# Population-scale long-read sequencing uncovers transposable elements contributing to gene expression variation and associated with adaptive signatures in *Drosophila melanogaster*

Gabriel E. Rech[1], Santiago Radío[1], Sara Guirao-Rico[1], Laura Aguilera[1], Vivien Horvath[1], Llewellyn Green[1], Hannah Lindstadt[1], Véronique Jamilloux[2], Hadi Quesneville[2] and Josefa González[1]

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain
[2]URGI, INRA, Université Paris-Saclay, Versailles, France

**Corresponding author:**

Josefa González

josefa.gonzalez@ibe.upf-csic.es

## ABSTRACT

High quality reference genomes are crucial to understanding genome function, structure and evolution. The availability of reference genomes has allowed us to start inferring the role of genetic variation in biology, disease, and biodiversity conservation. However, analyses across organisms demonstrate that a single reference genome is not enough to capture the global genetic diversity present in populations. In this work, we generated 32 high-quality reference genomes for the well-known model species *D. melanogaster* and focused on the identification and analysis of transposable element variation as they are the most common type of structural variant. We showed that integrating the genetic variation across natural populations from five climatic regions increases the number of detected insertions by 58%. Moreover, 26% to 57% of the insertions identified using long-reads were missed by short-reads methods. We also identified hundreds of transposable elements associated with gene expression variation and new TE variants likely to contribute to adaptive evolution in this species. Our results highlight the importance of incorporating the genetic variation present in natural populations to genomic studies, which is essential if we are to understand how genomes function and evolve.

## Introduction

Despite their crucial role and high prevalence in most eukaryotic genomes, transposable elements (TEs) and other structural variants (SVs) remain largely understudied, mainly as a consequence of the limitations of high throughput sequenced reads, tightly restricted to short-reads in the last decades (Audano *et al.* 2019; De Coster & Van Broeckhoven 2019; Huddleston & Eichler 2016). Short-reads not only limited the annotation of SVs to what inference methods were able to identify (Chaisson *et al.* 2019; Chakraborty *et al.* 2019; Chakraborty *et al.* 2018; Kou *et al.* 2020; Mahmoud *et al.* 2019; Zhou *et al.* 2019), but also required a reference genome to map the reads, which has at least three major drawbacks: (i) the information about the genetic background and genomic context of the SVs are usually lost (Chaisson *et al.* 2019); (ii) the analyses are biased to what is possible to identify using a specific reference genome (Audano *et al.* 2019; Ballouz *et al.* 2019; Yang *et al.* 2019); and (iii) repetitive sequences in the reference genome are not well characterized when they are longer than the sequenced reads (Treangen & Salzberg 2011). In the particular case of TEs, the limitations of using short-reads are exacerbated even further for two reasons: sequence divergence of the copies, and their extremely repetitive nature (Goerner-Potvin & Bourque 2018). Such a complexity has severely restricted inter- and intra-species TE dynamics studies, a crucial aspect that needs to be addressed in order to better understand the organization, function, and evolution of genomes (Barron *et al.* 2014).

During the last years, technological developments in DNA sequencing read length have lead not only to an improvement in the quality and completeness of reference genomes (Chaisson *et al.* 2015; Du & Liang 2019; Jain *et al.* 2018; Jiao *et al.* 2017; Miga *et al.* 2020; Solares *et al.* 2018), but also to a significant rise in the number of high-quality genomes for multiple individuals of the same species, opening a new era in comparative population genomics (Mitsuhashi & Matsumoto 2020; Sakamoto *et al.* 2020). The ability of long-reads to span repetitive regions of the genome, together with the relative low price of generating sequences for several individuals, has opened up the possibility of resolving and comparing previously absent or misassembled regions in the genome (Alonge *et al.* 2020; Audano *et al.* 2019; Chakraborty *et al.* 2019; Levy-Sakin *et al.* 2019; Liu *et al.* 2020), which can lead to a significant improvement in our ability to study TE structure, activity and dynamics in different organisms (Jiao *et al.* 2017; Michael *et al.* 2018; Shahid & Slotkin 2020).

*Drosophila melanogaster* represents one of the best model animals for studying TEs, not only for having one of the best annotated eukaryotic genomes (Gramates *et al.* 2017; Thurmond *et al.* 2018), but also for containing several active TE families (Lerat *et al.* 2019). Interestingly, even in such a well-studied organism, long-read sequencing approaches have made novel insights into the evolutionary dynamics of TEs (Chakraborty *et al.* 2019; Ellison & Cao 2019; Mohamed *et al.* 2020). However, these studies do not take full advantage of the variability present in the populations analyzed, as they mainly use standard homology-based approaches (e.g. *RepeatMasker* and *RepBase*) for annotating and analyzing TEs, which limits their analysis to TE families already present in the available libraries.

Here, we used long-read sequences to generate high quality genome assemblies for 32 *D. melanogaster* natural strains collected mainly in Europe from populations located in five different climatic regions and belonging to three of the five main climate types (Figure 1). We used this new genomic resource for the *de novo* reconstruction and manual curation of a library of consensus TE sequences that account for the variability observed in natural populations. Genome annotations performed with this manually curated library of TEs not only outperformed the current *D. melanogaster* gold-standard TE annotation (FlyBase), but also showed significant improvements regarding the state-of-the-art short-read-based methods for TE annotation. Furthermore, a joint in-depth analysis of TE copies annotated in the 32 newly sequenced genomes, 14 additional worldwide high-quality genomes, and the reference genome, revealed that analyzing 20 genomes is sufficient to recover most of the common genetic variation in out-of-Africa *D. melanogaster* natural populations; revealed hundreds of TEs associated with changes in expression of their nearby genes; and. allowed to identify 31% more TEs with evidence of positive selection compared with the previous most extensive analysis

## RESULTS

### Thirty-two highly complete *D. melanogaster* genomes in terms of genes and transposable elements

In order to access as much TE diversity as possible in natural populations of *D. melanogaster*, we performed sequencing and *de novo* genome assembly of 32 strains using long-read sequencing technologies (Figure 1, Table 1, Table S1 and Table S2, Supplementary File S1.1). Most of these strains —24 out of 32— were collected from 11 geographical locations across Europe, while the other eight strains were originally taken from Raleigh, North Carolina, USA (Huang et al 2014). These 12 populations represent five different climatic regions belonging to three main climatic types: arid, temperate, and cold (Figure 1; Table S1). Long-read sequencing resulted in 458.7 Gb, representing a theoretical average coverage of 82X (ranging from 45X to 123X) and average read length > 5.6Kb, which has been previously shown to be sufficient for generating highly contiguous genome assemblies in other Drosophila species (Miller *et al.* 2018; Table S2). Genome assembly, polishing, deduplication and contaminant removal resulted in genomes with a number of contigs ranging from 153 to 1,185 (average 367), genome sizes from 136.6Mb to 151.3Mb (average 142Mb), N50 values from 400Kb to 18.9Mb (average 3.8Mb) complete BUSCO scores between 96.1% and 99%, and per base quality values (QV scores) between 37.2 and 52.9 (Table 1 and Supplementary File S1.2-3). Although the high variability, these results are comparable with genomes previously obtained using similar sequencing and assembling strategies (Miller *et al.* 2018). Note that differences in sequencing coverage did not explain the observed differences in genome size or TE content across genomes (Figure S1). Similarly, differences in read length and N50 values do not correlate with differences in genome size, TE content, or BUSCO scores (Figure S1).

After reference-guided scaffolding using the ISO1 reference genome, on average >90% of the contigs mapped to major chromosomal arms, which contained >98.5% of the bases in the *de novo* assembled genomes (Table S3A). Scaffolding also significantly increased CUSCO scores (percentage of contiguously assembled piRNA clusters; (Wierzbicki *et al.* 2020): average CUSCO score increased from 64.1% at the contig level to 93.7% at the scaffolding level (sc.CUSCO; Table 1, Table S3B and Supplementary File S1.5). The detectability of a cluster was inversely correlated with its size (Pearson´s correlation = -0.47; Table S3B, Figure S2 and Supplementary File S1.5). In addition to high sc.CUSCO values, the scaffolded genomes also showed a high level of completeness, covering on average around 95% of ISO1 major chromosomal arms (Table 1, Figure S3) and with an average of 99.75% of the protein coding genes successfully transferred (Table S3A).

To quantify the accuracy of the TE sequences generated with long-read sequencing, we used our pipeline (from base calling to genome scaffolding) to process the ONT long-reads available for the reference genome (Solares *et al.* 2018). The newly assembled reference genome was 147.8Mb with a complete BUSCO score of 96% (Table 1). We identified 1,842 orthologous TE insertions between our assembly and the FlyBase reference genome, with 99.9% pairwise identity suggesting that our pipeline produces highly accurate TE sequences (Table S4).

Overall, we generated 32 *de novo D. melanogaster* assembled genomes from 12 geographically diverse populations that are contiguous and complete in terms of gene and TE content.

## A new manually curated library of consensus sequences allowed the annotation of 58% more TE copies in the high-quality *D. melanogaster* reference genome

In order to accurately annotate TE copies in the 32 *de novo* assembled genomes of *D. melanogaster*, we implemented a TE annotation strategy involving, as a first step, the generation of a manually curated TE (MCTE) library. The MCTE library was built using the *REPET TEdenovo* pipeline for the *de novo* prediction of consensus sequences representative of TE families (Flutre *et al.* 2011). Because the library required extensive manual curation, we focused on 13 genomes that represent the 12 geographical locations in our analysis (Table 1). Overall, the *TEdenovo* pipeline reconstructed 28,009 consensus sequences. After manual curation (Supplementary File S1.6), the MCTE library ended up with 165 consensus sequences, which are 34 more sequences than the ones present in the Berkeley Drosophila Genome Project (BDGP) dataset for *D. melanogaster* (Kaminker *et al.* 2002) (Table S5). The MCTE library sequences are representative of 146 TE families (13 of them represented by more than one consensus sequence), including three new families (see below).

The second step of the annotation process used the *TEannot* pipeline of *REPET* to annotate all the TEs present in each one of the 32 genomes and the reference ISO1 genome using the MCTE library. The euchromatic region analyzed ranged from 100.1Mb to 103.9Mb (Table 1), which is a slightly larger region than previous similar analysis (e.g. 94.5Mb in Chakraborty *et al.* 2019). As a proof of concept, we compared the euchromatic TE annotation performed with *REPET* with the current TE annotation available in FlyBase, which is considered the gold-standard (Gramates *et al.* 2017). We found that all but two families in FlyBase were present in the *REPET* annotations: *frogger* and *gypsy3*, with only one copy each annotated in FlyBase. *REPET* most likely fails to detect the *frogger* copy because it is nested in a *copia1* insertion, while the only copy of *gypsy3* is annotated in the heterochromatin and thus not included in our *REPET* annotations. When considering only those families present in both annotations, we observed no significant differences in the number of copies between REPET and FlyBase annotations (FDR p-value > 0.05, $X^2$ test, Figure 2A, Table S6A), with the exception of the INE-1 elements, for which REPET annotated a larger number of copies than FlyBase (FDR p-value < 0.0001, $X^2$ test, Table S6A). At the genomic coordinates level, ~85% of the FlyBase copies were overlapping with *REPET* copies (95% reciprocal minimum breadth of coverage; Figure 2B, Table S6B). Moreover, overall sensitivity and specificity of *REPET* annotation when comparing with FlyBase were 99.44% and 99.29%, respectively (calculated according to Quesneville *et al.* (2005); Table S6C). Thus, overall the annotation of the reference genome performed with the MCTE library was able to reproduce with high accuracy the FlyBase TE annotation, the current gold-standard TE annotation in *D. melanogaster* (Thurmond *et al.* 2018).

However, while the number of copies and the coordinates of TEs from families present in both annotations were very similar, our annotation strategy allowed us to annotate 468 copies from 28 TE families not present in the FlyBase annotation. While most of them correspond to known TE families, such as *LARD*, *Kepler* and *THARE*, 27 copies correspond to three new TE families (see below). Moreover, 15 copies belong to families such as *gypsy10*, *BS4* and *ZAM,* which according to FlyBase were only present in the heterochromatic regions, but we found them in euchromatic regions as well (Table S6A, Figure 2A). Although most of the new TE copies annotated only with *REPET* were small insertions, we also identified 50 insertions larger than 2Kb (Figure 2C, Supplementary File S1.7).

We further compared the number of TEs annotated in the 13 genomes with the previously available *D. melanogaster* BDGP library and with the MCTE library (Table S6D, Supplementary File S1.7). We found that 42% to 44% of the copies annotated using the MCTE library were not annotated by the BDGP library.

Overall, by creating a library that contains the TE diversity of 13 *D. melanogaster* strains from 12 geographical locations, we were able to identify TE copies from 25 known families not previously annotated in the reference euchromatic genome, and from three new families (see below). In total, 58% more insertions were annotated in the euchromatic reference genome using the MCTE library (1,301 FlyBase vs 2,059 *REPET*), and 42-44% more copies were identified using the MCTE library compared with the BDGP library when analyzing 13 other genomes.

## The new manually curated TE library allowed the identification of three new families in *D. melanogaster*, two of them also present in other Drosophila species

Three consensus sequences in the MCTE library that failed to be assigned to any known family in the BDGP or the *RepBase* database were further analyzed using *PASTEC* (Hoede *et al.* 2014). These new consensus sequences were classified as a Miniature Inverted Repeat Transposable Element (*MITE*), a Terminal Repeat Retrotransposons in Miniature (*TRIM*), and a Terminal Inverted Repeat (*TIR*) element (Figure 3A).

Numerous Bari-like *MITEs* (Palazzo *et al.* 2016) and Mariner-like *MITEs* (Wallau *et al.* 2014) have been previously described in *D. melanogaster*. However, the *MITE* consensus sequence showed no significant alignments with any previously described *MITEs* (nucleotide identity percentage < 50%), suggesting that they belong to a new undescribed *MITE* family. On average, more than eight *MITE* copies were found in each *D. melanogaster* strain. Identified copies were of variable length (Figure S4A) and highly similar (average identity >89%, Figure S5). Moreover, the consensus sequence of the new *MITE* family showed no significant similarities with TEs identified in other five Drosophila genomes (Table S7, Supplementary File S1.8), suggesting that this element could have invaded the *D. melanogaster* genome recently.

