



**HAL**  
open science

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

Gabriel Rech, Santiago Radío, Sara Guirao-Rico, Laura Aguilera, Vivien Horvath, Llewellyn Green, Hannah Lindstadt, Véronique Jamilloux, Hadi Quesneville, Josefa González

### ► To cite this version:

Gabriel Rech, Santiago Radío, Sara Guirao-Rico, Laura Aguilera, Vivien Horvath, et al.. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. *Nature Communications*, 2022, 13 (1), pp.1-16. 10.1038/s41467-022-29518-8 . hal-03807727

**HAL Id: hal-03807727**

**<https://hal.science/hal-03807727>**

Submitted on 31 Jan 2023



**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

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

Gabriel E. Rech<sup>1</sup>, Santiago Radío<sup>1</sup>, Sara Guirao-Rico<sup>1</sup>, Laura Aguilera <sup>1</sup>, Vivien Horvath<sup>1</sup>, Llewellyn Green<sup>1</sup>, Hannah Lindstadt<sup>1</sup>, Véronique Jamilloux<sup>2</sup>, Hadi Quesneville<sup>2</sup> & Josefa González <sup>1</sup>✉

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 generate 32 high-quality reference genomes for the well-known model species *D. melanogaster* and focus on the identification and analysis of transposable element variation as they are the most common type of structural variant. We show 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 identify 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.

<sup>1</sup>Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), 08003 Barcelona, Spain. <sup>2</sup>Université Paris-Saclay, INRAE, URGI, 78026 Versailles, France. ✉email: [josefa.gonzalez@ibe.upf-csic.es](mailto:josefa.gonzalez@ibe.upf-csic.es)

Despite their crucial role and high prevalence in most eukaryotic genomes, transposable elements (TEs) and other structural variants (SVs) remain largely understudied. This is mainly a consequence of the limitations of high throughput sequencing read length, tightly restricted to short-reads in the last decades<sup>1–3</sup>. Short-reads not only limited the annotation of SVs to what inference methods were able to identify<sup>4–9</sup>, 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<sup>4</sup>; (ii) the analyses are biased to what is possible to identify using a specific reference genome<sup>3,10,11</sup>; and (iii) repetitive sequences in the reference genome are not well characterized when they are longer than the sequenced reads<sup>12</sup>. 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<sup>13</sup>. 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<sup>14</sup>.

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<sup>15–20</sup>, 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<sup>21,22</sup>. 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<sup>3,8,23–25</sup>, which can lead to a significant improvement in our ability to study TE structure, activity and dynamics in different organisms<sup>20,26,27</sup>.

*Drosophila melanogaster* represents one of the best model animals for studying TEs, not only for having one of the best annotated eukaryotic genomes<sup>28,29</sup>, but also for containing several active TE families<sup>30</sup>. Interestingly, even in such a well-studied organism, long-read sequencing approaches have made novel insights into the evolutionary dynamics of TEs<sup>8,31,32</sup>. 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.

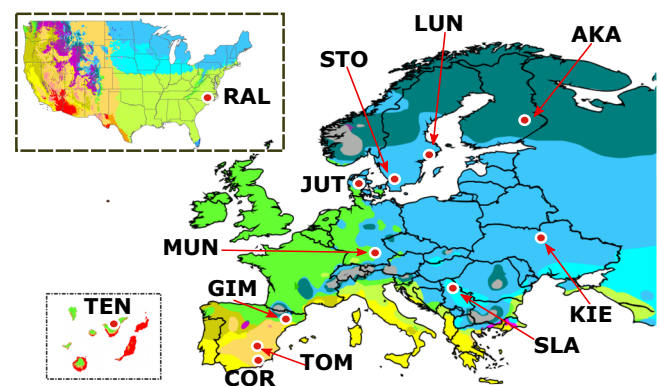
Here, we used long-read sequences to generate high quality genome assemblies for 32 *D. melanogaster* natural strains collected mainly in Europe from populations located in five different climatic regions and belonging to three of the five main climate types (Fig. 1). We used this new genomic resource for the de novo construction 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 compared with 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; identified 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<sup>33</sup>.

## 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 (Fig. 1, Table 1, Supplementary Data 1 and 2, Supplementary Note 1). These 32 strains were collected from 12 geographical locations: 24 strains were collected from 11 European locations and eight strains were collected in a North American population<sup>34</sup>. These 12 populations represent five different climatic regions belonging to three main climatic types: arid, temperate, and cold (Fig. 1; Supplementary Data 1). Long-read sequencing resulted in 458.7 Gb, representing a theoretical average coverage of 82X (ranging from 45X to 123X) and average read length > 5.6 Kb, which has been previously shown to be sufficient for generating highly contiguous genome assemblies in other *Drosophila* species<sup>35</sup>; Supplementary Data 2).

Genome assembly, polishing, deduplication and contaminant removal resulted in genomes with a number of contigs ranging from 153 to 1185 (average 367), genome sizes from 136.6 Mb to 151.3 Mb (average 142 Mb), N50 values from 400 Kb to 18.9 Mb (average 3.8 Mb) 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 Notes 2–4). CUSCO scores, i.e., percentage of contiguously assembled piRNA clusters<sup>36</sup>, range from 35.3% to 84.7% (average 64.1%; Table 1). The detectability of a cluster was inversely correlated with its size (Pearson's correlation = -0.47; Supplementary Data 3b, Supplementary Fig. 1 and Supplementary Note 5). Although the high variability, these results are comparable with genomes previously obtained using similar sequencing and assembling strategies<sup>35</sup>. Note that differences in sequencing coverage did not explain the observed differences in genome size or TE content across genomes (Supplementary Fig. 2). Similarly, differences in read length and N50 values do not correlate with differences in genome size, TE content, or BUSCO scores (Supplementary Fig. 2).

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



**Fig. 1 Geographical location of the 12 *D. melanogaster* natural populations analyzed 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 from each location. Map colors represent different climatic regions according to the Köppen climate classification (Supplementary Data 1).

**Table 1 Summary of assembly metrics of the 32 genomes sequenced in this work.**

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 <sup>a,b</sup>	Akka, Finland	164	142.7	18.9	98.7%	0.50%	51.04	82.35%	94.12%	96.30%	100.1
AKA-018 <sup>c</sup>	Akka, Finland	162	136.7	2.3	98.4%	0.70%	37.63	72.94%	92.94%	93.50%	100.9
COR-014 <sup>a,b</sup>	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 <sup>c</sup>	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 <sup>c</sup>	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 <sup>c</sup>	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 <sup>c</sup>	Gimenells, Spain	383	140	1.2	98.4%	0.80%	40.56	45.88%	87.06%	94.10%	101.2
GIM-024 <sup>a,b,c</sup>	Gimenells, Spain	316	142.3	6.8	99.0%	0.50%	50.77	77.65%	94.12%	95.20%	100.2
JUT-008 <sup>c</sup>	Jutland, Denmark	330	148.5	9.6	98.4%	0.50%	49.52	80.00%	96.47%	93.60%	101.5
JUT-011 <sup>a,b</sup>	Jutland, Denmark	184	138.4	4	98.7%	0.50%	44.94	70.59%	98.82%	96.50%	100.8
KIE-094 <sup>a,b</sup>	Kiev, Ucraina	343	143.8	3.8	98.7%	0.80%	48.78	75.29%	96.47%	96.20%	101.9
LUN-004 <sup>a,b</sup>	Lund, Sweden	314	138.1	2	98.7%	0.60%	44.24	62.35%	96.47%	96.30%	101.1
LUN-007 <sup>c</sup>	Lund, Sweden	360	142.4	1.1	98.0%	0.60%	39.91	52.94%	95.29%	94.10%	102.1
MUN-008 <sup>c</sup>	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 <sup>c</sup>	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 <sup>a</sup>	Munich, Germany	217	142	7.8	98.50%	0.60%	NA	77.65%	92.94%	96.60%	100.7
MUN-020 <sup>c</sup>	Munich, Germany	324	138.1	1.3	97.10%	1.10%	40.93	48.24%	82.35%	93.80%	101.2
RAL-059 <sup>c</sup>	Raleigh, USA	688	143.5	0.8	98.10%	0.90%	43.25	51.76%	94.12%	93.20%	101.7
RAL-091 <sup>c</sup>	Raleigh, USA	887	145.1	0.5	97.50%	1.00%	44.04	57.65%	92.94%	92.80%	103.9
RAL-176 <sup>c</sup>	Raleigh, USA	1185	151.3	0.4	97.10%	0.80%	46.62	43.53%	88.24%	92.70%	102.9
RAL-177 <sup>a,b,c</sup>	Raleigh, USA	188	141.9	14.6	97.40%	0.40%	46.70	84.71%	96.47%	95.70%	100.7
RAL-375 <sup>a,b,c</sup>	Raleigh, USA	179	141.2	13.5	96.10%	0.40%	44.86	82.35%	96.47%	96.10%	100.7
RAL-426 <sup>c</sup>	Raleigh, USA	500	137	0.7	97.60%	0.50%	38.04	51.76%	90.59%	93.50%	102.0
RAL-737 <sup>c</sup>	Raleigh, USA	469	147.8	1.5	97.40%	0.50%	42.11	70.59%	95.29%	93.20%	102.1
RAL-855 <sup>c</sup>	Raleigh, USA	332	144.4	3.9	97.00%	0.40%	41.78	78.82%	97.65%	93.40%	102.2
SLA-001 <sup>a,b</sup>	Slankamen, Serbia	432	143.7	0.8	97.90%	0.80%	38.45	58.82%	97.65%	96.60%	103.0
STO-022 <sup>a,b</sup>	Stockholm, Sweden	153	142.4	3.1	98.10%	0.70%	36.00	71.76%	96.47%	96.90%	102.5
TEN-015 <sup>a,b</sup>	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 <sup>a,c</sup>	Tomelloso, Spain	219	136.6	1.9	98.10%	0.80%	41.75	61.18%	85.88%	94.10%	101.3
ISO1-Sold	Reference Genome	518	147.8	3.4	96.00%	0.50%	42.92	77.65%	91.76%	97.57%	101.9

