Bioinformatics, 36(4), 2020, 1191–1197
doi: 10.1093/bioinformatics/btz727
Advance Access Publication Date: 3 October 2019
Original Paper

Genetics and population analysis

T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data

María Bogaerts-Márquez[1], Maite G. Barrón[1], Anna-Sophie Fiston-Lavier [2], Pol Vendrell-Mir[3], Raúl Castanera[3], Josep M. Casacuberta[3] and Josefa González[1,*]

[1]Institute of Evolutionary Biology (CSIC-Universitat Pompeu Fabra), Paseo Maritimo Barceloneta 37–49, Barcelona, Spain, [2]Institut des Sciences de l'Evolution de Montpellier (UMR 5554, CNRS-UM-IRD-EPHE), 11 Université de Motpellier, Place Eugène Bataillon, Montpellier, France and [3]Center for Research in Agricultural Genomics, CRAG (CSIC-IRTA-UAB-UB), Campus UAB, Cerdanyola del Vallès, Barcelona, Spain

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on March 28, 2019; revised on September 16, 2019; editorial decision on September 18, 2019; accepted on September 25, 2019

Abstract

Motivation: Transposable elements (TEs) constitute a significant proportion of the majority of genomes sequenced to date. TEs are responsible for a considerable fraction of the genetic variation within and among species. Accurate genotyping of TEs in genomes is therefore crucial for a complete identification of the genetic differences among individuals, populations and species.

Results: In this work, we present a new version of *T-lex*, a computational pipeline that accurately genotypes and estimates the population frequencies of reference TE insertions using short-read high-throughput sequencing data. In this new version, we have re-designed the *T-lex* algorithm to integrate the BWA-MEM short-read aligner, which is one of the most accurate short-read mappers and can be launched on longer short-reads (e.g. reads >150 bp). We have added new filtering steps to increase the accuracy of the genotyping, and new parameters that allow the user to control both the minimum and maximum number of reads, and the minimum number of strains to genotype a TE insertion. We also showed for the first time that *T-lex3* provides accurate TE calls in a plant genome.

Availability and implementation: To test the accuracy of *T-lex3*, we called 1630 individual TE insertions in *Drosophila melanogaster*, 1600 individual TE insertions in humans, and 3067 individual TE insertions in the rice genome. We showed that this new version of *T-lex* is a broadly applicable and accurate tool for genotyping and estimating TE frequencies in organisms with different genome sizes and different TE contents. *T-lex3* is available at Github: https://github.com/GonzalezLab/T-lex3.

Contact: josefa.gonzalez@ibe.upf-csic.es

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Structural variants, such as insertions, deletions and inversions, are often ignored when analyzing genomic variation due to the technical limitations associated with short-read high-throughput sequencing (Hoban *et al.*, 2016; Villanueva-Cañas *et al.*, 2017). However, it has become apparent that structural variants are a considerable source of genomic variation. For example, in the human genome structural variants outnumber single-base-pair differences (Alkan *et al.*, 2011). Transposable elements (TEs) are a type of structural variant that are present in virtually all genomes sequenced to date, where they represent a substantial proportion of the genome content—ranging from ~3% in yeast to ~80% in maize and wheat (Bleykasten-Grosshans and Neuvéglise, 2011; Mascher *et al.*, 2017; Schnable *et al.*, 2009). Species differ not only in the total TE genome content, but also in the diversity of TEs, and in the proportion of active and inactive TE copies (Guio and González, 2019). The transposition activity of TEs can generate a significant amount of genetic variation. TEs can alter fitness-related traits or cause diseases when they disrupt genes or affect their level of expression (Chuong *et al.*, 2017; Elbarbary *et al.*, 2016). TEs can also generate mutations after they have lost the capacity to transpose by acting as substrates for ectopic recombination or by facilitating template switching during repair of replication errors (Campbell *et al.*, 2014; Lee *et al.*, 2007; Startek *et al.*, 2015). All these TE-induced mutations contribute to the genetic variation within

Table 1. Characteristics of computational tools used to annotate and detect transposable element insertions

Software	Reference/ Non-reference	Polymorphic TE detection	TE classes tested	Runs with pooled data	TSD detection	Species tested	Reference
RelocaTE and CharacTErizer	Non-reference and reference	Yes, but only for non-reference	*mPing*	No	No	*Oryza sativa*	Robb *et al.* (2013)
STEAK	Non-reference and reference	Yes	*HK2*	No	No	Human	Santander *et al.* (2017)
MELT	Non-reference and reference	Yes	Alu, L1 and SVA	Yes	Yes	Human, *D.melanogaster*	Gardner *et al.* (2017)
TEMP	Non-reference and reference	Yes	All	Yes	No	*D.melanogaster*	Zhuang *et al.* (2014)
PoPoolationTE2	Non-reference and reference	Yes	All	Yes	No	*D.melanogaster*	Kofler *et al.* (2016)
TIDAL	Non-reference and reference	No	All	Yes	No	*D.melanogaster*	Rahman *et al.* (2015)
T-lex3	Reference	Yes	All	Yes	Yes	*D.melanogaster*, human, *Oryza sativa*	This study

and among species. Thus, accurate genotyping of TEs in genomes is crucial for complete identification of the genetic differences among individuals, populations and species.

There are many different computational tools that take advantage of the wealth of short-read high-throughput sequencing data available to discover and estimate population frequencies of TE insertions (Goerner-Potvin and Bourque, 2018; Rishishwar *et al.*, 2017). A non-exhaustive list of tools that identify non-reference and reference insertions shows that some characteristics are commonly shared among tools, such as the ability to detect polymorphic insertions (Table 1). However, other characteristics are more specific, such as the detection of the target site duplications (TSD), which helps to reconstruct the pre-integration site and thus identify misannotated TE insertions (Table 1).