Regarding the new *TRIM* element, while the consensus sequence showed the typical *TRIM* structure (less than 1,000bp, with LTRs sequences between 100bp - 250bp, Figure 3A), no similarities with any known TE in public databases was found. Notably, this sequence was not the only *TRIM* element in the MCTE library since other *TRIM* consensus sequence showing similarities with a *Kepler* element was also found. Most copies of the new *TRIM* element have the size of the consensus sequence (Figure 3B), however we found relatively low similarity among the copies (average identity 77%, Figure S5) and evidences that the element is present in at least another Drosophila species (*D. pseudoobscura*, Table S7, Supplementary File S1.8), suggesting that this *TRIM* element could represent the remains of an ancestral TE family.

Finally, the newly identified *TIR* element showed 51% sequence similarity to the internal domain of *EnSpm-1_JC*, a *TIR* element from the *Jatropha curcas* genome (Kojima & Jurka 2011) (Figure 3A and Figure S4B). Moreover, while the consensus sequence did not actually contain the inverted repeats at the ends (*TIRs*), we found 31%-43% of the copies annotated in each of the 32 genomes to contain degraded inverted repeats in the 1kb flanking regions (Supplementary File S1.9). Besides, average copy identity per genome was low (68%, Figure S5) and most copies were truncated representations of the consensus (Figure 3B). These results, coupled with the similarity showed by the new *TIR* element against TE consensus sequences from *D. virilis* and *D. bipectinata* (Table S7, Supplementary File S1.8), suggested an ancient origin for this element.

Thus, even in a well-studied species as *D. melanogaster*, the *de novo* TE annotation and manual curation using a long-read strategy in a geographically diverse panel of strains allowed the identification of three new TE families. Copies from two of these families (*TRIM* and *TIR* elements) showed low levels of similarity suggesting that they are old insertions; while copies of the new *MITE* family were highly similar suggesting that it might have recently transposed.

## Short-read methods failed to detect up to 57% of the insertions detected by long-read based annotation

Besides comparing our TE annotations with those available in FlyBase, we also wanted to investigate how *de novo* annotations based on long-read sequencing assemblies compare with annotations based on short-read sequencing. Previous estimates suggested that short-reads failed to find 36%-38% of the TE insertions annotated based on long-reads (Chakraborty *et al.* 2019; Chakraborty *et al.* 2018). To estimate this percentage in our genomes, we compared the results obtained with the MCTE library in long-reads using *REPET*, and in short-reads using two different tools: *TEMP* (Zhuang *et al.* 2014) and *TIDAL* (Rahman *et al.* 2015) (Table S8). For this comparison, we focused on 11 of the most complete genomes representative of the geographic variability of our samples and included in the previous subset of 13 genomes used to build the MCTE library (Table 1, Supplementary File S1.10).

The total number of TE insertions detected by each software was more similar for *REPET* and *TEMP* (6,632 and 7,430, respectively) than for *TIDAL* (9,066) (Table S8A). The number of TE insertions detected both by *REPET* and *TIDAL* (4,041) is higher that the number of TE insertions detected by *REPET* and *TEMP* (3,254). The overlap of the insertions detected both by *TIDAL* and *TEMP* is higher (4,786), probably because the methodologies of these two software are more similar (Table S8A).

To estimate the false negative rate of *TEMP* and *TIDAL* and the false positive rate of *REPET*, we performed manual inspection for 300 TE insertions annotated by *REPET*. When comparing the TE annotations between *REPET* and *TEMP*, 120 TEs (40%) were correctly annotated by the two software, while 170 (57%) TEs annotated by REPET were missed by *TEMP* (Table S8B). When comparing *REPET* and *TIDAL* annotations, 212 TEs (71%) were correctly annotated by the two software, while 78 TEs (26%) were correctly annotated by REPET and missed by TIDAL (Table S8B). Finally, 10 of the 300 TEs annotated by REPET, were false positives as we could not confirm their presence using Blast (see Material and Methods).

Additionally, we performed manual inspection of 50 TEs that were identified by *TEMP*/*TIDAL* but were not identified by *REPET* (Table S8C). None of these insertions were present in the genome assemblies. For these TEs, we could not distinguish whether they were *REPET* false negatives or *TEMP*/*TIDAL* false positives. However, the majority of these insertions (39/50) have a frequency estimate <20% according to *TEMP*, suggesting that they could be false positives (Zhuang et al 2014). For the 11 TEs with frequencies >20% we cannot discard that these correspond to *REPET* false negatives as *REPET* is run on the assembled genomes that contain a single haplotype, while software based on short-reads allow the interrogation of all the haplotypes present in a given sample (Table S8C).

Thus overall and depending on the tool, short-read tools fail to annotate 26% to 57% of the TEs annotated using long-read tools, while *REPET* false positive rate was 3%.

## TE content is similar across *D. melanogaster* strains while TE activity varies

When comparing TE annotations for the 32 genomes plus the reference genome (ISO1), we observed low variation among strains regarding both TE content (percentage of the euchromatic genome occupied by TEs, average=3.56%, SD=0.3%) and number of TE copies (average=2,016, SD=69.6) (Table S9A). The coefficient of variation for the number of non-reference insertions across populations was similar to previous estimates (7% vs 9% in Chakraborty *et al.* 2019). As previously described, TE variation across populations did not reflect the geographical or environmental origin of the populations (Figure 4A; see Material and Methods; Lerat *et al.* (2019).

At the TE order level, and in agreement with previous studies (Lerat *et al.* 2019), we found *LTRs* to be the most abundant, representing near 60% of all TE content (Table S9B, Figure S6A), while the number of TE copies was more evenly distributed among the five main orders (*Helitrons*, *LARDs*, *LINEs*, *LTRs* and *TIRs*) (Table S9B, Figure S6B). Also in agreement with previous observations, *INE-1* superfamily showed the largest number of copies among Class II DNA elements (Thomas *et al.* 2014) and *Gypsy* and *Pao* elements were the most abundant among the LTRs (Lerat *et al.* 2019; Linheiro & Bergman 2012) Table S9C). Moreover, while no overall significant differences in abundance were found at the superfamily level (Pearson's $X^2$ test of independence = 575.44, p-value=0.4987, Figure 4B, Table S9C), genome pairwise comparisons were significant for the MUN-009 and ISO1 pair of strains ($X^2$ test, adjusted p-value=0.03, Figure 4C), mainly due to the *P* superfamily overrepresentation in MUN-009 compared with the ISO1 genome (Figure 4D). This observation was also confirmed by the analysis at the family level, where MUN-009 was found to contain 60 copies of the *P-element*, while this element is absent from the ISO1 genome (Figure S7 and Table S9D; (Anxolabéhère *et al.* 1988). *P-elements* were indeed among the most variable families in the 33 genomes (Figure S8, Table S9D).

We used the percentage of sequence identity between individual TE copies and the family consensus sequence as a proxy for the age of the insertions. As expected, we found *INE-1* and *LARD* elements to be the oldest superfamilies in all genomes (Kalendar *et al.* 2004; Kapitonov & Jurka 2003), while copies of the *I*, *TcMar-pogo*, *Copia* and *Pogo* superfamilies showed the highest values of identity with the consensus, suggesting they are relatively young, as also previously described (Bucheton *et al.* 1992; Lerat *et al.* 2019) (Figure 4E and Figure S9). Moreover, some superfamilies showed a large variability in identity such as *R1*, *Jockey* and *Gypsy*, indicating that they contain both young and old members (Figure 4E and Figure S9). Genome pairwise comparisons in the distribution of identity values per genome showed significant differences between some pairs of genomes (Figure S10A). Notably, such differences seem to be mainly caused by members of the *Jockey* and *Gypsy* superfamilies (Figure S10B).

Our results, together with previous studies in Drosophila populations, suggest a scenario in which while natural variation in TE abundance between populations exist, certain families tend to be either abundant or rare in most populations (Lerat *et al.* 2019; Rahman *et al.* 2015). Moreover, while almost no significant differences were observed between genomes in the number of TE copies (Figure 4C), we did find pairwise differences in the identity of the copies (Figure S10A), particularly among members of two superfamilies, *Jockey* and *Gypsy* (Figure 4E; Figure S10B), suggesting a population specific behavior regarding TE activity as previously described in both European (Lerat *et al.* 2019) and North American strains (Adrion *et al.* 2017).

### 20 genomes allow the identification of the vast majority of TEs that are common in out-of-Africa natural populations

To investigate how the number of genomes analyzed affects the total number of unique TE copies identified and the estimation of their population frequencies, we identified orthologous insertions by comparing the annotations obtained using *REPET* in 47 genomes: the 32 genomes sequenced in this work, the ISO1 reference genome, and the 14 genomes reported by Chakraborty *et al.* (2019) collected in Africa (2), Europe (2), North America (4), North Atlantic Ocean (1), South America (2), and Asia (3) (Table S10 and S11). On average, 2,016 euchromatic TE copies were annotated per genome (ranging from 1,883 to 2,178, Table S9A), and for 97% of them (on average) orthologous relationships of the insertion flanking regions in the ISO1 reference genome were determined (Table S11A; Supplementary File S1.11). Overall, we annotated 28,947 TEs across the 47 genomes (Table S10). As expected, the site frequency spectrum of TE insertions showed an excess of rare variants compared with SNP variants (Figure S11; Cridland *et al.* (2013).

We classified the 28,947 TEs in three frequency classes: rare (present in <10% of the genomes), common (present in ≥10% and ≤95%) and fixed (present in >95%) and calculated the number of TEs detected in each frequency class starting with the analysis of only five genomes and adding one genome at a time until the total 47 genomes available (see Material and Methods). As expected, we found that as the number of genomes analyzed increased, the number of rare TEs also increased in a linear fashion, as each genome contributes a similar number of rare TEs to the population (Figure 5A and Table S11B). On the other hand, the number of fixed TEs was very similar regardless of the number of genomes considered, and the small variations seen were probably due to errors in either the TE transfer, TE annotation, or genome assemblies (Figure 5A). Finally, we observed that the number of common TEs is more variable depending on the number of genomes considered, and this number stabilizes around 800-900 TEs. The overlap of common TEs considering 10, 20, 30, 40 and 47 strains showed that most of the common TEs (785; 74%) were present in all the subsets (Figure 5B). By increasing the number of genomes analyzed from 10 to 20, the number TEs identified as common decreased (Figure 5B). Besides the core set of 785 common TEs detected in all the subsets, additional 112 TEs were detected as common when analyzing 10 genomes, while only 36 additional TEs were detected as common when analyzing 20 genomes, and 27 additional TEs when analyzing more than 20 strains (Figure 5B). These results suggest that 20 genomes are enough to accurately identify most common TEs in populations, which is the subset of TEs expected be enriched for candidate adaptive mutations (Rech *et al.* 2019).

To determine whether the geographical origin of the strains affects the total number of TE copies identified and their frequency classification, we analyzed genomes according to the continental origin of the sequenced strain: North America, Europe and All populations (Table S11A). Most of the TE insertions were only identified in either Europe or North America (Figure 5C). However, most of these were rare, reflecting the increase in the number of genomes analyzed rather than a geographical effect. On the other hand, if we focused on the common TE insertions, 127 insertions were unique to North America and 103 to Europe (Figure 5C; Table S11C). While some of these insertions were classified as fixed in the other continent, 70 of the common TEs only found in Europe were absent in North America, while 47 of the common TEs found only in North America were absent in Europe (Table S11D). These common TEs that are specific to a particular geographic region are good candidates to have a role in local adaptation. However, the number of TEs was too small to identify enriched biological processes in the genes nearby these TE insertions in these continents.

Overall, our results suggest that the analysis of 20 genomes accurately identifies most common and fixed TEs in a diverse set of populations. Still, because a proportion of the common TEs identified were continent specific, analyzing populations from other continents should lead to the identification of additional common TE insertions.

## Hundreds of *de novo* annotated TEs are associated with the expression of nearby genes

To determine whether TE insertions were associated with the level of expression of nearby genes, we looked for significant associations between cis-eQTLs and TE insertions using RNA-Seq data available for 20 of the strains in our dataset (Table 1, Table S2C). We focused on TE insertions located in high recombination regions as those insertions are more likely to be causal mutations. We identified 503 significant associations (adjusted p-value <0.05), including 481 genes and 472 TEs, the majority of them annotated in this work for the first time (470; Table S12A). Also, most of them (433 out of 472; 91.7%) were present at low frequencies in populations (≤ 5%) suggesting that their effect on gene expression could be deleterious. These TEs were enriched for members of the *P* superfamily and for the *P-element*, *transib1*, *Gypsy-2_Dsim*, *412* and *Doc* families ($X^2$ test, p-value < 0.05, Table S12B). Genes located nearby these TEs were not significantly overrepresented for any biological process, molecular function or cellular component nor any metabolic pathways (Jassal *et al.* 2020; Mi *et al.* 2018). Contrary to previous results, we found a similar number of low frequent TEs associated with gene up- and down-regulation (214 vs 258, respectively; Table S12A; (Cridland *et al.* 2013). *Gypsy-2_sim, 1360*, *Copia* and *Blood* were enriched only nearby up-regulated genes, while *transib1* and *Doc* were only enriched nearby down-regulated genes (Table S12C-D).

We manually curated the TE annotations that showed an adjusted p-value <0.01, and we confirmed 13 significant associations involving 13 genes and 14 TEs, as the *Ten-a* gene had two nearby TEs in linkage disequilibrium that were identified as the top variants (Figure 6 and Table 2; see Material and Methods). Several of the 13 most significant genes are involved in response to stimulus and could be candidates to play a role in the adaptation to new environments (Table 2). For example, *Cyp6a17*, is involved in temperature preference behavior (Kang *et al.* 2011) and it is located within a genomic region harboring several insecticide resistance genes from the *cyp* family (Carareto *et al.* 2014). Manual curation of this region revealed that strains with the TE insertion also had a triplication of the *Cyp6a17* gene that could also contribute to the increased level of expression found in strains with the TE insertion. *Gr64a*, is a gustatory receptor gene required for the behavioral responses to multiple sugars (glucose, sucrose, and maltose) (Jiao *et al.* 2007). Furthermore, other genes may be important for their role in neurogenesis (*pde9*, *ppk*) (Day *et al.* 2005) and synaptic organization (*Ten-a*, *dpr8*) (Cheng *et al.* 2019) (Table 2).