Genomes were sequenced using ONT except MUN-016 and TOM-007 that were sequenced using PacBio.

Additional information on the strains can be found in Supplementary Data 1 and on the sequencing in Supplementary Tables S2 and S3. Besides contig-CUSCO scores (sc.CUSCO) are also given (the later values are higher as expected if the piRNA flanking regions are present in the assembled genomes).

<sup>a</sup>The 13 strains used in the construction of the de novo MCTE library.

<sup>b</sup>The 11 strains used in the comparison of TE annotations using REPET, TIDAL and TEMP.

<sup>c</sup>The 20 strains used in the cis-eQTL analysis.

<sup>d</sup>Genome assembled using long-read sequencing data of the *D. melanogaster* reference genome provided in Solares et al.<sup>18</sup>.

de novo assembled genomes (Supplementary Data 3a). The scaffolded genomes also showed a high level of completeness, covering on average around 95% of ISO1 major chromosomal arms (Table 1, Supplementary Fig. 3) and with an average of 99.75% of the protein coding genes successfully transferred (Supplementary Data 3a).

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<sup>18</sup>. The newly assembled reference genome was 147.8 Mb with a complete BUSCO score of 96% (Table 1). We identified 1842 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 (Supplementary Data 4). We also used the pipeline applied in Berlin et al.<sup>37</sup> to annotate TEs in an ISO-1 assembly based on PacBio sequencing, to annotate TEs in our ISO-1 assembly based on the Soares et al.<sup>18</sup> ONT reads. We found that 18% more TE insertions were annotated when using the Berlin et al.<sup>37</sup> assembly, suggesting that besides TE annotation pipelines, sequencing and assembly strategies can also influence the annotation of TEs in genomes.

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<sup>38</sup>. 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 Note 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*<sup>39</sup> (Supplementary Data 5). 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 100.1 Mb to 103.9 Mb (Table 1), which is a slightly larger region than in previous similar analysis (e.g., 94.5 Mb in Charkraborty et al.<sup>8</sup>). 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<sup>28</sup>. 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 *copial* 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, Fig. 2a, Supplementary Data 6a), 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, Supplementary Data 6a). At the genomic coordinates level, ~85% of the FlyBase copies were overlapping with *REPET* copies (95% reciprocal minimum breadth of coverage; Fig. 2b, Supplementary Data 6b). 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.<sup>40</sup>; Supplementary Data 6c). 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*<sup>29</sup>.

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 (Supplementary Data 6a, Fig. 2a). Although most of the new TE copies annotated only with *REPET* were small insertions, we also identified 50 insertions larger than 2 Kb (Fig. 2c, Supplementary Note 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 (Supplementary Data 6d, Supplementary Note 7). We found that 42–44% of the copies annotated using the MCTE library were not annotated by the BDGP library.

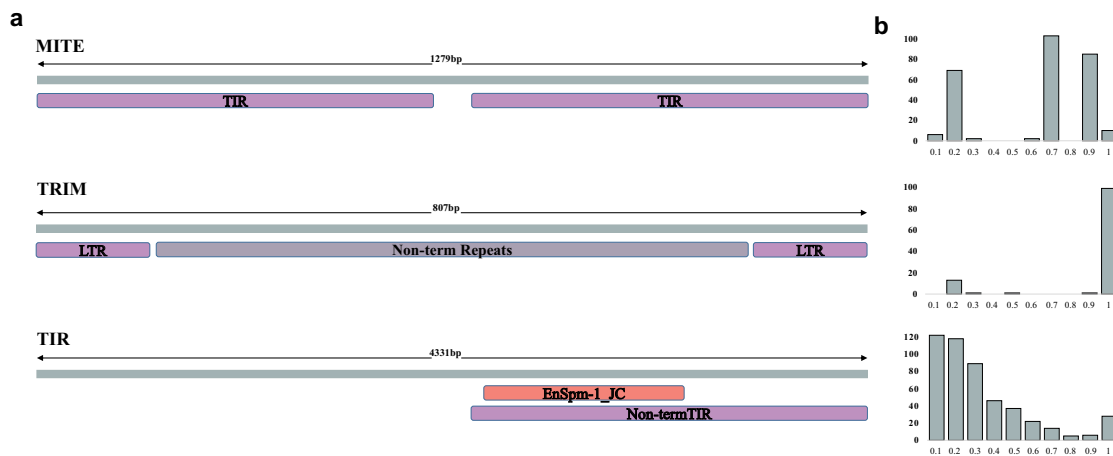
Overall, by creating a library that contains the TE diversity of 13 *D. melanogaster* strains from 12 geographical locations, we were able to identify TE copies from 25 known families not previously annotated in the reference euchromatic genome, and from three new families (see below). In total, 58% more insertions were annotated in the euchromatic reference genome using the MCTE library (1301 FlyBase vs 2059 *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*<sup>41</sup>. These new consensus sequences were classified as a Miniature Inverted Repeat Transposable Element (*MITE*), a Terminal Repeat Retrotransposon in Miniature (*TRIM*), and a Terminal Inverted Repeat (*TIR*) element (Fig. 3a).

Numerous Bari-like *MITEs*<sup>42</sup> and Mariner-like *MITEs*<sup>43</sup> have been previously described in *D. melanogaster*. However, the *MITE* consensus sequence identified in this work showed no significant alignments with any previously described *MITEs* (nucleotide identity percentage  $<50%$ ), suggesting that it belongs 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 (Supplementary Fig. 4a) and highly similar (average identity  $>89%$ , Supplementary Fig. 5). Moreover, the consensus sequence of the new *MITE* family showed no significant similarities with TEs identified in other five *Drosophila* genomes (Supplementary Data 7, Supplementary Note 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 1000 bp, with LTRs sequences between 100 bp and 250 bp, Fig. 3a), no





**Fig. 3** Three new TE families in *D. melanogaster*. **a** Schematic representation of the structural features detected by *PASTE*C 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.

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* (Supplementary Data 8b). 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* (Supplementary Data 8b). Finally, 10 of the 300 TEs annotated by *REPET*, were false positives as we could not confirm their presence using Blast (see Methods).

Additionally, we performed manual inspection of 50 TEs that were identified by *TEMP/TIDAL* but were not identified by *REPET* (Supplementary Data 8c). 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<sup>45</sup>. 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 (Supplementary Data 8c).