In 2011, we developed a tool for fast and accurate frequency estimation of reference TE insertions called *T-lex* (Fiston-Lavier *et al.*, 2011). The first update of this software, *T-lex2* (Fiston-Lavier *et al.*, 2015), allowed the user to work with individual strain and pooled whole genome sequencing data. In addition, we added a *TE-TSD detection* module that allowed us to correct the genome annotations for 65 *Drosophila melanogaster* reference TE insertions. We also showed that *T-lex2* provides accurate TE calls in *D.melanogaster*, with 99.14% specificity and 89.58% sensitivity, and in the human genome, with 97.65% specificity and 93.26% sensitivity.

In this work, we implemented a new version of this broadly applicable and flexible tool, which can now be used with the latest short-read high-throughput sequencing technologies. We have redesigned the *T-lex* algorithm to integrate the *BWA-MEM* short-read aligner: one of the most accurate short-read mappers available (Li and Durbin, 2009). This aligner has no read length limitation, is currently being maintained by its developers, and it is easy to install as the package is available for most *Linux* distributions and OSX. In addition, we have added extra filtering steps and parameters, and fixed some bugs present in *T-lex2*. We demonstrate that *T-lex3*, now available at *Github* (https://github.com/GonzalezLab/T-lex3), can provide accurate TE genotyping and frequency estimation using *D.melanogaster*, human and *Oryza sativa* datasets.

2 Materials and methods

T-lex3, as its previous version, is composed of five modules. Briefly, the *TE-analysis* module analyses whether the sequences flanking each reference TE insertion are part of a segmental duplication, or whether they contain repetitive regions, as both features are known to affect the accuracy of TE calls. The *TE-presence detection* module and the *TE-absence detection* module are two independent modules that detect both the presence and the absence of the reference TE based on the analysis of the junction sequences of the TE insertion. When the requirements to call a TE as *present* or *absent* are not met, these modules return a *no data* call. The *TE-combine* module combines the results of the two detection modules to determine whether the TE insertion is *present*, *absent*, *polymorphic* or *no data*. Finally, the *TE-TSD detection* module analyzes the read alignments of the *TE-absence detection* module to identify the Target Site Duplication (TSD) of the TE insertion. In the new version of *T-lex* described here, *T-lex3*, we have updated and improved three of these modules (Fig. 1), and we have also implemented other more general changes as described below. All these changes have been summarized in Table 2.

2.1 TE-presence detection module

We have changed the mapper used in the *TE-presence detection* module: while *T-lex2* used *MAQ* (Li *et al.*, 2008), *T-lex3* uses *BWA-MEM* (Li and Durbin, 2009) (Table 2). *MAQ* has two main limitations: (i) it cannot process reads > 127 bp, while the newest *Illumina* technology produces longer reads (150 and 300 bp) and (ii) it is no longer updated by the developers. We chose *BWA-MEM* as the new mapper because it accepts reads > 127 bp and it performs mapping with high accuracy (Hatem *et al.*, 2013). In addition, the *bam* files generated by *BWA-MEM* can be visualized using tools such as *IGV* that are extensively used in the community (Robinson *et al.*, 2017). Because *T-lex2* used many of the output files of *MAQ* to call a TE as *present* or *absent*, we have implemented in *T-lex3* pipeline several *Perl* and *AWK* scripts, and included other tools such as *SAMtools* (Li *et al.*, 2009) to generate these files. Furthermore, we have added an extra filtering step: *T-lex3* now filters using the CIGAR string information all the reads with an alignment match < 95% compared with the reference genome. Because this percentage is calculated based on the read length specified by the user, this filter also allows to remove shorter reads in those cases in which reads of different lengths are available for a given strain or pool. This allows fixing a bug in *T-lex2* code, which previously considered all reads to be of the size specified by the user leading to incorrect TE calls.

After the mapping is completed, the *TE-presence detection* module analyzes the junction regions to call the TE as *present* or *absent*. The size of the junction region is based on the *–limp* and *–buffer* parameters (by default 15 and 60, respectively). To consider a TE as *present*, three conditions had to be met in at least one of the two junction regions: (i) at least one read mapped in the junction, (ii) > 15 bp mapped inside the TE (*-limp* parameter) and (iii) 95% identity of the sequence inside the TE. There was a bug in *T-lex2*: if two conditions were fulfilled in one of the junction regions and the other condition was fulfilled in the other junction region, the TE was considered as *present*. *T-lex3* now requires that all three conditions are met in at least one of the junction regions (Fig. 2A, Table 2).

