Particular sequence characteristics induce bias in the detection of polymorphic transposable element insertions

Marie Verneret[1,2], Van Anthony Le[1], Thomas Faraut[3], Jocelyn Turpin[2], Emmanuelle Lerat[1*]

[1] Universite Claude Bernard Lyon 1, LBBE, UMR5558, CNRS, VAS, Villeurbanne, F-69622, France.

[2] IVPC UMR754, INRAE, Universite Claude Bernard Lyon 1, EPHE, Université PSL, Lyon, F-69007, France

[3] GenPhySE, Universite de Toulouse, INRAE, INPT, ENVT, 31326, Castanet Tolosan, France

*corresponding author

Abstract

Transposable elements (TEs) have an important role in genome evolution but are challenging for bioinformatics detection due to their repetitive nature and ability to move and replicate within genomes. New sequencing technologies now enable the characterization of nucleotide and structural variations within species. Among them, TE polymorphism is critical to identify as it may influence species adaptation or trigger diseases. Despite the development of numerous bioinformatic programs, identifying the most effective tool is challenging due to non-overlapping results and varying efficiency across studies. Benchmarking efforts have highlighted the limitations of these tools, often evaluated on either real or simulated data. However, real data may be incomplete or contain unannotated TEs, while simulated data may not accurately reflect real genomes. This study introduces a simulation method generating data based on real genomes to control all genomic parameters. Evaluating several TE polymorphic detection tools using data from *Drosophila melanogaster* and *Arabidopsis thaliana*, our study investigates factors like copy size, sequence divergence, and GC content that influence detection efficiency. Our results indicate that only a few programs perform satisfactorily and that all are sensitive to TE and genomic characteristics, that may differ according to the species considered. Using *Bos taurus* population data as a case study to identify polymorphic LTR-retrotransposon insertions, we found low-frequency insertions particularly challenging to detect due to a high number of false positives. Increased sequencing coverage improved sensitivity but reduced precision. Our work underscores the importance of selecting appropriate tools and thresholds according to the specific research questions.

Introduction

Recognized as being among the most important players in the evolution of genomes, transposable elements (TEs) represent a real challenge for bioinformatics approaches to detect them. TEs are repeated sequences present in almost all eukaryotic genomes. They have the ability to move and replicate, forming different families of similar but not always identical sequences. Several types have been described, depending on their structure and their mode of transposition, varying both in genomic distribution and in sequence length (Wicker et al. 2007). For example, LTR-retrotransposons represent sequences of approximately 10 kb but DNA transposons such as MITEs (Miniature Inverted Repeats Transposable Elements) span only a few hundred base pairs. Moreover, TEs are not randomly distributed in the genome since their insertion patterns reflect a balance between selection pressure against their deleterious effects and genetic drift (Bourque et al. 2018). As a consequence, TEs are likely to be found inserted into each other constituting nested insertions, which are particularly difficult to automatically identify (Bergman and Quesneville 2007). In addition, their proportion in genomes can vary greatly, ranging from a few percent as for example in the honeybee (Weinstock et al. 2006) to the major part of the genome as in maize (Schnable et al. 2009). Over the past twenty years, different bioinformatic tools have been developed allowing their annotation in assembled genomes (Lerat 2019). However, the rapid development of new sequencing technologies has made it possible to access numerous data from different individuals or populations in order to characterize the nucleotide and structural variations within a given species. Indeed, a reference genome for a given species is not sufficient to reflect the overall diversity of individuals. In particular, although TEs are generally regulated in a genome to prevent their activity, certain TE families can nevertheless continue to transpose throughout the life of an individual or may be reactivated due to some stress (Di Stefano 2022). It has been proposed that in Drosophila, the transposition rate is comparable to that of the nucleotide mutation rate (Adrion et al. 2017). More recently, according to the TE family, the transposition rate has been shown to be higher with an average of $4.93 \times 10{-}9$ insertions per site per generation corresponding to a new insertion in each new embryo (Wang et al. 2023). In human, the most active TEs have a transposition rate of one insertion every 20 births (Cordaux and Batzer 2009). We can thus expect to find variations in the TE insertion pattern between individuals, which constitutes the TE polymorphism. Polymorphic TEs are particularly important to identify since they represent insertions that may be at the basis of species/population adaptation or triggering diseases. For example, numerous polymorphic TEs have been detected in sub-populations of the Chinese white poplar (*Populus tomentosa)* some of them being under positive selection while inserted in genes involved in stress, defense and immune

responses (Zhao et al. 2022). In humans, a specific polymorphic TE insertion is associated with the development of the Fukuyama type congenital muscular distrophy (Kobayashi et al. 1998).

In order to search for polymorphic insertions, bioinformatics tools have been developed to answer specific questions and on particular organisms such as Drosophila, human or some plants (Lerat 2019). All these methods follow similar principles in their functioning which consist first in mapping sequenced reads to a reference genome and a set of reference TE sequences. Then two approaches, that can be combined, have been proposed to detect the presence/absence of TEs. The first is to consider discordant read pairs with one read mapping uniquely on a genomic location and the other mapping on different sequences of a same TE family. The second approach considers split reads, *i.e.*, reads overlapping a junction between the genome and a TE insertion, with a part of the read mapping uniquely on the genome while the other part maps on several TE sequences. More than twenty programs have been developed during the past ten years (for an exhaustive list, see https://tehub.org/), which makes it difficult for users to determine which program is the most appropriate or the most efficient. In particular, the results of these programs are often not entirely overlapping (Ewing 2015; Lerat et al. 2019). This makes it more difficult to identify true positives, especially in the case of new insertions. Some attempts haven been previously made to benchmark all these programs (Nelson et al. 2017; Rishishwar et al. 2017; Vendrell-Mir et al. 2019, Chen et al. 2023). These works showed that all these programs generally are not as efficient as indicated in their original publication. However, these evaluations were made either on partial real data or on fully simulated data, or were targeting only particular TE types. A problem with real data is that they may be only partial or may contain unannotated TE insertions that can blur the results. However, using only simulated data is also problematic since it usually does not reflect in a realistic manner a real genome. It therefore remains a certain number of unanswered questions to explain the lack of performance of certain tools, in connection with some peculiarities of the studied genome.

In this study, we have developed a simulation tool to produce data based on real genomes to allow the complete control of all genomic parameters. Using data generated for *Drosophila melanogaster* and *Arabidopsis thaliana*, we evaluated several TE polymorphic detection tools and investigated different characteristics like the copy size, the sequence divergence, the distance between copies, the GC content of the surrounding genomic regions, the Target Site Duplicate (TSD) size or the TE family that could explain why some insertions are better detected than others. Our results show that only very few of the different tested programs give satisfactory results and that all programs are sensitive to TE and genomic sequence characteristics that slightly differ according to the species considered. As an application case, we used *Bos taurus* real population data to identify polymorphic LTR-retrotransposon insertions. Low-frequency insertions appeared to be

more challenging to detect due to a high proportion of false positives. Increasing sequencing coverage improved the sensitivity but at the expense of precision. Our study emphasizes the importance of selecting appropriate tools and thresholds depending on the scientific questions asked.

Material and Methods

Genomic data used for simulation

The sequence of the *Drosophila melanogaster* 2L chromosome in GenBank format was obtained from the NCBI GenBank database (accession number: NT_033779). The chromosome sequence is 23,513,712 bp long in which 3,519 genes and 919 TEs are annotated. For *Arabidopsis thaliana*, a GenBank file of the chromosome 1 was generated using the TAIR10 gene and transposable element (TE) annotation in gff format available from the Arabidopsis Information Resource website (https://www.arabidopsis.org/). The chromosome sequence is 30,427,671 bp long in which 7,509 genes and 7,135 TEs are annotated. The sequence of the chromosome 25 from *Bos taurus* was obtained from the GenBank database (version ARS-UCD1.3, accession number: GCF_002263795.2). The chromosome, that is 42,350,435 bp long, contains 1,006 genes but no TEs have been previously annotated. We thus determined the position of endogenous retroviruses (ERV) using RepeatMasker version 2.0.3 with cattle ERV consensus sequences from Repbase version 29.03 (https://www.girinst.org/). ERV insertions from four ERV families were used for the simulations: two class I ERV families (ERV1-1_BT and BtERVF2) and two class II ERV families (ERV2-2_BT and ERV2-3_BT).

Simulation tool *replicaTE*

We have developed a simulation tool based on real data. This tool is implemented as several python3 scripts that need to be run successively, using as a starting point a GenBank file (Figure 1). In summary, we consider three types of sequences (genes, TEs and intergenic regions). The genes are cleaned up from any TE insertions. Intergenic regions are simulated to remove any misannotated TEs and based on the real intergenic regions (%GC, length). The characteristics of TEs (number of copies, size of copies, %divergence, etc.) are used to simulate new TE sequences. These TE sequences are randomly assigned to the intergenic regions. Finally, all three parts are reassembled to create a complete simulated genome and a deleted simulated genome in which half of the TE insertions are removed. The tool is available as a git repository (https://github.com/e-lerat/replicaTE).

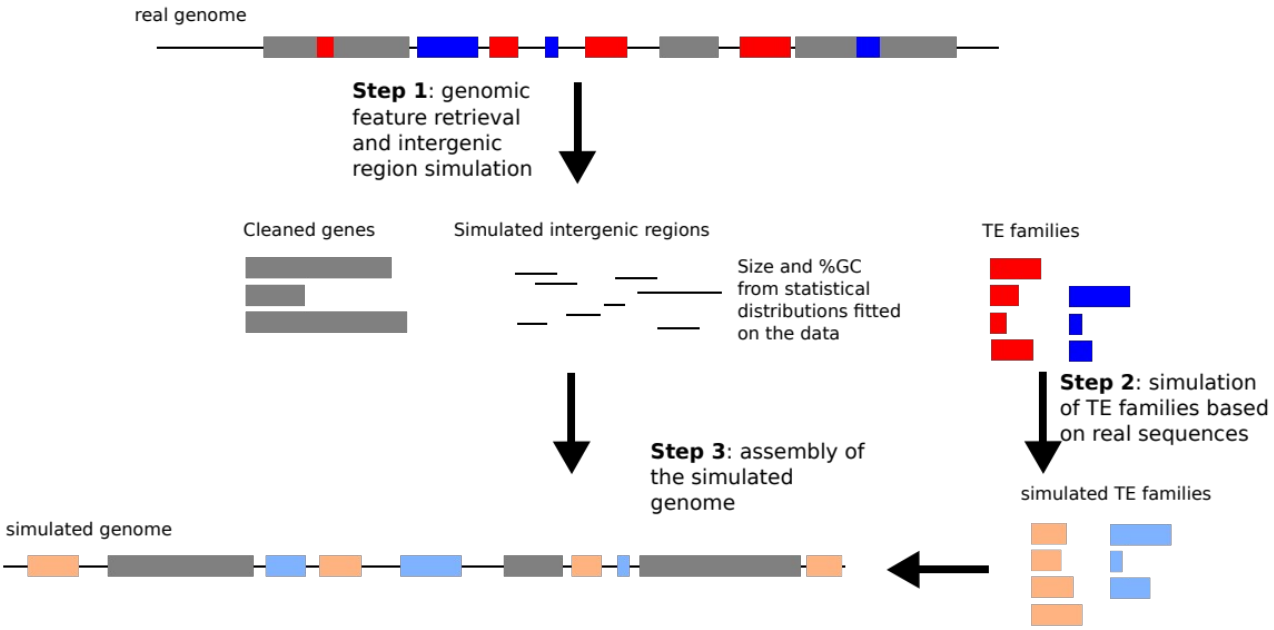


Figure 1: Workflow of the simulation tool *ReplicaTE*. Each step corresponds to the different scripts. In gray are indicated the genes, in blue and red are indicated the TE insertions, a given color corresponding to a given TE family. The simulated TE copies are represented in orange and light blue.

deleTE.py

This script allows to get the characteristics of each element (genes, intergenic regions, TEs) for the next steps and to generate a simulated genome without TEs. It takes as an input file a GenBank file from which it will extract the annotations. It outputs multiple files which can then be used as input by the other codes. The size of the simulated intergenic regions are drawn from an exponentiated Weibull distribution constructed from the computed gene density (number of genes per Mb) with a minimal size of 200 pb. The GC content of the simulated intergenic regions are drawn from a truncated normal law fitted on the observed %GC of the genome, with values comprised between 0 and 100%.

generaTE.py

This script generates TE copies based on different characteristics (copy number, length, Target Site Duplication (TSD) length, strand). It attributes an intergenic region to each copy to be inserted into with the possibility to have nested insertions. For each family, a pool of copy sizes is drawn in a truncated exponential law, with values between 80 bp and 102.5% of the largest sequence of the family to take into account potential small insertions, called the “ancestral” sequence. The sequence divergences of the copies compared to the “ancestral” sequence are drawn from a truncated normal law distribution, with values comprised between 0 and 20% (mean = 10 and standard deviation =

4). By default, the copy number corresponds to the observed copy number in the real genome. It is also possible to simulate the copy number. In that case, it is randomly drawn from an exponentiated Weilbull distribution fitted on the data. The TSD have a length between 0 and 8 bp and are attributed for a given family when the option is specified.

inseraTE.py

This script associates the cleaned genes, the simulated TEs and the simulated intergenic regions to produce a genomic sequence. The TE copies are randomly inserted into their attributed intergenic region. The insertion can be 'normal' or 'nested' (inserted into a previous TE) and multiple nested events can arise. The simulated genome is provided in fasta format. A deleted version of the simulated genome is also generated, in which half of the TE copies are not present.

Short-read simulation

The different tested tools all use short-read sequences as an input. We thus have generated short reads based on the simulated genomes using the program ART Version 2.1.8 (Huang et al. 2012). This program produces theoretical reads expected by an NGS technique on a given genome. For this analysis, we generated paired-end Illumina reads of 150 bp (with a fragment size of 300 bp) using three different coverages (10X, 50X and 100X). Only 15X short reads were produced for *B. taurus* to reflect the landscape of the real cattle data coverage in the public databases.

Statistical analyses

All statistical tests were performed using the R software version 3.6.3 (2020-02-29) (R Core Team 2017). The programs were evaluated according to different metrics described below.

Recall (sensitivity): it corresponds to the proportion of True Positives (TP) among all the TE insertions present in the reference genome. It is computed as: $\frac{TP}{TP+FN}$

Precision: it corresponds to the proportion of good answers among the predicted TE insertions. It is computed as: $\frac{TP}{TP+FP}$

F-score: it corresponds to the harmonic mean of the recall and the precision. It is computed as:

$$2\frac{recall\,.\,precision}{recall+precision}$$

To compute these different metrics, it is necessary to assess the number of TPs among the identified TE insertions proposed by each program. We thus considered an insertion to be a TP when the program associates the same TE family name and a position that is close to the real position, with a certain margin of error. More specifically, we considered four different margins of error to determine whether the position was correct or not which are 5 bp, 20 bp, 100 bp and 150 bp. The False Negatives (FN) correspond to insertions present in the reference dataset that were not detected by the program and the False Positives (FP) correspond to predicted insertions that do not correspond to insertions present in the reference dataset.

False positive rate estimation in real data

Endogenous retroviruses (ERV) insertion detection was performed on 10 samples of short-read data from various individuals of *Bos taurus* (in prep). The same protocol as above was applied to compute the proportion of TPs, FPs and FPs among the identified insertions. We considered insertions as TPs if the program associated the same ERV family name with a position within 20 bp of error margin. For the reference insertions, the ERV annotation of the *B. taurus* assembly was considered as a reference. For the new insertions identified, a variant calling analysis was performed on long-read data from the same samples using the pbsv variant caller with default parameters (https://github.com/PacificBiosciences/pbsv). The variant output file was considered as a reference for the classification of TP and FP new insertions.

Results

Chromosome simulation and evaluation approach

The simulation tool *replicaTE* was used on the chromosome 2L of *D. melanogaster* and on the chromosome 1 of *A. thaliana* (all generated files are available as supplementary data). The first script, *deleTE.py*, produces different output files. Among them, the “gene_clean_tab.csv” file contains the real genes without any annotated internal TE insertions. The “intergenic_sim_tab.csv” file contains the simulated intergenic regions with their length and %GC. The “stat_TEs_tab.csv” contains a sequence corresponding to the longest TE sequence (considered as the “ancestral” TE sequence) of a given family that will be used to generate all simulated TE copies and the number of copies for each family, that corresponds to the real number of annotated copies in the considered chromosome. These two last files are used in the second script, *generaTE.py*, to simulate the TE copies. It produces a fasta file containing the simulated sequences (“simulated_TEs.fas”) and a text file (“param_TEs_tab.csv”) containing all the information regarding each TE family (length of each

copy, sequence divergence of each copy compared to the “ancestral” TE sequence, the associated intergenic region, the strand and the TSD size). These two files, in addition to the “gene_clean_tab.csv” and the “stat_TEs_tab.csv” files, are used in the third script *inseraTE.py.* It produces, among other files, the two simulated genomes in fasta format and the files “annot_TEs.tsv” and “annot_TEs_del_1” containing all information regarding each TE copy (positions, length, divergence, insertion type (nested or not), strand, TSD size, distances to the closest TE insertions, and the GC content of the flanking genomic regions).

For each chromosome, we thus have all information about the inserted copies in addition to their precise positions. These different parameters will be used to evaluate the tested programs. In particular, we will be able to determine if particular factors relative to the TE sequences (size, distance to other copies, divergence, TSD size) or to their genomic region of insertion (%GC) may have an influence on whether they are correctly detected or not by the tested programs.

In our evaluation approach, the complete simulated chromosome and the deleted simulated chromosome can be used alternatively as reference genome or as sample genome in order to evaluate the possibility to identify reference / absent insertions or new insertions. Indeed, as described on the Figure 2, when using the complete simulated genome as a reference, the short reads will be generated using the deleted simulated genome, in which half of the TE insertions are missing. This will allow us to evaluate the capacity of the programs to detect both reference and absent insertions. Alternatively, if the deleted simulated genome is used as a reference and the complete simulated genome is used to generate short reads, then it will allow us to evaluate the capacity of the programs to detect new TE insertions.

