TrEMOLO: Accurate transposable element allele frequency estimation using long-read sequencing data combining assembly and mapping-based approaches

Mourdas Mohamed [(1)]*, François Sabot [(2,3)]*, Marion Varoqui [(1)], Bruno Mugat [(1)], Katell Audouin [(4)], Alain Pélisson [(1)], Anna-Sophie Fiston-Lavier [(5,6)]°, Séverine Chambeyron [(1)]°

[(1)]Institute of Human Genetics, UMR9002, CNRS and Université de Montpellier, Montpellier, France

[(2)]DIADE, University of Montpellier, CIRAD, IRD, Montpellier, France

[(3)] IFB - Southgreen Bioversity, CIRAD, INRAE, IRD, Montpellier, France

[(4)] TAGC, UMR 1090 INSERM, Marseille, France

[(5)]ISEM, Université Montpellier, CNRS, IRD, CIRAD, EPHE, Montpellier, France
[(6)]Institut Universitaire de France (IUF)

*co-first authors

°co-last and corresponding authors

Mourdas Mohamed: mourdas.mohamed@igh.cnrs.fr
François Sabot: francois.sabot@ird.fr ORCID: 0000-0002-8522-7583
Marion Varoqui: marion.varoqui@igh.cnrs.fr
Bruno Mugat: bruno.mugat@igh.cnrs.fr
Katell Audouin: katell.audouin@etu.univ-amu.fr
Alain Pélisson: alain.pelisson@igh.cnrs.fr
Anna-Sophie Fiston-Lavier: anna-sophie.fiston-lavier@umontpellier.fr ORCID: 0000-0002-7306-6532
Séverine Chambeyron: severine.chambeyron@igh.cnrs.fr ORCID: 0000-0003-2775-6556

Abstract

Transposable Element MOnitoring with LOng-reads (TrEMOLO) is a new software that combines assembly- and mapping-based approaches to robustly detect genetic elements called transposable elements (TEs). Using high- or low-quality genome assemblies, TrEMOLO can detect most TE insertions and deletions and estimate their allele frequency in populations. Benchmarking with simulated data revealed that TrEMOLO outperforms other state-of-the-art computational tools. TE detection and frequency estimation by TrEMOLO were validated using simulated and experimental datasets. Therefore, TrEMOLO is a comprehensive and suitable tool to accurately study TE dynamics. TrEMOLO is available under GNU GPL3.0 at https://github.com/DrosophilaGenomeEvolution/TrEMOLO.

Keywords

Genome, software, transposable element calling, long-read DNA sequencing, nanopore sequencing, transposable element allelic frequency, haplotypes, structural variation

Background

Transposable Elements (TEs) are endogenous mobile elements that can move within their host genome and increase their copy number, autonomously or not (1). They are present in almost all genomes that have been analyzed, and are represented by various structures and mobility mechanisms. Their size ranges from less than 100 bp (e.g. for MITEs (2)) to more than 20kb (3,4). In large genomes, they represent more than 90% of nuclear DNA (5), but they whatever are a major component of smaller genomes. For instance, the *Drosophila melanogaster* genome is only 180Mb long, but almost ~15% of it is made of TEs (6). However, their identification, annotation, and dynamic tracking are complex due to their intrinsic repetitive nature. Indeed, identifying

53 insertions in a newly assembled genome is a hard and complex task. Many tools have been
54 developed to tackle this issue (*e.g.*, REPET (7), RepeatMasker (8), EDTA (9)), with more or less
55 success, and requiring all a large amount of computational resources. Moreover, tracking the
56 presence or absence of a given insertion among individuals or through generations was, up to
57 recently, limited to heavy wet-laboratory approaches, with a low throughput (10–12).

58 Sequencing technology advances in the last 15 years, especially Illumina short-read sequencing,
59 have allowed sequencing, on a daily basis, not only individuals within a species or subspecies, but
60 also within whole populations. Nowadays, it is not uncommon to sequence the same population
61 many times over generations to study its genetic evolution (13). One of the main findings provided
62 by the availability of massive sequencing data is that TEs are the most variable component of
63 genomes, even more than previously suspected (14). Therefore, many tools have been developed
64 to use short-read sequencing data to identify potential TE polymorphisms among samples (*e.g.,* T-
65 lex (15–17); McClintock (18)) or within a population (*e.g.,* DNApipeTE (19); PoPoolationTE2 (20)).
66 They rely on three main mapping features: split reads, unpaired reads, and depth of coverage (14).
67 However, short-read sequencing data are not suitable for the optimal detection of structural
68 variations (SV), especially SV due to TE insertion/deletion. Indeed, as short-reads are shorter (100
69 to 250 bases) than the mean length of repeats and TEs, they do not span the entire length of the
70 variation and thus cannot be used to correctly identify SV, leading to high false-positive rates
71 (40%) and false-negative results (21,22). Conversely, Long Read (LR) sequencing technologies,
72 such as Pacific Biosciences and Oxford Nanopore Technologies (ONT), allow obtaining reads that
73 are >20kb long (and even 100kb and more for ONT). Therefore, a single read may contain a
74 complete SV or TE insertion and enough surrounding sequence data to unambiguously anchor the
75 variation. However, the longer read length is counterbalanced by a lower sequencing quality (still
76 improving) compared with short-read sequencing and fewer dedicated tools for TE detection. The
77 current methods to detect TE variations using LR sequencing data (and also most of the short-read
78 sequencing data-based tools) generally rely on the presence/absence of TE insertions annotated
79 in a reference genome (*e.g.,* LorTE (23)), or on the detection of abnormally mapped reads upon

the reference genome that can be linked to a TE (*e.g.,* TELR (24)). However, some tools also allow the partial detection, assembly and annotation of non-reference TE insertions (e.g. TLDR) (Transposons from Long DNA Reads) (25).

In addition, unlike short-read, LR sequencing allows generating high-quality genome assemblies with higher levels of contiguity, and consequently to compare genomes directly and to extract more efficiently large and complex SVs (26). Indeed, the number of available almost complete genomes is increasing due to these LR sequencing technologies (13,27), and even telomere-to-telomere assemblies can be obtained (28). Moreover, some reliable tools are available for genome-to-genome comparisons (MUMmer (29), minimap2 (30)) and for identifying complex SV between two assemblies more efficiently than with mapping-based approaches. However, assembly methods generally rely on a layout-and-consensus step when solving the multiple possibilities at a given locus (*i.e.,* choosing one of all possible haplotypes in the graph, usually the most represented one), and thus may mask some minor populational/somatic variations.

In this paper, we present the Transposable Element MOvement detection using LOng-reads (TrEMOLO) software that combines the advantages offered by LR sequencing (*i.e.,* highly contiguous assembly and unambiguous mapping) to identify TE insertion (and deletion) whatever the frequency at which they segregate in a population. TrEMOLO (available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO) is a SnakeMake pipeline that can be used in script mode with dependencies already installed, or within a Singularity container to ensure a higher reproducibility and to facilitate future updates. The pipeline is designed so that the inclusion of additional samples is straightforward. We demonstrated its efficiency using simulated and experimental LR sequencing data from a highly polymorphic *Drosophila* population. TrEMOLO detected TE insertions that could not be identified with other approaches, with a very low error rate. Therefore, TrEMOLO provides an unprecedented accurate picture of TE variability compared with other tools.

106

107 # Results

108 ## TrEMOLO design

109 TrEMOLO is a suite of Python 3 (31) scripts that follow the SnakeMake (32) rules. It takes as input
110 a TE library, a reference genome sequence, single or pooled LR DNA sequencing data, and a
111 genome assembly generated from these sequencing data. In the sequel, TE insertions from the
112 major haplotype belonging to the assembly will be called INSIDER TE insertions. The minor
113 frequency haplotype TE insertions are not included in the assembled genome and will be called
114 OUTSIDER TE insertions. The complete tool includes three internal modules (Fig.1 and Fig. S1):
115 (1) the ***INSIDER TE detection module*** identifies TE insertions and deletions present only in the
116 newly assembled genome when compared to the reference genome, by aligning these two
117 genomes and parsing the pairwise alignment; (2) the ***OUTSIDER TE detection module*** searches
118 for the additional TE insertions (those that are carried by rare haplotypes and are therefore not
119 present in the assembly), by mapping, LRs to the genome assembly generated using the same
120 data, and detecting TE variants by parsing the mapping results (Fig.1 and Fig.S1); (3) the ***TE***
121 ***analysis module*** computes from the previous modules outputs the frequency of each TE variant
122 using the LR sequencing data and looks for target site duplications (TSDs) that occur upon TE
123 integration as a result of staggered DNA double-strand breaks (33). A detailed manual on how to
124 run TrEMOLO with all its options is available at TrEMOLO github
125 (https://github.com/DrosophilaGenomeEvolution/TrEMOLO).

126

127 ***The INSIDER TE detection module.*** This module detects the so-called INSIDER TEs. These are
128 the non-reference TE insertions/deletions that are found frequently enough among the reads to be
129 incorporated in the assembly (generally the TE polymorphisms of the major haplotype). It takes as
130 input a TE library, a reference genome, and the genome assembly of the studied biological

131 sample, all in fasta format. This assembly must be generated with the input LR sequencing data to
132 be compatible with the second module. Whole-genome pairwise alignment is performed between
133 the reference and assembled genomes using minimap2 (30), and is then parsed with *Assemblytics*
134 (34) for SV identification (Fig. S1). This analysis returns a list of SVs from the pairwise alignment of
135 the assembly against the reference genome with their positions. This list is then compared with a
136 TE library to keep only SV belonging to already known TE families (Fig. 1, left panel, and Fig. S1).
137 Before providing the final list of detected TE insertions/deletions, the alignment data are scanned
138 to filter out sequences that do not fit the parameters (considered as low-quality TEs),
139 corresponding here to 80% of the size and 80% of identity (Methods).