For the *TE-presence detection* module to call a TE insertion as *absent*, *T-lex2* required < 65 Ns (missing data) in the sequence mapped to the junction, otherwise the TE call was *no data*. In *T-lex3*, we have modified this step and introduced a

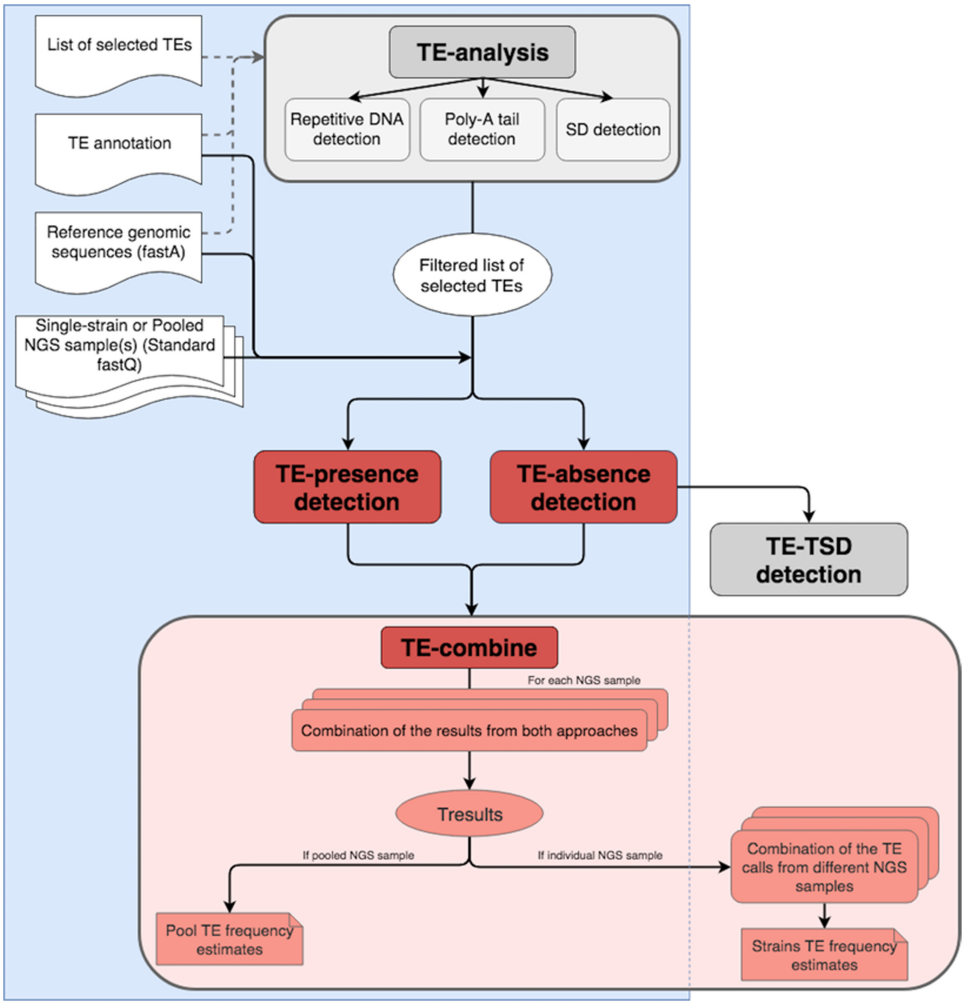

Fig. 1. *T-lex3* pipeline. *T-lex3* modules re-designed or improved are highlighted in darker colour. The vertical box indicates the modules that are run by default when *T-lex3* is launched. To run the *TE-TSD detection* module the *TE-absence detection* module results are required. For individually sequenced strains, the individual *Tresults* files for each strain are needed before running the combination of the TE calls from different high-throughput sequencing samples. The discontinuous lines in grey indicate the recommended path when the program is run for the first time with a given reference genome. For subsequent runs, we recommend running the pipeline with the filtered list of selected TEs

new one before calling a TE insertion as either *absent* or *no data* (Fig. 2B). To call a TE insertion as *absent*, *T-lex3* now requires < 70 Ns (missing data) in the sequence mapping to the junction, and 20 of the other nucleotides in that sequence must map to the flanking region with $\geq 95\%$ identity. If this step is not fulfilled, an additional step requires that 10 nucleotides in the junction region (five nucleotides inside and five outside the TE sequence) are not *Ns* (missing data) to consider the TE insertion as *absent*. If this condition is not met, the TE call is *no data* (Fig. 2C). This extra step is necessary because we detected some cases in which there was a read mapping to the junction that only had sequence identity within the TE leading to incorrect *absent* calls.

2.2 TE-absence detection module

Minor changes affecting the default parameters were made within this module to adapt it to longer size reads. The parameter *–lima*, defines the minimum number of non-repeated nucleotides in each side of the insertion site in the junction region. In *T-lex3*, this parameter is 10 by default instead of five. The parameter *-v* defines the minimum read length mapping in each side of the insertion site in the junction regions. In *T-lex3* this parameter is 20 instead of 15 (Table 2).

2.3 TE-combine module

We have implemented several changes in the *TE-combine* module. The first step of this module is the combination of the results from the *TE-presence* and *TE-absence detection* modules to generate a *Tresults* file that is used to estimate the TE frequencies (Fig. 1). *T-lex2* considered a TE insertion to be *absent/polymorphic* when the *TE-presence detection* module gave a *no data* call and the *TE-absence detection* module gave an *absent* call. Similarly, *T-lex2* considered a TE to be *present/polymorphic* when the *TE-presence detection* module gave a *present* call and the *TE-absence detection* module gave a *no data* call. Manual curation of several of these calls revealed that most of these cases should be considered as *no data*. Thus, in *T-lex3* we considered that when we have a call from only one of the two modules, the combination of the results from the two approaches is *no data* (Table 3).

The *TE-combine* module takes into account whether the data comes from pooled or from individual strains to estimate the TE frequencies. When the TE frequency is estimated from pool-seq data, we have added two new parameters and we have fixed two bugs (Table 2). The new parameters *–minR* and *-maxR* allow the user to define the minimum and the maximum number of reads needed to estimate the TE frequency. Requiring a minimum and maximum number of reads allows false positives to be discarded (very low number of reads to make an accurate TE call), and to discard regions with an excess of coverage due to non-unique mapping or spurious reads (very high number of reads to make an accurate TE call).

