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

Mohamed Mourdas [(1)]*, Sabot François[(2,3)]*, Varoqui Marion[(1)] , Mugat Bruno[(1)], Audouin Katell[(4)], Pélisson Alain[(1)], Fiston-Lavier Anna-Sophie[(5,6)]°, Chambeyron Séverine[(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 estimate their allele frequency. TE detection and frequency estimation by TrEMOLO were validated using simulated and experimental datasets. Benchmarking with simulated data revealed that TrEMOLO outperforms other computational tools. Therefore, TrEMOLO is a comprehensive and suitable tool to accurately study TE dynamics. TrEMOLO is available under a GPL3.0 at https://github.com/DrosophilaGenomeEvolution/TrEMOLO.

Keywords

Bioinformatics software, transposable element, long-read DNA sequencing

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 for MITEs (2) to more than 20kb (3,4). In large genomes, they represent more than 90% of nuclear DNA (5), and they are a major component also of smaller genomes. For instance, the *Drosophila melanogaster* genome is only 180Mb long, but almost ~35% of it is made of TEs (6). However, their identification, annotation, and dynamic tracking are complex due to their intrinsic repetitive nature. Indeed, identifying insertions in a newly assembled genome is a hard and complex task. Many tools have been developed to tackle this issue (*e.g.*, REPET (7), RepeatMasker (8), EDTA (9)). Moreover, tracking the presence or absence of a given insertion among individuals or through generations was, up to recently, limited to heavy wet-laboratory approaches, with a limited throughput (10–12).

Sequencing technology advances in the last 15 years, especially Illumina Short Read (SR) sequencing, have allowed sequencing, on a daily basis, not only individuals within a species or subspecies, but also whole populations. Currently, the same population can be resequenced more than once to identify changes across generations (13). One of the main findings provided by the availability of massive sequencing data is that TEs are the most variable component of genomes, even more than previously suspected (14). Therefore, many tools have been developed to use SR sequencing data to identify potential TE polymorphisms among samples (*e.g.,* T-lex (15–17); McClintock (18)) or within a population (*e.g.,* DNApipeTE (19); PoPoolationTE2 (20)). They rely on three main mapping features: split reads, unpaired reads, and depth of coverage (14). However, SR sequencing data are not suitable for the optimal detection of structural variations (SV), especially SV due to TE insertion/deletion. Indeed, as SR

are shorter (100 to 250 bases) than the mean length of repeats and TEs, they cannot be used to correctly identify SV, leading to high false-positive rates (40%) and false-negative results (21,22). Conversely, Long Read (LR) sequencing technologies, such as Pacific Biosciences and Oxford Nanopore Technologies (ONT), allow obtaining reads >20kb long (and even 100kb and more for ONT). Therefore, a single read may contain a complete SV or TE insertion and enough surrounding sequence data to unambiguously anchor the variation. However, the longer read length is counterbalanced by a lower sequencing quality (still improving) compared with SR sequencing and fewer dedicated tools for TE detection. The current methods to detect TE variations using LR sequencing data (and also most of the SR sequencing data-based tools) generally rely on the presence/absence of TE insertions annotated in a reference genome (*e.g.,* LorTE (23)), or on the detection of reads that abnormally map to the reference genome and that can be linked to a TE (*e.g.,* TELR (24)). Finally, non-referenced TE insertions can be detected, assembled, and annotated using the Transposons from Long DNA Reads (TLDR) software (25).

In addition, unlike SR, LR sequencing allow generating high-quality genome assemblies with higher levels of contiguity, and consequently to compare genomes directly and to extract more efficiently large and complex SV. Indeed, the number of available almost complete genomes is increasing due to these LR sequencing technologies (13,26), and even telomere-to-telomere assemblies can be obtained (27). Moreover, some reliable tools can be used for genome-to-genome comparisons (MUMmer (28), minimap2 (29)) 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 will 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) variations at the global and population level, for TE detection and frequency estimation. TrEMOLO (available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO) is a SnakeMake pipeline that can be used as it is, or with Singularity to ensure higher reproducibility, and to facilitate future updates. This pipeline is designed to be flexible, thus simplifying the streamlined addition of new samples. We demonstrated its efficiency using simulated LR sequencing data 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 offers a more precise picture of TE variability compared with other tools.

Results

TrEMOLO design

TrEMOLO is a suite of Python 3 (30) scripts that follow the SnakeMake (31) rules. It takes as input a reference genome sequence, single or pooled LR DNA sequencing data, and a genome assembly generated from these sequencing data. The complete tool includes three modules (Fig.1 and Fig. S1): (1) the ***INSIDER TE detection module*** identifies TE insertions and deletions present in the newly assembled genomescompared with the reference by aligning the two genomes and parsing the pairwise alignment; (2) the ***OUTSIDER TE detection module*** looks for TE insertions that have not been incorporated into the genome assembly (second or minor haplotype). In this module, LR are mapped to the genome assembly generated using the same data, and then TE variants are detected by parsing the mapping results; (3) the ***TE analysis module*** computes the frequency of each TE variant using the LR sequencing data and looks for target site duplications (TSDs) that occur upon TE integration as a result of staggered DNA

double-strand breaks (32). A detailed manual on how to run TrEMOLO with all its options is available at TrEMOLO github (https://github.com/DrosophilaGenomeEvolution/TrEMOLO).

The INSIDER TE detection module. This module takes as input a reference genome and the genome assembly of the studied biological sample. This assembly must be generated with the input LR sequencing data to be compatible with the second module. Pairwise alignment is performed between the reference and assembled genome using minimap2 (29), and is then parsed with *Assemblytics* (33) for SV identification (Fig. S1). Before providing the final list of detected TE insertions, the alignment data are scanned to filter out low-quality TE detections (Methods). This analysis returns a list of SVs from the pairwise alignment of the assembly against the reference genome with their positions. This list is then compared with a TE library to keep only insertions belonging to already known TE families (Fig. 1, left panel, and Fig. S1). This module detects the so-called INSIDER TEs. These are the non-reference TE insertions that are found frequently enough among the reads incorporated in the assembly (generally TEs present in the major haplotype of the sample).

***The OUTSIDER TE detection module*.** This module takes as input a genome assembly and the LR sequencing data used to build this assembly. The genome assembly and reads are aligned using minimap2 (29). Reads that contain a TE insertion not retained in the genome assembly are identified as unmapped (or partially mapped) reads of the output SAM file that contains the TE sequences identified based on the TE library (Fig. 1 right panel, Fig. S1). As done in the INSIDER module, low-quality TEs are filtered out and then, TE insertions are listed in a BED file. This module allows defining the so-called OUTSIDER TEs. These TEs were not included in the assembly because of their lower frequency compared with the non-insertion haplotype.

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

(a) For the INSIDER TE insertions, the two 30 bp-flanking regions of the TE insertion are aligned using the Boyer-Moore exact algorithm (34) to detect short motifs shared on each side of the insertion without mismatches. For the OUTSIDER TE insertions, the most informative single LR (the one with the highest TE identity compared with the canonical TE sequence and the longest TE sequence) spanning the TE insertion is selected. Then, like for INSIDER TE insertions, the sequences flanking the TE insertion site are extracted and aligned. However, to handle the LR error rate, the algorithm can accept mismatches (default 1). For both INSIDER and OUTSIDER TE insertions, the candidate TSDs are then filtered out if the detected shared motifs are not at the insertion junction. Moreover, for OUTSIDER TE insertions, candidates where the single-copy target site is not present in the genome assembly are removed. This step also allows better defining the extremities of the inserted TE.

(b) To compute the insertion frequency, the number of reads showing the TE insertion (including clipped reads for the OUTSIDER TE insertions, see below) is divided by the total number of reads at the same position, with or without TE.

OUTPUT data. A BED file that contains the newly inserted TE positions, TSD sequences, and frequencies is generated based on the assembled genome coordinates. TrEMOLO also lists the putative INSIDER TE deletion events, and integrates the OUTSIDER TE insertions in the genome assembly to construct a pseudo-genome fasta sequence.

TrEMOLO benchmarking using simulated datasets

To assess TrEMOLO, we compared the results obtained with TrEMOLO and with TLDR (25) and TELR (24), two recently developed computational tools dedicated to TE detection using LR

sequencing data. The *ab initio* approach of these tools is similar to the one used with SR sequencing data: they use LR data (mapped to a reference sequence, through a BAM file for TLDR (25) or directly from the mapping tool output for TELR. Insertions are detected based on discordant mappings (*i.e.* clipped reads or partially mapped reads or both). Finally, like for SR-based approaches, the non-mapped parts of the reads are aligned against a library of canonical TE sequences. Then, discordant mapped reads associated with TE matches are selected, clustered, and analyzed to identify the TE insertion site in the reference genome.

We generated a simulated dataset in which 1,100 TE insertions with different allele frequencies (simulated sample S1; see Methods) were added to a reference *D. melanogaster* genome called G0-F100 (13). We mapped the simulated reads to this reference genome using TLDR (25), TELR (24), and the TrEMOLO OUTSIDER module with minimap2 (29) (here called “TrEMOLO_NA”, see Methods). We also launched the TrEMOLO software (OUTSIDER and INSIDER modules) with the G0-F100 reference genome. For the OUTSIDER module, as a proxy of the assembly, we used this reference genome that contained 100 highly frequent TEs (segregating at 75% in the simulated reads). As TLDR parameters are fixed (25), we decided to use the same parameters for the four methods/tools (TrEMOLO, TrEMOLO_NA, TLDR, and TELR): 200 bp as minimum TE insertion length and 80% as minimum TE sequence identity. To limit false-positive TE detections, we also set at 15kb the largest TE insertion length.

Using these simulated datasets, TrEMOLO and TrEMOLO_NA detected more TE insertions (949 and 925 among the 1,100 insertions, respectively) than the other two tools (866 and 14 for TLDR and TELR, respectively). Moreover, sensitivity was higher for TrEMOLO and TrEMOLO_NA (84.6% and 83.3%, respectively): 931 and 917 true-positive results compared with 813 (73.9%) and 12 (1%) for TLDR and TELR, respectively. TrEMOLO and TrEMOLO_NA were also the most specific: 18 and 8 false-positive results compared with 53 and 2 for TDLR and TELR, respectively

(Fig. 2). The difference between TrEMOLO and TrEMOLO_NA was relevant, and it might more important if a more different reference sequence than the tested sample were used.