140 ***The OUTSIDER TE detection module.*** This module allows to retrieve OUTSIDER TEs, i.e. those
141 TE polymorphisms that are low frequency in the sample and not included in the genome assembly.
142 This module takes as input a TE library in fasta format, a genome assembly (fasta) and the LR
143 sequencing data in fastq used to build this assembly. The reads are aligned to the assembled
144 genome using minimap2 (30). Reads that are flagged as partially or non-mapping in the output
145 SAM file are candidates for carrying OUTSIDER TEs. Candidate reads are validated if the
146 corresponding sequences are present in the provided TE library (Fig. 1 right panel, Fig. S1). As for
147 the INSIDER module, low-quality TEs are filtered out and then, TE insertions are listed in a BED
148 file.

149 ***The TE analysis module.*** This module is dedicated to the analysis of the newly inserted TEs (a) to
150 detect hallmarks of the transposition mechanism (*i.e.* TSDs) and (b) to estimate the allele
151 frequency of each TE insertion.

152 (a) For the INSIDER TE insertions, the 30 bp flanking both extremities of the TE insertion are
153 aligned against each other using the Boyer-Moore exact algorithm (35) to detect short motifs
154 shared on each side of the insertion without mismatches. For the OUTSIDER TE insertions, the
155 most informative single LR (with the higher BLAST score) spanning the TE insertion is selected.
156 Then, like for INSIDER TE insertions, the sequences flanking the TE insertion site are extracted

157 and aligned. However, to handle the LR error rate, the algorithm can accept mismatches (default
158 1). If the putative TSDs are located more than 2pb away from the detected 5' and 3' ends of the TE
159 insertions, they are filtered out because it could be a simple duplication present in the genome. In
160 addition, for OUTSIDER TE insertions, if the putative TSD sequence is not found as a unique
161 sequence at the putative insertion locus in the assembled genome, it is also filtered out and not
162 considered as the sequence of the empty site. Finally, retained TSDs allow to better define the 5'
163 and 3' ends of the inserted TE.

164 (b) To compute the insertion frequency, the number of reads with the TE insertion is divided by
165 the total number of reads spanning the same position, either with or without TE.

166 **OUTPUT data.** Several BED files are produced containing information on the inserted or deleted
167 TE insertions, their nature and their location on the reference and assembled genomes, as well as
168 the computed frequencies for each of them.

169

170 Using simulated datasets to benchmark the ability of TrEMOLO to annotate new TE
171 insertions on a reference genome

172 We compared the results obtained by TrEMOLO, TELR (24) and TLDR (25), two recently
173 developed computational tools dedicated to TE detection using LR sequencing data. The rationale
174 of these tools is similar to the one used with short-read sequencing data: they use LR data mapped
175 to a reference sequence, through a BAM file for TLDR (25) or directly from the mapping tool output
176 for TELR. Insertions are detected based on discordant mappings (*i.e.* clipped reads and partially
177 mapped reads). Finally, as for short-read-based approaches, the non-mapped parts of the reads
178 are aligned against a library of canonical TE sequences. Then, discordant mapped reads
179 associated with TE matches are selected, clustered, and analyzed to validate or invalidate the TE
180 insertion sites in the reference genome.

181 We generated a simulated reads dataset (sample S1, see Methods and Fig. S2) in which 1100 TE
182 insertions with different allele frequencies were added to the *D. melanogaster* unmasked genome
183 called G0-F100 previously published (13). For that, we first generated LR dataset with
184 DeepSimulator on G0-F100 (sb1 at a depth of 10x). Then we used a G0-F100 pseudo genome
185 containing 100 TE insertions to simulate the so-called INSIDER TE sequencing dataset (sb2 at a
186 depth of 45x). To obtain the OUTSIDER TE simulated sequencing dataset (sb3 at a depth of 5x),
187 we launched DeepSimulator on the G0-F100 INSIDER pseudo genome into which we had inserted
188 1000 additional TEs. Finally, the three LR datasets were mixed together (sb1+sb2+sb3) to obtain
189 the sample S1 (see Methods and Fig. S2; simulated data are available at
190 https://doi.org/10.23708/N447VS).

191 We mapped this simulated reads dataset to the unmasked G0-F100 genome using TELR (24),
192 TLDR (25) and the TrEMOLO OUTSIDER module (here called "TrEMOLO_NA", see Methods) with
193 minimap2 (30). We also launched the whole TrEMOLO pipeline (both OUTSIDER and INSIDER
194 modules) with the G0-F100 genome as reference and as the assembled genome, we used the G0-
195 F100 genome with the 100 highly frequent TE insertions. As the TLDR parameters are fixed (25),
196 we decided to use the same parameters for the four methods/tools (TrEMOLO, TrEMOLO_NA,
197 TLDR, and TELR): 200 bp as minimum TE insertion length and 80% as minimum TE sequence
198 identity. To limit false-positive TE detections, we also set at 15kb the largest TE insertion length.

199 Using these simulated datasets, we observed that out of the 1100 simulated insertions, TELR and
200 TLDR only detected 12 (1%) and 928 (84.4%) TE true-positive insertions, respectively. By contrast,
201 TrEMOLO_NA found 1064 (96.7%) TE true-positive insertions, illustrating the greater sensitivity of
202 the TrEMOLO mapping algorithm (Fig. 2). Unlike TLDR and TELR tools, TrEMOLO also relies on
203 LR based genome assemblies, which enables it to retrieve and use the information of the most
204 frequent TE polymorphims. When using a complete TrEMOLO approach, *i.e.* by taking advantage
205 of the comparison of two whole-genome assemblies, the detection rate even rose up to 97.3%.

206 Moreover, only 80 (7%) and 95 (8%) TE false-positive insertions were detected in TrEMOLO_NA
207 and TrEMOLO respectively, which is the same order of magnitude as the TLDR false positive
208 insertion rate (11%), and remains very low compared to the significant gain of true-positive
209 insertions detected by TrEMOLO.

210

211 **TrEMOLO experimental validation**

212 To validate TrEMOLO insertion/deletion detection accuracy with real biological samples, we took
213 advantage of a previously described *D. melanogaster* laboratory line (36) in which the knockdown
214 of Piwi, a protein essential for TE silencing, can be induced in adult follicle cells (somatic cells of
215 the ovary). This transient somatic Piwi knockdown allows TE derepression in follicle cells and their
216 integration as new proviruses in the progeny genome (36). We performed TE derepression for 73
217 successive generations to obtain the G73 line that contains thousands of newly integrated TEs in
218 its genome (13,36). We used genomic DNA extracted from G0-F100 (before Piwi knockdown) and
219 G73 (G73-SRE; see Methods) fly samples to check by genomic PCR, 12 of the 2334 newly
220 integrated TEs detected by TrEMOLO in a previous G73 sequencing dataset (G73 depth 183x)
221 (13).This small sample is representative of both OUTSIDER (mostly from the gtwin and ZAM
222 families) and INSIDER TE insertions, i.e. insertions detected by TrEMOLO in either a minority or a
223 majority of the initial G73 (183x) LRs, respectively. Using G73-SRE DNA, genomic PCR yielded
224 amplicons of the expected sizes, whereas no amplification was observed when the G0-F100
225 reference genomic DNA was used as a template (Fig. 3a, upper panels and lower left panel).
226 Moreover, to control the DNA sample quality, we also checked the presence of four TEs shared by
227 both genomes (Fig. 3a, lower right panel). We also experimentally validated the excision of a *hobo*
228 DNA transposon in G73 that was present in the G0-F100 genome (Fig. 3b; 3c). Altogether, these
229 data illustrated the efficacy of TrEMOLO to detect both insertions and excisions whatever their
230 frequency in the population.

231 **Impact of the genome assembly quality on TrEMOLO efficiency**

232 To assess the impact of the genome assembly quality on TE insertion/deletion detection, we ran
233 TrEMOLO using the G73 genome assembled with different tools and after different polishing
234 rounds, using the G0-F100 assembly as reference. We used Flye (37), Shasta (38) and Wtdbg2
235 (39) followed by Racon (40) polishing steps and a scaffolding by RaGOO (41) (Methods). As
236 previously reported (13), four rounds of Racon polishing (40) increased the BUSCO score (Table
237 1), suggesting a better-quality assembly.

238 **Table 1** : Comparison of the TE insertion/deletion detection efficiency by TrEMOLO using genome
239 assemblies of different quality.
240

Assembly process	Contigs / scaffolds number	N50 (bp)	BUSCO	Assembly total size (bp)	Number of TE insertions	Number of INSIDER TE insertions	Number of OUTSIDER TE insertions	Number of INSIDER TE deletions
Flye + Racon (4 runs) + Ragoo*	30	27 282 734	98%	148 869 489	2334	60	2274	49
Flye + Racon (1 run) + Ragoo*	31	27 316 827	96.9%	149 145 078	2339	54	2285	49
Shasta + Ragoo*	15	27 256 021	93%	151 203 880	2344	94	2250	56
Wtdbg2 + Ragoo*	22	34 296 233	92.6%	171 989 387	2579	159	2420	109

241 * Ragoo was run with default parameters.

242

243 Using genomes assembled with the recently updated Flye and Shasta assemblers led to the
244 detection of similar numbers of INSIDER and OUTSIDER TEs, with a slightly higher number of
245 INSIDER TEs for the Shasta-assembled genome (Table 1). When using a genome assembly
246 generated with Wtdbg2, which has not been updated for two years and thus has not incorporated
247 the recent changes in ONT sequencing data (*e.g.*, new base calling inducing a much lower error
248 rate) and has a lower BUSCO score (indicating a lower quality of assembly), TrEMOLO detected
249 roughly the same number of OUTSIDER TE insertions as the more recent assemblers. Note
250 however, that the INSIDER module produced at least twice as much insertions and deletions
251 (among which many of them were probably false positives) probably because it was based on a
252 low-quality assembly.

253 Altogether, we conclude that a draft genome assembly, even obtained with a not-so-recent
254 assembler version (with a lower BUSCO score indicating a lower assembly quality), is of enough
255 quality however to allow the identification of TE polymorphisms by TrEMOLO.