There was a bug in the formula for the TE frequency estimation in *T-lex2*: TE frequency = (NP)/(NP+NA), where NP is the total number of reads supporting the presence and NA is the total number

Table 2. Summary of changes introduced in the new *T-lex* version

Module	Steps	Features/parameters	*T-lex2*	*T-lex3*
TE-presence detection	Mapping	Mapper (read length limit)	MAQ (127bp)	BWA-MEM (70bp to 1Mb)
	Filtering of mapping output	Read identity	No filter	Filters reads with < 95 % alignment match
		Read length	No filter	Filters reads shorter than read length specified by the user
	TE call	*Present* call	At least one condition must be fulfilled in one flanking region	All conditions must be fulfilled for the same flanking region
		Absent call	<65 *Ns* (missing data) in the junction region	<70 *Ns* (missing data) in the junction region, and ≥ 95% identity in at least 20 bp of the flanking region
		Absent call	No filter	<10 *Ns* (missing data) in the junction region
TE-absence detection	TE call	*–lima* parameter	5	10
		-v parameter	15	20
TE-combine	TE call	Combination of the TE calls from the absence and presence modules	Absent/polymorphic Present/polymorphic	No data
	Pool Frequency Estimation	*–minR* parameter	No	Yes
		-maxR parameter	No	Yes
		Frequency estimate takes into account whether information is available for one or for the two junction regions	No	Yes
		Considers reads before the calling steps	Yes	No
	Individual strains TE frequency estimates	*–minP* parameter	No	Yes
		File name for the combination of different individual strains	*Tlex_output*	*Tfreqs_output*
General	Warnings about running process		No	Yes
	Available test dataset		No	Yes
	Available in Github		No	Yes

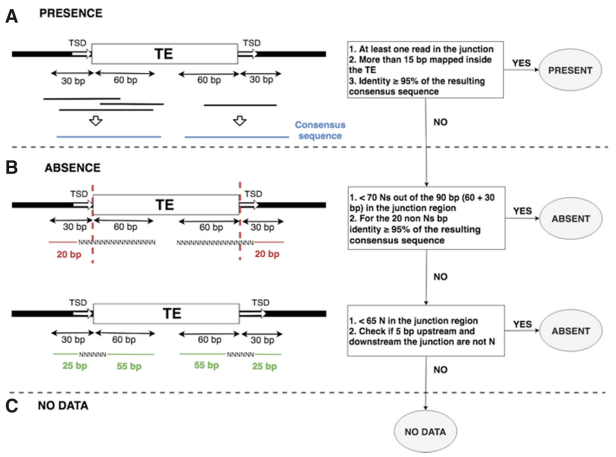

Fig. 2. *TE calling steps of the T-lex3 TE-presence detection module.* Example of the three steps of *T-lex3* TE-presence detection module to call a TE present, absent or no data for 100 bp reads with default parameters

of reads supporting the absence. This formula did not take into account whether there were reads providing evidence for the presence for only one or for the two flanking regions. In *T-lex3* this formula is only used if there is information for only one of the flanking regions. If there are reads mapping on each of the two flanking regions, the formula is: TE frequency = [(NPR+NPL)/2]/(NPR+NPL+NA), where NPR and NPL are the total number of reads supporting the presence at the right and left flanking regions, respectively.

Table 3. Changes in final TE calls between *T-lex2* and *T-lex3* according to the individual *TE-presence* and *TE-absence detection* modules

TE-presence module	*TE-absence module*	*T-lex2*	*T-lex3*
Present	*present*	*present*	*present*
Present	*absent*	*polymorphic*	*polymorphic*
Present	*no data*	*present/poly*	*no data*
Absent	*present*	*no data*	*no data*
Absent	*absent*	*absent*	*absent*
Absent	*no data*	*no data*	*no data*
no data	*present*	*no data*	*no data*
no data	*absent*	*absent/poly*	*no data*
no data	*no data*	*no data*	*no data*

The other bug was in the estimation of the frequency of pooled samples: *T-lex2* considered the reads that the presence module maps in the flanking region before the calling steps. As such, there could be reads that mapped to the flanking regions even if the TE was finally called as absent. In *T-lex3*, these reads are not considered to estimate the frequency.

For the TE frequency estimation from individual strains data, we have added one new parameter and we have fixed one bug. The new parameter *–minP* allows the user to define a minimum number of strains with a *present*, *absent* or *polymorphic* result required to estimate the frequency of a given TE. Thus, it allows the user to discard TE frequencies estimates based on a low number of strains, as this result

might not be representative of the frequency of that TE in the population. Finally, we have changed the name of the folder where the combination of the individual strains results (*Tresults*) and the file with the frequencies (*Tfreqs*) are saved. *T-lex2* named this folder *Tlex_output*, which was problematic due to the similarity with the individual strains folder name. In *T-lex3*, this file is now called *Tfreqs_output*.

Besides the changes in the three modules described above, *T-lex3* now provides more detailed standard output that informs the user about the running process. For example, in the *TE-presence detection* module the program now informs the user when each step of the mapping process starts, and whether any of the mapping steps has failed. If a mapping step fails, the program stops. We have uploaded the pipeline to *Github* along with a test dataset that allows the user to check whether the program is properly installed. An updated user manual that provides all the information required to run and interpret *T-lex3* results can also be found at *Github* (https://github.com/GonzalezLab/T-lex3).