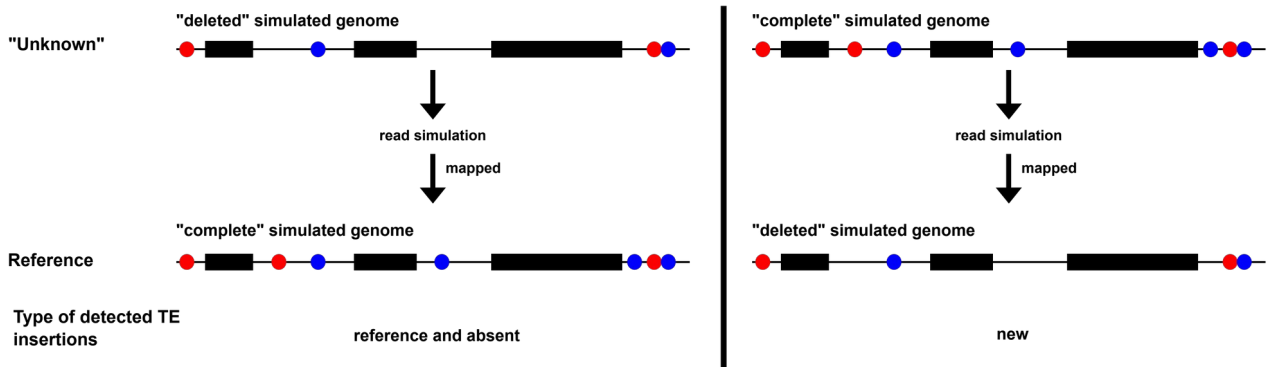


Figure 2: Evaluation approach. The black rectangles correspond to genes. The red and blue circles correspond to TE copies from two different families. The reads simulated on the “deleted” simulated genome will be mapped to the “complete” simulated genome, which will allow to identify reference and/or absent TE insertions. The reads simulated on the “complete” simulated genome will be mapped to the “deleted” simulated genome, which will allow to identify new TE insertions.

Using the *D. melanogaster* and *A. thaliana* data, we generated two simulated chromosomes for each species. For Drosophila, the complete simulated chromosome, which is 23,298,325 bp long, contains 790 TEs whereas the deleted simulated chromosome contains 400 TEs. In the case of *A. thaliana*, the complete simulated chromosome, which is 37,444,832 bp long, contains 6,324 TEs whereas the deleted simulated chromosome contains 3,132 TEs. Knowing exactly the positions and name of each insertion, it is thus possible to compute the number of True Positives (TP), False Positives (FP) and False Negatives (FN) for each program allowing to determine their efficiency. Additionally, since we have all information about the different insertions for which we can control all associated parameters (size, distance, %GC etc.), it will be possible to compare the characteristics of the TP to those of the FN that could indicate any detection bias in the tested programs.

Tests of the polymorphic TE detection programs

More than 20 programs were proposed during the last 10 years to identify polymorphic TE insertions. However, many of them were not possible to evaluate in this analysis. Some programs were no longer available to be retrieved. Other programs were not flexible about the reference genome that can be used, unless modifying significantly the source code. We also did not test T-lex3 (Bogaerts-Márquez et al. 2020) since it cannot detect new TE insertions but only presence/absence of annotated TE insertions in a reference genome.

We finally tested 14 programs for which it was possible to use customized reference genomes (TEMP (Zhuang et al. 2014), TEMP2 (Yu et al. 2021), ngs_te_mapper (Linheiro and Bergman 2012), ngs_te_mapper2 (Han et al. 2021), PoPoolationTE (Kofler et al. 2012), PoPoolationTE2 (Kofler et al. 2016), RetroSeq (Keane et al. 2013), RelocaTE (Robb et al. 2013), RelocaTE2 (Chen et al. 2017), TEBreak (Schauer et al. 2018), TEFLoN (Adrion et al. 2017), TE-locate (Platzer et al. 2012), TEPID (Stuart et al. 2016), and Jitterbug (Hénaff et al. 2015)). Excepted TEPID and Jitterbug, the programs were used through the pipeline McClintock (Nelson et al. 2017, Chen et al. 2023). The programs are designed to find new insertions (when compared to a reference genome) and, excepted Jitterbug, shared insertions (between a reference genome and a genome under investigation). The programs were developed on particular organisms but sometimes tested on several of them (human, Drosophila, Arabidopsis, rice, mouse and Daphnia).

Detection of reference insertions

We first evaluated the capacities of the programs to identify reference insertions, that is to say, insertions present in the reference genome and in the genome from which the reads are produced.

For the *D. melanogaster* simulated genomes, it represents 400 insertions and in *A. thaliana*, it represents 3,133 insertions. On Figure 3, the total number of reference insertions found by each program is represented for each species, independently of the identification of true positives (TP). For TEPID, this number has been estimated by subtraction since the program provides information about the absence of reference insertions. As we can see, globally, the increase of coverage has little influence on the total number of reference insertions detected, excepted for PopoolationTE and PopoolationTE2, which do not find many insertions at 10X. For both species, the TEMP and TEMP2 programs, which have the same results, find far more reference insertions than expected. Other programs find more than expected reference insertions but to a lesser extent in *A. thaliana* (TEPID) and in *D. melanogaster* (TEFLoN for all coverage and PoPoolationTE2 for coverage 50X and 100X). PopoolationTE2, TEFLoN and ngs_te_mapper2 (for *A. thaliana*) and PopoolationTE, ngs_te_mapper2 and TEPID (for *D. melanogaster)* find a number of reference insertions close to what is expected. All the other programs find no or few reference insertions.

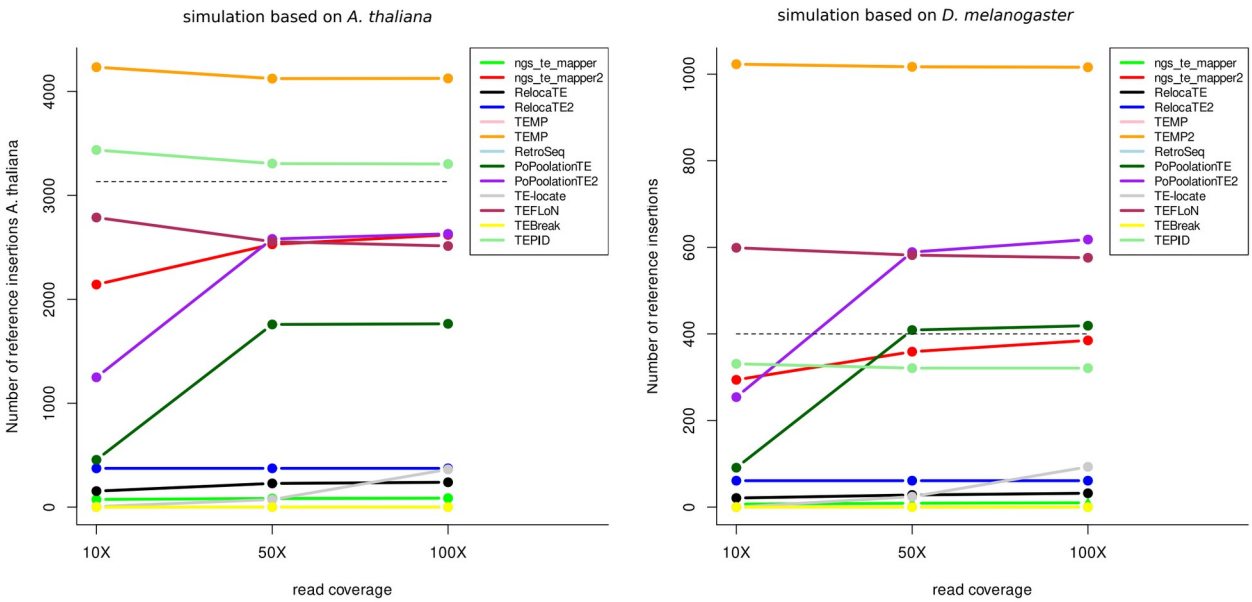


Figure 3: Number of reference insertions detected by each program. On the left panel is represented the number of reference insertions found for *A. thaliana* and on the right panel for *D. melanogaster*, for the three different read coverages. The dashed lines correspond to the expected number of reference insertions in each species.

We then determined among all these insertions the number of False Positives (FP), False Negatives (FN) and True Positives (TP) in order to compute various metrics to evaluate the programs. TPs were identified according to both the capacity of the program to identify the right TE family and according to the localization prediction with several margin of errors (see Material and

Methods). On Figure 4 are represented the different metrics for both species using a localization prediction with a margin of error of 20 bp (see supplementary figures S1, S2 and S3 for all cutoffs).

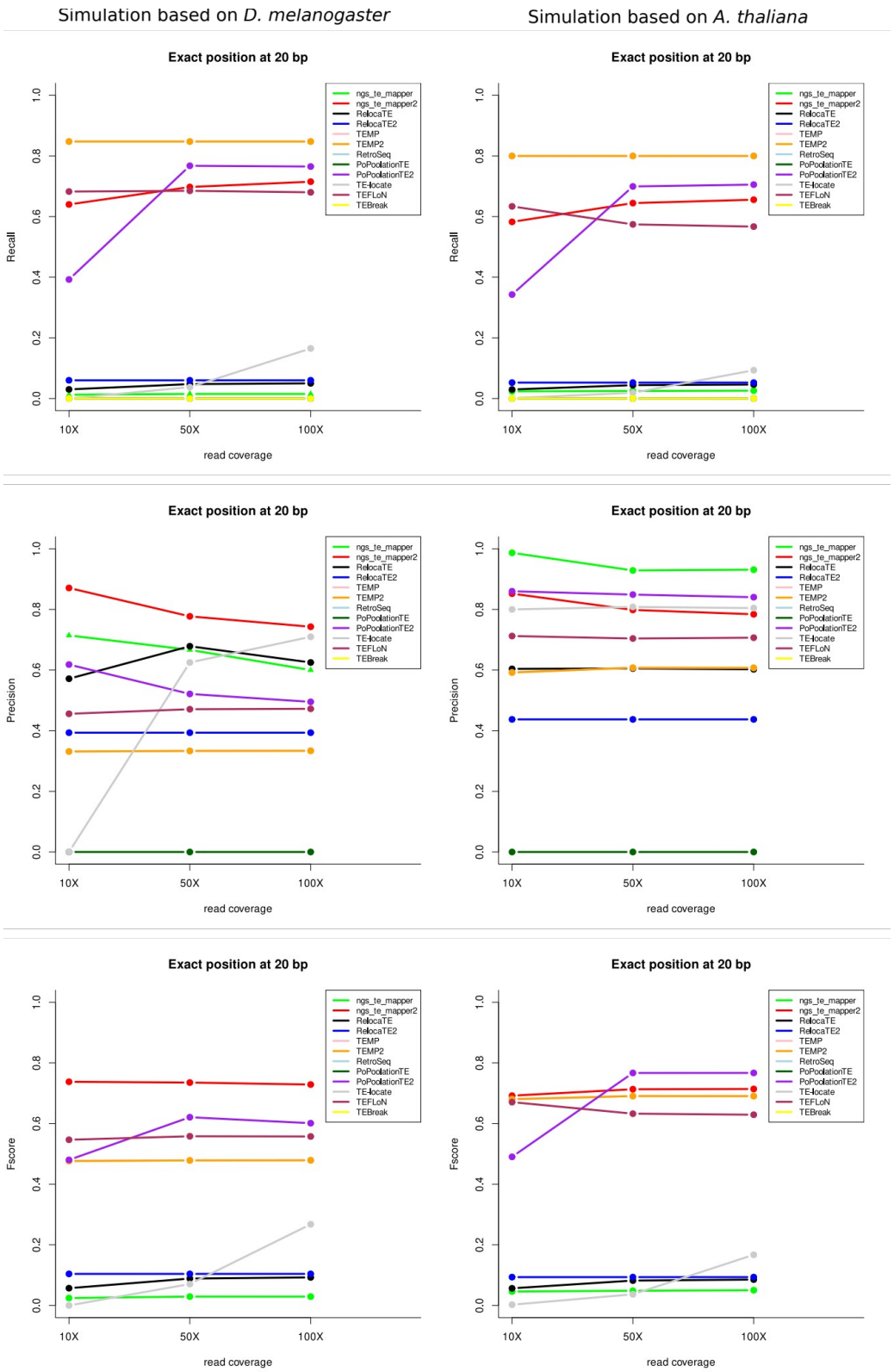


Figure 4: Evaluation metrics for the two species for the three read coverages for the reference insertions with a precision localization of 20 bp.

We retained this particular margin of error since at 5 bp all programs do not perform well whereas at 100 bp and 150 bp the efficiency of the programs is not improved. The *recall* metrics indicates the number of good answers among all the possible answers. In our case, it indicates for each program the number of TPs among all the TE insertions that should be detected. For both species, five programs give the best results for this metrics: TEMP, TEMP2, PoPoolationTE2 (starting at 50X), ngs_te_mapper2 and TEFLoN, with *recall* values of more than 0.5. The other programs find few or no TP among all the TE insertions that can be found given a localization window of 20 bp. The *precision* gives the number of good answers among all the results proposed by the programs. According to the species, the tools have not the same results. For *D. melanogaster*, ngs_te_mapper2 has the best results for this metrics, whereas it is ngs_te_mapper for *A. thaliana*. In order to take into account both metrics, we computed the *Fscore*. For both species, five programs give the best results: ngs_te_mapper2, PoPoolationTE2, TEMP, TEMP2, and TEFLoN. However, according to the species, the best program is not the same: ngs_te_mapper2 performs better for *D. melanogaster* when it is PoPoolationTE2 for *A. thaliana*.

We observed the overlap of TPs between the top four programs for each species (Figure 5). The results showed that 66.8% for *D. melanogaster* and 61.9% for *A. thaliana* of the TPs are found by the four programs. Among the remaining TPs, a majority is found in common by at least three programs. Only TEMP/TEMP2 find a significant proportion of unique TP (3.5% for *D. melanogaster* and 6.1% for *A. thaliana*).

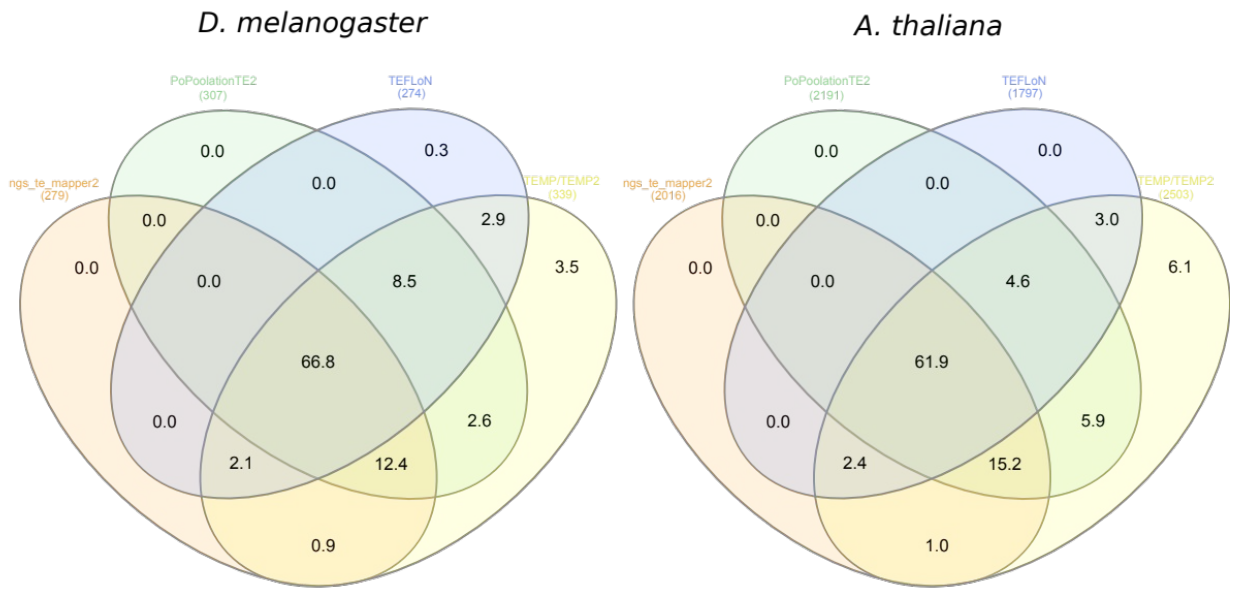


Figure 5: Percentage overlap of TP reference insertions found by the programs having the best *Fscores*.

Detection of new insertions

We then evaluated the capacity of the programs to find new insertions when compared to the reference genome. They correspond to 390 insertions in the simulated *D. melanogaster* chromosome and 3192 insertions in the simulated *A. thaliana* chromosome. All programs find a total number of new insertions inferior to what is expected (Figure 6). The sequence coverage has an impact on the total number of new insertions found for the majority of the programs. In particular, a coverage of 10X seems to be insufficient for most programs. Only ngs-te-mapper2 and TEBreak are not very impacted.

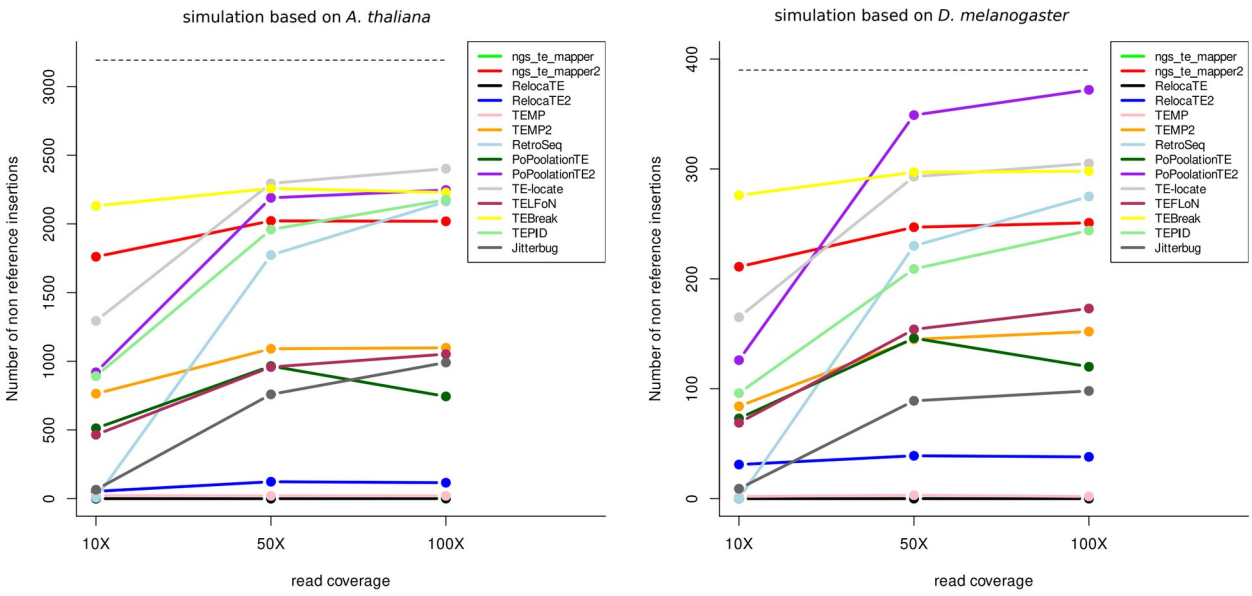


Figure 6: Number of new insertions detected by each program. On the left panel is represented the number of new insertions found for *A. thaliana* and on the right panel for *D. melanogaster*, for the three different read coverages. The dashed lines correspond to the expected number of new insertions in each species.