TrEMOLO experimental validation

To validate TrEMOLO detection accuracy with real biological samples, we took advantage of a previously described *D. melanogaster* laboratory line (35) in which the knockdown of Piwi, a protein essential for TE silencing, can be induced in adult follicle cells (somatic cells of the ovary). This transient somatic Piwi knockdown allows TE derepression in follicle cells and their integration as new proviruses in the progeny genome (35). We performed TE derepression for 73 successive generations to obtain the G73 line that contains thousands of newly integrated TEs in its genome (13,35). We used genomic DNA extracted from G0 (before Piwi knockdown) and G73 fly samples for LR DNA sequencing and TE detection by. Here, we validated by genomic PCR some of the newly integrated TEs detected by TrEMOLO in the G73 samples, compared with G0 samples. Genomic PCR yielded amplicons of the expected sizes for the nine randomly tested OUTSIDER TEs, detected in a minority of the support reads by TrEMOLO, but not present in the G0 reference genome (Fig. 3 upper panel). Similarly, we detected three INSIDER TEs, identified by TrEMOLO in the majority of the support reads, specifically in G73 but not G0 genomic DNA samples (Fig. 3 lower left panel). To control the sample quality, we also checked the presence of four TEs shared by the G0 and G73 genomes (Fig. 3, lower right panel).

Impact of the genome assembly quality on TrEMOLO efficiency

To assess the impact of the genome assembly quality on TE detection, we ran TrEMOLO using the G73 genome assembled with different tools and after different polishing rounds, using the G0-

F100 assembly as reference. We used Flye (36), Shasta (37) and Wtdbg2 (38) followed by Racon (39) polishing steps and scaffolding by RaGOO (40)(Methods). As previously reported (13), four rounds of Racon polishing (39) increased the BUSCO score (Table 1), suggesting a better-quality assembly.

Table 1: Comparison of the TE detection efficiency by TrEMOLO using genome assemblies of different quality.

Assembly process	Contig / scaffold number	N50 (bp)	BUSCO score	Assembly total size (bp)	INSIDER TE number	OUTSIDER TE number	TrEMOLO results (novel TE insertions, number)
Flye + Racon (4 rounds) + Ragoo*	30	27,282,734	98%	148,869,489	27	1,038	1,065
Flye + Racon (1 round) + Ragoo*	31	27,316,827	96.9%	149,145,078	29	1,100	1,129
Shasta + Ragoo*	15	27,256,021	93%	151,203,880	59	1,016	1,075
Wtdbg2 + Ragoo*	22	34,296,233	92.6%	171,989,387	68	442	505

* RaGOO was run with default parameters.

Using genomes assembled with the recently updated Flye and Shasta assemblers led to the detection of similar numbers of INSIDER and OUTSIDER TEs, with higher numbers of INSIDER TEs for the Shasta-assembled genome (Table 1). Conversely, when using a genome assembly generated with Wtdbg2, which has not been updated for two years and thus has not incorporates the recent changes in ONT sequencing data (*e.g.*, new base calling inducing a much lower error rate), TrEMOLO detected many INSIDER TE insertions (among which many were false positives), but only ~45% of OUTSIDER TE insertions. Altogether, we conclude that a draft genome assembly, obtained with a recent assembler version, is of enough quality to allow the identification of almost all TE movements by TrEMOLO.

Validation of TE frequency estimated by TrEMOLO using simulated data

To validate the accuracy of TrEMOLO frequency estimation, we used simulated datasets with known frequencies of INSIDER and OUTSIDER TEs (see Methods). For OUTSIDER TEs, we wanted also to determine whether taking into account clipped reads would improve the accuracy of TrEMOLO frequency calculation (Fig. 4a). We first observed that the TrEMOLO frequency module slightly overestimated the frequency of INSIDER TEs in the simulated datasets (Fig. 4b). Moreover, estimation of the OUTSIDER TE frequency was very accurate only when the clipped reads present in the samples were taken into account (Fig. 4c). Therefore, we decided to keep this parameter for the next frequency computation.

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

To validate the TrEMOLO frequency module accuracy, we extracted again long DNA molecule from G73 flies (Methods) and analyzed the extracted DNA by ONT sequencing and droplet digital polymerase chain reaction (ddPCR). This latter approach allows the absolute quantification using forward primers located in the left side of the TE landing site and reverse primers located either

in the TE or on the right side of the landing site (Fig. 5). To increase the detection specificity, we also designed fluorescent probes. The primer and probe sequences are listed in Table S2. We extracted these sequences from the TrEMOLO analysis of the G73 genome. We selected nine of the thousand TE insertions, as representatives of different TEs and frequencies. For three of them, the low frequencies estimated in the first library were confirmed by the second library, as indicated by the low percentages obtained by ddPCR (0.6, 2.2, and 9.6%) and by the absence of supporting reads in the G73-SRE library (see Methods). For the other six, the ddPCR estimates fitted very well with the TrEMOLO estimates, whatever their frequency.

Table 2: TE frequency estimates using ddPCR and TrEMOLO.

TE insertion	ddPCR frequency estimation (%)	TreMOLO frequency estimation (%)
412-2R-14285430	82.6	93.44
blood-2L 18128567	2.2	NA
ZAM-3L_13591033	0.6	NA
ZAM-2L_16741915	9.6	NA
ZAM-2L_16868329	13.6	12.24 (18.09)
ZAM-2R_18388126	17.3	12.16 (18.31)
gtwin-3L_163676	26.5	11.25 (25.68)
gtwin-3L_6082767	10.4	3.61 (6.67)

roo-2L 13882751	59.2	65.22

NA = TE insertions with not enough reads in G73-SRE to be detected by TrEMOLO (percentages were calculated by taking into account the clipped read information and the fully mapped reads).

Impact of sequencing depth on TE identification and frequency estimation

Using the initial G73 dataset, we generated different sequence subsets using Filtlong (41) (see Methods), resulting in seven “samples” of reads for the same DNA with 24 (i.e. original dataset, 183x; G73), 19 (G73_F19), 15 (G73_F15), 12 (G73_F12), 10 (G73_F10), 5 (G73_F5), and 2 (20x; (G73_F2)) Gbases of data. We then launched TrEMOLO with these datasets, and compared the number of OUTSIDER TEs detected by mapping these reads to the G73 assembly (Table 3).

Table 3: Impact of coverage depth on the detection of OUTSIDER TE insertions

Subsampling	G73	G73_F19	G73_F15	G73_F12	G73_F10	G73_F5	G73_F2
Gbases (% of total)	24 (100%)	19 (79%)	15 (62%)	12 (50%)	10 (42%)	5 (21%)	2 (8%)
Coverage depth	183x	145x	115x	92x	76x	38x	20x
Detected OUTSIDER TEs (% of the original set)	1038 (100%)	1024 (99%)	972 (94%)	940 (91%)	906 (87%)	713 (69%)	475 (46%)

Using the initial dataset (total read depth = 183x), TrEMOLO detected 1,038 OUTSIDER TEs. The percentage of detected OUTSIDERs is relative to this dataset.

The depth of coverage started to have a major impact at around 76x, with only 87% of OUTSIDER TEs recovered (Table 3). Decreasing the coverage depth from 183x to 92x reduced by only ~10% the number of OUTSIDER TEs detected. Below 76x of coverage depth, TrEMOLO detection performance decreased more rapidly (69% and 46% of detected TE insertions with 38x and 20x depth). Nevertheless, even with a quite low coverage (20x), TrEMOLO could still identify 46% of OUTSIDER TEs.

Among the 1,038 original OUTSIDER TE insertions, only 465 were detected in all subsamples. We used these 465 TEs to compute the TE frequencies in each sub-dataset in order to estimate, the frequency computation bias linked to the sequencing depth.

The accuracy of OUTSIDER TE frequency estimation was acceptable when using datasets with sequencing depth ≥76x (DNA extracted from ~ 100 flies, *i.e.,* 200 haploid genomic samples), with a decrease of 12% maximum compared with the initial frequency computed at 183x (Fig. 6). However, for very “low depth” datasets, compared with the number of samples in the population (here 38x and less, meaning 0.19x per sample or less), the accuracy frequency decreased quite rapidly with an almost systematic underestimation by up to 50% compared with the original dataset. Thus, we recommend a sequencing depth corresponding to a minimum of 1x per haplotype in case of polymorphic populations.

Discussion

Accurate detection of TEs with different frequencies in a population or in a single sample at the somatic level is important for researchers in evolution, ecology, agronomy, and medicine. However, it is still a challenge due TE repetitive nature. Most of the current approaches to identify TE insertions are based on SR mapping to a reference genome, sometimes quite distant from the studied samples. Working with distantly related genomes may drastically bias the mapping results and thus the subsequent analyses, for instance by providing high false positive rates (42). Therefore, TE detection tools based on mapping approaches allow the identification of low-frequency TE insertions (or from the non-major haplotype), but not all of them, only when using very high-depth SR sequencing data. The availability of low-cost LR sequencing technologies (ONT) open the possibility to accurately detect new TE insertions. Improving the sequencing quality allows obtaining accurate and (almost) complete genome assemblies in a short amount of time. For instance, a high-quality *Drosophila* genome can be assembled in 24 hours. However, most of these assemblers provide a single haplotype sequence as output, generally the most prevalent. Finally, most TE detection tools rely on the identification of TE fragments as source of SV *ab initio*.

We designed TrEMOLO, a comprehensive computational tool with the aim of combining the main haplotype information (assembly step) followed by the recovery of minor haplotypes by LR sequence mapping to the newly assembled genome to detect all SV types without *a priori.* It is only after this naive detection that the SV are annotated as possible TEs or not. This approach allows to run TrEMOLO using different TE databases and also to detect new TE families. TrEMOLO accuracy in TE identification and the correction system used for TSD detection allow predicting the insertion site within a 2-base pair window. Therefore, TrEMOLO is currently one of the most efficient tools for TE detection.

Moreover, TrEMOLO allows the estimation of the insertion frequency for each detected TE copy with high accuracy (when using data with sequencing depth >50x). Importantly, TrEMOLO

performance is weakly influenced by the assembly quality and sequencing depth variations (up to a certain point). Therefore, it can be used for different projects (e.g. pooled DNA sequencing, population analysis, human genetics, tumor analysis) that require the fast and precise identification of TE insertions, even at very low frequency within a sample. Indeed, our data showed that a single insertion can be unambiguously identified through a unique read, as long as the read is long enough to harbor the whole TE and its surrounding insertion site.