## Most of the insertions with signatures of selection in their flanking regions were non-reference insertions

In order to identify TEs likely to play a role in adaptation we looked for evidence of positive selection in the TE flanking regions. We used SNPs alleles as a proxy to identify genomic regions undergoing selective sweeps and then we explored whether such a sweep was linked to a nearby TE insertion. We applied three haplotype-based statistics: iHS (Voight *et al.* 2006), iHH12 (Garud *et al.* 2015; Torres *et al.* 2018) and nSL (Ferrer-Admetlla *et al.* 2014). We defined a SNPs to have a significant iHS, iHH12 or nSL values when, after normalizing by frequency and chromosome location, the normalized values were >95th percentile of the distribution of values for SNPs falling in neutral introns (see Material and Methods). We then looked for candidate adaptive TE insertions in linkage disequilibrium with each significant SNP, and located <1kb from the significant SNP (see Material and Methods). We considered as candidate adaptive TEs those present at high population frequency and located in regions with recombination rates >0 (see Material and Methods and Rech *et al.* (2019). Among the 746 candidate adaptive TEs, we found 19 TEs co-occurring with SNPs showing evidence of selective sweeps (Table S13A). Among these 19 TE insertions, two correspond to an *Accord* element inserted in the *Cypg6g1* gene that is duplicated in some genomes (Table S14). These two insertions are part of an allelic series previously associated with phenotypic variation, in which the more derived the allele is, the greater the level of insecticide resistance (Daborn *et al.* 2002; Schmidt *et al.* 2010). We discarded the presence of other structural variants linked to our 18 candidate adaptive TEs that could also be driving positive selection (Table S14). Moreover, our set of candidate adaptive TEs was enriched for signatures of selection compared with the whole dataset of TEs present at >5% population frequency (the minimum frequency required to calculate the selection statistics; $X^2$ test, *p-value* = 0.0081). Given the small number of genomes analyzed, strong selection appears to be acting on these 18 insertions as exemplified by the *Accord* insertion (Daborn *et al.* 2002; Schmidt *et al.* 2010). However, further functional validation is needed before arriving at any conclusive evidence on the functional role of these TEs. Note that for one of these 18 insertions, we found significant association with the level of expression of the nearby gene in whole-body non-stress conditions (Figure 6).

We next performed GO enrichment analysis with all the genes located nearby candidate adaptive TE insertions identified so far in *D. melanogaster*, including 84 TEs reported in Rech *et al.* (2019), five other insertions recently described by Bogaerts-Márquez *et al.* (2021), and the 18 TEs identified in this work, including the previously described *Accord* insertion (107 insertions in total). Biological process GO terms analysis identified clusters enriched for response to stimulus, behavior, and development and morphogenesis as the ones showing the highest enrichment scores (Figure 7, Table S15). Pigmentation was also among the significant clusters, as has been previously described (Rech et al 2019). Several gene list enrichments, including regulatory miRNAs and transcription factors, confirmed that genes located nearby these candidate adaptive TEs are enriched for response to stimulus (biotic and abiotic factors), development, behavior*,* (olfactory and locomotor), and energy metabolism (fatty acid and glucose) functions (Figure 7 and Table S15).

The 107 candidate adaptive TEs identified so far in *D. melanogaster* (Table S16A) were enriched for TEs belonging to the *BS* and *Rt1b* families of the LINE order and to the *1360*, *S-element, pogo* and *transib2* families of the TIR order (Table S16B). Finally, regarding gene body location, we found that the subset of candidate adaptive TEs was slightly enriched for TEs inserted in 5'UTR and promotors, although the differences were not statistically significant (Table S16C).

## Discussion

Despite the increasing evidence showing TEs as an important source of genomic structural variation and gene regulation, we are just starting to understand the genome-wide role of these abundant and active components of the genome. The main reasons for this gap in our genomic knowledge are the methodological challenges intrinsic to TEs repetitive nature. New high throughput long-read sequencing technologies that allow to span repetitive regions of the genome, and cutting-edge computational tools offer us now the opportunity to systematically include TE analysis as part of genomics studies. Some works have already demonstrated this, proving that even in an extensively studied biological model organism like *D. melanogaster* we can still identify new and interesting biological properties in which TEs are involved (Chakraborty *et al.* 2019; Ellison & Cao 2019; Mohamed *et al.* 2020). In this work, we go a step further by not only using long-read sequencing to generate whole genome assemblies of 32 natural *D. melanogaster* strains collected from 12 populations located in three climate types (Figure 1 and Table S1), but by also taking into account the genetic variability present in these genomes to create a new *D. melanogaster* TE library. We proved that the use of this library —together with a comprehensive TE annotation strategy— not only improves the current gold standard annotation in the well-studied fruit-fly genome (Figure 2), but also allows the identification of new TE families (Figure 3) and outperforms state-of-the-art methods for TE annotation using short-reads. Our results also showed that reference genomes consisting of a haplotype-collapse representation are likely to miss some TE insertions as they do not incorporate polymorphisms. Future development of haplotype-resolved *de novo* assemblies should improve variant calling in long-read genomes (De Coster *et al.* 2021). Moreover, the availability of even longer reads together with the improvement of computational analysis should help to characterize nested and highly complex variation in the near future (De Coster *et al.* 2021).

Improving the annotation of TEs in genome sequences is the first necessary step to accurately evaluate the role of this abundant an active component in genome function and evolution. We identified 472 TEs associated with nearby gene expression variation (Figure 6 and Table S12). While previous genome-wide studies reported an association of TE insertions with reductions of gene expression, our data provide evidence for associations with both up- and down-regulation of nearby genes, in line with a recent analysis on the role of TEs in immune-related genes (Cridland *et al.* 2015; Ullastres *et al.* 2021). TE annotations in genomes from arid, temperate and cold climates should allow us to test whether TEs have been involved in adaptation to different environmental conditions. Moreover, the new TE library was also used to annotate 14 other high-quality *D. melanogaster* genomes, which allowed us to analyze the frequency distribution of TE insertions in a total of 47 genomes (Figure 5). We identified 746 TE insertions present at high population frequencies (≥10% and ≤95%) in genomic regions with recombination rates >0. Eighteen of these common TE insertions were associated with signatures of selection at the DNA sequence level, including the well-known *Accord* insertion in *Cyp6g1* associated with increased resistance to insecticides, and represent 31% more candidate adaptive TE insertions compared with the previous most extensive analysis (Table 3; (Daborn *et al.* 2002; Schmidt *et al.* 2010; Rech *et al.* 2019). The joint analysis of all the *D. melanogaster* TE insertions showing evidence of positive selection identified so far confirmed that development and response to stimulus are among the most frequent biological processes shaped by TE insertions, together with behavior and pigmentation (Table 3, Figure 7; Rech *et al.* 2019).

Overall, given the growing evidence of the importance of TE insertions in genome evolution and function, in addition to their relevance in several human diseases, the approach reported here provides a framework for studying TE dynamics, evolution and the functional implications of TEs in natural population using long-read sequencing. A critical step, was the manual curation of the TE libraries and annotations, a noteworthy effort that allows us to fine-tune the TE annotation strategy to reduce false positives and retain most of the true copies only. We expect that the increasing shift towards the use of long-read sequencing together with comprehensive integration of natural variation in the TE analyses will keep helping to elucidate the role of these active and abundant members of the genome.

## Methods

### Sequenced strains

We sequenced the genomes of 32 *D. melanogaster* strains originally collected from natural populations. All the samples represent either isofemale or inbred stocks from such natural populations (Table S1). 24 strains were obtained from 11 European natural populations and the remaining eight are RAL strains from the DGRP, obtained from North Carolina Bloomington (Figure 1, Table S1). All flies were reared on standard fly food medium in a 12:12 h light/dark cycle at 25 °C.

### DNA extraction and long-read sequencing

We sequenced two strains (MUN-016 and TOM-007) using Pacific Biosciences (PacBio) technology and the remaining 30 using Oxford Nanopore Technologies (ONT) and Illumina technologies. DNA for PacBio sequencing was extracted from 400 female flies, using the Gentra Puregene Tissue Kit (Qiagen) following manufacturer's instructions. Briefly, 400 flies from each strain were mechanically homogenized in 24ml of lysis buffer (proteinase K added) and incubated overnight at 55ºC, and DNA was precipitated with isopropanol after RNAse treatment and protein precipitation. Finally, DNA was resuspended in 1,6ml of Hydration Solution. DNA concentration was measured using a Nanodrop® spectrophotometer. Most DNA samples for ONT sequencing were extracted from 100 female flies from each strain using the Blood and Cell Culture DNA Mini Kit (Qiagen) following manufacturer's instructions with small modifications (Table S2; Supplementary File S1.1).

PacBio libraries were prepared using 20Kb SMRTbell and were delivered to Macrogen Inc. Korea to be sequenced using the PacBio RSII System. ONT libraries were constructed using the Ligation Sequencing Kit (SQK-LSK108 or SQK-LSK109) following manufacturer's instructions (Table S2; Supplementary File S1.1) and were sequenced *in house* using the MinION device. Basecalling of ONT reads was performed using the *Albacore* Sequencing Pipeline Software (v.2.2). The quality of the long-read sequencing was assessed using *NanoPlot* (v.1.19) (De Coster *et al.* 2018).

### Short-read sequencing

The previously extracted DNA used for ONT sequencing was also sequenced using short-read Illumina sequencing either by Macrogen Inc. Korea (TruSeq DNA PCR-free kit, 350bp insert libraries, 150bp pair-end sequencing) or by the Genomics Unit of the Center for Genomic Regulation (gDNA-PCR free, HiSeq 2500, 125bp pair-end) (Table S2C).

### Genome assemblies

We performed *de novo* genome assembly of the 32 strains sequenced with long-read sequencing technologies. For PacBio sequences, we used *Canu* (v.1.7) (Koren *et al.* 2017) for building draft genome assemblies followed by *FinisherSC* (v.2.1) (Lam *et al.* 2015) for improving contig continuity. We then aligned PacBio reads to the draft assembly using *pbalign* (SMRT Link v.5.0.1) and used *quiver* (SMRT Link v.5.0.1) to obtain the consensus sequences (polished assembly). PacBio-related programs were all run using default parameters (Figure S12A). For ONT genomes, we also started with *Canu* (v.1.7) (Koren *et al.* 2017) with default options for building raw *de novo* assemblies. We then applied *Racon* (v.1.0) (Vaser *et al.* 2017), *Nanopolish* (v.0.10.1) (https://github.com/jts/nanopolish) and *Pilon* (v.1.22) (Walker *et al.* 2014) for obtaining final polished assemblies (Figure S12B, Supplementary File S1.2).

### Genome deduplication, decontamination and scaffolding

Since we detected that raw *de novo* genome assembly sizes positively correlated with BUSCO Duplicates (besides repetitive content) (Supplementary File S1.3, Figures S13-S15), we evaluated whether levels of heterozygosity might also be involved in determining genome size. Heterozygosity levels in the sequenced strains were evaluated using the short-reads sequences by first calling SNPs against the ISO1 genome following the *GATK* (v.4.0) (McKenna *et al.* 2010) best practices for variant discovery (Van der Auwera *et al.* 2013). Then, we used the *bcftools stats* (v.1.9) (Li 2011) for calculating the percentage of heterozygous SNPs at each genome and we found a positive correlation between the estimated heterozygosity and the raw assembly size (Supplementary File S1.3, Figures S16). Genomes showing levels of heterozygosity > 0.2 were deduplicated (removing alleles-contigs- present twice in the genome) using *purge_haplotigs* (v.1.0.1) (Roach *et al.* 2018) (Figures S17, Table S3, Supplementary File S1.3).

After deduplication, we evaluated contigs for putative contaminations using *MUMmer* (v.4.0) (Marçais *et al.* 2018). Briefly, we attempted to align all contigs to the *D. melanogaster* hologenome (Kapun *et al.* 2020) plus the *D. simulans* genome. We considered as putative contaminant contigs showing matches with identities >98% and overlapping >95% of the contig length. We identified putative contaminant contigs in seven genomes (COR-018, LUN-004, MUN-016, MUN-020, RAL-737, TEN-015, TOM-007) (Table S3). Once we removed the putative contaminant contigs, we performed a reference-guided scaffolding of the contigs using *RaGOO* (v.1.02) (Alonge *et al.* 2019), which uses *minimap2* (v.2.9) (Li 2018) for aligning contigs to the ISO1 reference genome for ordering and orienting contigs into pseudomolecules. In order to determine whether the scaffolds were covering most of the major chromosomal arms in ISO1, we mapped back the scaffolded genomes to the ISO1 genome using *MUMmer4* (v.4.0) (Marçais *et al.* 2018) (Table S3).

### Assembly quality

Quality of the assemblies was evaluated by estimating completeness, continuity and accuracy. Completeness and accuracy were calculated using *BUSCO* (v.3.0.2) (Waterhouse *et al.* 2017) for the Diptera lineage (diptera_odb9), consisting on 2,799 genes. Continuity and completeness were estimated by aligning the polished genome assemblies to the *Drosophila melanogaster* strain ISO1 reference genome release 6 (Hoskins *et al.* 2015). We first masked simple repeats in both genomes using *RepeatMasker* (v.3.0) (www.repeatmasker.org) and then used *MUMmer* (v.3.0) (Kurtz *et al.* 2004) for genome alignment. The quality of the genomes in the context of TEs was evaluated using *CUSCO* (downloaded on May 6, 2020) (Cluster BUSCO; Wierzbicki *et al.* 2020) based on the flanking sequences for 85 out of the 142 annotated piRNA clusters of *D. melanogaster* (Brennecke *et al.* 2007) (Table S3B, Supplementary File S1.5). QV scores were estimated according to Solares *et al.* (2018) using both SNPs and INDELs called from the mapping of Illumina short-reads over the de-novo assembled genomes.