We then determined among all the new references that are detected which ones are TP according to the same rational presented above and in material and methods. On Figure 7 is represented the different metrics for both species using a correct localization prediction at 20 bp (see supplementary figures S4, S5, and S6 for all cutoffs). Globally, the recall for each program, and for both species, is not very high, meaning that many TPs are missed by the programs. Three of the programs give the best results considering 50X of coverage (TEBreak, PoPoolationTE2 and ngs_te_mapper2). The *precision* metric on the contrary shows that for most of programs, TP are numerous among all the results produced, especially for *D. melanogaster*. The *Fscore* shows similar results between the two species with four programs having the best results: TEBreak, ngs_te_mapper2, PopoolationTE2 (starting at 50X) and RetroSeq (starting at 50X).

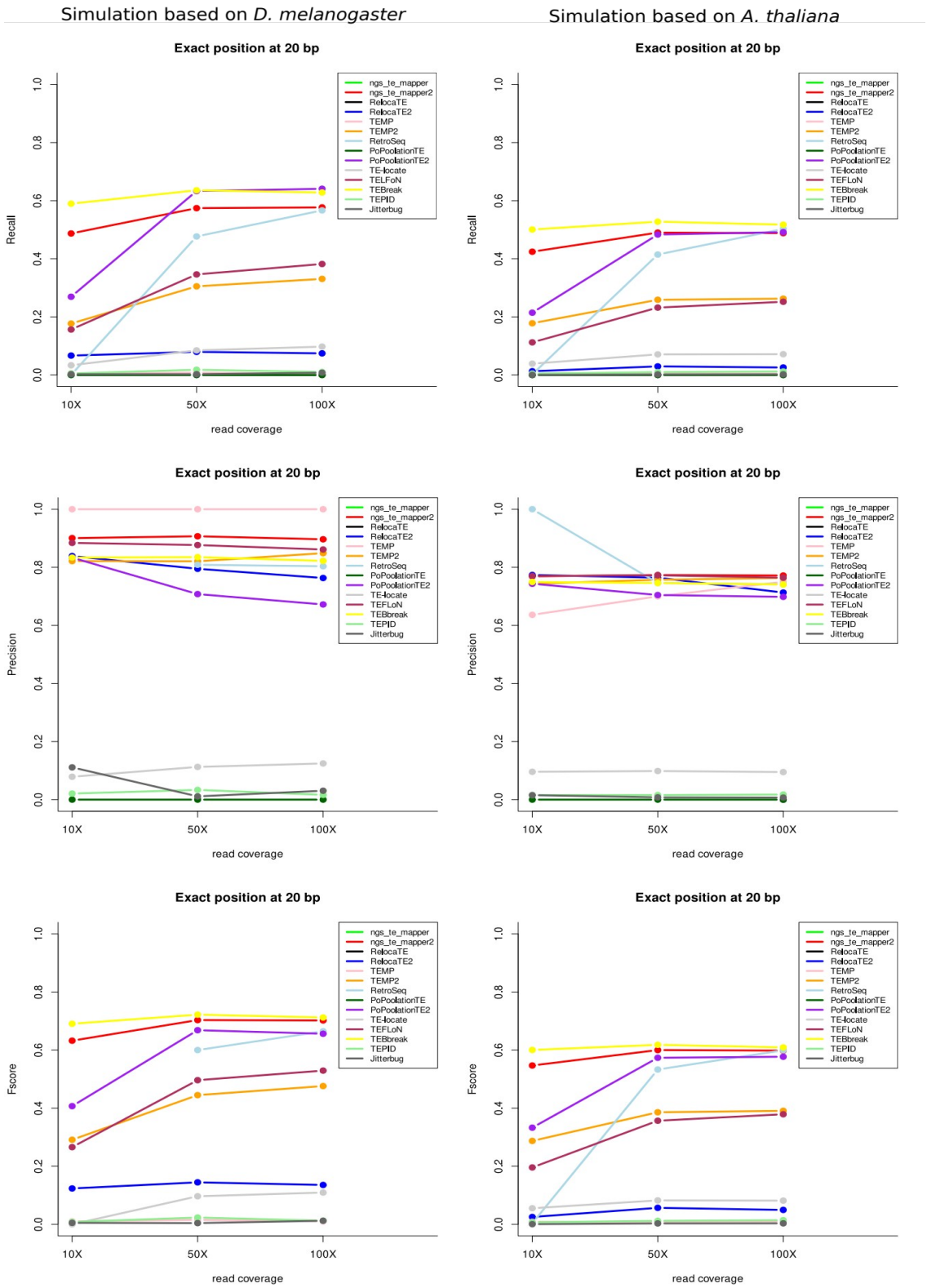


Figure 7: Evaluation metrics for the two species for the three read coverages for the new insertions with a precision localization of 20 bp.

The overlapping of the TP detected by the six best programs accounts for only 10.9% of the TP for *D. melanogaster* and 10.5% of the TP for *A. thaliana* (Figure 8). PopoolationTE2 and ngs_te_mapper2 each identify 6.5% and 4.4% unique TP in *D. melanogaster*, and 2.1% and 2.9% respectively in *A. thaliana*. It should be noted that 11.4% of TP are found by all the programs except TEFLoN in *A. thaliana*. For *D. melanogaster*, 10.6% of TP are found in common for all programs except TEMP/TEMP2.

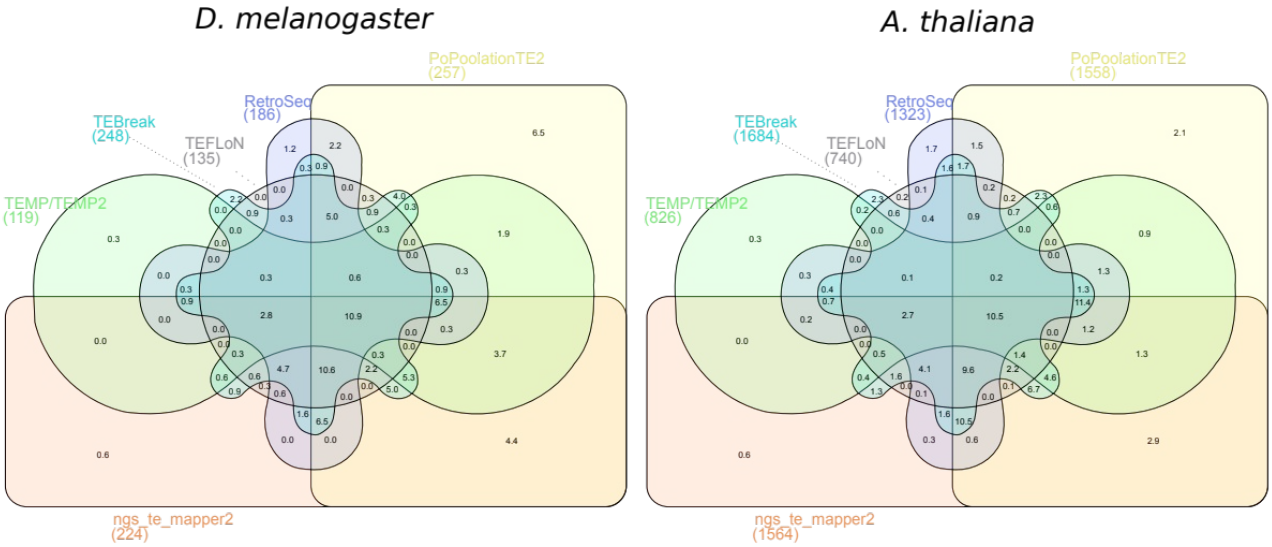


Figure 8: Percentage overlap of TP new insertions found by the programs having the best *Fscores*.

Comparison of the characteristics between True Positives and False Negatives

Since we know with accuracy the characteristics of all insertions present in the two reconstructed chromosomes for both species, it is possible to determine whether some of them may have an impact on the fact that an insertion is detected or not by the programs. We considered the results of the programs having the best *Fscores* for a coverage of 50X and with enough identified TPs considering a localization precision of 20bp to allow statistical analyses without bias.

First, we considered the reference insertions in both species (Table 1 and Table 2, Wilcoxon tests). The results show that the TPs have significantly smaller sizes than FNs for all programs (expected for ngs_te_mapper2 with *D. melanogaster*). Moreover, the distance to the closest TE insertions is also important since it is significantly larger for TPs when compared to FNs, for all programs and for the two species. Additionally, in *A. thaliana*, the %GC of the flanking regions of TPs are significantly more GC rich than those around FNs. To summarize, the programs better detect reference insertions that are small and largely distant from other TE insertions.

Table 1: characteristics of TPs vs FNs for reference insertions for *D. melanogaster* (400 reference insertions in total)

		ngs_te_mapper2 (279 TPs)		**PoPoolationTE2 (307 TPs)**		**TEFLoN (274 TPs)**		**TEMP /TEMP2 (339 TPs)**	
		TP	FN	TP	FN	TP	FN	TP	FN
Mean insertion size (bp)		1,126.5	1,800.4	1,096.5	2,105.7	990.6	2,072	1,086.2	2,702.3
		NS		*0.011116*		*1.696e-05*		*8.091e06*	
%divergence to the reference		9.593	11.092	10.070	9.958	10.054	10.02	10.05	10.013
		0.002068		*NS*		*NS*		*NS*	
Mean distance to the nearest insertions (bp)		31,427	21,329	32,127.3	15,918	30,561.1	23,630	30,332.1	17,416
		1.627e-06		*2.254e-10*		*0.0008189*		*5.21e-05*	
%GC of flanking regions	5’	40.29	40.81	40.39	40.57	40.40	40.54	40.57	40.03
		NS		*NS*		*NS*		*NS*	
	3’	40.57	40.45	40.68	39.82	40.61	39.65	40.81	39.76
		NS		*0.02122*		*NS*		*NS*	
Mean TSD size (bp)		3.62	3.967	3.713	3.761	3.752	3.664	3.735	3.667
		NS		*NS*		*NS*		*NS*	

Table 2: characteristics of TPs vs FNs for reference insertions for *A. thaliana* (3,133 reference insertions in total)

		ngs_te_mapper2 (2,019 TPs)		**PoPoolationTE2 (2,191 TPs)**		**TEFLoN (1,799 TPs)**		**TEMP /TEMP2 (2,506 TPs)**	
		TP	FN	TP	FN	TP	FN	TP	FN
Mean insertion size (bp)		978	1894.7	1043	1917	950.7	1781.8	1013	2465
		<2.2e-16		*<2.2e-16*		*<2.2e-16*		*<2.2e-16*	
%divergence to the reference		9.70	10.33	9.91	9.93	9.81	10.1	9.88	10.1
		1.435e-05		*NS*		*NS*		*NS*	
Mean distance to the nearest insertions (bp)		5,684	2,455.5	5,494	2,296.3	5,525	3,202.5	5,156	2,059
		<2.2e-16		*<2.2e-16*		*<2.2e-16*		*<2.2e-16*	
%GC of flanking regions	5’	34.91	34.15	34.97	33.96	34.93	34.35	34.74	34.58
		0.04939		*0.006899*		*0.005606*		*NS*	
	3’	35.14	33.89	35.11	33.59	35.09	34.28	34.90	33.85
		5.771e-06		*2.559e-06*		*0.002174*		*0.03816*	
Mean TSD size (bp)		4.02	4.17	4.08	4.08	4.24	3.86	4.08	4.05

	NS	*NS*	*0.0003743*	*NS*

In the case of the new insertions (Table 3 and Table 4, Wilcoxon tests), the results show slightly different characteristics. For all species and almost of programs, the percentage of divergence of TP compared to its ancestral sequence is significantly lower than for the FN. Again, the distance of TP to the closest TEs is larger than for the FN, especially for *A. thaliana* but also for *D. melanogaster* for three programs (TEFLoN, RetroSeq, and TEBreak). Also, the size of TSD is significantly larger for TP than for FN for both species and for most of the programs. Finally, in *A. thaliana*, the %GC of the flanking regions of TP are significantly more GC rich than those around FN. To summarize, the programs better detect new insertions that are not too divergent from the consensus TE used to identify them (so likely to be recent insertions), largely distant from other TE insertions and with specific TSD size.

Table 3: characteristics of TPs vs FNs for new insertions for *D. melanogaster* (390 new insertions in total)

		ngs_te_mapper2 (224 TPs)		**PoPoolationTE2 (257 TPs)**		**TEFLoN (135 TPs)**		**TEMP2 (119 TPs)**		**RetroSeq (186 TPs)**		**TEBreak (248 TPs)**	
		TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
Mean insertion size (bp)		1,257.8	1,518	1,439	1,230.5	636.9	1,756.6	1,503	1,308.5	1,061.1	1,649	1,277.4	1,527
		NS		*NS*		*3.875e-08*		*NS*		*0.01601*		*NS*	
%divergence to the reference		9.22	11.12	9.841	10.387	10.12	9.975	7.112	11.311	9.47	10.537	9.740	10.53
		4.227e-07		*NS*		*NS*		*< 2.2e-16*		*0.003103*		*0.02933*	
Mean distance to the nearest insertions (bp)		27,946	26,612	27,967	26,238	30,095	25,937	28,538	26,869	30,054	24,930	30,147	22,513
		NS		*NS*		*0.03794*		*NS*		*0.0136*		*0.001241*	
%GC of flanking regions	5’	40.59	40.50	40.67	40.29	40.70	40.48	40.74	40.47	40.72	40.39	40.71	40.23
		NS		*NS*		*NS*		*NS*		*NS*		*NS*	
	3’	40.15	40.80	40.38	40.50	40.21	40.53	40.19	40.52	40.25	40.58	40.37	40.50
		NS		*NS*		*NS*		*NS*		*NS*		*NS*	
Mean TSD size (bp)		4.045	3.03	3.642	3.561	3.956	3.433	3.899	3.489	3.774	3.468	3.903	3.106
		1.63e-06		*NS*		*0.002307*		*NS*		*0.04023*		*5.961e-05*	

Table 4: characteristics of TPs vs FNs for new insertions for *A. thaliana* (3,192 new insertions in total)

		ngs_te_mapper2 (1,564 TP)		**PoPoolationTE2 (1,558 TP)**		**TEFLoN (740 TP)**		**TEMP2 (826 TP)**		**RetroSeq (1,323 TP)**		**TEBreak (1,684 TP)**	
		TP	FN	TP	FN	TP	FN	TP	FN	TP	FN	TP	FN
Mean insertion size (bp)		1,358.1	1,326.8	1,351.1	1,334	1,187	1,388.9	1,272.4	1,366.5	1,313.0	1,363	1,370.4	1,310.6
		NS		*NS*		*NS*		*NS*		*NS*		*NS*	
%divergence to the reference		9.589	10.43	9.668	10.355	10.101	9.995	7.149	11.022	9.53	10.37	9.765	10.304
		5.138e-10		*2.619e-07*		*NS*		*< 2.2e-16*		*6.297e-10*		*7.778e-05*	
Mean distance to the nearest insertions (bp)		5,498	3,853.5	5,682	3,684.3	5,402	4,435	5,280	4,442	5,345	4,173	5,366	3,870
		< 2.2e-16		*< 2.2e-16*		*8.975e-09*		*3.313e-09*		*< 2.2e-16*		*< 2.2e-16*	
%GC of flanking regions	5’	35.43	34.03	35.44	34.04	35.43	34.61	35.36	34.61	35.61	34.12	35.47	33.84
		3.734e-12		*5.535e-10*		*0.0001875*		*0.004747*		*4.639e-14*		*1.087e-15*	
	3’	35.23	34.24	35.14	34.32	35.18	34.60	35.27	34.54	35.20	34.41	35.19	34.23
		1.352e-07		*6.72e-06*		*0.002374*		*0.0003581*		*1.399e-06*		*1.81e-07*	
Mean TSD size (bp)		4.503	3.568	4.15	3.907	4.789	3.795	4.063	4.013	4.044	4.013	4.404	3.603
		< 2.2e-16		*0.01874*		*< 2.2e-16*		*NS*		*NS*		*7.709e-15*	

Application case: detection of endogenous retroviruses polymorphic insertions in real cattle population data

Although a comprehensive understanding of TEs could have an agricultural interest improving animal breeding, few TE studies have been conducted on livestock species and more particularly on cattle. We decided to use cattle as a mammalian genome example to study a subpart of the TEs, the endogenous retroviruses (ERV) insertions. We propose hereafter a workflow to perform such an analysis.

Find the best configuration on simulated data

The study of simulated *D. melanogaster* and *A. thaliana* chromosomes has shown that the performance of the programs to detect polymorphic TE insertions were different depending on the studied species. In order to choose the best tool to use, the same pipeline as before has been applied to *Bos taurus* to detect polymorphic insertions in short-read data. Two simulated chromosomes were generated using *ReplicaTE* from the chromosome 25. In this chromosome, 899 random CDS sequences were extracted and 900 intergenic regions were generated. The obtained complete

simulated chromosome is 25,638,271 bp long including 936 ERVs whereas the deleted simulated chromosome contains 474 ERVs. The deleted simulated chromosome was used as a reference and the complete simulated chromosome was used to generate simulated short reads in order to evaluate the capacity of the programs to detect the 462 reference insertions and 474 new insertions. We determined among the detected insertions the number of False Positives (FP), False Negatives (FN) and True positives (TP) to compute the same metrics as for *D. melanogaster* and *A. thaliana* (see material and methods section).

On figure 9 is represented the *Fscore* metric for the detection of ERV reference and new insertions using each of the tested program (see supplementary figure S7 for *recall* and *precision* metrics). Similar results are found in cattle compared to the other species but the best programs slightly differed. For the reference insertions, TEMP2 and TEFLoN give the best results with a *Fscore* higher than 0.80. For the new insertions, TEFLoN and TEBreak are the two programs giving the best results with respectively a *Fscore* of 0.82 and 0.66. In conclusion TEFLoN appears to be the most performing tool to use on *B. taurus* data.

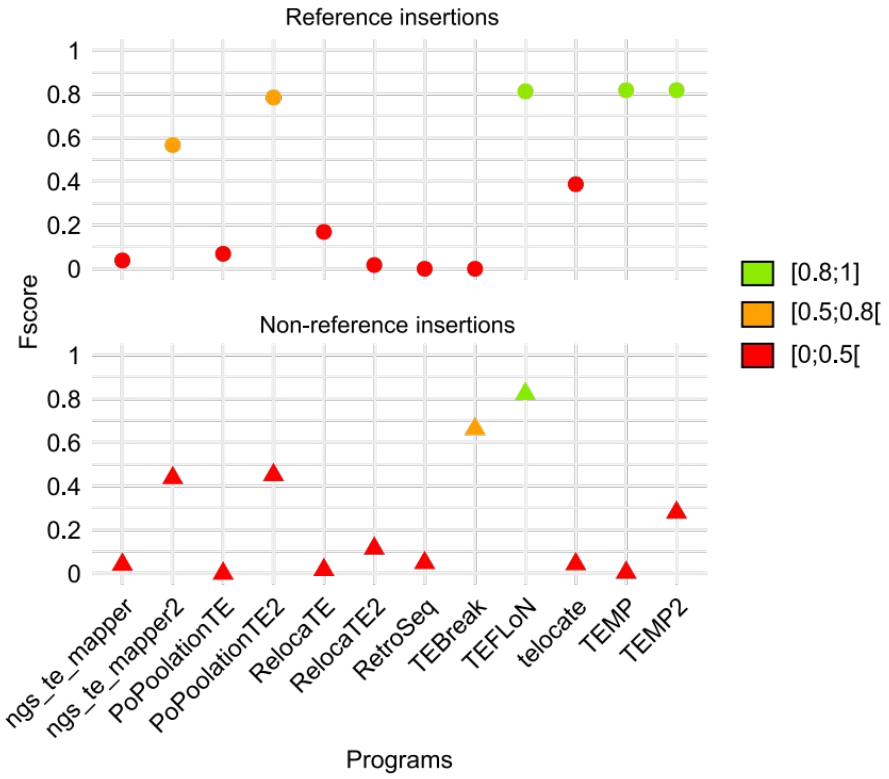


Figure 9: **Performance evaluation of McClintock programs on *Bos taurus* simulated data**. The detection of the reference and new insertions is represented in the upper and lower panels respectively. The performance of the programs was evaluated with the *Fscore* metric.

