Rapid evolution at the 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*

[#] *Corresponding author*

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

ABSTRACT

Background: 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. The conservation of this arrangement 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.

Results: We leveraged population genomic data to analyze nearly 10,000 HTT insertions in 85 *D. melanogaster* genomes and compared their variation to TE families evolving under more typical dynamics. We find that distinct 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. However, occasional 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. We further find that some HTT families may be activated by the erosion of whole telomeres, suggesting the existence of HTT-specific host control mechanisms.

Conclusions: Even the persistent telomere localization of HTTs may not reflect domestication, but rather a highly successful evolutionary strategy to reduce their impact on the host genome. We propose that HTT evolution is driven by multiple processes including domestication, genetic conflict, niche specialization, and telomere instability, with the latter two being previously underappreciated and likely predominant.

KEYWORDS (3-10 key words)

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

BACKGROUND

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 [1]. The presence of dispersed repeats further provides substrates for ectopic recombination, permitting large scale and potentially lethal genome arrangements [2]. The resulting conflict between TEs and the host genome can 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 [3,4]. 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 [5].

TEs can further alter these dynamics by adopting strategies that mitigate the mutational burden they impose on the genome without sacrificing their replicative success [6]. 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 [7]. Arthropod R-elements are an extreme example, as they only insert in the highly repeated ribosomal DNA (rDNA) genes [8]. 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 [9]. 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, 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 [10]. In Diptera, the telomerase

gene was lost [11] 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) [12,13]. HTTs have several unique features compared to other non-LTR retrotransposons that reflect their specialized function [14–16]. First, they are thought to only insert at the ends of chromosomes, resulting in head-to-tail tandem arrays. Second, they rely upon each other at several steps in their life cycles. *HeT-A* and *TART* 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 [17,18]. Additionally, *HeT-A* may provide telomere specificity to *TART* and *TAHRE* through its Gag-like protein encoded by ORF1 [19]. Finally, *HeT-A* is non-autonomous [20–22] — 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, with new families and subfamilies having evolved across the genus and even among closely related species of the *melanogaster* subgroup [23,24]. In addition, features typical of a family in one species, for example 3'-promoters or terminal repeats, may be absent in other species [15]. 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 microevolutionary scale in ways distinct from more typical TEs. Yet, a comprehensive population genomic study of HTT variation remains to be undertaken, which could provide the power to delineate the forces shaping their contemporary evolution. We consider four potential forces shaping HTT variation at the micro-evolutionary scale.

First, selection acting at the level of the organism has likely played an important role in both the evolution and dynamics of the HTTs, in ways distinct from more typical TEs. The fact that the HTTs comprise the Drosophila telomeres should favor their long-term persistence and transposition activity. Organismal selection may also favor increased levels of domestication. In addition, selection might promote new HTT variants that increase organismal fitness in other ways, for example by reducing telomere instability or by having pleiotropic beneficial effects on chromosome structure. Further, unlike

other TEs, the genome is under pressure to regulate rather than wholly suppress their activity. Coordinating telomere elongation to prevent either complete erosion or overextension is a feature of more typical telomerase-based systems [25,26], and mutational studies suggest that Drosophila has recruited the piRNA and heterochromatin maintenance pathways typically involved in TE suppression into this role [27]. Host factors that interact with the telomeres, such as the capping proteins that assemble onto them and the subtelomeric sequence which directly abut them, may have also evolved properties that regulate HTT activity [28].

Second, selection can act below the level of the organism—that is, on the HTTs themselves because they are active transposons and thus replicative entities in their own right. Both the continual erosion of the telomeres and the potential fitness impacts of excessively long telomeres imply that space in the telomere is a limiting resource for which all HTT families are competing. HTT variants that are better able to increase in copy number will outcompete those that cannot. One simple means of increasing in copy number is for the HTT to escape regulation by the host genome, and thus conflict between the HTTs and genome (or among the HTTs) is an ever present possibility. Indeed, like their feral cousins, studies using mutations in HTT regulators suggest that they are capable of taking over given the opportunity [27]. Similar to the rapid evolution of the Drosophila HTTs [23,29], some telomere-associated proteins show high evolutionary rates and signatures of positive selection [28,30]. These evolutionary analyses led Lee et al. to propose that HTTs are in genetic conflict with their Drosophila hosts [30]. Complete resolution of conflict with the genome likely requires somehow separating the reverse transcription machinery that extends the telomere from the HTTs [9,31].

Third, the dynamic and unstable nature of telomeres has the potential to shape HTT evolution. Telomeres continually experience terminal erosion due to the end-replication problem. In addition, their tandemly arrayed structure facilitates the amplification and deletion of sequence through unequal exchange events, and chromosome breaks near the terminus can cause complete loss of a telomere [32–34]. The expansions, contractions, and deletions of telomeric sequence also have the potential to heighten the rate at which polymorphisms fix within HTT families and to increase the potential for the

extinction of lineages. The net effect of these processes is to facilitate rapid sequence change at telomeres in a manner driven by mutational processes rather than selection.

Fourth, the trade-off between reducing their impact on the genome versus inhabiting an unstable locus has likely also shaped HTT evolution. Compared to much of the genome, the telomere is a “safe harbor” that allows the HTTs to accumulate without disrupting essential host genes and also minimizes the potential effects of ectopic recombination [6]. The tension between this and the inherent instability of the telomeres disposing HTTs to high rates of deletion suggests that the persistence of this strategy reflects the outcome of an evolutionary trade-off. Indeed, some of the peculiar features of HTTs, such as 3’-end promoters, are likely adaptations to cope with this instability. Another possible outcome of evolutionary trade-offs (as well as of genetic conflict with other HTTs) would be to escape the telomere. Surprisingly, despite transposing to chromosome ends, *TART* and *TAHRE* each retain a potentially functional endonuclease domain [35], which raises the possibility that the HTTs could generate double-strand breaks and insert randomly as do other non-LTR retrotransposons. While limited surveys suggest that HTTs exclusively insert at telomeres in *D. melanogaster,* the *melanogaster* subgroup species *D. rhopaloa* has many copies outside the telomere [24].

The extent to which the HTTs vary among species should reflect the compounded effects and interaction among these evolutionary processes playing out over millions of years. However, the timescales involved and the complexity of this array of processes makes it challenging to understand which processes are responsible for which particular aspects of HTT rapid evolution. A clear picture of within-species HTT variation offers the opportunity to understand how these processes play out over shorter timescales where their impacts on HTT evolution can be more clearly distinguished. Here, we leverage available population genomic data and the ConTExt pipeline [36] to comprehensively analyze HTT sequence and copy number polymorphism in the Drosophila Global Diversity Lines (GDL) [37], encompassing nearly ten thousand HTT insertions in 85 strains of *Drosophila melanogaster*. We compare the HTTs to other TE families, including those that are similarly restricted to unstable genomic loci. Within this framework we ask whether the patterns of HTT copy number, organization, and sequence variation require explanations beyond the standard conflict between HTTs and the host

genome and, if so, whether this reflects their remarkable symbiosis with the genome or is instead the natural consequence of occupying the unstable telomeres.

RESULTS

Telomeres are highly dynamic in Drosophila melanogaster

Drosophila melanogaster telomeres are comprised of three non-LTR retrotransposon families *HeT-A*, *TAHRE*, and *TART*, which in our analyses are further subdivided into five *HeT-A* subfamilies (*HeT-A*, *HeT-A1*, *HeT-A-2*, *HeT-A3*, *HeT-A5*) and three *TART* subfamilies (*TART-A*, *TART-B1*, *TART-C*) (see methods for details). For clarity we use the terms "*HeT-A* family" and "TART" when referring to the families. Telomere length has previously been reported to vary by two orders of magnitude [38]. However, that survey was based on qPCR estimates of abundance using a primer set that was designed to detect a single consensus *HeT-A*, in a single North American population, so the absolute size range of telomeres in *D. melanogaster* remains unknown.

To gain a more comprehensive picture of the composition of *D. melanogaster* telomeres, we consider how the size and composition of telomeres vary among strains in terms of the sequence abundance and copy number of all HTT families. We define sequence abundance as the number of nucleotides each HTT family contributes to the genome; this provides an estimate of the total telomere length that reflects both the copy number and the lengths of the HTT insertions. We independently estimate copy number from the junctions we identify with ConTExt [36], which provides a more direct picture of HTT dynamics than does telomere length alone. In brief, ConTExt aligns repeat derived reads to repeat consensus sequences and uses paired-end information to identify the junctions between repeats and neighboring sequence. In the case of HTT insertions, the neighboring sequence should generally be another HTT insertion or subtelomeric sequence, and all such junctions should involve the intact 3'-end of a HTT. The range of 5'-truncations resulting from incomplete reverse transcription, which is frequent in all non-LTR retrotransposons, provides additional power to distinguish independent insertions, allowing us to identify insertions and estimate TE copy number within the telomeres.

The median total telomere length in the GDL is 400 kb with a median of 95 HTT insertions per genome, corresponding roughly to 50 kb and 12 insertions per each of the eight telomeres (Figure 1A, Table S2). The majority of insertions belong to the *HeT-A* family (median of 58 insertions) with medians of 15 and 17 insertions belonging to *TAHRE* and the *TART* family, respectively (Figure S1A); the relative abundance of *HeT-A* and *TART* family elements is roughly equal to previous estimates [12,39]. The approximately equal abundance of *TAHRE* and *TART* family elements contrasts with an analysis of the 4R telomere in the reference strain, which contains 8 *TART* family but no *TAHRE* elements [39]. This may reflect a peculiar structure of the 4R telomere, because analysis of 3 other telomeres in the reference strain identified 4 *TAHRE* elements [12]. Other strains also show *TAHRE* on a subset of telomeres [40]. The majority of HTT insertions are incomplete, with about 80% of *HeT-A* family and *TAHRE* insertions and 93% of *TART* family insertions being truncated to various degrees, though this proportion varies considerably among strains (Figure 1B). There is some population structure to telomere length, with the estimate of mean telomere length in the Beijing population being 100-140kb longer than the other four populations (Ithaca, Netherlands, Tasmania and Zimbabwe), corresponding to 25-40 more insertions per strain on average. An increase in the number of *HeT-A* family elements drives this, especially the *HeT-A*1, *HeT-A*2, and *HeT-A*5 subfamilies (Figure S1).

Telomere length displays appreciable variation among individuals, with an interquartile range of 200kb and 56 HTT insertions (Table S2). In addition, there is clear potential for extreme gains and losses of telomere sequences, which are observed as significant outliers. Several strains have much shorter total telomere length, with the shortest (T14) being 143 kb and having 43 insertions (average of 17 kb or 5 insertions per telomere). Some of the shortest strains are among those that contain unhealed telomere deficiencies (see “Terminal deletions are common” below). At the opposite extreme, in I01 the telomeres have expanded to 1.7 Mb of sequence, corresponding to 673 insertions (average of 212 kb or 84 insertions per telomere). The range of telomere lengths is similar to that found in an earlier study of several strains (including the long-telomere strain *GIII*) that used a different approach (Southern blotting) [39].

HTT family composition is similar in long and short telomere strains

Comparative description of telomere composition across individuals is scarce. We described above that the Beijing population has elongated telomeres, and that this is driven by expansions of particular *HeT-A* subfamilies. Telomere elongation must to some extent be under host control, and such expansions are suggestive of escape from host regulation. If a specific HTT escapes host control and drives telomere elongation across lines, this subfamily should be over-represented in long telomere strains compared to short telomere strains. We analyzed telomere composition by comparing HTT family and subfamily copy number and relative proportions between long (90th percentile, > 163 HTT insertions) and short (10th percentile, < 54 insertions) telomere strains. At the family level, there is no apparent difference in telomere composition between short and long telomere strains (Figure 1C). As the copy number of *HeT-A* family, *TAHRE* and *TART* elements increases in long telomere strains, so does their relative proportion, resulting in similar telomere composition. However, some differences arise at the subfamily level (Figure 1D). In long telomere strains, the copy number of *HeTA-1*, *HeTA-5*, *TART-B1* and *TART-C* increases less substantially than other sub-families, resulting in a reduced representation in long telomere strains. Altogether, these observations suggest that telomere length is under host control, that HTT sub-families respond differently to a more or less transposition-permissive state, and that long telomere strains do not result from runaway transposition of specific HTT subfamilies.

Copy number expansions are not specific to HTTs

The pronounced copy number expansions of HTTs we observed in some strains raises the question of whether this is specific to the HTTs. We found that large TE expansions are common in GDL strains for both HTTs and non-telomeric TEs (Figure 2A). Interestingly, in line I01, which has the greatest number of HTTs, *Copia* and *MDG3* (both LTR retrotransposons) have also considerably increased in copy number (Table S1). Total TE copy number is also highest in strain I01, but across all lines there is no strong general tendency for strains with non-HTT copy number expansions to have longer telomeres. Three non-HTT TE families, *Gypsy*, *Gypsy1*, and *Zam*, that were defined as inactive under our criteria (see methods) are present in few copies in most strains but expanded to 78, 20 and 37 copies,

respectively, in single strains. For the most part, the copy number expansions all occurred in different strains, and we observe no general relationship between the occurrence of TE expansions and HTT copy number, with the exception of the aforementioned strain I01 (Figure 2B). It is noteworthy that in this strain, all of the *HeT-A* subfamilies and *TAHRE* have expanded considerably in copy number, but not the *TART* elements, suggesting differential responsiveness among the HTTs to whatever caused the expansion. We suspect that most or all of the expansions observed occurred during propagation in the laboratory, because in nature the new insertions resulting from such an expansion would be at low population frequency and rapidly become unlinked from each other due to independent assortment and recombination. In contrast, new insertions can more easily reach high population frequency in the laboratory due either to the small population size or to the fixation of permissive variants in piRNA clusters or piRNA pathway genes. We conclude that like any other active and selfish TE, HTTs are primed to take over given the opportunity.

HTT copy number variation reflect its highly dynamic environment

Given that HTT abundances are highly dynamic across the GDL panel, we wanted to assess comprehensively how the relationship between average copy number and variability compares between HTTs and other TEs. We expected that the magnitude of between-line copy number variation ought to increase with mean copy number. Active elements provide a reasonable baseline because they are in conflict with the genome via transposition. However, the tandem nature of the telomeres additionally permits copy number expansions and contractions through unequal exchange, which may heighten the variability in copy number. Two other TE families, the *R1* and *R2* elements, reside in a similarly dynamic environment: the tandemly arrayed ribosomal RNA genes (rDNA). The *R1* and *R2* elements therefore provide a comparison of the HTTs with TEs whose copy number evolves both by transposition and unequal exchange, but which provide no essential function.

We sought to leverage population genetic theory for the relationship between the mean and variance of TE copy number distributions to interpret the empirical relationship we observed in the GDL. [4] highlight that the variance in TE copy number should depend only on the mean and variance of the

site frequency spectrum of TE insertions and the degree of correlation in the presence/absence of distinct insertions (linkage disequilibrium). Following their approach, we consider that whatever dynamics and processes underlie the copy number evolution of a TE family within a species, each occupiable locus contains an insertion in some fraction of individuals in a population. The set of loci containing insertions in an individual genome is sampled from this larger population of insertions, each insertion allele being present with probability equal to its population frequency. Assuming that the number of *de novo* insertions per generation is small relative to the number of preexisting insertions, the copy number in an individual genome is then the sum of Bernoulli trials with potentially different success probabilities. If these trials are independent and have identical population frequencies, TE copy number will be Poisson distributed and the variance will be equal to the mean. If they are independent and but the population frequencies vary, this can be described with a Poisson-binomial distribution, and the variance will be less than the mean. In reality the presence/absence of insertions may not be independent of each other due to a variety of factors, including genetic linkage, selection, and demographic history and population structure, collectively termed linkage disequilibrium, which will increase the variance of TE copy number. These factors are synthesized in the mean-variance relation

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

where $\mu$is the population mean copy number, T is the number of occupiable loci, $Var(x_i)$is the variance in population frequency of TE insertions and $D_{ij}$is the linkage disequilibrium of insertions [4].