### TE sequence accuracy based on long-read sequences

Incremental updates to the ONT base-calling algorithm has been reported to improve read accuracy (Wick *et al.* 2019). To test whether the ONT base-calling algorithm used in this work affected the TE sequence accuracy, we assembled ONT long-reads available for the reference genome (Solares *et al.* 2018) using our pipeline (Figure S12B). We annotated TE copies using the MCTE library and we identified 1,842 orthologous TEs comparing with the ISO1 reference genome TE annotation, which represents >83% of the TEs annotated in Solares *et al.* (2018)´s genome and > 89% of the TEs annotated in the ISO1 reference genome. For every TE pair, we performed global pairwise alignments using MAFFT v.7.4 aligner (parameters: *mafft --globalpair --thread 4 --reorder --adjustdirection --auto*). For each pair we then calculated the pairwise identity in two ways: considering and not considering gaps in the alignment. Average gap-ignorant identity was 99.9% and gap-aware identity was 98.9%. Some TE families showed more variability than others but in most cases this variability was explained by individual TE insertions.

### Construction of the Manually Curated TE (MCTE) library

We used the *REPET* package (v.2.5) (Flutre *et al.* 2011; Hoede *et al.* 2014; Quesneville *et al.* 2005) for performing TE annotations using a manually curated TE (MCTE) library of consensus sequences. Briefly, *REPET* is composed of two main pipelines, *TEdenovo* dedicated to *de novo* detection of TE families and *TEannot* for the annotation and analysis of TEs in genomic sequences. For the creation of the MCTE library, we first run the *TEdenovo* pipeline (default parameters) on 13 genomes (representatives of the geographic distribution of the strains) (Table 1). The manual curation of the identified consensuses consisted in three main procedures: removal of redundant sequences, the manual identification of potentially artifactual sequences, and the classification of consensuses into families (Supplementary File S1.6). Redundant sequences (consensus sequences present in more than one genome) were removed by first running *PASTEC* (v2.0) with default options (Hoede *et al.* 2014). We also performed similarity clustering, multiple sequence alignments (MSA) of the clusters and generated consensus sequences for each MSA in order to obtain a consensus sequence representative of all the genomes (Supplementary File S1.6). We manually explored the consensus sequences and their copies using the *plotCoverage* tool from *REPET* and discarded consensuses showing mainly a high number of small copies. The assignation of the consensus sequences into families was performed using *BLAT* (v.35) (Kent 2002) against the curated canonical sequences of Drosophila TEs from the Berkeley Drosophila Genome Project (BDGP) (v.9.4.1) (https://fruitfly.org/p_disrupt/TE.html). When no matches were found, we used *RepeatMasker* (v.4) (Smit 2015) with the release *RepBaseRepeatMaskerEdition-20181026* of the *RepBase* (Bao *et al.* 2015) (Supplementary File S1.6).

### TE Annotation

We use the MCTE library as input for the *TEannot* pipeline to annotate each of the 32 genomes and the ISO1 reference genome. The pipeline was run with default parameters. We annotated TE copies only in the euchromatic regions of the genome since heterochromatic regions are gene-poor (Smith *et al.* 2007) and its assembly and annotation usually require specific methods and extensive curation (Chakraborty *et al.*

2019; Khost *et al.* 2017). In this work, we determined the euchromatic regions using the recombination rate calculator (RRC) (Fiston-Lavier *et al.* 2010) available at *http://petrov.stanford.edu/cgi-bin/recombination-rates_updateR5.pl*. Such coordinates were originally calculated based on the release 5 of *D. melanogaster* genome so we converted them to release 6 coordinates using the *coord_converter.pl* script from FlyBase (Gramates *et al.* 2017), resulting in the following regions: 2L:530,000..18,870,000; 2R:5,982,495..24,972,477; 3L:750,000..19,026,900; 3R:6,754,278..31,614,278; X:1,325,967..21,338,973. In order to determine the coordinates of the euchromatic regions in each scaffolded genome, we mapped scaffolds to the euchromatic region of the ISO1 genome using *MUMmer* (v3.0) (Kurtz *et al.* 2004). We then determined the coordinates in the scaffolded genomes by parsing *MUMmer*´s output and extracting the coordinates mapping at the boundaries of the euchromatic region of the ISO1 genome. After running the *TEannot* pipeline over the euchromatic regions of each genome, we performed a post-annotation filtering step consisting in the removal of TE copies <100bp, as *REPET* cannot accurately annotate these copies, and copies whose length overlapped >80% with satellite annotations.

Multiple sequence alignments of TE insertions for manual curation were performed with *MUSCLE* (v.3.5) using *Geneious* (v.10.0.2) for alignment and visualization (https://www.geneious.com). Identity values between TE copies and the consensus were obtained from *REPET TEannot* pipeline.

## Comparison with short-read-based TE annotations

We compared *REPET* TE annotations on the *de novo* assembled genomes using the MCTE library with the annotations performed by two short-read-based TE annotation software: *TEMP* (v.1.05) (Zhuang et al. 2014) and *TIDAL* (v.1.0) (Rahman et al. 2015). To make the comparison unbiased regarding the TE library, we also used the MCTE library for *TEMP* and *TIDAL*. We considered 11 strains representative of the geographic variability and with the best quality assembled genomes (Table 1). We used *BEDtools* (v.2.18) (Quinlan & Hall 2010) to find the overlapping TEs copies predicted by the three different methods (*REPET*, *TIDAL* and *TEMP*) in the 11 strains in a family-aware fashion. To estimate TEMP and TIDAL false negative rate and REPET false positive rate, manual inspection was performed for 300 of the 712 *de novo* insertions in the COR-014 genome. To do this, we identified the region where each of these TEs was annotated according to

REPET/TEMP/TIDAL and we aligned this region against the ISO1 reference genome to find out if a *de novo* insertion truly exists. We also used Blast to search for sequence similarities of such genomic region with (i) a database that contains all the individual TE copies identified in our genomes; and (ii) all Flybase´s 'Transposons - all annotated elements (NT)' . If REPET identified a TE not annotated by TEMP/TIDAL we considered it as TEMP/TIDAL false negative. If a TE was annotated by REPET but we could not find sequence similarities with any of the TE databases by Blast, we considered it as a REPET false positive. Additional 50 TEs annotated by TEMP/TIDAL but not by REPET were also manually curated following the same procedure.

## TE orthology identification

To identify orthologous TEs, we first transferred the TE coordinates from each strain to the ISO1 reference genome. Briefly, we used a similarity and synteny approach based on *minimap2* (v.2.9) (Li 2018) mapping of the TE sequence and its flanking regions to the ISO1 genome and the coordinates of genes as anchored synteny sequences (see Supplementary File S1.11 for details). To transfer the TEs, we took into account whether its flanking region mapped unequivocally or not, whether it mapped completely or partially, whether it was a tandem or nested TE, among others. Then, based on the information of the alignment and characteristics of the transfer, we defined each of the TEs as either reliable or unreliable, being the latter ones discarded from the transfer. Finally, once all the reliable TEs of each strain were transferred to the reference, the orthologous TEs were defined (Supplementary File S1.11, Figures S18-S20). To avoid false positives, we only used those TEs for which more than half of the orthologous TEs were larger than 120bp. All scripts used for the TE transfer are available at www.github.com/sradiouy/deNovoTEsDmel.

After determining the presence/absence of TEs, we classified them in three frequency classes: rare (TEs present in <10% genomes), fixed (TEs present in >95% of the genomes) and common (TEs present in ≥10% and ≤95%). We then calculated the number of TEs for each frequency class considering different number of genomes, starting from 5 up to 47. We estimated the mean and standard deviation of the number of TEs in each frequency class by randomly choosing genomes (30 iterations). Then, we intersected the different sets of common TEs considering 10, 20, 30, 40 and 47 strains using *UpSetR* (v.1.3) (Conway *et al.* 2017) and also established different sets of TEs based on the geographical origin of the genomes and compare them using *VennDiagram* (v.1.6) (Chen & Boutros 2011). For determining the location of the TE insertion regarding annotated genes, we used *annotatr*.

**TE eQTL analysis**

In order to identify polymorphic TEs significantly associated with the expression levels of nearby genes, we analyzed available whole-body RNA-Seq data from 12 European and 8 American strains (Table1, Table S2C). Briefly, RNA-Seq data was trimmed using the *fastp* package (v.0.20) (Chen *et al.* 2018) with default parameters. Expression levels were quantified by applying the *salmon* package (v.1.0.0) (Patro *et al.* 2017) against the ENSEMBL (Dm.BDGP6.22.9) transcripts. Obtained transcripts per million (TPM) were summed up to gene level and *rlog* normalized using *DESeq2* (v.1.28.1) (Love *et al.* 2014). eQTL analysis was performed using the *QTLTools* package (v.1.2) (Delaneau *et al.* 2017) taking into account the population structure (Figure S21, S22). Putative cis-eQTL were searched within a 1Kb window around each gene using the *cis* module in *QTLTools*. We used the nominal pass to evaluate the significance of the association of the gene expression level to TE insertions. The genotype table was created with a custom script. Finally, we performed a permutation pass (100,000 permutation) to adjust for multiple testing. Overall, we evaluated 12,281 eGenes-TE involving 4,709 genes and 9,676 TEs. We focused on TEs located in high recombination regions and we considered significant eGenes-TE associations when the nominal p-value and the associated adjusted p-value were significant (<0.05). Manual inspection of the 15 TEs that were the top variant and the most significant associations (adjusted p-value <0.01) confirmed that they were correctly annotated in all the genomes (300/300 correct calls) except for an *INE-1* element that was removed from the analysis as it was fixed in all the genomes analyzed (7/20 correct calls) and a *Blastopia* insertion that was miss annotated in one of the strains (19/20 correct calls).

**Selection Analysis**

We looked for evidences of selection in genomic regions targeted by TEs insertions using *selscan* (v.1.2.0a) (Szpiech & Hernandez 2014) and Single Nucleotide Polymorphisms (SNPs) as a proxy (Supplementary File S1.12). We looked for evidences of incomplete soft or hard selective sweeps in the 46 *D. melanogaster* genomes (the 32 sequenced in this work plus the 14 genomes sequenced by Chakraborty *et al.* 2019). SNPs were called using the *GATK* (v.4.0) (McKenna *et al.* 2010) *HaplotypeCaller* best practices for variant discovery (Van der Auwera *et al.* 2013) and the haplotype phasing was performed using *SHAPEIT4* (v.4.1) (Delaneau *et al.* 2019). Initial SNP calling resulted in 5,578,437 SNPs, from which we kept only biallelic SNPs using the GATK command *SelectVariants* (parameters -*select-type SNP --restrict-alleles-to BIALLELIC*). Finally, we also removed SNPs with *missing data* in at least one genome, resulting in a total of 2,797,589 SNPs (available at http://dx.doi.org/10.20350/digitalCSIC/13708). Genetic positions and recombination maps (Comeron *et al.* 2012) were obtained from FlyBase (https://wiki.flybase.org/wiki/FlyBase:Maps, last updated June 15, 2016). Three statistics were calculated in *selscan*: iHS (Voight *et al.* 2006), iHH12 (Garud *et al.* 2015; Torres *et al.* 2018) and nSL (Ferrer-Admetlla *et al.* 2014). iHS and nSL statistics are both aimed to identify incomplete sweeps, where the selected allele is not fixed in the sample, and the main difference is that nSL is more robust to recombination rate variations, which increases the power to detect soft sweeps. iHH12 has been developed for the detection of both hard and soft sweeps, with more power than iHS to detect soft sweeps (Szpiech & Hernandez 2014). After obtaining results from each statistic, we normalized them using the *norm* package in 10 frequency bins across each chromosome. We considered iHS, iHH12 and nSL normalized values to be statistically significant for a given SNP if they were greater than the 95th percentile of the distribution of normalized values for SNPs falling within the first 8–30 base pairs of small introns (<=65bp) which are considered to be neutrally evolving (Parsch *et al.* 2010) (Table S13B). In order to identify TEs putatively linked to the selective sweeps, we analyzed the co-occurrence (in the same strains) of the allele showing signatures of a selective sweep and a nearby TE (<1Kb). We focused only on those TEs more likely to have a role in adaptation: First, from the 28,365 transferred TEs, we selected those at frequencies ≥10% and ≤95% and inserted in regions with recombination rates >0, as these insertions are more likely to play a role in adaptive evolution rather than being linked to the causal mutation (Rech *et al.* 2019), resulting in 902 TEs. From those, we also discarded TEs belonging to the INE-1 and the LARD families, since those represent very old TE families likely to have reach high frequencies neutrally, ending up with a set of 746 TEs. We considered TEs in this 746 dataset as likely to be enriched for candidate adaptive TEs (Rech *et al.* 2019). We then looked whether any of these 746 TEs was nearby a SNP showing significant values at some of the haplotype-based selection test. Finally, for each SNP-TE pair we stablished criteria of ‘co-occurrence’ by requesting certain number of the strains containing both the SNP allele undergoing a selective sweep and the nearby TE: for TEs present in 5-6 strains we request at least 4 of the strains to contain both the allele undergoing a selective sweep and the nearby TE and for TEs present in ≥7 strains we request the majority of strains to contain both the significant SNP and the nearby TE. In all cases, we also requested the TE to be absent in 100% of strains that do not contain the significant SNP allele (Table S13A).

To discard that other CNVs could be linked to the identified 18 TEs associated with signatures of selection, we identified using the *Structural Variants and MUmmer* (SVMU) tool the presence of CNVs in the 1kb regions flanking these insertions (Chakraborty *et al.* 2019).