Conclusions

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

Methods

TrEMOLO pipeline preprocessing

INSIDER TE identification: to compare the reference genome and a new assembly, TrEMOLO uses the same methodology as RaGOO (v1.1; 22) up to the SV identification step. Then, SV are compared to the TE database (Bergman's laboratory, https://github.com/bergmanlab/transposons) with BLASTN (2.9.0) (--outfmt 6) (43). The output file is filtered based on the size and identity percentages listed in parameters. DNA fragments that appear to be part of the same sequence are grouped. For INSERTION, Repeat_expansion, Tandem_expansion, sequences are retrieved from the query (new assembly). For DELETION, Repeat_contraction, Tandem_contraction, sequences are recovered from the reference (Fig. S1)

OUTSIDER TE identification: the LRs and the genome assembled from them are aligned using minimap2 (29) and SVs are identified using Sniffles (44) / SVIM (45). For Sniffles, the reads covering the SV are recovered from the SAM/BAM file using as anchor point the Sniffles provided position and a 30bp-window with SAMtools (46). For SVIM, only the support reads provided by SVIM are recovered. The SV reads and sequences are compared to the TE database, and filtered as done for INSIDER TEs, but only the best support reads are kept among those provided by Sniffles/SVIM or the SAM parsing (Fig. S1).

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

TE frequency estimation: two different approaches are used to estimate the INSIDER and OUTSIDER TE frequencies in the samples. For INSIDER TEs, from the original BAM file only reads aligned to the putative TE positions (with a 100bp window 5' and 3') are retained using SAMtools view and the -F 2048 -L option. Then, clipped reads are removed to keep only the reads aligned on the putative TE position within a 5bp window (5' and 3'). The mean depth of 30bp at the 5' and 3' of the putative insertion is computed with SAMtools depth. Then, the TE frequency is computed as (number of reads showing the insertion)/(total number of reads).

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

Simulated datasets

A previously assembled reference *D. melanogaster* genome, called G0-F100 (13) was used for the LR simulation by DeepSimulator (version v1.5) (47) (-B 2 -K 10 -l 20 000) with a depth of 10X (subsample 1 (sb1)). In parallel, two fake genomes containing INSIDER TEs only or both INSIDER TEs and OUTSIDER TEs were built as follows. For INSIDER TEs, 100 TE sequences were artificially and randomly inserted in the G0-F100 assembly genome. They correspond to the canonical sequences of 11 TE families of different lengths (from one to 13kb), each one providing five (ZAM and gtwin) or ten copies (Doc, roo, copia, F-element, hopper, Helena, HB, TART-A and gypsy). Then, DeepSimulator (version v1.5) (47)(-B 2 -K n -l 20000) was used to simulate LRs with a mean length of 20kb and depth of 30, 40 and 45x (subsamples 2 (sb2)). For OUTSIDER TEs, 1000 TE sequences belonging to the same families as the simulated INSIDER TEs and also

to the Idefix family (200, 200, 50, 50, 50, 50, 50, 50, 50, 50, 100, and 100 copies, respectively) were inserted randomly in the G0-F100 genome assembly containing the INSIDER TEs. Then, DeepSimulator (version v1.5; 28) (47) (-B 2 -K **n** -l 20000) was used to simulate LRs with a mean length of 20kb mean length and depth of 5, 10 and 20X (subsamples 3 (sb3)).

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

For the INSIDER TE frequency estimation, the sb1 and sb2 subsamples were used together with a new sb4 subsample simulated with a depth of 1X and minimum to maximum length of 10 to 50kb to obtain IS1 (10X sb1 + 45X sb2 + 1X sb4), IS2 (10X sb1 + 40X sb2 + 1X sb4) and IS3 (10X sb1 + 30X sb2 + 1X sb4).

Tools comparison

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

Genome assembly

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

Datasets

The *D. melanogaster* reference genome (dm6) and annotations (Release 32) were downloaded from FlyBase (https://flybase.org/). The LR dataset of the *D. melanogaster* laboratory G0-F100 and G73 lines previously published (Mohamed et al., 2020) is available at ENA (https://www.ebi.ac.uk/ena) under the accession number ERP122844.

Benchmark of *Drosophila* genome assemblies

Genome assemblies were performed using the raw LR data and CulebrONT for assembly and polishing (48). Flye v2.8 (36) was used with standard parameters except for the --nano-raw option. One to four rounds of Racon v1.4.3 with minimap2 v2.17 (29) were performed using the standard conditions for ONT data, and with medaka v1.2 (standard conditions). Scaffolding was performed using RaGOO v1.1 (40) with standard options except for output structural variations with the -s option and the corresponding reference. For raw assemblies, wtdbg2 v2.5 (38) and shasta v0.1.0 (37) were used (standard conditions).

Robustness of the frequency computation dataset

Data from the G73 assembly (183x; Mohamed et al, 2020) were degraded in different datasets using FiltLong (41) and the following options: -t 19000000000 --min_length 6000, -t 15000000000, -t 12000000000, -t 10000000000, -t 5000000000, -t 2000000000).

Genomic DNA extraction for G73-SRE LR sequencing

Genomic DNA was extracted from 50 mg of G0 or G73 male flies using the Nanobind Tissue Big DNA Kit (Circulomics SKU NB-900-701-01) with Insect buffer (PL1) as lysis buffer. Briefly, samples were homogenized in 200 µl CT buffer with a pellet pestle, centrifuged at 16,000 x 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 2min. Pellets were resuspended in 200µl PL1 lysis buffer (Circulomics Aux. Kit) supplemented with 20µl proteinase K and incubated in a thermomixer at 55°C and 900 rpm for 1 h. 20µl RNase A was

added and incubation was continued for 15 min. Samples were centrifuged at 16,000 x g for 5 min, and supernatants passed through a 70µm filter and collected in a 1.5 ml DNA LoBind microcentrifuge tube. 50 µl BL3, a Nanobind disk and 400 µl isopropanol were added to the samples. Incubation and washes were done according to the manufacturers' instructions. Elution was performed with 50 µl EB buffer. The genomic DNA concentration was estimated with a Nanodrop spectrophotometer (Thermo Fisher Scientific) and a Qubit Fluorometer with the Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific), according to the manufacturer's instructions. To reduce SRs <25 kb, genomic DNA samples with concentration adjusted between 50-150 ng/µl in 60 µl were treated with the SRE Kit, 25 kb cutoff (Circulomics SKU SS-100-101-01), according to the manufacturer's instructions, and eluted in 50 µl EB.

PCR amplification of long DNA molecules

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

Library preparation for LR sequencing

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

Droplet digital PCR for TE insertion frequency determination

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

For droplet generation, 20 µl of each mixture was transferred to the center line of a DG8 droplet generator cartridge (Bio-Rad Laboratories). Then, 70 µl of droplet generation oil for the probe (Bio-Rad Laboratories) was injected in the cartridge bottom line and droplets were produced using a QX200 Droplet Generator (Bio-Rad Laboratories). Droplets (40 µl) were transferred to a 96-well plate and endpoint PCR was performed on a T100 Thermal Cycler (Bio-Rad Laboratories) using the following cycling protocol: enzyme activation at 95 °C for 10 min, followed by 40 cycles of 95 °C for 30 s and 60 °C for 1 min (ramp rate settings of 2°C/sec), followed by a droplet stabilization step at 98 °C for 10 min, and a final hold step at 4 °C. Droplet measurement was performed on a QX200 Droplet Reader (Bio-Rad Laboratories) and data analyzed with the QuantaSoft software (Bio-Rad Laboratories). Two different fluorescent signals were measured: i) amplification of the TE inserted in a given locus sequence (TE_{in}) or amplification of the same locus without TE (TE_{out}) in the FAM channel, and ii) amplification of the *RpL32* housekeeping gene in the HEX channel. The proportion of TE_{in} or TE_{out} to *Rpl32* was calculated as follows: $TE_{in}/Rpl32$ *or* $TE_{out}/Rpl32$. Each value was expressed as copies/µl. Frequency was calculated with the following formula: Frequency in %= ($TE_{in}/Rpl32$x100) / ($TE_{in}/Rpl32$ + $TE_{out}/Rpl3$).

Declarations

Author Contributions

F.S. M.M. and A.S.F.L 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.S.F.L, 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.

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.

Acknowledgments:

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.

Availability of data and materials

The LR sequencing data used for this study were deposited at ENA (https://www. ebi.ac.uk/ena) under the accession numbers (ERP138838 under validation by ENA) and TrEMOLO is available at https://github.com/DrosophilaGenomeEvolution/TrEMOLO under the GNU Public License v3.

Conflicts of Interest: The authors declare no competing interests.

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. Mob DNA. 2020 Jul 3;11(1):23.

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 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 MGG. 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 Cell Mol Biol. 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. Bioinforma Oxf Engl. 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 Res. 2011 Mar;39(6):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 Genet. 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. Mob 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 [Internet]. bioRxiv; 2022 [cited 2022 Jul 20]. p. 2022.01.04.471818. Available from: https://www.biorxiv.org/content/10.1101/2022.01.04.471818v1

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

26. 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.

27. 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.

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

29. Li H. New strategies to improve minimap2 alignment accuracy. Bioinforma Oxf Engl. 2021 Oct 8;btab705.

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

31. Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, et al. Sustainable data analysis with Snakemake. F1000Research. 2021;10:33.

32. 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.

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

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

35. 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.

36. 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.

37. 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.

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

39. 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.

40. 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 Biol. 2019 Oct 28;20(1):224.

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

42. 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.

43. 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.

44. 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.

45. Heller D, Vingron M. SVIM: structural variant identification using mapped long reads. Bioinforma Oxf Engl. 2019 Sep 1;35(17):2907–15.

46. 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.

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

48. 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. The required INPUTs are: a genome in fasta format and long-read data in fastq format to launch the OUTSIDER TE detection module. If, besides the reference genome and the LR DNA sequencing data, a genome assembly is provided, both detection modules can be started.

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 (TrEMOLO, TrEMOLO_NA, TLDR, and TELR).

Figure 3: Validation of TrEMOLO specificity for TE detection. PCR performed using long DNA molecules extracted from the G0 (control) and G73 Drosophila lines. The name of the TE identified by TrEMOLO is indicated followed by the chromosome and the genomic coordinate (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 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 frequency estimated by TrEMOLO in the simulated datasets with and without clipped reads. Dashed lines represent the real frequency in the simulated datasets. In the S1 sample, TE frequency could not be estimated without clipped reads.

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. TrEMOLO frequency estimation variation of the 465 common OUTSIDER TEs using datasets of different coverage depths. Violin plot of the difference (positive or negative) 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.

Figure S1. Schematic representation of TrEMOLO method for TE detection. The INSIDER TE module (red) detects TE insertions from genome assemblies and the OUTSIDER TE module detects TE insertions by mapping reads on these assemblies (blue).