2.4 Datasets

To compare the performance between *T-lex2* and *T-lex3* we ran several datasets with both versions. For *D.melanogaster*, we selected 10 pool-seq datasets from the DrosEU consortium (Kapun *et al.*, 2018), and 30 individual strains: 19 from the DGRP dataset (Huang *et al.*, 2014) and 11 from the DPGP2 dataset (Pool *et al.*, 2012; Supplementary Additional File S1). We also ran *T-lex3* for several datasets that failed to give results with *T-lex2*: three pooled datasets from Italy, Austria (Bastide *et al.*, 2013), and Portugal (Kofler *et al.*, 2012), 86 individual strains from Lyon, France (Lack *et al.*, 2016), and 32 strains from Winters, CA (Campo *et al.*, 2013; Supplementary Additional File S4). For all runs, we used a dataset of 1630 reference TE insertions as described in Rech *et al.* (2019).

For humans, we used the same dataset as in *T-lex2*: the child genome of the trio human dataset NA12878 (SRR622457) with the reference genome NCBI36/hg18 (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/bigZips/chromFa.zip) (Stewart *et al.*, 2011). We analyzed 1600 insertions out of the 2010 that (i) have genomic coordinates with confidence intervals smaller or equal to 1, and (ii) have been analyzed in the CEU trio dataset.

In this work, we also run for the first time *T-lex3* with a rice genome (MH63) (Zhang *et al.*, 2016). We first annotated in the *Oryza sativa* Nipponbare assembly (Sasaki and International Rice Genome Sequencing, 2005) 1600 MITE and 1553 LTR insertions. To detect MITE families, we run *MITE-hunter* (Han and Wessler, 2010) and combined the results with the high-quality predictions available in the *PMITE* database (Chen *et al.*, 2014). Only families with TSDs were considered. Clustering at 90% was performed to remove redundancy using *cd-hit* (Fu *et al.*, 2012) to produce a final library. *RepeatMasker* (http://www.repeatmasker.org/) was run to annotate all regions having significant homology with any of the MITE families. Only the annotations corresponding to full-length elements (length equal to consensus length ± 20%) were considered.

LTR insertions were identified with *LTRharvest* (Ellinghaus *et al.*, 2008) on the Nipponbare assembly (Sasaki and International Rice Genome Sequencing, 2005) using default parameters. The potential to encode for proteins was investigated using *hmmscan* (Eddy, 2011), and only elements potentially coding for proteins containing typical retrotransposon conserved domains (e.g. reverse transcriptase and integrase) were retained for further analyses.

2.5 Sensitivity, specificity and accuracy estimations

T-lex3 sensitivity, specificity and accuracy in the three datasets analyzed were estimated as: (i) sensitivity = number of correct presence calls/total number of presence calls in the validation dataset; (ii) specificity = number of correct absence calls/total number of absence calls in the validation dataset; and (iii) accuracy = (number of correct presence calls + number of correct absence calls)/(total number of presence calls + total number of absence calls in the validation dataset). These calculations were performed using a *Python* script.

For *D.melanogaster*, we used the same individual PCR frequencies used in *T-lex2* (Fiston-Lavier *et al.*, 2015), except those for one strain (*RAL-730*) for which the raw reads were no longer available.

For humans, we used the information based on PCRs and/or mapping algorithms for 1600 insertions available in Stewart *et al.* (2011).

For rice, validation was done as follows: we annotated LTRs and MITE insertions in the MH63 genome (Zhang *et al.*, 2016) following the strategy described above for the Nipponbare genome. We compared the orthologous loci containing LTR and MITEs to determine if they were present in both genomes (orthologous insertions), or only in Nipponbare (Nipponbare-specific). To this end, the 500 bp flanking regions of every LTR and MITE insertions in Nipponbare were mapped onto the rice MH63 assembly using BBmap (https://sourceforge.net/projects/bbmap/). The distribution of the distances between the MH63 regions aligning to the two Nipponbare flanks (D) followed closely that of the length of the MITEs and LTR insertions in Nipponbare (L) (Supplementary Additional File S2), which is to be expected if the insertions are present in both genomes (orthologous), except for an additional peak centered at distance 0, which should correspond to Nipponbare specific insertions. The manual inspection of some of these loci confirmed this assumption. The insertions were then classified based on the comparison of the distance between the two aligned flanks (D) and the length of the original Nipponbare element (L), as follows: when $1.33xL > D < 0.66xL$ the insertion was considered as being present in both genomes (orthologous insertion); when $D < 50$ bp (LTRs) or $D < 15$ bp (MITEs), the insertion was considered as Nipponbare specific.

2.6 *T-lex3* processing times

For an individual *D.melanogaster* genome of 134 Mb, using a 18.6x sequencing dataset of 73 bp reads, *T-lex3* processing time was ~ 10.5 hours, while for *T-lex2* processing time was 14 hours. Both runs were performed on a standard computer with a 2.8 GHz Intel Core i5 with 16 GB RAM memory, and both versions were executed with a single thread.

In humans (genome size 3 Gb), *T-lex3* processing time was 72.5 hours for a 3.1× sequencing dataset of 101 bp reads, for a dataset of 1600 TEs. In rice (genome size 362 Mb), *T-lex3* processing time was 38 hours for a 23.7× sequencing dataset of 100 bp read length, for a dataset of 1584 TEs. Both runs with human and rice genomes were performed with a single thread on a *Linux* Cluster using two different nodes: one with 16 cores (126 Gb of RAM) Intel® Xeon® CPU ES-2630 v3 with 2.40 GHz and one node with 48 cores (512 GB of RAM) Intel® Xeon® CPU E5-4640(4660) v4 with 2.10 Ghz.