We examined the mean-variance relationship of TE families in the GDL, considering two different estimates of TE copy number (see Methods) which gave similar results. We found that inactive TE families are generally underdispersed relative to a Poisson expectation (Figures 3A, B and S2 A,B). This is consistent with theory, as older insertions should have more variation in their population frequencies than recent insertions, leading to an underdispersed copy number distribution (due to $Var(x_i)T > 0$). In contrast, active TEs not only show greater copy number variation than inactive TEs, but are generally overdispersed with respect to the Poisson expectation. This greater than expected variability may reflect linkage disequilibrium or population structure in the success of the TE family's

invasion. We consider continued copy number evolution in the lab stocks (itself a form of population structure) a likely explanation. But compared to other active elements, HTT copy number is even more variable (Figure 3A, B and S2A, B), which could reflect either heightened conflict or instability of the genomic regions they occupy, coupled with high linkage disequilibrium ($4\sum_{i<j} D_{ij} > 0$). That is, while other TE families are dispersed across the genome, all HTT insertions are clustered at eight tandem arrays, such that many insertions will be transmitted or lost together. Notably, the R-elements also display elevated variability comparable to the HTTs (Figure 3A, B and S2 A,B). We suggest therefore that the similarly heightened copy number variation displayed by the HTTs and R-elements reflects both the dynamic nature and tight linkage of the genomic environments caused by the tandem array organization characteristic of these families.

The high sequence diversity of HTTs is driven by active variants

Sequence evolution of TEs determines the properties of their encoded proteins and internal regulatory sequences as well as their potential to escape from host suppression. HTTs are highly divergent among species, which has been suggested to reflect continued conflict with their host genome despite domestication [24,30]. We propose that the instability of their genomic niche might also contribute to HTT sequence variation. The heightened copy number variation we observe for the HTTs (and *R*-elements) ought to increase the stochasticity with which lineages go extinct, allowing sequence polymorphism to fix more rapidly in them. We sought, therefore, to understand how the within-species sequence diversity of the HTTs compares to that of other active TEs, especially the rDNA-restricted *R*-elements.

We find that the HTTs display high levels of sequence diversity, though several other active families of TEs have comparable levels (Figure 3C). Such sequence diversity could reflect heterogeneity among the transposition-competent elements. Alternatively, the active elements might be relatively homogeneous, but distinct in sequence from older inactive insertions that are relics of prior invasions. This alternative scenario applies to TEs such as *I*-element and *Hobo*, which have invaded *D. melanogaster* multiple times and left both recently active and degenerate copies from past invasions

[41,42]. The shape of a phylogeny from complete sequences of all TEs in the GDL would easily distinguish these scenarios [43]. However, we cannot reconstruct individual elements, much less their phylogeny, with short reads. To circumvent this challenge, we developed an approach to leverage our observation that the copy number variation of active TE families tends to be overdispersed and that of inactive families tends to be underdispersed (Figure 3A, B). We reasoned that high diversity at positions where both the major and the most abundant minor alleles display overdispersed copy number variation is suggestive of heterogeneity among the active elements of the family. On the other hand, if the major allele is overdispersed but the minor allele is underdispersed, this instead suggests that the active elements are homogeneous at that position, with a distinct population of inactive copies that are divergent from the active lineage.

Consistent with a population of heterogeneous active elements, we find that both the minor and major alleles of the HTTs display the same mean-variance relationship we observed for HTT copy number and tend to be overdispersed (Figure 3C, D, Figure S3). In contrast, for most of the other TE families with high sequence diversity including *Hobo* and *I*-element, the variance of the minor allele tends to be shifted down from that of the major allele and is generally underdispersed (Figure 3E, F). We suggest that this pattern of sequence diversity of non-HTTs is largely driven by degenerate inactive copies, with only a small fraction of positions in most elements reflecting divergence among active elements. *TAHRE* is one exception to the tendency of the HTTs to show mainly overdispersed minor alleles, as it appears to have a subset of minor alleles displaying underdispersion, suggesting the presence of inactive copies outside of the telomere (Figure 3G; see also "*HTTs insertions are restricted to the telomeres*" below). To investigate whether this heightened diversity observed in the HTTs is characteristic of TEs residing within unstable genomic regions, we consider the sequence diversity of the *R*-elements. However, while *R1* also shows high levels of diversity, this is likely driven by the presence of hundreds of tandemly arrayed *R1*-elements evolving independently of the active elements [36,44]. A cleaner comparison is with *R2*, which does not form tandem arrays. Unlike the HTTs, *R2* displays considerable variation consistent with inactive and degenerate copies, which likely reflects accretion from the edge of the rDNA array. This accretion of degenerate and fragmented copies of repeat units is

commonly observed at the edge of satellite arrays [45]. The absence of accreted HTT relics may simply reflect the action of telomere erosion, which readily deletes telomeric sequence, including sequence accreted into the subtelomere. Nonetheless, *R2* displays lower levels of sequence diversity among active copies, which may indicate that even for TEs within unstable niches, the HTTs are evolving rapidly. We conclude that the high sequence diversity of HTTs is driven by active variants whose appearance and maintenance is favored by the unstable genomic niche they occupy.

HTT transposition is restricted to the telomeres

While HTT transposition has been apparently restricted to the telomeres for millions of years, the retention of intact endonuclease ORFs in *TART* and *TAHRE* suggests that HTTs could potentially transpose elsewhere in the genome [35]. Transposition of HTTs outside the telomere might reflect selfish HTT behavior that is in conflict with the host [24]. Non-telomeric *HeT-A* family and *TART* sequences are found on the heterochromatic Y chromosome of *D. melanogaster* and other Drosophila [46–49]. However, Berloco et al. [49] noted that "these non-telomeric tandem repeats differ from the telomeric elements in at least two respects: they are interspersed with other repetitive sequences and they are not always oriented in the same direction" and postulated that chromosomal rearrangements rather than non-telomeric insertions were responsible. These structural differences suggest that non-telomeric HTT copies may result from gene conversion, capture by other TEs, or rearrangements, rather than by non-telomeric insertion [46,47].

There are several fragments of HTTs found throughout the euchromatic arms of the *D. melanogaster* reference genome; their short lengths (several hundred base pairs or smaller) suggest that they either arose from gene conversion or represent ancient insertions that have largely degraded. Non-telomeric insertions were recently reported as common, however, in the *melanogaster* subgroup species *D. rhopaloa* [24]. We therefore used ConTExt to search the *D. melanogaster* GDL for low-frequency HTT insertions in euchromatic regions as well as in other repeats. Because sequencing artifacts could easily create the appearance of non-telomeric insertions, we required that both the 3' and 5' junctions be identifiable for junctions outside the subtelomere. We found no junctions beyond the subtelomeres that

satisfied these criteria, other than short fragments, some of which are present in the reference genome. We conclude that non-telomeric euchromatic insertions of HTT elements are negligible in *D. melanogaster*.

However, several of our observations suggest there may be copies of *TAHRE* outside the telomere. First, a subset of *TAHRE* SNPs vary less in copy number than expected of alleles found in active elements, suggesting the possibility of inactive copies outside the telomeres (Figure 3G). Second, we find that few strains lack full-length *TAHRE* insertions in contrast to the other HTTs (Figure 1D), suggesting there may be tandem *TAHRE* copies that are rarely lost. Finally, we observe an R1 inserted into *TAHRE* sequence which has reached moderate population frequency (12/15) in the Beijing population. Given the instability of the telomeres, we think it unlikely that such structures could be present in so many genomes if they involved telomeric *TAHRE*.

We therefore searched for non-telomeric *TAHRE* elements in unmapped heterochromatic sequence missing from the reference genome. As the unmapped contigs in the reference genome are too short to determine whether *TAHRE*-containing contigs are surrounded by telomeric repeats or instead embedded in non-telomeric repeats, we searched for heterochromatic *TAHRE* insertions in the PacBio long-read assembly (GCA_000778455.1) of the reference strain ISO1. We identified one 140kb scaffold (JSAE01000744) which contains three tandem *TAHRE* insertions embedded among many other TEs. These are highly similar to each other, though 5% divergent from the *TAHRE* consensus. Both the 3'- and 5'-ends of two insertions are evident and they form tandem junctions. These *TAHREs* likely correspond to the full-length *TAHRE* that we found above in most strains. Nearly all of the underdispersed *TAHRE* minor alleles (which led us to search for non-telomeric *TAHRE*s) correspond to polymorphisms present in these heterochromatic elements (Figure 3G), indicating that these are also the cause of the underdispersed alleles and providing validation for our inference that underdispersion reflects older insertions. Some polymorphisms within these non-telomeric copies are present in more than three copies and display overdispersed variation more consistent with active elements, suggesting that some extant active *TAHRE*s are closely related to the lineage that gave rise to these tandem *TAHRE*. Both the *gag* and *Pol* ORFs are interrupted by several nested TE insertions that are present in

most (82/85) strains. Tandem TEs indeed can form by multiple, independent insertions at the same locus, but generally constitute a minority of total insertions and often form at strong consensus insertion sites [36]. If these tandem *TAHRE* elements reflect true non-telomeric insertions, at least two independent *TAHRE* insertions must have occurred at this locus. Given that we observed no single insertions outside of the telomeres, we consider double insertion outside of the telomere to be unlikely (although it is possible that non-telomeric *TAHRE* insertions occurred more frequently in the past). We instead propose that these *TAHRE* elements originated as insertions within heterochromatin-adjacent telomeres, XR or 4L being likely candidates, and were subsequently pushed into the adjacent heterochromatin by accretion from the edge of the array.

While selection on the genome might have favored the evolution of factors that constrain the HTTs to the telomeres, there are unlikely to be similar constraints on the *R*-elements. Rather, their restricted localization to the rDNA is probably a property purely intrinsic to the *R*-element transposition machinery. Thus, the extent to which we detect *R*-element insertions outside of their niche both validates our detection method and provides insight into the extent to which niche specificity can be maintained by TEs alone without external control by the genome. We identified no instances of *R2* elements inserted outside of rDNA units in the GDL data. By contrast, we found 33 insertions of *R1* outside of rDNA units where both the intact 3'-end and (often truncated) 5' junctions were identifiable. Most such junctions were restricted to single strains, though one insertion in the 3L pericentric heterochromatin was present in 70/85 strains, suggesting it occurred more distantly in the past. Overall, while this corresponds to only 0.34% of all *R1* elements in the GDL, it nonetheless reflects more insertion site promiscuity outside of its typical niche than we observe for the HTTs. On the whole, though, the tight restriction of the *R*-elements and HTTs to their respective niches suggests it is a property of the elements themselves and that external control may not be required to maintain highly specific insertion sites.

Terminal deficiencies are common

The high degree of copy number variability displayed by the HTTs appears to be shaped by the dynamic nature of the telomere. We sought, therefore, to further characterize the stability of the

telomeres. Beyond the action of unequal exchange between telomeres to amplify and delete sequence, there are other possible mechanisms by which telomeric sequence may turn over. First, the end replication problem directly causes telomere erosion, which in some cases may reduce the chromosome ends to the subtelomeric TAS arrays or even into centromere-proximal unique sequence. Second, DNA double strand breaks in the subtelomere may result in sudden terminal deficiencies.

We consider three lines of evidence to identify deficiencies. First, to identify the subset of terminal deficiencies that extend through the HTT array and subtelomeric repeats into unique sequence, we searched for a sharp loss of (homozygous) or reduction in (heterozygous) read depth over telomere-proximal unique sequence. Second, deficiencies which have been healed by new HTT insertions can be detected as junctions between the 3'-ends of HTT elements and subtelomeric unique sequence [50]. Such insertions onto the end of a broken chromosome can be readily distinguished from chromosome-internal insertions, which would produce junctions with unique sequence involving both the 3' and 5' ends of the HTT. Considering any subtelomere that met either of these criteria, we found evidence for 56 terminal deficiencies in 48/85 strains (Figure 4A). As a third and independent test, we looked for concomitant deletion of the subtelomeric satellite. The left arms of chromosomes 2 and 3 harbor arrays of the same subtelomeric satellite, TAS-L. If we have identified true terminal deficiencies of 3L, the TAS array should be lost from those chromosomes and TAS-L copy number reduced. Consistent with true terminal deficiencies on 3L, in strains with HTT-3L junctions the TAS-L copy number is roughly half that of strains without detectable deficiencies (Figure 4B).

Of the 56 terminal deficiencies, 14 display evidence of heterozygosity, suggesting these are not fixed in the fly stocks and perhaps arose after collection. The remaining 42/56 deficiencies are all consistent with homozygous deficiencies, as expected in inbred strains. Of the 56 deficiencies, 34 showed evidence of an HTT-subtelomere junction, indicating that the lost telomere has been replaced by a new HTT array. All but five of those 34 deficiencies extend sufficiently far into unique sequence that we could confirm a clear drop in coverage after the HTT-subtelomere junction. For 22/56 deficiencies, however, we observe only a loss of subtelomeric unique sequence based on read depth but no evidence of repair in the form of an HTT junction. Some of these may reflect our failure to detect a junction that is

truly present, for example due to variation in read counts or repeat-masked sequence (note that five putatively unhealed deficiencies sit at the edge of masked sequence, Figure 4A middle panel). However, total telomere length is generally lower in these strains with putatively unhealed deficiencies compared to strains where a HTT junction is identifiable (Figure 4C), consistent with truly missing a telomere.

At least 39 of the deficiencies breakpoints are independently derived, involving different sequence coordinates. Most (44/56) are found on 3L and often delete the most distal gene, *mthl8,* but a few delete the next annotated gene *CG43149*; a similar high rate of 3L terminal deficiencies was discovered by [33]. We note that the density of deficiencies is twice as high distal to *mthl8* than it is between *mthl8* and *CG43149,* suggesting that deletions of *mthl8* nonetheless have some deleterious effect even if not as serious as loss of *CG43149* (Figure 4A, middle panel). Only 9 and 3 deficiencies are present on 2R and 3R, respectively, but the distribution of these are also limited by genes within several kilobases of the telomere, with few extending beyond *CG30429* on 2R and beyond *Map205* on 3R. We observed no deficiencies on 2L, likely because there is only 400 bp of non-repetitive sequence between the gene *lethal (2) giant larvae* and the TAS array. Mutations in *l(2)gl* are homozygous lethal; therefore terminal deletions removing *l(2)gl* would not be expected in the GDL lines we sampled because they have been inbred to homozygosity, but they are common on heterozygous chromosomes in natural populations and in heterozygous lab stocks [51–54]. Taken together, our analysis reveals that terminal deletions occur at high frequency on all chromosomes, and also highlights the fitness impacts of terminal erosion and deletion.

Patterns of HTT insertions suggest different families are active under different conditions

The presence of distinct families and divergent subfamilies of TEs is a common feature of Drosophila telomeres and opens up the potential for specialization of elements and differential regulation by the host genome. We investigated two possible cases. First, we hypothesized that HTT specialization could affect patterns of HTT interspersion. For example, competition among elements that is resolved by limiting their expression to different cell types or development stages would result in finding adjacent copies less frequently than expected. Conversely, cooperation among elements that requires their co-

expression for successful transposition would result in an increased frequency of interspersed insertions. We found that elements of the same subfamily tend to neighbor each other, suggesting that they tend to be reverse transcribed multiple times at a given chromosome end (Figure 5A, B). We also observed that elements of the *TAHRE* and *HeT-A* families are generally interspersed as would be expected under random ordering. *TART* elements, however, are less commonly found adjacent to *HeT-A* family and *TAHRE* insertions than expected. Further, the tendency of *TART* elements to be found adjacent to their own subfamily is much stronger than that of *HeT-A* subfamilies and *TAHRE*. The interspersion of *HeT-A* family and *TAHRE* elements contrasted with the low frequency of junctions of either of them with *TART* may be a consequence of the *HeT-A* family relying upon the expression of *TAHRE*'s polymerase for reverse transcription, implying that the *HeT-A* family and *TAHRE* might mobilize together [40]. It further suggests that *TART* and *TAHRE* may not be active at the same time. To confirm these observations, we examined the assembled telomeres in the release 6 reference genome [55], and observed several instances where multiple full-length copies of the same element are in tandem (Figure 5C). We conclude that the pattern of HTT arrays are not random, and that this likely reflects the dependence of the *HeT-A* family on *TAHRE'*s retrotransposition activity as well as possible specialization of *TART* versus *TAHRE*. The three *TART-A* elements in tandem on the X chromosome telomere, however, share terminal repeats, which is consistent with recombination between the 3' and 5' terminal repeats [56] and the principal mechanism by which TEs with terminal repeats generate tandems in Drosophila [36]. Thus the higher rate of self-adjacency observed for *TART* may reflect recombination between termini and/or multiple independent insertions.