### TE genomic location

TE´s overlapping genes or located nearby genes were determined using the following criteria: (i) we considered only protein-coding genes from FlyBase gene annotation *r6.31* (13,939 genes);. (ii) to determine the gene location (3´UTR, 5´UTR, CDS, INTRON, PROMOTER) we considered the position regarding the longest transcript only; (iii) promoter regions were considered as the region 1Kb upstream of the TSS; (iv) 3´UTR, 5´UTR, CDS, INTRON coordinates were obtained from the header of the *fasta* files available at FlyBase (http://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.31_FB2019_06/fasta/); (v) only the closest gene (<1Kb) to the TE was considered; (vi) when a TE overlapped (distance = 0) with more than one gene, all overlapping genes were considered. This is also true for the (rare) case in which the distance to more than one gene is exactly the same; and (vii) when no gene was found at <1Kb, the TE was classified as ‘Intergenic’.

### Enrichment analysis:

GO enrichment analyses for list of genes nearby candidate TEs were performed using DAVID functional annotation cluster tool (v.6.8) (Huang da *et al.* 2009; Huang *et al.* 2008) using all *D. melanogaster* protein-coding genes from FlyBase gene annotation *r6.31* as a background. In addition, we also used the online version of FlyEnrichr (Chen *et al.* 2013; Kuleshov *et al.* 2016) to analyze enrichments regarding four gene-set libraries: 1) *Anatomy GeneRIF Predicted*: list of genes with predicted GeneRIF terms involved in fly´s bodily structures (Gene Reference into Function: https://www.ncbi.nlm.nih.gov/gene/about-generif). 2) *Allele LoF Phenotypes from FlyBase*: FlyBase's allele phenotypic dataset. Loss of function phenotypes and gene sets with alleles producing those phenotypes. 3) *Putative Regulatory miRNAs from DroID*: DroID's (http://www.droidb.org/) putative miRNA targets dataset and 4) *Transcription Factors from DroID*: DroID's (http://www.droidb.org/) transcription factor-gene interactions datasets. We report only terms with an adjusted p-value <0.05.

## Data Availability

All scaffolded assemblies and the raw data (long and short read sequencing) have been deposited in NCBI under the BioProject accession PRJNA559813. The VCF file containing SNP callings for 46 *D. melanogaster* genomes used for testing positive selection evidences is available at http://dx.doi.org/10.20350/digitalCSIC/13708. Fasta sequences for the *D. melanogaster* Manually Curated Transposable Elements (MCTE) library are available at http://dx.doi.org/10.20350/digitalCSIC/13765. Recombination rates according to Fiston-Lavier *et al.* (2010) and Comeron *et al.* (2012) for *D. melanogaster* genome release 6 are available at http://dx.doi.org/10.20350/digitalCSIC/13766. BED files containing Transposable Element (TE) annotations for 47 Drosophila melanogaster genomes: http://dx.doi.org/10.20350/digitalCSIC/13894.

## Code Availability

All scripts and codes have been deposited to GitHub and freely accessible from https://github.com/gabyrech/deNovoTEsDmel and www.github.com/sradiouy/deNovoTEsDmel.

## Acknowledgments

We would like to thank DrosEU researchers for sharing with us the strains sequenced in this manuscript (Table S1). We also thank Miriam Merenciano, Anna Ullastres, and Lain Guio for helping create the inbred strains. This project has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (H2020-ERC-2014-CoG-647900). SR is funded by the MICINN/FSE/AEI (PRE2018-084755) and VH was funded by the Generalitat de Catalunya (FI2017_B00468). DrosEU is funded by an ESEB Special Topic Network award.

## Author contributions

G.E.R.: data acquisition, analysis and data interpretation, drafted and revised the manuscript. S.R. and S.G-R: data acquisition, analysis and data interpretation, drafted and revised the manuscript. L.A., V.H., L.G and H.L.: data acquisition and revised the manuscript. V.J. and H.Q.: data analysis and revised the manuscript. JG: conception and design of the work, analysis and interpretation of data, drafted and revised the manuscript.

## References

Adrion JR, Song MJ, Schrider DR, Hahn MW, Schaack S (2017) Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. *Genome Biol Evol* **9**, 1329-1340.

Alonge M, Soyk S, Ramakrishnan S*, et al.* (2019) Fast and accurate reference-guided scaffolding of draft genomes. *bioRxiv*, 519637.

Alonge M, Wang X, Benoit M*, et al.* (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell* **182**, 145-161.e123.

Anxolabéhère D, Kidwell MG, Periquet G (1988) Molecular characteristics of diverse populations are consistent with the hypothesis of a recent invasion of Drosophila melanogaster by mobile P elements. *Mol Biol Evol* **5**, 252-269.

Audano PA, Sulovari A, Graves-Lindsay TA*, et al.* (2019) Characterizing the Major Structural Variant Alleles of the Human Genome. *Cell* **176**, 663-675.e619.

Ballouz S, Dobin A, Gillis JA (2019) Is it time to change the reference genome? *Genome Biology* **20**, 159.

Bao W, Kojima KK, Kohany O (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA* **6**, 11.

Barron MG, Fiston-Lavier AS, Petrov DA, Gonzalez J (2014) Population genomics of transposable elements in Drosophila. *Annu Rev Genet* **48**, 561-581.

Bogaerts-Márquez M, Guirao-Rico S, Gautier M, González J (2021) Temperature, rainfall and wind variables underlie environmental adaptation in natural populations of Drosophila melanogaster. *Molecular ecology* **30**, 938-954.

Brennecke J, Aravin AA, Stark A*, et al.* (2007) Discrete small RNA-generating loci as master regulators of transposon activity in Drosophila. *Cell* **128**, 1089-1103.

Bucheton A, Vaury C, Chaboissier MC*, et al.* (1992) I elements and the Drosophila genome. *Genetica* **86**, 175-190.

Carareto CM, Hernandez EH, Vieira C (2014) Genomic regions harboring insecticide resistance-associated Cyp genes are enriched by transposable element fragments carrying putative transcription factor binding sites in two sibling Drosophila species. *Gene* **537**, 93-99.

Chaisson MJP, Huddleston J, Dennis MY*, et al.* (2015) Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608-611.

Chaisson MJP, Sanders AD, Zhao X*, et al.* (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nature Communications* **10**, 1784.

Chakraborty M, Emerson JJ, Macdonald SJ, Long AD (2019) Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nature Communications* **10**, 4872.

Chakraborty M, VanKuren NW, Zhao R*, et al.* (2018) Hidden genetic variation shapes the structure of functional elements in Drosophila. *Nature Genetics* **50**, 20-25.

Chen EY, Tan CM, Kou Y*, et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* **14**, 128.

Chen H, Boutros PC (2011) VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinformatics* **12**, 35.

Chen S, Zhou Y, Chen Y, Gu J (2018) fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884-i890.

Cheng S, Ashley J, Kurleto JD*, et al.* (2019) Molecular basis of synaptic specificity by immunoglobulin superfamily receptors in Drosophila. *Elife* **8**.

Comeron JM, Ratnappan R, Bailin S (2012) The Many Landscapes of Recombination in Drosophila melanogaster. *PLOS Genetics* **8**, e1002905.

Conway JR, Lex A, Gehlenborg N (2017) UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940.

Cridland JM, Macdonald SJ, Long AD, Thornton KR (2013) Abundance and distribution of transposable elements in two Drosophila QTL mapping resources. *Mol Biol Evol* **30**, 2311-2327.

Cridland JM, Thornton KR, Long AD (2015) Gene Expression Variation in Drosophila melanogaster Due to Rare Transposable Element Insertion Alleles of Large Effect. *Genetics* **199**, 85-93.

Daborn PJ, Yen JL, Bogwitz MR*, et al.* (2002) A single p450 allele associated with insecticide resistance in Drosophila. *Science* **297**, 2253-2256.

Day JP, Dow JA, Houslay MD, Davies SA (2005) Cyclic nucleotide phosphodiesterases in Drosophila melanogaster. *Biochem J* **388**, 333-342.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C (2018) NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666-2669.

De Coster W, Van Broeckhoven C (2019) Newest Methods for Detecting Structural Variations. *Trends in Biotechnology* **37**, 973-982.

De Coster W, Weissensteiner MH, Sedlazeck FJ (2021) Towards population-scale long-read sequencing. *Nature Reviews Genetics* **22**, 572-587.

Delaneau O, Ongen H, Brown AA*, et al.* (2017) A complete tool set for molecular QTL discovery and analysis. *Nature Communications* **8**, 15452.

Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET (2019) Accurate, scalable and integrative haplotype estimation. *Nature Communications* **10**, 5436.

Du H, Liang C (2019) Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nature Communications* **10**, 5360.

Ellison CE, Cao W (2019) Nanopore sequencing and Hi-C scaffolding provide insight into the evolutionary dynamics of transposable elements and piRNA production in wild strains of Drosophila melanogaster. *Nucleic Acids Research* **48**, 290-303.

Ferrer-Admetlla A, Liang M, Korneliussen T, Nielsen R (2014) On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol Biol Evol* **31**, 1275-1291.

Fiston-Lavier AS, Singh ND, Lipatov M, Petrov DA (2010) Drosophila melanogaster recombination rate calculator. *Gene* **463**, 18-20.

Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS One* **6**, e16526.

Garud NR, Messer PW, Buzbas EO, Petrov DA (2015) Recent Selective Sweeps in North American Drosophila melanogaster Show Signatures of Soft Sweeps. *PLOS Genetics* **11**, e1005004.

Goerner-Potvin P, Bourque G (2018) Computational tools to unmask transposable elements. *Nature Reviews Genetics* **19**, 688-704.

Gramates LS, Marygold SJ, Santos Gd*, et al.* (2017) FlyBase at 25: looking to the future. *Nucleic Acids Research* **45**, D663-D671.

Hoede C, Arnoux S, Moisset M*, et al.* (2014) PASTEC: an automatic transposable element classification tool. *PLoS One* **9**, e91929.

Hoskins RA, Carlson JW, Wan KH*, et al.* (2015) The Release 6 reference sequence of the Drosophila melanogaster genome. *Genome Res* **25**, 445-458.

Huang da W, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13.

Huang DW, Sherman BT, Lempicki RA (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols* **4**, 44-57.

Huddleston J, Eichler EE (2016) An Incomplete Understanding of Human Genetic Variation. *Genetics* **202**, 1251-1254.

Jain M, Koren S, Miga KH*, et al.* (2018) Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology* **36**, 338-345.

Jassal B, Matthews L, Viteri G*, et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res* **48**, D498-d503.

Jiao Y, Moon SJ, Montell C (2007) A Drosophila gustatory receptor required for the responses to sucrose, glucose, and maltose identified by mRNA tagging. *Proc Natl Acad Sci U S A* **104**, 14110-14115.

Jiao Y, Peluso P, Shi J*, et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524-527.

Kalendar R, Vicient CM, Peleg O*, et al.* (2004) Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes. *Genetics* **166**, 1437.

Kaminker JS, Bergman CM, Kronmiller B*, et al.* (2002) The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol* **3**, Research0084.

Kang J, Kim J, Choi K-W (2011) Novel Cytochrome P450, cyp6a17, Is Required for Temperature Preference Behavior in Drosophila. *PLoS One* **6**, e29800.

Kapitonov VV, Jurka J (2003) Molecular paleontology of transposable elements in the Drosophila melanogaster genome. *Proc Natl Acad Sci U S A* **100**, 6569-6574.

Kapun M, Barrón MG, Staubach F*, et al.* (2020) Genomic analysis of European Drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Molecular Biology and Evolution*.

Kent WJ (2002) BLAT--the BLAST-like alignment tool. *Genome Res* **12**, 656-664.

Khost DE, Eickbush DG, Larracuente AM (2017) Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in Drosophila melanogaster. *Genome Res* **27**, 709-721.

Kojima KK, Jurka J (2011) Crypton transposons: identification of new diverse families and ancient domestication events. *Mob DNA* **2**, 12.

Koren S, Walenz BP, Berlin K*, et al.* (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**, 722-736.

Kou Y, Liao Y, Toivainen T*, et al.* (2020) Evolutionary genomics of structural variation in Asian rice (Oryza sativa) domestication. *Molecular Biology and Evolution*.

Kuleshov MV, Jones MR, Rouillard AD*, et al.* (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97.

Kurtz S, Phillippy A, Delcher AL*, et al.* (2004) Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12.

Lam K-K, LaButti K, Khalak A, Tse D (2015) FinisherSC: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* **31**, 3207-3209.

Lerat E, Goubert C, Guirao-Rico S*, et al.* (2019) Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. *Molecular ecology* **28**, 1506-1522.

Levy-Sakin M, Pastor S, Mostovoy Y*, et al.* (2019) Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nature Communications* **10**, 1025.

Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993.

Li H (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100.

Linheiro RS, Bergman CM (2012) Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster. *PLoS One* **7**, e30008.

Liu Y, Du H, Li P*, et al.* (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell* **182**, 162-176.e113.

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550-550.

Mahmoud M, Gobet N, Cruz-Dávalos DI*, et al.* (2019) Structural variant calling: the long and the short of it. *Genome Biology* **20**, 246.

Marçais G, Delcher AL, Phillippy AM*, et al.* (2018) MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology* **14**, e1005944.

McKenna A, Hanna M, Banks E*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303.

Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD (2018) PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Research* **47**, D419-D426.

Michael TP, Jupe F, Bemm F*, et al.* (2018) High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications* **9**, 541.

Miga KH, Koren S, Rhie A*, et al.* (2020) Telomere-to-telomere assembly of a complete human X chromosome. *Nature*.

Miller DE, Staber C, Zeitlinger J, Hawley RS (2018) Highly Contiguous Genome Assemblies of 15 Drosophila Species Generated Using Nanopore Sequencing. *G3 (Bethesda)* **8**, 3131-3141.

Mitsuhashi S, Matsumoto N (2020) Long-read sequencing for rare human genetic diseases. *Journal of Human Genetics* **65**, 11-19.

Mohamed M, Dang NT, Ogyama Y*, et al.* (2020) A Transposon Story: From TE Content to TE Dynamic Invasion of Drosophila Genomes Using the Single-Molecule Sequencing Technology from Oxford Nanopore. *Cells* **9**.