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

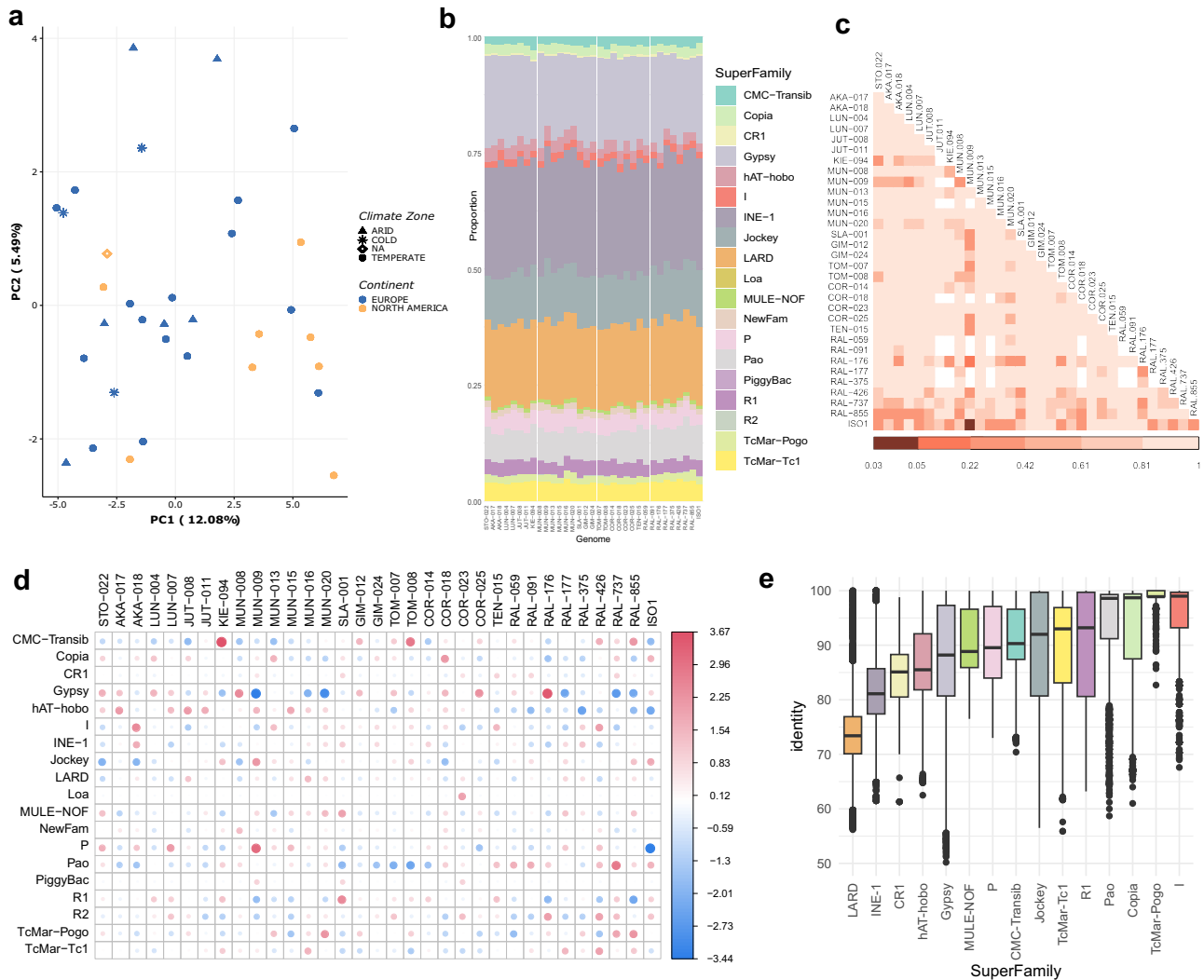
**TE content is similar across *D. melanogaster* strains while TE activity varies.** When comparing TE annotations for the 32 genomes plus the reference genome (ISO1), we observed low variation among strains regarding both TE content (percentage of the euchromatic genome occupied by TEs, average = 3.56%, SD = 0.3%) and number of TE copies (average = 2016, SD = 69.6) (Supplementary Data 9a). 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.<sup>8</sup>). As previously described, TE variation across populations did not reflect the geographical or environmental origin of the populations<sup>30</sup> (Fig. 4a; see Methods).

At the TE order level, and in agreement with previous studies<sup>30</sup>, we found *LTRs* to be the most abundant, representing near 60% of all TE content (Supplementary Data 9b, Supplementary Fig. 6a), while the number of TE copies was more evenly distributed among the five main orders (*Helitrons*, *LARDs*, *LINEs*, *LTRs* and *TIRs*)

(Supplementary Data 9b, Supplementary Fig. 6b). Also in agreement with previous observations, *INE-1* superfamily showed the largest number of copies among Class II DNA elements<sup>47</sup> and *Gypsy* and *Pao* elements were the most abundant among the *LTRs*<sup>30,48</sup> (Supplementary Data 9c). 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, Fig. 4b, Supplementary Data 9c), genome pairwise comparisons were significant for the MUN-009 and ISO1 pair of strains ( $X^2$  test, adjusted  $p$ -value = 0.03, Fig. 4c), mainly due to the *P* superfamily overrepresentation in MUN-009 compared with the ISO1 genome (Fig. 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<sup>49</sup> (Supplementary Fig. 7 and Supplementary Data 9d). *P-elements* were indeed among the most variable families in the 33 genomes (Supplementary Fig. 8, Supplementary Data 9d).

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<sup>50,51</sup>, 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<sup>30,52</sup> (Fig. 4e and Supplementary Fig. 9). Moreover, some superfamilies showed a large variability in identity such as *RI*, *Jockey* and *Gypsy*, indicating that they contain both young and old members (Fig. 4e and Supplementary Fig. 9). Genome pairwise comparisons in the distribution of identity values per genome showed significant differences between some pairs of genomes (Supplementary Fig. 10a). Notably, such differences seem to be mainly caused by members of the *Jockey* and *Gypsy* superfamilies (Supplementary Fig. 10b).

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<sup>30,46</sup>. Moreover, while almost no significant differences were observed between genomes in the number of TE copies (Fig. 4c), we did find pairwise differences in the identity of the copies (Supplementary Fig. 10a), particularly among members of two superfamilies, *Jockey* and *Gypsy* (Fig. 4e; Supplementary Fig. 10b), suggesting a population specific behavior regarding TE activity as previously described in both European<sup>30</sup> and North American strains<sup>53</sup>.



**Fig. 4 TE annotations at the superfamily level.** **a** Principal component analysis based on TE insertions polymorphisms grouped by continent (colors) 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 colors of the matrix squares represent adjusted (FDR) *p*-values of the two-sided 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. The boxplot shows median (the horizontal line in the box), 1st and 3rd quartiles (lower and upper bounds of box, respectively), minimum and maximum (lower and upper whiskers, respectively). Number of copies analyzed per superfamily are given in Supplementary Data 9c.