Second, given that loss of subtelomeric sequence is associated with increased HTT copy number, we were particularly interested in the possibility that certain families are more involved in the recovery of lost telomeres than others, which would result in an enrichment near the base of the telomere. On the autosomes, tandemly repeated telomere-associated sequences (TAS-L and TAS-R) sit at the base of each telomere. An HTT insertion adjacent to a TAS-L or TAS-R repeat therefore represents the first insertion after the HTT array eroded completely. Similarly, if an HTT has healed a terminal deficiency extending beyond the TAS array, it will form a junction with unique subtelomeric

sequence (as described above) [33]. In both cases, if an HTT family is particularly activated upon the complete erosion of a telomere, it will be enriched next to these subtelomeric sequences relative to the other HTT families. We find that *TART* family elements are enriched at the base of the telomere and *TAHRE* is depleted, with the *TART* enrichment mostly driven by *TART-C* (Figure 5D; posterior mean = 1.65-fold, 95% CI=1.25-2.1). Among the *HeT-A* subfamilies, *HeT-A* shows evidence of weaker enrichment (posterior mean = 1.3-fold, 95% CI=1.03-1.57), which at the family level is obscured by the other *HeT-A* subfamilies. *TAHRE* is clearly underrepresented at the base (posterior mean = 0.33-fold, 95% CI=0.2-0.47) as is the subfamily *Het-A2* (posterior mean = 0.51-fold, 95% CI=0.34-0.68). These data suggest that the HTTs differ in their propensity to heal completely eroded telomeres, and some may be especially activated by telomere loss.

DISCUSSION

Drosophila telomeres are maintained by three rapidly evolving non-LTR retrotransposable elements that represent an exceptional case of transposable element domestication, as complete elements fulfill an essential function for their host by protecting chromosome ends. HTTs and host-associated proteins evolve and turn over rapidly, which may result from incomplete HTT domestication and ongoing genomic conflict with their host [24,30]. Here, we leveraged the ConTExt pipeline and available population genomic data to investigate HTT variation in *D. melanogaster* at an unprecedented scale. We provide evidence that the telomere represents a unique and highly dynamic genomic niche for HTTs that also contributes to rapid evolution of HTTs and host-associated proteins. The high rate of HTT sequence evolution due to the telomere's rapid telomere could require host capping proteins keep up, which might create the appearance of an arms race.

ConTExt provides a comprehensive picture of Drosophila telomeres

We performed a population survey of HTT copy number, sequence diversity, and organization, using comparisons with other TE families to place our analyses in context. Using ConTExt, our pipeline for accessing repeat variation in short read data, we examined nearly 10,000 HTT insertions across 85

D. melanogaster genomes, greatly expanding upon previous efforts that relied on qPCR and Southern blots [38,39]. We inferred a range of properties of HTT variation despite using short-read data that prevented us from reconstructing entire telomeres, or even individual TEs, similar to our previous inference of structural variation among TEs [36]. The large number of genomes analyzed also provided insights from population genetics and known differences between active and inactive TE families to draw conclusions that typically have required the construction of phylogenies from individual sequences (e.g. see [43,57,58]). Here, we successfully distinguished polymorphism residing in recently active versus older TEs by considering the population genetics of TE copy number distributions. Using this strategy, we also detected the presence of inactive *TAHRE* insertions outside the telomere and correctly predicted which polymorphisms were present within these non-telomeric copies. This approach may be useful in future population genomic surveys of transposable elements.

The data we analyzed derive from pools of about 50 whole females of varying ages from inbred lab stocks [37], with some implications for interpretation. First, the strain-specific TE copy number expansions we observed highlight that TE copy number evolution likely continues during inbreeding, as in the wild new insertions resulting from a copy number expansion would be separated by recombination, increasing the population mean but not resulting in individual copy number outliers. Without this, it would not have been possible for us to identify TEs escaping repression or differences in their susceptibility to repression. However, we emphasize that this issue is not particular to our analysis but is instead relevant to any analysis using strains that have been maintained as stocks for some time, including those used to construct reference genomes. Second, telomere length might differ between sexes, across tissues and over an individual's lifespan. However, somatic elongation is unlikely to affect our estimates as these junctions should be specific to individual flies rather than present in the complete pool of sequenced flies and thus generally below detection limits due to their low coverage. In contrast, age-related telomere loss could produce systematic biases across all pooled individuals. However, germline erosion is reported to be on the order of 70-bp per generation in Drosophila [59,60] so any somatic or age-dependent effects would have to be more severe than these reported rates of erosion to impact our conclusions.

The telomere is a highly dynamic genomic niche

The principal prediction of using selfish genetic elements to maintain the telomeres rather than telomerase is that the conflicting interests of the HTTs and the genome should make it harder for Drosophila to maintain its optimal telomere length. We found *D. melanogaster* telomeres, averaging 50kb, to be somewhat longer than is typical of telomerase based systems (generally less than 20kb). Drosophila may be particularly vulnerable to excessive elongation as telomeric TEs gain a reproductive advantage by doing so. Consistent with prior estimates, we also found telomere length to be quite variable among strains. This variation likely reflects that the monomers for Drosophila telomeres are 6-10kb-long TEs rather than ~6 bp repeats in telomerase-based organisms. But there are species with even longer telomerase-maintained telomeres such as lab mice and raptors, and telomere-length variation is similar in *C. elegans* and *A. thaliana* and even greater in some birds [61–63]. We suggest that idiosyncratic species-specific factors rather than different mechanisms of telomere formation account for observed variation in telomere length among organisms.

In addition to cycles of elongation and erosion, other processes likely contribute to telomere length variation. The tandem arrangement of HTTs in telomeres means that unequal exchange can amplify and delete their sequence. The heightened tendency for TART subfamilies to be tandemly repeated compared to other HTTs without terminal repeats is consistent with the action of such a recombination-based mechanism. We also confirmed that the loss of even entire telomeres is an astoundingly common event, occurring in more than half of the strains surveyed (48/85).

We were particularly surprised to discover evidence that 22/56 of the terminal deficiencies appear to be unhealed, that is, are lacking a terminal HTT element. Chromosomes stably lacking an HTT array have been observed in lab stocks [59,64], but to our knowledge have not been previously reported in stocks recently derived from the wild, suggesting that such chromosomes may also be common in natural populations. The prevalence of these unhealed deficiencies suggests that in some cases elongation of lost telomeres is slow to occur. This could be exacerbated in strains that have lower than average rates of HTT transposition, such that terminal deficiencies may be both more prone to occur and remain unhealed longer. Consistent with this possibility, the median total telomere length in strains with

unhealed homozygous deficiencies is 65% that of strains with healed deficiencies, somewhat lower than the 12.5% difference expected from missing a single telomere (Figure 4C).

Telomere instability contributes to the rapid evolution of HTTs.

Residing in a region as prone to sequence turnover by erosion, unequal exchange, and sudden deficiencies as the telomeres presents unique challenges for the HTTs compared with other TE families. Indeed the HTTs have evolved a variety of strategies to cope with this instability, for example 3'-promoters and near terminal repeats (Pardue and DeBaryshe 2011). The tendency we observed for HTTs to insert multiple times in tandem might reflect another such strategy, as it protects the more proximal insertions from the immediate erosion faced by the most distal.

Beyond favoring HTT adaptations, telomere instability may also explain the extent of HTT variation. Consistent with this, we found that the HTTs are among the most variable TEs in copy number across *D. melanogaster* genomes, comparable only to the R-elements, which are the only other *D. melanogaster* TEs to insert at tightly linked, unstable tandem arrays that are susceptible to unequal exchange. We suggest that HTTs, by virtue of inserting at telomeres, are thus predisposed to a much higher probability of being lost than typical TEs, with implications for their evolution. First, this likely contributes to the variability in the complement of telomeric TEs among Drosophila species, by increasing the stochasticity with which particular HTT lineages fix or go extinct across the phylogeny. Second, simulation studies suggest that high rates of TE deletion may drive non-neutral sequence evolution by heightening the strength of selection on transpositional activity [65], as the birth rate of new insertions must compensate for the heightened death rate of preexisting insertions for the family to remain viable. If there are fitness impacts on the organism of excessive telomere elongation [66], which additional observations of telomere trimming suggest may be the case [67,68], this heightened selection on transposition may also intensify conflict among the HTTs as they compete for limited space and transposition machinery.

Our finding that diversity among HTTs largely reflects divergence among active elements stands in contrast to other active TE families whose high diversity is mostly driven by divergent inactive relics

from prior invasions. This distinctive property of HTTs is likely explained by the high-turnover rate of telomeric sequence. Whereas old insertions may persist for millions of years elsewhere in the genome, they are likely rapidly lost at the telomere. Yet, this high turnover of telomere sequence may not be sufficient to completely explain their sequence diversity. Considering both total sequence diversity and that restricted to active elements, the HTTs are still much more diverse than *R2*, which is also subject to a high turnover rate. This might be a consequence of *R2* elements averaging twice the copy number of individual HTTs subfamilies, as new alleles should reach intermediate frequency more slowly in larger populations of elements. But while the copy number of individual *HeT-A* subfamilies is less than that of *R2*, the total *HeT-A* family is more abundant than *R2* and has diversified into five distinct lineages that display evidence of ongoing diversification. Acknowledging that our argument here is based on contrasting HTTs with only a single other TE family, it does hint that some dynamic is at play leading to higher rates of diversification in the HTTs compared to *R2*. Competition for limited space among HTTs or conflict with the host genome's regulatory mechanisms are likely candidates. We can more firmly conclude that HTT sequence diversity is driven by recently active elements and that this is due to the high rate of turnover at the telomere.

HTT copy number expansions suggest conflict

While the instability of their niche is likely what distinguishes much of HTT copy number and sequence variability from other active TEs, as bona fide TEs genomic conflict may also be important in explaining HTT evolution [30]. We stress that such conflict may occur at two levels: among HTTs, as they compete for limited space, and between HTTs and their host. Our observation of runaway expansions of both HTTs and non-telomeric TEs highlights the action of host factors in repressing TE activity and demonstrates that absent such repression HTTs are no more restrained in their replication than typical TEs. Any HTT that escapes host repression will have a reproductive advantage over other HTTs, though perhaps only in the short-term as telomeres populated by hyperactive HTTs might be detrimental to the genome. In the longest telomere strains, no one subfamily predominated, suggesting general misregulation as the principle cause of excessive elongation rather than recurrent escape from

suppression of any particular HTT subfamily. However, some subfamilies were moderately over- or underrepresented in the long telomere strains compared to the short strains, suggesting some differences in susceptibility to host regulation among the HTTs. Further, the observation that not only are telomeres longer in the Beijing strains but that this results from expansions of particular *HeT-A* subfamilies, but not *TAHRE* or *TART*, suggests that this variable susceptibility may be actively evolving across the species, as would be expected under conflict.

Is telomere restriction domestication or an evolutionary strategy?

The persistent localization of HTTs to the chromosome ends for tens of millions of years is frequently interpreted to reflect domestication by their Drosophila host, but it might instead reflect an evolutionary strategy of the HTTs to reduce their fitness impact on the genome [6]. Studies of reference strains of multiple species in the *melanogaster* subgroup found that HTTs are entirely telomere-restricted in most species, though an ancestrally telomeric TE has promiscuous non-telomeric localization in *D. rhopoloa* [24]. While the large number of TEs included in our study offered the potential to detect lower rates of insertion outside their niches than studies focused on single reference strains, we found no evidence of HTTs actively inserting outside of the telomere in *D. melanogaster*; our discovery of several non-telomeric *TAHRE* elements suggests that their origin is not recent. We did, however, observe a low rate of *R1* insertions outside of the rDNA, which concords with prior observations that non-rDNA *R*-element transposition occurs at low rates across the genus, with a few species showing high rates [69].

Inserting only at gene-poor but unstable loci like the telomeres and rDNA arrays must entail a tradeoff between the reduced fitness cost imposed on the host genome and the heightened rate of loss the TEs themselves incur. The long term localization of the R-elements to the rDNA array suggests this tradeoff can be stable even in the absence of domestication. The lower frequency of HTTs escaping from their niche compared to *R* elements might reflect domestication, with the fitness benefits they afford the genome further shifting the balance of the trade off in the HTT's favor. In contrast, the more frequent escape of *R*-elements could reflect competition among themselves [70] or perhaps a higher fitness cost of rDNA insertions in some species. Again, though, a more neutral explanation is also plausible. While

endonuclease activity is a necessary step in the insertion of *R*-elements, the HTTs probably do not require a double-strand break to insert at the chromosome ends. While they apparently retain endonuclease domains, these are possibly non-functional, and endonuclease-deficient LINE1 elements have been observed to insert at chromosome ends in a fashion similar to the HTTs [71]. Escape from the telomeres may then require reacquisition of a functional endonuclease, thus locking telomeric TEs more tightly into their insertion site than the *R*-elements, which need only to relax the sequence specificity of their endonuclease to escape the rDNA. We therefore suggest that the long-term telomere restriction of the HTTs need not be the consequence of host-driven domestication, rather evolutionary strategies and neutral factors locking the HTTs into their insertion site may be sufficient explanations.

Host regulation of the HTTs

Complete domestication would imply that the activity of the HTTs is completely under the genome's control. While such control of *bona fide* TEs is likely impossible as long as some of the HTTs retain autonomous transposition activity, the importance of maintaining telomeres in the face of this instability should place pressure on the host genome to control the activity of telomeric TEs. Mutations in a number of genes involved in general TE regulation, primarily the heterochromatin maintenance and piRNA pathways, are associated with excessive HTT expression and telomere elongation. Our observations are consistent with excessive telomere elongation largely resulting from host misregulation, and the observation that *Copia* and *Mdg3* also expanded considerably in the longest telomere strain is particularly suggestive of defects in piRNA repression. The reliance of the HTTs upon each other for telomere localization and transposition machinery may simplify their regulation by the genome. We observed expansions of all *HeT-A* subfamilies and *TAHRE*, but not the *TART* elements in the longest telomere strain, which is consistent with the *HeT-A* family being more sensitive to piRNA repression than *TART* as suggested by [72]. Similarly, we found that *TART* elements are segregated from *HeT-A* family elements and *TAHRE* in the telomere. This might entirely reflect tandem expansions by terminal repeat recombination in *TART*, but different patterns of expression could also contribute. For example, in piRNA mutants, *TART* expression is most evident in nurse cells during later oogenesis while *HeT-A* family and

TAHRE upregulation occurs in the early oocyte [23,40,73]. This coexpression phenotype suggests that *TAHRE* controls the non-autonomous *HeT-A* family, as proposed by others [12]. The nearly fixed heterochromatic *TAHRE* we identified might then reflect the separation of a piRNA source from the HTT elements that it regulates, potentially one step toward the imposition of tighter host control on the transposase upon which the predominant HTTs, the *HeT-A* family, rely.

The enrichment of *TART* elements and strong depletion of *TAHRE* elements at the base of telomeres further suggests differences in responsiveness among the HTTs. This may reflect regulatory differences or functional specialization among the HTTs, with *TART* elements being particularly responsive or *TAHRE* particularly unresponsive to complete telomere erosion. Consistent with this, distinct patterns have been observed in mutant backgrounds that contain pre-existing terminal deficiencies. Mutations in *SuVar(205)* (which encodes the heterochromatin and telomere-associated protein HP1) elevate rates of healing by both the *HeT-A* family and *TART*, while disruption of piRNA silencing by mutations in *spn-E* leads only to *TART* insertions at the deficiency [73,74].

CONCLUSIONS

Together our results highlight the complexity of HTT evolution and the variety of forces at work. Telomere instability clearly drives adaptations, but it also shapes HTT copy number and sequence variation in ways that distinguish them from other TEs. Copy number and sequence variability should not therefore be ascribed solely to either conflict or domestication without adequately considering the impact of this remarkably unstable niche. In fact, while the long-term telomere localization of the HTTs is usually ascribed to domestication, we suggest that it may instead reflect a highly successful evolutionary strategy on the part of the HTTs to reduce their deleterious impact on the host genome. Nonetheless, our results are consistent with prior observations that host sequences regulate the HTTs and further suggest that their susceptibility to regulation varies among the HTTs and is actively evolving, consistent with ongoing conflict between these genomic endosymbionts and their host genome.

METHODS

The repeat index and nomenclature

We used the manually curated repeat index described in [36], which contains consensus sequences for the known *D. melanogaster* TE families as well as satellite repeats, including the left and right telomere associated sequences (TAS-L and TAS-R). The telomeric TE sequences in this index are somewhat distinct from the Repbase [75] entries. First, the long perfect near-terminal repeats (PNTR) found in the *TART* elements are given entries separate from the internal sequences, similar to how the terminal repeats of LTR retrotransposons are given their own entries separate from the internal sequence in Repbase. Second, the index includes four *HeT-A* family entries in addition to the Repbase consensus (Table S1), as many of the *HeT-A* family insertions in the reference genome are quite divergent from the consensus sequence (< 85% identity) (Figure S4A). To ensure reads derived from *HeT-A* family insertions would align to the repeat index, we extracted all insertions annotated as *HeT-A* in the UCSC genome browser Repeat Masker track for the release 6 *D. melanogaster* genome. We then performed all pairwise alignments using BLASTn and constructed a graph by construing insertions as nodes and connected all pairs of sequence that shared >90% identity with edges [36]. We partitioned this graph into communities of homologous insertions using the Louvain algorithm [76]. This identified four *HeT-A* family communities, one of which appeared to contain subcommunities. We split the large community into two subcommunities by reclustering with the alignment cutoff set at 95% identity. From each community we manually chose a full-length element as representative and added these to the repeat index, labelling them *HeT-A1* to *HeT-A5*. However, we removed *HeT-A4* due to high sequence similarity with *Het-A3*. We refer to the original *HeT-A* sequence from RepBase as "*HeT-A*" [77]. The final set of *Het-A* family sequences all have less than 85% identity to each other, with most pairs having roughly 80% (Figure S1A). For TEs that have internal repeats, we removed them by scanning along each consensus, and for each 70-mer in the consensus masking all subsequent 70-mers within 5 mismatches.