Palazzo A, Lovero D, D'Addabbo P, Caizzi R, Marsano RM (2016) Identification of Bari Transposons in 23 Sequenced Drosophila Genomes Reveals Novel Structural Variants, MITEs and Horizontal Transfer. *PLoS One* **11**, e0156014.

Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P (2010) On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. *Mol Biol Evol* **27**, 1226-1234.

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419.

Quesneville H, Bergman CM, Andrieu O*, et al.* (2005) Combined Evidence Annotation of Transposable Elements in Genome Sequences. *PLOS Computational Biology* **1**, e22.

Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841-842.

Rahman R, Chirn GW, Kanodia A*, et al.* (2015) Unique transposon landscapes are pervasive across Drosophila melanogaster genomes. *Nucleic Acids Res* **43**, 10655-10672.

Rech GE, Bogaerts-Márquez M, Barrón MG*, et al.* (2019) Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLOS Genetics* **15**, e1007900.

Roach MJ, Schmidt SA, Borneman AR (2018) Purge Haplotigs: Synteny Reduction for Third-gen Diploid Genome Assemblies. *bioRxiv*.

Sakamoto Y, Sereewattanawoot S, Suzuki A (2020) A new era of long-read sequencing for cancer genomics. *Journal of Human Genetics* **65**, 3-10.

Schmidt JM, Good RT, Appleton B*, et al.* (2010) Copy number variation and transposable elements feature in recent, ongoing adaptation at the Cyp6g1 locus. *PLoS Genet* **6**, e1000998.

Shahid S, Slotkin RK (2020) The current revolution in transposable element biology enabled by long reads. *Current Opinion in Plant Biology* **54**, 49-56.

Smit A, Hubley, R & Green, P. (2015) *RepeatMasker Open-4.0.* http://www.repeatmasker.org

Smith CD, Shu S, Mungall CJ, Karpen GH (2007) The Release 5.1 annotation of Drosophila melanogaster heterochromatin. *Science (New York, N.Y.)* **316**, 1586-1591.

Solares EA, Chakraborty M, Miller DE*, et al.* (2018) Rapid Low-Cost Assembly of the <em>Drosophila melanogaster</em> Reference Genome Using Low-Coverage, Long-Read Sequencing. *G3: Genes|Genomes|Genetics* **8**, 3143-3154.

Szpiech ZA, Hernandez RD (2014) selscan: An Efficient Multithreaded Program to Perform EHH-Based Scans for Positive Selection. *Molecular Biology and Evolution* **31**, 2824-2827.

Thomas J, Vadnagara K, Pritham EJ (2014) DINE-1, the highest copy number repeats in Drosophila melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (Helentrons). *Mob DNA* **5**, 18.

Thurmond J, Goodman JL, Strelets VB*, et al.* (2018) FlyBase 2.0: the next generation. *Nucleic Acids Research* **47**, D759-D765.

Torres R, Szpiech ZA, Hernandez RD (2018) Human demographic history has amplified the effects of background selection across the genome. *PLoS Genet* **14**, e1007387.

Treangen TJ, Salzberg SL (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics* **13**, 36-46.

Ullastres A, Merenciano M, González J (2021) Regulatory regions in natural transposable element insertions drive interindividual differences in response to immune challenges in Drosophila. *Genome Biology* **22**, 265.

Van der Auwera GA, Carneiro MO, Hartl C*, et al.* (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**, 11.10.11-33.

Vaser R, Sovic I, Nagarajan N, Sikic M (2017) Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746.

Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A Map of Recent Positive Selection in the Human Genome. *PLOS Biology* **4**, e72.

Walker BJ, Abeel T, Shea T*, et al.* (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS One* **9**, e112963.

Wallau GL, Capy P, Loreto E, Hua-Van A (2014) Genomic landscape and evolutionary dynamics of mariner transposable elements within the Drosophila genus. *BMC Genomics* **15**, 727.

Waterhouse RM, Seppey M, Simao FA*, et al.* (2017) BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol*.

Wick RR, Judd LM, Holt KE (2019) Performance of neural network basecalling tools for Oxford Nanopore sequencing. *Genome Biology* **20**, 129.

Wierzbicki F, Schwarz F, Cannalonga O, Kofler R (2020) Generating high quality assemblies for genomic analysis of transposable elements. *bioRxiv*, 2020.2003.2027.011312.

Yang X, Lee W-P, Ye K, Lee C (2019) One reference genome is not enough. *Genome Biology* **20**, 104-104.

Zhou Y, Minio A, Massonnet M*, et al.* (2019) The population genetics of structural variants in grapevine domestication. *Nature Plants* **5**, 965-979.

Zhuang J, Wang J, Theurkauf W, Weng Z (2014) TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Research* **42**, 6826-6838.

**Figure 1. Geographical location of the 12 *D. melanogaster* natural populations analysed in this work**.

The 32 sequenced and assembled genomes correspond to strains obtained from: Tenerife, Spain: TEN (1), Munich, Germany: MUN (6), Gimenells, Spain: GIM (2), Raleigh, USA: RAL (8), Cortes de Baza, Spain: COR (4), Tomelloso, Spain: TOM (2), Jutland, Denmark: JUT (2), Stockholm, Sweden: STO (1), Lund, Sweden: LUN (2), Slankamen, Serbia: SLA (1), Kiev, Ukraine: KIE (1) and Akka, Finland: AKA (2). In brackets, the number of genomes sequenced at that location. Map colours represent different climatic regions according to the Köppen climate classification.

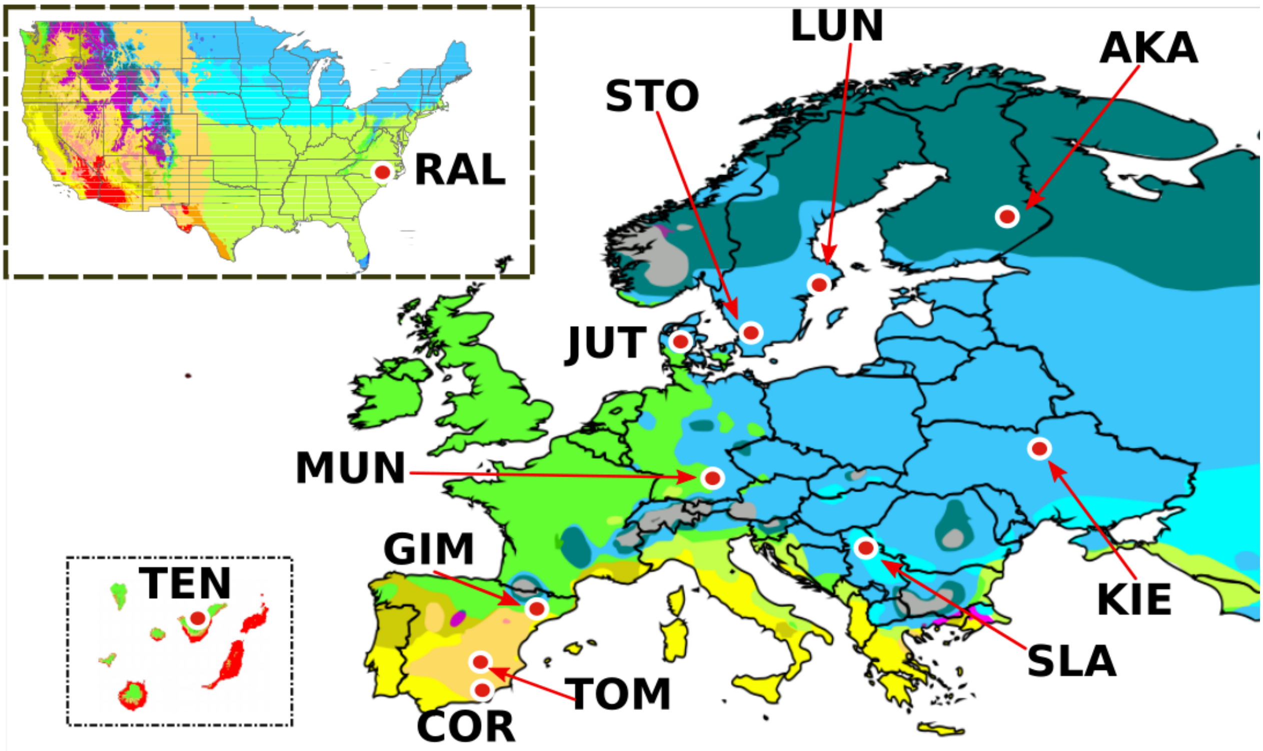

**Figure 2**. **Comparison between TE annotation in FlyBase and the TE annotation performed using *REPET* with the MCTE library.**
**A)** Number of TE copies per family. **B)** Overlapping of TE annotations considering that the copies were from the same family and that they were overlapping at least 95% of their lengths (breadth of coverage). TEs shorter than 100bp, from the INE-1 family and nested TEs were excluded from the analysis. **C)** Distribution of number of TE copies of different sizes.

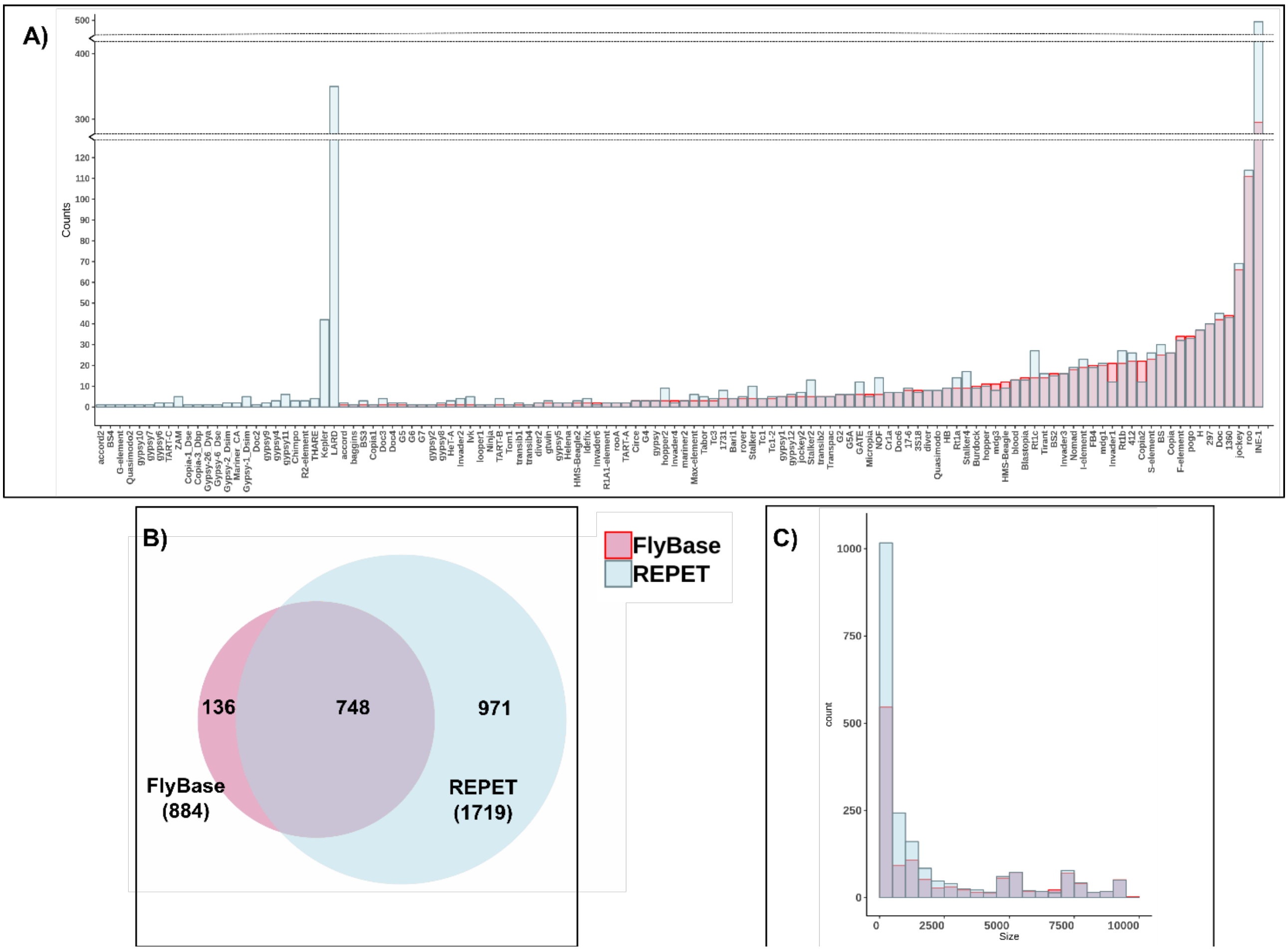

**Figure 3. Three new TE families in *D. melanogaster*. A)** Schematic representation of the structural features detected by *PASTEC* in the consensus sequences of the three new families identified in this study. **B)** Length ratio (size as proportion of the consensus) distribution for TE copies annotated in the 32 genomes with each of the three new consensus sequences.