256

257 **Validation of TE frequency estimated by TrEMOLO using simulated data**

258 To validate the accuracy of TrEMOLO frequency estimation, we used simulated datasets with
259 known frequencies of INSIDER and OUTSIDER TEs (see Methods; Fig. 4a). The frequency is
260 computed by the number of reads supporting the insertion divided by the total number of reads
261 spanning the region of insertion. For OUTSIDER TEs, we count as supporting reads, the reads
262 with the complete insertion as well as the clipped ones (Fig. 4a). TrEMOLO frequency module
263 slightly overestimated the frequency of INSIDER TEs for the simulated datasets (Fig. 4b). For
264 OUTSIDER TEs, whatever is the frequency, TrEMOLO obtained an accurate estimation of the
265 frequency (Fig. 4c).

266

267 **Experimental validation of TE frequency estimated by TrEMOLO using droplet digital**
268 **polymerase chain reaction**

269 To validate the TrEMOLO frequency module accuracy, we re-extracted long DNA molecules from
270 G73 flies (Methods) and analyzed the extracted DNA by ONT sequencing with Short-Reads
271 Eliminator kit (G73-SRE library) and by droplet digital polymerase chain reaction (ddPCR). This
272 latter approach allows the absolute quantification using forward primers located in the left side of
273 the TE landing site and reverse primers located either in the TE or on the right side of the landing
274 site (Fig. 5). To increase the detection specificity, we also designed fluorescent probes. The primer
275 and probe sequences are listed in Table S2. We selected nine of the 2234 TE insertions, as
276 representatives of different TEs families and of frequencies in the initial G73 library (depth of 183x)
277 (13). We ran TrEMOLO (parameter: 80% as minimum TE sequence identity and 10% as minimum

278 TE insertion length) and estimated their frequency in the new G73-SRE library having a depth of
279 82x (Table 2).

280 **Table 2 :** TE frequencies estimated using ddPCR and TrEMOLO.
281

	Frequency estimation (%)	
Name of the TE and its coordinates on G73 (183x) assembly	ddPCR G73-SRE	TrEMOLO G73-SRE (82x)
ZAM-3L_13591033	0.6	NA
blood-2L_18128567	2.2	37.68
ZAM-2L_16741915	9.6	17.3
gtwin-3L_6082767	10.4	6.58
ZAM-2L_16868329	13.6	17.89
ZAM-2R_18388126	17.3	18.31
gtwin-3L_163676	26.5	27.63
roo-2L_13882749	59.2	57.95
412-2R_14285430	82.6	93.44

282 NA= TE insertions with no reads containing the insertion in G73-SRE (82x).

283

284 For two of them, the low frequencies estimated by ddPCR (0.6% and 2.2%) do not allow an
285 accurate estimation of their frequency by TrEMOLO. In the G73-SRE library (depth 82x) there are
286 no reads supporting the ZAM-3L_13591033 insertion, while we detected it using the initial library
287 G73 that has shorter reads and a higher depth (183x). Concerning the blood-2L_18128567
288 insertion, the TrEMOLO frequency estimation is not congruent with the ddPCR one. This
289 overestimation of the frequency by TrEMOLO (37.68%) illustrates that TrEMOLO estimation starts
290 to be accurate when the insertion TE frequency is around 10% in the population. Indeed, for the

291 other six TE insertions tested, the ddPCR estimates fitted very well with the TrEMOLO estimates,
292 whatever their frequency from 10% to 82%.

293

294 **Impact of sequencing depth on TE calling and frequency estimation**

295 The total amount of the initial G73 sequencing is averaging a depth of 183x, fairly representative
296 most of the 200 haploid genomes of the 100 males used for the DNA extraction. Indeed, the
297 probability of sequencing one out of the 200 haplotypes from this sample (i.e. the depth per
298 haplotype) was (1:200)x183=0.91. From this dataset, we generated several sequence subsets
299 using Filtlong (42) (see Methods), resulting in seven subsamples of reads (ranging from 145x to
300 15x of depth) (Table S3). We then launched TrEMOLO with these datasets, and compared the
301 number of OUTSIDER TE insertions detected by mapping these reads to the G73 assembly (Table
302 S3 and Fig. 6).

303 Using the initial dataset G73 (183x), TrEMOLO detected 2334 TEs of which 2274 are OUTSIDER
304 TE insertions. The percentage of detected OUTSIDERs is relative to this dataset. Decreasing the
305 depth from 183x to 76x slightly reduced the number of OUTSIDER TEs detected. Then, TrEMOLO
306 detection performance decreased more rapidly. Nevertheless, even with the lowest depth (15x),
307 TrEMOLO could still identify 47% (1064) of OUTSIDER TE insertions. As expected, these
308 insertions were significantly more frequent in the 183x dataset than the 1210 ones that were lost
309 during the down-sampling (Fig. S3, p=0,05 Wilcoxon rank test). Thus, in case of highly
310 polymorphic populations, such as ours, we would recommend a sequencing depth corresponding
311 to a minimum of 0.7x per haplotype (corresponding here to our 145x dataset) to ensure an optimal
312 recovery of low-frequency transpositions.

313 Among the 2334 original TE insertions, only 820 OUTSIDER TEs were detected in all subsamples
314 (Fig. 7). To test the impact of the sequencing depth on the accuracy of frequency estimates, we
315 computed the frequencies of these 820 shared TE insertions in each subsample. The accuracy

316 of OUTSIDER TE insertion frequency estimation was acceptable when using datasets
317 with sequencing depth ≥76x, with a decrease of 12% maximum compared with the
318 initial frequency computed at 183x (Fig. 7). However, for very "low depth" datasets, (here
319 0.19x per haplotype or less), the accuracy frequency decreased quite rapidly with an almost
320 systematic underestimation by up to 70% compared with the original dataset.

321

322 **Discussion**

323 Accurate detection of TEs with different frequencies in a population or in a single sample at the
324 somatic level is important for researchers in evolution, ecology, agronomy, and medicine.
325 However, it is still a challenge due to the repetitive nature of the TEs. Most of the current
326 approaches to identify TE insertions are based on short-read mapping to a reference genome,
327 sometimes quite distant from the studied samples. Working with distantly related genomes may
328 drastically bias the mapping results and thus the subsequent analyses, for instance by providing
329 high false positive rates (43). Therefore, TE detection tools based on mapping approaches allow
330 the identification of low-frequency TE insertions (or from the non-major haplotype), but not all of
331 them, only when using very high-depth short-read sequencing data. The availability of low-cost LR
332 sequencing technologies (ONT) opens the possibility to accurately detect new TE insertions.
333 Improving the sequencing quality allows to obtain accurate and (almost) complete genome
334 assemblies in a short amount of time for small sized ones. For instance, a high-quality *Drosophila*
335 genome can be assembled in 24 hours (44). However, most of these assemblers provide a single
336 haplotype sequence as output, generally the most prevalent. Finally, most TE detection tools rely
337 on the identification of TE fragments as source of SV *ab initio*, through a global genome annotation
338 without any comparison with other already annotated sequence.

339 We designed TrEMOLO, a comprehensive computational tool with the aim of combining the main
340 haplotype information (assembly step) followed by the recovery of minor haplotypes by LR

341 sequence mapping to the newly assembled genome to detect all SV types without *ab initio* TE
342 detection. It is only after this naive detection that the SVs are annotated as possible TEs or not.
343 This approach allows to run TrEMOLO using different TE databases. TrEMOLO accuracy in TE
344 identification and the correction system used for TSD detection allow predicting the insertion site
345 within a 2-base pair window (provided, of course, that the extremities of the TE family are precisely
346 defined in the consensus sequence). Therefore, TrEMOLO is currently one of the most efficient
347 tools for TE detection. Moreover, TrEMOLO allows the estimation of the insertion frequency for
348 each detected TE copy with high accuracy (when using data with sequencing depth >50x).
349 Importantly, TrEMOLO performance is weakly influenced by the assembly quality and sequencing
350 depth variations (up to a certain point). In terms of parameters, TrEMOLO in its standard approach
351 will use the previously published 80/80/80 rule for TE recovery (1). Indeed, it relies on 80% of
352 similarity to and 80% in length of the reference TE sequence. However, as shown in the tests
353 before, we could decrease the length coverage down to 10% and still recover true positive
354 variations. This length parameter can thus be modified for e.g. identifying partially transposed
355 elements (for insertions of 5'-truncated LINEs), or insertions of abnormal copies (internally deleted
356 TEs still able to transpose). In the same way, decreasing the level of similarity would allow to
357 detect more distantly related copies from the consensus/reference copy in the database. However,
358 in this case, the decrease of the stringency of detection will have a stronger effect on the quality of
359 results (improving the false positive rates) (45), and we do not recommend it. The best approach
360 would be to add in the reference TE file the targeted sequences instead.

361 In terms of execution times and memory usage, it depends a lot on the considered species.
362 However, as an example, the analyses presented here for simulated data were run on a single
363 laptop computer with 16 Gb of RAM and an 12-cores i7 Intel processor, in less than 10 hours, and
364 thus can be deployed in almost any lab.

365 Therefore, TrEMOLO can be used for different projects (e.g. pooled DNA sequencing, population
366 analysis, human genetics, tumor analysis) that require the fast and precise identification of TE
367 insertions, even at very low frequency within a sample. Indeed, our data showed that a single

368 insertion can be unambiguously identified through a unique read, as long as the read is long
369 enough to harbor the whole TE and its surrounding insertion site.

370 Conclusions

371 TrEMOLO allows the precise analysis and study of haplotype-specific TE insertions and a better
372 understanding of the true impact of TE evolution in the wider context of their host genome.