Categorizing TE families as active

We categorize TE families as active or inactive based on their percent identity and population frequency as summarized in [78]. We consider a family as putatively active (or recently so) if the mean pairwise identity between individual insertions in the reference genome was greater than 95% [78,79] and the average population frequency of insertions was less than 0.4 [80]. We additionally consider the telomeric TEs and the two *R*-elements as being active, as well as include *P*-element, *Hobo*, and *I*-element. While this approach is likely to misclassify some truly active elements as inactive, the two categories should be fairly representative of the active and inactive families in *D. melanogaster*.

Sequence data and read mapping

Sequencing data come from a previously published study in which pools of ~50 females from 85 wild-derived *Drosophila melanogaster* lines covering five continents were sequenced with Illumina 100nt paired-end reads to an average depth of 12.5X [37] (NCBI BioProject PRJNA268111).

We employed the ConTExt pipeline to organize these data and discover structural and sequence variation within the repetitive portions of these genomes [36]. Briefly, ConTExt aligns repeat-derived reads to the corresponding repeat consensus sequences (using Bowtie2 v. 2.1.0 [81]). Mixture modeling is used to infer the set of underlying structures that generated the set of discordantly aligned read pairs in each sample, and ConTExt subsequently clusters these junctions to determine which structures are present in multiple samples. We modified the pipeline here to infer sequence variation within repeat families from the aligned reads (described below). Additionally, rather than align reads first to individual insertions and then collapse these alignments onto consensus sequences [36], here we aligned reads directly to the consensus sequences using permissive alignment parameters (`--score-min L,0,-2.5 -L 11 -N 1 -i S,1,.5 -D 100 -R 5`).

Filtering ambiguous alignments

While Bowtie2's mapping quality summarizes the ambiguity of alignments, it heavily penalizes divergence from the reference sequences and is defined in a rather opaque manner involving nested

conditionals. Because we expect some reads to derive from TE insertions diverged from the consensus, it is undesirable to penalize reads with no secondary alignments but which are nonetheless diverged from the consensus. We instead filter ambiguous alignments by directly considering the primary and secondary alignment scores, which we convert to percent identities assuming all penalties were due to mismatches. We use these to define a score *M* reflecting the distance between the primary (*AS*) and secondary (*XS*) hits:

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

This score summarizes the distance between the primary and secondary alignments and is penalized by the divergence of best alignment from the consensus, but it does so in a predictable fashion. If a secondary alignment was reported by Bowtie2, we required this score to be greater than 0.05 for the alignment to be considered unambiguous and included in the analysis. If the primary hit perfectly matches the consensus, the secondary alignment must be more than 5% diverged from the consensus for the read to be included in the analysis. If the primary alignment is 10% divergent, the secondary alignment must be more than 20% diverged from the consensus for the read to be included in the analysis (Figure S4B). Finally, we exclude any read whose primary alignment is more than 80% diverged from the consensus.

Overview of statistical analyses

We describe the details of each model employed in the subsequent sections, but outline the general logic of our analyses here. We approached the analyses of the data from the perspective of Bayesian parameter estimation, seeking to define probability distributions over all possible values of our statistics of interest to guide our interpretations. For models with non-uniform priors we include plots comparing the marginal densities of the prior and posteriors (Figure S2). To accomplish model-checking, we employ posterior predictive simulations to evaluate the ability of our models to account for features of the observed data as a means of model checking. When the failure of a model to account for features of the data is an observation of interest, we report the posterior predictive p-value [82], which unlike the

frequentist p-value is conditioned upon the set of models most consistent with the data rather than a predefined null-hypothesis.

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

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

For Bayesian modelling we employ PyMC3 (v. 3.6). For each model, we draw samples from the posterior distribution in two chains to assess convergence, sampling continuous random variables with the No-U-Turn Sampler initialized with jitter+adapt_diag as implemented in PyMC3. The No-U-Turn Sampler (NUTS) is the current state of the art in Monte Carlo sampling, capable of efficiently exploring high-dimensional parameter spaces and drawing uncorrelated samples from the posterior. While superior to the Metropolis algorithm in most regards, it is unable to sample discrete random variables due to its reliance on gradient information. Consequently, we represent copy number in our analyses with positive

continuous random variables rather than with discrete variables, preferring greater confidence in the posterior sample afforded by NUTS. We note that treating copy number as continuous is not dissimilar from the commonly used approach of estimating copy number by dividing the observed read counts by the expected read depth. Similarly, when we incorporate outlier detection in our models we marginalize out the binary labels representing the outlier status of individual observations to avoid the need for Gibbs sampling. We retain the interpretability provided by these labels by using the posterior distribution to reconstruct the probability that each observation reflects an outlier.

Modelling read depth

To infer the copy number of multicopy sequence, we need to know the read depth distribution of single copy sequences. We estimate this from coverage of the two major autosomes in the reference genome, using the same filtering steps described in [36]. As in [36] we consider the coverage of sequence not by the reads themselves, but rather by the interval between read pairs. However, here we model the read depth of single copy sequence with negative binomial distributions, allowing the mean and overdispersion to vary with %GC. To better handle junctions with extreme %GC, we model the relationship between the mean and overdispersion of %GC in each strain *j* with library-specific functions:

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

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

where *g* is the expected %GC of read pairs spanning a position and the βs are the coefficients of each quadratic function.

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

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

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

Estimating copy number from read depth

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

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

Estimating the breakpoints of terminal deficiencies from read depth

We identified terminal deficiencies by visually examining the coverage of subtelomeric sequence for obvious loss of read depth (Supplemental File S1). To estimate the breakpoint of a terminal deficiency based on read depth, we sought to identify a sequence coordinate beyond which the copy number, estimated as the ratio between the observed and expected coverage, drops below 1. For each of the autosome arms, we considered coverage of the most telomere-proximal 100 kb sequence and thinned it by retaining only every hundredth position, to account for autocorrelation in the coverage signal. We then removed any position that has more than five repeat-masked nucleotides within 100 base pairs to reduce

the impact of repeat masking on coverage. At each position, we divided the observed read count by the expected read count given the local GC-content. We then fit a step function to these copy number estimates, *x*, where the expected copy number differs from 1 on the telomere-proximal side of a breakpoint, *c*. We estimate this breakpoint by minimizing the sum of squared residuals (SSR), defined for positions on the centromere proximal side of the breakpoint as

$$SRR_{cent} = \sum_i (x_i - 1)^2$$

and on the telomere-proximal side as

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

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

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

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

If we had evidence of an HTT-subtelomere junction at a breakpoint, we used the sequence coordinates of the junction as the deficiency breakpoint. In estimating the number of independently derived deficiencies, we considered any pair of deficiencies within 200 nt nucleotides of each other as potentially the same allele and collapsed them.

Analysis of truncated HTTs

HTTs, like all non-LTR retrotransposons, are frequently 5'-truncated due to incomplete reverse transcription, in addition to being susceptible to telomere end erosion. While ConTExt does not permit the reconstruction of HTT insertions, we can determine the extent to which each insertion is 5'-truncated.

A full-length insertion must have the intact 5'-end, though the presence of the 5'-end does not guarantee that there are not deletions within the insertion. We note, however, that we observe few junctions consistent with internal deletions within HTTs. Given the high rate of truncation shown in Figure 1B, strains containing only truncated insertions of a particular family might be common. But because such elements are unlikely to encode for functional proteins, and only the autonomous *TAHRE* and *TART* elements encode reverse transcriptase, some full-length elements are likely maintained by selection. We found that 33% of strains lack any full-length *TART* elements, but only 2 out of 85 strains lack full-length *TAHRE*. To assess the likelihood of this observation, we model the number of full-length elements (see Methods) and then used posterior predictive simulations to ask whether we observe fewer strains without any full-length *TAHRE* elements than would be expected under the fitted model. In 99% of the simulations, the fitted model predicted more strains entirely missing full-length *TAHRE* than we observed. One interpretation is that selection acts on the genome to maintain the reverse transcriptase encoded by *TAHRE*. Alternatively, in most genomes there may be a full-length *TAHRE* insertion in tandem with another HTT that is rarely deleted, perhaps because it is located outside of the telomere. We present in the Results evidence in favor of this alternative explanation, but stress that these *TAHRE* elements are old and likely non-functional.

Interpreting sequence variation within repeats

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

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

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

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

Correcting for additional biases in read counts over junctions

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

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

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

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

Fitting this model, we find that the read count over these junctions is 53% (95% credible interval 43 - 67%) of what we expected given the read depth distributions observed for single copy sequence. We incorporate this bias (and our uncertainty about its exact degree) into our copy number estimates of junctions between HTTs and other repeats, calibrating their read counts against this subset of HTT junctions with known copy number. When we use these corrected copy number estimates to infer the total amount of telomeric sequence in each strain we find a strong concordance with our estimates based on consensus coverage without MapQ filtering (r^2=.93, Figure S4C).

Identifying TE copies from their junctions

We identify and count the numbers of euchromatic TE insertions as described in [36], by identifying the 3' and/or 5' junctions the TE forms with its insertion site. For *R*-elements, we estimate the copy number from the read count of the junction between their 3'-end and the rDNA. As these junctions all have the same sequence coordinate and are present in multiple copies rather than being a set of single and low-copy junctions like the HTT-HTT junctions, we estimate their copy as the ratio between the observed and expected read count (assuming a single copy structure with the same %GC) without needing to employ partial pooling to avoid overestimating their copy number. As described in the next section, we employ partial pooling to estimate the copy number of HTT insertions, as we expect there might be both single and multi-copy HTT-HTT junctions.

However, as we estimated the copy number of TE insertions (by counting junctions) in unique sequence differently from that of the HTTs and the *R*-elements (by also considering read counts), we worried that some differences among these categories could reflect differences in the behavior of these estimators rather than true biological differences. To rule this out, we also estimated copy number from the read depth near the 3'-ends of each family's consensus, estimating copy number as the average sequence abundance over a 1 kb interval ending 500 bp from the element's 3'-end. Considering only the 3'-end helps limit the impacts of 5'-truncation on estimated copy number. While this ensures the copy numbers of all TE families are estimated in the same way, it does not distinguish TE-derived satellites,

such as the large *R1*- and *Bari1*-derived satellite arrays, from dispersed TEs. We are, however, able to distinguish these in the copy number estimates based on junctions rather than read depth.

Modelling copy number of HTT-HTT junctions

As particular junctions might be present in multiple copies, we estimated the copy numbers of HTT junctions based on their read counts. For a junction *i* in sample *j*, we model its read depth x_{ij} as being negative binomially distributed

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

Here μ_j and $\alpha_j$are the library-specific functions that describe how the read depth and overdispersion vary with the expected %GC of read pairs spanning the junction, and $g_{ij}$and λ are the aforementioned bias. Finally, $n_{ij}$is the quantity of interest: the underlying copy number of the junction.

If we estimated this copy number for each junction independently with weakly informative priors, we would upwardly bias our copy-number estimates. This is because *a priori* we should expect that most HTTs junctions are truly single-copy and that higher than expected read counts often reflect single-copy structures. But, as we do not know ahead of time what fraction are single-copy, it is challenging to encode this belief in a prior. Instead, we employ partial pooling, inferring the underlying copy number distribution at the same time we estimate the copy numbers of the individual junctions.

The straightforward approach to modelling this would be

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

where *m* and *s* are the mean and standard deviation of the copy number distribution. However, sampling $log_{10}n_{ij}$ directly leads to pathological sampling behavior, as is common with hierarchical models [83]. So, we instead employ an equivalent parameterization by instead sampling an offset, ξ_{ij}, which we transform into the copy number as

$$log_{10}n_{ij} = m + s\xi_{ij}, \xi_{ij} \sim Normal(0,1)$$

which leads to efficient sampling behavior.

We then define hyperpriors over the mean and variance of this copy number distribution (Figure S2 A, B)

$$m \sim Normal(1,1.5)$$

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

Modelling telomere length and copy number distributions

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

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

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

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

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

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

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

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

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

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

$$z_i \sim Bernoulli(p)$$

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

However, as binary variables are challenging to sample we marginalize these labels out of the model [84],

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

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

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

and average over all values of θ sampled from the posterior.

When modelling telomere length we describe our prior beliefs (Figure S5 C, D) 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 1kb or as large 100Mb. 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 may 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

$$x_{ij} \sim Multinomial(N = n_j, p = p_j)$$

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*,

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

$\lambda_i$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

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

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

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

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

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

$$\lambda \sim Dirichlet(\alpha_0 .. \alpha_m = \mathbf{1})$$

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

Modelling the proportion of full-length elements

We estimate the length of insertions by considering the coordinates of an element's 5'-end in an HTT-HTT junction, and define an element as being full-length if its estimated end is within 300 bp of the expected 5'-end of its consensus sequence. For elements of the TART family we require the junction to involve the PNTR rather than the internal sequence. As the counts are overdispersed, for each family we model the number of full-length insertions, k_i out of all N_i insertions in strain *i* with a beta-binomial distribution. For interpretability, we estimate mean, q, and concentration, ν, parameters that respectively describe the expected binomial rate of full-length elements and the variability around this which drives the overdispersion. We define the priors in this model as

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

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

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

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

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

To assess the extent to which the model accounts for the fraction of strains lacking any full-length elements, we employ posterior predictive simulations to obtain a posterior p-value that describes the probability of observing as few or fewer strains without full-length elements present in the GDL given the posterior distribution.

Modelling the dispersion of TE copy number

For each TE family we compute the index of dispersion of their copy number as the ratio between the empirical variance and mean of the observed copy number distribution

$$IoD = \frac{\sigma^2}{\mu}$$

and for each of the following categories of transposable elements, we model the logarithm of this statistic as arising from a normal distribution

$$log_{10} IoD \sim Normal(\mu, \sigma)$$

and define our priors (Figure S2 I, J) as

$$\mu \sim Normal(0,3)$$

$$\sigma \sim HalfCauchy(\beta = 2)$$

DECLARATIONS

Availability of data and materials

Sequencing data come from a previously published study (NCBI BioProject PRJNA268111) [37]. ConTExt is located at https://github.com/LaptopBiologist/ConTExt. The PacBio assembly (GCA_000778455.1) is described in reference [85].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by NIH grants R01-GM074737 and R01-GM119125 to D.A.B. A.M.D.C. was supported by postdoctoral fellowships from the FRQ-S, NSERC, and the Lawski foundation. The funding bodies had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author's contributions

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

Acknowledgements

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

FIGURE LEGENDS

Figure 1. Telomeres are highly dynamic in *Drosophila melanogaster*. (A) Telomere length distribution (in kb) as estimated from HTT sequence abundance for each strain, grouped by population. Filled circles represent outlier strains. B: Beijing, I: Ithaca, N: Netherlands, T: Tasmania, Z: Zimbabwe. (B) Proportion of full-length elements per subfamily per strain. (C) Telomere composition depicted by proportion of total telomere length per HTT subfamily as estimated from copy number. White corresponds to short telomere strains (bottom 10%), grey to long telomere strains (top 10%). (D) Telomere composition depicted by proportion of total telomere length per HTT family as estimated from copy number. White and grey are same as in (C).

Figure 2. Comparing HTT variability to other TEs within *Drosophila melanogaster*. (A) Copy number of selected TEs per strain as estimated from junctions. Each dot indicates the copy number of a TE family in a single strain, and red dots indicate strains with extreme copy-number expansions. TEs are grouped as: Left, TEs with copy number outliers; middle, HTTs; right, all other active TEs. (B) HTT copy number in strains where no other TE family has undergone a copy number expansion (–) compared to strains where other TE families have expanded in copy number (+).