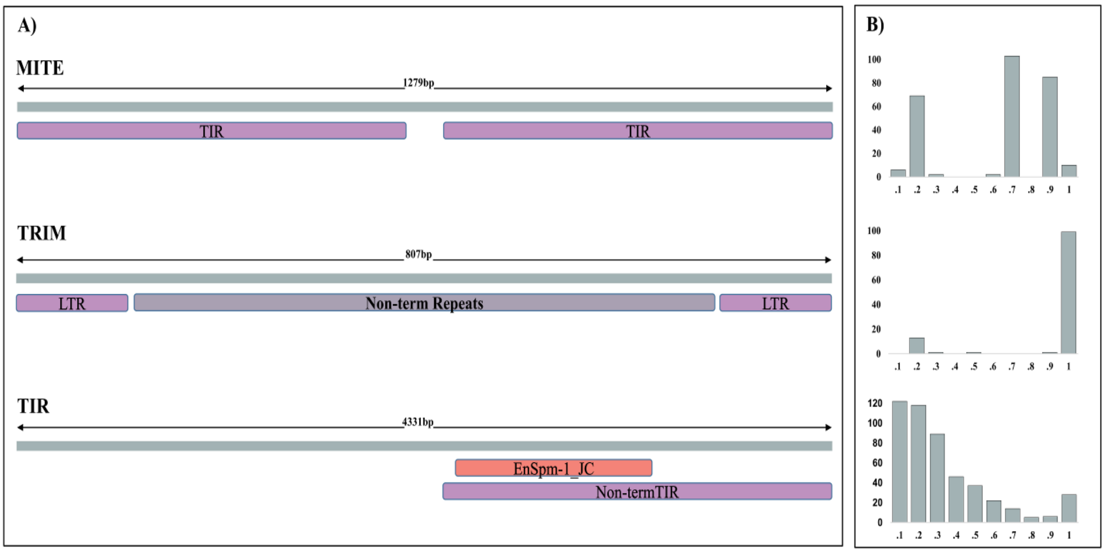

**Figure 4**. TE annotations at the superfamily level. **A**) Principal component analysis based on TE insertions polymorphisms grouped by continent (colours) and climatic zoned (shapes). **B)** The proportion of TE copies annotated for each superfamily. **C)** Per genome pairwise comparisons in the proportion of copies annotated at the superfamily level. The colours of the matrix squares represent adjusted (FDR) p-values of the Chi Square test. Only one significant result was observed (adjusted p-value=0.03) between ISO1 and MUN-009. **D)** Representation of the Pearson residuals (r) for each cell (pair Superfamily-genome). Cells with the highest residuals contribute the most to the total Chi Square score. Positive values in cells (red) represent more copies than the expected, while negative residuals (blue) represent fewer copies than the expected (does not imply statistical significance). **E)** Distribution of TE insertion identity values classified by superfamily and considering all genomes together.

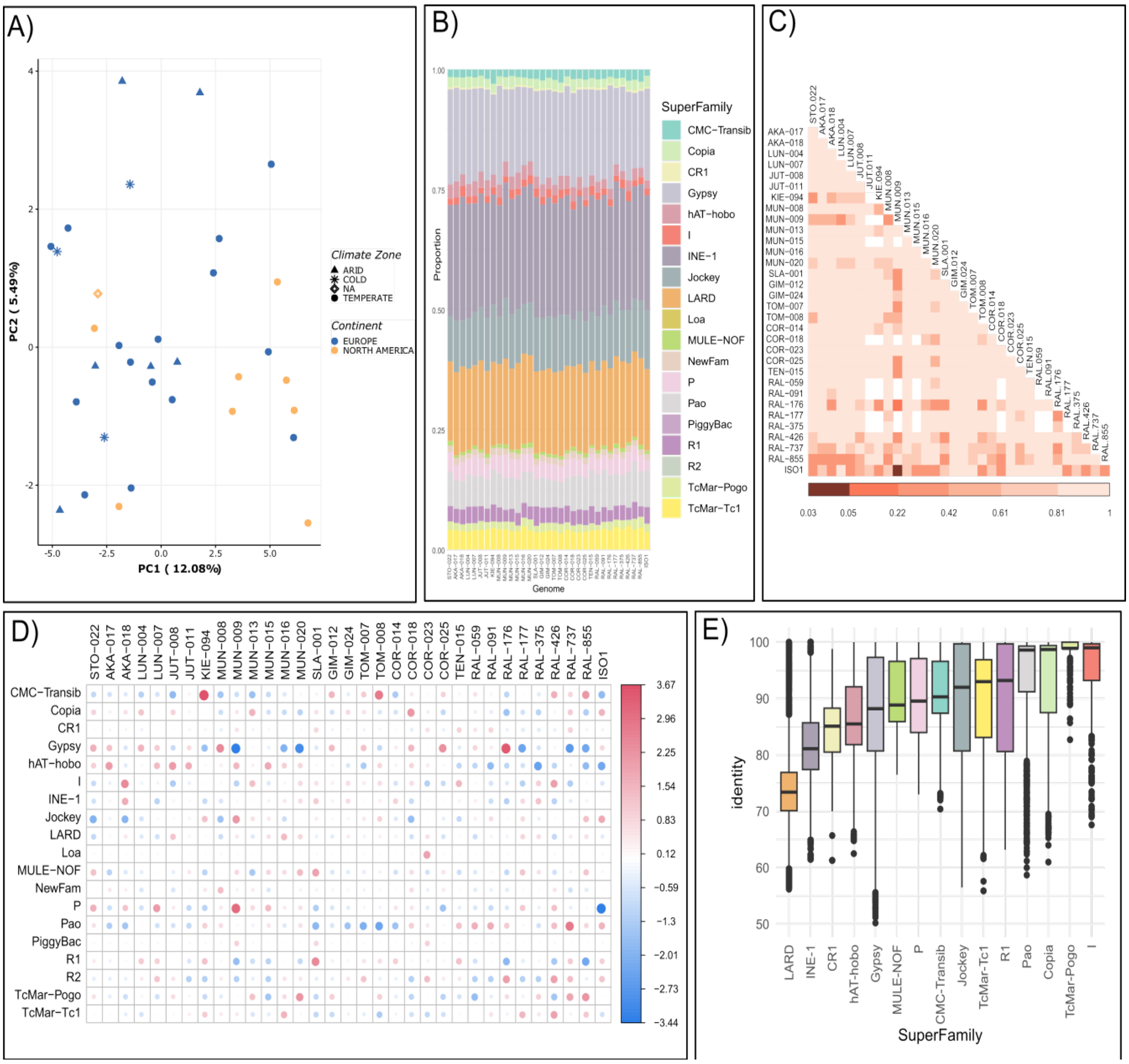

**Figure 5**. TE classification according to three frequency classes: rare (present in <10 of the strains), common (present in ≥10 and ≤95% of the strains) and fixed (present in >95% of the strains). **A)** Number of TEs and their classification according to their frequency in the population using from 5 to 47 strains. The standard deviation was calculated by taking 30 random samples of strains for each case. **B)** Intersection of the different sets of common TEs identified taking into account 10, 20, 30, 40 and 47 strains at random. **C)** Venn diagrams depicting the intersection of orthologous TEs defined by geographic origin. The ALL diagram represents all TEs regardless their frequency class, while the rare, common and fixed diagrams are defined by the TEs of each of the classes in each set.

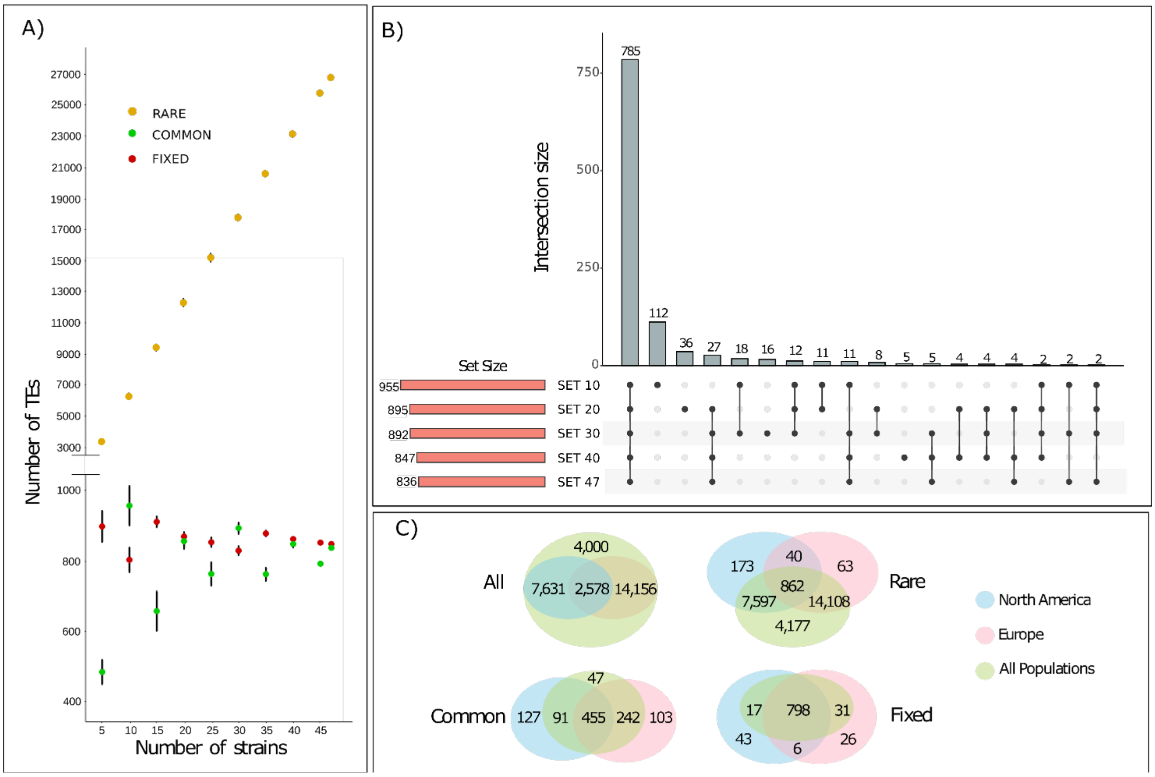

**Figure 6. Gene expression levels in strains with and without TE insertions.**
Gene expression levels in strains without (grey) and with (red) the 13 TE insertions with the most significant association according to our eQTL analysis, and for the *3L_14050243_14050245_pogo* insertion with evidence of selection (last plot). The name of the TE insertions and the genomic location regarding the associated gene is provided.

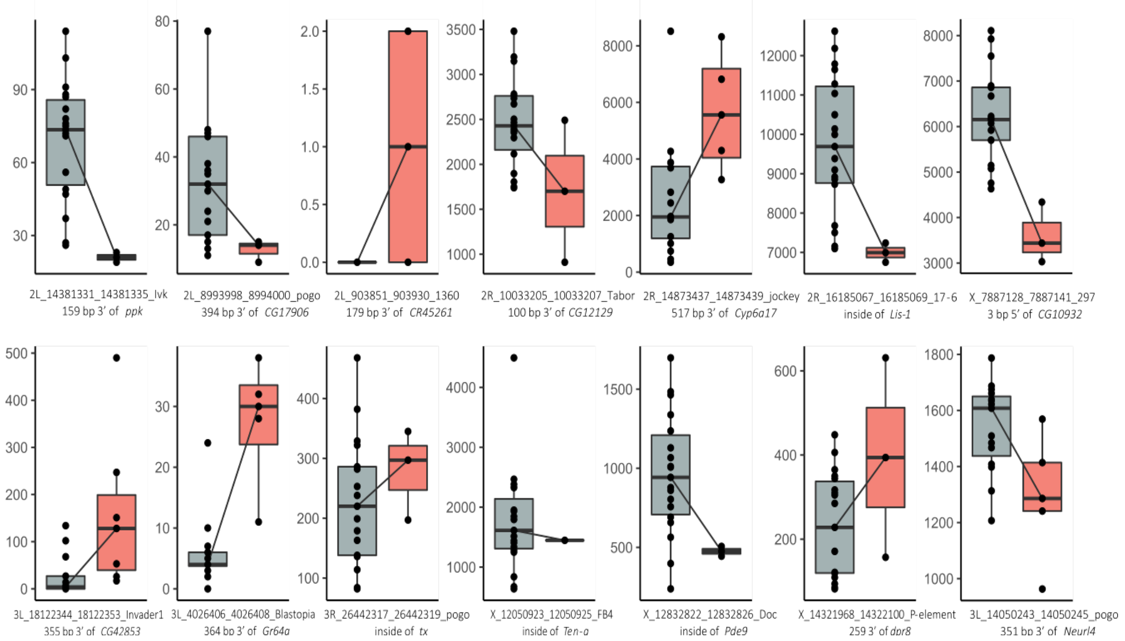

**Figure 7**. **Significantly enriched terms for genes nearby 107 TEs showing evidence of selection**.

Each panel shows significant enriched terms using different approaches. **A)** DAVID GO Biological Process: Horizontal axis represents DAVID enrichment score. Only significant (score >1.3) and non-redundant clusters are shown. **B-E)** FlyEnrichr results when using different libraries: **B)** Anatomy GeneRIF Predicted, **C)** Allele LoF Phenotypes from FlyBase, **D)** Putative Regulatory miRNAs from DroID and **E)** Transcription Factors from DroID. Only statistically significant terms are shown (adjusted p-value < 0.05). Horizontal axis represents the *Enrichr* Combined Score. For Regulatory miRNAs and Transcription Factors, putative biological functions or phenotypes associated were assigned based on FlyBase gene summaries. Bar colours indicate similar biological functions as specified at the bottom of the figure.