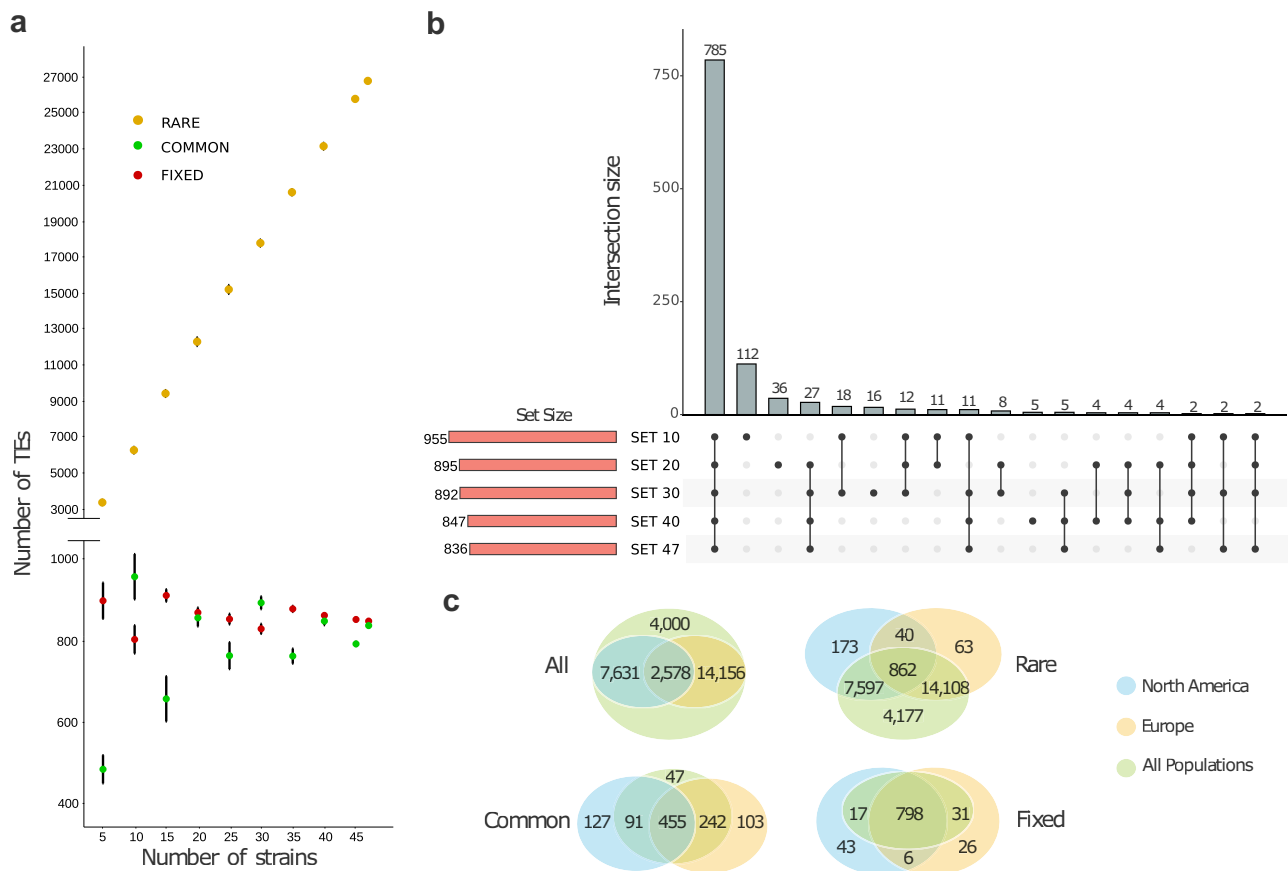
**20 genomes allow the identification of the vast majority of TEs that are common in out-of-Africa natural populations.**

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.<sup>8</sup> collected in Africa (2), Europe (2), North America (4), North Atlantic Ocean (1), South America (2), and Asia (3) (Supplementary Data 10 and 11). On average, 2016 euchromatic TE copies were annotated per genome (ranging from 1883 to 2178, Supplementary Data 9a), and for 97% of them (on average) orthologous relationships of the insertion flanking regions in the ISO1 reference genome were determined (Supplementary Data 11a; Supplementary Note 11). Overall, we annotated 28,947 TEs across the 47 genomes (Supplementary Data 10). As

expected, the site frequency spectrum of TE insertions showed an excess of rare variants compared with SNP variants<sup>54</sup> (Supplementary Fig. 11).

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 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 (Fig. 5a and Supplementary Data 11b). 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 (Fig. 5a). Finally, we observed that the number of





**Fig. 5 TE classification according to three frequency classes: rare (present in <10 of the strains), common (present in  $\geq 10$  and  $\leq 95\%$  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. Data are presented as median values  $\pm$  standard deviation. **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.**

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 (Fig. 5b). By increasing the number of genomes analyzed from 10 to 20, the number TEs identified as common decreased (Fig. 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 (Fig. 5b). These results suggest that 20 genomes are enough to accurately identify most common TEs in populations, which is the subset of TEs expected to be enriched for candidate adaptive mutations<sup>33</sup>.

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 (Supplementary Data 11a). Most of the TE insertions were only identified in either Europe or North America (Fig. 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 (Fig. 5c; Supplementary Data 11c). 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 (Supplementary Data 11d). 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<sup>55–57</sup> (Table 1, Supplementary Data 2c). 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; Supplementary Data 12a). Also, most of them (433 out of 472; 91.7%) were present at low frequencies

in populations ( $\leq 5\%$ ) 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  $< 0.05$ , Supplementary Data 12b). Genes located nearby these TEs were not significantly overrepresented for any biological process, molecular function or cellular component nor any metabolic pathways<sup>58,59</sup>. Contrary to previous results, we found a similar number of low frequent TEs associated with gene up- and down-regulation<sup>54</sup> (214 vs 258, respectively; Supplementary Data 12a; *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 (Supplementary Data 12c–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 (Fig. 6 and Table 2; see 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<sup>60</sup> and it is located within a genomic region harboring several insecticide resistance genes from the *cyp* family<sup>61</sup>. 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)<sup>62</sup>. Furthermore, other genes may be important for their role in neurogenesis (*pde9*, *ppk*<sup>63</sup>) and synaptic organization (*Ten-a*, *dpr8*<sup>64</sup>, Table 2).

**Most of the insertions with signatures of selection in their flanking regions were de novo annotated 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*<sup>65</sup>, *iHH12*<sup>66,67</sup> and *nSL*<sup>68</sup>. We defined a SNP to have a significant *iHS*, *iHH12* or *nSL* values when, after normalizing by frequency and chromosome location, the normalized values were  $>95^{\text{th}}$  percentile of the distribution of values for SNPs falling in neutral introns (see Methods). We then looked for candidate adaptive TE insertions in linkage disequilibrium with each significant SNP, and located  $< 1$  kb from the significant SNP (see Methods). We considered as candidate adaptive TEs those present at high population frequency and located in regions with recombination rates  $> 0$  (see Methods and Rech et al.<sup>33</sup>). Among the 746 candidate adaptive TEs, we found 19 TEs co-occurring with SNPs showing evidence of selective sweeps (Supplementary Data 13a). Among these 19 TE insertions, two correspond to an *Accord* element inserted in the *Cyp6g1* gene that is duplicated in some genomes (Supplementary Data 14). 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<sup>69,70</sup>. We discarded the presence of other structural variants linked to our 18 candidate adaptive TEs that could also be driving positive selection (Table 3 and Supplementary Data 14). 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

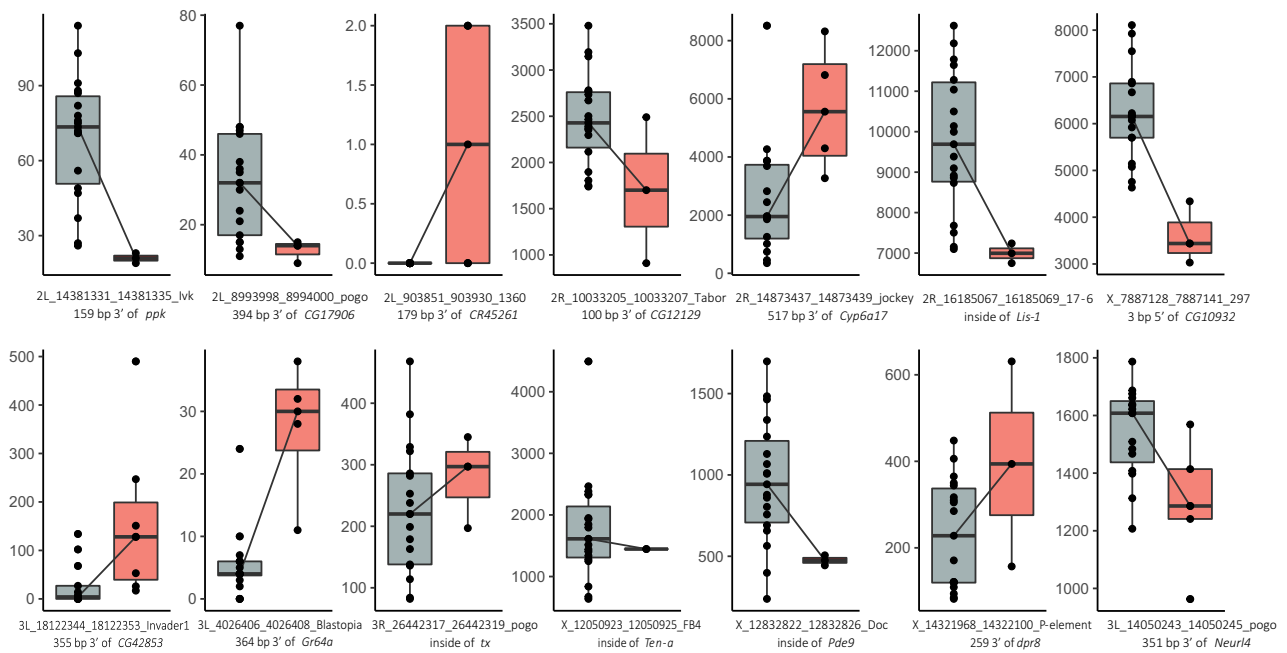
insertions as exemplified by the *Accord* insertion<sup>69,70</sup>. However, further functional validation is needed before arriving at any conclusive evidence on the functional role of these TEs. Note that for one of these 18 insertions, we found significant association with the level of expression of the nearby gene in whole-body non-stress conditions (Fig. 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<sup>33</sup>, five other insertions recently described by Bogaerts-Márquez et al.<sup>71</sup>, and the 18 TEs identified in this work, including the previously described *Accord* insertion (107 insertions in total). Biological process GO term analysis identified clusters enriched for response to stimulus, behavior, and development and morphogenesis as the ones showing the highest enrichment scores (Fig. 7, Supplementary Data 15). Pigmentation was also among the significant clusters, as has been previously described (Rech et al.<sup>33</sup>). 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 (Fig. 7 and Supplementary Data 15).

