The “lethal giant larvae”-the most frequent second chromosome lethal in natural
1576 population of *D. melanogaster*. Drosoph. Inf. Serv. 53: 179–179.

1577 Green M. M., and S. H. Shepherd, 1979 Genetic instability in *Drosophila melanogaster*: the induction of
1578 specific chromosome 2 deletions by MR elements. Genetics 92: 823–832.

1579 Grenier J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, *et al.*, 2015 Global diversity lines
1580 - a five-continent reference panel of sequenced *Drosophila melanogaster* strains. G3 5: 593–603.

1581 Hogg D. W., J. Bovy, and D. Lang, 2010 Data analysis recipes: Fitting a model to data. arXiv [astro-
1582 ph.IM].

1583 Hoskins R. A., J. W. Carlson, K. H. Wan, S. Park, I. Mendez, *et al.*, 2015 The Release 6 reference
1584 sequence of the *Drosophila melanogaster* genome. Genome Res. 25: 445–458.

1585 Huang W., L. Li, J. R. Myers, and G. T. Marth, 2012 ART: a next-generation sequencing read simulator.
1586 Bioinformatics 28: 593–594.

1587 Jangam D., C. Feschotte, and E. Betrán, 2017 Transposable Element Domestication As an Adaptation to
1588 Evolutionary Conflicts. Trends Genet. 33: 817–831.

1589 Jurka J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, *et al.*, 2005 Repbase Update, a
1590 database of eukaryotic repetitive elements. Cytogenet. Genome Res. 110: 462–467.

1591 Kaminker J. S., C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, *et al.*, 2002 The transposable
1592 elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 3:
1593 research0084.1-0084.20.

1594 Ke N., and D. F. Voytas, 1997 High frequency cDNA recombination of the saccharomyces

1595 retrotransposon Ty5: The LTR mediates formation of tandem elements. Genetics 147: 545–556.

1596 Kelleher E. S., and D. A. Barbash, 2013 Analysis of piRNA-mediated silencing of active TEs in
1597 *Drosophila melanogaster* suggests limits on the evolution of host genome defense. Mol. Biol. Evol.
1598 30: 1816–1829.

1599 Kelleher E. S., D. A. Barbash, and J. P. Blumenstiel, 2020 Taming the Turmoil Within: New Insights on
1600 the Containment of Transposable Elements. Trends Genet. 36: 474–489.

1601 Kern A. D., and D. J. Begun, 2008 Recurrent deletion and gene presence/absence polymorphism:
1602 telomere dynamics dominate evolution at the tip of 3L in *Drosophila melanogaster* and *D. simulans*.
1603 Genetics 179: 1021–1027.

1604 Khan H., A. Smit, and S. Boissinot, 2006 Molecular evolution and tempo of amplification of human LINE-
1605 1 retrotransposons since the origin of primates. Genome Res. 16: 78–87.

1606 Kofler R., A. J. Betancourt, and C. Schlötterer, 2012 Sequencing of pooled DNA samples (Pool-Seq)
1607 uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. PLoS
1608 Genet. 8: e1002487.

1609 Lange T. de, 2009 How telomeres solve the end-protection problem. Science 326: 948–952.

1610 Langley C. H., B. P. Lazzaro, W. Phillips, E. Heikkinen, and J. M. Braverman, 2000 Linkage disequilibria
1611 and the site frequency spectra in the *su(s)* and *su(w(a))* regions of the *Drosophila melanogaster* X
1612 chromosome. Genetics 156: 1837–1852.

1613 Langmead B., and S. L. Salzberg, 2012 Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:
1614 357–359.

1615 Lee Y. C. G., and C. H. Langley, 2010 Transposable elements in natural populations of *Drosophila*
1616 *melanogaster.* Philos. Trans. R. Soc. Lond. B Biol. Sci. 365: 1219–1228.

1617 Lee Y. C. G., C. Leek, and M. T. Levine, 2017 Recurrent Innovation at Genes Required for Telomere
1618 Integrity in Drosophila. Mol. Biol. Evol. 34: 467–482.

1619 Le Rouzic A., T. S. Boutin, and P. Capy, 2007 Long-term evolution of transposable elements. Proc. Natl.
1620 Acad. Sci. U. S. A. 104: 19375–19380.

1621 Levis R. W., 1989 Viable deletions of a telomere from a Drosophila chromosome. Cell 58: 791–801.

1622 Levis R. W., R. Ganesan, K. Houtchens, L. A. Tolar, and F. M. Sheen, 1993 Transposons in place of
1623 telomeric repeats at a Drosophila telomere. Cell 75: 1083–1093.

1624 Long E., C. Evans, J. Chaston, and J. A. Udall, 2018 Genomic Structural Variations Within Five
1625 Continental Populations of Drosophila melanogaster. G3 8: 3247–3253.

1626 Losada A., J. P. Abad, and A. Villasante, 1997 Organization of DNA sequences near the centromere of
1627 the *Drosophila melanogaster* Y chromosome. Chromosoma 106: 503–512.

1628 Markova D. N., S. M. Christensen, and E. Betrán, 2020 Telomere-Specialized Retroelements in
1629 Drosophila : Adaptive Symbionts of the Genome, Neutral, or in Conflict? Bioessays 42: 1900154.

1630 Mason J. M., R. C. Frydrychova, and H. Biessmann, 2008 Drosophila telomeres: an exception providing
1631 new insights. Bioessays 30: 25–37.

1632 Mason J. M., T. A. Randall, and R. C. Frydrychova, 2016 Telomerase lost? Chromosoma 125: 65–73.

1633 Maxwell P. H., J. M. Belote, and R. W. Levis, 2006 Identification of multiple transcription initiation,
1634 polyadenylation, and splice sites in the *Drosophila melanogaster TART* family of telomeric
1635 retrotransposons. Nucleic Acids Res. 34: 5498–5507.

1636 McAllister B. F., and J. H. Werren, 1999 Evolution of tandemly repeated sequences: What happens at
1637 the end of an array? J. Mol. Evol. 48: 469–481.

1638 McGurk M. P., and D. A. Barbash, 2018 Double insertion of transposable elements provides a substrate

1639 for the evolution of satellite DNA. Genome Res. 28: 714–725.

1640 Mechler B. M., W. McGinnis, and W. J. Gehring, 1985 Molecular cloning of *lethal(2)giant larvae*, a
1641 recessive oncogene of *Drosophila melanogaster*. EMBO J. 4: 1551–1557.

1642 Morrish T. A., J. L. Garcia-Perez, T. D. Stamato, G. E. Taccioli, J. Sekiguchi, *et al.*, 2007 Endonuclease-
1643 independent LINE-1 retrotransposition at mammalian telomeres. Nature 446: 208–212.

1644 Okazaki S., K. Tsuchida, H. Maekawa, H. Ishikawa, and H. Fujiwara, 1993 Identification of a
1645 pentanucleotide telomeric sequence,(TTAGG) n, in the silkworm *Bombyx mori* and in other insects.
1646 Mol. Cell. Biol. 13: 1424–1432.

1647 Okazaki S., H. Ishikawa, and H. Fujiwara, 1995 Structural analysis of TRAS1, a novel family of telomeric
1648 repeat-associated retrotransposons in the silkworm, *Bombyx mori*. Mol. Cell. Biol. 15: 4545–4552.