373

374 Methods

375 TrEMOLO pipeline preprocessing

376 INSIDER TE identification: to compare the reference genome and a new assembly, TrEMOLO
377 uses the same methodology as RaGOO (v1.1) (41) up to the SV identification step. Then, SVs are
378 compared to the TE database (Bergman's laboratory, https://github.com/bergmanlab/drosophila-
379 transposons/) with BLASTN (2.9.0) (--outfmt 6) (46). The output file is filtered based on the size
380 and identity percentages listed in parameters (80% in standard for each of them). DNA fragments
381 that appear to be part of the same sequence are grouped. For INSERTION, Repeat_expansion,
382 Tandem_expansion, sequences are retrieved from the query (new assembly). For DELETION,
383 Repeat_contraction, Tandem_contraction, sequences are recovered from the reference (Fig. S1)

384 OUTSIDER TE identification: the LRs are aligned on the genome resulting from their assembly
385 using minimap2 (v.2.24) (30) and SVs are identified using Sniffles (47). Reads covering the SV are
386 recovered from the SAM/BAM file using as anchor point the Sniffles provided position and a 30bp-
387 window with SAMtools (48). The SV reads and sequences are compared to the TE database, and
388 filtered as done for INSIDER TEs, but only the best support reads are kept among those provided
389 by Sniffles or the SAM parsing (Fig. S1).

390 TSD detection: both TE flanking sequences are recovered to check whether there are duplicate
391 sequences at the TE junction (see Results for more details). For an OUTSIDER TE, if TrEMOLO
392 detects duplicated sequences farther than the initially identified TE junctions, and if they fit to the
393 TSD, then the OUTSIDER TE position is corrected accordingly.

394 TE frequency estimation: two different approaches are used to estimate the INSIDER and
395 OUTSIDER TE frequencies in the samples. For INSIDER TEs, from the original BAM file only
396 reads aligned to the putative TE positions (with a 100bp window 5' and 3') are retained using
397 SAMtools view and the -F 2048 -L option (removing any reads with these FLAGS value (see SAM
398 specifications) and overlapping a provided BED file, respectively). Then, clipped reads are
399 removed to keep only the reads aligned on the putative TE position within a 5bp window (5' and
400 3'). The mean depth of 30bp at the 5' and 3' of the putative insertion is computed with SAMtools
401 depth. Then, the TE frequency is computed as (number of reads showing the insertion)/(total
402 number of reads).

403 For OUTSIDER TEs, first, reads not aligned to the putative TE insertion site are removed from the
404 original BAM file (SAMtools view with the -l option). Then, the clipping tags in the CIGAR strings for
405 the remaining reads are modified by adding an insertion tag of the TE size plus 10 bp to the read
406 sequence (*i.e.,* the same number of "N" at the corresponding position with a fake quality of 8).
407 Then, Sniffles is run again to obtain the true number of reads associated with a given TE. After
408 recovering the local depth, as done for INSIDER TEs, frequency is computed as (number of reads
409 showing the insertion)/(total number of reads).

410

411 **Simulated datasets**

412 A previously assembled reference *D. melanogaster* genome, called G0-F100 (13) was used for the
413 LR simulation by DeepSimulator (version v1.5) (49) (-B 2 -K 10 -l 20 000) with a mean length of
414 20kb and a depth of 10X (subsample 1 (sb1)) (Fig. S2). In parallel, two fake genomes containing

415 INSIDER TE insertions only or both INSIDER and OUTSIDER TE insertions were built as follows.
416 For INSIDER TE insertions, 100 TE sequences were artificially and randomly inserted in the G0-
417 F100 assembly genome. They correspond to the canonical sequences of 11 TE families of
418 different classes and different lengths (from one to 13kb), each one providing five (ZAM and gtwin)
419 or ten new insertions (Doc, roo, copia, F-element, hopper, Helena, HB, TART-A and gypsy). Then,
420 DeepSimulator (version v1.5) (47) (-B 2 -K n -l 20000) was used to simulate LRs with a mean
421 length of 20kb and depth of 30, 40 and 45x (subsamples 2 (sb2)). For OUTSIDER TE insertions,
422 1,000 TE sequences belonging to the same families as the simulated INSIDER TEs and also to the
423 Idefix family (200, 200, 50, 50, 50, 50, 50, 50, 50, 50, 100, and 100 insertions, respectively) were
424 inserted randomly in the G0-F100 genome assembly containing the INSIDER TE insertions. Then,
425 DeepSimulator (version v1.5) (49) (-B 2 -K n -l 20000) was used to simulate LRs with a mean
426 length of 20kb mean length and depth of 5, 10 and 20X (subsamples 3 (sb3)).

427 These three LR subsamples were mixed together (total depth of 60X) to test TrEMOLO sensitivity
428 at different levels: S1: (10X sb1 +45X sb2 + 5X sb3), S2: (10X sb1 + 40X sb2 + 10X sb3), and S3:
429 (10X sb1 + 30X sb2 + 20X sb3) (Fig. S2).

430 For the INSIDER TE frequency estimation, the sb1 and sb2 subsamples were used together with a
431 new sb4 subsample simulated with a depth of 1X and minimum to maximum length of 10 to 50kb
432 to obtain IS1 (10X sb1 + 30X sb2 + 1X sb4), IS2 (10X sb1 + 40X sb2 + 1X sb4) and IS3 (10X sb1
433 + 45X sb2 + 1X sb4). Simulated data set are available at https://doi.org/10.23708/N447VS.

434

435 **Tools comparison**

436 For TLDR (v1 - February 2021; (25)), the specified parameters were: --min_te_len 200 --
437 max_te_len 15000 (minimum of 200 nt and maximum of 15 000 nt in size for detected TEs). The
438 same parameters were used for TrEMOLO and TrEMOLO_NA. For TELR (v1 - August 2021; (24)),
439 the settings were --minimap2_family --assembler flye -x ont --aligner minimap2.

440

441 **Genome assembly**

442 Raw ONT reads for the G73-SRE sample were QC checked using Nanoplot v1.10.2
443 (https://github.com/wdecoster/NanoPlot) for sequencing run statistics. Reads with QC <7 were
444 removed by the sequence provider (Montpellier Genomix, Montpellier, France) before QC. For all
445 ONT sequencing data, the genome was assembled as previously published (13), except that a
446 newer version of Flye (v2.9) was used.

447

448 **Datasets**

449 The LR dataset of the *D. melanogaster* laboratory G0-F100 and G73 lines previously published
450 (13) is available at ENA (https://www.ebi.ac.uk/ena) under the accession number ERP122844.
451 Simulated data set are available at https://doi.org/10.23708/N447VS.

452

453 **Benchmark of *Drosophila* genome assemblies effect**

454 Genome assemblies were performed using the raw LR data and CulebrONT for assembly and
455 polishing (50). Flye v2.9 (37) was used with standard parameters except for the --nano-raw option.
456 Four rounds of Racon v1.4.3 with minimap2 v.2.24 (30) were performed using the standard
457 conditions for ONT data, and with medaka v1.2 (standard conditions). Scaffolding was performed
458 using RaGOO v1.1 (41) with standard options except for output structural variations with the -s
459 option and the corresponding reference. For raw assemblies, Wtdbg2 v2.5 (39) and Shasta v0.1.0
460 (38) were used (standard conditions).

461

462 **Impact of sequencing depth on TE calling and frequency estimation**

463 A previously published G73 long read sequencing library (13) was downsized to smaller samples
464 using FiltLong (42) and the following options: -t 19000000000, -t 15000000000, -t 12000000000, -t
465 10000000000, -t 5000000000, -t 2000000000). This subsampling resulted in a series of
466 sequencing datasets with more or less degraded depths ranging from 183x to 15x.

467

468 **Genomic DNA extraction and Short-Read Eliminator kit (SRE)**

469 Genomic DNA was extracted from 50 mg of G0-F100 and G73 male flies using the Nanobind
470 Tissue Big DNA Kit (Circulomics SKU NB-900-701-01) with Insect buffer (PL1) as lysis buffer.
471 Briefly, samples were homogenized in 200 µl CT buffer with a pellet pestle, centrifuged at 16,000 x
472 g at 4°C for 2min. Pellets were washed in 1 ml CT buffer and centrifuged at 16,000 x g at 4°C for
473 2min. Pellets were resuspended in 200µl PL1 lysis buffer (Circulomics Aux. Kit) supplemented with
474 20µl proteinase K and incubated in a thermomixer at 55°C and 900 rpm for 1 h. 20µl RNase A was
475 added and incubation was continued for 15 min. Samples were centrifuged at 16,000x g for 5 min,
476 and supernatants passed through a 70µm filter and collected in a 1.5 ml DNA LoBind
477 microcentrifuge tube. 50 µl BL3, a Nanobind disk and 400 µl isopropanol were added to the
478 samples. Incubation and washes were done according to the manufacturers' instructions. Elution
479 was performed with 50 µl EB buffer. The genomic DNA concentration was estimated with a
480 Nanodrop spectrophotometer (Thermo Fisher Scientific) and a Qubit Fluorometer with the Qubit
481 dsDNA BR Assay Kit (Thermo Fisher Scientific), according to the manufacturer's instructions. To
482 get rid of reads <25 kb, genomic DNA samples with concentration adjusted between 50-150 ng/µl
483 in 60 µl were treated with the SRE Kit, 25 kb cutoff (Circulomics SKU SS-100-101-01), according
484 to the manufacturer's instructions, and eluted in 50 µl EB.

485

486 **Library preparation for G73-SRE LR sequencing (depth 82x)**

487 Libraries were prepared on 1µg genomic SRE DNA with the SQK-LSK110 Kit (Nanopore) following
488 the manufacturer's instructions. 12 µl of each library, 37.5µl of sequencing buffer II (SBII), and 25.5
489 µl of loading beads II (LBII) were loaded onto the R9.4.1 flow cell. Base-calling was performed
490 using the Guppy v5 hac model.

491
492

493 **PCR amplification of long DNA molecules**

494 PCR was performed with 60 ng G0-F100 or G73 genomic DNA in a final volume of 20µl containing
495 10µl Platinum Hot-Start Green Master Mix 2X (Invitrogen 14001-012) and 0.2 µM final
496 concentration of each primer (see Table S1 for sequences). Annealing was performed at 60°C for
497 15 sec and elongation was adapted to the amplicon size (15 sec/kb). Between 3 and 5 µl of each
498 PCR reaction were loaded on a 1% agarose gel with EtBr.