Figure 3. HTT copy number and sequence variation in the context of other TEs. (A) Scatter plot showing the relationship between mean copy number as estimated from junction data of TE families (log scale) and their variance (log scale) across the GDL. Designations of active and inactive TEs are from prior estimates of sequence divergence and population frequency as described in the Methods. Solid line represents the expected relationship under the assumption of little variation in population frequency and low linkage disequilibrium among insertions. Shaded regions summarize the distributions of mean and variance for inactive (grey), active (yellow), and the HTT+*R*-elements (purple) TE families, covering two standard deviations of bivariate Gaussians matched to the moments of the data they are approximating. (B) Boxplots depicting the posterior samples of the mean index of dispersion for each of the four categories of the TEs. (C) Average pairwise divergence among all copies of each active TE family in the GDL, estimated from the depth of reads supporting each possible allele. The contribution of positions where the minor allele displays overdispersed copy number variation, suggesting a variant in an active element, is indicated in red. The contribution of positions where the minor allele displays underdispersed copy number variation, suggesting a variant in an inactive element, is indicated in blue. (D-F) The mean-variance relationships of *HeT-A* (D), *hobo* (E) , and *I*-element (F) broken down by the copy number of the major and minor alleles. Each dot reflects the observed mean and variance of the copy number of the major (blue) and minor (gold) alleles of positions with >0.1 sequence diversity. For reference, the shaded regions are re-plotted from Figure 3A, though in 1D the purple region is covered by the dots. G) The mean-variance relationship of TAHRE's minor alleles, colored by whether the minor allele is found in the

heterochromatic TAHRE insertions (blue) or is likely telomeric (red). Shaded regions are re-plotted from Figure 3A.

Figure 4. Terminal deficiencies are frequent and tend to be healed by *TART*. (A) Locations of terminal deficiencies (triangles) and elements involved (see colour legend; "none" indicates a terminal deficiency lacking any HTT-subtelomere junction). Thin and thick black bars represent UTRs and exons respectively, thin lines represent introns. Top, Chromosome 2R; middle, Chromosome 3L; bottom, Chromosome 3R. X-axis scales are in kilobases; note that the telomeres are to the right for 2R and 3R, and to the left for 3L. (B) TAS-L copy-number boxplot in strains with (+) or without (–) deficiencies on 3L. (C) Boxplots comparing the total telomere length of strains with homozygous deficiencies and heterozygous deficiencies without observed HTT-subtelomere junctions (unhealed) to the telomere length of strains with healed deficiencies and without deficiencies. Strains with healed deficiencies (whether homozygous or heterozygous) will have eight telomeres whereas those with homozygous unhealed deficiencies willwill have only seven. Note the discontinuity (the double slash marks) on the Y-axis to accommodate the extremely long 1.75 MB telomeres of the I01 strain.

Figure 5. HTT insertions tend to be enriched adjacent to themselves. (A) The observed frequency with which two HTTs neighbor each other relative to the expected frequency ($\log_2$). (B) Boxplots of the posterior distributions describing the degree to which elements tend to neighbor themselves ($\log_2$). The whiskers reflect the 95% credible intervals. (C) A visualization of the HTT subfamilies, depicted as thick bars, in the X chromosome telomere of the release 6 reference. We depict alignments with at least 90% identity to the consensus; if a region is homologous to two elements, we assign it to the element with the greatest homology, which was only an issue due to some homology between *TART-A* and *TART-C*. The upward ticks indicate the 3'-end of an element and the downward ticks the 5'-end; a full length insertion has both ticks. TART-A_PNTR is the TART-A Perfect Near-Terminal Repeat. (D) Top: The observed proportion which with each subfamily is found anywhere in the telomere (white) or is the first HTT found at the base of a telomere (grey). The error bars are 95% credible intervals computed analytically for

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

Supplementary tables

Table S1. HTT reference insertion position. The positions of the insertions in the Release 6 *Drosophila melanogaster* reference genome used as representative sequences for the *HeT-A* subfamilies used in these analyses.

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

Supplementary figures

Fig S1. The copy number of the HTT subfamilies in each strain, broken down by population. A) *HeT-A*, B) *HeT-A1*, C) *HeT-A2*, D) *HeT-A3*, E) *HeT-A5*, F) *TAHRE*, G) *TART-A*, H) *TART-B*, I) *TART-C*

Fig S2. Mean-variance relationship of copy number estimated from 3'-end read depth. (A) Scatter plot showing the relationship between mean copy number as estimated from the 3'-end of TE consensus sequences (log scale) and their variance (log scale) across the GDL. Designations of active and inactive TEs are from prior estimates of sequence divergence and population frequency as described in the Methods. Solid line represents the expected relationship under the assumption of little variation in population frequency and low linkage disequilibrium among insertions. (B) Boxplots depicting distributions of indices of dispersion for each of the four categories of the TE families based on these copy number estimates.

Fig S3. Minor and major allele mean-variance relationship of all HTTs. A) *HeT-A*, B) *HeT-A1*, C) *HeT-A2*, D) *HeT-A3*, E) *HeT-A5*, F) *TAHRE*, G) *TART-A*, H) *TART-B*, I) *TART-C.*

Figure S4. Details on interpreting alignments to telomeric transposable elements (A) Percent identities of all >300 bp insertions annotated as *HeT-A* in the UCSC genome browser repeat-masker track of the release 6 reference genome telomeres. Percent identities are based on a ClustalOmega multiple sequence alignment with default parameters. The RepBase consensus sequence and the representatives we chose for each *HeT-A* subfamily in our repeat index are indicated as such. The white cells indicate no alignment between a pair of sequences and all involve alignments with the two shortest *HeT-A* fragments. B) A depiction of how we define ambiguous alignments. The black line indicates how diverged the second best aligned must be for a primary alignment with a given percent identity to be considered unambiguous. C) A comparison of total telomere length estimated from the mapping quality filtered read count of junctions and from coverage of HTT consensus sequences without mapping quality filtering.

Figure S5. Visualization of prior and posterior distributions. A comparison of the prior (red) and posterior (blue) probability density functions for model parameters. The hyperparameters used in the partially pooled copy number estimates, the mean (A) and standard deviation (B) of the distribution describing the copy numbers of all HTT junctions in the GDL. The population specific means (C) and standard deviations (D) of the total telomere length distributions. The population-specific means (E) and variances (F) of the total HTT copy number distributions and the influence of terminal deficiencies on these means (G) and variances (H). The mean (I) and standard deviations (J) of the log of the index of dispersion for the four categories of TE families.

Supplementary File 1. Visualization of read coverage on chromosomes with evidence for terminal deficiencies. Each blue dot is the estimated copy number at a given position, determined as the ratio

between the observed and expected read depth. The black lines below the plots indicate which regions of the reference are repeat-masked (and should show little or no read coverage) The orange trendlines are a 10 kb rolling window average of estimated copy number and the red lines indicate the average copy number estimate to the left and right of the inferred deficiency breakpoint. The vertical red line in some plots indicates the location of an HTT-subtelomere junctions. Below each plot we note the breakpoint coordinate we estimated based on read depth as well as the percent of the coverage prior to the breakpoint remaining in the deleted regions, from which we infer the heterozygosity or homozygosity of the deficiency.

References

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19:199.

2. Deininger PL, Moran JV, Batzer MA, Kazazian HH Jr. Mobile elements and mammalian genome evolution. Curr Opin Genet Dev. 2003;13:651–8.

3. Charlesworth B, Langley CH. The population genetics of Drosophila transposable elements. Annu Rev Genet. 1989;23:251–87.

4. Charlesworth B, Charlesworth D. The population dynamics of transposable elements. Genet Res . Cambridge University Press; 1983;42:1–27.

5. Le Rouzic A, Boutin TS, Capy P. Long-term evolution of transposable elements. Proc Natl Acad Sci U S A. 2007;104:19375–80.

6. Cosby RL, Chang NC, Feschotte C. Host–transposon interactions: conflict, cooperation, and cooption. Genes Dev. genesdev.cshlp.org; 2019;33:1098–116.

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

8. Eickbush TH. R2 and related site-specific non-long terminal repeat retrotransposons. Mobile DNA II. American Society of Microbiology; 2002. p. 813–35.

9. Jangam D, Feschotte C, Betrán E. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends Genet. 2017;33:817–31.

10. de Lange T. How telomeres solve the end-protection problem. Science. 2009;326:948–52.

11. Mason JM, Randall TA, Frydrychova RC. Telomerase lost? Chromosoma. 2016;125:65–73.

12. Abad JP, De Pablos B, Osoegawa K, De Jong PJ, Martín-Gallardo A, Villasante A. TAHRE, a novel telomeric retrotransposon from Drosophila melanogaster, reveals the origin of Drosophila telomeres. Mol Biol Evol. 2004;21:1620–4.

13. Levis RW, Ganesan R, Houtchens K, Tolar LA, Sheen FM. Transposons in place of telomeric repeats at a Drosophila telomere. Cell. 1993;75:1083–93.

14. Mason JM, Frydrychova RC, Biessmann H. Drosophila telomeres: an exception providing new insights. Bioessays. 2008;30:25–37.

15. Pardue M-L, DeBaryshe PG. Retrotransposons that maintain chromosome ends. Proc Natl Acad Sci U S A. 2011;108:20317–24.

16. Arkhipova IR. Telomerase, Retrotransposons, and Evolution. In: Lue NF, Autexier C, editors. Telomerases. Hoboken, NJ, USA: John Wiley & Sons, Inc.; 2012. p. 265–99.

17. Maxwell PH, Belote JM, Levis RW. Identification of multiple transcription initiation, polyadenylation, and splice sites in the Drosophila melanogaster TART family of telomeric retrotransposons. Nucleic Acids Res. 2006;34:5498–507.

18. Danilevskaya ON, Arkhipova IR, Pardue ML, Traverse KL. Promoting in Tandem: The Promoter for Telomere Transposon HeT-A and Implications for the Evolution of Retroviral LTRs. Cell. 1997;88:647–

55.

19. Fuller AM, Cook EG, Kelley KJ, Pardue M-L. Gag proteins of Drosophila telomeric retrotransposons: collaborative targeting to chromosome ends. Genetics. 2010;184:629–36.

20. Biessmann H, Kasravi B, Bui T, Fujiwara G, Champion LE, Mason JM. Comparison of two active HeT-A retroposons of Drosophila melanogaster. Chromosoma. 1994;103:90–8.

21. Biessmann H, Champion LE, O'Hair M, Ikenaga K, Kasravi B, Mason JM. Frequent transpositions of Drosophila melanogaster HeT-A transposable elements to receding chromosome ends. EMBO J. Wiley Online Library; 1992;11:4459–69.

22. Biessmann H, Valgeirsdottir K, Lofsky A, Chin C, Ginther B, Levis RW, et al. HeT-A, a transposable element specifically involved in "healing" broken chromosome ends in Drosophila melanogaster. Mol Cell Biol. 1992;12:3910–8.

23. Villasante A, de Pablos B, Méndez-Lago M, Abad JP. Telomere maintenance in Drosophila: rapid transposon evolution at chromosome ends. Cell Cycle. 2008;7:2134–8.

24. Saint-Leandre B, Nguyen SC, Levine MT. Diversification and collapse of a telomere elongation mechanism. Genome Res. 2019;29:920–31.

25. Zhao Z, Pan X, Liu L, Liu N. Telomere length maintenance, shortening, and lengthening. J Cell Physiol. 2014;229:1323–9.

26. Stewart JA, Chaiken MF, Wang F, Price CM. Maintaining the end: roles of telomere proteins in end-protection, telomere replication and length regulation. Mutat Res. 2012;730:12–9.

27. Shpiz S, Kalmykova A. Control of Telomere Length in Drosophila. In: Li B, editor. Reviews on Selected Topics of Telomere Biology. InTech; 2012.

28. Raffa GD, Ciapponi L, Cenci G, Gatti M. Terminin: a protein complex that mediates epigenetic maintenance of Drosophila telomeres. Nucleus. 2011;2:383–91.

29. Villasante A, Abad JP, Planelló R, Méndez-Lago M, Celniker SE, de Pablos B. Drosophila telomeric retrotransposons derived from an ancestral element that was recruited to replace telomerase. Genome Res. 2007;17:1909–18.

30. Lee YCG, Leek C, Levine MT. Recurrent Innovation at Genes Required for Telomere Integrity in Drosophila. Mol Biol Evol. 2017;34:467–82.

31. Arkhipova IR. Telomerase, retrotransposons, and evolution. Telomerases: Chemistry, Biology and Clinical Applications. John Wiley and Sons, Inc; 2012;265:299.

32. Begun DJ, Aquadro CF. Evolution at the tip and base of the X chromosome in an African population of Drosophila melanogaster. Mol Biol Evol. 1995;12:382–90.

33. Kern AD, Begun DJ. Recurrent deletion and gene presence/absence polymorphism: telomere dynamics dominate evolution at the tip of 3L in Drosophila melanogaster and D. simulans. Genetics. 2008;179:1021–7.

34. Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM. Linkage disequilibria and the site frequency spectra in the su(s) and su(w(a)) regions of the Drosophila melanogaster X chromosome. Genetics. 2000;156:1837–52.

35. Levin HL, Moran JV. Dynamic interactions between transposable elements and their hosts. Nat Rev Genet. 2011;12:615–27.

36. McGurk MP, Barbash DA. Double insertion of transposable elements provides a substrate for the evolution of satellite DNA. Genome Res. 2018;28:714–25.

37. Grenier JK, Arguello JR, Moreira MC, Gottipati S, Mohammed J, Hackett SR, et al. Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. G3 . 2015;5:593–603.

38. Wei KH-C, Reddy HM, Rathnam C, Lee J, Lin D, Ji S, et al. A Pooled Sequencing Approach Identifies a Candidate Meiotic Driver in Drosophila. Genetics. 2017;206:451–65.

39. George JA, DeBaryshe PG, Traverse KL, Celniker SE, Pardue M-L. Genomic organization of the Drosophila telomere retrotransposable elements. Genome Res. 2006;16:1231–40.

40. Shpiz S, Kwon D, Uneva A, Kim M, Klenov M, Rozovsky Y, et al. Characterization of Drosophila telomeric retroelement TAHRE: transcription, transpositions, and RNAi-based regulation of expression. Mol Biol Evol. 2007;24:2535–45.

41. Boussy IA, Itoh M. Wanderings of hobo: a transposon in Drosophila melanogaster and its close relatives. Genetica. 2004;120:125–36.

42. Bucheton A, Vaury C, Chaboissier MC, Abad P, Pélisson A, Simonelig M. I elements and the Drosophila genome. Genetica. 1992;86:175–90.

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

44. Roiha H, Glover DM. Duplicated rDNA sequences of variable lengths flanking the short type I insertions in the rDNA of Drosophila melanogaster. Nucleic Acids Res. 1981;9:5521–32.

45. McAllister BF, Werren JH. Evolution of tandemly repeated sequences: What happens at the end of an array? J Mol Evol. 1999;48:469–81.

46. Danilevskaya O, Lofsky A, Kurenova EV, Pardue ML. The Y chromosome of Drosophila melanogaster contains a distinctive subclass of Het-A-related repeats. Genetics. 1993;134:531–43.

47. Agudo M, Losada A, Abad JP, Pimpinelli S, Ripoll P, Villasante A. Centromeres from telomeres? The centromeric region of the Y chromosome of Drosophila melanogaster contains a tandem array of telomeric HeT-A- and TART-related sequences. Nucleic Acids Research. 1999;27:3318–24.

48. Losada A, Abad JP, Villasante A. Organization of DNA sequences near the centromere of the Drosophila melanogaster Y chromosome. Chromosoma. 1997;106:503–12.

49. Berloco M, Fanti L, Sheen F, Levis RW, Pimpinelli S. Heterochromatic distribution of HeT-A- and TART-like sequences in several Drosophila species. Cytogenet Genome Res. 2005;110:124–33.

50. Biessmann H, Mason JM, Ferry K, d'Hulst M, Valgeirsdottir K, Traverse KL, et al. Addition of telomere-associated HeT DNA sequences "heals" broken chromosome ends in Drosophila. Cell. 1990;61:663–73.

51. Roegiers F, Kavaler J, Tolwinski N, Chou Y-T, Duan H, Bejarano F, et al. Frequent unanticipated alleles of lethal giant larvae in Drosophila second chromosome stocks. Genetics. 2009;182:407–10.

52. Mechler BM, McGinnis W, Gehring WJ. Molecular cloning of lethal(2)giant larvae, a recessive oncogene of Drosophila melanogaster. EMBO J. 1985;4:1551–7.

53. Green MM, Shepherd SH. Genetic instability in Drosophila melanogaster: the induction of specific chromosome 2 deletions by MR elements. Genetics. 1979;92:823–32.

54. Golubovsky MD. The “lethal giant larvae”-the most frequent second chromosome lethal in natural population of D. melanogaster. Drosoph Inf Serv. 1978;53:179–179.