1649 Orgel L. E., and F. H. Crick, 1980 Selfish DNA: the ultimate parasite. Nature 284: 604–607.

1650 Pardue M.-L., and P. G. DeBaryshe, 2008 Drosophila telomeres: A variation on the telomerase theme.
1651 Fly 2: 101–110.

1652 Pardue M.-L., and P. G. DeBaryshe, 2011 Retrotransposons that maintain chromosome ends. Proc. Natl.
1653 Acad. Sci. U. S. A. 108: 20317–20324.

1654 Raffa G. D., L. Ciapponi, G. Cenci, and M. Gatti, 2011 Terminin: a protein complex that mediates
1655 epigenetic maintenance of Drosophila telomeres. Nucleus 2: 383–391.

1656 Roegiers F., J. Kavaler, N. Tolwinski, Y.-T. Chou, H. Duan, *et al.*, 2009 Frequent unanticipated alleles of
1657 *lethal giant larvae* in Drosophila second chromosome stocks. Genetics 182: 407–410.

1658 Roiha H., and D. M. Glover, 1981 Duplicated rDNA sequences of variable lengths flanking the short type
1659 I insertions in the rDNA of *Drosophila melanogaster*. Nucleic Acids Res. 9: 5521–5532.

1660 Saint-Léandre B., S. C. Nguyen, and M. T. Levine, 2019 Diversification and collapse of a telomere

1661 elongation mechanism. Genome Res. 29: 920–931.

1662 Saint-Léandre B., and M. T. Levine, 2020 The Telomere Paradox: Stable Genome Preservation with
1663 Rapidly Evolving Proteins. Trends Genet. 36: 232–242.

1664 Servant G., and P. L. Deininger, 2015 Insertion of Retrotransposons at Chromosome Ends: Adaptive
1665 Response to Chromosome Maintenance. Front. Genet. 6: 358.

1666 Shpiz S., D. Kwon, A. Uneva, M. Kim, M. Klenov, *et al.*, 2007 Characterization of Drosophila telomeric
1667 retroelement *TAHRE*: transcription, transpositions, and RNAi-based regulation of expression. Mol.
1668 Biol. Evol. 24: 2535–2545.

1669 Shpiz S., and A. Kalmykova, 2012 Control of Telomere Length in Drosophila, in *Reviews on Selected*
1670 *Topics of Telomere Biology*, edited by Li B. InTech.

1671 Siriaco G. M., G. Cenci, A. Haoudi, L. E. Champion, C. Zhou, *et al.*, 2002 Telomere elongation (Tel), a
1672 new mutation in Drosophila melanogaster that produces long telomeres. Genetics 160: 235–245.

1673 Stage D. E., and T. H. Eickbush, 2007 Sequence variation within the rRNA gene loci of 12 Drosophila
1674 species. Genome Research 17: 1888–1897.

1675 Stage D. E., and T. H. Eickbush, 2009 Origin of nascent lineages and the mechanisms used to prime
1676 second-strand DNA synthesis in the R1 and R2 retrotransposons of *Drosophila*. Genome Biol. 10:
1677 R49.

1678 Stewart J. A., M. F. Chaiken, F. Wang, and C. M. Price, 2012 Maintaining the end: roles of telomere
1679 proteins in end-protection, telomere replication and length regulation. Mutat. Res. 730: 12–19.

1680 Sultana T., A. Zamborlini, G. Cristofari, and P. Lesage, 2017 Integration site selection by retroviruses
1681 and transposable elements in eukaryotes. Nat. Rev. Genet. 18: 292–308.

1682 Villasante A., J. P. Abad, R. Planelló, M. Méndez-Lago, S. E. Celniker, *et al.*, 2007 Drosophila telomeric

1683 retrotransposons derived from an ancestral element that was recruited to replace telomerase.
1684 Genome Res. 17: 1909–1918.

1685 Villasante A., B. de Pablos, M. Méndez-Lago, and J. P. Abad, 2008 Telomere maintenance in
1686 Drosophila: rapid transposon evolution at chromosome ends. Cell Cycle 7: 2134–2138.

1687 Walter M. F., C. Jang, B. Kasravi, J. Donath, B. M. Mechler, *et al.*, 1995 DNA organization and
1688 polymorphism of a wild-type Drosophila telomere region. Chromosoma 104: 229–241.

1689 Walter M. F., M. R. Biessmann, C. Benitez, T. Török, J. M. Mason, *et al.*, 2007 Effects of telomere length
1690 in Drosophila melanogaster on life span, fecundity, and fertility. Chromosoma 116: 41–51.

1691 Wei K. H.-C., H. M. Reddy, C. Rathnam, J. Lee, D. Lin, *et al.*, 2017 A Pooled Sequencing Approach
1692 Identifies a Candidate Meiotic Driver in Drosophila. Genetics 206: 451–465.

1693 Weissensteiner M. H., A. W. C. Pang, I. Bunikis, I. Höijer, O. Vinnere-Petterson, *et al.*, 2017 Combination
1694 of short-read, long-read, and optical mapping assemblies reveals large-scale tandem repeat arrays
1695 with population genetic implications. Genome Res. 27: 697–708.

1696 Zhao Z., X. Pan, L. Liu, and N. Liu, 2014 Telomere length maintenance, shortening, and lengthening. J.
1697 Cell. Physiol. 229: 1323–1329.