3 Results

We have improved and updated *T-lex*, a pipeline that genotypes and estimates reference TE insertion frequencies using individually or pooled short-read high-throughput sequencing data. Briefly, *T-lex3* now uses *BWA-MEM* instead of *MAQ* as the mapping algorithm to detect the presence of insertions. We have added several extra filtering steps and parameters, we have fixed bugs in the pipeline, and we have also introduced some more general changes that overall improves the user experience (Table 2).

3.1 *T-lex3* genotypes and estimates TE population frequencies in datasets that could not be analyzed with *T-lex2* in *D.melanogaster*

We have run *T-lex3* in three datasets that gave good results with *T-lex2* (>60% of *present*, *polymorphic* or *absent* calls): 19 individual strains from the DGRP dataset (Huang *et al.*, 2014), 11 individual strains from the DPGP dataset (Pool *et al.*, 2012), and 10 pooled datasets (Kapun *et al.*, 2018) (Supplementary Additional File S1). The 19 individual strains from the DGRP dataset had a variable number of *no data* calls: from 146 in RAL-757 to 585 in RAL-399 (Supplementary Additional File S1A and B). *T-lex3* substantially reduced the number of *no data* calls. The maximum number of *no data* calls is now 374 in RAL-399 with some strains having as few as 83 (RAL-738, Fig. 3A and Supplementary Additional File S1A and B). The number of *no data* calls according to *T-lex2* in Zambia strains was smaller, ranging from

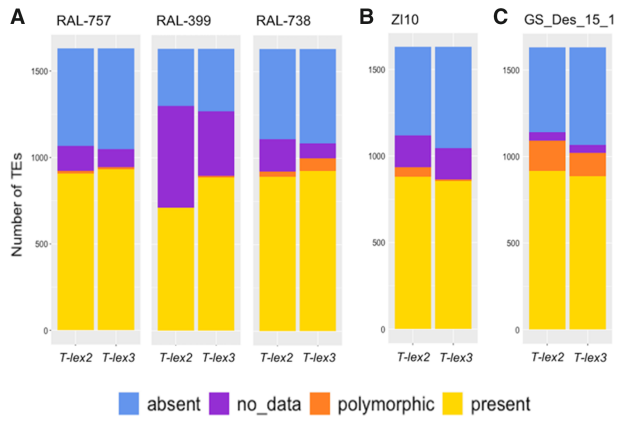

Fig. 3. Comparison between *T-lex2* and *T-lex3* results. Each column represents results for one strain. (**A**) Three different DGRP individually sequenced strains: RAL-757, RAL-399 and RAL-738; (**B**) one Zambian individually sequenced strain: ZI10 and (**C**) one pool-seq European sample: GS_Des_15_1

156 to 344, and this number was also substantially reduced in *T-lex3*: 128 to 181 (Fig. 3B, Supplementary Additional File S1C and D). For pool-seq datasets, we found that the correlation between the frequency estimates obtained using the two versions of *T-lex* was very high (Pearson correlation r^2 = 0.92–0.97, Fig. 3C and Supplementary Additional File S1E and F), as expected since the number of *no data* calls was similar between the two versions.

To estimate the sensitivity, specificity and accuracy of *T-lex3*, we used the same DGRP strains previously used to estimate these parameters for *T-lex2* (see Section 2). We found that both sensitivity, specificity and accuracy are high: 100, 93.33 and 97.14%, respectively (Table 4, Supplementary Additional File S3A), with similar results among different TE classes (Supplementary Additional File S3B). Thus, *T-lex3* increases the number of *present*, *polymorphic* and *absent* calls with a very similar sensitivity and slightly higher specificity compared with *T-lex2*. As this DGRP dataset contains strains with different read lengths, we also estimated sensitivity and specificity using different *T-lex3* parameters. We found that specificity decreased and the number of *no data* calls increased when sub-optimal parameters are used (Supplementary Additional File S3C). We then estimated the specificity for this same dataset using TIDAL, a method that identifies the absence of reference TE insertions and also discovers non-reference TE insertions (Rahman *et al.*, 2015). The specificity of TIDAL was 64.36%.

Finally, we ran *T-lex3* with five datasets that previously failed to give results: two individual strains datasets and three pooled datasets. For all five datasets, *T-lex3* is able to estimate population frequencies for the majority of TE insertions tested (Supplementary Additional File S4). As an example, for 78 of the 86 strains collected in Lyon, the number of *no data* calls with *T-lex2* was >40%. For this same dataset, *T-lex3* only returns >40% of *no data* calls for 1 of the 86 strains (Fig. 4, Supplementary Additional File S4A).

3.2 *T-lex3* provides accurate TE genotyping and frequency estimations in humans

We also tested the performance of *T-lex3* in a human dataset: the child genome of the trio human dataset in Stewart *et al.* (2011) (Supplementary Additional File S5A). Out of the 1600 insertions tested, *T-lex3* returns *present*, *absent* or *polymorphic* calls for the majority of them (82.5%: 1320/1600). Sensitivity and specificity were high: 99.35% and 87.78%, respectively, and accuracy was 91.81% (Table 4, Supplementary Additional File S5B), with similar results for the different families tested (Supplementary Additional File S5C). We estimated the specificity in this same dataset using the two algorithms described in Stewart *et al.* (2011): the read-pair (RP) algorithm and the split-read (SR) algorithm. The specificity using RP was 99.27% and using SR was 88.89%.