Figure 1

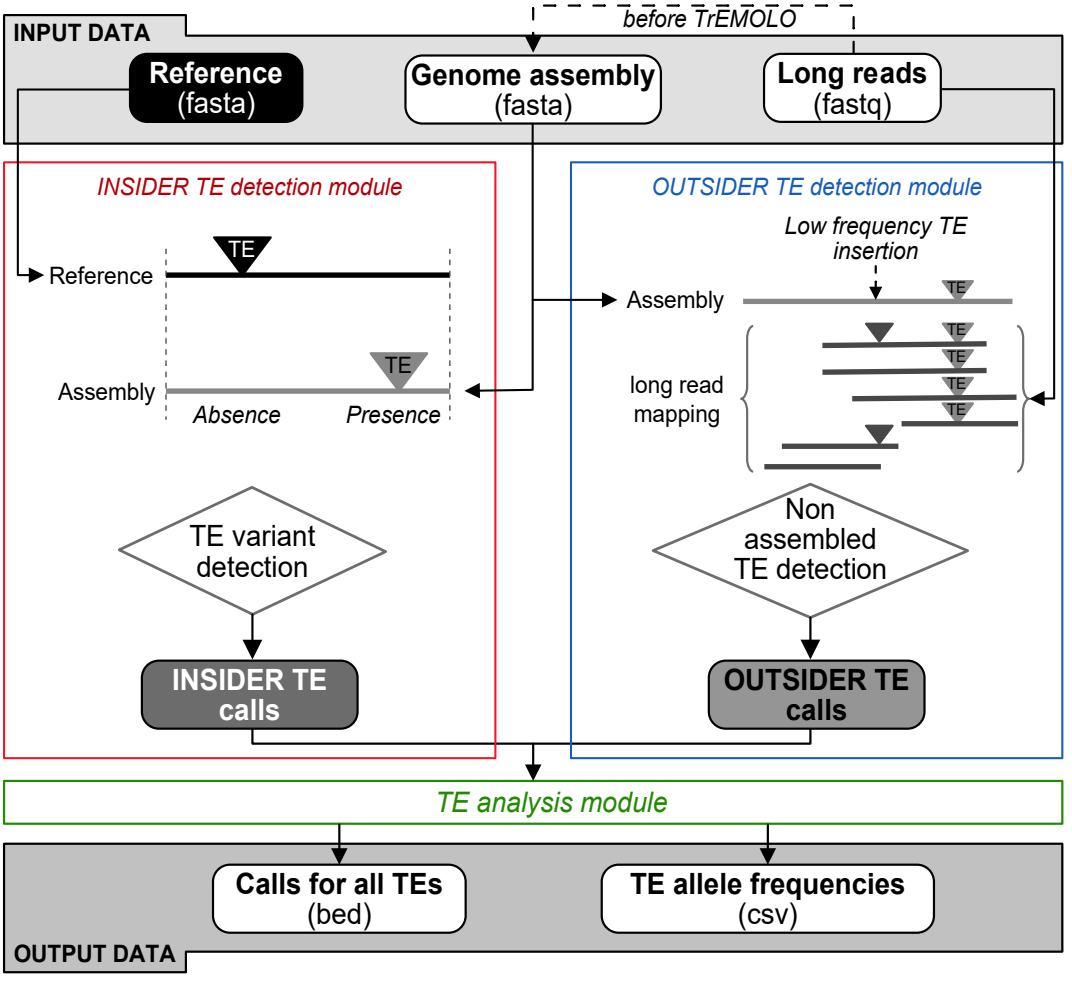

INPUT DATA
before TrEMOLO
Reference (fasta)
Genome assembly (fasta)
Long reads (fastq)
INSIDER TE detection module
Reference
TE
Assembly
TE
Absence
Presence
TE variant detection
INSIDER TE calls
OUTSIDER TE detection module
Low frequency TE insertion
Assembly
TE
long read mapping
TE
Non assembled TE detection
OUTSIDER TE calls
TE analysis module
Calls for all TEs (bed)
TE allele frequencies (csv)
OUTPUT DATA

Figure 2

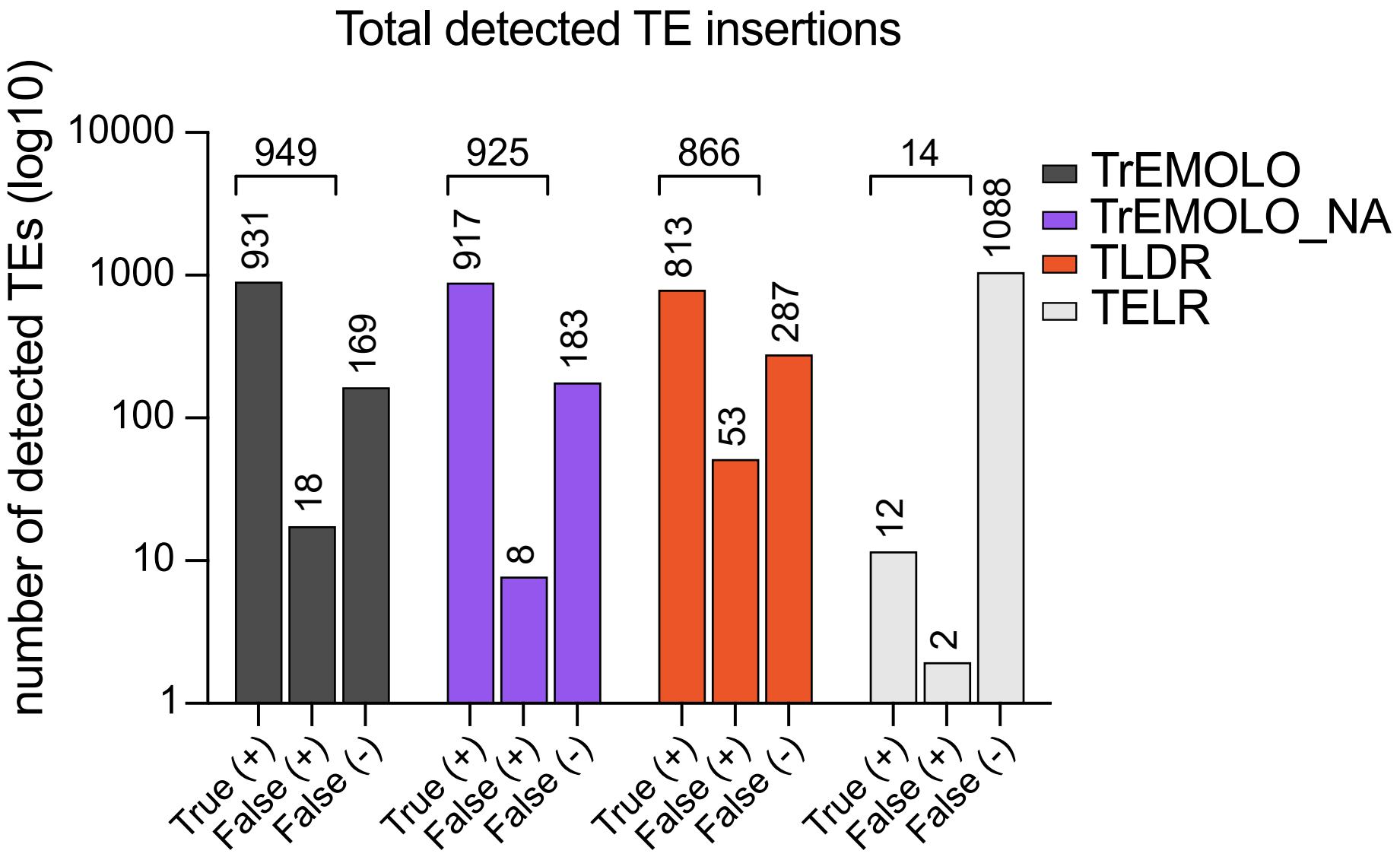

Total detected TE insertions
number of detected TEs (log10)
10000
1000
100
10
1
949
925
866
14
931
18
169
917
8
183
813
53
287
12
2
1088
TrEMOLO
TrEMOLO_NA
TLDR
TELR
True (+)
False (+)
False (-)
True (+)
False (+)
False (-)
True (+)
False (+)
False (-)
True (+)
False (+)
False (-)

Figure 3

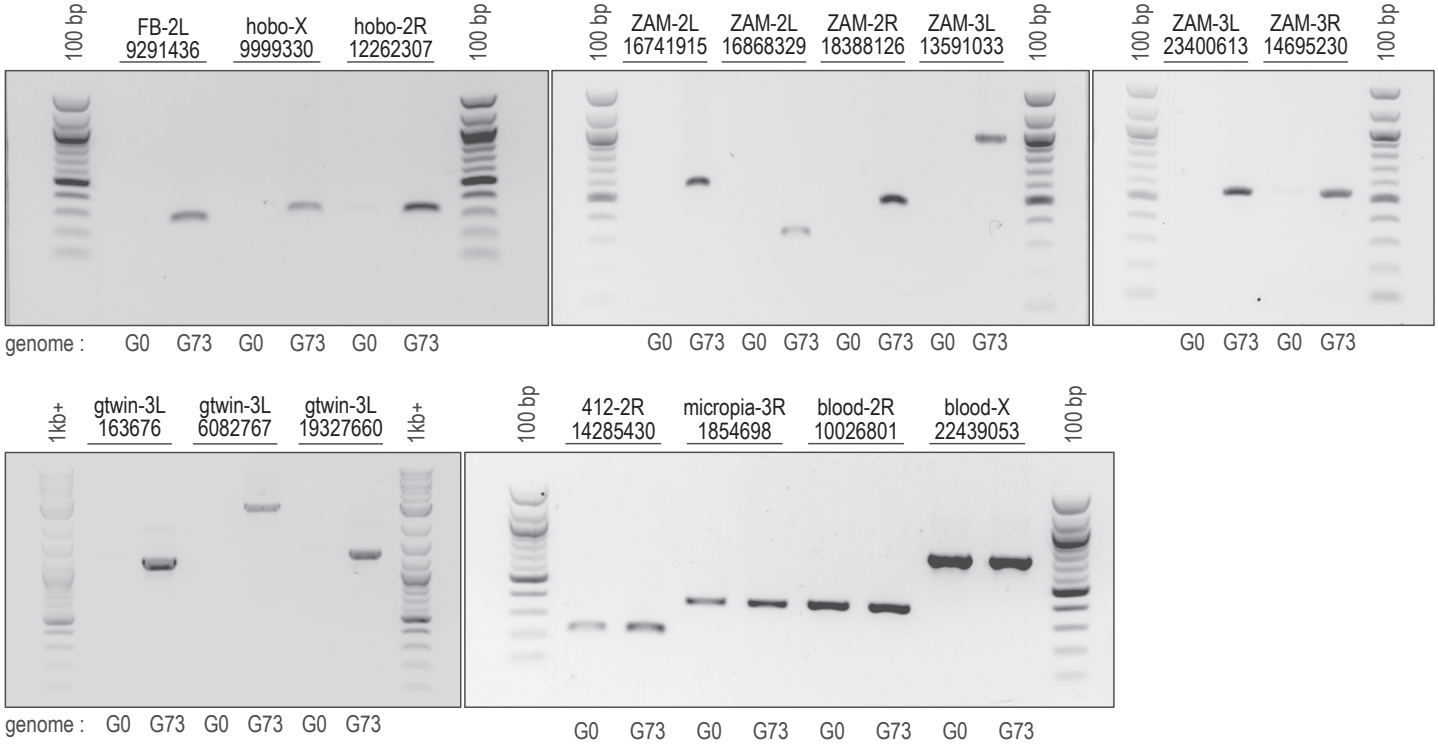

100 bp
FB-2L 9291436
hobo-X 9999330
hobo-2R 12262307
100 bp
genome : G0 G73 G0 G73 G0 G73
100 bp
ZAM-2L 16741915
ZAM-2L 16868329
ZAM-2R 18388126
ZAM-3L 13591033
100 bp
G0 G73 G0 G73 G0 G73 G0 G73
100 bp
ZAM-3L 23400613
ZAM-3R 14695230
100 bp
G0 G73 G0 G73
1kb+
gtwin-3L 163676
gtwin-3L 6082767
gtwin-3L 19327660
1kb+
genome : G0 G73 G0 G73 G0 G73
100 bp
412-2R 14285430
micropia-3R 1854698
blood-2R 10026801
blood-X 22439053
100 bp
G0 G73 G0 G73 G0 G73 G0 G73