1698

Figure 1

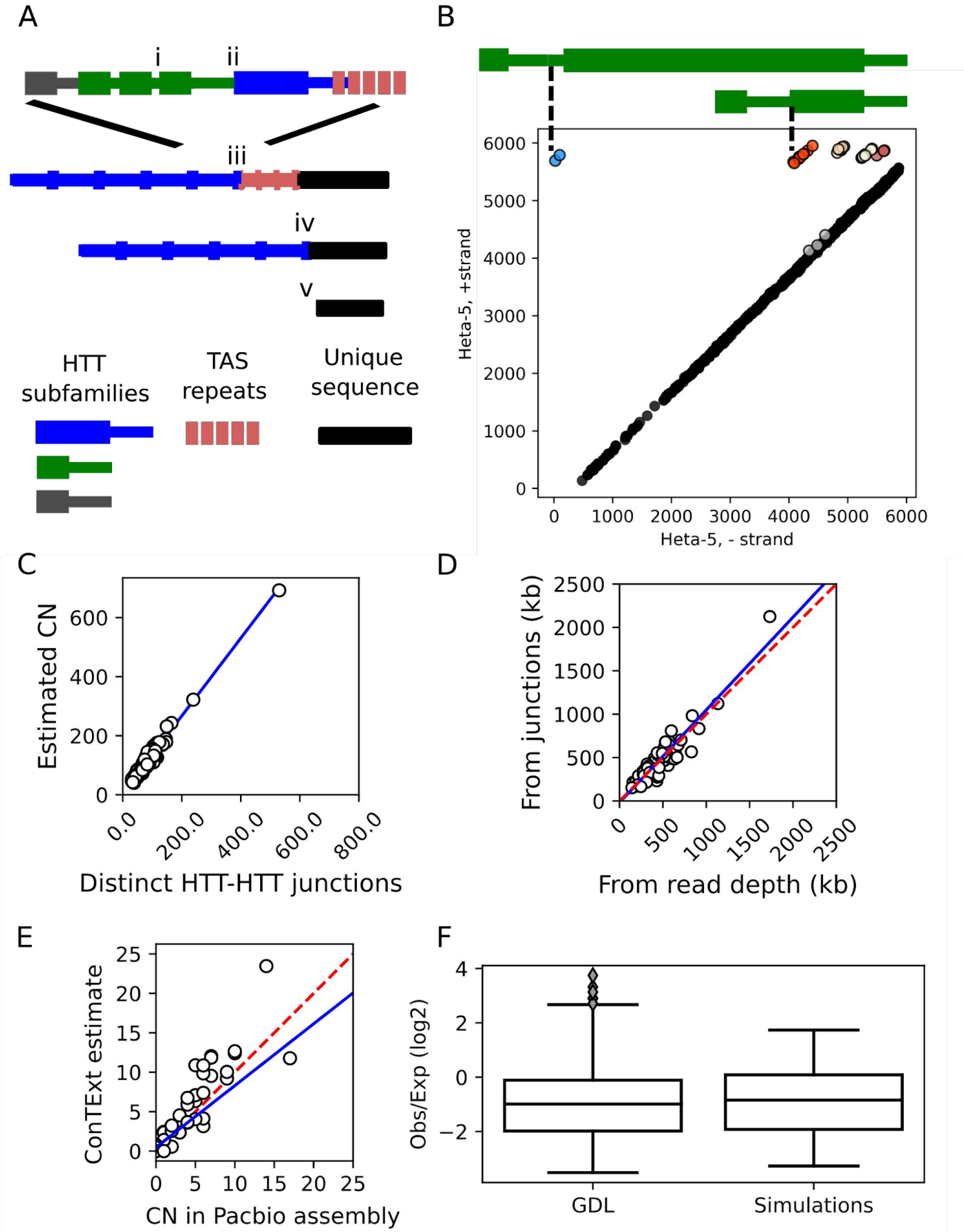

A
i
ii
iii
iv
v
HTT subfamilies
TAS repeats
Unique sequence
B
Heta-5, +strand
Heta-5, - strand
0
1000
2000
3000
4000
5000
6000
C
Estimated CN
Distinct HTT-HTT junctions
0.0
200.0
400.0
600.0
800.0
D
From junctions (kb)