Repbase consensus sequences are largely used for TE annotation using RepeatMasker. For LTR-retrotransposons, the LTR sequences and the internal part are often split into two separate sequences. Different types of input sequences have been evaluated to detect ERV insertions in *B. taurus* simulated data using TEFLoN (Figure 10A): i) only the LTR sequences, ii) only the internal sequences, iii) the LTR and internal sequences separately, iv) the LTR and internal sequences concatenated for each ERV family sequence, and v) the LTR, the internal and the concatenated family sequences together to test redundancy. On the figure 10B is represented the *Fscore* metric for each input sequences (see supplementary figure S8 for *recall* and *precision* metrics). The use of the internal part alone is not working well contrary to other configurations involving both LTR and internal parts. The use of internal and LTR parts separately gives satisfying results for reference insertions but is less efficient for new insertions detection. The input giving the best results is the one with the concatenated family sequences.

We used four ERV families to generate the simulated chromosomes from ERV class I and II clades. The figure 10C shows how the different ERV families were detected by TEFLoN. Reference ERVs insertions from class I families are better recognized that class II but it is not the case for new insertions. Each family seems to have its own detection characteristics that might correspond to sequence characteristics identified for *D. melanogaster* and *A. thaliana*.

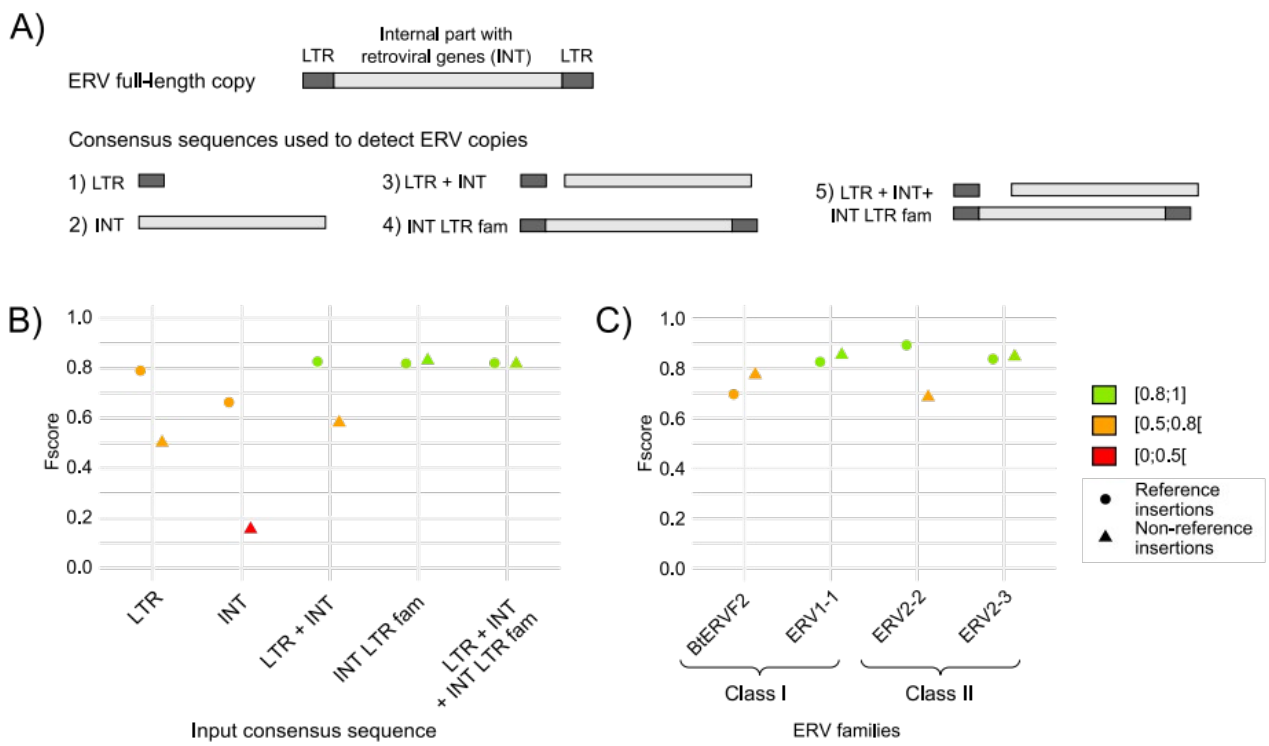


Figure 10: **Impact of ERV input consensus sequences and ERV families on insertion detection with TEFLoN on *Bos taurus* simulated data.** A) ERV copy genomic structure and the different input consensus sequences tested on cattle simulated data to detect ERV copies, B) *Fscores* (Performance) of TEFLoN with the

different input consensus sequences, C) *Fscores* (Performance) of TEFLoN using the detail of each ERV family when using the consensus sequences labeled “INT LTR fam”. The detection of reference insertions is represented with circles and the detection of new insertions with triangles.

Detection of insertion polymorphism in real population data

We used the previously selected tool TEFLoN to analyze 10 samples of cattle short-reads. The detected insertions were compared to the ERV annotation of the reference assembly and to the output of a variant calling analysis performed on long-read data from the same samples. On the figure 11 is represented the *Fscore* obtained for these samples and the correlation between the tool performance and the sample short-read depth sequencing. More than 80% of the expected insertions were detected on average in the 10 samples. ERV insertions also present in the reference genome were significantly better recognized than the new insertions (Wilcoxon test, $p = 1.1e-05$). Among the insertions common with the reference, almost no FP was identified. For new insertions, almost a hundred of FP were detected representing from 30 to 40% of the new insertions detected in each sample. The tool performances were also more homogeneous between the samples for the detection of reference insertions than for the new ones mainly due the short-read coverage differences across samples. A higher coverage improved the detection of insertions but also increased the detection of FPs (see supplementary figure S9). Furthermore, samples a coverage lower than 10X had a drop in detection rates compared to the others. It appears that 10X is the minimum coverage to detect reliably a sufficient number of ERV insertions. Finally, the comparison between the analysis on simulated and real data showed better results in detecting reference insertions in real data compared to simulated data, with median *Fscores* of 0.97 and 0.81 respectively. On the contrary, TEFLoN was less effective in identifying novel insertions in real data compared to simulated data with median *Fscore* of 0.75 and 0.82 respectively (Figure 11).

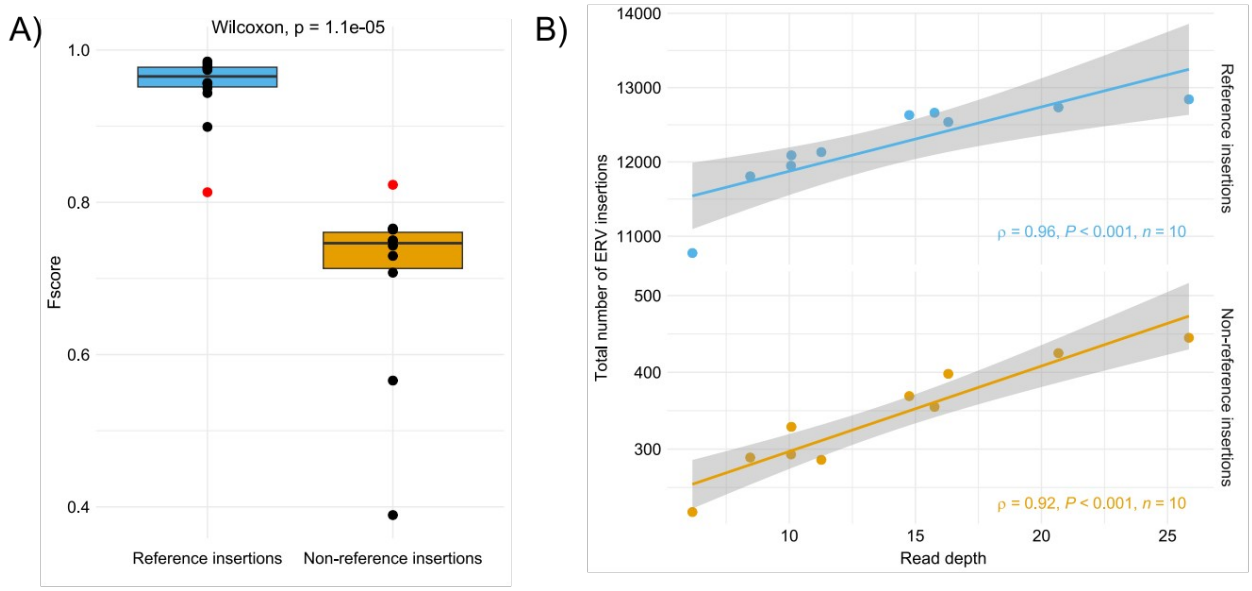


Figure 11: **Detection of ERV insertions in 10 *Bos taurus* samples with TEFLoN.** A) *Fscore* for the detection of ERV insertions in the 10 samples. The red dots indicate the *Fscore* obtained with TEFLoN on cattle simulated data, B) Impact of the short-read depth sequencing on the number of detected insertions. Depth was computed on trimmed reads mapped on the cattle reference genome.

Discussion

In this work, we have developed an approach to simulate TE insertions from a known biological context. The data obtained made it possible to test in a reliable and controlled manner 14 programs for the detection of polymorphic TEs. For the first time, it is possible to show why certain insertions are better detected than others by the different programs.

Generally, full data simulation approaches are often used to test TE polymorphic detection programs. They make it possible to perfectly control all the information. The major problem is that often these simulated data do not completely reflect the biological reality. In order to overcome this problem, we have proposed here an approach that uses real data as a starting point and simulates sequences using the biological information of the organism of interest. We thus made the choice to completely simulate the intergenic regions in order to free ourselves from possible bad TE annotations. However, these intergenic regions are not randomly generated sequences. In particular, the %GC of these sequences must correspond to what is observed in the analyzed genome. The GC content may be an important factor influencing the detection since it may be a caveat in steps of mapping (Donato et al. 2021). Similarly, the reinserted sequences of the TEs are not the true sequences but come from a sequence representative of each family contained in the genome. This allows us to control not only the position of the insertions but also to know with accuracy other information that may play a role on whether an insertion is detected or not. Thus, among the

parameters which are controlled, the size of the insertions, the sequence divergence with respect to the reference element, the distance to the closest TEs and the size of TSD are perfectly known for each insertion. It thus allows us to shed light on precise sequence characteristics rather than limiting tests on specific type of TEs, which is sometimes an approach used to benchmark TE polymorphic tools (Vendrell-Mir et al. 2019). Our approach seems to be a good compromise between the use of complete simulation and real but partial biological data representing a small portion of a chromosome. However, a number of improvements can be considered. Currently, only one chromosome is simulated. It could be interesting to simulate several chromosomes and in particular, to generate populations of chromosomes in order to mimic what can be observed in a natural population. Additionally, the tool is currently limited regarding the number of TE insertions that can be inserted. Thus, for a human chromosome for example, the tool works only with a limited number of TE families.

Our results show that all the programs tested here are far from obtaining results as good as announced in their original publication. For some of them, read coverage strongly impacts the ability to find new insertions, as has already been shown (Rishishwar et al. 2017; Vendrell-Mir et al. 2019) but this is only true up to 50X coverage from which a plateau is reached. Moreover, the results are not as good when one is interested in the reference insertions (present in the reference genome) or the new insertions (present only in the read samples). Globally, the *Fscores* are better in the first case. However, the values obtained for the best programs do not indicate exceptional performance. Indeed, for reference insertions, the best programs ngs-te mapper2 and popoolationTE2 have *Fscores* below 0.8. The other programs (PopoolationTE2, TEFLoN, TEMP, and TEMP2) show values around 0.6. For new insertions, the best programs (TEBreak, ngs_te_mapper2, popoolationTE2 and RetroSeq) have values hovering around 0.6. It is important to note that some programs are more successful in finding reference insertions than new insertions, and *vice versa*. This is the case for TEBreak and RetroSeq which give better results to find new insertions. TEFLoN, TEMP and TEMP2 show poorer performance in finding new insertions compared to reference insertions. Overall ngs_te_mapper2 and PoPoolationTE2 give consistent results for the two types of insertions. If we compare the results for the two species, there are some notable differences for the detection of reference insertions. Ngs-te-mapper2 gives better results with *D. melanogaster* while the best program is PoPoolationTE2 (at 50X and 100X) for *A. thaliana*. In the case of new insertions, all programs give comparable results for the two species, although working a little less well in the case of *A. thaliana*.

Given that the programs produce many false positives (FP), an approach allowing to optimize the identification of the true positives (TP), in the absence of comparison, is to use several

tools at the same time to retain only the insertions detected in common. This approach has been used for the analysis of many natural populations of *D. melanogaster* (Lerat et al. 2019). However, the two tools used showed little overlap in their results. We observed the overlap between the TP for the best programs identified in this work for reference insertions and new insertions. The proportion of common insertions correctly found by all the programs is quite high in the case of reference insertions since it is almost 70% considering four programs. This proportion is much lower in the case of new insertions with less than 11% for six programs. However, the proportion reach 48.1% for *D. melanogaster* and 58,5% for *A. thaliana* when considering only the results common to TEBreak, ngs_te_mapper2 and PoPoolationTE2, the three programs giving the best results in our benchmark. This remains lower that for the reference insertions. Thus, as proposed by Vendrell Mir et al. (Vendrell-Mir et al. 2019), an approach consisting of using several tools at the same time to optimize the number of TPs must be limited to a few tools at a time. Even with this method, it is important to take into account that some information will be inevitably lost and that the number of polymorphic TE insertions will be underestimated.

With our approach, it was possible to compare the characteristics of the True Positives (TP) compared to those of the False Negatives (FN), *i.e.* the insertions which are missed by the programs. The goal was to determine if there are biases inherent in the sequences preventing their detection. Between reference and new insertions, some differences appeared. Especially, reference insertions correctly detected tend to be smaller than those not detected by the programs. That would indicate that degraded or small size type of TEs will be better detected as reference insertions. On the contrary, the new insertions are better detected when their divergence compared to a reference element is low. Then, recent new insertions will be better detected. Although this could be enough to identify recent events, it remains that some of the new insertions may be ancient. These particular insertions would be missed by the different programs. In their original manuscript, almost all programs acknowledge the fact that they cannot detect nested insertions. This is confirmed by our analysis for both types of insertions since TP present significantly larger distances to the nearest TEs than FN. Globally, the same bias appear between the two explored species. However, for *A. thaliana*, we observed that the GC content of genomic regions surrounding the insertions also play a role in whether they are detected or not by the program. This species has globally AT rich intergenic regions (DeRose-Wilson and Gaut 2007). We observed that the insertions are better detected when the genomic regions are less AT rich. Since TEs are known to be also AT rich (Lerat et al. 2002; Boissinot 2022), they may be better identified when their base composition is more different from the surrounding genomic regions. We also observed for the detection of new insertions that the size

of the TSD is important. Since these sequences may not be well conserved, it may prevent the detection of many insertions.

The case study provided here, focusing on *B. taurus*, allowed us to identify important criteria that should be considered before performing studies on polymorphic TEs in real population data. The choice of the program is crucial and depends on the analyzed species. Therefore, it is essential to first perform tests on simulated data build with specific elements from the interest species to identify the most suitable tool(s) to use. The different programs were all used through the McClintock pipeline (Nelson et al. 2017, Chen et al. 2023) which has a significant advantage to allow the use of multiple tools simultaneously, prevents difficulties in program installation and ensures standardized results. It is also important to carefully select the type of consensus sequences, especially for LTR-retrotransposons. For these elements, usually the LTR and the internal parts are separated in distinct consensus sequences. The re-association of the LTR sequences and the internal parts of a given family is thus necessary and require an in-depth annotation of the reference genome.

Here, we demonstrated the importance of testing a tool also on real data before launching a large-scale population analysis. Even though our study was limited to 10 samples, the genomic characteristics and TE content reflected the reality. The results obtained on real data were different compared to the simulated data, with a better detection of the reference insertions but a less effective identification of the new insertions. This difference is mainly due to the total number of ERV insertions. In the simulated data, half of the total insertions were new insertions not present in the reference, whereas they constituted approximately 2% of the insertions in the real data. It appears that detecting new insertions is easier when they represent a larger fraction of the genome of interest.

We showed that new insertions were overall more challenging to detect than the reference ones. Moreover, assessing new insertions absent from the reference genome in real samples is challenging because we do not know what to expect making it difficult to determine whether an insertion is a true or false positive. In our analysis, we used variant calling results obtained from long-reads sequencing data. However, this approach might also miss some insertions, raising questions about its reliability as a reference. Nevertheless, it provides results from two distinct methodologies, ensuring the identification of TPs, even if some are missed.

In conclusion, most of the tested tools do not achieve extraordinary results. There are several biases that prevent them from detecting certain insertions. In addition, the FP rate is particularly high for some tools. Therefore, it is advisable to use a small number of programs simultaneously to optimize the detection of real insertions while keeping a critical perspective on the results.

Acknowledgments

This work was performed using the computing facilities of the CC LBBE/PRABI and of the IFB-cloud. We thank the SeqOccin projet and the Get-Plage platform (https://get.genotoul.fr/) for sharing the bovine data set, and Mekki Boussaha (G2B team, INRAE Jouy-en-Josas) for sharing the associated alignments computed by his team. We thank Caroline Leroux (IVPC) and Vincent Navratil (PRABI-Doua) for useful discussions about this work. JT and TF were supported by INRAE GoatRetrovirome grant for this project. MV PhD fellowship was funded by ANR, grant ANR-22-CE35-0002-01.