55. Hoskins RA, Carlson JW, Wan KH, Park S, Mendez I, Galle SE, et al. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res. 2015;25:445–58.

56. Ke N, Voytas DF. High frequency cDNA recombination of the saccharomyces retrotransposon Ty5: The LTR mediates formation of tandem elements. Genetics. 1997;147:545–56.

57. Volff J-N, Körting C, Schartl M. Multiple Lineages of the Non-LTR Retrotransposon Rex1 with Varying Success in Invading Fish Genomes. Molecular Biology and Evolution. 2000;17:1673–84.

58. Malik HS, Eickbush TH. Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res. 2001;11:1187–97.

59. Biessmann H, Carter SB, Mason JM. Chromosome ends in Drosophila without telomeric DNA sequences. Proc Natl Acad Sci U S A. 1990;87:1758–61.

60. Levis RW. Viable deletions of a telomere from a Drosophila chromosome. Cell. 1989;58:791–801.

61. Cook DE, Zdraljevic S, Tanny RE, Seo B, Riccardi DD, Noble LM, et al. The Genetic Basis of Natural Variation in Caenorhabditis elegans Telomere Length. Genetics. 2016;204:371–83.

62. Fulcher N, Teubenbacher A, Kerdaffrec E, Farlow A, Nordborg M, Riha K. Genetic architecture of natural variation of telomere length in Arabidopsis thaliana. Genetics. 2015;199:625–35.

63. Gomes NMV, Shay JW, Wright WE. Telomere biology in Metazoa. FEBS Lett. 2010;584:3741–51.

64. Fanti L, Giovinazzo G, Berloco M, Pimpinelli S. The heterochromatin protein 1 prevents telomere fusions in Drosophila. Mol Cell. 1998;2:527–38.

65. Kijima TE, Innan H. Population genetics and molecular evolution of DNA sequences in transposable elements. I. A simulation framework. Genetics. 2013;195:957–67.

66. Walter MF, Biessmann MR, Benitez C, Török T, Mason JM, Biessmann H. Effects of telomere length in Drosophila melanogaster on life span, fecundity, and fertility. Chromosoma. 2007;116:41–51.

67. Pickett HA, Reddel RR. The role of telomere trimming in normal telomere length dynamics. Cell Cycle. 2012;11:1309–15.

68. Li JSZ, Miralles Fusté J, Simavorian T, Bartocci C, Tsai J, Karlseder J, et al. TZAP: A telomere-associated protein involved in telomere length control. Science. 2017;355:638–41.

69. Stage DE, Eickbush TH. Origin of nascent lineages and the mechanisms used to prime second-strand DNA synthesis in the R1 and R2 retrotransposons of Drosophila. Genome Biol. 2009;10:R49.

70. Ye J, Pérez-González CE, Eickbush DG, Eickbush TH. Competition between R1 and R2 transposable elements in the 28S rRNA genes of insects. Cytogenet Genome Res. 2005;110:299–306.

71. Morrish TA, Garcia-Perez JL, Stamato TD, Taccioli GE, Sekiguchi J, Moran JV. Endonuclease-

independent LINE-1 retrotransposition at mammalian telomeres. Nature. 2007;446:208–12.

72. Ryazansky S, Radion E, Mironova A, Akulenko N, Abramov Y, Morgunova V, et al. Natural variation of piRNA expression affects immunity to transposable elements. PLoS Genet. 2017;13:e1006731.

73. Savitsky M, Kwon D, Georgiev P, Kalmykova A, Gvozdev V. Telomere elongation is under the control of the RNAi-based mechanism in the Drosophila germline. Genes Dev. 2006;20:345–54.

74. Savitsky M, Kravchuk O, Melnikova L, Georgiev P. Heterochromatin protein 1 is involved in control of telomere elongation in Drosophila melanogaster. Mol Cell Biol. 2002;22:3204–18.

75. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110:462–7.

76. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks [Internet]. Journal of Statistical Mechanics: Theory and Experiment. 2008. p. P10008. Available from: http://dx.doi.org/10.1088/1742-5468/2008/10/p10008

77. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6:11.

78. Kelleher ES, Barbash DA. Analysis of piRNA-mediated silencing of active TEs in Drosophila melanogaster suggests limits on the evolution of host genome defense. Mol Biol Evol. 2013;30:1816–29.

79. Kaminker JS, Bergman CM, Kronmiller B, Carlson J, Svirskas R, Patel S, et al. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. Genome Biol. 2002;3:RESEARCH0084.

80. Kofler R, Betancourt AJ, Schlötterer C. Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genet. 2012;8:e1002487.

81. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

82. Gelman A. Two simple examples for understanding posterior p-values whose distributions are far from uniform. Electron J Stat. The Institute of Mathematical Statistics and the Bernoulli Society; 2013;7:2595–602.

83. Betancourt MJ, Girolami M. Hamiltonian Monte Carlo for Hierarchical Models [Internet]. arXiv [stat.ME]. 2013. Available from: http://arxiv.org/abs/1312.0906

84. Hogg DW, Bovy J, Lang D. Data analysis recipes: Fitting a model to data [Internet]. arXiv [astro-ph.IM]. 2010. Available from: http://arxiv.org/abs/1008.4686

85. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33:623–30.

Figure 1

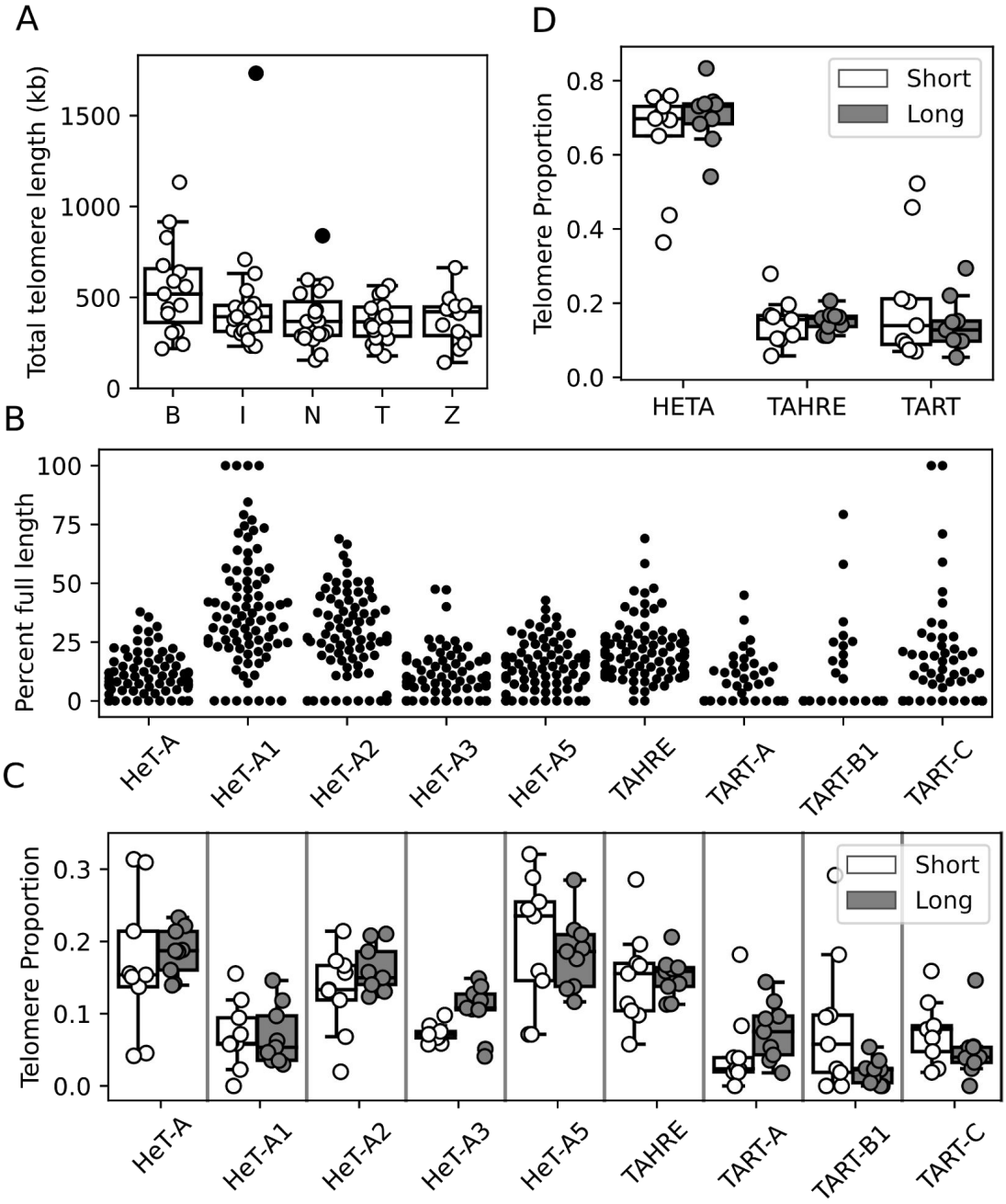
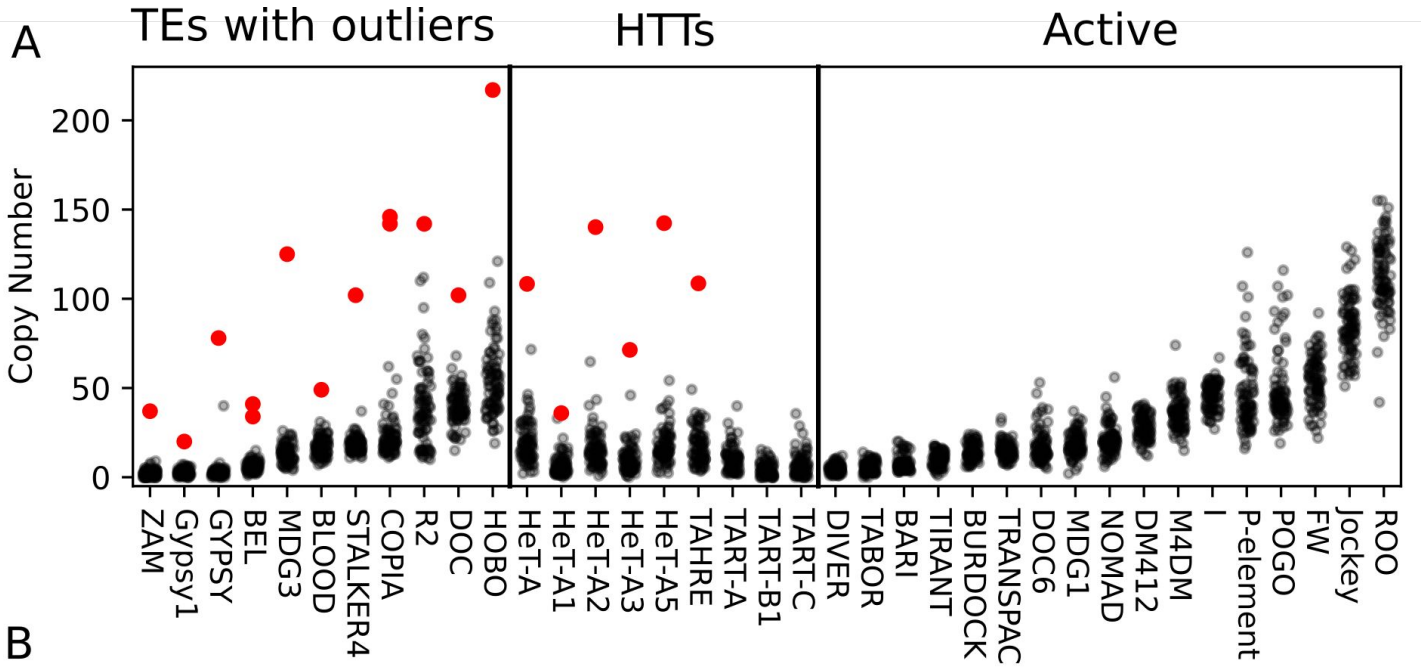

A
Total telomere length (kb)
1500
1000
500
0
B
I
N
T
Z
D
Telomere Proportion
0.8
0.6
0.4
0.2
0.0
Short
Long
HETA
TAHRE
TART
B
Percent full length
100
75
50
25
0
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C
C
Telomere Proportion
0.3
0.2
0.1
0.0
Short
Long
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C

Figure 2


A
TEs with outliers
HTTs
Active
Copy Number
0
50
100
150
200
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
BURDOCK
TRANSPAC
DOC6
MDG1
NOMAD
DM412
M4DM
I
P-element
POGO
FW
Jockey
ROO

B

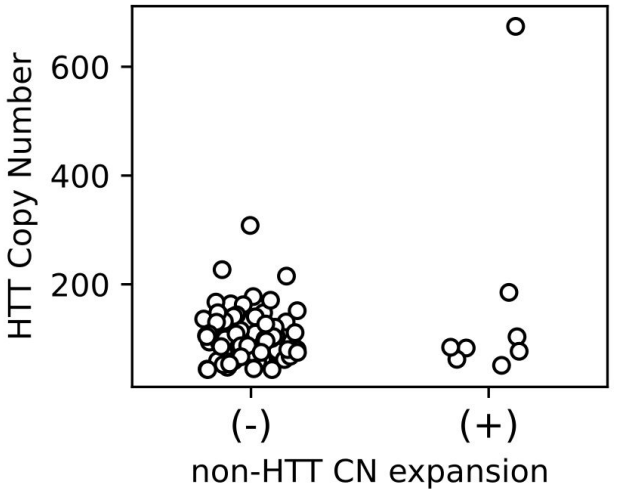

HTT Copy Number
200
400
600
(-)
(+)
non-HTT CN expansion

Figure 3

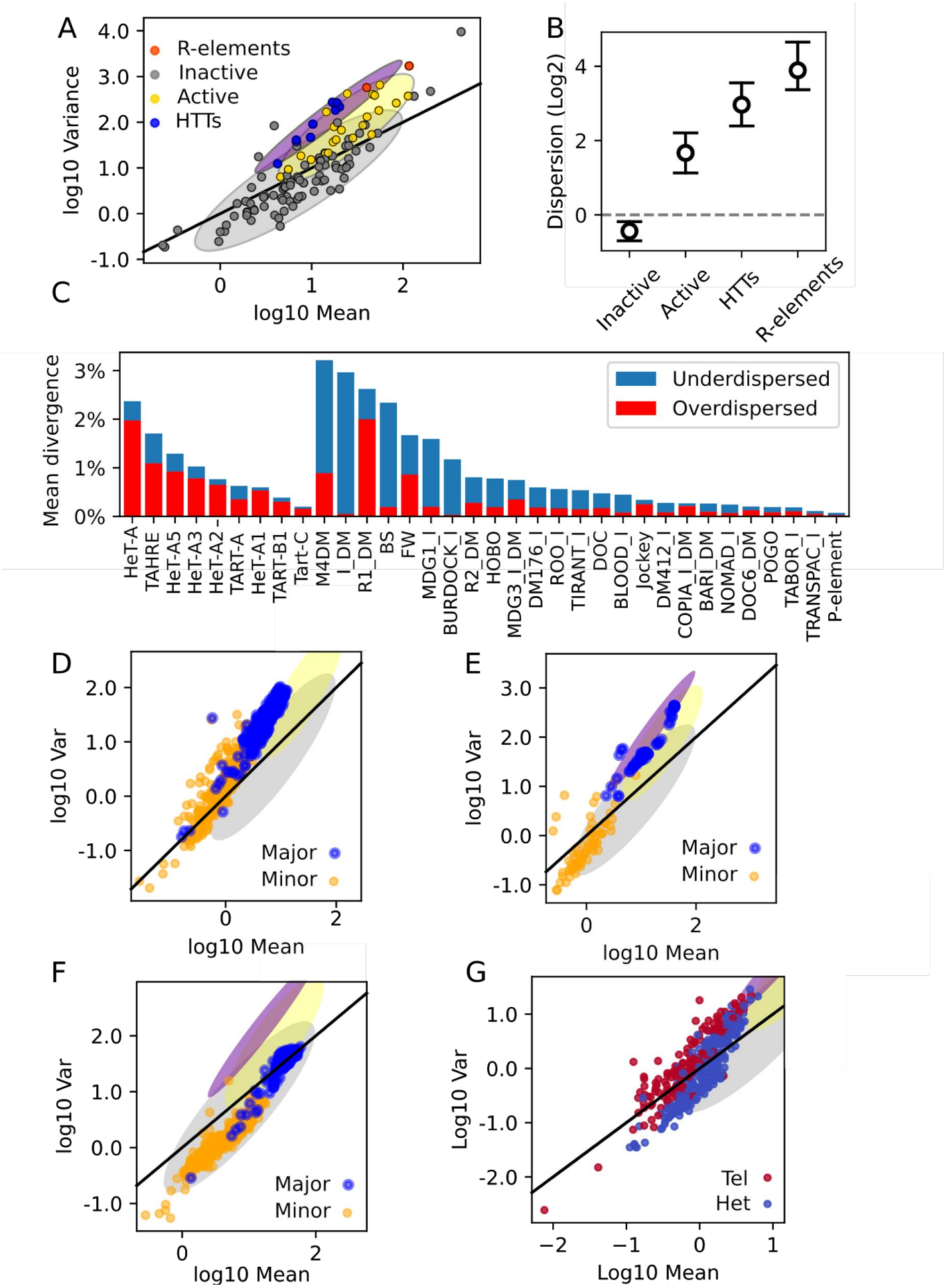

A
R-elements
Inactive
Active
HTTs
log10 Variance
log10 Mean
B
Dispersion (Log2)
Inactive
Active
HTTs
R-elements
C
Mean divergence
Underdispersed
Overdispersed
HeT-A
TAHRE
HeT-A5
HeT-A3
HeT-A2
TART-A
HeT-A1
TART-B1
Tart-C
M4DM
I_DM
R1_DM
BS
FW
MDG1_I
BURDOCK_I
R2_DM
HOBO
MDG3_I_DM
DM176_I
ROO_I
TIRANT_I
DOC
BLOOD_I
Jockey
DM412_I
COPIA_I_DM
BARI_DM
NOMAD_I
DOC6_DM
POGO
TABOR_I
TRANSPAC_I
P-element
D
log10 Var
log10 Mean
Major
Minor
E
log10 Var
log10 Mean
Major
Minor
F
log10 Var
log10 Mean
Major
Minor
G
Log10 Var
Log10 Mean
Tel
Het