The 107 candidate adaptive TEs identified so far in *D. melanogaster* (Supplementary Data 16a) 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 (Supplementary Data 16b). Finally, regarding gene body location, we found that the subset of candidate adaptive TEs was slightly enriched for TEs inserted in 5'UTR and promoters, although the differences were not statistically significant (Supplementary Data 16c).

## 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<sup>8,31,32</sup>. 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 (Fig. 1 and Supplementary Data 1), 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 (Fig. 2), but also allows the identification of new TE families (Fig. 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<sup>72</sup>. Moreover, the availability of even longer reads together with the improvement of computational analysis



**Fig. 6 Gene expression levels in strains with and without TE insertions.** Gene expression levels in strains without (gray) 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. In total, the expression levels of 20 strains are plotted. The boxplot shows median (the horizontal line in the box), 1st and 3rd quartiles (lower and upper bounds of box, respectively), minimum and maximum (lower and upper whiskers, respectively).

**Table 2 TEs showing the highest significance values in their association with the expression of a nearby gene (adjusted *p*-value  $\leq 0.01$ , defined by an approximation method based on the beta distribution using QTLtools).**

TE ID	Freq.	Gene symbol	Gene expression	Biological process
2L_903851_903930_1360	0.30	<i>CR45261</i>	Up	-
2L_8993998_8994000_pogo	0.15	<i>CG17906</i>	Down	-
2L_14381331_14381335_lvk	0.10	<i>ppk</i>	Down	Behavior, Response to stimulus
2R_10033205_10033207_Tabor	0.10	<i>CG12129</i>	Down	-
2R_14873437_14873439_jockey	0.20	<i>Cyp6a17</i>	Up	Response to stimulus, Behavior (thermosensory)
2R_16185067_16185069_17-6	0.10	<i>Lis-1</i>	Down	Development, Reproduction, Transport/localization, Cell organization/biogenesis, cell cycle/proliferation, Response to stimulus
3L_4026406_4026408_Blastopia	0.20	<i>Gr64a</i>	Up	Response to stimulus, Nervous system process
3L_18122344_18122353_Invader1	0.35	<i>CG42853</i>	Up	-
3R_26442317_26442319_pogo	0.15	<i>tx</i>	Down	Development, Gene expression
X_7887128_7887141_297	0.15	<i>CG10932</i>	Down	Small molecule metabolism
X_12832822_12832826_Doc	0.10	<i>Pde9</i>	Down	Response to stimulus, Signaling
X_14321968_14322100_P-element	0.10	<i>dpr8</i>	Up	Nervous system process, Cell organization/biogenesis
X_12050923_12050925_FB4	0.05	<i>Ten-a</i>	Down	Development, Cell organization/biogenesis, Response to stimulus
X_12050923_12050925_FB4.t1	0.05	<i>Ten-a</i>	Down	Development, Cell organization/biogenesis, Response to stimulus

Note that for *Ten-a* gene there were two TEs with equal nominal *p*-value.

should help to characterize nested and highly complex variation in the near future<sup>72</sup>.

Improving the annotation of TEs in genome sequences is the first necessary step to accurately evaluate the role of this abundant active component in genome function and evolution. We identified 472 TEs associated with nearby gene expression variation (Fig. 6 and Table 2 and Supplementary Data 12). 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<sup>73,74</sup>. 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 (Fig. 5). We identified 746 TE insertions present at high population frequencies ( $\geq 10\%$  and  $\leq 95\%$ ) in genomic regions with recombination rates  $>0$ . Eighteen of these

**Table 3 Eighteen candidate adaptive TE insertions showing evidence of selection identified in this work.**

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%	<i>sprt</i>	Intron	NA
2R_12185376_12185380_accord	nSL	62%	<i>Cyp6g1</i>	Promoter	response to insecticide
2R_14078395_14078397_hopper	nSL	11%	<i>Prosap</i>	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
			<i>Neur14</i>	Promoter	NA
3L_2426710_2426713_pogo	nSL	19%	<i>Svil</i>	Intron	NA
3L_3798612_3798621_1360	nSL	30%	CG32264	Intron	NA
3R_20502048_20502058_Doc	nSL	28%	<i>Dic2</i>	Promoter	NA
			CG46441	Promoter	NA
3R_21385503_21385506_pogo	nSL	19%	-	Intergenic	-
3R_29952746_29952748_Invader4	nSL	23%	<i>Tkr99D</i>	Intron	olfactory behavior; detection of chemical stimulus
X_15012530_15012533_mdg3	nSL	60%	<i>hiw</i>	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%	<i>Tbh</i>	3UTR	aggressive behavior; behavioral response to ethanol; flight behavior; learning; ovulation
3L_18931204_18931207_F-element	nSL	15%	CG32204	Intron	NA

Biological process information according to FlyBase.

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<sup>33,69,70</sup> (Table 3). 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<sup>33</sup> (Table 3 and Fig. 7).

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 genome components.

## 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 (Supplementary Data 1). 24 strains were obtained from 11 European natural populations and the remaining eight are RAL strains from the DGRP, obtained from North Carolina, US (Fig. 1, Supplementary Data 1). 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 *D. melanogaster* 5–10 day-old female flies, using the Genra Puregene Tissue Kit (Qiagen) following manufacturer's instructions. Briefly, 400 flies from each strain were mechanically homogenized in 24 ml of lysis

buffer (proteinase K added) and incubated overnight at 55 °C, and DNA was precipitated with isopropanol after RNase treatment and protein precipitation. Finally, DNA was resuspended in 1.6 ml of Hydration Solution. DNA concentration was measured using a Nanodrop® spectrophotometer. Most DNA samples for ONT sequencing were extracted from 100 *D. melanogaster* 5–10 day-old female flies from each strain using the Blood and Cell Culture DNA Mini Kit (Qiagen) following manufacturer's instructions with small modifications (Supplementary Data 2; Supplementary Note 1).

PacBio libraries were prepared using 20 Kb SMRTbell and were sequenced using the PacBio RSII System by Macrogen Inc. Korea. ONT libraries were constructed using the Ligation Sequencing Kit (SQK-LSK108 or SQK-LSK109) following manufacturer's instructions (Supplementary Data 2; Supplementary Note 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)<sup>75</sup>.

**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, 350 bp insert libraries, 150 bp pair-end sequencing) or by the Genomics Unit of the Center for Genomic Regulation (gdNA-PCR free, HiSeq 2500, 125 bp pair-end) (Supplementary Data 2c).

**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)<sup>76</sup> for building draft genome assemblies followed by *FinisherSC* (v.2.1)<sup>77</sup> for improving contig continuity. We then aligned PacBio reads to the draft assembly using *pbalgn* (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 (Supplementary Fig. 12a). For ONT genomes, we also started with *Canu* (v.1.7)<sup>76</sup> with default options for building raw de novo assemblies. We then applied *Racon* (v.1.0)<sup>78</sup>, *Nanopolish* (v.0.10.1) (<https://github.com/jts/nanopolish>) and *Pilon* (v.1.22)<sup>79</sup> for obtaining final polished assemblies (Supplementary Fig. 12b, Supplementary Note 2).