References

Adrion JR, Song MJ, Schrider DR, et al (2017) Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. Genome Biology and Evolution 9:1329–1340. https://doi.org/10.1093/gbe/evx050

Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Briefings in Bioinformatics 8:382–392. https://doi.org/10.1093/bib/bbm048

Bogaerts-Márquez M, Barrón MG, Fiston-Lavier A-S, et al (2020) T-lex3: an accurate tool to genotype and estimate population frequencies of transposable elements using the latest short-read whole genome sequencing data. Bioinformatics 36:1191–1197. https://doi.org/10.1093/bioinformatics/btz727

Boissinot S (2022) On the Base Composition of Transposable Elements. Int J Mol Sci 23:4755. https://doi.org/10.3390/ijms23094755

Bourque G, Burns KH, Gehring M, et al (2018) Ten things you should know about transposable elements. Genome Biology 19:199. https://doi.org/10.1186/s13059-018-1577-z

Chen J, Basting PJ, Han S, et al (2023) Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast. Mobile DNA 14:8. https://doi-org/10.1186/s13100-023-00296-4

Chen J, Wrightsman TR, Wessler SR, Stajich JE (2017) RelocaTE2: a high resolution transposable element insertion site mapping tool for population resequencing. PeerJ 5:e2942. https://doi.org/10.7717/peerj.2942

Cordaux R, Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nature reviews Genetics 10:691–703. https://doi.org/10.1038/nrg2640

DeRose-Wilson LJ, Gaut BS (2007) Transcription-related mutations and GC content drive variation in nucleotide substitution rates across the genomes of Arabidopsis thaliana and Arabidopsis lyrata. BMC Evol Biol 7:1–12. https://doi.org/10.1186/1471-2148-7-66

Di Stefano L (2022) All Quiet on the TE Front? The Role of Chromatin in Transposable Element Silencing. Cells 11:2501. https://doi.org/10.3390/cells11162501

Donato L, Scimone C, Rinaldi C, et al (2021) New evaluation methods of read mapping by 17 aligners on simulated and empirical NGS data: an updated comparison of DNA- and RNA-Seq data from Illumina and Ion Torrent technologies. Neural Comput & Applic 33:15669–15692. https://doi.org/10.1007/s00521-021-06188-z

Ewing AD (2015) Transposable element detection from whole genome sequence data. Mobile DNA 6:24. https://doi.org/10.1186/s13100-015-0055-3

Han S, Basting PJ, Dias GB, et al (2021) Transposable element profiles reveal cell line identity and loss of heterozygosity in Drosophila cell culture. Genetics 219:iyab113. https://doi.org/10.1093/genetics/iyab113

Hénaff E, Zapata L, Casacuberta JM, Ossowski S (2015) Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. BMC Genomics 16:768. https://doi.org/10.1186/s12864-015-1975-5

Huang W, Li L, Myers JR, Marth GT (2012) ART: a next-generation sequencing read simulator. Bioinformatics 28:593–594. https://doi.org/10.1093/bioinformatics/btr708

Keane TM, Wong K, Adams DJ (2013) RetroSeq: Transposable element discovery from next-generation sequencing data. Bioinformatics 29:389–390. https://doi.org/10.1093/bioinformatics/bts697

Kobayashi K, Nakahori Y, Miyake M, et al (1998) An ancient retrotransposal insertion causes Fukuyama-type congenital muscular dystrophy. Nature 394:388–392. https://doi.org/10.1038/28653

Kofler R, Betancourt AJ, Schlötterer C (2012) Sequencing of pooled DNA samples (Pool-Seq) uncovers complex dynamics of transposable element insertions in Drosophila melanogaster. PLoS Genetics 8:. https://doi.org/10.1371/journal.pgen.1002487

Kofler R, Gómez-Sánchez D, Schlötterer C (2016) PoPoolationTE2: Comparative Population Genomics of Transposable Elements Using Pool-Seq. Molecular Biology and Evolution 33:2759–2764. https://doi.org/10.1093/molbev/msw137

Lerat E (2019) Repeat in Genomes: How and Why You Should Consider Them in Genome Analyses? In: Ranganathan S, Nakai K, Schönbach C, Gribskov M (eds) Encyclopedia of Bioinformatics and Computational Biology. Elsevier Inc., pp 210–220

Lerat E, Capy P, Biémont C (2002) Codon usage by transposable elements and their host genes in five species. Journal of Molecular Evolution 54:625–637. https://doi.org/10.1007/s00239-001-0059-0

Lerat E, Goubert C, Guirao-Rico S, et al (2019) Population-specific dynamics and selection patterns of transposable element insertions in European natural populations. Mol Ecol 28:1506–1522. https://doi.org/10.1111/mec.14963

Linheiro RS, Bergman CM (2012) Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. PLoS ONE 7:. https://doi.org/10.1371/journal.pone.0030008

Nelson MG, Linheiro RS, Bergman CM (2017) McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3 (Bethesda) 7:2763–2778. https://doi.org/10.1534/g3.117.043893

Platzer A, Nizhynska V, Long Q (2012) TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data. Biology 1:395–410. https://doi.org/10.3390/biology1020395

R Core Team (2017) R: A Language and Environment for Statistical Computing. https://www.r-project.org

Rishishwar L, Mariño-Ramírez L, Jordan IK (2017) Benchmarking computational tools for polymorphic transposable element detection. Brief Bioinform 18:908–918. https://doi.org/10.1093/bib/bbw072

Robb SMC, Lu L, Valencia E, 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, Md) 3:949–57. https://doi.org/10.1534/g3.112.005348

Schauer SN, Carreira PE, Shukla R, et al (2018) L1 retrotransposition is a common feature of mammalian hepatocarcinogenesis. Genome Res 28:639–653. https://doi.org/10.1101/gr.226993.117

Schnable PS, Ware D, Fulton RS, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science (New York, NY) 326:1112–5. https://doi.org/10.1126/science.1178534

Stuart T, Eichten SR, Cahn J, et al (2016) Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife 5:. https://doi.org/10.7554/eLife.20777

Vendrell-Mir P, Barteri F, Merenciano M, et al (2019) A benchmark of transposon insertion detection tools using real data. Mob DNA 10:53. https://doi.org/10.1186/s13100-019-0197-9

Wang Y, McNeil P, Abdulazeez R, et al (2023) Variation in mutation, recombination, and transposition rates in Drosophila melanogaster and Drosophila simulans. Genome Res gr.277383.122. https://doi.org/10.1101/gr.277383.122

Weinstock GM, Robinson GE, Gibbs RA, et al (2006) Insights into social insects from the genome of the honeybee Apis mellifera. Nature 443:931–949. https://doi.org/10.1038/nature05260

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nature reviews Genetics 8:973–982. https://doi.org/10.1038/nrg2165-c4

Yu T, Huang X, Dou S, et al (2021) A benchmark and an algorithm for detecting germline transposon insertions and measuring de novo transposon insertion frequencies. Nucleic Acids Res 49:e44. https://doi.org/10.1093/nar/gkab010

Zhao Y, Li X, Xie J, et al (2022) Transposable Elements: Distribution, Polymorphism, and Climate Adaptation in Populus. Front Plant Sci 13:814718. https://doi.org/10.3389/fpls.2022.814718

Zhuang J, Wang J, Theurkauf W, Weng Z (2014) TEMP: A computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Research 42:6826–6838. https://doi.org/10.1093/nar/gku323

Supplementary data

Supplementary files: Output files produced by *ReplicaTE* on the two species *D. melanogaster* and *A. thaliana*.

Figure S1: Recall metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the reference insertions according to the different precision localization tested.

Figure S2: Precision metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the reference insertions according to the different precision localization tested.

Figure S3: *Fscore* metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the reference insertions according to the different precision localization tested.

Figure S4: Recall metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the new insertions according to the different precision localization tested.

Figure S5: Precision metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the new insertions according to the different precision localization tested.

Figure S6: *Fscore* metrics for *D. melanogaster* and *A. thaliana* for the three read coverages for the new insertions according to the different precision localization tested.

Figure S7: Recall and precision metrics of the different tested programs on *Bos taurus* simulated data

Figure S8: Impact of the structure of the ERV input consensus sequences (panel A) and ERV families (panel B) on recall and precision metrics using TEFLoN on *Bos taurus* simulated data The detection of reference insertions is represented with circles and the detection of new insertions with triangles.

Figure S9: Performance of TEFLoN and impact of read coverage in the detection of ERV insertions in 10 *Bos taurus* samples. A) Recall and precision metrics, B) Number of TP and FP according to the short-read depth sequencing. Depth was computed on trimmed reads mapped on the cattle reference genome.

Figure 1: Workflow of the simulation tool *ReplicaTE*. Each step corresponds to the different scripts. In gray are indicated the genes, in blue and red are indicated the TE insertions, a given color corresponding to a given TE family. The simulated TE copies are represented in orange and light blue.

Figure 2: Evaluation approach. The black rectangles correspond to genes. The red and blue circles correspond to TE copies from two different families. The reads simulated on the "deleted" simulated genome will be mapped to the "complete" simulated genome, which will allow to identify reference and/or absent TE insertions. The reads simulated on the "complete" simulated genome will be mapped to the "deleted" simulated genome, which will allow to identify new TE insertions.

Figure 3: Number of reference insertions detected by each program. On the left panel is represented the number of reference insertions found for *A. thaliana* and on the right panel for *D. melanogaster,* for the three different read coverages. The dashed lines correspond to the expected number of reference insertions in each species.

Figure 4: Evaluation metrics for the two species for the three read coverages for the reference insertions with a precision localization of 20 bp.

Figure 5: Percentage overlap of TP reference insertions found by the programs having the best Fscores.

Figure 6: Number of new insertions detected by each program. On the left panel is represented the number of new insertions found for *A. thaliana* and on the right panel for *D. melanogaster,* for the three different read coverages. The dashed lines correspond to the expected number of new insertions in each species.

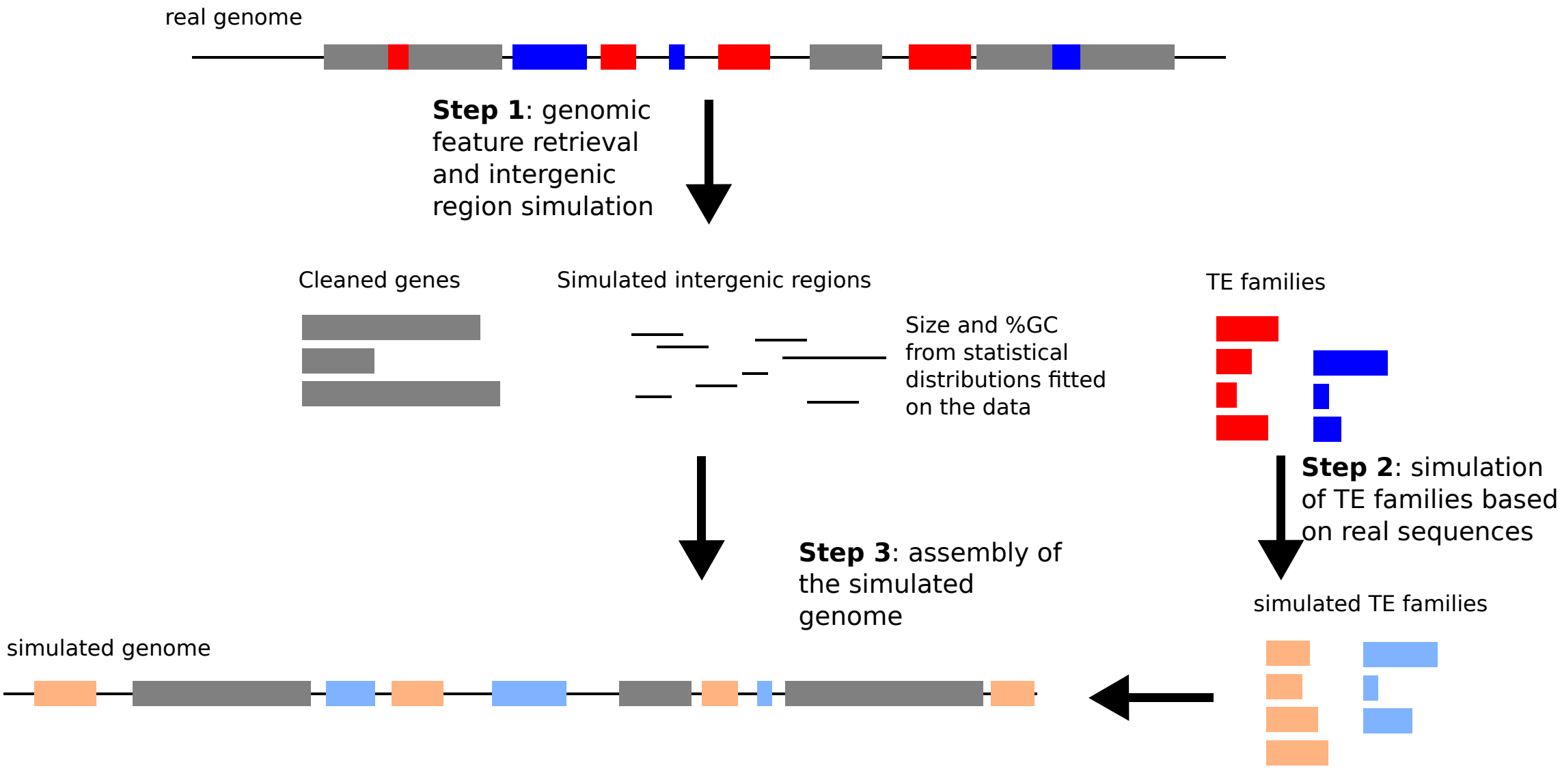
Figure 7: Evaluation metrics for the two species for the three read coverages for the new insertions with a precision localization of 20 bp.

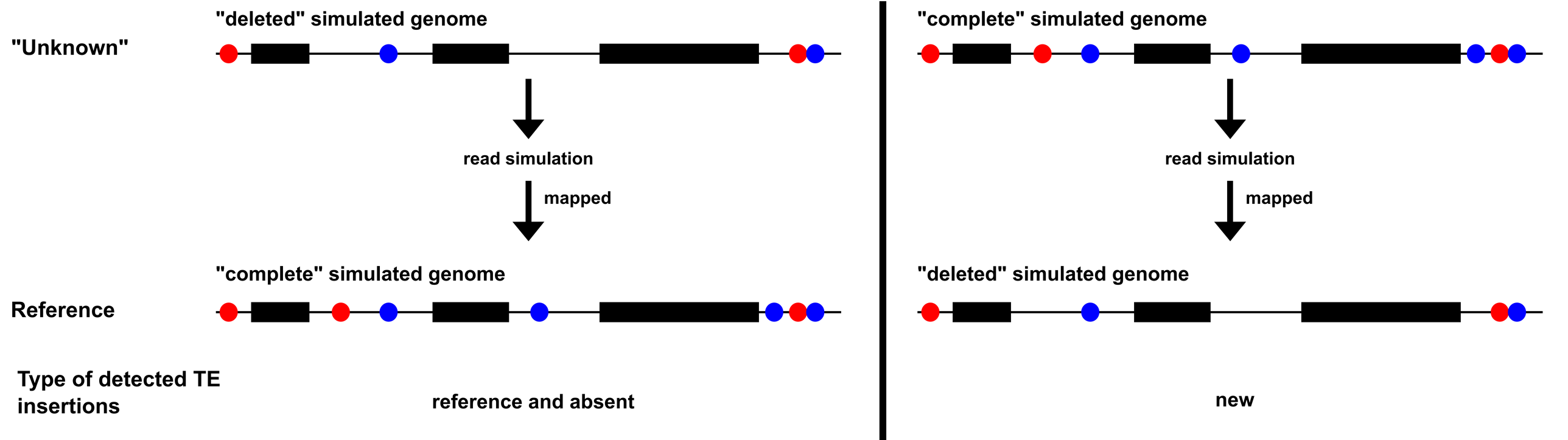
Figure 8: Percentage overlap of TP new insertions found by the programs having the best *Fscores*.

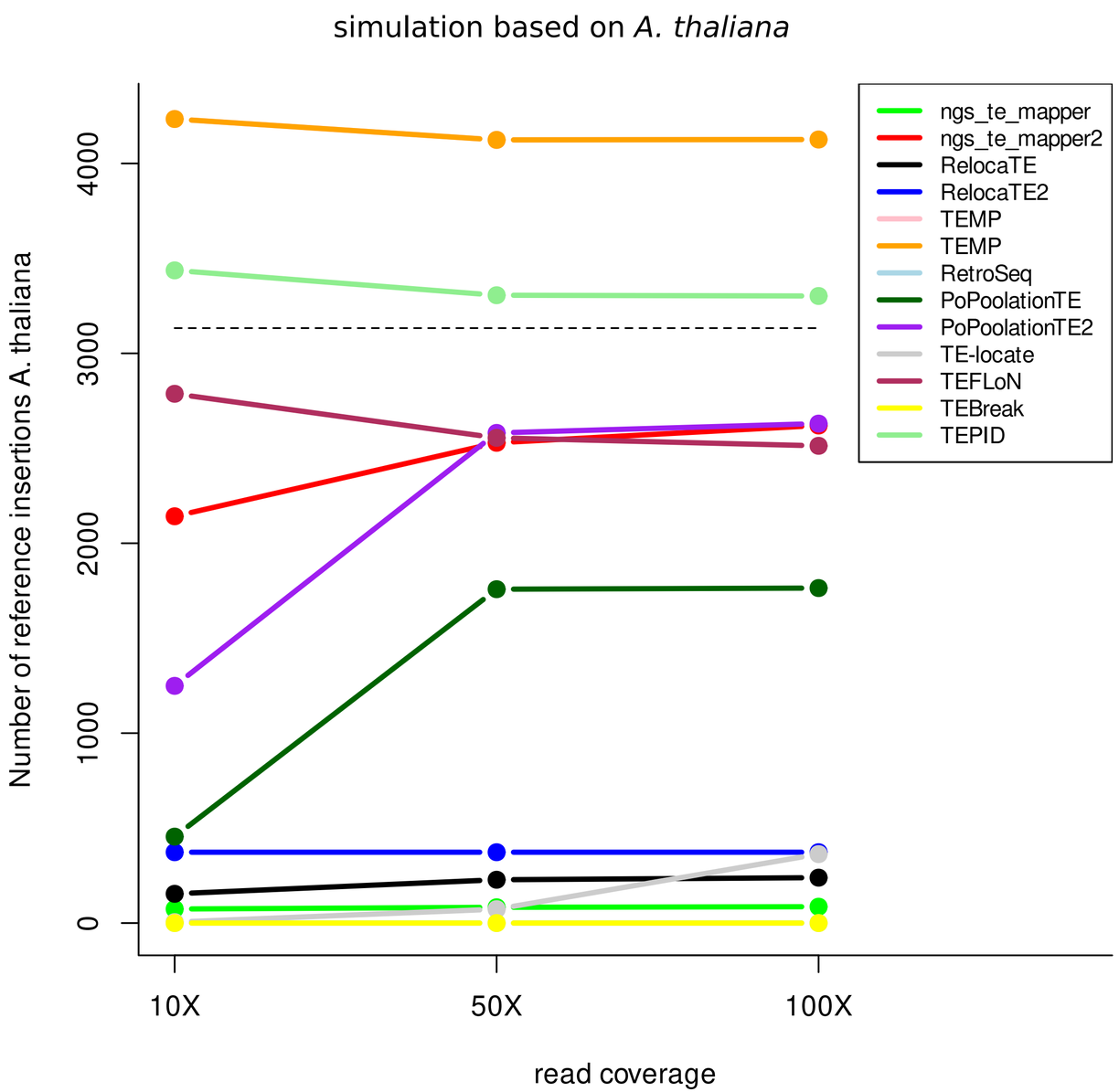
Figure 9: Performance evaluation of McClintock programs on Bos taurus simulated data. The detection of the reference and new insertions is represented in the upper and lower panels respectively. The performance of the programs was evaluated with the Fscore metric.