Figure 4

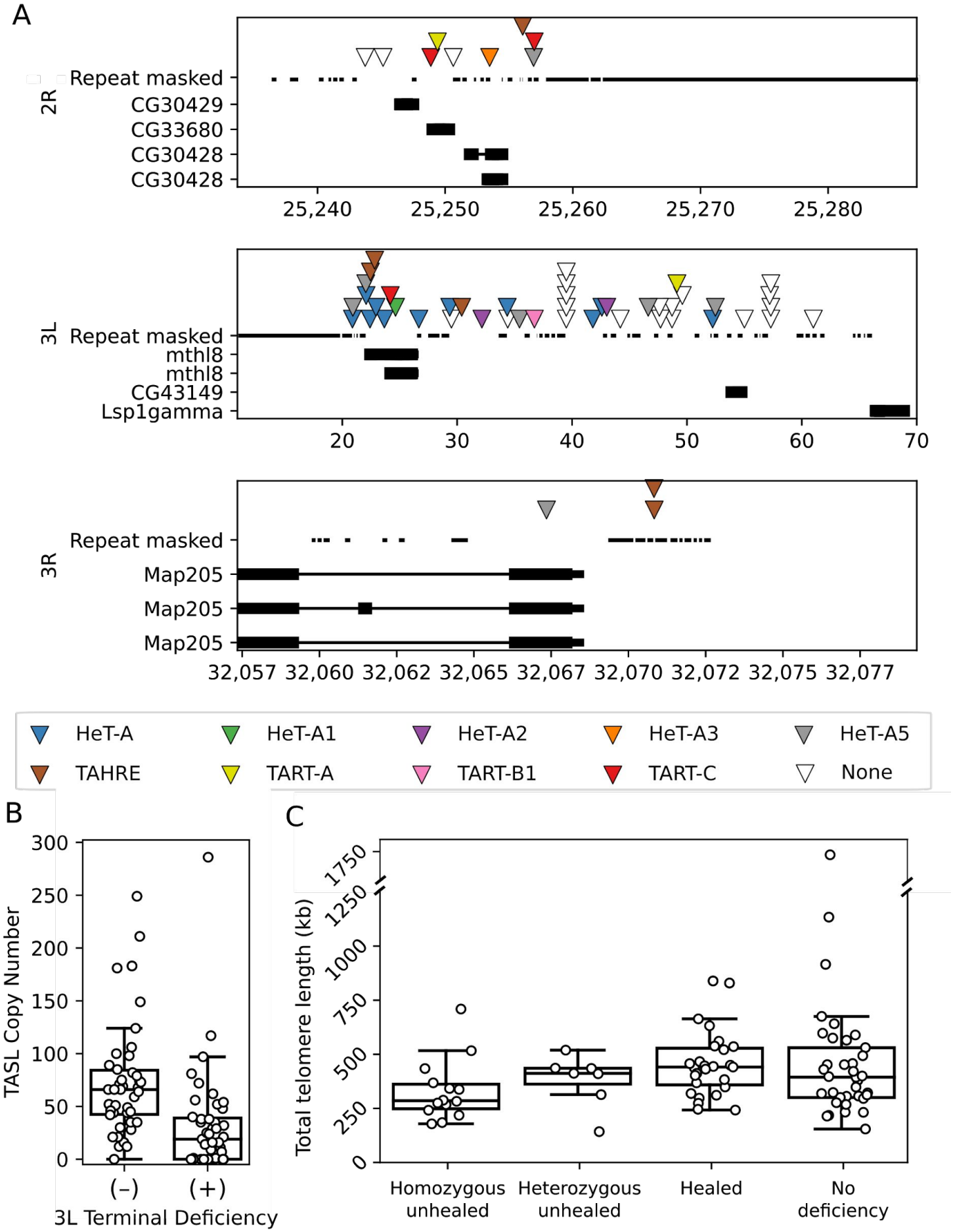

A
2R
Repeat masked
CG30429
CG33680
CG30428
CG30428
25,240
25,250
25,260
25,270
25,280
3L
Repeat masked
mthl8
mthl8
CG43149
Lsp1gamma
20
30
40
50
60
70
3R
Repeat masked
Map205
Map205
Map205
32,057
32,060
32,062
32,065
32,067
32,070
32,072
32,075
32,077
HeT-A
HeT-A1
HeT-A2
HeT-A3
HeT-A5
TAHRE
TART-A
TART-B1
TART-C
None
B
TASL Copy Number
300
250
200
150
100
50
0
(–)
(+)
3L Terminal Deficiency
C
Total telomere length (kb)
1750
1250
1000
750
500
250
0
Homozygous unhealed
Heterozygous unhealed
Healed
No deficiency

Figure 5

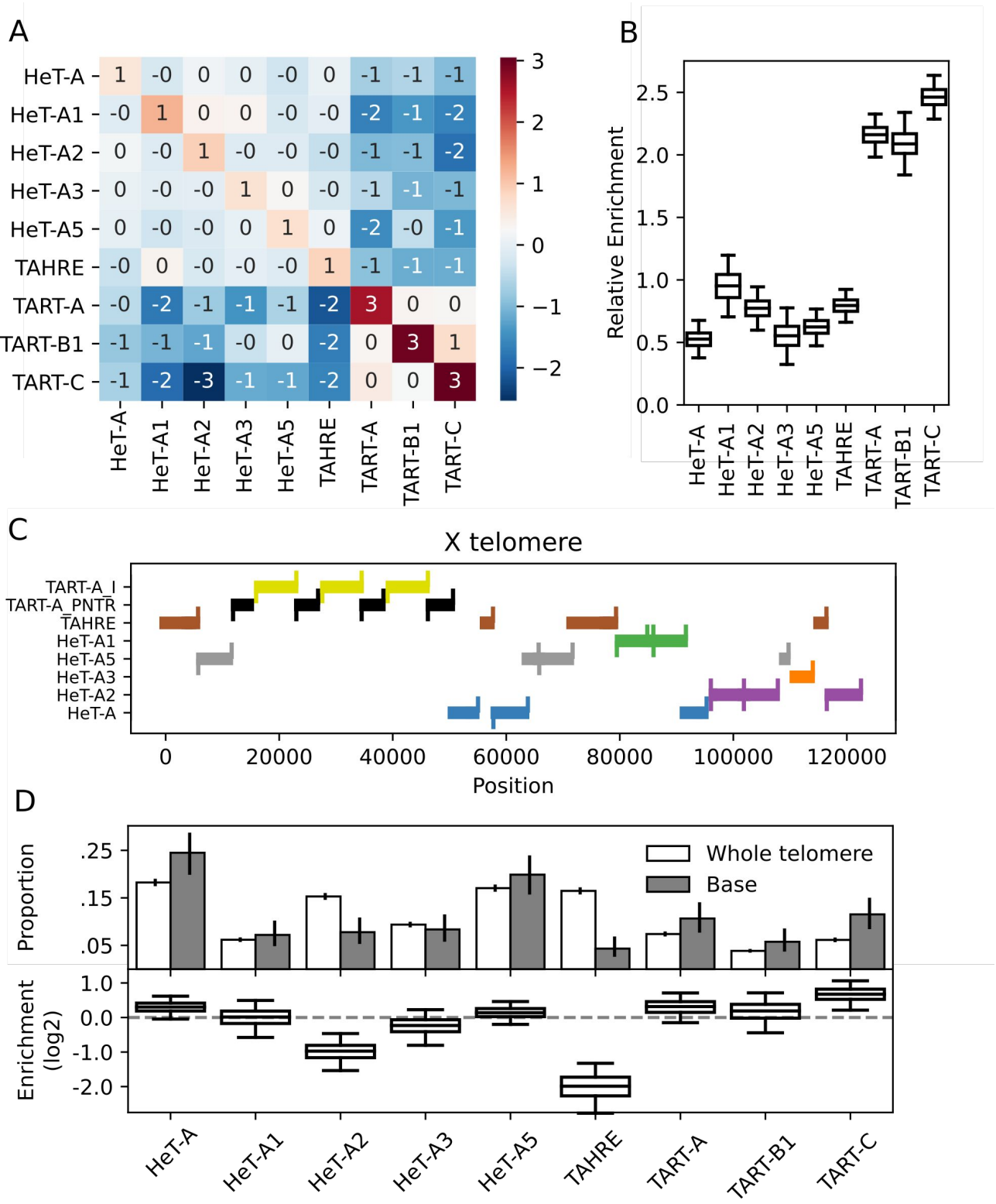

A
HeT-A 1 -0 0 0 -0 0 -1 -1 -1
HeT-A1 -0 1 0 0 -0 -0 -2 -1 -2
HeT-A2 0 -0 1 -0 -0 -0 -1 -1 -2
HeT-A3 0 -0 -0 1 0 -0 -1 -1 -1
HeT-A5 0 -0 -0 0 1 0 -2 -0 -1
TAHRE -0 0 -0 -0 -0 1 -1 -1 -1
TART-A -0 -2 -1 -1 -1 -2 3 0 0
TART-B1 -1 -1 -1 -0 0 -2 0 3 1
TART-C -1 -2 -3 -1 -1 -2 0 0 3
HeT-A HeT-A1 HeT-A2 HeT-A3 HeT-A5 TAHRE TART-A TART-B1 TART-C
3 2 1 0 −1 −2
B
Relative Enrichment
0.0 0.5 1.0 1.5 2.0 2.5
HeT-A HeT-A1 HeT-A2 HeT-A3 HeT-A5 TAHRE TART-A TART-B1 TART-C
C
X telomere
TART-A_I
TART-A_PNTR
TAHRE
HeT-A1
HeT-A5
HeT-A3
HeT-A2
HeT-A
0 20000 40000 60000 80000 100000 120000
Position
D
Proportion
.25 .15 .05
Whole telomere
Base
Enrichment (log2)
1.0 0.0 -1.0 -2.0
HeT-A HeT-A1 HeT-A2 HeT-A3 HeT-A5 TAHRE TART-A TART-B1 TART-C

Figure S1

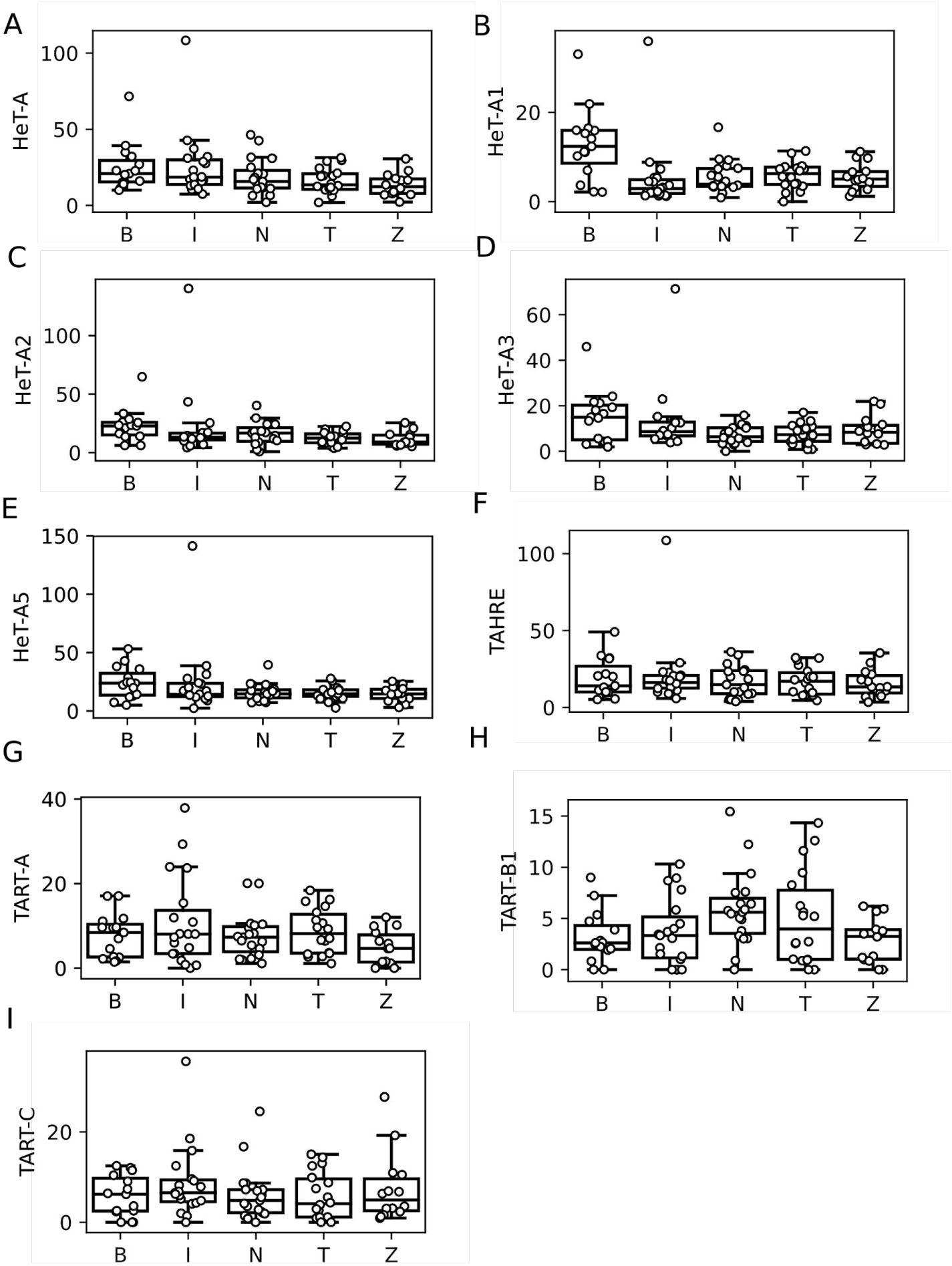

A
HeT-A
100
50
0
B I N T Z
B
HeT-A1
20
0
B I N T Z
C
HeT-A2
100
50
0
B I N T Z
D
HeT-A3
60
40
20
0
B I N T Z
E
HeT-A5
150
100
50
0
B I N T Z
F
TAHRE
100
50
0
B I N T Z
G
TART-A
40
20
0
B I N T Z
H
TART-B1
15
10
5
0
B I N T Z
I
TART-C
20
0
B I N T Z