From read depth (kb)
0
500
1000
1500
2000
2500
E
ConTExt estimate
CN in Pacbio assembly
0
5
10
15
20
25
F
Obs/Exp (log2)
−2
0
2
4
GDL
Simulations

Figure 2

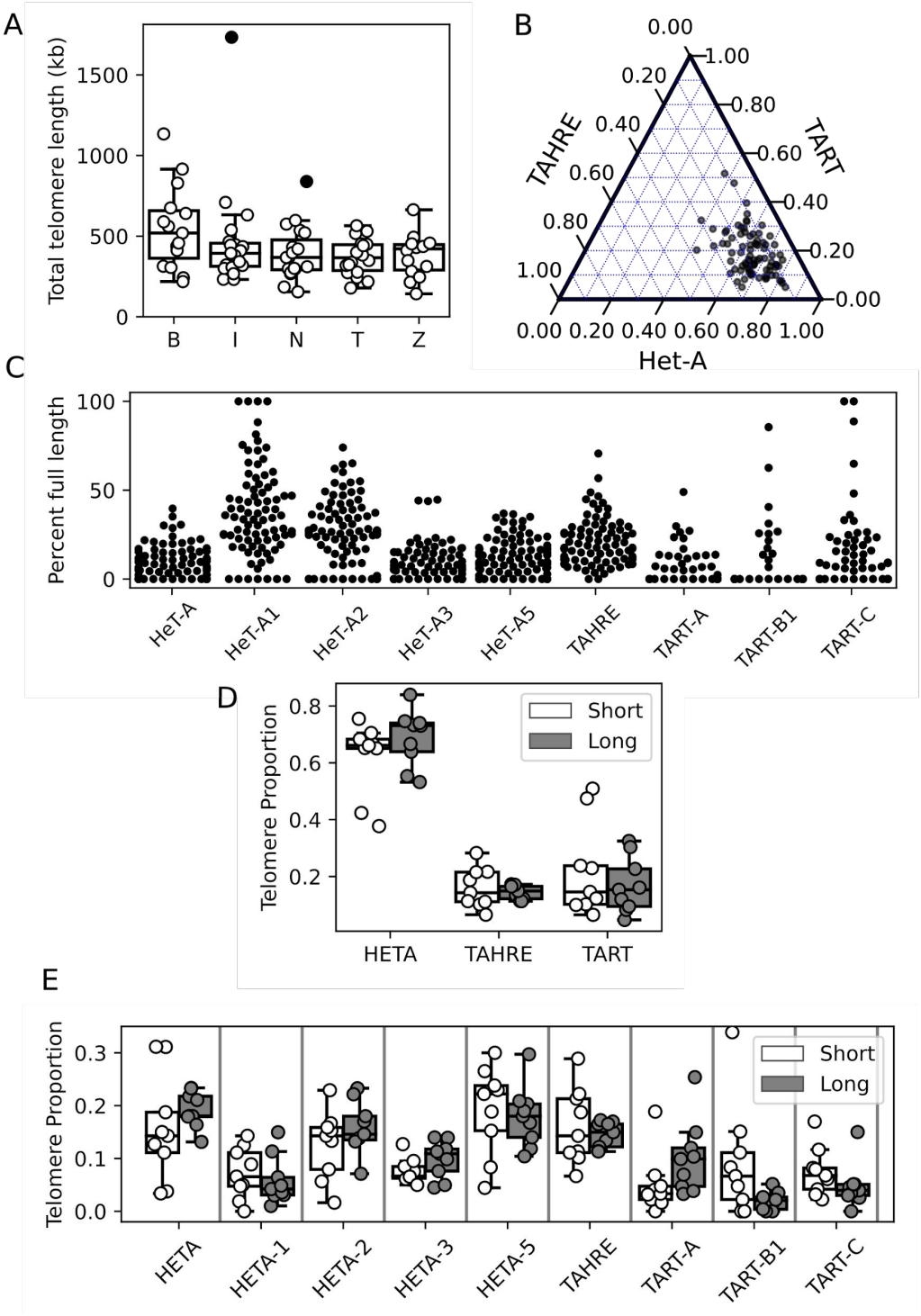
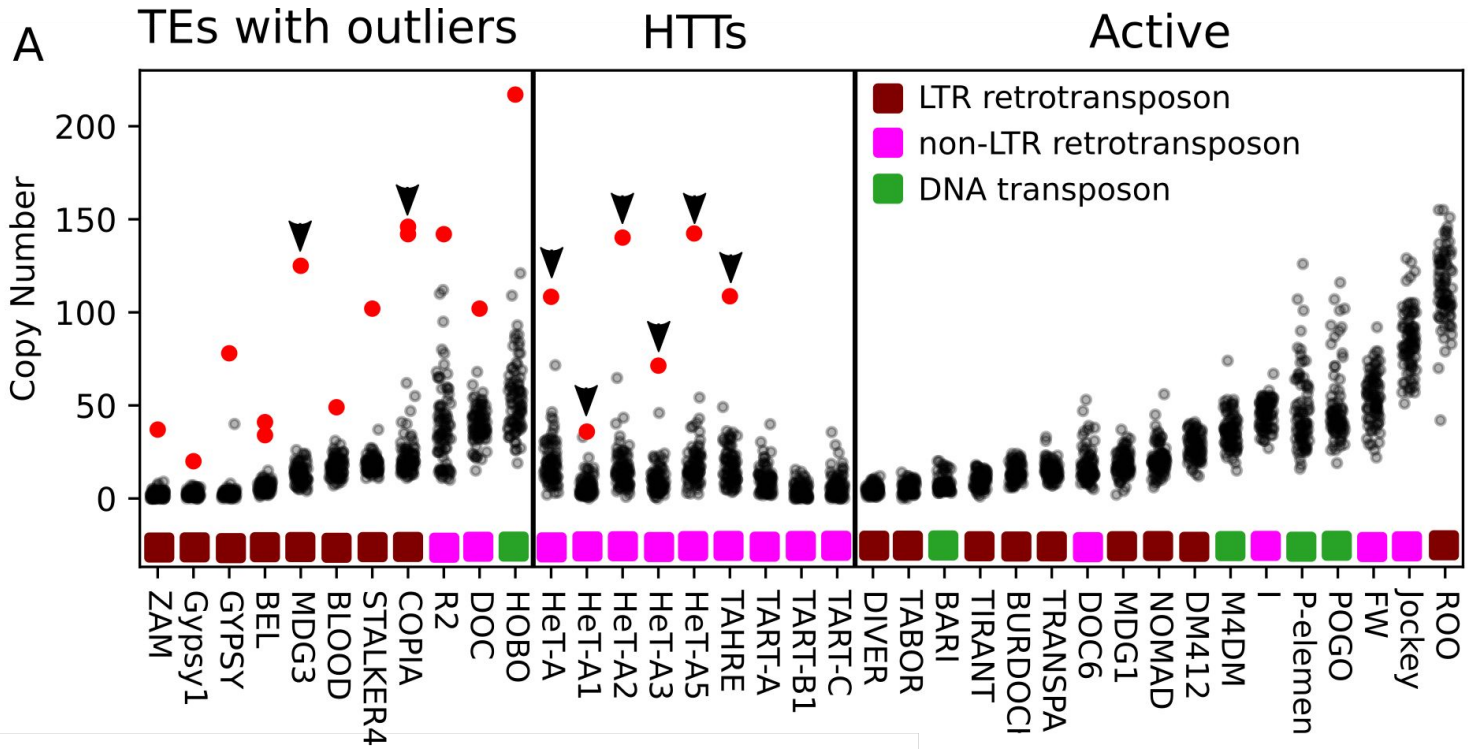
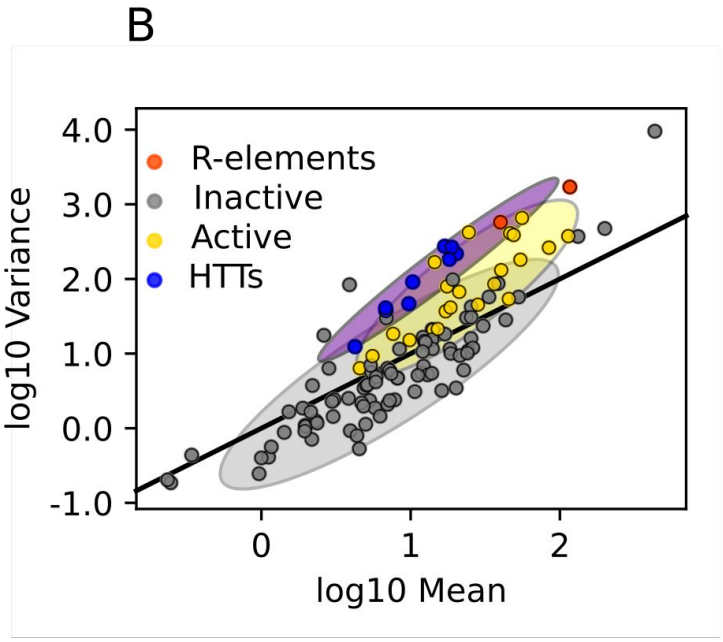