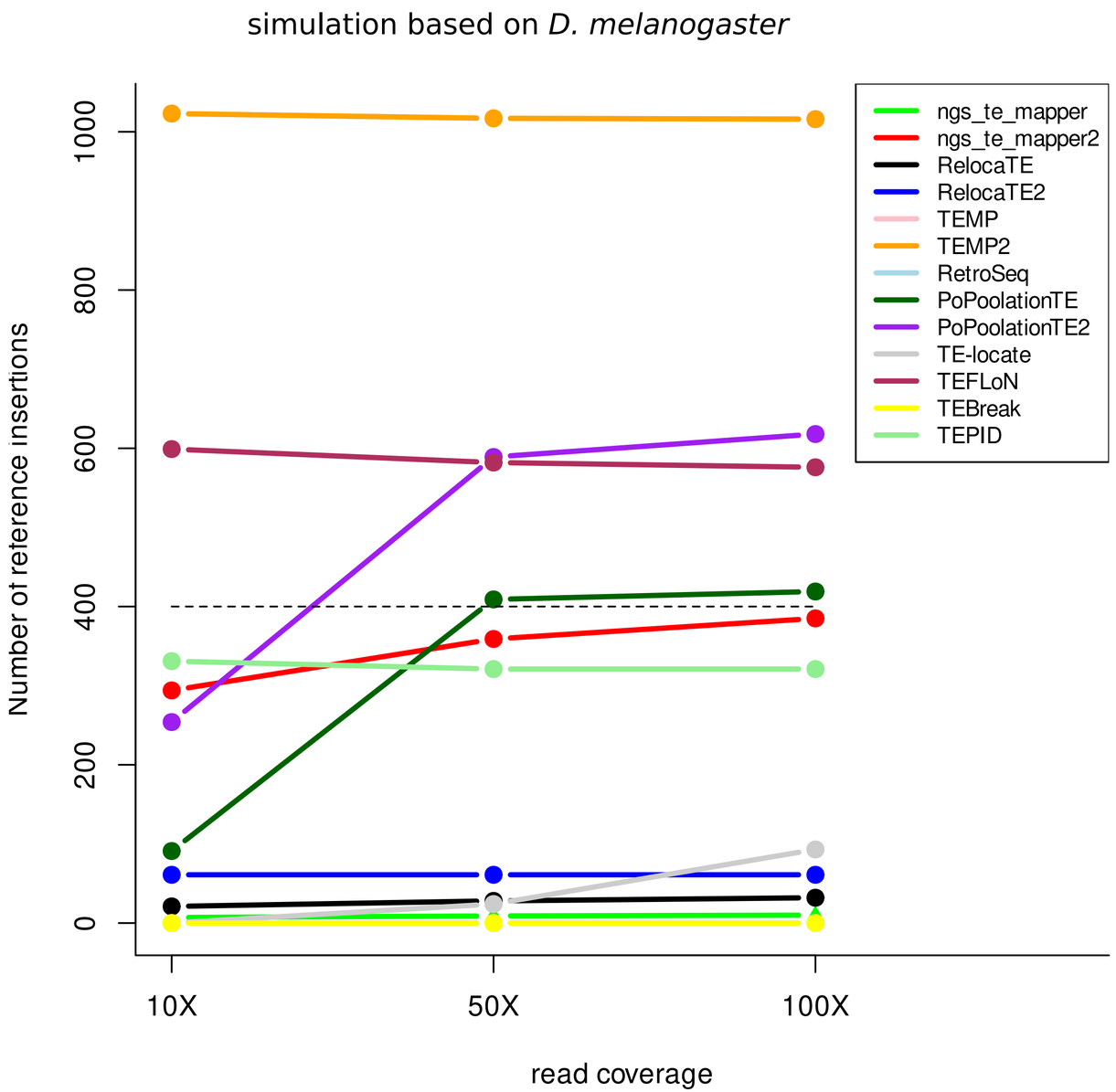
Figure 10: Impact of ERV input consensus sequences and ERV families on insertion detection with TEFLoN on *Bos taurus* simulated data. A) ERV copy genomic structure and the different input consensus sequences tested on cattle simulated data to detect ERV copies, B) *Fscores* (Performance) of TEFLoN with the different input consensus sequences, C) *Fscores* (Performance) of TEFLoN using the detail of each ERV family when using the consensus sequences labeled "INT LTR fam". The detection of reference insertions is represented with circles and the detection of new insertions with triangles.

Figure 11: Detection of ERV insertions in 10 Bos taurus samples with TEFLoN. A) Fscore for the detection of ERV insertions in the 10 samples. The red dots indicate the *Fscore* obtained with TEFLoN on cattle simulated data, B) Impact of the short-read depth sequencing on the number of detected insertions. Depth was computed on trimmed reads mapped on the cattle reference genome.


real genome
Step 1: genomic feature retrieval and intergenic region simulation
Cleaned genes
Simulated intergenic regions
Size and %GC from statistical distributions fitted on the data
TE families
Step 2: simulation of TE families based on real sequences
simulated TE families
Step 3: assembly of the simulated genome
simulated genome

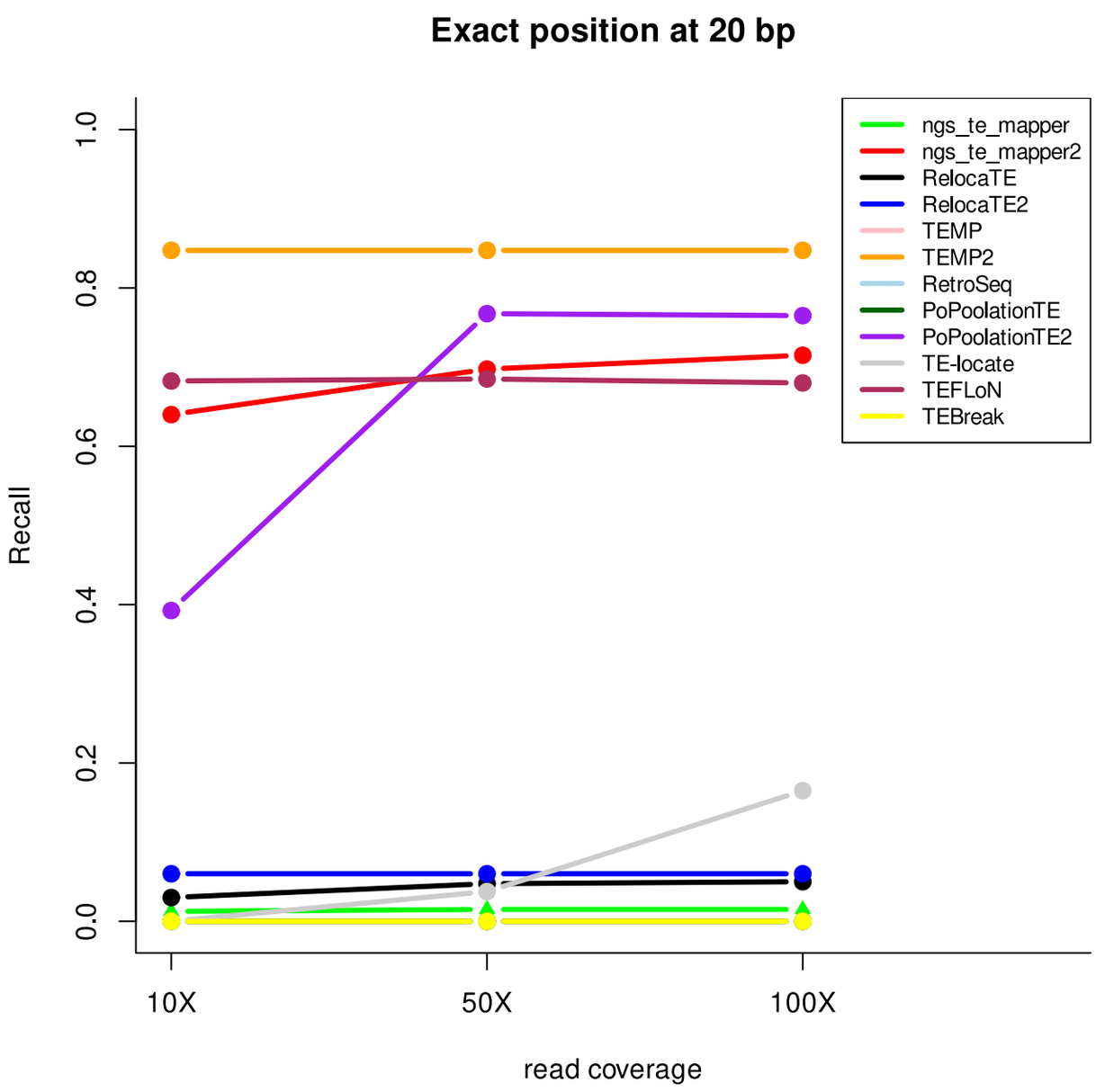

"Unknown"
"deleted" simulated genome
read simulation
mapped
"complete" simulated genome
Reference
Type of detected TE insertions
reference and absent
"complete" simulated genome
read simulation
mapped
"deleted" simulated genome
new

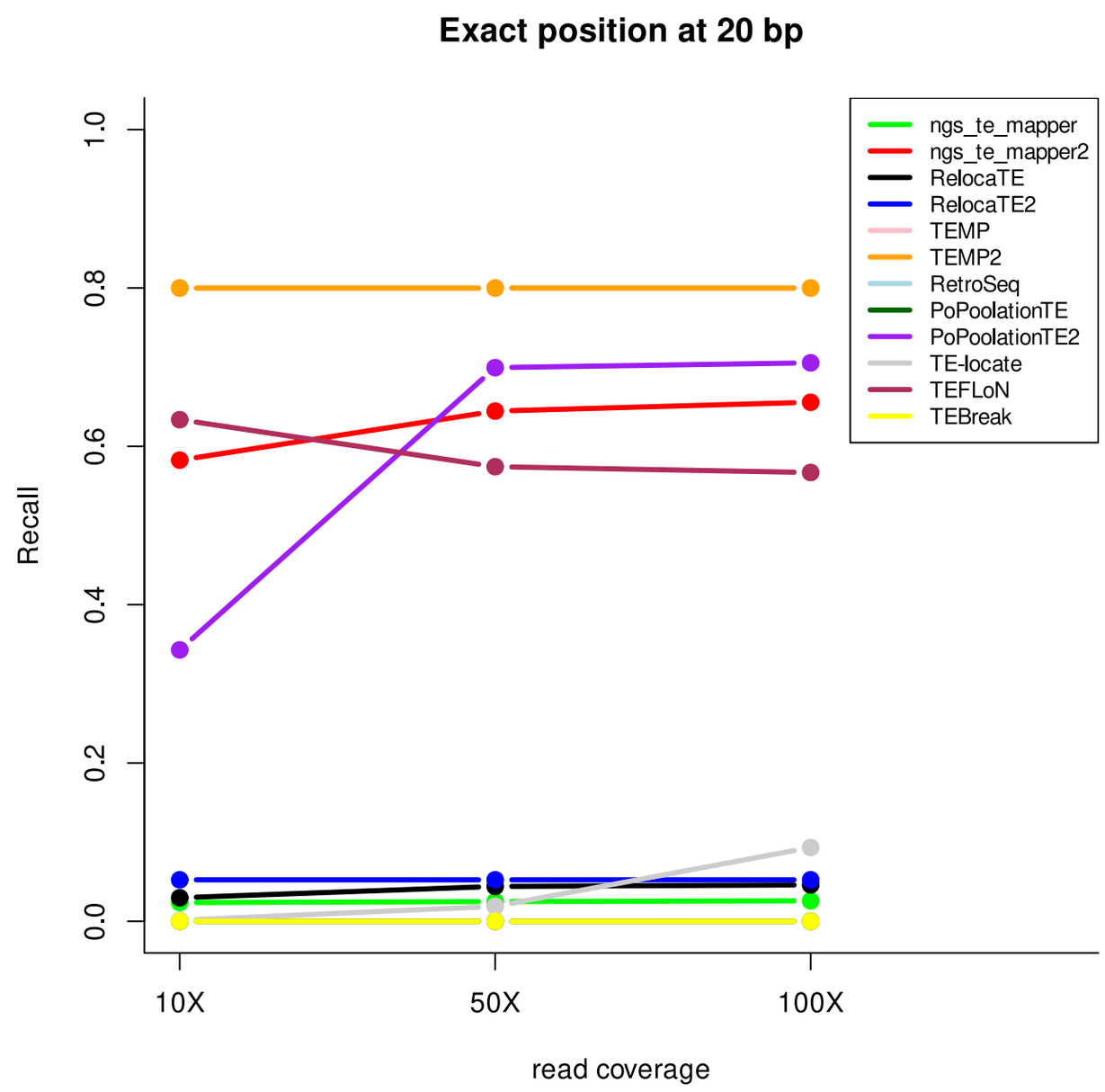

simulation based on *A. thaliana*
Number of reference insertions A. thaliana
0
1000
2000
3000
4000
10X
50X
100X
read coverage
ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBreak
TEPID

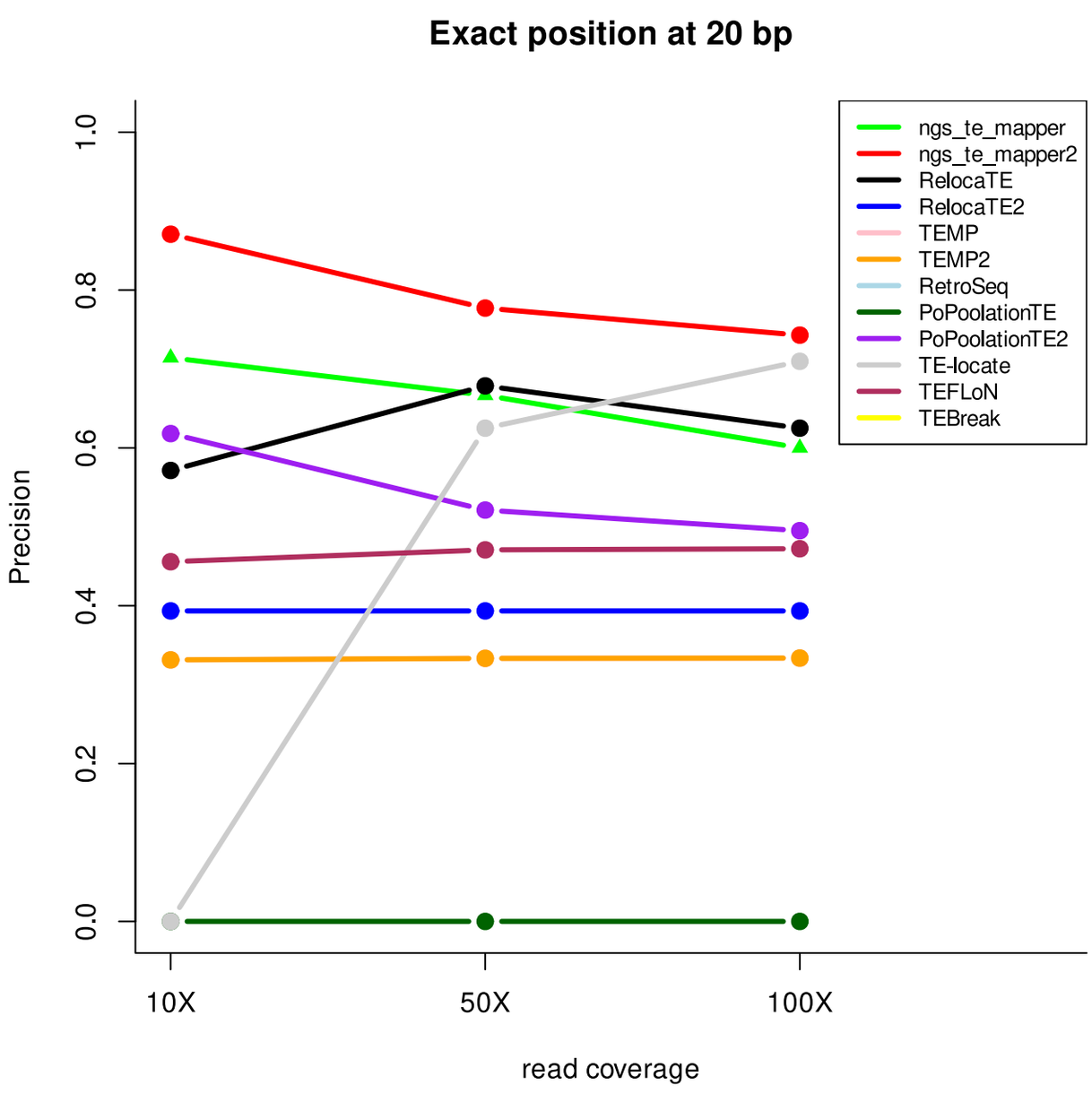

simulation based on *D. melanogaster*
Number of reference insertions
0
200
400
600
800
1000
10X
50X
100X
read coverage
ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBreak
TEPID