499

500 **Droplet digital PCR for TE insertion frequency determination**

501 Between 5 and 15 ng genomic SRE DNA was combined with 10 µl of 2x ddPCR Supermix for
502 probes (No dUTP) (Bio-Rad), 1µl 20x primers/probe sets (1x primers/probe set includes 450 nM
503 each primer and 250 nM probe) for the targets (see design in Table S2), 1µl 20x primers/probe set
504 for the housekeeping reference gene (*RpL32*), 1µl restriction enzyme (*Hind*III, diluted 1:10 in
505 diluent B (NEB)) and nuclease-free water to a total volume of 20µl.

506 For droplet generation, 20 µl of each mixture was transferred to the center line of a DG8 droplet
507 generator cartridge (Bio-Rad Laboratories). Then, 70 µl of droplet generation oil for the probe (Bio-
508 Rad Laboratories) was injected in the cartridge bottom line and droplets were produced using a
509 QX200 Droplet Generator (Bio-Rad Laboratories). Droplets (40 µl) were transferred to a 96-well
510 plate and endpoint PCR was performed on a T100 Thermal Cycler (Bio-Rad Laboratories) using

511 the following cycling protocol: enzyme activation at 95 °C for 10 min, followed by 40 cycles of 95
512 °C for 30 s and 60 °C for 1 min (ramp rate settings of 2°C/sec), followed by a droplet stabilization
513 step at 98 °C for 10 min, and a final hold step at 4 °C. Droplet measurement was performed on a
514 QX200 Droplet Reader (Bio-Rad Laboratories) and data analyzed with the QuantaSoft software
515 (Bio-Rad Laboratories). Two different fluorescent signals were measured: i) amplification of the TE
516 inserted in a given locus sequence (TE_{in}) or amplification of the same locus without TE (TE_{out}) in
517 the FAM channel, and ii) amplification of the *RpL32* housekeeping gene in the HEX channel. The
518 proportion of TE_{in} or TE_{out} to *Rpl32* was calculated as follows: $TE_{in}/Rpl32$ *or* $TE_{out}/Rpl32$. Each value
519 was expressed as copies/µl. Frequency was calculated with the following formula: Frequency in
520 %= ($TE_{in}/Rpl32$x100) / ($TE_{in}/Rpl32$ + $TE_{out}/Rpl3$).

521

522 # Declarations

523 **Availability of data and materials**

524 The LR sequencing data used for this study were deposited at ENA (https://www. ebi.ac.uk/ena)
525 under the accession numbers (ERP138838 under validation by ENA) and *the benchmarking*
526 *datasets and the command lines are available at https://dataverse.ird.fr/dataverse/tremolo_data* .

527 - Project name: TrEMOLO
528 - Project home page: https://github.com/DrosophilaGenomeEvolution/TrEMOLO
529 - Archived version: https://doi.org/10.23708/2FYBUL
530 - Operating system(s): Linux
531 - Programming language: Python/SnakeMake
532 - Other requirements: Singularity or minimap2/assemblytics/samtools
533 - License: GNU GPL v3

534 Any restrictions to use by non-academics: license acceptation

Competing interests: The authors declare no competing interests.

Funding

This research was funded by the Agence Nationale de la Recherche, grant number (**ANR-19-CE12-0012** -Top53) to S.C., by the MITI CNRS to A.S.F.L.. M.V. was funded by CNRS-Japon PhD Joint Program.

Authors' contributions

F.S., M.M., A.S.F.L and S.C. designed the methodology; M.M. and F.S. encoded software; M.V. tested the software; F.S., B.M. and S.C. supervised M.M., K.A. and B.M. respectively; Biological experiments: B.M and K.A.; F.S., A.P. and S.C. wrote the manuscript. Funding acquisition: A.S.F.L, S.C.. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We thank Dr Marie Fablet for her comments on the manuscript; Elisabetta Andermarcher for manuscript editing; We acknowledge the ISO 9001 certified IRD itrop HPC (member of the South Green Platform) at IRD Montpellier for providing HPC resources that have contributed to the research results reported in this paper (URLs: https://bioinfo.ird.fr/ and http://www.southgreen.fr); the Genotoul platform (https://genotoul.fr/) and (https://www.france-bioinformatique.fr/) for providing calculation time on their servers.

References

1. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007 Dec;8(12):973–82.

2. Jiang N, Feschotte C, Zhang X, Wessler SR. Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol. 2004 Apr;7(2):115–9.

3. Macas J, Neumann P. Ogre elements--a distinct group of plant Ty3/gypsy-like retrotransposons. Gene. 2007 Apr 1;390(1–2):108–16.

4. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007 Apr 1;390(1–2):3–17.

5. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature. 2020 Dec;588(7837):277–83.

6. Mérel V, Boulesteix M, Fablet M, Vieira C. Transposable elements in Drosophila. In: Mobile DNA. In press.

7. Flutre T, Duprat E, Feuillet C, Quesneville H. Considering Transposable Element Diversification in De Novo Annotation Approaches. PLOS ONE. 2011 Jan 31;6(1):e16526.

8. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2: automated genomic discovery of transposable element families [Internet]. Genomics; 2019 Nov [cited 2022 Jul 20]. Available from: http://biorxiv.org/lookup/doi/10.1101/856591

9. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019 Dec 16;20(1):275.

10. Roy NS, 최지영, Lee SI, Soo KN. Marker utility of transposable elements for plant genetics, breeding, and ecology: a review. Genes and Genomics. 2015;37(2):141–51.

11. Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BB, et al. Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet. 1997 Feb 27;253(6):687–94.

12. Flavell AJ, Knox MR, Pearce SR, Ellis TH. Retrotransposon-based insertion polymorphisms (RBIP) for high throughput marker analysis. Plant J. 1998 Dec;16(5):643–50.

13. Mohamed M, Dang NTM, Ogyama Y, Burlet N, Mugat B, Boulesteix 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. 2020 Jul 25;9(8):1776.

14. Goerner-Potvin P, Bourque G. Computational tools to unmask transposable elements. Nat Rev Genet. 2018 Sep 19;

15. Bogaerts-Márquez M, Barrón MG, Fiston-Lavier AS, Vendrell-Mir P, Castanera R, Casacuberta JM, et al. T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. Bioinformatics. 2020 Feb 15;36(4):1191–7.

16. Fiston-Lavier AS, Barrón MG, Petrov DA, González J. T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Res. 2015 Feb 27;43(4):e22.

17. Fiston-Lavier AS, Carrigan M, Petrov DA, González J. T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. Nucleic Acids Research. 2011 Mar;39(6):e36–e36.

18. Nelson MG, Linheiro RS, Bergman CM. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole Genome Shotgun Sequencing Data. G3: Genes, Genomes, Genetics. 2017 Jan 1;g3.117.043893.

19. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De novo assembly and annotation of the Asian tiger mosquito (Aedes albopictus) repeatome with dnaPipeTE from raw genomic reads and comparative analysis with the yellow fever mosquito (Aedes aegypti). Genome Biol Evol. 2015 Mar 11;7(4):1192–205.

20. Kofler R, Gómez-Sánchez D, Schlötterer C. PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. Mol Biol Evol. 2016 Aug 2;msw137.

21. Vendrell-Mir P, Barteri F, Merenciano M, González J, Casacuberta JM, Castanera R. A benchmark of transposon insertion detection tools using real data. Mobile DNA. 2019 Dec;10(1):1–19.

22. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in Drosophila. Nat Commun. 2022 Apr 12;13(1):1948.

23. Disdero E, Filée J. LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. Mob DNA. 2017;8:5.

24. Han S, Dias GB, Basting PJ, Viswanatha R, Perrimon N, Bergman CM. Local assembly of long reads enables phylogenomics of transposable elements in a polyploid cell line. Nucleic Acids Research. 2022 Sep 26;gkac794.

25. Ewing AD, Smits N, Sanchez-Luque FJ, Faivre J, Brennan PM, Richardson SR, et al. Nanopore Sequencing Enables Comprehensive Transposable Element Epigenomic Profiling. Molecular Cell. 2020 Nov;S1097276520307310.

26. Lin J, Jia P, Wang S, Ye K. Comparison and benchmark of long-read based structural variant detection strategies [Internet]. Bioinformatics; 2022 Aug [cited 2022 Dec 8]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.08.09.503274

27. He K, Minias P, Dunn PO. Long-Read Genome Assemblies Reveal Extraordinary Variation in the Number and Structure of MHC Loci in Birds. Genome Biol Evol. 2021 Feb 3;13(2):evaa270.

28. Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022 Apr;376(6588):44–53.

29. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biology. 2004 Jan 30;5(2):R12.

30. Li H. New strategies to improve minimap2 alignment accuracy. Alkan C, editor. Bioinformatics. 2021 Dec 7;37(23):4572–4.

31. Welcome to Python.org [Internet]. Python.org. [cited 2022 Jul 20]. Available from: https://www.python.org/

32. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Res. 2021;10:33.

33. Hedges DJ, Deininger PL. Inviting instability: Transposable elements, double-strand breaks, and the maintenance of genome integrity. Mutat Res. 2007 Mar 1;616(1–2):46–59.

34. Nattestad M, Schatz MC. Assemblytics: a web analytics tool for the detection of variants from an assembly. Bioinformatics. 2016 Oct 1;32(19):3021–3.

35. Boyer RS, Moore JS. A fast string searching algorithm. Commun ACM. 1977 Oct 1;20(10):762–72.

36. Barckmann B, El-Barouk M, Pélisson A, Mugat B, Li B, Franckhauser C, et al. The somatic piRNA pathway controls germline transposition over generations. Nucleic Acids Res. 2018 Oct 12;46(18):9524–36.

37. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019 May;37(5):540–6.

38. Shafin K, Pesout T, Lorig-Roach R, Haukness M, Olsen HE, Bosworth C, et al. Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. Nat Biotechnol. 2020 Sep;38(9):1044–53.

39. Ruan J, Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods. 2020 Feb;17(2):155–8.

40. Vaser R, Sović I, Nagarajan N, Šikić M. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res. 2017 Jan 5;27(5):737–46.

41. Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ, et al. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology. 2019 Oct 28;20(1):224.