A
Total telomere length (kb)
1500
1000
500
0
B
I
N
T
Z
B
TAHRE
TART
Het-A
0.00
0.20
0.40
0.60
0.80
1.00
C
Percent full length
100
50
0
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C
D
Telomere Proportion
Short
Long
HETA
TAHRE
TART
E
Telomere Proportion
0.3
0.2
0.1
0.0
Short
Long
HETA
HETA-1
HETA-2
HETA-3
HETA-5
TAHRE
TART-A
TART-B1
TART-C

Figure 3


A
TEs with outliers
HTTs
Active
Copy Number
200
150
100
50
0
LTR retrotransposon
non-LTR retrotransposon
DNA transposon
ZAM
Gypsy1
GYPSY
BEL
MDG3
BLOOD
STALKER4
COPIA
R2
DOC
HOBO
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C
DIVER
TABOR
BARI
TIRANT
BURDOCI
TRANSPA
DOC6
MDG1
NOMAD
DM412
M4DM
I
P-elemen
POGO
FW
Jockey
ROO


B
R-elements
Inactive
Active
HTTs
log10 Variance
4.0
3.0
2.0
1.0
0.0
-1.0
0
1
2
log10 Mean

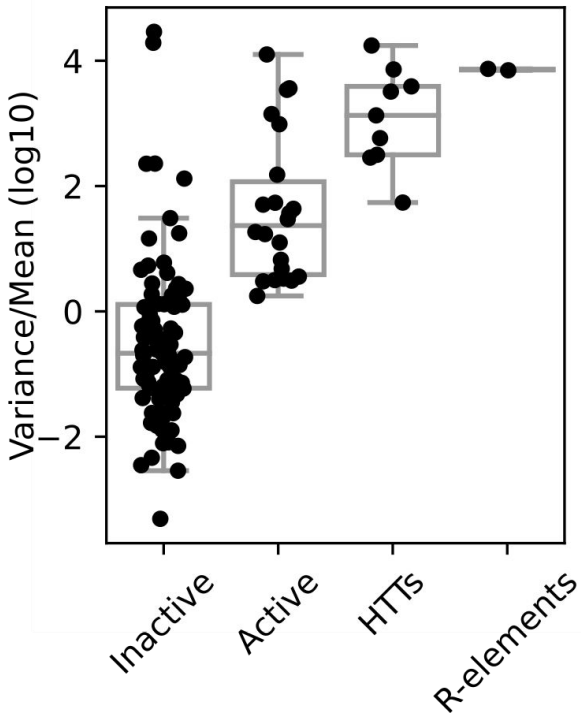

C
Variance/Mean (log10)
4
2
0
−2
Inactive
Active
HTTs
R-elements

Figure 4

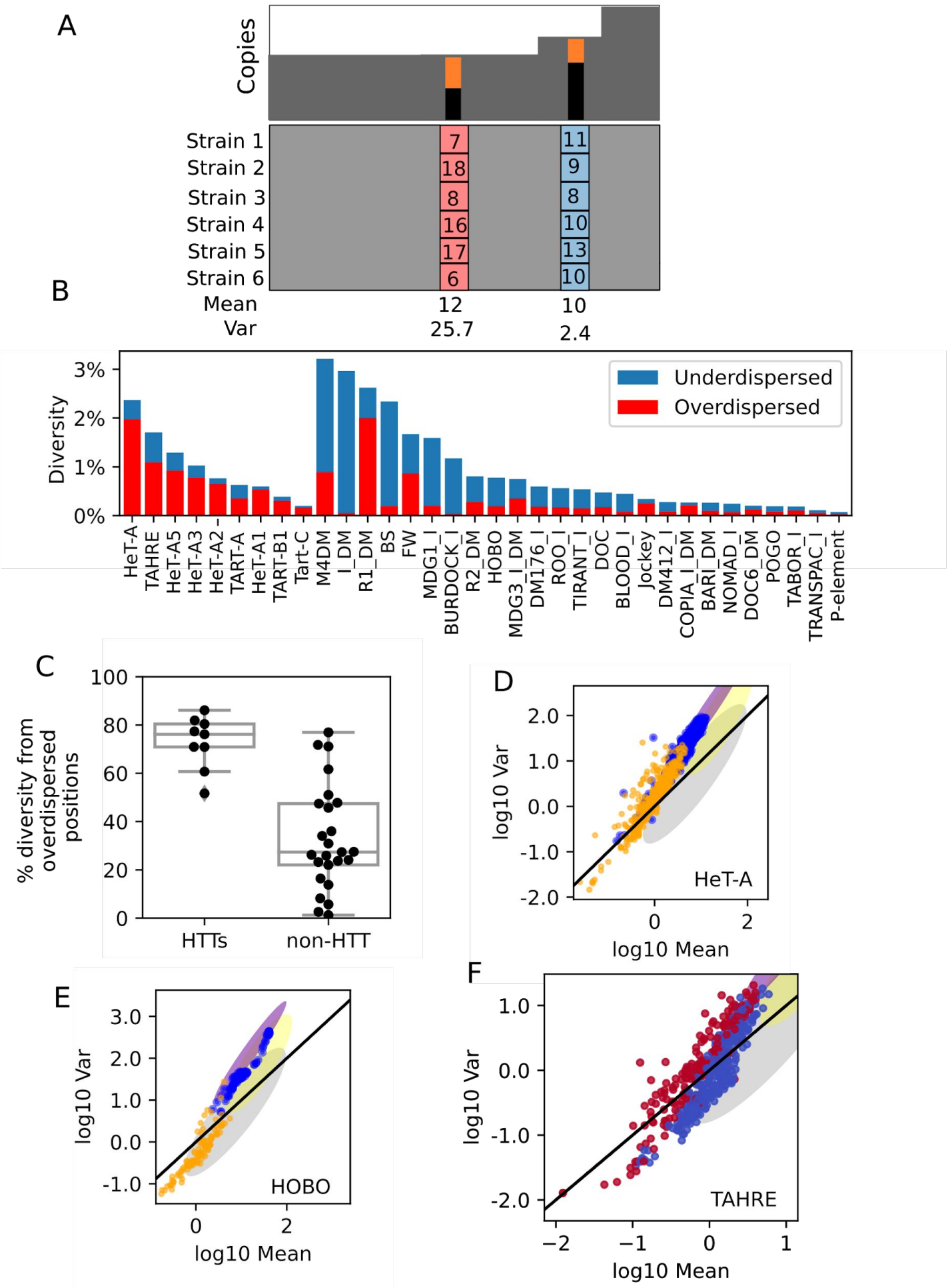

A
Copies
Strain 1
Strain 2
Strain 3
Strain 4
Strain 5
Strain 6
Mean
Var
7
18
8
16
17
6
11
9
8
10
13
10
12
25.7
10
2.4
B
Diversity
3%
2%
1%
0%
Underdispersed
Overdispersed
HeT-A
TAHRE
HeT-A5
HeT-A3
HeT-A2
TART-A
HeT-A1
TART-B1
Tart-C
M4DM
I_DM
R1_DM
BS
FW
MDG1_I
BURDOCK_I
R2_DM
HOBO
MDG3_I_DM
DM176_I
ROO_I
TIRANT_I
DOC
BLOOD_I
Jockey
DM412_I
COPIA_I_DM
BARI_DM
NOMAD_I
DOC6_DM
POGO
TABOR_I
TRANSPAC_I
P-element
C
% diversity from overdispersed positions
100
80
60
40
20
0
HTTs
non-HTT
D
log10 Var
log10 Mean
HeT-A
E
HOBO
F
TAHRE