Figure 3

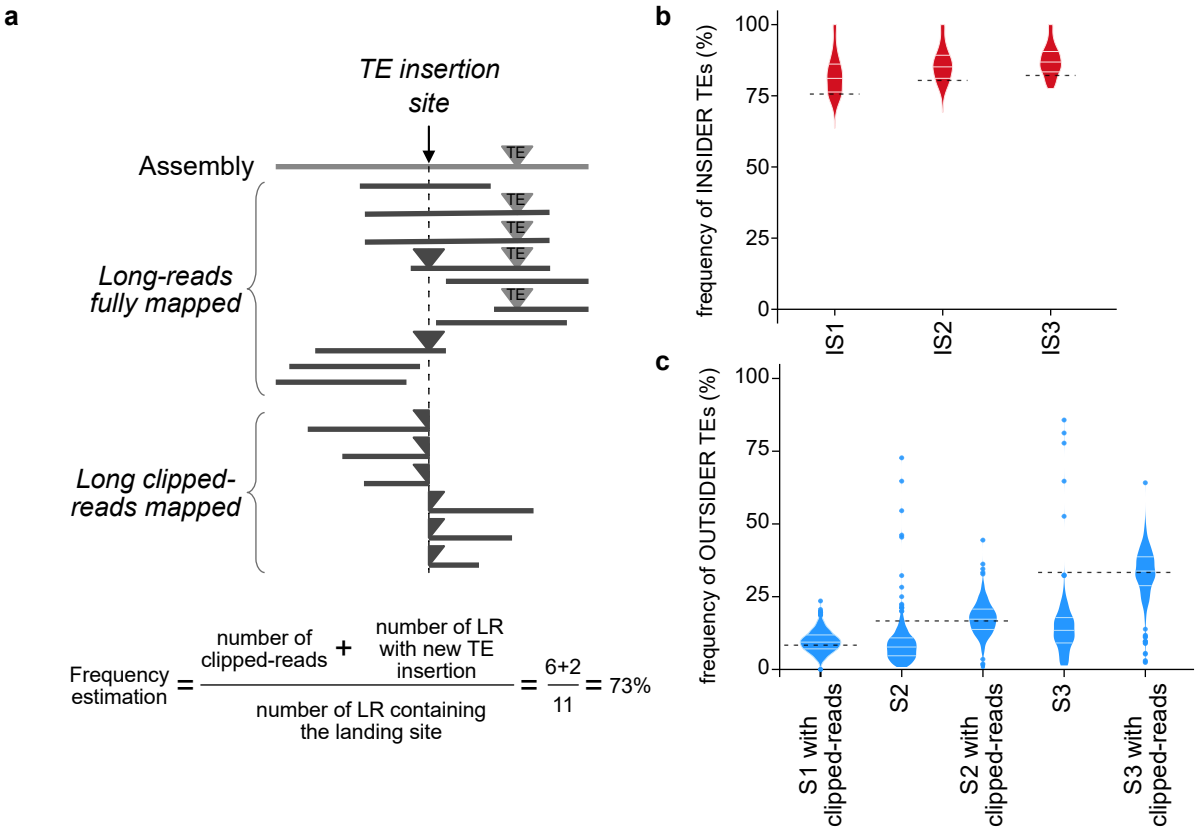

a
TE insertion site
Assembly
TE
Long-reads fully mapped
Long clipped-reads mapped
Frequency estimation = (number of clipped-reads + number of LR with new TE insertion) / number of LR containing the landing site = (6+2)/11 = 73%
b
frequency of INSIDER TEs (%)
100
75
50
25
0
IS1
IS2
IS3
c
frequency of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

Figure 5

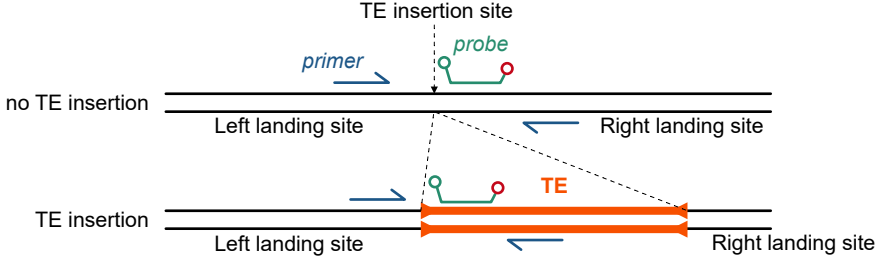

TE insertion site
primer
probe
no TE insertion
Left landing site
Right landing site
TE
TE insertion
Left landing site
Right landing site

Figure 6

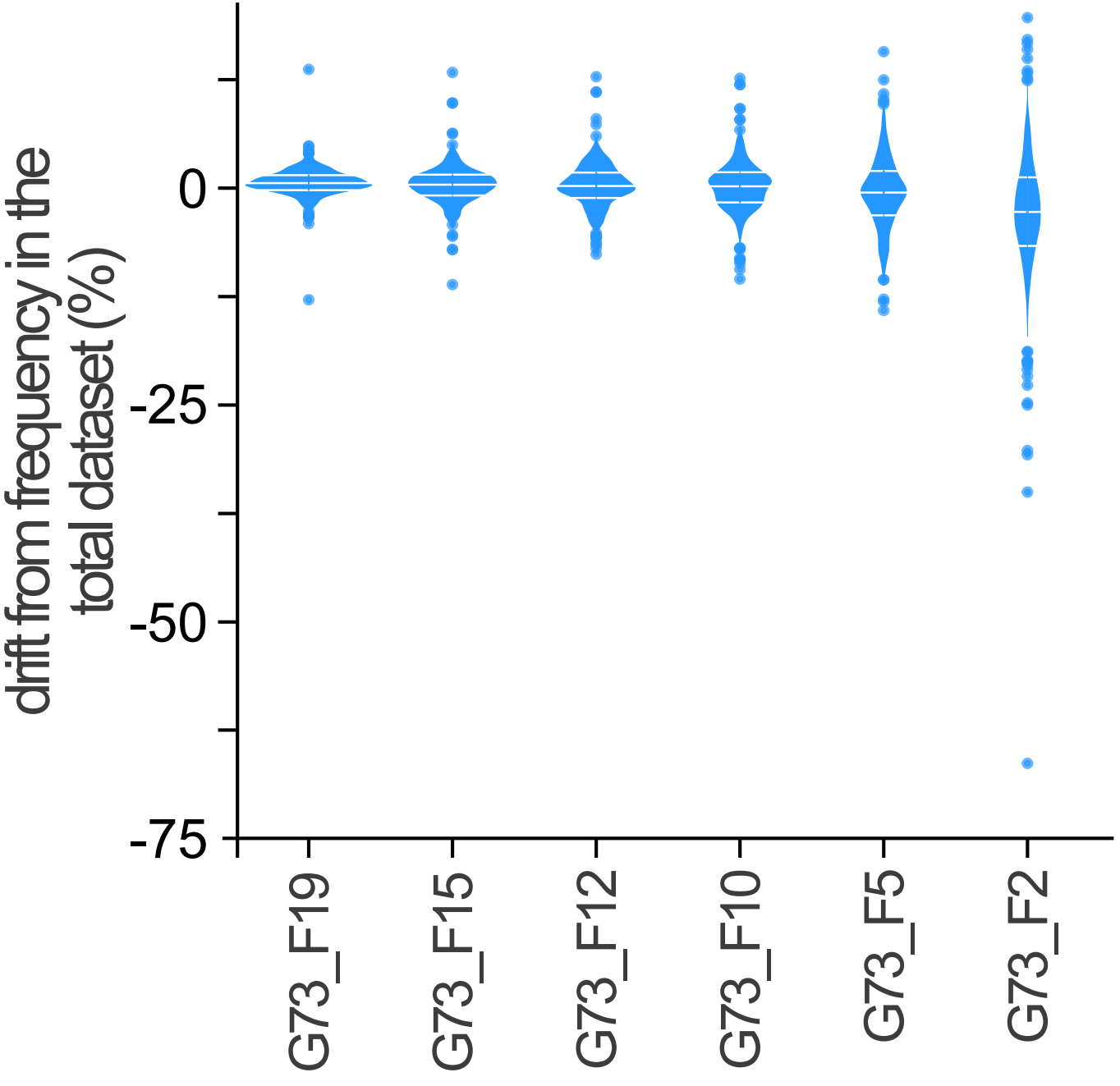

drift from frequency in the total dataset (%)
0
-25
-50
-75
G73_F19
G73_F15
G73_F12
G73_F10
G73_F5
G73_F2