42. Wick R. rrwick/Filtlong [Internet]. 2022 [cited 2022 Jul 20]. Available from: https://github.com/rrwick/Filtlong

43. Valiente-Mullor C, Beamud B, Ansari I, Francés-Cuesta C, García-González N, Mejía L, et al. One is not enough: On the effects of reference genome for the mapping and subsequent analyses of short-reads. PLoS Comput Biol. 2021 Jan 27;17(1):e1008678.

44. Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, et al. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. G3 (Bethesda). 2018 03;8(10):3143–54.

45. Gotea V, Veeramachaneni V, Makałowski W. Mastering seeds for genomic size nucleotide BLAST searches. Nucleic Acids Research. 2003 Dec 1;31(23):6935–41.

46. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009 Dec 15;10(1):421.

47. Sedlazeck FJ, Rescheneder P, Smolka M, Fang H, Nattestad M, Haeseler A von, et al. Accurate detection of complex structural variations using single-molecule sequencing. Nat Methods. 2018 Jun;15(6):461–8.

48. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009 Aug 15;25(16):2078–9.

49. Li Y, Wang S, Bi C, Qiu Z, Li M, Gao X. DeepSimulator1.5: a more powerful, quicker and lighter simulator for Nanopore sequencing. Bioinformatics. 2020 Apr 15;36(8):2578–80.

50. Orjuela J, Comte A, Ravel S, Charriat F, Vi T, Sabot F, et al. CulebrONT: a streamlined long reads multi-assembler pipeline for prokaryotic and eukaryotic genomes [Internet]. bioRxiv; 2021 [cited 2022 Jul 20]. p. 2021.07.19.452922. Available from: https://www.biorxiv.org/content/10.1101/2021.07.19.452922v1

Figure legends

Figure 1. Overview of the TrEMOLO pipeline. Three main modules: (1) the INSIDER TE detection module in red, (2) the OUTSIDER TE detection module in blue, and (3) the TE analysis module in green.-

Figure 2: TrEMOLO benchmarking using simulated LR sequencing data. The number of true-positive (True (+)), false-positive (False (+)) and false-negative (False (-)) results obtained using the simulated data (1100 known TE insertions) and four methods/tools (TELR, TLDR, TrEMOLO_NA and TrEMOLO).

Figure 3: Experimental validation of TrEMOLO specificity of TE insertion/deletion detection. a) PCR performed using long DNA molecules extracted from the G0-F100 (control) and G73 Drosophila lines (genomic DNA extracted using Short Read Eliminator kit: G73-SRE sample). The name of the TE family identified by TrEMOLO is indicated followed by the chromosome arm and the genomic coordinate in the G73 (183x) assembly. The frequencies determined in the G73 (depth 183x) of each TE insertion tested by PCR are indicated as percentages in blue. b) Illustration of the primers used to detect a hobo excision c) PCR performed using long DNA molecules extracted from the G0-F100 (control) and G73 Drosophila lines. The primer sequences and amplicon expected sizes are in Table S1.

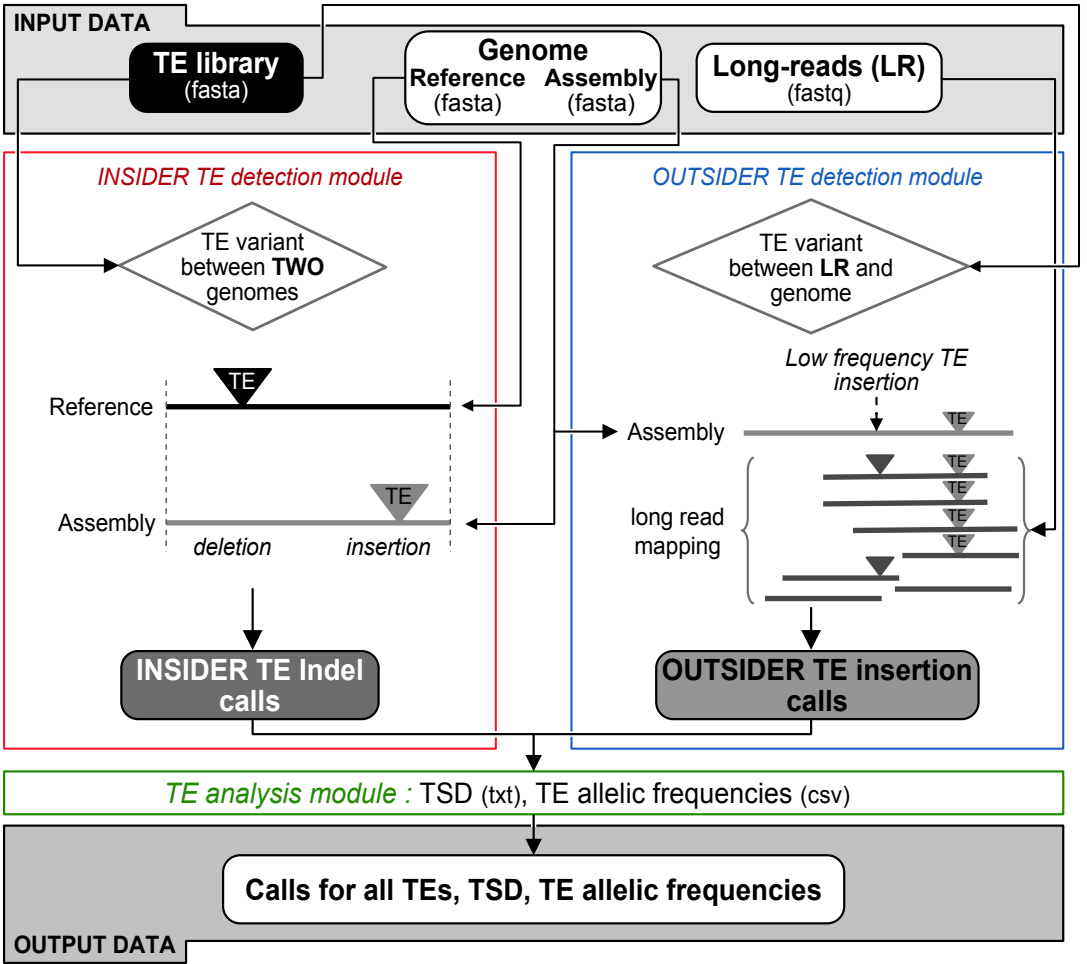
Figure 4: Validation of TrEMOLO frequency estimation. a) Rationale of frequency estimation using fully mapped and clipped reads. b) Violin plots of the INSIDER TE insertion frequencies estimated by TrEMOLO using the simulated datasets. Dashed lines represent the real frequency in the simulated datasets. c) Violin plots of the OUTSIDER TE insertion frequency estimated by TrEMOLO in the simulated datasets with clipped reads. Dashed lines represent the real frequency in the simulated datasets.

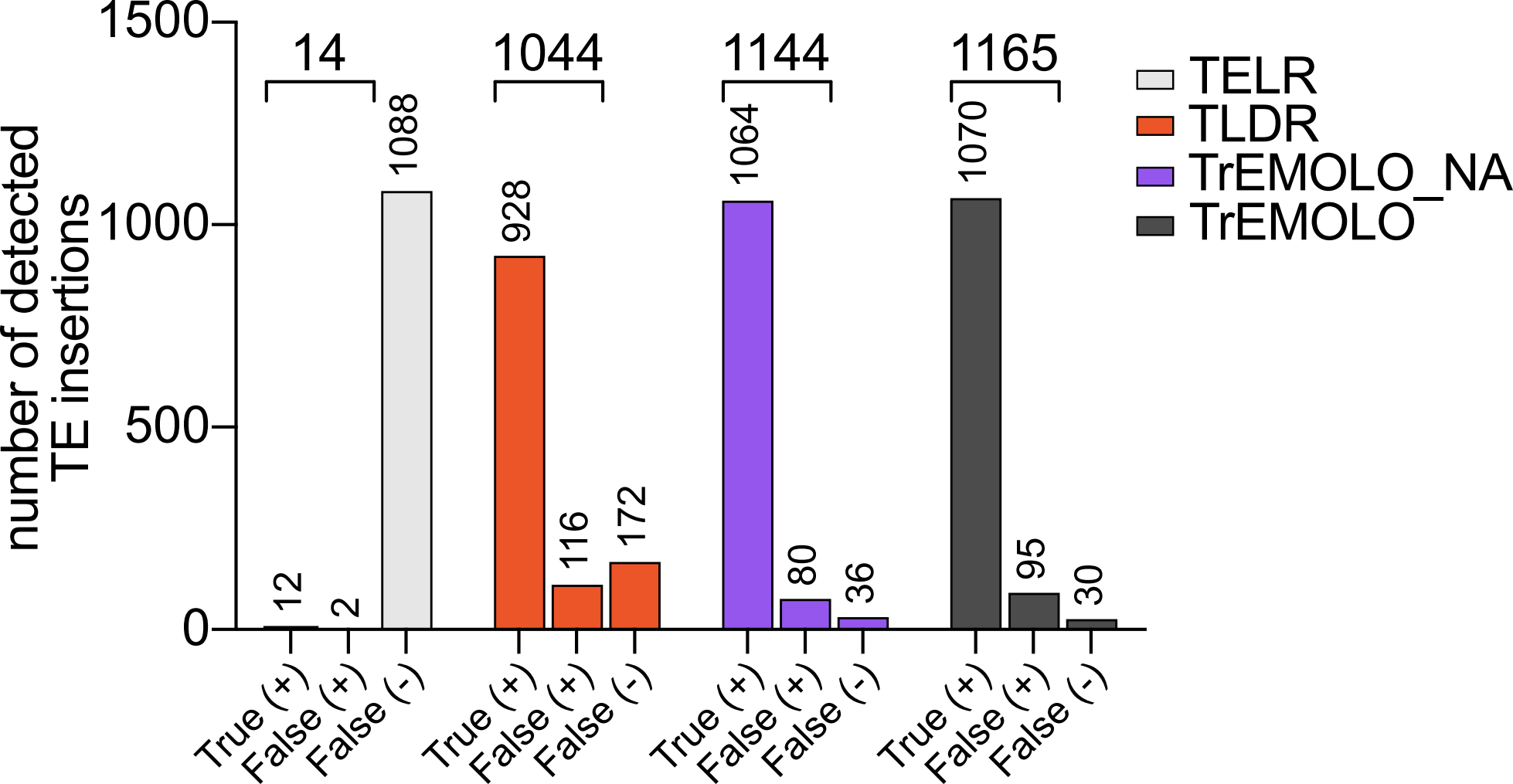
Figure 5. Example of primer pair and probe design for ddPCR quantification of the "empty" (no TE insertion) or the "full" (TE insertion) haplotype. Primers and probes for each tested TE are listed in Table S2.