Figure 5

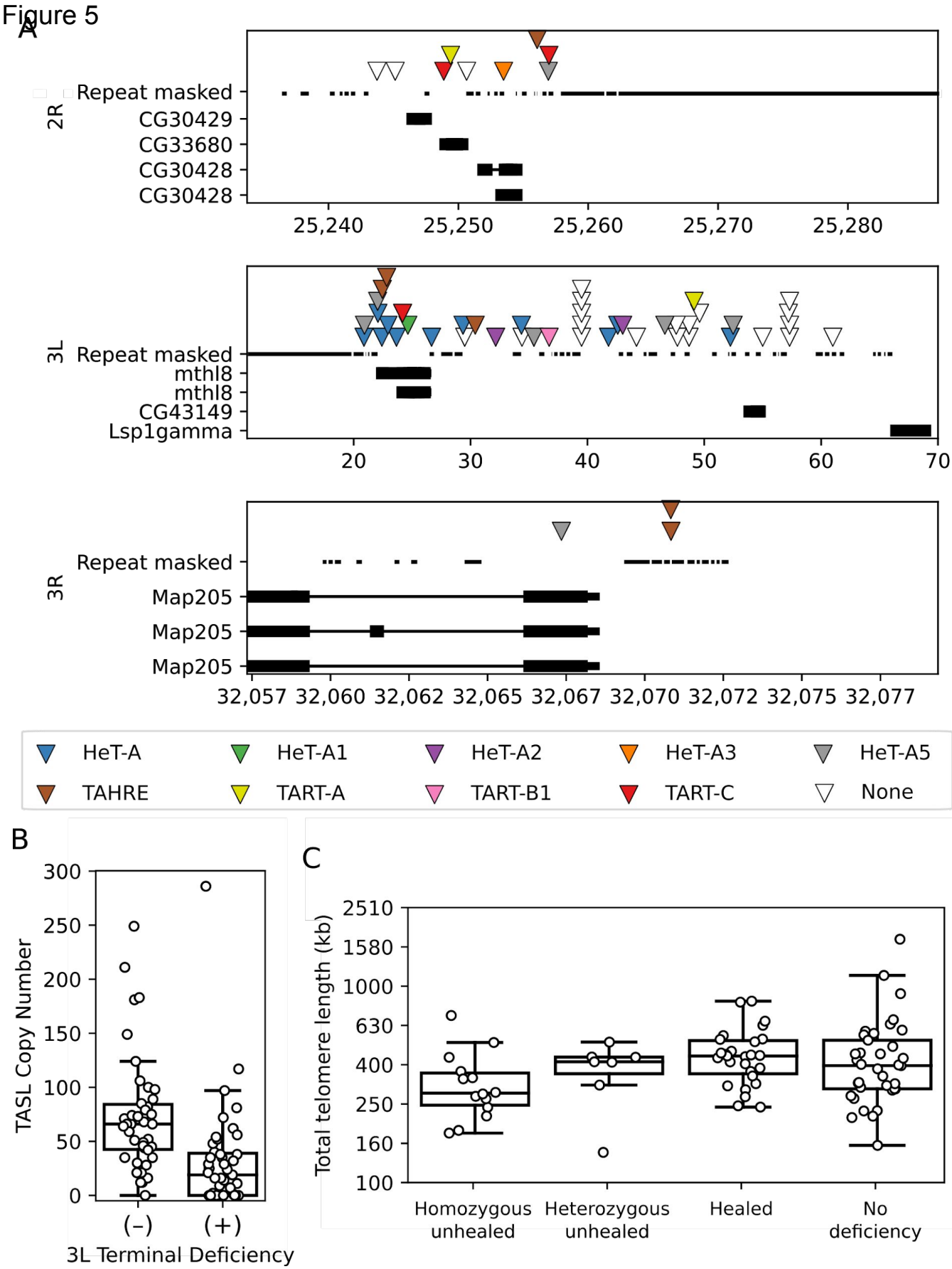

A
2R
Repeat masked
CG30429
CG33680
CG30428
CG30428
25,240
25,250
25,260
25,270
25,280
3L
Repeat masked
mthl8
mthl8
CG43149
Lsp1gamma
20
30
40
50
60
70
3R
Repeat masked
Map205
Map205
Map205
32,057
32,060
32,062
32,065
32,067
32,070
32,072
32,075
32,077
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C
None
B
TASL Copy Number
300
250
200
150
100
50
0
(–)
(+)
3L Terminal Deficiency
C
Total telomere length (kb)
2510
1580
1000
630
400
250
160
100
Homozygous unhealed
Heterozygous unhealed
Healed
No deficiency

Figure 6

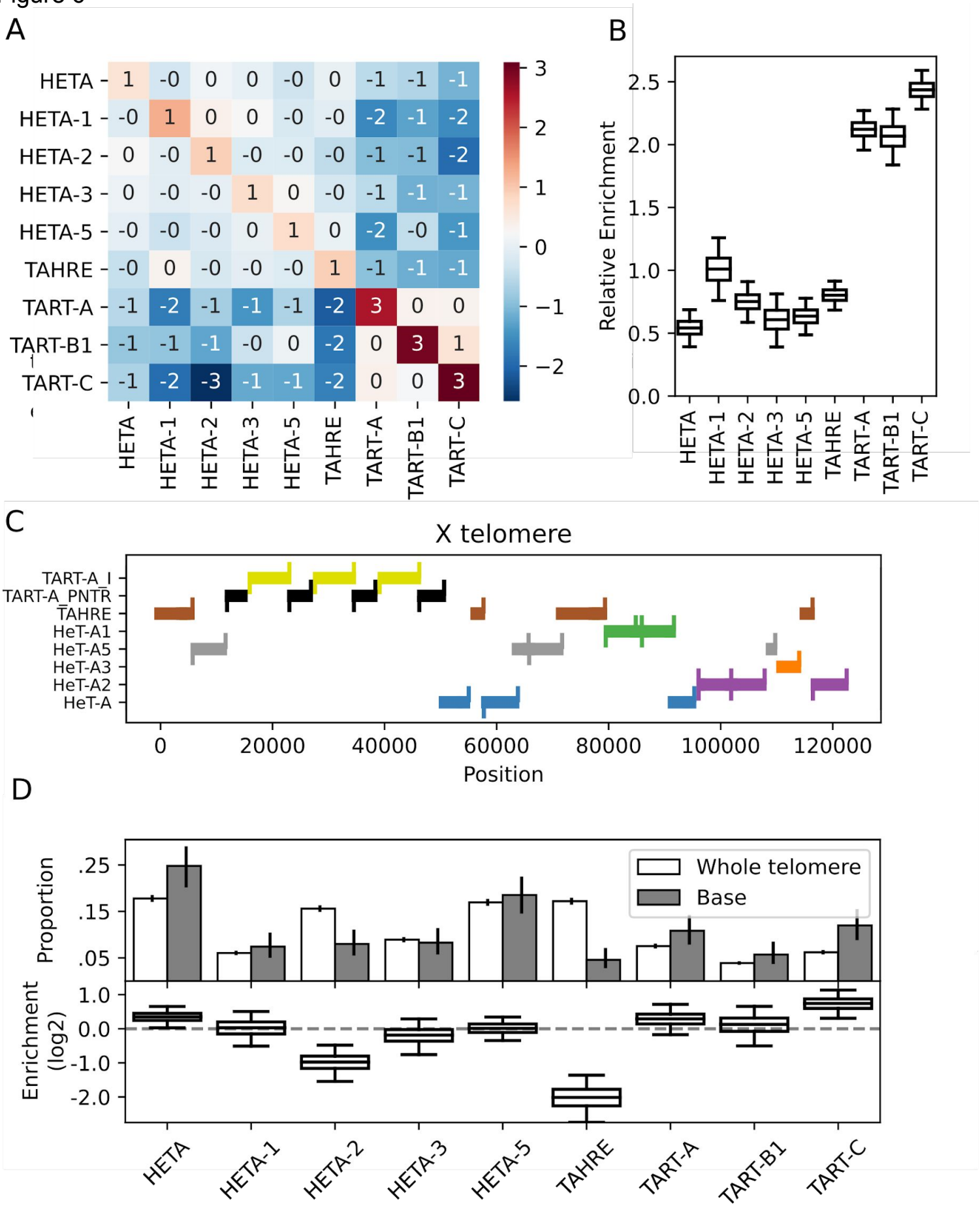

A
HETA
HETA-1
HETA-2
HETA-3
HETA-5
TAHRE
TART-A
TART-B1
TART-C
B
Relative Enrichment
C
X telomere
TART-A_I
TART-A_PNTR
TAHRE
HeT-A1
HeT-A5
HeT-A3
HeT-A2
HeT-A
Position
D
Proportion
Whole telomere
Base
Enrichment (log2)