Figure S1

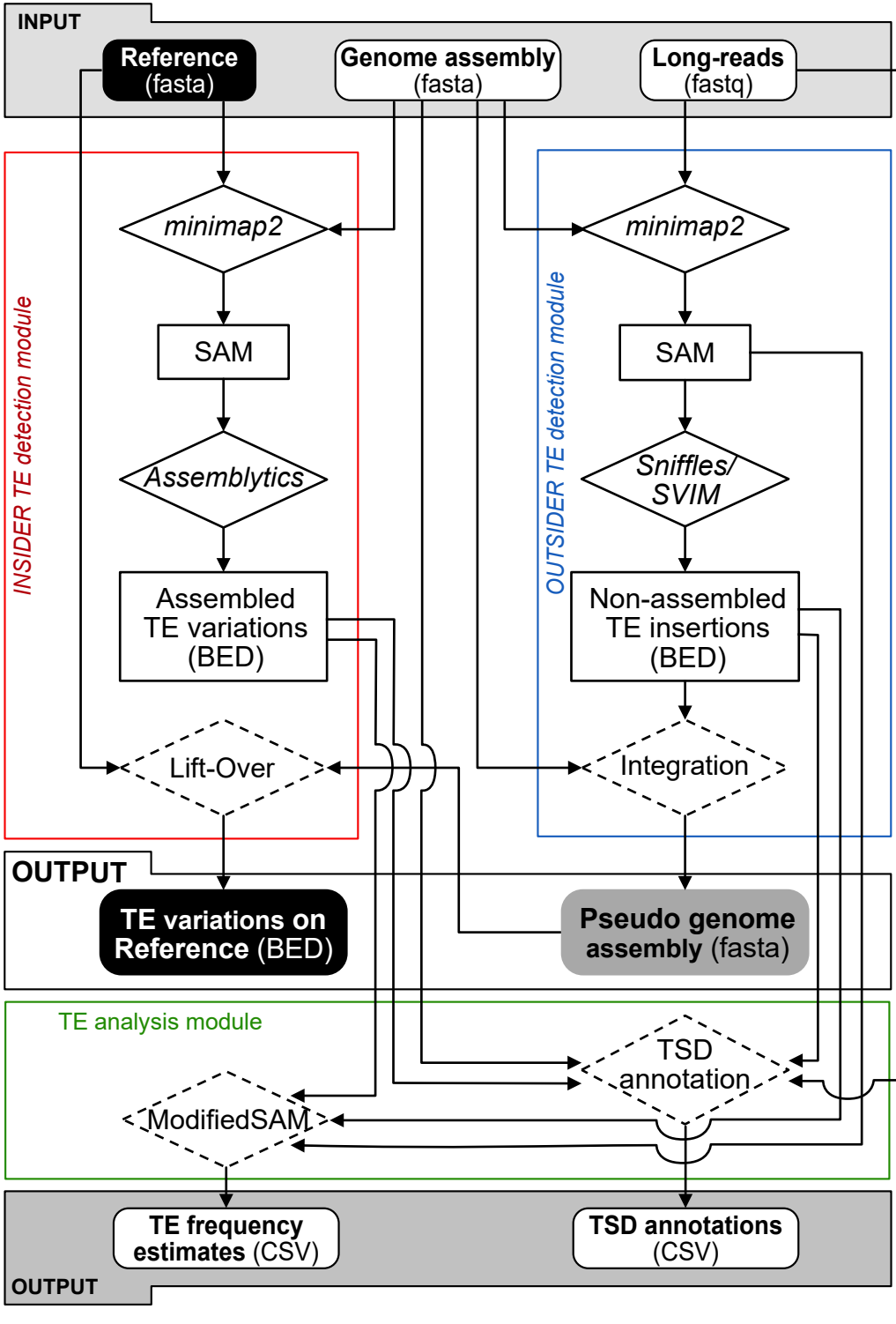
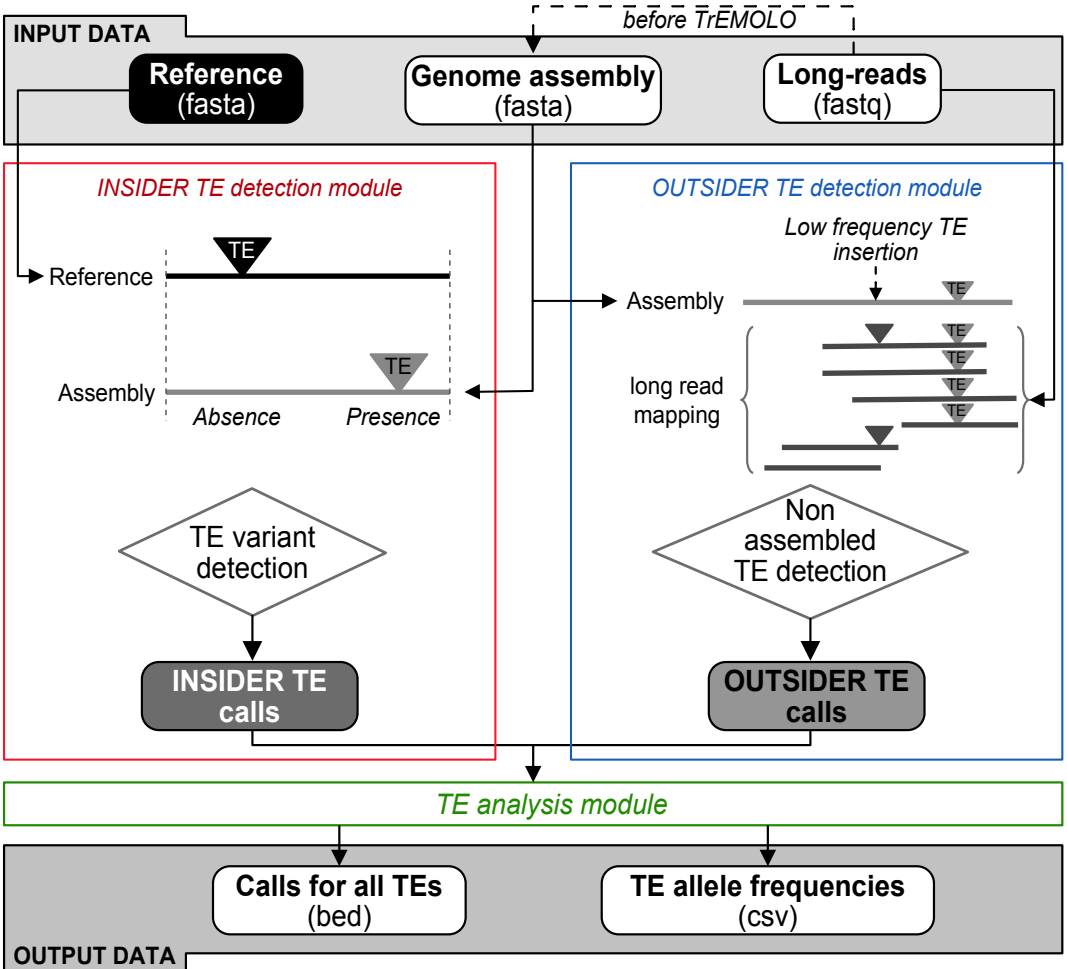

INPUT
Reference (fasta)
Genome assembly (fasta)
Long-reads (fastq)
INSIDER TE detection module
minimap2
SAM
Assemblytics
Assembled TE variations (BED)
Lift-Over
OUTSIDER TE detection module
minimap2
SAM
Sniffles/ SVIM
Non-assembled TE insertions (BED)
Integration
OUTPUT
TE variations on Reference (BED)
Pseudo genome assembly (fasta)
TE analysis module
TSD annotation
ModifiedSAM
TE frequency estimates (CSV)
TSD annotations (CSV)
OUTPUT

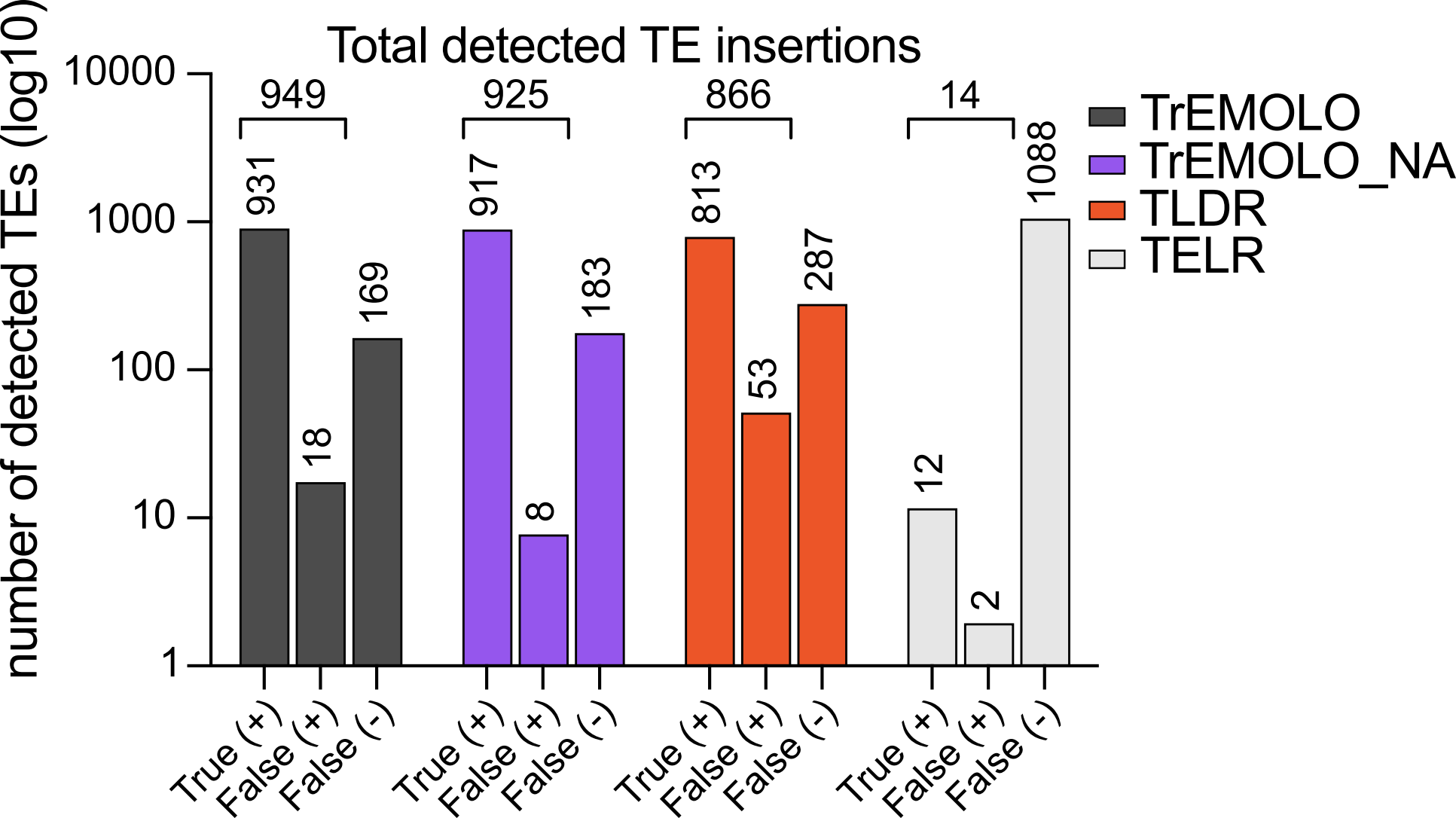
Table S1 : primer sequences used for TE detection in Figure 3.