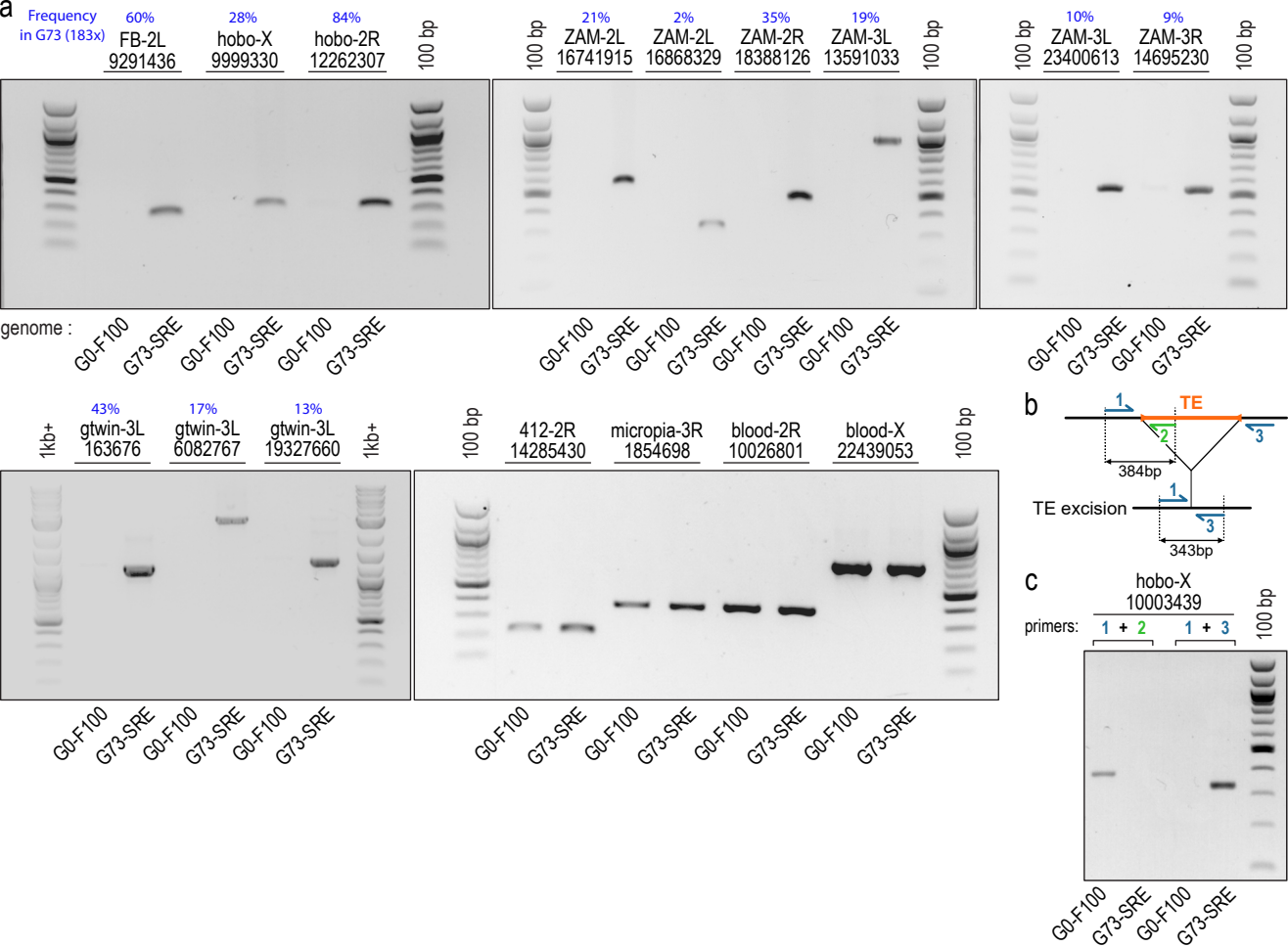
Figure 6. OUTSIDER TEs detection depends on depth of sequencing.

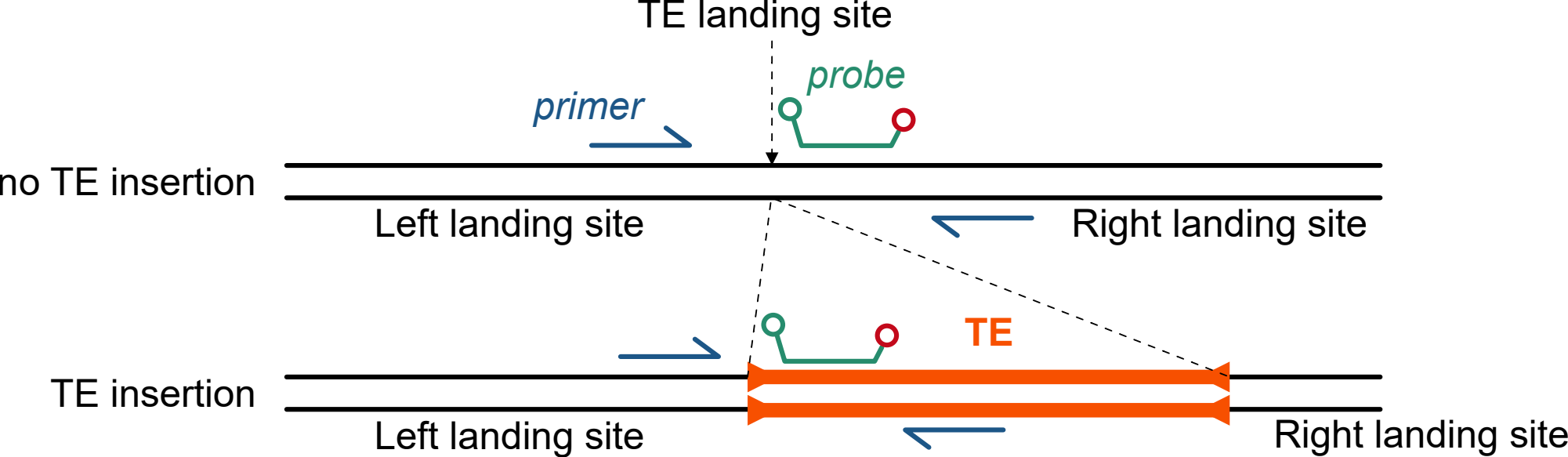
G73 sequencing reads averaging a 183x depth were subsampled and the OUTSIDER TrEMOLO module was run to detect OUTSIDER TE insertions at each depth.

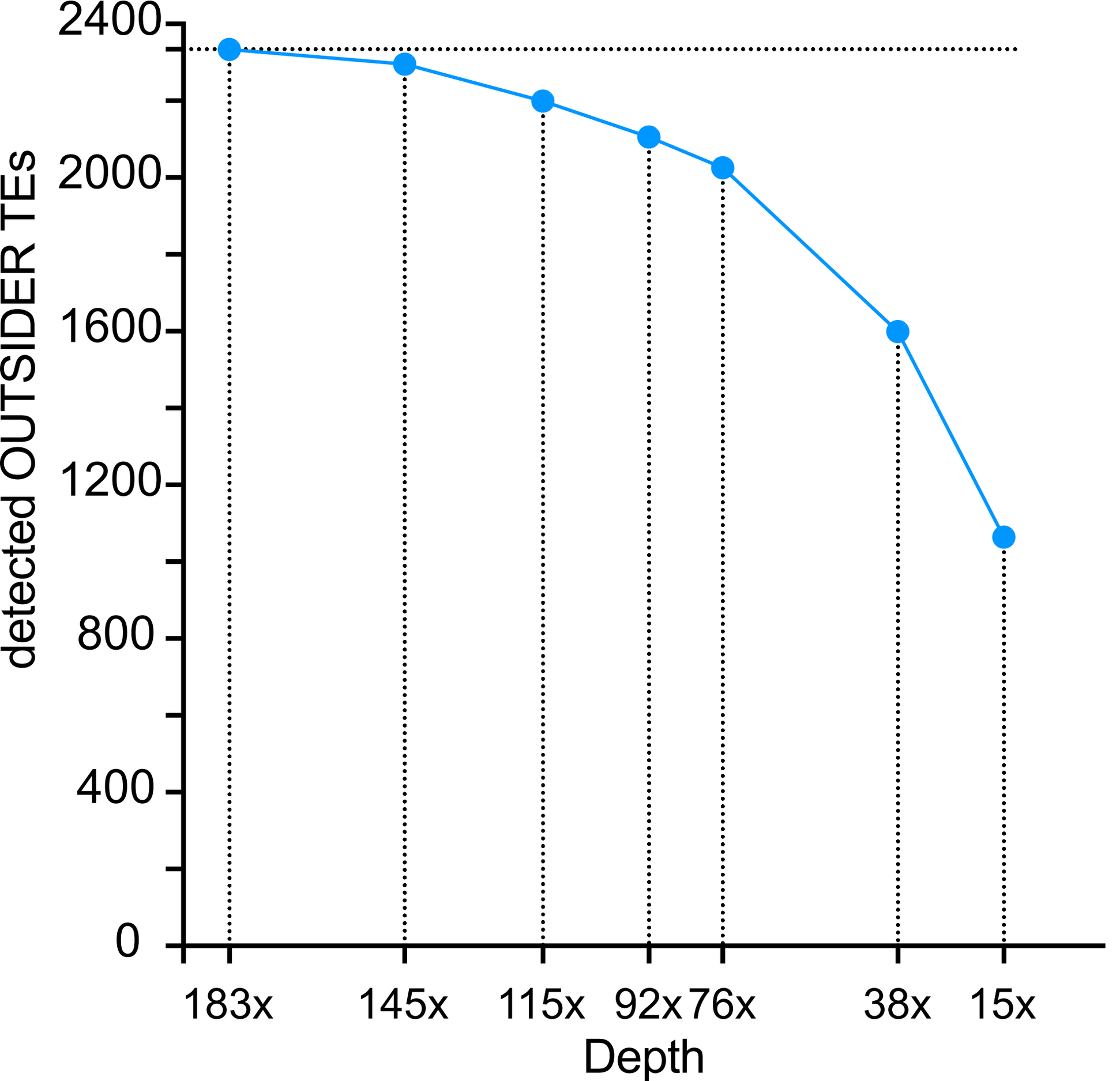
Figure 7. TrEMOLO frequency estimation variation of the 1064 common OUTSIDER TEs using datasets of different depths. Violin plot of the difference (positive or negative ratio) between the frequency computed for each common insertion in a given dataset (different sequencing depths) and the frequency of the same OUTSIDER TE insertion computed in the total dataset.