Figure S1

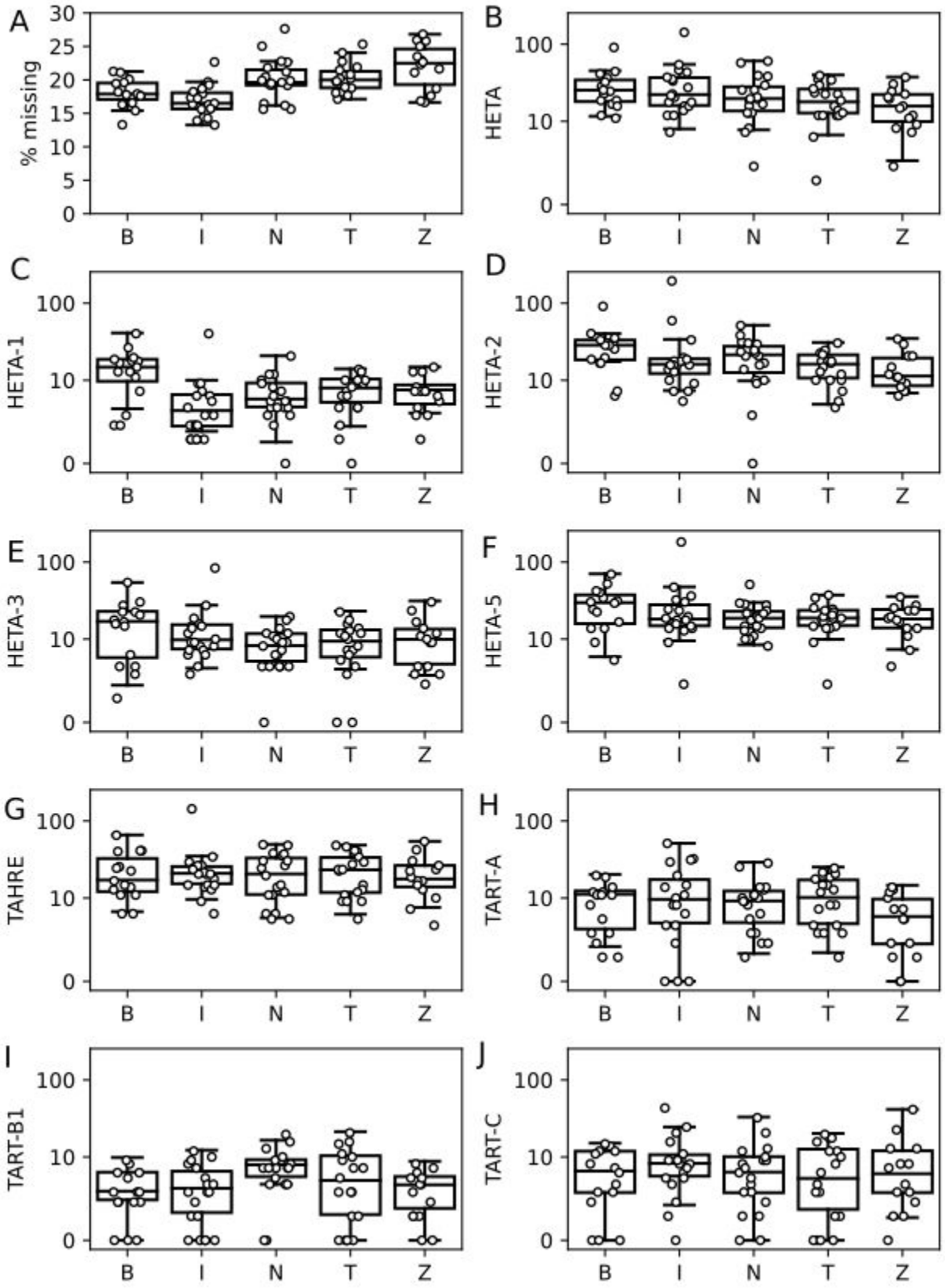

A
% missing
B
HETA
C
HETA-1
D
HETA-2
E
HETA-3
F
HETA-5
G
TAHRE
H
TART-A
I
TART-B1
J
TART-C
B
I
N
T
Z

Figure S2

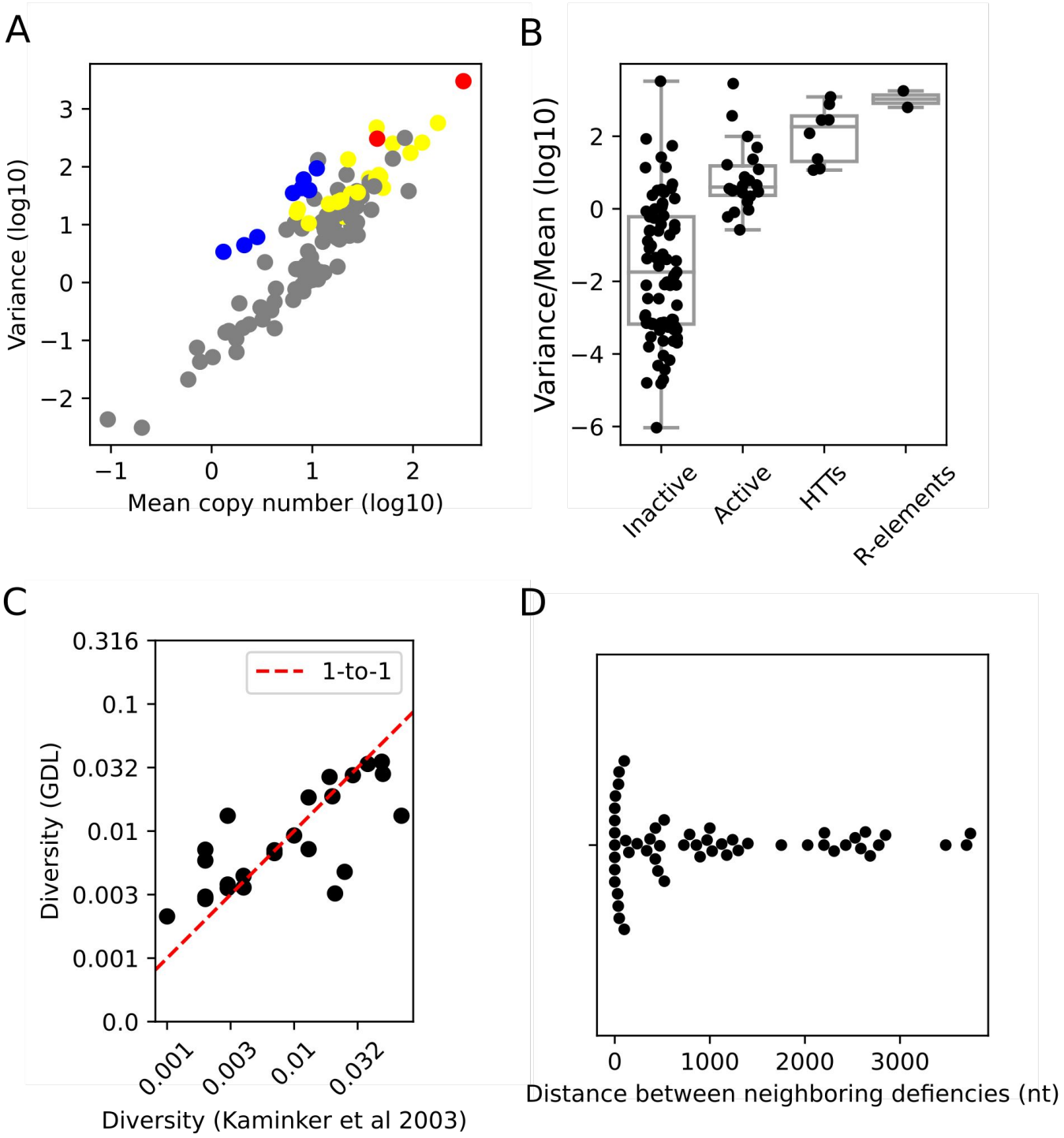
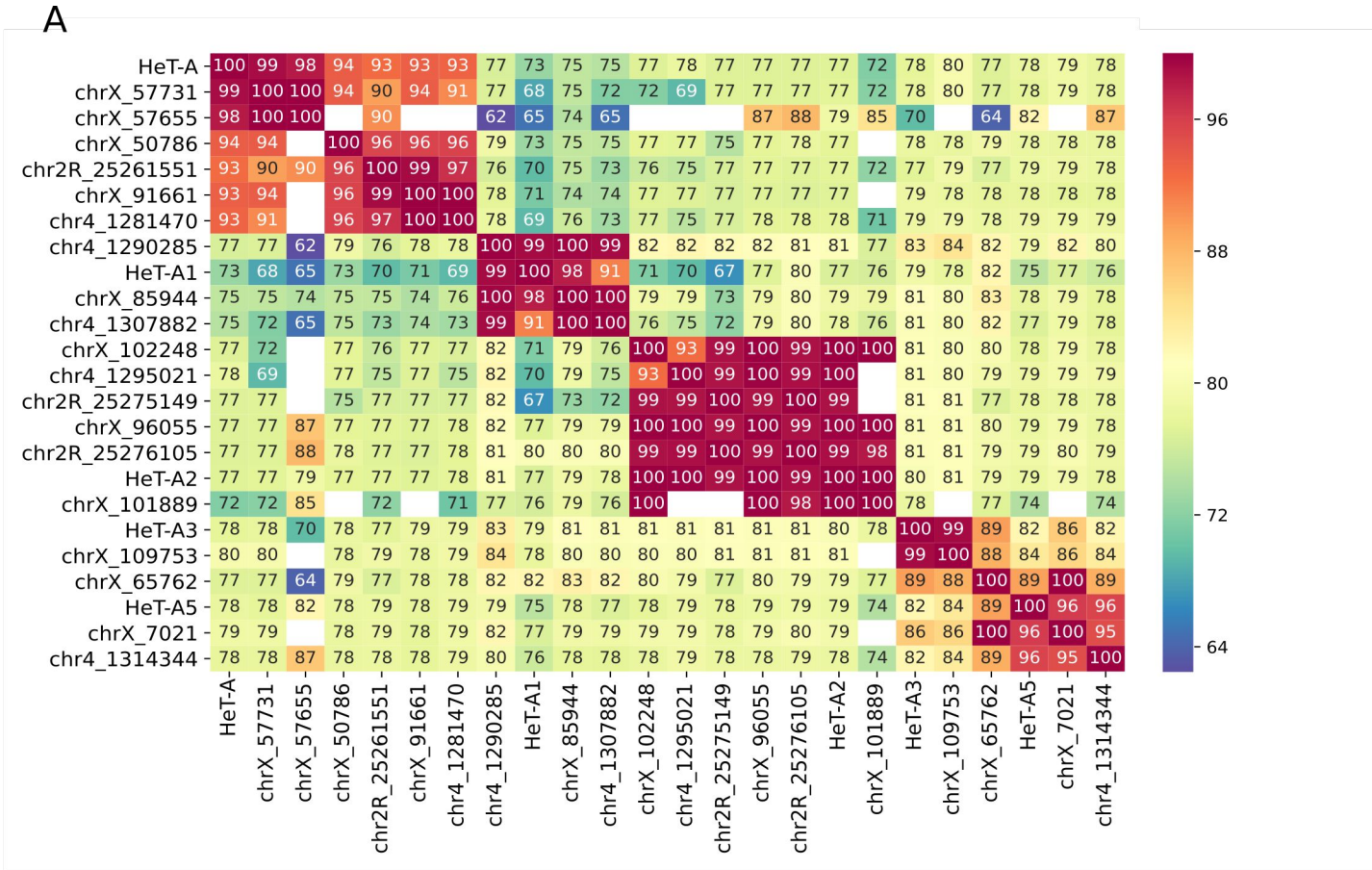
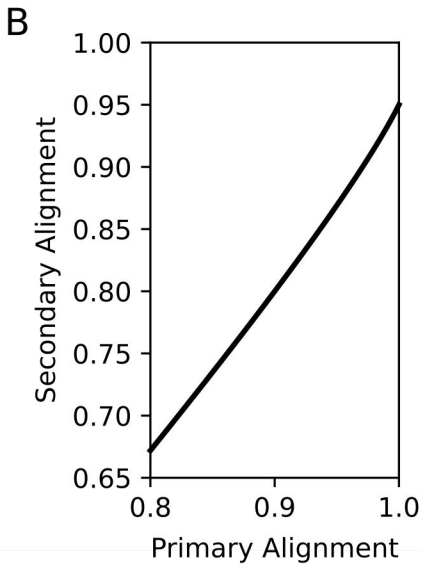
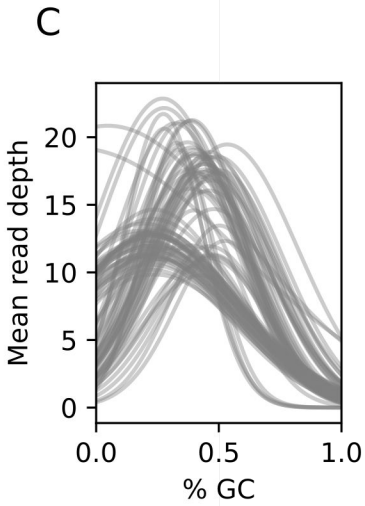
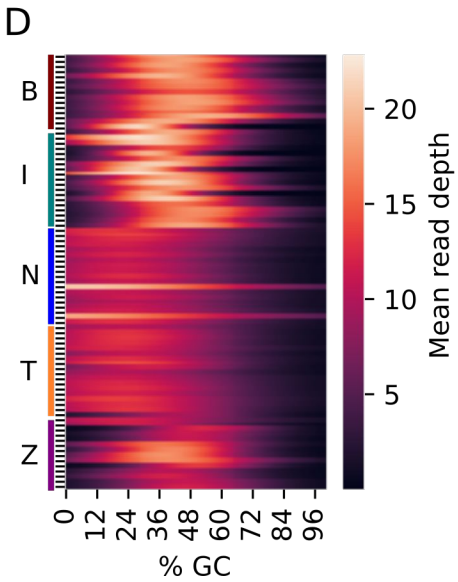

A
Variance (log10)
Mean copy number (log10)
B
Variance/Mean (log10)
Inactive
Active
HTTs
R-elements
C
Diversity (GDL)
Diversity (Kaminker et al 2003)
1-to-1
D
Distance between neighboring defiencies (nt)

Figure S3


A
HeT-A
chrX_57731
chrX_57655
chrX_50786
chr2R_25261551
chrX_91661
chr4_1281470
chr4_1290285
HeT-A1
chrX_85944
chr4_1307882
chrX_102248
chr4_1295021
chr2R_25275149
chrX_96055
chr2R_25276105
HeT-A2
chrX_101889
HeT-A3
chrX_109753
chrX_65762
HeT-A5
chrX_7021
chr4_1314344
96
88
80
72
64


B
Secondary Alignment
Primary Alignment
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.8
0.9
1.0


C
Mean read depth
% GC
20
15
10
5
0
0.0
0.5
1.0


D
B
I
N
T
Z
% GC
0
12
24
36
48
60
72
84
96
Mean read depth
20
15
10
5