Figure S2

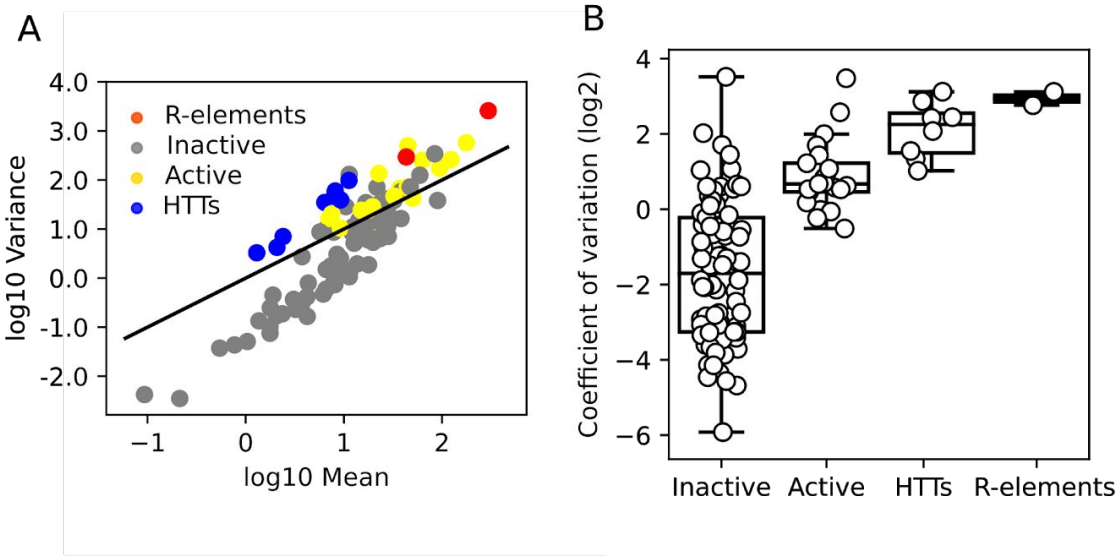

A
R-elements
Inactive
Active
HTTs
log10 Variance
log10 Mean
4.0
3.0
2.0
1.0
0.0
-1.0
-2.0
−1
0
1
2
B
Coefficient of variation (log2)
4
2
0
−2
−4
−6
Inactive
Active
HTTs
R-elements

Figure S3

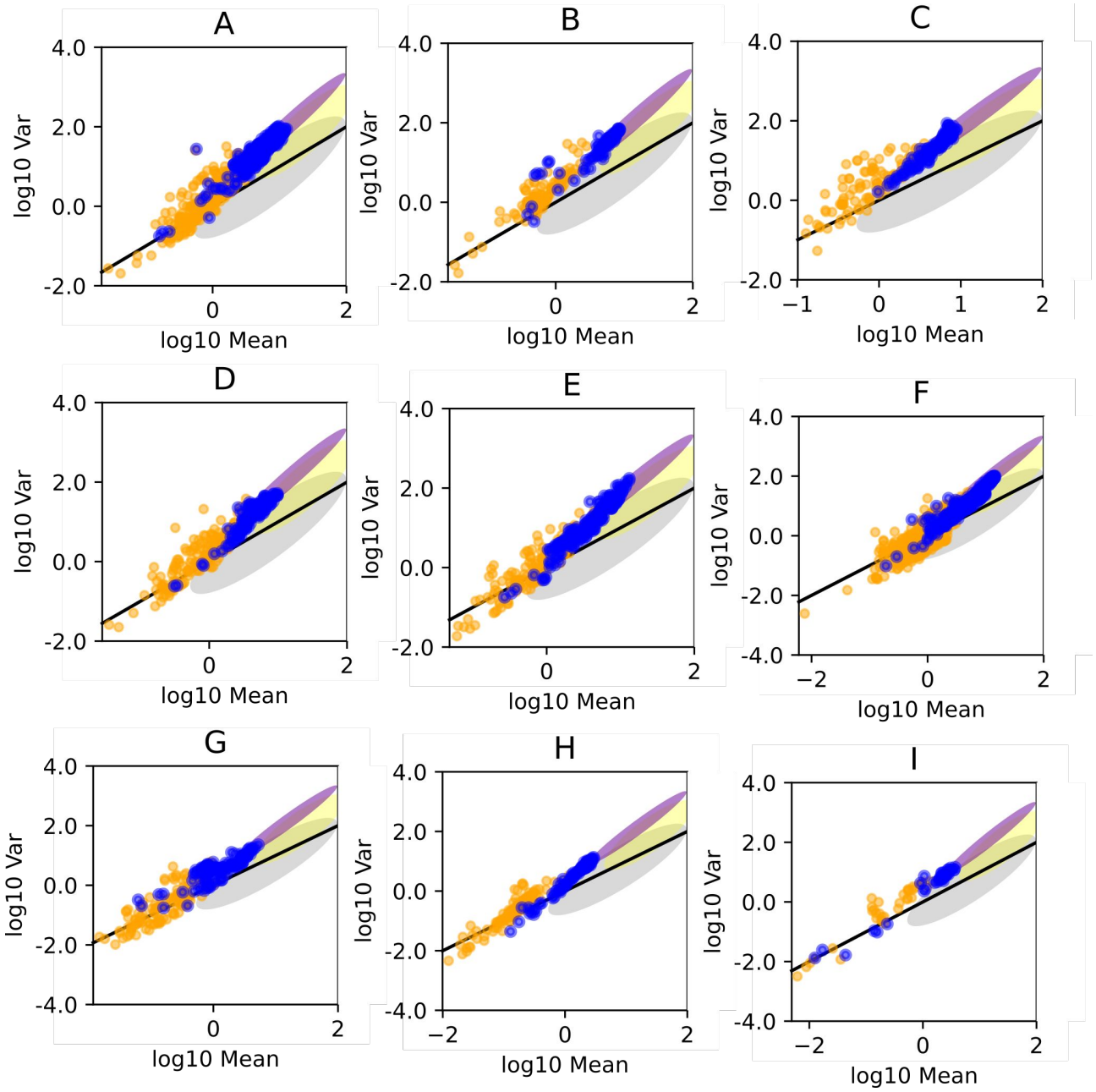


Figure S4

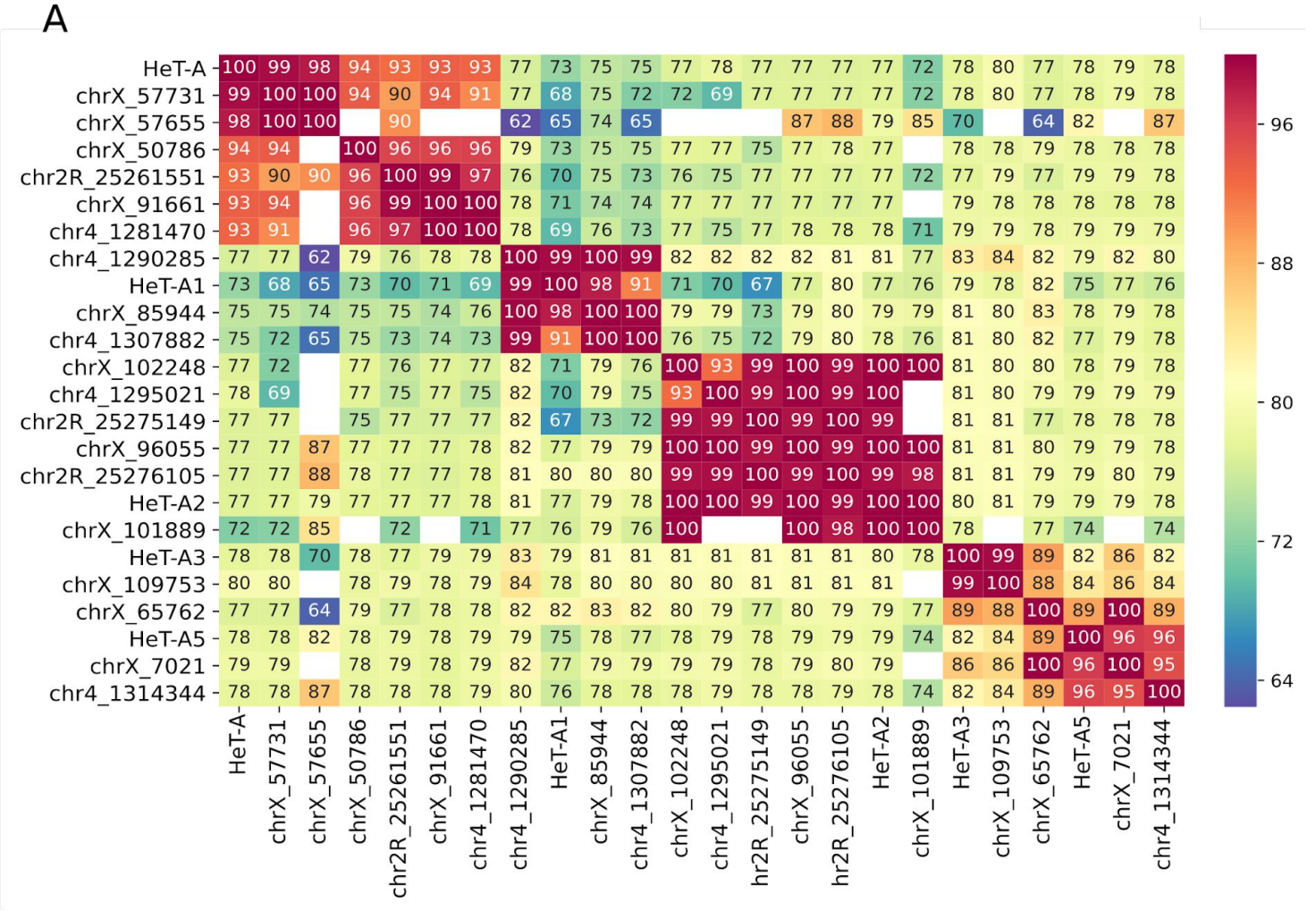
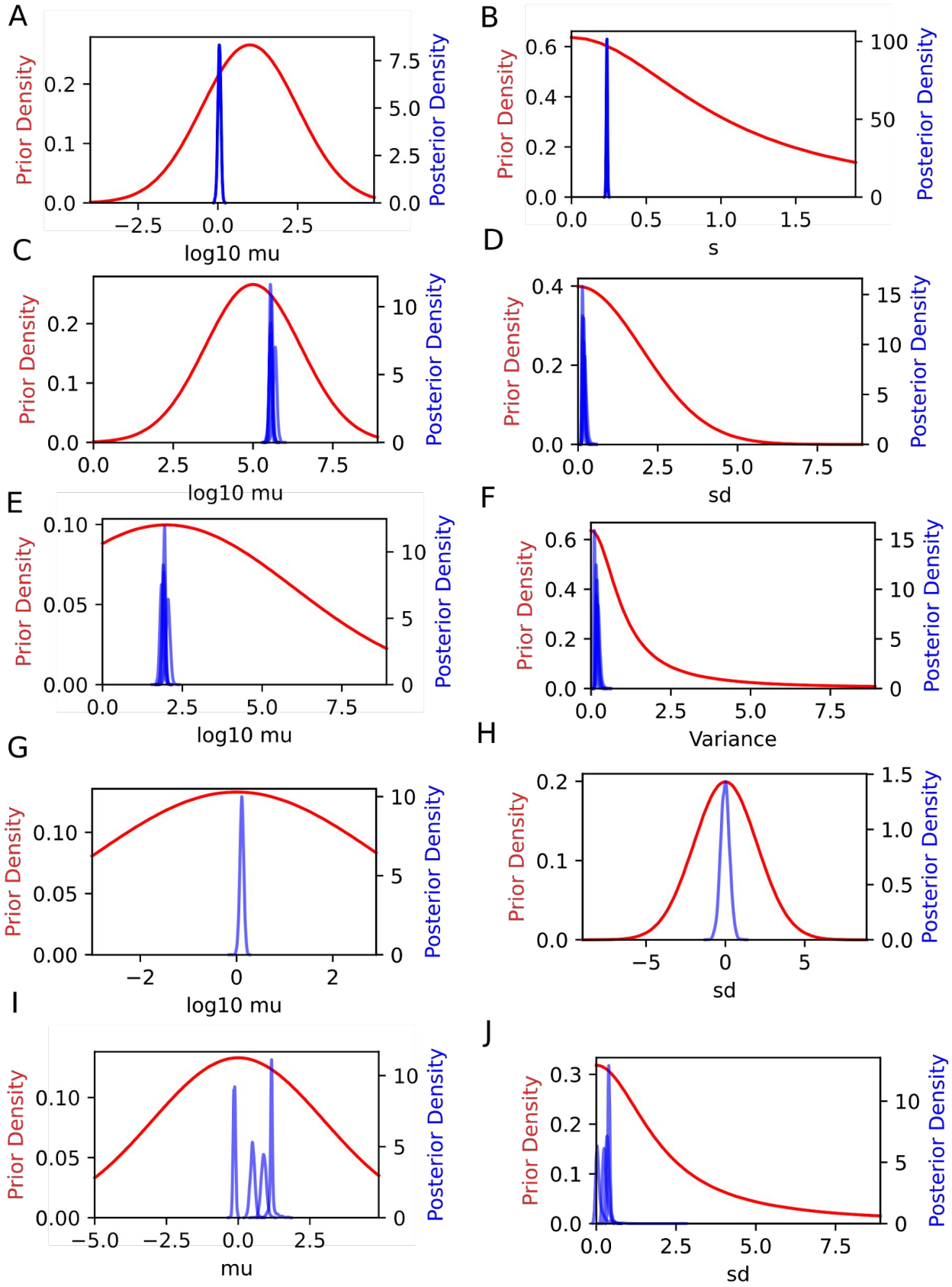

A
HeT-A 100 99 98 94 93 93 93 77 73 75 75 77 78 77 77 77 77 72 78 80 77 78 79 78
chrX_57731 99 100 100 94 90 94 91 77 68 75 72 72 69 77 77 77 77 72 78 80 77 78 79 78
chrX_57655 98 100 100 90 62 65 74 65 87 88 79 85 70 64 82 87
chrX_50786 94 94 100 96 96 96 79 73 75 75 77 77 75 77 78 77 78 78 79 78 78 78
chr2R_25261551 93 90 90 96 100 99 97 76 70 75 73 76 75 77 77 77 77 72 77 79 77 79 79 78
chrX_91661 93 94 96 99 100 100 78 71 74 74 77 77 77 77 77 77 79 78 78 78 78 78
chr4_1281470 93 91 96 97 100 100 78 69 76 73 77 75 77 78 78 78 71 79 79 78 79 79 79
chr4_1290285 77 77 62 79 76 78 78 100 99 100 99 82 82 82 82 81 81 77 83 84 82 79 82 80
HeT-A1 73 68 65 73 70 71 69 99 100 98 91 71 70 67 77 80 77 76 79 78 82 75 77 76
chrX_85944 75 75 74 75 75 74 76 100 98 100 100 79 79 73 79 80 79 79 81 80 83 78 79 78
chr4_1307882 75 72 65 75 73 74 73 99 91 100 100 76 75 72 79 80 78 76 81 80 82 77 79 78
chrX_102248 77 72 77 76 77 77 82 71 79 76 100 93 99 100 99 100 100 81 80 80 78 79 78
chr4_1295021 78 69 77 75 77 75 82 70 79 75 93 100 99 100 99 100 81 80 79 79 79 79
chr2R_25275149 77 77 75 77 77 77 82 67 73 72 99 99 100 99 100 99 81 81 77 78 78 78
chrX_96055 77 77 87 77 77 77 78 82 77 79 79 100 100 99 100 99 100 100 81 81 80 79 79 78
chr2R_25276105 77 77 88 78 77 77 78 81 80 80 80 99 99 100 99 100 99 98 81 81 79 79 80 79
HeT-A2 77 77 79 77 77 77 78 81 77 79 78 100 100 99 100 99 100 100 80 81 79 79 79 78
chrX_101889 72 72 85 72 71 77 76 79 76 100 100 98 100 100 78 77 74 74
HeT-A3 78 78 70 78 77 79 79 83 79 81 81 81 81 81 81 81 80 78 100 99 89 82 86 82
chrX_109753 80 80 78 79 78 79 84 78 80 80 80 80 81 81 81 81 99 100 88 84 86 84
chrX_65762 77 77 64 79 77 78 78 82 82 83 82 80 79 77 80 79 79 77 89 88 100 89 100 89
HeT-A5 78 78 82 78 79 78 79 79 75 78 77 78 79 78 79 79 79 74 82 84 89 100 96 96
chrX_7021 79 79 78 79 78 79 82 77 79 79 79 79 78 79 80 79 86 86 100 96 100 95
chr4_1314344 78 78 87 78 78 78 79 80 76 78 78 78 79 78 78 79 78 74 82 84 89 96 95 100
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
hr2R_25275149
chrX_96055
hr2R_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
OLS fit
One--to-one
From junctions (kb)
From read depth (kb)
2250
2000
1750
1500
1250
1000
750
500
250
0
0
500
1000
1500
2000

Figure S5


A
Prior Density
Posterior Density
log10 mu
B
s
C
log10 mu
D
sd
E
log10 mu
F
Variance
G
log10 mu
H
sd
I
mu
J
sd