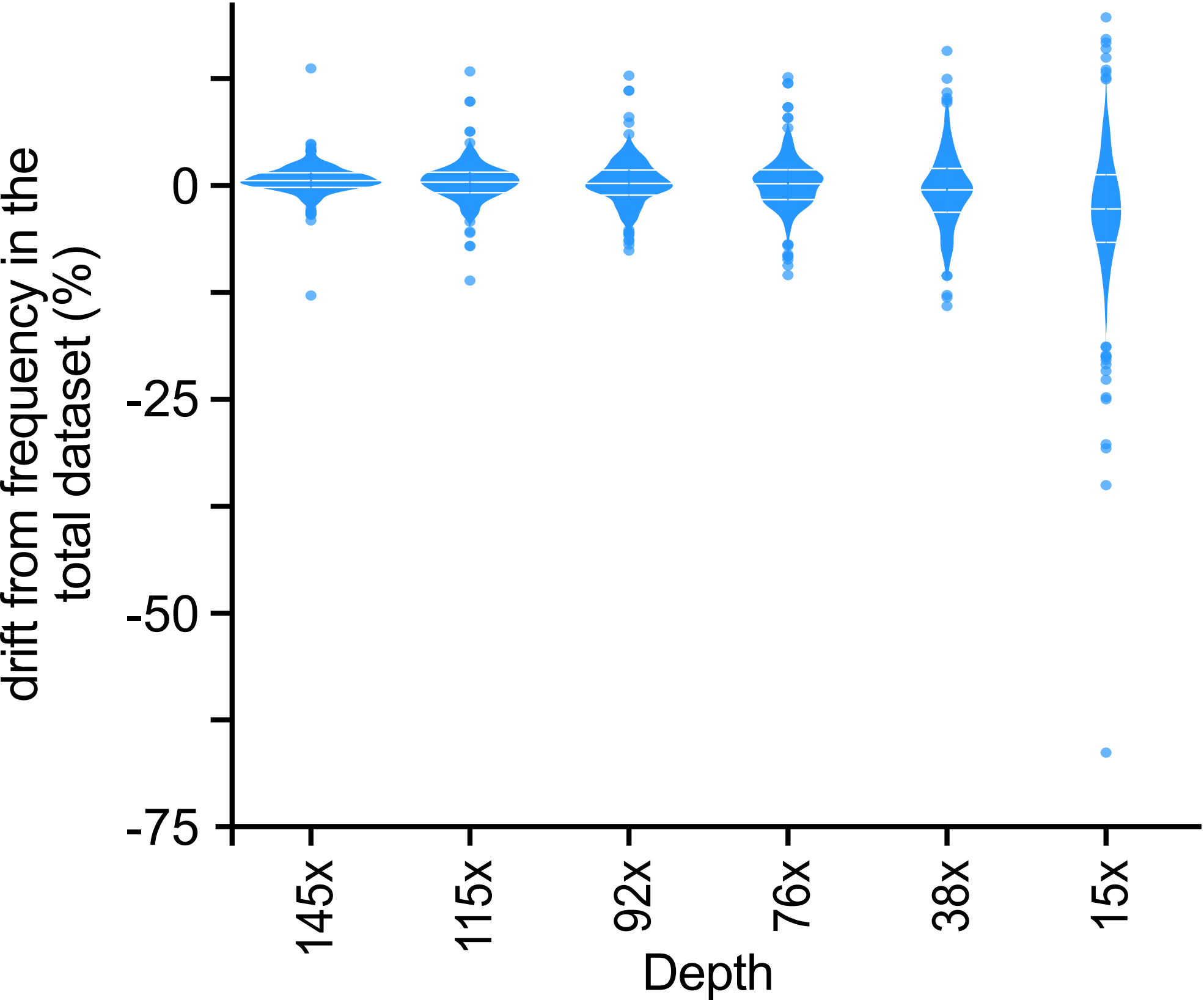

INPUT DATA
TE library
(fasta)
Genome
Reference
(fasta)
Assembly
(fasta)
Long-reads (LR)
(fastq)
INSIDER TE detection module
TE variant between TWO genomes
Reference
TE
Assembly
TE
deletion
insertion
INSIDER TE Indel calls
OUTSIDER TE detection module
TE variant between LR and genome
Low frequency TE insertion
Assembly
TE
long read mapping
OUTSIDER TE insertion calls
TE analysis module : TSD (txt), TE allelic frequencies (csv)
Calls for all TEs, TSD, TE allelic frequencies
OUTPUT DATA

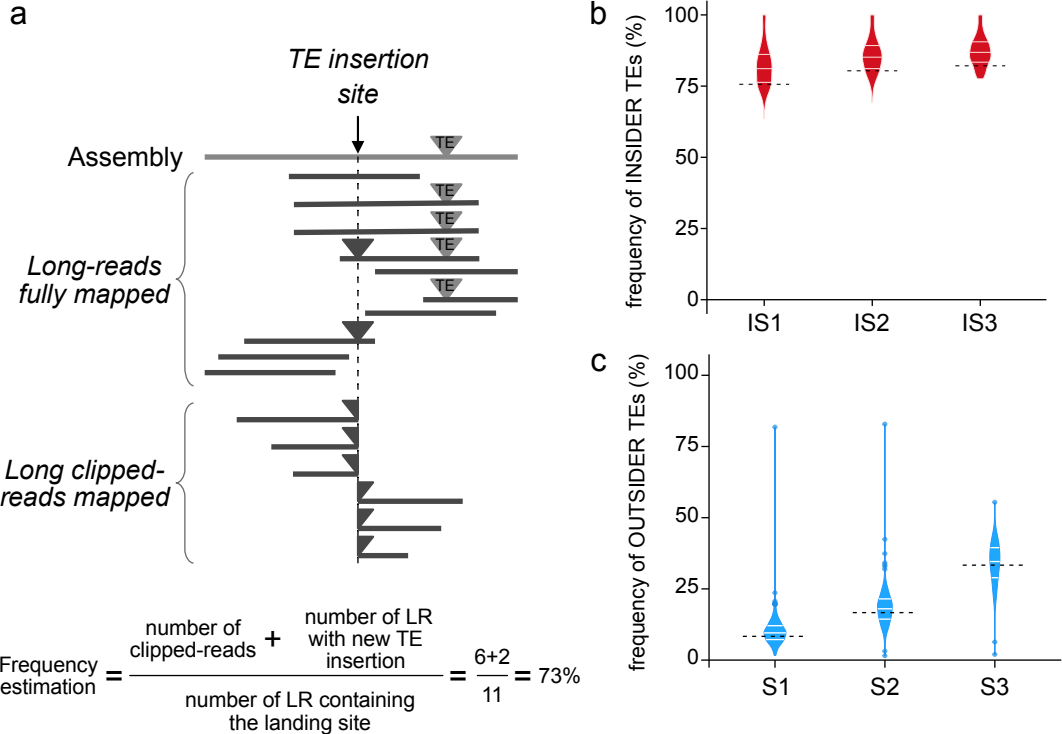

number of detected TE insertions
1500
1000
500
0
14
1044
1144
1165
12
2
1088
928
116
172
1064
80
36
1070
95
30
True (+)
False (+)
False (-)
TELR
TLDR
TrEMOLO_NA
TrEMOLO


a
Frequency in G73 (183x)
60%
FB-2L
9291436
28%
hobo-X
9999330
84%
hobo-2R
12262307
100 bp
21%
ZAM-2L
16741915
2%
ZAM-2L
16868329
35%
ZAM-2R
18388126
19%
ZAM-3L
13591033
10%
ZAM-3L
23400613
9%
ZAM-3R
14695230
genome :
G0-F100
G73-SRE
1kb+
43%
gtwin-3L
163676
17%
gtwin-3L
6082767
13%
gtwin-3L
19327660
412-2R
14285430
micropia-3R
1854698
blood-2R
10026801
blood-X
22439053
b
1
TE
2
3
384bp
TE excision
343bp
c
hobo-X
10003439
primers:
1 + 2
1 + 3


TE landing site
probe
primer
no TE insertion
Left landing site
Right landing site
TE
TE insertion
Left landing site
Right landing site


2400
2000
1600
1200
800
400
0
detected OUTSIDER TEs
183x
145x
115x
92x
76x
38x
15x
Depth


drift from frequency in the total dataset (%)
0
-25
-50
-75
145x
115x
92x
76x
38x
15x
Depth


a
TE insertion site
Assembly
TE
Long-reads fully mapped
Long clipped-reads mapped
Frequency estimation = (number of clipped-reads + number of LR with new TE insertion) / number of LR containing the landing site = (6+2)/11 = 73%
b
frequency of INSIDER TEs (%)
100
75
50
25
0
IS1
IS2
IS3
c
frequency of OUTSIDER TEs (%)
100
75
50
25
0
S1
S2
S3