**Genome deduplication, decontamination and scaffolding.** Besides repetitive content, we found that raw de novo genome assembly sizes positively correlated with BUSCO Duplicates (Supplementary Note 3, Supplementary Figs. 13–15). Thus, 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)<sup>80</sup> best practices for variant discovery<sup>81</sup>. Then, we used the *bctools stats* (v.1.9)<sup>82</sup> 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 Note 3, Supplementary Fig. 16). Genomes showing levels of heterozygosity >0.2 were deduplicated



**Fig. 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. 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 (Fisher test (two sided) 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 colors indicate similar biological functions as specified at the bottom of the figure.

(removing alleles-contigs- present twice in the genome) using *purge\_haplotigs* (v.1.0.1)<sup>83</sup>; Supplementary Fig. 17, Supplementary Data 3, Supplementary Note 3).

After deduplication, we evaluated contigs for putative contaminations using *MUMmer* (v.4.0)<sup>84</sup>. Briefly, we attempted to align all contigs to the *D. melanogaster* hologenome<sup>85</sup> plus the *D. simulans* genome. We considered as putative contaminant, those 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) (Supplementary Data 3). Once we removed the putative contaminant contigs, we performed a reference-guided scaffolding of the contigs using *RaGOO* (v.1.02)<sup>86</sup>, which uses *minimap2* (v.2.9)<sup>87</sup> 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)<sup>84</sup>; Supplementary Data 3).

**Assembly quality.** Quality of the assemblies was evaluated by estimating completeness, accuracy and continuity. Completeness and accuracy were calculated using *BUSCO* (v.3.0.2)<sup>88</sup> for the Diptera lineage (diptera\_odb9), consisting on 2799 genes. Continuity and completeness were estimated by aligning the polished genome assemblies to the *Drosophila melanogaster* strain ISO1 reference genome release 6<sup>89</sup>. We first masked simple repeats in both genomes using *RepeatMasker* (v.3.0) ([www.repeatmasker.org](http://www.repeatmasker.org)) and then used *MUMmer* (v.3.0)<sup>90</sup> 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.<sup>36</sup> based on the flanking sequences for 85 out of the 142 annotated piRNA clusters of *D. melanogaster*<sup>91</sup> Supplementary Data 3b, Supplementary Note 5). QV scores were estimated according to Solares et al.<sup>18</sup> 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<sup>92</sup>. 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<sup>18</sup> using our pipeline (Supplementary Fig. 12b). We annotated TE copies using the MCTE library and we identified 1842 orthologous TEs comparing with the ISO1 reference genome TE annotation, which represents >83% of the TEs annotated in Solares et al.<sup>18</sup> 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)<sup>38,40,41</sup> 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 consensus sequences consisted in three main procedures: removal of redundant sequences, the manual identification of potentially artifactual sequences, and the classification of consensus sequences into families (Supplementary Note 6). Redundant sequences (consensus sequences present in more than one genome) were removed by first running *PASTE*C (v2.0) with default options<sup>41</sup>. 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 Note 6). We manually explored the consensus sequences and their copies using the *plotCoverage* tool from *REPET* and discarded consensus sequences showing mainly a high number of small copies. The assignment of the consensus sequences into families was performed using *BLAT* (v.35)<sup>93</sup> 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](https://fruitfly.org/p_disrupt/TE.html)). When no matches were found, we used *RepeatMasker* (v.4)<sup>94</sup> with the release *RepBaseRepeatMaskerEdition-20181026* of the *RepBase*<sup>95</sup> Supplementary Note 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<sup>96</sup> and its assembly and annotation usually require specific methods and extensive curation<sup>8,97</sup>. In this work, we determined the euchromatic regions using the recombination rate calculator (RRC)<sup>98</sup> available at [http://petrov.stanford.edu/cgi-bin/recombination-rates\\_updateR5.pl](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<sup>28</sup>, 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)<sup>90</sup>. 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 <100 bp, 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)<sup>45</sup> and *TIDAL* (v.1.0)<sup>46</sup>. 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)<sup>99</sup> to find the overlapping TE 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) 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)<sup>87</sup> mapping of the TE sequence and its flanking regions to the ISO1 genome and the coordinates of genes as anchored synteny sequences (see Supplementary Note 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 Note 11, Supplementary Figs. 18–20). To avoid false positives, we only used those TEs for which more than half of the orthologous TEs were larger than 120 bp. All scripts used for the TE transfer are available at [www.github.com/sradiouy/deNovoTEsDmel](http://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)<sup>100</sup> and also established different sets of TEs based on the geographical origin of the genomes and compare them using *VennDiagram* (v.1.6)<sup>101</sup>. For determining the location of the TE insertion regarding annotated genes, we used *annotatr* (R package version 1.20.0).

**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<sup>56,57</sup> and 8 American strains<sup>55</sup> (Table 1, Supplementary Data 2c). Briefly, RNA-Seq data was trimmed using the *fastp* package (v.0.20)<sup>102</sup> with default parameters. Expression levels were quantified by applying the *salmon* package (v.1.0.0)<sup>103</sup> 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)<sup>104</sup>. eQTL analysis was performed using the *QTLtools* package (v.1.2)<sup>105</sup> taking into account the population structure (Supplementary Figs. 21 and 22). Putative cis-eQTL were searched within a 1 Kb 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 4709 genes and 9676 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).

**Positive selection analysis.** We looked for evidences of selection in genomic regions targeted by TE insertions using *selscan* (v.1.2.0a)<sup>106</sup> and Single Nucleotide Polymorphisms (SNPs) as a proxy (Supplementary Note 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.<sup>8</sup>). SNPs were called using the *GATK* (v.4.0)<sup>80</sup> *HaplotypeCaller* best practices for variant discovery<sup>81</sup> and the haplotype phasing was performed using *SHAPEIT4* (v.4.1)<sup>107</sup>. 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 <https://doi.org/10.20350/digitalCSIC/13708>). Genetic positions and recombination maps<sup>108</sup> were obtained from FlyBase (<https://wiki.flybase.org/wiki/FlyBase:Maps>, last updated June 15, 2016). Three statistics were calculated in *selscan*: iHS<sup>65</sup>, iHH12<sup>66,67</sup> and nSL<sup>68</sup>. 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<sup>106</sup>. 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 95<sup>th</sup> percentile of the distribution of normalized values for SNPs falling within the first 8–30 base pairs of small introns (≤65 bp) which are considered to be neutrally evolving<sup>109</sup> (Supplementary Data 13b). 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 (<1 Kb). 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<sup>33</sup>, 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<sup>33</sup>. 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 established 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 (Supplementary Data 13a).

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 1 kb regions flanking these insertions (Supplementary Data 14)<sup>8</sup>.

**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 1 Kb region 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/](http://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.31_FB2019_06/fasta/)); (v) only the closest gene (<1 Kb) 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 <1 Kb, 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)<sup>110,111</sup>

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<sup>112,113</sup> 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.

**Reporting summary.** Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

All scaffolded assemblies and the raw data (long and short read sequencing) have been deposited in NCBI database under the BioProject accession [PRJNA559813](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA559813). The VCF file containing SNP callings for 46 *D. melanogaster* genomes used for testing positive selection evidences is available at DIGITAL.CSIC repository (<https://doi.org/10.20350/digitalCSIC/13708>) Fasta sequences for the *D. melanogaster* Manually Curated Transposable Elements (MCTE) library are available at DIGITAL.CSIC repository (<https://doi.org/10.20350/digitalCSIC/13765>). The new consensus sequences are deposited in Dfam (Storer et al.<sup>114</sup>). Recombination rates according to Fiston-Lavier et al.<sup>98</sup> and Comeron et al.<sup>108</sup> for *D. melanogaster* genome release 6 are available at DIGITAL.CSIC repository (<https://doi.org/10.20350/digitalCSIC/13766>). BED files containing Transposable Element (TE) annotations for 47 *Drosophila melanogaster* genomes are available at DIGITAL.CSIC repository (<https://doi.org/10.20350/digitalCSIC/13894>).