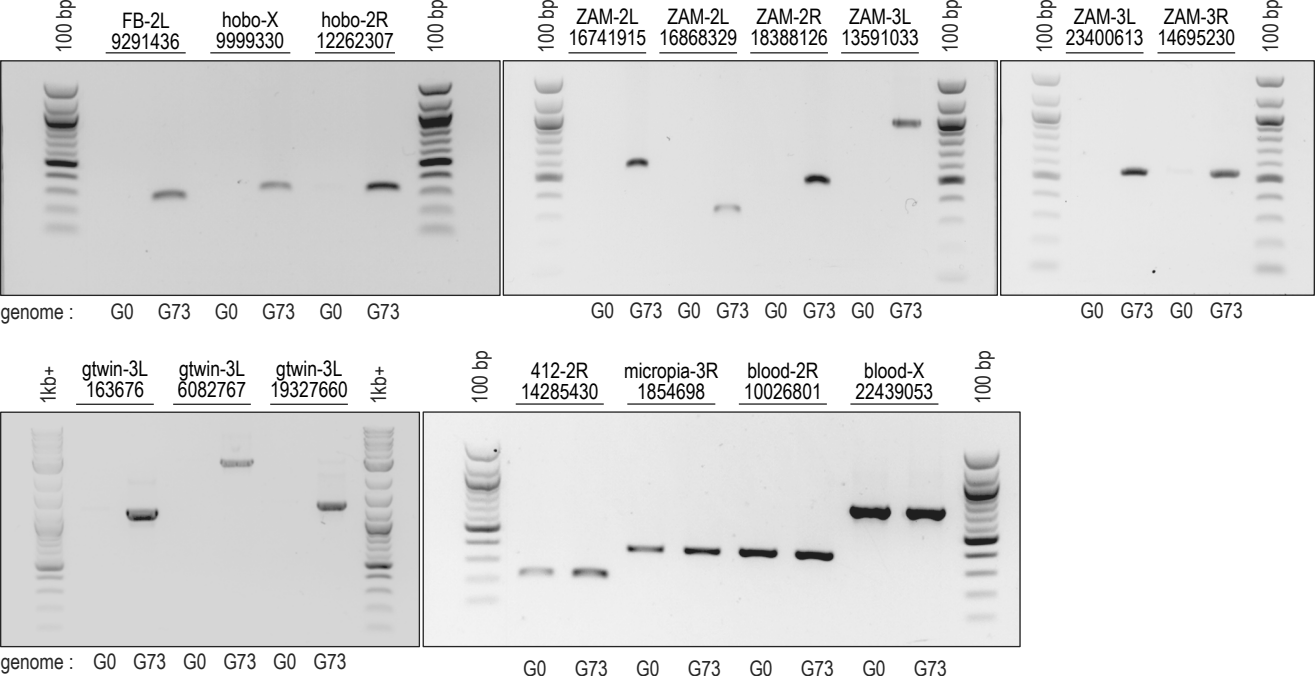
TE	**primer forward**	**primer reverse**
FB-2L-9291436	ATTTTGGCCACTTGTTGGAA	ACGTGGGCAAACGATGTTAT
hobo-X 9999330	GCCCCAGAGAAATTCAATCA	TTGTAAGTACCCGCGATGTG
hobo-2R 12262307	GACAGGCAAACAAGGAAAGC	AAAAACGGGGTGCCTAGAGT
412-2R-14285430	AAGATACACGGGCGATGAAG	GCAACAAAGTGCTGCCATAA
micropia-3R 1854698	CAAGTTTTCGGCCAATTCAT	ACACCCTGCATTTTCTGAGG
blood-2R 10026801	CCAAAATAGACGCATGCAGA	CGTGGGGATGCTGACTTAG
blood-X 22439053	AGGGTCCAGAAGAACACCAA	CTCACACCTGTTGTCGCTGT
ZAM-2L 16741915	GTCAAGGACTGAAATCCTTTTACT	TTCGCTCAGACAGAACAACG
ZAM-2L 16868329	GGAGAGCGGGCACTATTTTA	CCCGGTTCGATGTCTTCTAC
ZAM-2R 18388126	CTGCAAAGTTTGTTCGGAGA	TTCGCTCAGACAGAACAACG
ZAM-3L 13591033	TTGTGTGTCCTCGTTTCCAG	GCGGTTACGGCATTTTTAAG
ZAM-3L 23400613	CATCGTCTAGTACTCCCTGGAAA	TTCGCTCAGACAGAACAACG
ZAM-3R 14695230	TTGAATGCTGCCTGAGTGTC	TTCGCTCAGACAGAACAACG
gtwin-3L 163676	AGTGGAAACGATCCTCTGGA	CGGGCTCTAAATCAAACTCG
gtwin-3L 6082767	TGGTGTCCTTTGACTTGGTG	ACAGGGCCCCTACTTCAACT
gtwin-3L 19327660	TTGTAAGTACCCGCGATGTG	CGGGCTCTAAATCAAACTCG

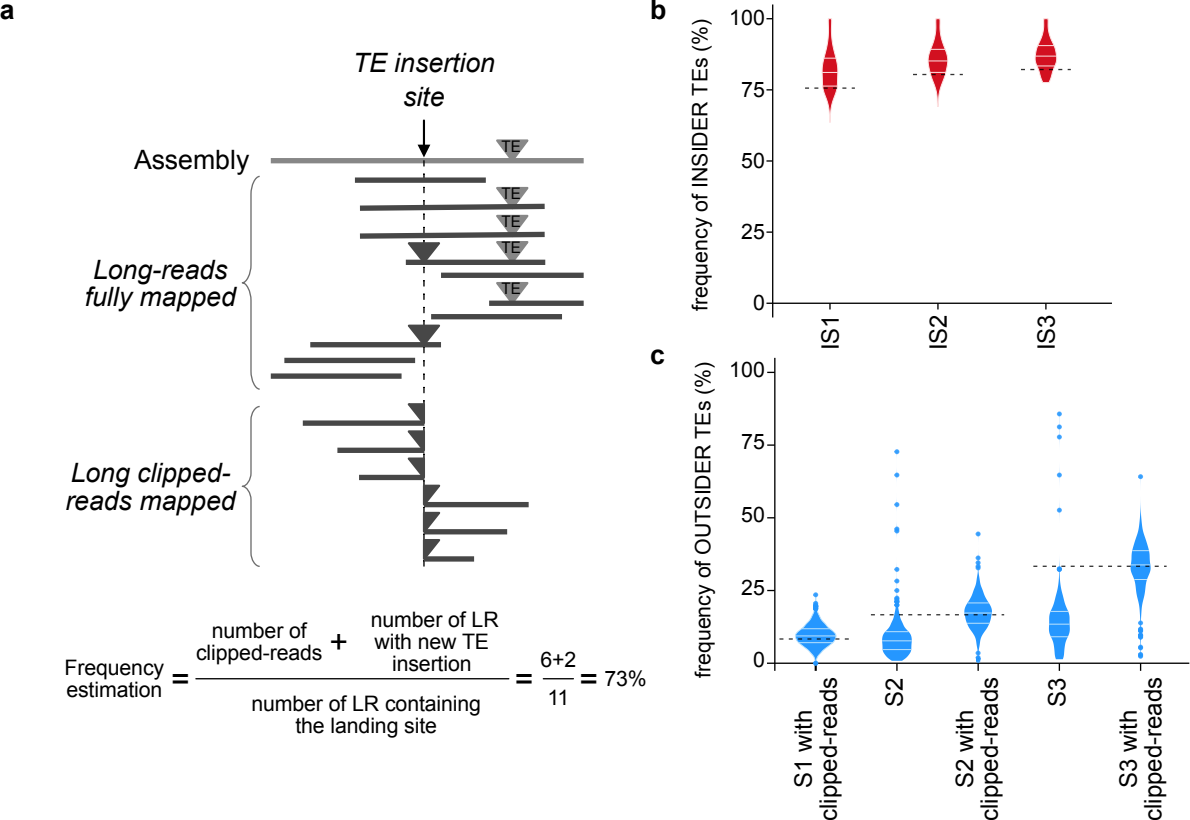
Table S2 : primer and probe sequences used for digital polymerase chain reaction (figure 5).