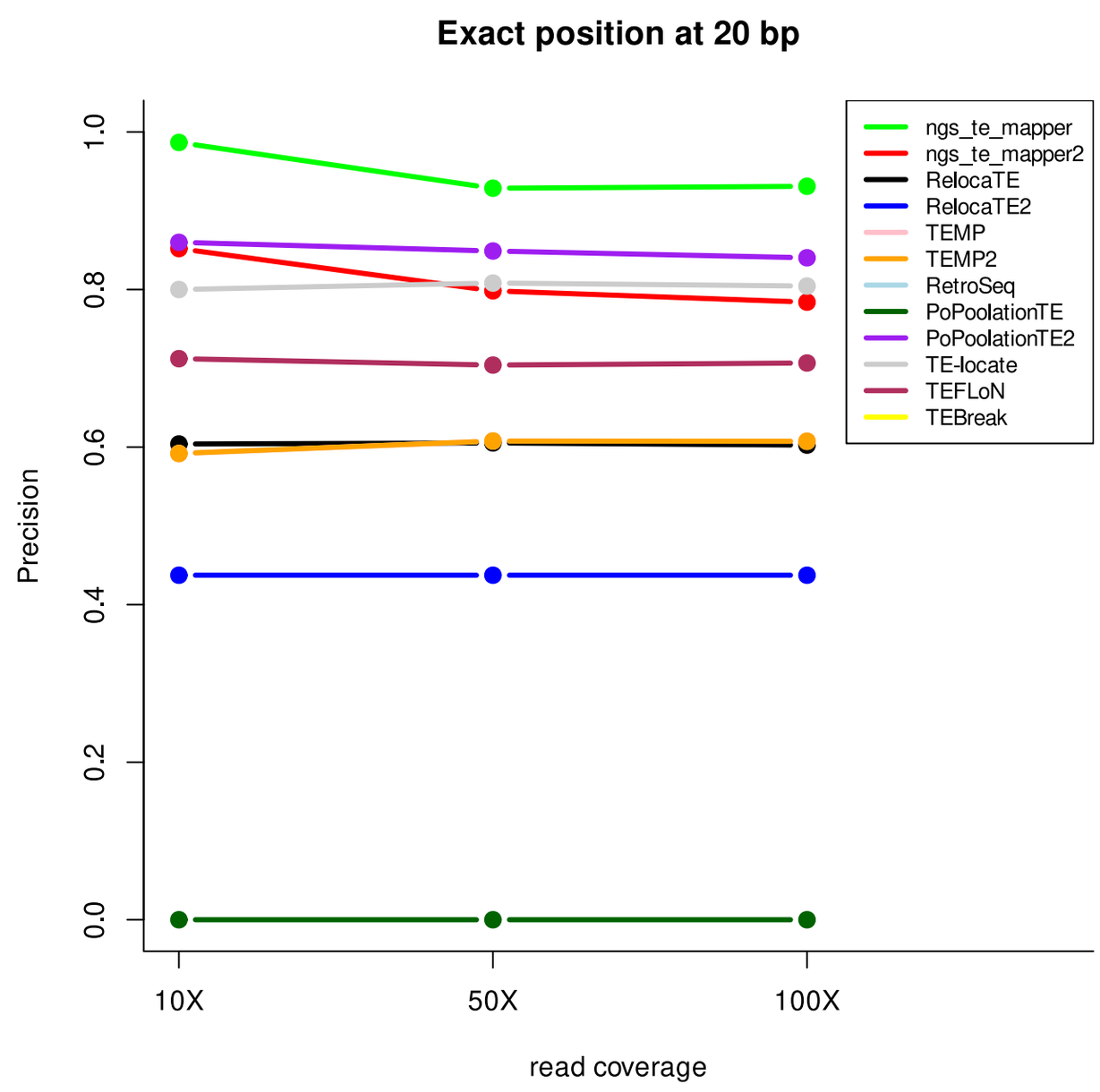
Simulation based on *D. melanogaster*

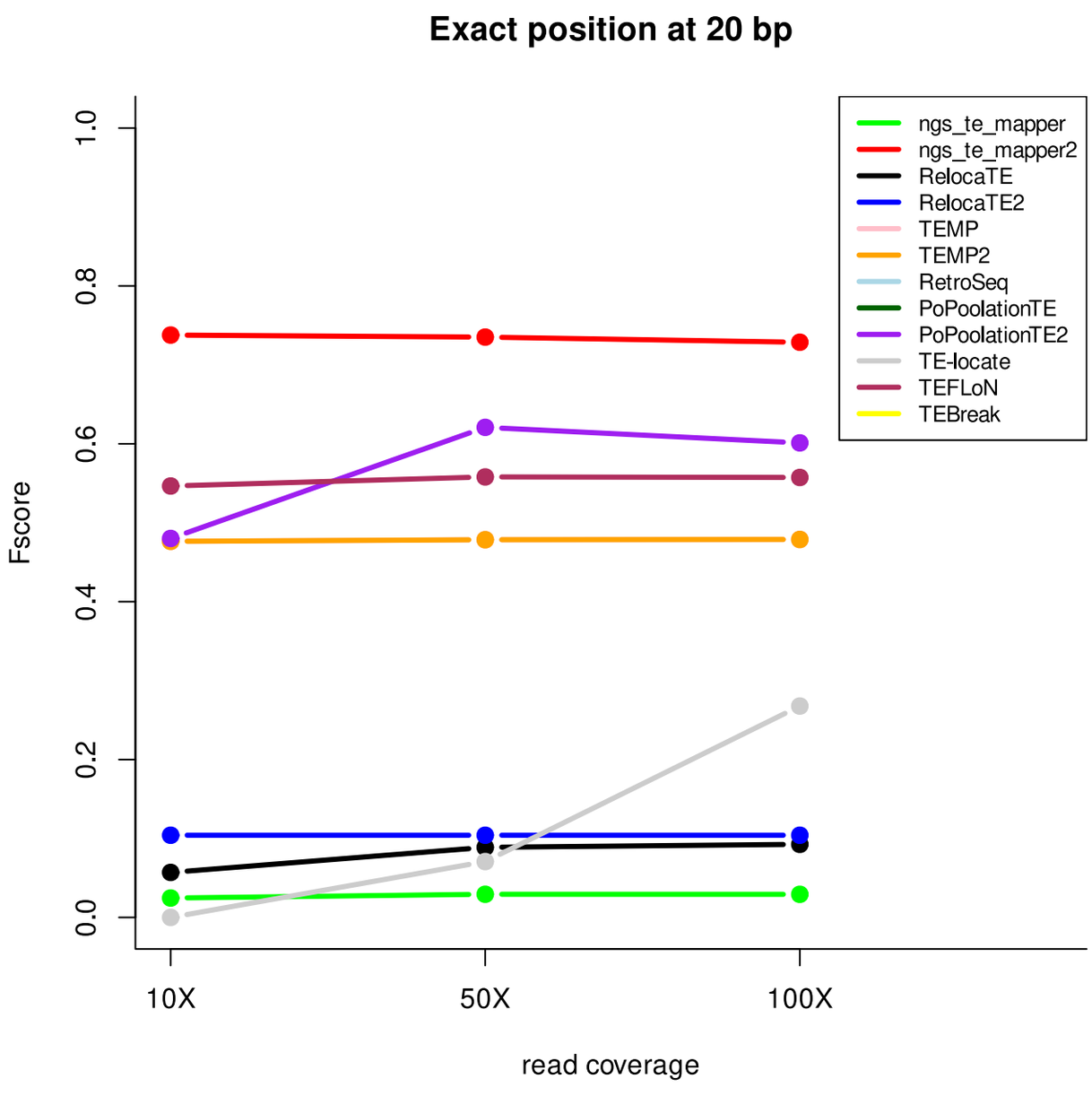
Simulation based on *A. thaliana*

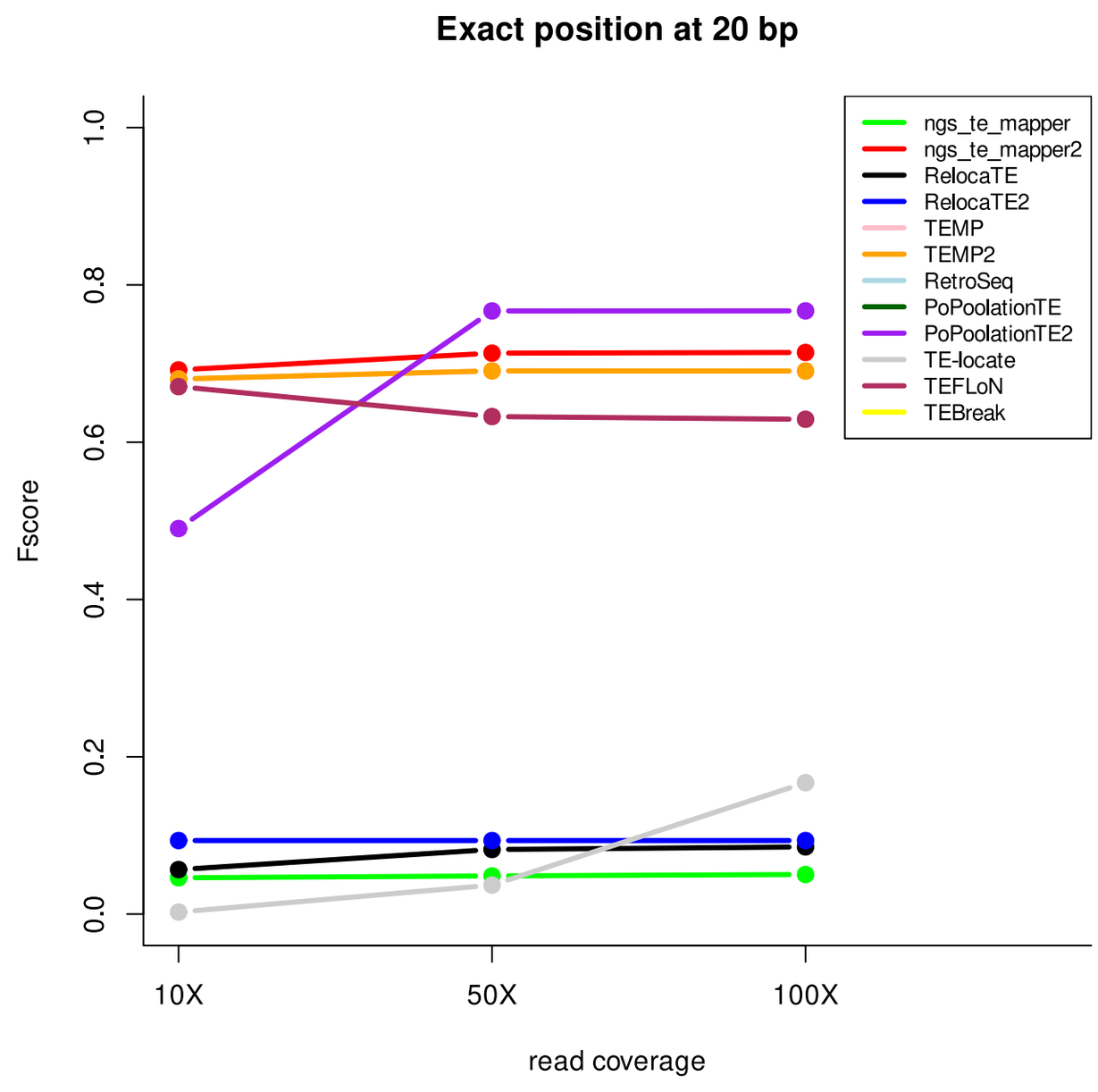


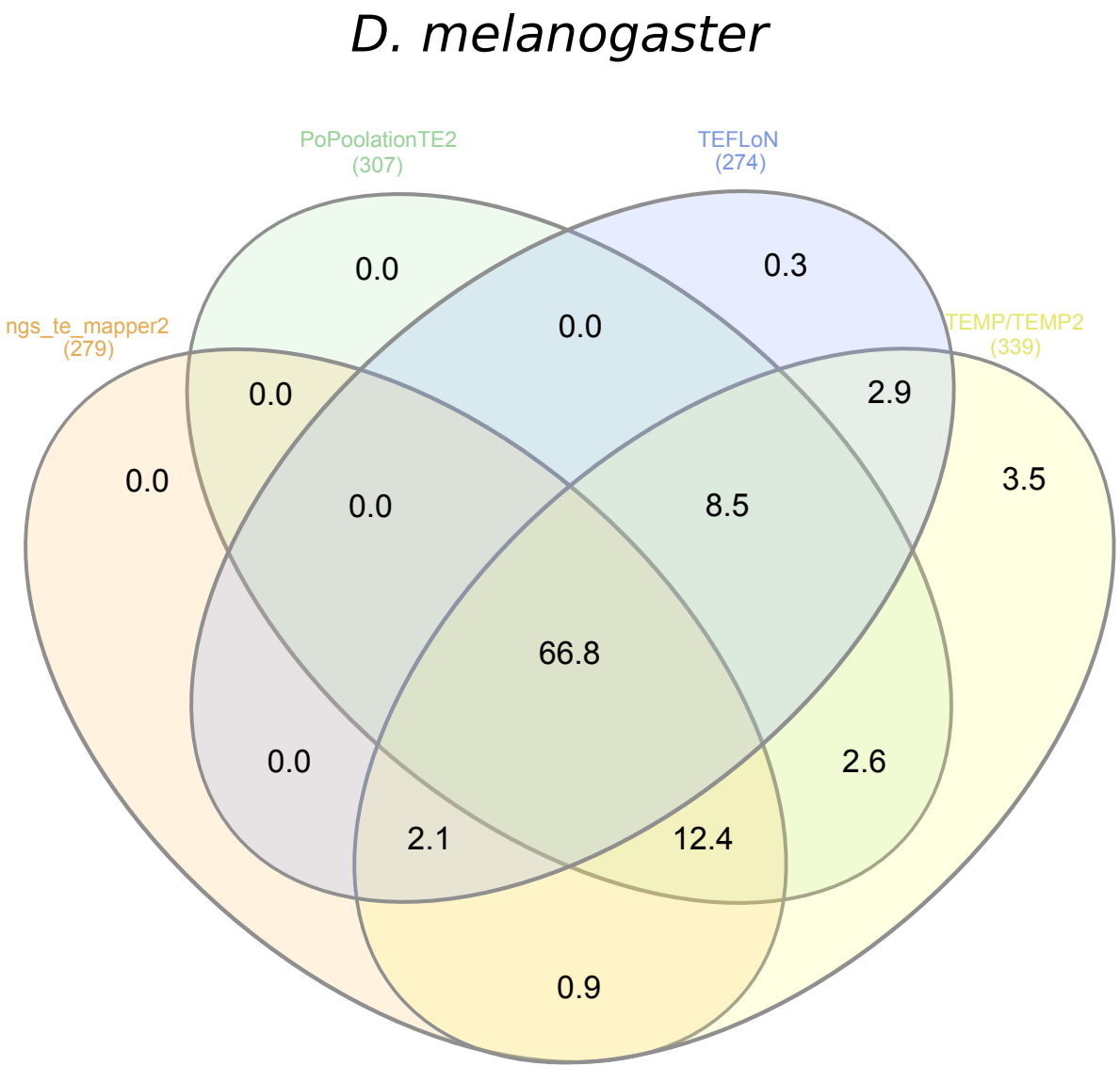


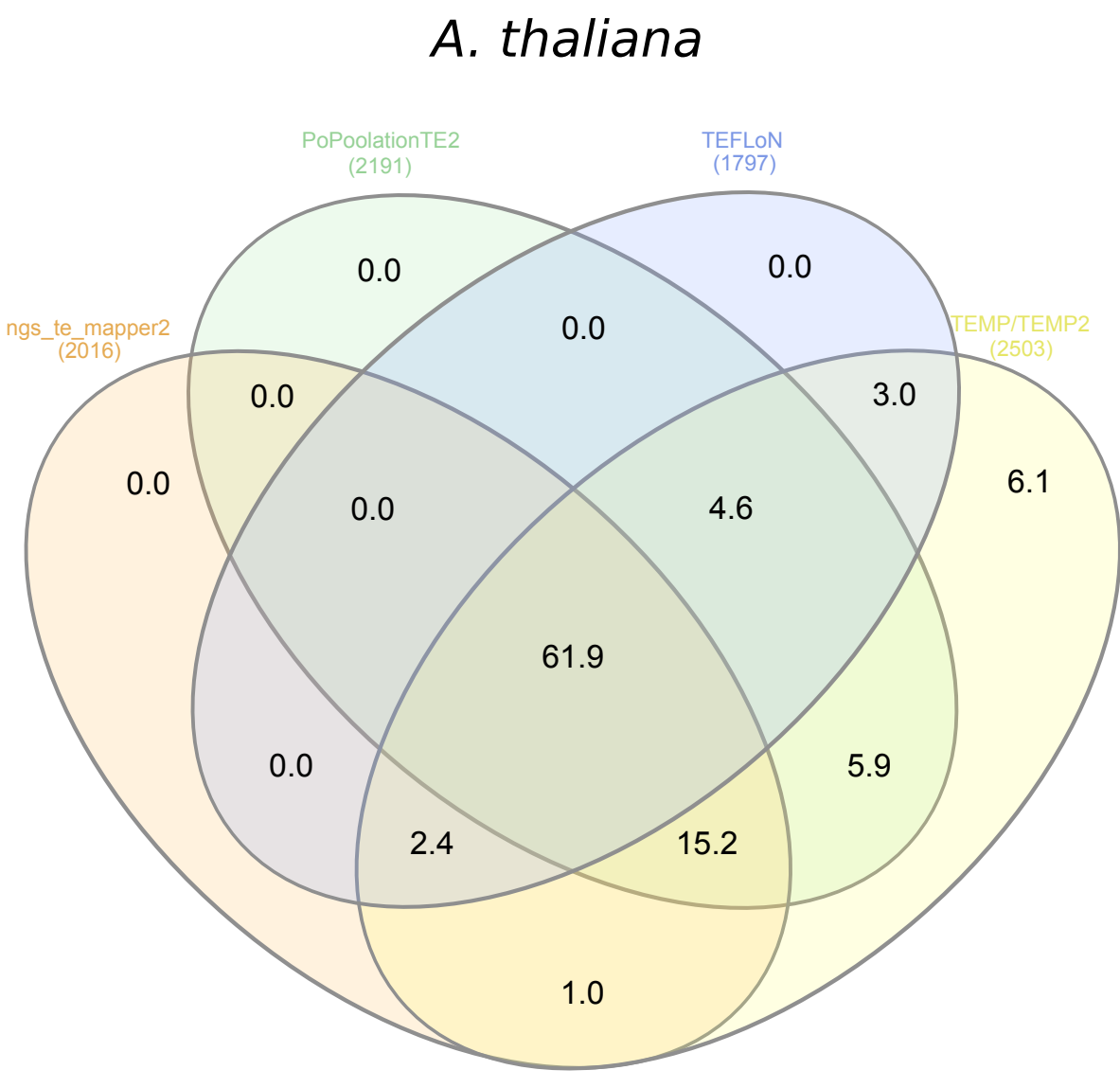


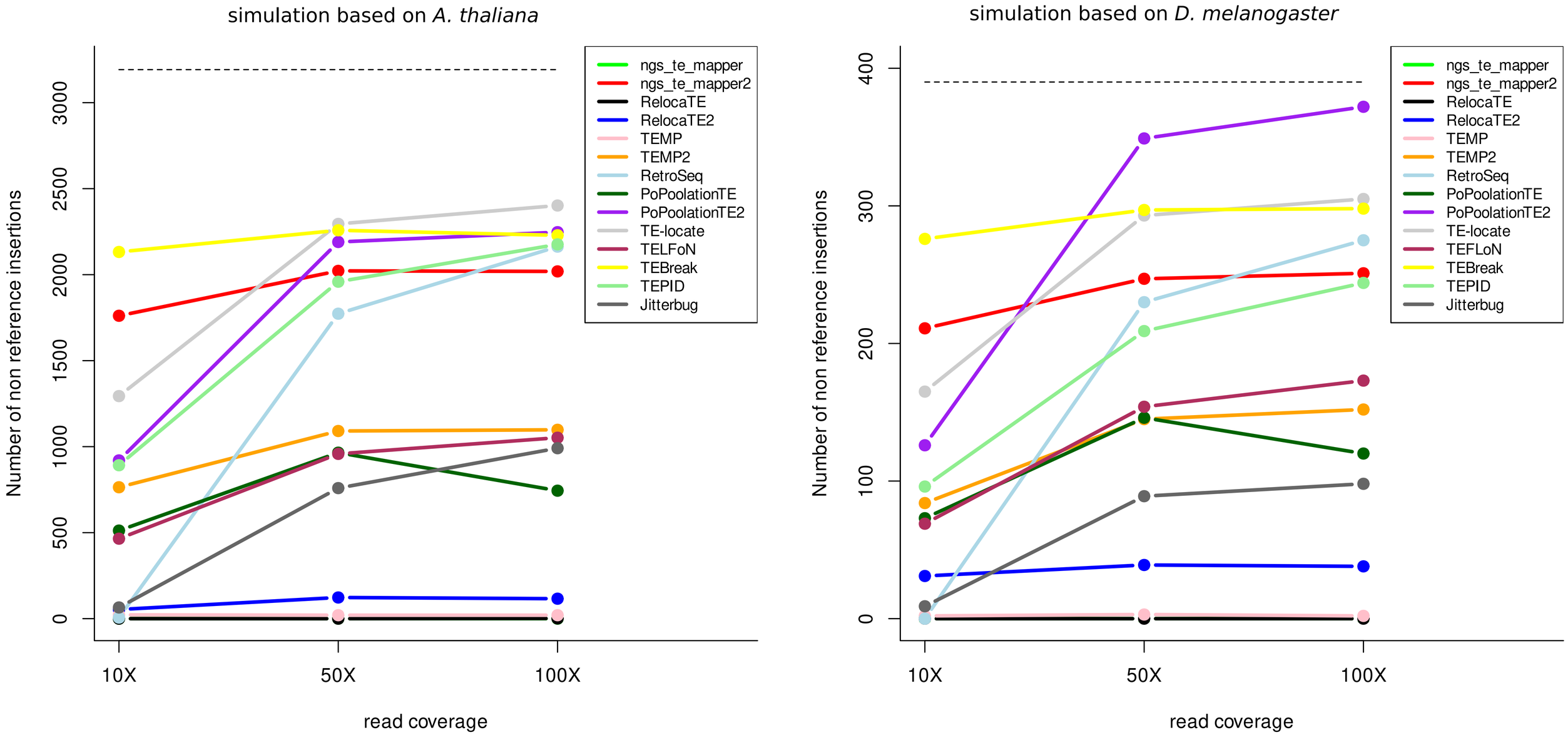







D. melanogaster
PoPoolationTE2
(307)
TEFLoN
(274)
ngs_te_mapper2
(279)
TEMP/TEMP2
(339)
0.0
0.3
0.0
0.0
2.9
0.0
0.0
8.5
3.5
66.8
0.0
2.6
2.1
12.4
0.9