## Code availability

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

Received: 22 March 2021; Accepted: 15 March 2022;

Published online: 12 April 2022

## References

- De Coster, W. & Van Broeckhoven, C. Newest methods for detecting structural variations. *Trends Biotechnol.* **37**, 973–982 (2019).
- Huddleston, J. & Eichler, E. E. An incomplete understanding of human genetic variation. *Genetics* **202**, 1251–1254 (2016).
- Audano, P. A. et al. Characterizing the major structural variant alleles of the human genome. *Cell* **176**, 663–675.e619 (2019).
- Chaisson, M. J. P. et al. Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.* **10**, 1784 (2019).
- Mahmoud, M., Gobet, N., Cruz-Dávalos, D. I., Mounier, N., Dessimoz, C. & Sedlazeck, F. J. Structural variant calling: The long and the short of it. *Genome Biol.* **20**, 246 (2019).
- Zhou, Y. et al. The population genetics of structural variants in grapevine domestication. *Nat. Plants* **5**, 965–979 (2019).
- Kou, Y. et al. Evolutionary genomics of structural variation in asian rice (*oryza sativa*) domestication. *Mol. Biol. Evol.* **37**, 3507–3524 (2020).
- Chakraborty, M., Emerson, J. J., Macdonald, S. J. & Long, A. D. Structural variants exhibit widespread allelic heterogeneity and shape variation in complex traits. *Nat. Commun.* **10**, 4872 (2019).
- Chakraborty, M. et al. Hidden genetic variation shapes the structure of functional elements in drosophila. *Nat. Genet.* **50**, 20–25 (2018).
- Yang, X., Lee, W.-P., Ye, K. & Lee, C. One reference genome is not enough. *Genome Biol.* **20**, 104–104 (2019).
- Ballouz, S., Dobin, A. & Gillis, J. A. Is it time to change the reference genome? *Genome Biol.* **20**, 159 (2019).
- Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2011).
- Goerner-Potvin, P. & Bourque, G. Computational tools to unmask transposable elements. *Nat. Rev. Genet.* **19**, 688–704 (2018).
- Barron, M. G., Fiston-Lavier, A. S., Petrov, D. A. & Gonzalez, J. Population genomics of transposable elements in drosophila. *Annu Rev. Genet.* **48**, 561–581 (2014).
- Du, H. & Liang, C. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun.* **10**, 5360 (2019).
- Jain, M. et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat. Biotechnol.* **36**, 338–345 (2018).

17. Miga, K. H. et al. Telomere-to-telomere assembly of a complete human x chromosome. *Nature* **585**, 79–84 (2020).
18. Solares, E. A. et al. Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3* **8**, 3143–3154 (2018).
19. Chaisson, M. J. P. et al. Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
20. Jiao, Y. et al. Improved maize reference genome with single-molecule technologies. *Nature* **546**, 524–527 (2017).
21. Mitsuhashi, S. & Matsumoto, N. Long-read sequencing for rare human genetic diseases. *J. Hum. Genet.* **65**, 11–19 (2020).
22. Sakamoto, Y., Sereewattanawoot, S. & Suzuki, A. A new era of long-read sequencing for cancer genomics. *J. Hum. Genet.* **65**, 3–10 (2020).
23. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e113 (2020).
24. Alonge, M. et al. Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* **182**, 145–161.e123 (2020).
25. Levy-Sakin, M. et al. Genome maps across 26 human populations reveal population-specific patterns of structural variation. *Nat. Commun.* **10**, 1025 (2019).
26. Shahid, S. & Slotkin, R. K. The current revolution in transposable element biology enabled by long reads. *Curr. Opin. Plant Biol.* **54**, 49–56 (2020).
27. Michael, T. P. et al. High contiguity arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nat. Commun.* **9**, 541 (2018).
28. Gramates, L. S. et al. Flybase at 25: looking to the future. *Nucleic Acids Res.* **45**, D663–D671 (2017).
29. Thurmond, J. et al. Flybase 2.0: the next generation. *Nucleic Acids Res.* **47**, D759–D765 (2018).
30. Lerat, E. et al. Population-specific dynamics and selection patterns of transposable element insertions in european natural populations. *Mol. Ecol.* **28**, 1506–1522 (2019).
31. Mohamed, M. et al. A transposon story: from TE content to TE dynamic invasion of drosophila genomes using the single-molecule sequencing technology from oxford nanopore. *Cells* **9**, 1776 (2020).
32. Ellison, C. E. & Cao, W. 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 Res.* **48**, 290–303 (2019).
33. Rech, G. E. et al. Stress response, behavior, and development are shaped by transposable element-induced mutations in drosophila. *PLoS Genet.* **15**, e1007900 (2019).
34. Huang, W. et al. Natural variation in genome architecture among 205 drosophila melanogaster genetic reference panel lines. *Genome Res.* **24**, 1193–1208 (2014).
35. Miller, D. E., Staber, C., Zeitlinger, J. & Hawley, R. S. Highly contiguous genome assemblies 15 Drosoph. species generated using nanopore sequencing. *G3* **8**, 3131–3141 (2018).
36. Wierzbicki, F., Schwarz, F., Cannalunga, O. & Kofler, R. Novel quality metrics allow identifying and generating high-quality assemblies of piRNA clusters. *Mol. Ecol. Res.* **22**, 102–121 (2022).
37. Berlin, K. et al. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat. Biotechnol.* **33**, 623–630 (2015).
38. Flutur, T., Duprat, E., Feuillet, C. & Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS One* **6**, e16526 (2011).
39. Kaminker, J. S. et al. The transposable elements of the drosophila melanogaster euchromatin: a genomics perspective. *Genome Biol.* **3**, Research0084 (2002).
40. Quesneville, H. et al. Combined evidence annotation of transposable elements in genome sequences. *PLoS Computational Biol.* **1**, e22 (2005).
41. Hoede, C. et al. Pasted: an automatic transposable element classification tool. *PLoS One* **9**, e91929 (2014).
42. Palazzo, A., Lovero, D., D’Addabbo, P., Caizzi, R. & Marsano, R. M. Identification of bari transposons in 23 sequenced drosophila genomes reveals novel structural variants, mites and horizontal transfer. *PLoS One* **11**, e0156014 (2016).
43. Wallau, G. L., Capy, P., Loreto, E. & Hua-Van, A. Genomic landscape and evolutionary dynamics of mariner transposable elements within the drosophila genus. *BMC Genom.* **15**, 727 (2014).
44. Kojima, K. K. & Jurka, J. Crypton transposons: identification of new diverse families and ancient domestication events. *Mob. DNA* **2**, 12 (2011).
45. Zhuang, J., Wang, J., Theurkauf, W. & Weng, Z. Temp: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.* **42**, 6826–6838 (2014).
46. Rahman, R. et al. Unique transposon landscapes are pervasive across drosophila melanogaster genomes. *Nucleic Acids Res.* **43**, 10655–10672 (2015).
47. Thomas, J., Vадnagara, K. & Pritham, E. J. Dine-1, the highest copy number repeats in drosophila melanogaster are non-autonomous endonuclease-encoding rolling-circle transposable elements (helentrons). *Mob. DNA* **5**, 18 (2014).
48. Linheiro, R. S. & Bergman, C. M. Whole genome resequencing reveals natural target site preferences of transposable elements in drosophila melanogaster. *PLoS One* **7**, e30008 (2012).
49. Anxolabéhère, D., Kidwell, M. G. & Periquet, G. 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 (1988).
50. Kapitonov, V. V. & Jurka, J. Molecular paleontology of transposable elements in the drosophila melanogaster genome. *Proc. Natl Acad. Sci. USA* **100**, 6569–6574 (2003).
51. Kalendar, R. et al. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **166**, 1437 (2004).
52. Bucheton, A., et al. I elements and the drosophila genome. *Genetica* **86**, 175–190 (1992).
53. Adrion, J. R., Song, M. J., Schrider, D. R., Hahn, M. W. & Schaack, S. Genome-wide estimates of transposable element insertion and deletion rates in drosophila melanogaster. *Genome Biol. Evol.* **9**, 1329–1340 (2017).
54. Cridland, J. M., Macdonald, S. J., Long, A. D. & Thornton, K. R. Abundance and distribution of transposable elements in two drosophila qtl mapping resources. *Mol. Biol. Evol.* **30**, 2311–2327 (2013).
55. Everett, L. J. et al. Gene expression networks in the drosophila genetic reference panel. *Genome Res.* **30**, 485–496 (2020).
56. Green, L., Radio, S., Rech, G. E., Salces-Ortiz, J. & González, J. Natural variation in copper tolerance in *Drosophila melanogaster* is shaped by transcriptional and physiological changes in the gut. Preprint at <https://www.biorxiv.org/content/10.1101/2021.07.12.452058v1> (2021).
57. Horváth, V. et al. Basal and stress-induced expression changes consistent with water loss reduction explain desiccation tolerance of natural drosophila melanogaster populations. Preprint at <https://www.biorxiv.org/content/10.1101/2022.03.21.485105v1> (2022).
58. Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. Panther version 14: more genomes, a new panther go-slim and improvements in enrichment analysis tools. *Nucleic Acids Res.* **47**, D419–D426 (2018).
59. Jassal, B. et al. The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–d503 (2020).
60. Kang, J., Kim, J. & Choi, K.-W. Novel cytochrome p450, cyp6a17, is required for temperature preference behavior in drosophila. *PLoS One* **6**, e29800 (2011).
61. Carareto, C. M., Hernandez, E. H. & Vieira, C. 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 (2014).
62. Jiao, Y., Moon, S. J. & Montell, C. A drosophila gustatory receptor required for the responses to sucrose, glucose, and maltose identified by mrna tagging. *Proc. Natl Acad. Sci. USA* **104**, 14110–14115 (2007).
63. Day, J. P., Dow, J. A., Houslay, M. D. & Davies, S. A. Cyclic nucleotide phosphodiesterases in drosophila melanogaster. *Biochemical J.* **388**, 333–342 (2005).
64. Cheng, S. et al. Molecular basis of synaptic specificity by immunoglobulin superfamily receptors in drosophila. *Elife* **8**, e41028 (2019).
65. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
66. Garud, N. R., Messer, P. W., Buzbas, E. O. & Petrov, D. A. Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLoS Genet.* **11**, e1005004 (2015).
67. Torres, R., Szpiech, Z. A. & Hernandez, R. D. Human demographic history has amplified the effects of background selection across the genome. *PLoS Genet.* **14**, e1007387 (2018).
68. Ferrer-Admetlla, A., Liang, M., Korneliusen, T. & Nielsen, R. On detecting incomplete soft or hard selective sweeps using haplotype structure. *Mol. Biol. Evol.* **31**, 1275–1291 (2014).
69. Daborn, P. J. et al. A single p450 allele associated with insecticide resistance in drosophila. *Science* **297**, 2253–2256 (2002).
70. Schmidt, J. M. et al. Copy number variation and transposable elements feature in recent, ongoing adaptation at the cyp6g1 locus. *PLoS Genet* **6**, e1000998 (2010).
71. Bogaerts-Márquez, M., Guirao-Rico, S., Gautier, M. & González, J. Temperature, rainfall and wind variables underlie environmental adaptation in natural populations of drosophila melanogaster. *Mol. Ecol.* **30**, 938–954 (2021).
72. De Coster, W., Weissensteiner, M. H. & Sedlazeck, F. J. Towards population-scale long-read sequencing. *Nat. Rev. Genet.* **22**, 572–587 (2021).
73. Cridland, J. M., Thornton, K. R. & Long, A. D. Gene expression variation in drosophila melanogaster due to rare transposable element insertion alleles of large effect. *Genetics* **199**, 85–93 (2015).
74. Ullastres, A., Merenciano, M. & González, J. Regulatory regions in natural transposable element insertions drive interindividual differences in response to immune challenges in drosophila. *Genome Biol.* **22**, 265 (2021).
75. De Coster, W., D’Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. Nanopack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**, 2666–2669 (2018).



76. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
77. Lam, K.-K., LaButti, K., Khalak, A. & Tse, D. Finishersc: a repeat-aware tool for upgrading de novo assembly using long reads. *Bioinformatics* **31**, 3207–3209 (2015).
78. Vaser, R., Sovic, I., Nagarajan, N. & Sikic, M. Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res.* **27**, 737–746 (2017).
79. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963 (2014).
80. McKenna, A. et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
81. Van der Auwera, G. A. et al. From fastq data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr. Protoc. Bioinforma.* **43**, 11.10.11–33 (2013).
82. Li, H. A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
83. Roach, M. J., Schmidt, S. A. & Borneman, A. R. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinform.* **19**, 460 (2018).
84. Marçais, G. et al. Mummer4: a fast and versatile genome alignment system. *PLoS Computational Biol.* **14**, e1005944 (2018).
85. Kapun, M., et al. Genomic analysis of european drosophila melanogaster populations reveals longitudinal structure, continent-wide selection, and previously unknown DNA viruses. *Mol. Biol. Evol.* **37**, 2661–2678 (2020).
86. Alonge, M. et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. *Genome Biol.* **20**, 224 (2019).
87. Li, H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).
88. Waterhouse, R. M. et al. Busco applications from quality assessments to gene prediction and phylogenomics. *Mol. Biol. Evol.*, **35**, 543–548(2017).
89. Hoskins, R. A. et al. The release 6 reference sequence of the drosophila melanogaster genome. *Genome Res.* **25**, 445–458 (2015).
90. Kurtz, S. et al. Versatile and open software for comparing large genomes. *Genome Biol.* **5**, R12 (2004).
91. Brenneke, J. et al. Discrete small rna-generating loci as master regulators of transposon activity in drosophila. *Cell* **128**, 1089–1103 (2007).
92. Wick, R. R., Judd, L. M. & Holt, K. E. Performance of neural network basecalling tools for oxford nanopore sequencing. *Genome Biol.* **20**, 129 (2019).
93. Kent, W. J. Blat-the blast-like alignment tool. *Genome Res.* **12**, 656–664 (2002).
94. Smit, A. F. A., Hubley, R & Green, P. RepeatMasker open-4.0. <http://www.repeatmasker.org> (2015).
95. Bao, W., Kojima, K. K. & Kohany, O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 11 (2015).
96. Smith, C. D., Shu, S., Mungall, C. J. & Karpen, G. H. The release 5.1 annotation of drosophila melanogaster heterochromatin. *Science* **316**, 1586–1591 (2007).
97. Khost, D. E., Eickbush, D. G. & Larracuente, A. M. Single-molecule sequencing resolves the detailed structure of complex satellite DNA loci in drosophila melanogaster. *Genome Res.* **27**, 709–721 (2017).
98. Fiston-Lavier, A. S., Singh, N. D., Lipatov, M. & Petrov, D. A. Drosophila melanogaster recombination rate calculator. *Gene* **463**, 18–20 (2010).
99. Quinlan, A. R. & Hall, I. M. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
100. Conway, J. R., Lex, A. & Gehlenborg, N. Upsetr: an r package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
101. Chen, H. & Boutros, P. C. VennDiagram: a package for the generation of highly-customizable venn and euler diagrams in r. *BMC Bioinforma.* **12**, 35 (2011).
102. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
103. Patro, R., Duggal, G., Love, M. I., Izirary, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. methods* **14**, 417–419 (2017).
104. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol.* **15**, 550–550 (2014).
105. Delaneau, O. et al. A complete tool set for molecular qtl discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
106. Szpiech, Z. A. & Hernandez, R. D. Selscan: an efficient multithreaded program to perform ehh-based scans for positive selection. *Mol. Biol. Evolution* **31**, 2824–2827 (2014).
107. Delaneau, O., Zagury, J.-F., Robinson, M. R., Marchini, J. L. & Dermitzakis, E. T. Accurate, scalable and integrative haplotype estimation. *Nat. Commun.* **10**, 5436 (2019).
108. Comeron, J. M., Ratnappan, R. & Bailin, S. The many landscapes of recombination in drosophila melanogaster. *PLoS Genet.* **8**, e1002905 (2012).
109. Parsch, J., Novozhilov, S., Samadin-Peter, S. S., Wong, K. M. & Andolfatto, P. 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 (2010).
110. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1–13 (2009).
111. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2008).
112. Chen, E. Y. et al. Enrichr: interactive and collaborative html5 gene list enrichment analysis tool. *BMC Bioinforma.* **14**, 128 (2013).
113. Kuleshov, M. V. et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–97 (2016).
114. Storer, J., Hubley, R., Rosen, J., Wheeler, T. J. & Smit, A. F. The dfam community resource of transposable element families, sequence models, and genome annotations. *Mob. DNA* **12**, 2 (2021).

## Acknowledgements

We would like to thank DrosEU researchers for sharing with us the strains sequenced in this manuscript (Supplementary Data 1). 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). S.R. was 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.: design of the work, 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. J.G.: conception and design of the work, analysis and interpretation of data, drafted and revised the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-022-29518-8>.

**Correspondence** and requests for materials should be addressed to Josefa González.

**Peer review information** *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022