Table 4 *T-lex3* sensitivity, specificity and accuracy in *D.melanogaster*, humans and rice genomes

	D.melanogaster	Humans	Rice
Sensitivity	100% (120/120)	99.35% (458/461)	99.96% (2649/2650)
Specificity	93.33% (84/90)	87.78% (754/859)	92.11% (257/279)
Accuracy	97.14% (204/210)	91.81% (1, 212/1, 320)	99.2% (2, 855/2, 878)

Notes: In parenthesis, number of calls used to estimate the three parameters.

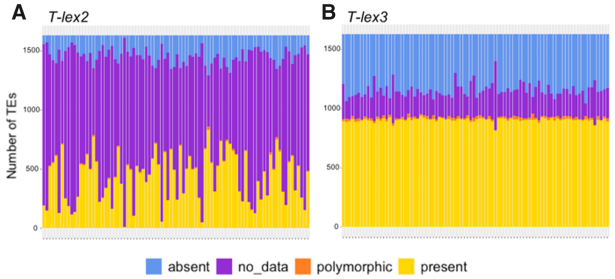

Fig. 4. *T-lex2* (**A**) and *T-lex3* (**B**) results comparison for 86 individual strains from Lyon (France). Each column represents results of one strain. Strains are placed in the same order as in Supplementary Additional File S4A

3.3 *T-lex3* accurately genotypes and estimates TE population frequencies in the rice genome

We tested *T-lex3* in a rice genome using our own annotated insertions (see Materials and Methods). Out of the 3067 insertions in our dataset, *T-lex3* returns *present*, *absent* or *polymorphic* calls for 2878 (93.84%) (Supplementary Additional File S6A). According to our validation dataset, sensitivity and specificity were 99.96% and 92.09% respectively while accuracy was 99.2% (Table 4, Supplementary Additional File S6B), with similar results for the different TE classes tested (Supplementary Additional File S6C). These results demonstrated that the *T-lex3* pipeline performance is also robust with rice genomes.

4 Discussion

Genome-wide genotyping of TE insertions is crucial in developing a complete catalog of genetic variants that can then be investigated for their potential role in adaptive evolution and/or disease. The repetitive nature of TEs complicates the correct genotyping of these variants and several computational pipelines have been designed to tackle this problem (Table 1). Ideally, these computational pipelines should be able to handle (i) different types of datasets, such as individual and pooled whole genome sequencing datasets; (ii) different lengths of short-reads; and (iii) a variety of TE insertions in diverse genomes (Table 1). In this work, we have improved and updated *T-lex*, a flexible and broadly applicable tool that provides accurate TE genotyping and frequency estimations for the different TE families present in a reference genome using both individual and pooled datasets. Compared to the previous *T-lex* version available, *T-lex3* is a more robust and stable tool, easier to install, and that runs with all the short-reads available, independent of their length. We showed that this new version of *T-lex* provides TE genotyping with higher sensitivity, specificity and accuracy than its previous versions, and similar or higher compared with other tools, while substantially increasing the number of TE reference insertions that can be analyzed. We also showed that *T-lex3* provides accurate calls not only in the *D.melanogaster* and human genomes as previous versions, but also in a plant genome: *Oryza sativa*. As such, *T-lex3* is one of the tools that is most broadly applicable to genomes differing not only in size but also in TE content (Table 1).

Being able to genotype and estimate population frequencies for the majority of TE insertions present in a genome is necessary to generate a

more complete picture of the existing genomic variation within and among individuals, populations and species. Although there are several available tools that identify non-reference TE insertions, they have a high rate of false positives (Rahman *et al.*, 2015; Rishishwar *et al.*, 2017). Thus, we are still limited to analyzing the TE insertions that have been annotated in the available reference genomes. Long-read sequencing technologies should allow the generation of new reference genomes and better *de novo* annotations of TE insertions (van Dijk *et al.*, 2018). These new long-read sequencing technologies are already being applied to model and non-model species (Miller *et al.*, 2018; Solares *et al.*, 2018). However, the number of new genome assemblies is still small as the cost of these technologies remains higher compared with short-read ones. Thus, once the coordinates of new TE insertions are defined based on the new genome-assemblies, *T-lex3* would be useful to estimate the frequencies of these new insertions in all the short-read datasets already available (Auton *et al.*, 2015; Guirao-Rico and González, 2019).

Acknowledgements

We thank members of the González Lab and Quentin Testard for comments on the manuscript, and Óscar de Arriba for helping improve the code.

Funding

This work was supported by the European Commission (H2020-ERC-2014-CoG-647900) and by the Secretaria d'Universitats i Recerca del Departament d'Economia i Coneixement de la Generalitat de Catalunya (GRC 2017 SGR 880). Work done at CRAG was partially funded by a grant from the Ministerio de Economia y Competitividad (AGL2016-78992-R).

Conflict of Interest: none declared.

References

Alkan,C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Auton,A. *et al.* (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

Bastide,H. *et al.* (2013) A genome-wide, fine-scale map of natural pigmentation variation in *Drosophila melanogaster*. *PLoS Genet.*, **9**, e1003534.

Bleykasten-Grosshans,C. and Neuvéglise,C. (2011) Transposable elements in yeasts. *CR Biol.*, **334**, 679–686.

Campbell,I.M. *et al.* (2014) Human endogenous retroviral elements promote genome instability via non-allelic homologous recombination. *BMC Biol.*, **12**, 74.

Campo,D. *et al.* (2013) Whole-genome sequencing of two North American *Drosophila melanogaster* populations reveals genetic differentiation and positive selection. *Mol. Ecol.*, **22**, 5084–5097.