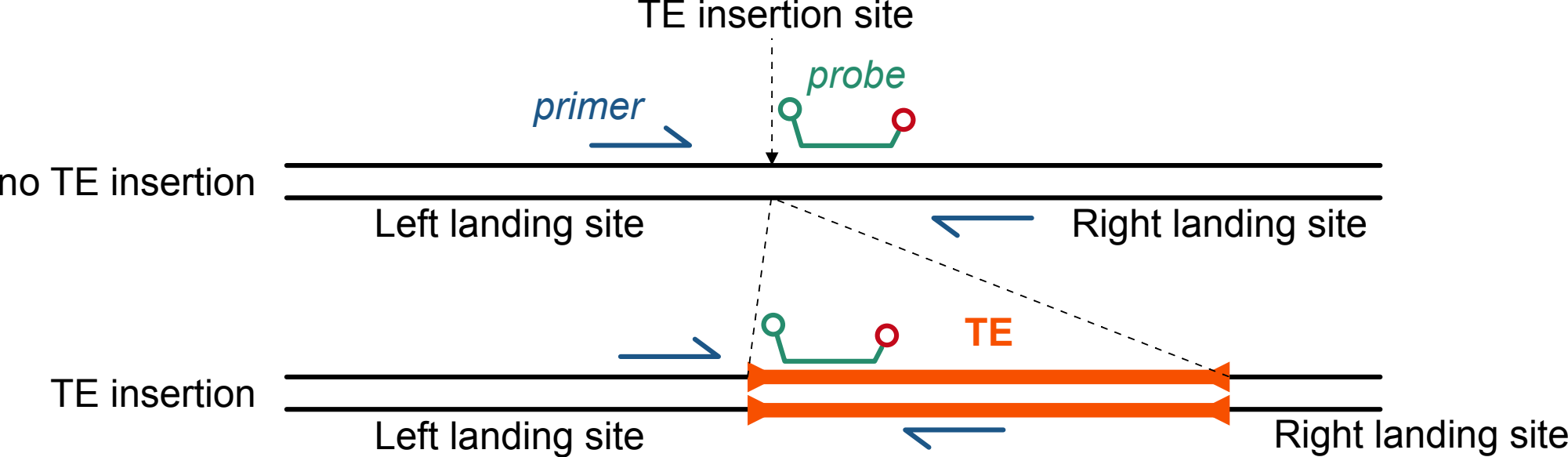
target	*detection*	*primers*	*seq*
412-2R_14285430	insertion	forward	TTGCTAAACTCCAAATTGCTGG
		reverse	AAGTGCTGCCATAAGTTAATATGC
		probe	56-FAM/TCAATGCGG/ZEN/TGGCCCCAGAAATAA/3IABkFQ
	no insertion	forward	TTGCTAAACTCCAAATTGCTGG
		reverse	ATCATAGAGAAACCAGTGCCAG
		probe	56-FAM/TCAATGCGG/ZEN/TGGCCCCAGAAATAA/3IABkFQ
blood-2L_18128567	insertion	forward	CCCAGAAGTGTTTCGGCAAT
		reverse	CACCCAACTGCAAGGAA
		probe	56-FAM/TATGCCACC/ZEN/AATTGAGAACACGT/3IABkFQ
	no insertion	forward	CCCAGAAGTGTTTCGGCAAT
		reverse	TGGCACGTTACACCTCAA
		probe	56-FAM/ACTCGTCCA/ZEN/CTCAGTGCACGTGCCA/3IABkFQ
roo-2L_13882751	insertion	forward	AAAACCGGAAAGCAACAGG
		reverse	CGGAGCCCAAAATTGTAAGTC
		probe	56-FAM/AGCAGTGAC/ZEN/ACCTTGACTTACATTTTG/3IABkFQ
	no insertion	forward	ACAGGCAAAATGTAAGTCAAGG
		reverse	GCACTAAACATTCGCCATTCAT
		probe	56-FAM/TCGAGTTTT/ZEN/GAATTCATTCCAAGCAG/3IABkFQ
gtwin-3L_163676	insertion	forward	TATTTCTAGTTCCCGCAAGCC
		reverse	CTGTTGCTGACCGTTCGT
		probe	56-FAM/ACGCGCCCA/ZEN/AACTGAGTTCAGCGC/3IABkFQ
	no insertion	forward	AGTGGAAACGATCCTCTGGA
		reverse	GGTGTCTTCAGCAACGTAGT
		probe	56-FAM/TGCCCACGA/ZEN/GCTTGGTCTCTGCCA/3IABkFQ
gtwin-3L_6082767	insertion	forward	CACTCGATGTGCGTATGTGT
		reverse	TATTGTCCCTGGTAGCAGCC
		probe	56-FAM/CCGGTGGTC/ZEN/TCCAGGCGGTGGA/3IABkFQ
	no insertion	forward	CTCGATGTGCGTATGTGTTTG
		reverse	TTATCGCAGACTTTCCCACC
		probe	56-FAM/CATTTTGTT/ZEN/TGCGGCTTTTTTGCG/3IABkFQ/
ZAM-2L_16741915	insertion	forward	AGAGGATTTCTCCAACCCCT
		reverse	GCAGCAAACACTTGTAGACG
		probe	56-FAM/CCTCCGGGG/ZEN/AGTCTTGCGGAGGT/3IABkFQ
	no insertion	forward	AAGAGGAAGAGCGAGAGGAT
		reverse	CCGAGAAAGCCTCGAAAGAA
		probe	56-FAM/TGGCGCGCG/ZEN/CAAGCGTAAGC/3IABkFQ
ZAM-2L_16868329	insertion	forward	CAGATGCATGTGCTCGATTG
		reverse	GTGGTGTATGGTACCGATGG
		probe	56-FAM/CGGCTTAAG/ZEN/CGGAGCCACTCCAATCCC/3IABkFQ
	no insertion	forward	idem forward with insertion
		reverse	GTATTCACAGCGTCGTTTCG
		probe	56-FAM/AGGCAGAGC/ZEN/GCGCCCACCCAT/3IABkFQ
ZAM-2R_18388126	insertion	forward	ATTCATCCATCCAGTCAGCC
		reverse	AAATTCTCCCAAGACGACCG
		probe	56-FAM/TGCACGCCG/ZEN/GGGCAAACTGG/3IABkFQ
	no insertion	forward	CAGTGTACGTACCCCTTTCC
		reverse	AAGTGAACCACACACACTCG
		probe	56-FAM/TGCACGCCG/ZEN/GGGCAAACTGG/3IABkFQ
ZAM-3L_13591033	insertion	forward	CTGTTGCTATGCATGTTGCG
		reverse	CCCGGTTCGATGTCTTCTAC
		probe	56-FAM/CCCCTCCCT/ZEN/CTAAGCCACCACGCC/3IABkFQ
	no insertion	forward	GCTTTTGCGCTTCGTCAATC
		reverse	GGCCCGAACACTTGTTAATG
		probe	56-FAM/CGCGCGCAT/ZEN/GCATTAAAGCTGCC/3IABkFQ
RpL32_gene	gene	forward	CACCAGTCGGATCGATATGC
		reverse	CATTTGTGCTGCAAGGAGAC
		probe	5HEX/TGGCACAAT/ZEN/CCTCGTTGGCACTCACCG/3IABkFQ

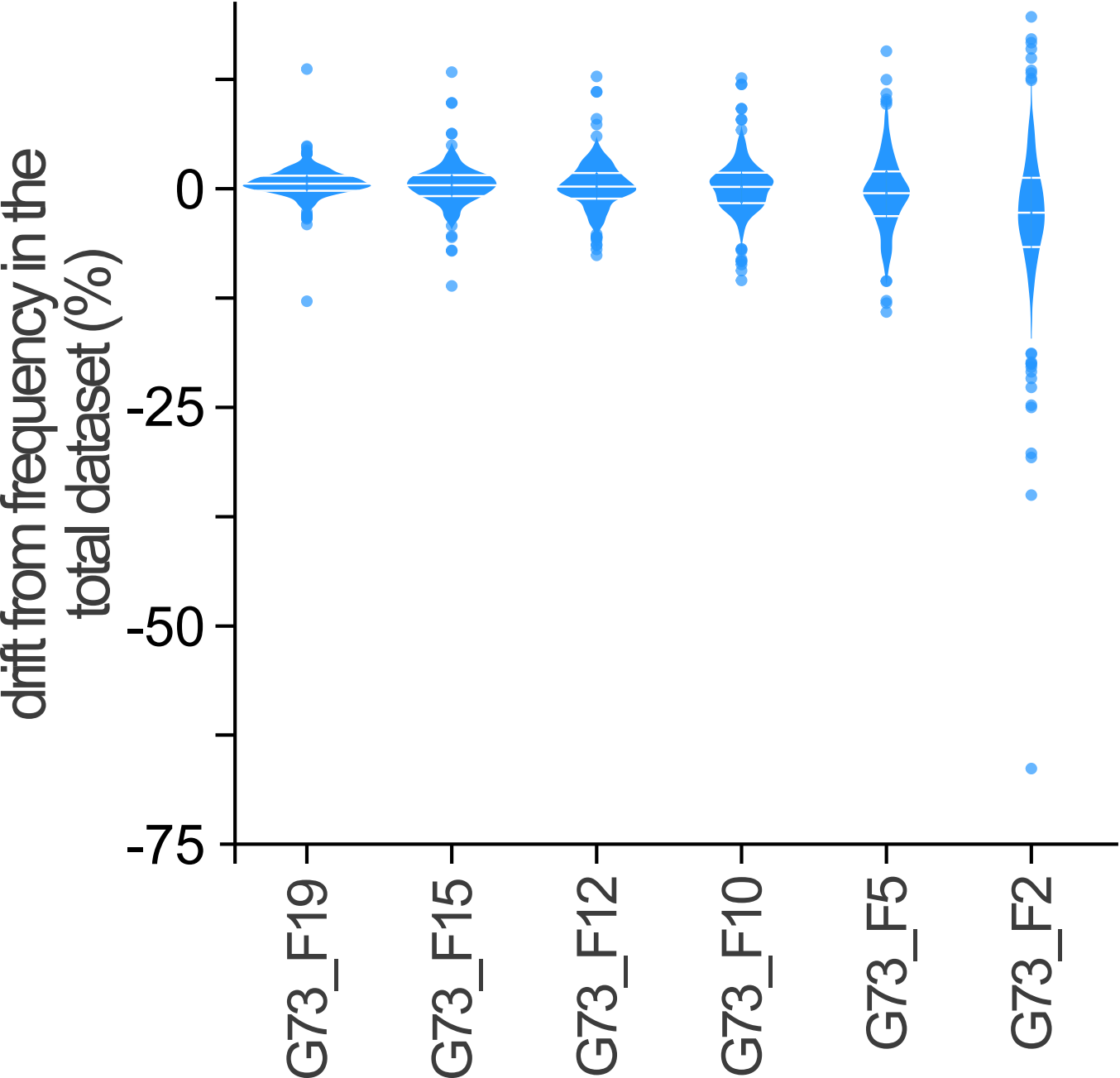

INPUT DATA
before TrEMOLO
Reference
(fasta)
Genome assembly
(fasta)
Long-reads
(fastq)
INSIDER TE detection module
Reference
TE
Assembly
TE
Absence
Presence
TE variant
detection
INSIDER TE
calls
OUTSIDER TE detection module
Low frequency TE
insertion
Assembly
TE
long read
mapping
TE
Non
assembled
TE detection
OUTSIDER TE
calls
TE analysis module
Calls for all TEs
(bed)
TE allele frequencies
(csv)
OUTPUT DATA


Total detected TE insertions
number of detected TEs (log10)
10000
1000
100
10
1
949
925
866
14
931
18
169
917
8
183
813
53
287
12
2
1088
TrEMOLO
TrEMOLO_NA
TLDR
TELR
True (+)
False (+)
False (-)


100 bp
FB-2L 9291436
hobo-X 9999330
hobo-2R 12262307
100 bp
genome : G0 G73 G0 G73 G0 G73
100 bp
ZAM-2L 16741915
ZAM-2L 16868329
ZAM-2R 18388126
ZAM-3L 13591033
100 bp
G0 G73 G0 G73 G0 G73 G0 G73
100 bp
ZAM-3L 23400613
ZAM-3R 14695230
100 bp
G0 G73 G0 G73
1kb+
gtwin-3L 163676
gtwin-3L 6082767
gtwin-3L 19327660
1kb+
genome : G0 G73 G0 G73 G0 G73
100 bp
412-2R 14285430
micropia-3R 1854698
blood-2R 10026801
blood-X 22439053
100 bp
G0 G73 G0 G73 G0 G73 G0 G73


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 with clipped-reads
S2
S2 with clipped-reads
S3
S3 with clipped-reads


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


drift from frequency in the total dataset (%)
0
-25
-50
-75
G73_F19
G73_F15
G73_F12
G73_F10
G73_F5
G73_F2