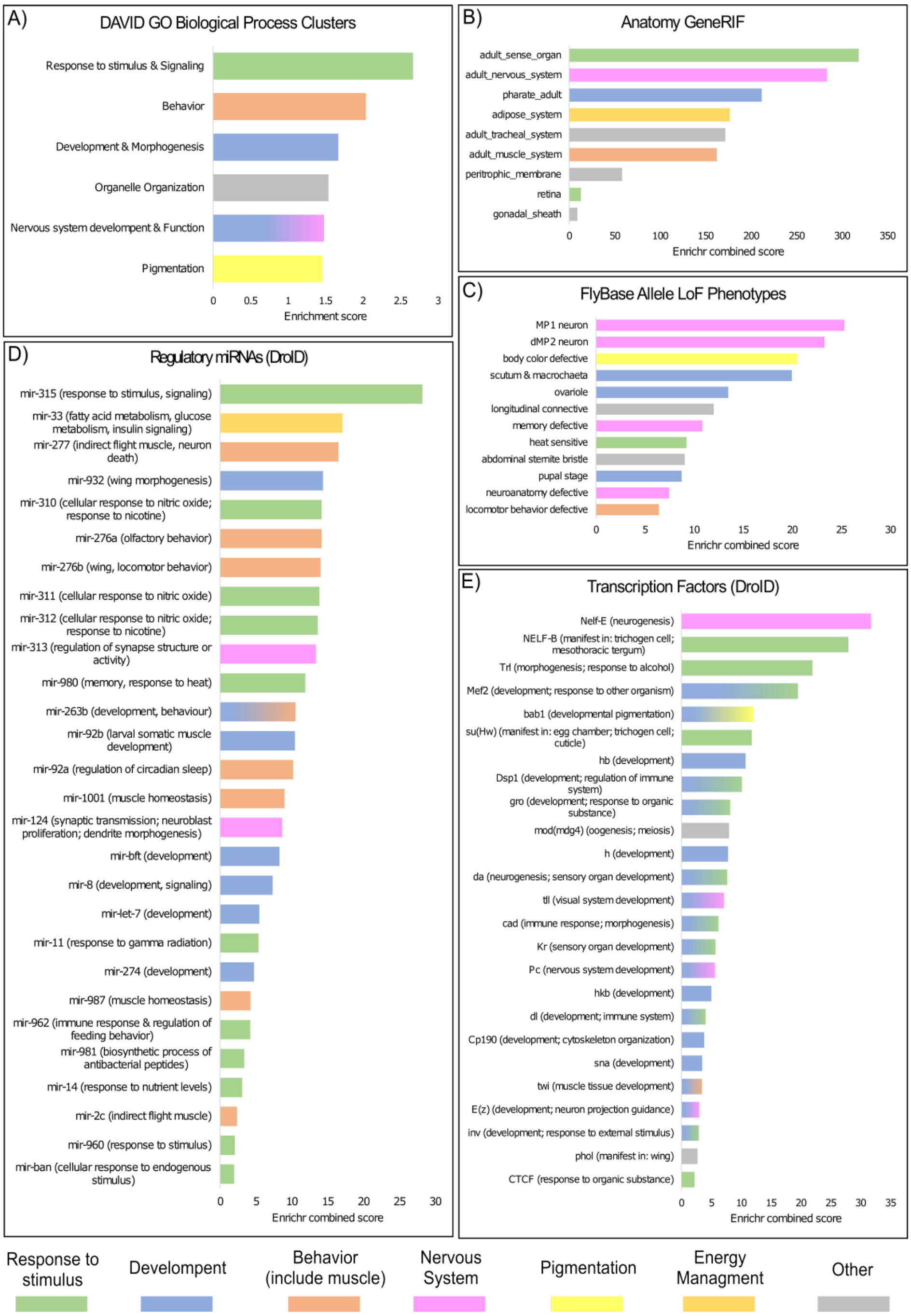

**Table 1.** Summary of assembly metrics of the 32 genomes sequenced in this work. Genomes were sequenced using ONT except MUN-016 and TOM-007 that were sequenced using PacBio. *Indicates the 13 strains used in the construction of the *de novo* MCTE library. $^{+}$Indicates the 11 strains used in the comparison of TE annotations using *REPET*, *TIDAL* and *TEMP*. $^{x}$Indicates the 20 strains used in the cis-eQTL analysis. $^{(S)}$ Genome assembled using long-read sequencing data of the *D. melanogaster* reference genome provided in Solares et al. (2018). Additional information on the strains can be found in Table S1 and on the sequencing in Tables S2 and S3.

| Strain | Location | Contigs | Genome Size | N50 (Mb) | BUSCO Complete | BUSCO Duplicate | QV | c.CUSCO | sc.CUSCO | Completeness (ISO1 aligned Bases) | Euchromatic Size (Mb) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| AKA-017*$^{+}$ | Akka, Finland | 164 | 142.7 | 18.9 | 98.7% | 0.50% | 51.04 | 82.35% | 94.12% | 96.30% | 100.1 |
| AKA-018$^{x}$ | Akka, Finland | 162 | 136.7 | 2.3 | 98.4% | 0.70% | 37.63 | 72.94% | 92.94% | 93.50% | 100.9 |
| COR-014*$^{+}$ | Cortes de Baza, Spain | 161 | 138.1 | 7.7 | 98.3% | 0.50% | 43.62 | 72.94% | 96.47% | 96.70% | 100.4 |
| COR-018$^{x}$ | Cortes de Baza, Spain | 402 | 143.5 | 0.9 | 98.0% | 1.00% | 38.47 | 55.29% | 96.47% | 94.30% | 103.3 |
| COR-023$^{x}$ | Cortes de Baza, Spain | 620 | 139.5 | 0.6 | 97.8% | 0.80% | 37.42 | 35.29% | 92.94% | 93.60% | 101.5 |
| COR-025$^{x}$ | Cortes de Baza, Spain | 377 | 143.4 | 0.7 | 98.1% | 1.00% | 37.83 | 57.65% | 92.94% | 94.00% | 102.7 |
| GIM-012$^{x}$ | Gimenells, Spain | 383 | 140 | 1.2 | 98.4% | 0.80% | 40.56 | 45.88% | 87.06% | 94.10% | 101.2 |
| GIM-024*$^{+\ x}$ | Gimenells, Spain | 316 | 142.3 | 6.8 | 99.0% | 0.50% | 50.77 | 77.65% | 94.12% | 95.20% | 100.2 |
| JUT-008$^{x}$ | Jutland, Denmark | 330 | 148.5 | 9.6 | 98.4% | 0.50% | 49.52 | 80.00% | 96.47% | 93.60% | 101.5 |
| JUT-011*$^{+}$ | Jutland, Denmark | 184 | 138.4 | 4 | 98.7% | 0.50% | 44.94 | 70.59% | 98.82% | 96.50% | 100.8 |
| KIE-094*$^{+}$ | Kiev, Ucrania | 343 | 143.8 | 3.8 | 98.7% | 0.80% | 48.78 | 75.29% | 96.47% | 96.20% | 101.9 |
| LUN-004*$^{+}$ | Lund, Sweden | 314 | 138.1 | 2 | 98.7% | 0.60% | 44.24 | 62.35% | 96.47% | 96.30% | 101.1 |
| LUN-007$^{x}$ | Lund, Sweden | 360 | 142.4 | 1.1 | 98.0% | 0.60% | 39.91 | 52.94% | 95.29% | 94.10% | 102.1 |
| MUN-008$^{x}$ | Munich, Germany | 250 | 142.2 | 1.1 | 97.5% | 0.90% | 37.76 | 68.24% | 94.12% | 94.10% | 101.7 |
| MUN-009 | Munich, Germany | 385 | 149.3 | 5.6 | 97.9% | 0.50% | 45.97 | 71.76% | 95.29% | 94.10% | 102.1 |
| MUN-013$^{x}$ | Munich, Germany | 406 | 138.4 | 1 | 98.2% | 0.50% | 39.28 | 49.41% | 90.59% | 93.80% | 101.9 |
| MUN-015 | Munich, Germany | 251 | 140 | 1.2 | 98.0% | 1.00% | 38.19 | 65.88% | 92.94% | 93.90% | 101.8 |
| MUN-016* | Munich, Germany | 217 | 142 | 7.8 | 98.50% | 0.60% | NA | 77.65% | 92.94% | 96.60% | 100.7 |
| MUN-020$^{x}$ | Munich, Germany | 324 | 138.1 | 1.3 | 97.10% | 1.10% | 40.93 | 48.24% | 82.35% | 93.80% | 101.2 |
| RAL-059$^{x}$ | Raleigh, USA | 688 | 143.5 | 0.8 | 98.10% | 0.90% | 43.25 | 51.76% | 94.12% | 93.20% | 101.7 |
| RAL-091$^{x}$ | Raleigh, USA | 887 | 145.1 | 0.5 | 97.50% | 1.00% | 44.04 | 57.65% | 92.94% | 92.80% | 103.9 |
| RAL-176$^{x}$ | Raleigh, USA | 1185 | 151.3 | 0.4 | 97.10% | 0.80% | 46.62 | 43.53% | 88.24% | 92.70% | 102.9 |
| RAL-177*$^{+x}$ | Raleigh, USA | 188 | 141.9 | 14.6 | 97.40% | 0.40% | 46.70 | 84.71% | 96.47% | 95.70% | 100.7 |
| RAL-375*$^{+x}$ | Raleigh, USA | 179 | 141.2 | 13.5 | 96.10% | 0.40% | 44.86 | 82.35% | 96.47% | 96.10% | 100.7 |
| RAL-426$^{x}$ | Raleigh, USA | 500 | 137 | 0.7 | 97.60% | 0.50% | 38.04 | 51.76% | 90.59% | 93.50% | 102.0 |
| RAL-737$^{x}$ | Raleigh, USA | 469 | 147.8 | 1.5 | 97.40% | 0.50% | 42.11 | 70.59% | 95.29% | 93.20% | 102.1 |
| RAL-855$^{x}$ | Raleigh, USA | 332 | 144.4 | 3.9 | 97.00% | 0.40% | 41.78 | 78.82% | 97.65% | 93.40% | 102.2 |
| SLA-001*$^{+}$ | Slankamen, Serbia | 432 | 143.7 | 0.8 | 97.90% | 0.80% | 38.45 | 58.82% | 97.65% | 96.60% | 103.0 |
| STO-022*$^{+}$ | Stockholm, Sweden | 153 | 142.4 | 3.1 | 98.10% | 0.70% | 36.00 | 71.76% | 96.47% | 96.90% | 102.5 |
| TEN-015*$^{+}$ | Tenerife, Spain | 329 | 140.5 | 1.1 | 97.90% | 1.00% | 40.30 | 61.18% | 94.12% | 96.20% | 102.0 |
| TOM-007 | Tomelloso, Spain | 222 | 139.5 | 3.2 | 98.20% | 0.70% | NA | 57.65% | 92.94% | 96.90% | 101.0 |
| TOM-008*$^{x}$ | Tomelloso, Spain | 219 | 136.6 | 1.9 | 98.10% | 0.80% | 41.75 | 61.18% | 85.88% | 94.10% | 101.3 |
| ISO1-Sol $^{(S)}$ | Reference Genome | 518 | 147.8 | 3.4 | 96.00% | 0.50% | 42.92 | 77.65% | 91.76% | 97.57% | 101.9 |

**Table 2.** TEs showing the highest significance values in their association with the expression of a nearby gene (adjusted p-value <= 0.01). Note that for *Ten-a* gene there were two TEs with equal nominal p-value.

| TE ID | Freq. | Gene symbol | Gene expression | Biological process |
|---|---|---|---|---|
| 2L_903851_903930_1360 | 0.30 | CR45261 | Up | - |
| 2L_8993998_8994000_pogo | 0.15 | CG17906 | Down | - |
| 2L_14381331_14381335_Ivk | 0.10 | ppk | Down | Behavior, Response to stimulus |
| 2R_10033205_10033207_Tabor | 0.10 | CG12129 | Down | - |
| 2R_14873437_14873439_jockey | 0.20 | Cyp6a17 | Up | Response to stimulus, Behavior (thermosensory) |
| 2R_16185067_16185069_17-6 | 0.10 | Lis-1 | Down | Development, Reproduction, Transport/localization, Cell organization/biogenesis, cell cycle/proliferation, Response to stimulus |
| 3L_4026406_4026408_Blastopia | 0.20 | Gr64a | Up | Response to stimulus, Nervous system process |
| 3L_18122344_18122353_Invader1 | 0.35 | CG42853 | Up | - |
| 3R_26442317_26442319_pogo | 0.15 | tx | Down | Development, Gene expression |
| X_7887128_7887141_297 | 0.15 | CG10932 | Down | Small molecule metabolism |
| X_12832822_12832826_Doc | 0.10 | Pde9 | Down | Response to stimulus, Signaling |
| X_14321968_14322100_P-element | 0.10 | dpr8 | Up | Nervous system process, Cell organization/biogenesis |
| X_12050923_12050925_FB4<br>X_12050923_12050925_FB4.t1 | 0.05<br>0.05 | Ten-a | Down | Development, Cell organization/biogenesis, Response to stimulus |

**Table 3.** Eighteen candidate adaptive TE insertions showing evidence of selection identified in this work. Biological process information according to FlyBase.

| **TE ID** | **Evidence of Selection** | **Freq** | **Gene Symbol** | **TE Location** | **Biological process (experimental evidence)** |
|---|---|---|---|---|---|
| 2L_14003409_14003462_Rt1a | nSL | 15% | - | Intergenic | - |
| 2L_8992666_8992668_pogo | nSL | 15% | CG9555 | Intron | NA |
| 2R_11394154_11394156_pogo | nSL | 17% | sprt | Intron | NA |
| 2R_12185376_12185380_accord | nSL | 62% | Cyp6g1 | Promoter | response to insecticide |
| 2R_14078395_14078397_hopper | nSL | 11% | Prosap | Intron | synaptic assembly at neuromuscular junction |
| 2R_18807888_18807894_BS | nSL | 62% | CG15096 | 3UTR | transmembrane transport |
| 3L_12863739_12863742_Transpac | nSL | 19% | CG10943 | Promoter | NA |
| 3L_14050243_14050245_pogo | nSL | 28% | CG6833 | Promoter | NA |
| | | | Neurl4 | Promoter | NA |
| 3L_2426710_2426713_pogo | nSL | 19% | Svil | Intron | NA |
| 3L_3798612_3798621_1360 | nSL | 30% | CG32264 | Intron | NA |
| 3R_20502048_20502058_Doc | nSL | 28% | Dic2 | Promoter | NA |
| | | | CG46441 | Promoter | NA |
| 3R_21385503_21385506_pogo | nSL | 19% | - | Intergenic | - |
| 3R_29952746_29952748_Invader4 | nSL | 23% | TkR99D | Intron | olfactory behavior; detection of chemical stimulus |
| X_15012530_15012533_mdg3 | nSL | 60% | hiw | Intron | autophagy; long-term memory; synapse organization; response to axon injury |
| X_20759991_20759993_BS3 | nSL | 57% | - | Intergenic | - |
| X_2431713_2431716_Doc | nSL | 13% | - | Intergenic | - |
| X_8027468_8027478_Doc6 | nSL | 26% | Tbh | 3UTR | aggressive behavior; behavioral response to ethanol; flight behavior; learning; ovulation |
| 3L_18931204_18931207_F-element | nSL | 15% | CG32204 | Intron | NA |