Chen,J. *et al.* (2014) P-MITE: a database for plant miniature inverted-repeat transposable elements. *Nucleic Acids Res.*, **42**, D1176–1181.

Chuong,E.B. *et al.* (2017) Regulatory activities of transposable elements: from conflicts to benefits. *Nat. Rev. Genet.*, **18**, 71–86.

Eddy,S.R. (2011) Accelerated profile HMM searches. *PLoS Comput. Biol.*, **7**, e1002195.

Elbarbary,R.A. *et al.* (2016) Retrotransposons as regulators of gene expression. *Science*, **351**, aac7247.

Ellinghaus,D. *et al.* (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, **9**, 18.

Fiston-Lavier,A.S. *et al.* (2015) T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.*, **43**, e22.

Fiston-Lavier,A.S. *et al.* (2011) T-lex: a program for fast and accurate assessment of transposable element presence using next-generation sequencing data. *Nucleic Acids Res.*, **39**, e36.

Fu,L. *et al.* (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, **28**, 3150–3152.

Gardner,E.J. *et al.* (2017) The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. *Genome Res.*, **27**, 1916–1929.

Goerner-Potvin,P. and Bourque,G. (2018) Computational tools to unmask transposable elements. *Nat. Rev. Genet.*, **19**, 688–704.

Guio,L. and González,J. (2019) New insights on the evolution of genome content: population dynamics of transposable elements in flies and humans. *Methods Mol. Biol.*, **1910**, 505–530.

Guirao-Rico,S. and González,J. (2019) Evolutionary insights from large scale resequencing datasets in *Drosophila melanogaster*. *Curr. Opin. Insect Sci.*, **31**, 70–76.

Han,Y. and Wessler,S.R. (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.*, **38**, e199.

Hatem,A. *et al.* (2013) Benchmarking short sequence mapping tools. *BMC Bioinformatics*, **14**, 184.

Hoban,S. *et al.* (2016) Finding the genomic basis of local adaptation: pitfalls, practical solutions, and future directions. *Am. Nat.*, **188**, 379–397.

Huang,W. *et al.* (2014) Natural variation in genome architecture among 205 *Drosophila melanogaster* Genetic Reference Panel lines. *Genome Res.*, **24**, 1193–1208.

Kapun,M. *et al.* (2018) Genomic analysis of European Drosophila populations reveals longitudinal structure and continent-wide selection. *bioRxiv*, 313759.

Kofler,R. *et al.* (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in *Drosophila melanogaster*. *PLoS Genet.*, **8**, e1002487.

Kofler,R. *et al.* (2016) PoPoolationTE2: comparative population genomics of transposable elements using Pool-Seq. *Mol. Biol. Evol.*, **33**, 2759–2764.

Lack,J.B. *et al.* (2016) A thousand fly genomes: an expanded drosophila genome nexus. *Mol. Biol. Evol.*, **33**, 3308–3313.

Lee,J.A. *et al.* (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell*, **131**, 1235–1247.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,H. *et al.* (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Mascher,M. *et al.* (2017) A chromosome conformation capture ordered sequence of the barley genome. *Nature*, **544**, 427–433.

Miller,D.E. *et al.* (2018) Highly contiguous genome assemblies of 15. *G3 (Bethesda)*, **8**, 3131–3141.

Pool,J.E. *et al.* (2012) Population genomics of Sub-Saharan *Drosophila melanogaster*: African Diversity and Non-African Admixture. *PLoS Genet.*, **8**, e1003080.

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

Rech,G.E. *et al.* (2019) Stress response, behavior, and development are shaped by transposable element-induced mutations in Drosophila. *PLoS Genet.*, **15**, e1007900.

Rishishwar,L. *et al.* (2017) Benchmarking computational tools for polymorphic transposable element detection. *Brief. Bioinform.*, **18**, 908–918.

Robb,S.M. *et al.* (2013) The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. *G3 (Bethesda)*, **3**, 949–957.

Robinson,J.T. *et al.* (2017) Variant review with the integrative genomics viewer. *Cancer Res.*, **77**, e31–e34.

Santander,C.G. *et al.* (2017) STEAK: a specific tool for transposable elements and retrovirus detection in high-throughput sequencing data. *Virus Evol.*, **3**, 23.

Sasaki,T. and International Rice Genome Sequencing,P. (2005) The map-based sequence of the rice genome. *Nature*, **436**, 793–800.

Schnable,P.S. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.

Solares,E.A. *et al.* (2018) Rapid low-cost assembly of the *Drosophila melanogaster* reference genome using low-coverage, long-read sequencing. *G3 (Bethesda)*, **8**, 3143–3154.

Startek,M. *et al.* (2015) Genome-wide analyses of LINE-LINE-mediated nonallelic homologous recombination. *Nucleic Acids Res.*, **43**, 2188–2198.

Stewart,C. *et al.* (2011) A comprehensive map of mobile element insertion polymorphisms in humans. *PLoS Genet.*, **7**, e1002236.

van Dijk,E.L. *et al.* (2018) The third revolution in sequencing technology. *Trends Genet.*, **34**, 666–681.

Villanueva-Cañas,J.L. *et al.* (2017) Beyond SNPs: how to detect selection on transposable element insertions. *Methods Ecol. Evol.*, **8**, 728–737.

Zhang,J. *et al.* (2016) Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl. Acad. Sci. USA*, **113**, E5163–5171.

Zhuang,J. *et al.* (2014) TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.*, **42**, 6826–6838.