A. thaliana
PoPoolationTE2
(2191)
TEFLoN
(1797)
ngs_te_mapper2
(2016)
TEMP/TEMP2
(2503)
0.0
0.0
0.0
0.0
3.0
0.0
0.0
4.6
6.1
61.9
0.0
5.9
2.4
15.2
1.0


simulation based on A. thaliana
Number of non reference insertions
0
500
1000
1500
2000
2500
3000
10X
50X
100X
read coverage
ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TELFoN
TEBreak
TEPID
Jitterbug
simulation based on D. melanogaster
Number of non reference insertions
0
100
200
300
400
10X
50X
100X
read coverage
ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBreak
TEPID
Jitterbug

Simulation based on *D. melanogaster*

Simulation based on *A. thaliana*

Exact position at 20 bp

Recall

read coverage

10X 50X 100X

0.0 0.2 0.4 0.6 0.8 1.0

ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TELFoN
TEBreak
TEPID
Jitterbug

Exact position at 20 bp

Recall

read coverage

10X 50X 100X

0.0 0.2 0.4 0.6 0.8 1.0

ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBbreak
TEPID
Jitterbug

Exact position at 20 bp

Precision

read coverage

10X 50X 100X

0.0 0.2 0.4 0.6 0.8 1.0

ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBbreak
TEPID
Jitterbug

Exact position at 20 bp

Precision

read coverage

10X 50X 100X

0.0 0.2 0.4 0.6 0.8 1.0

ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBbreak
TEPID
Jitterbug

Exact position at 20 bp

Fscore

read coverage

10X 50X 100X

0.0 0.2 0.4 0.6 0.8 1.0

ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBbreak
TEPID
Jitterbug

Exact position at 20 bp

Fscore

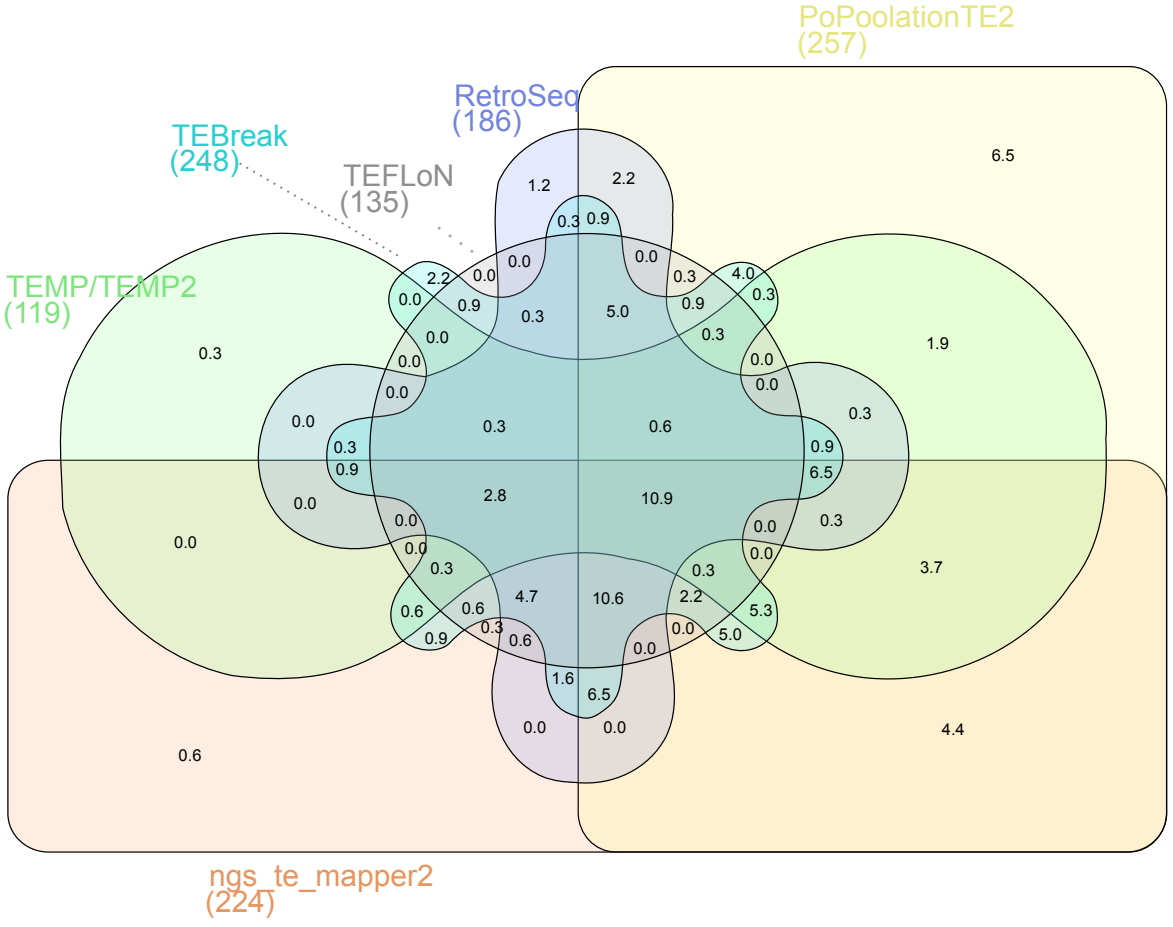
read coverage

10X 50X 100X

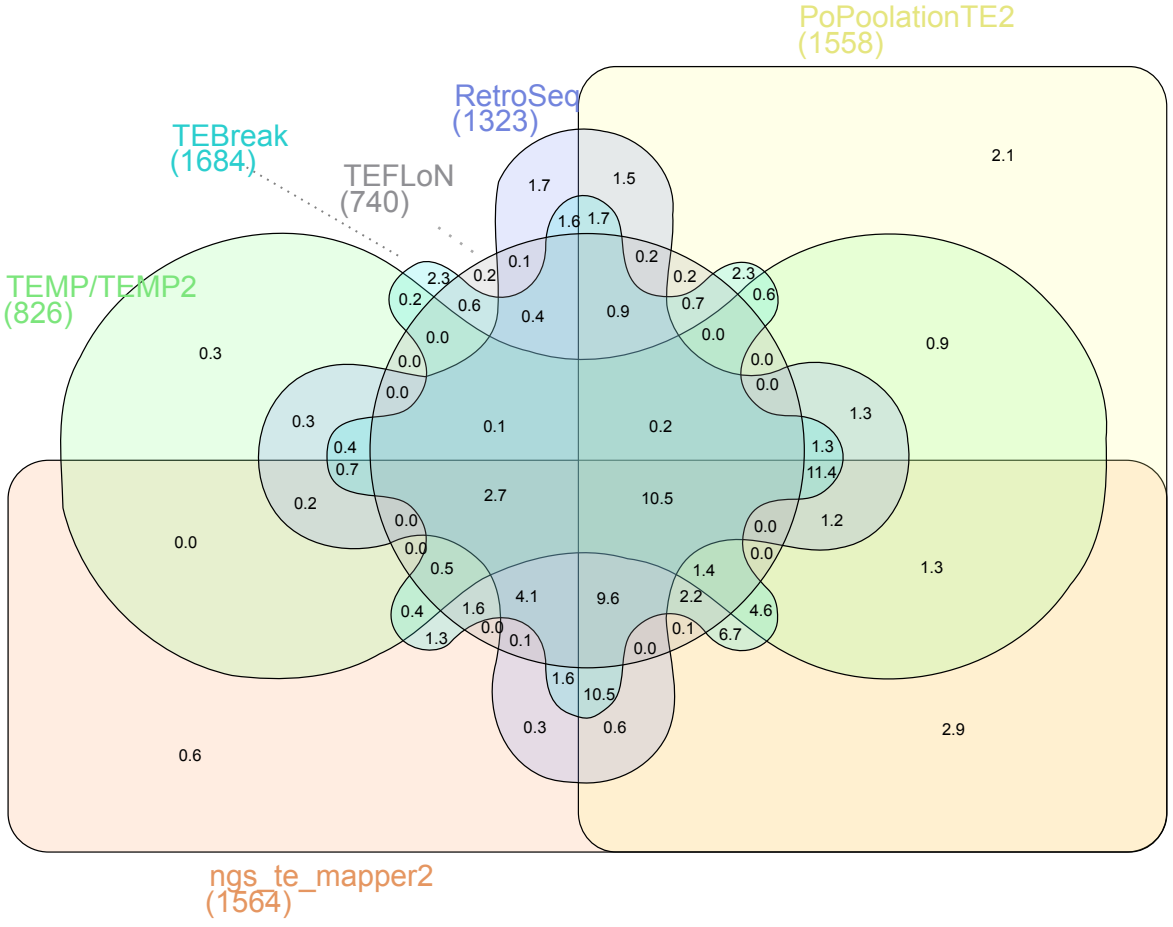
0.0 0.2 0.4 0.6 0.8 1.0

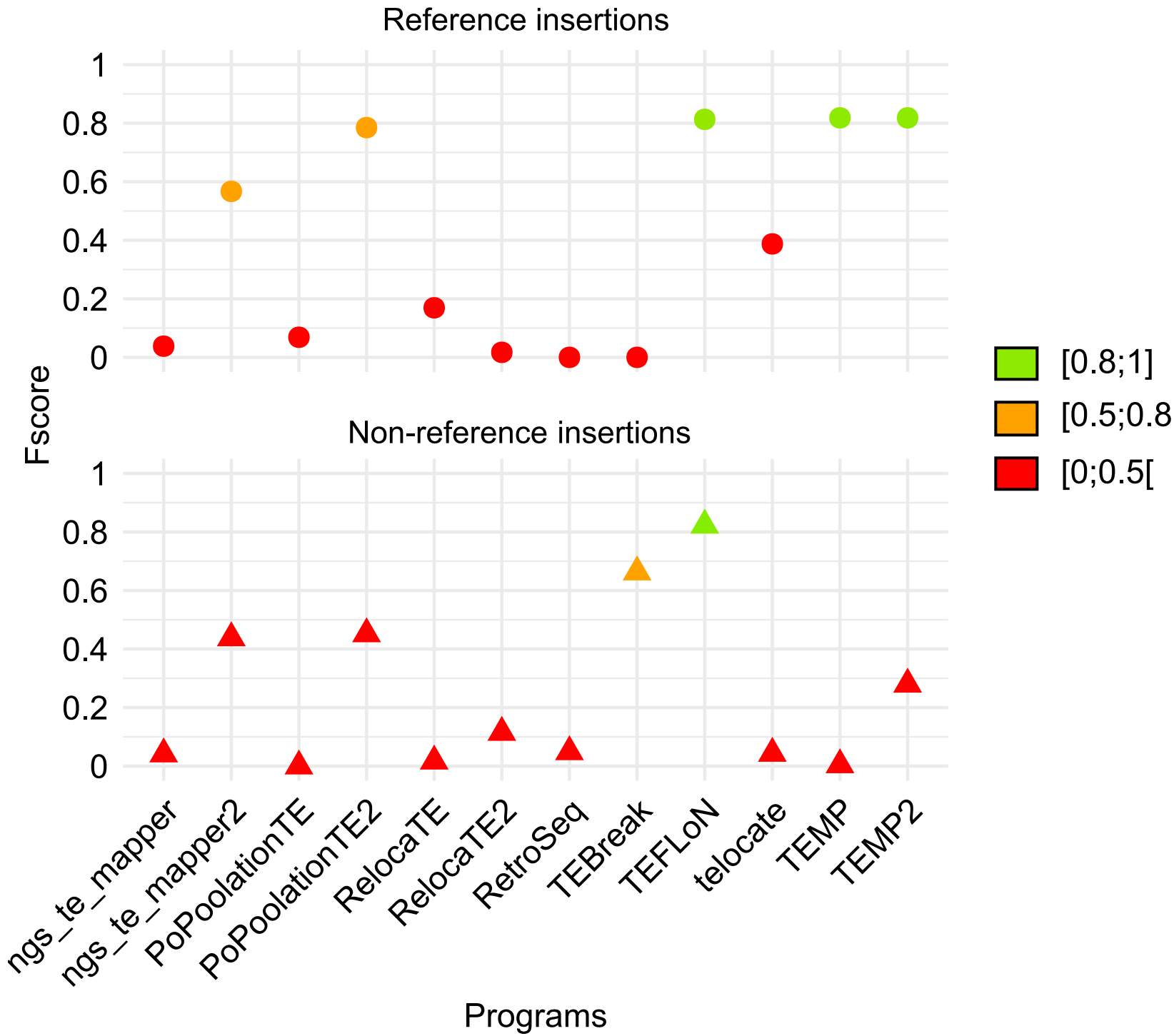
ngs_te_mapper
ngs_te_mapper2
RelocaTE
RelocaTE2
TEMP
TEMP2
RetroSeq
PoPoolationTE
PoPoolationTE2
TE-locate
TEFLoN
TEBbreak
TEPID
Jitterbug

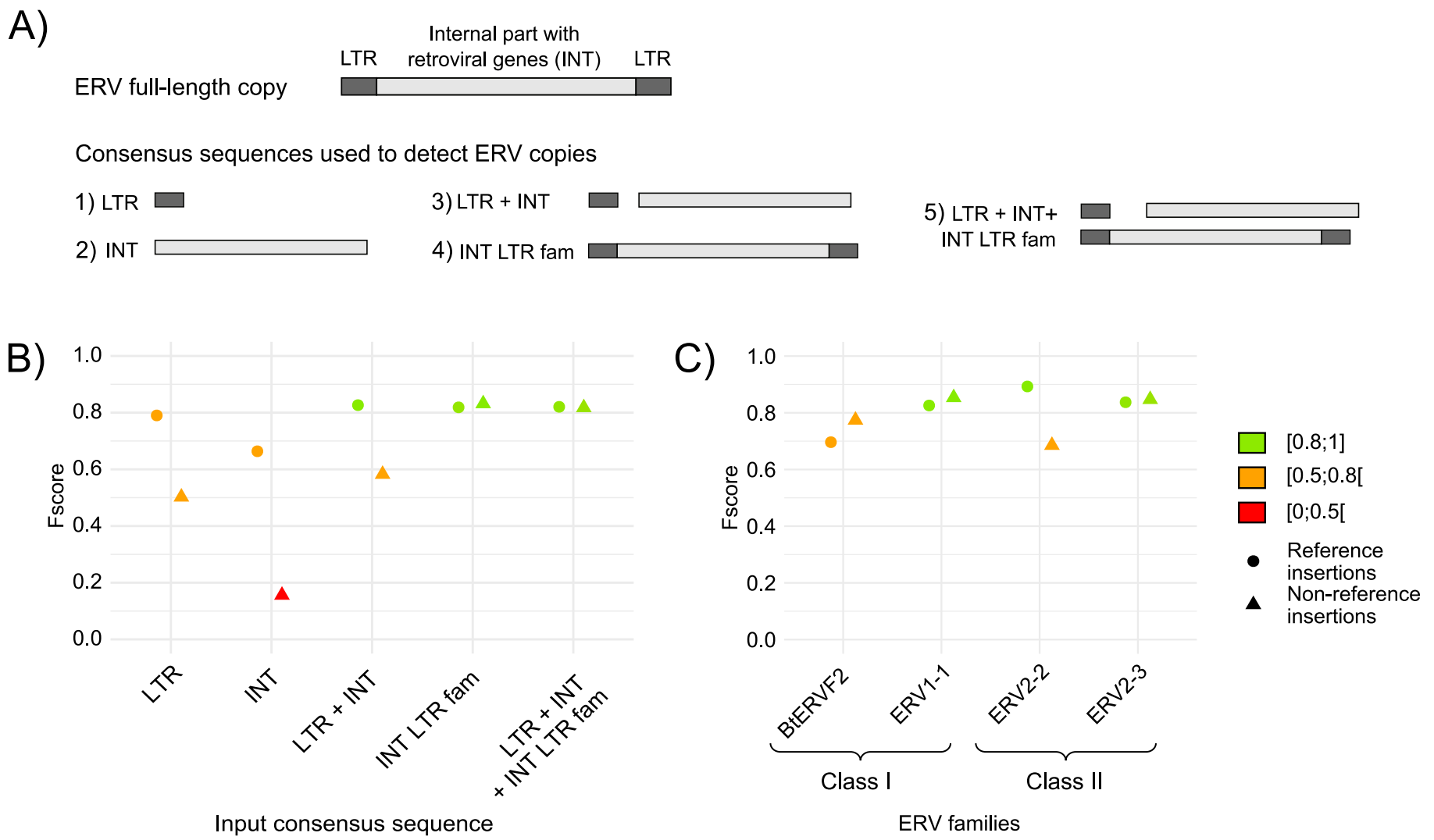
D. melanogaster

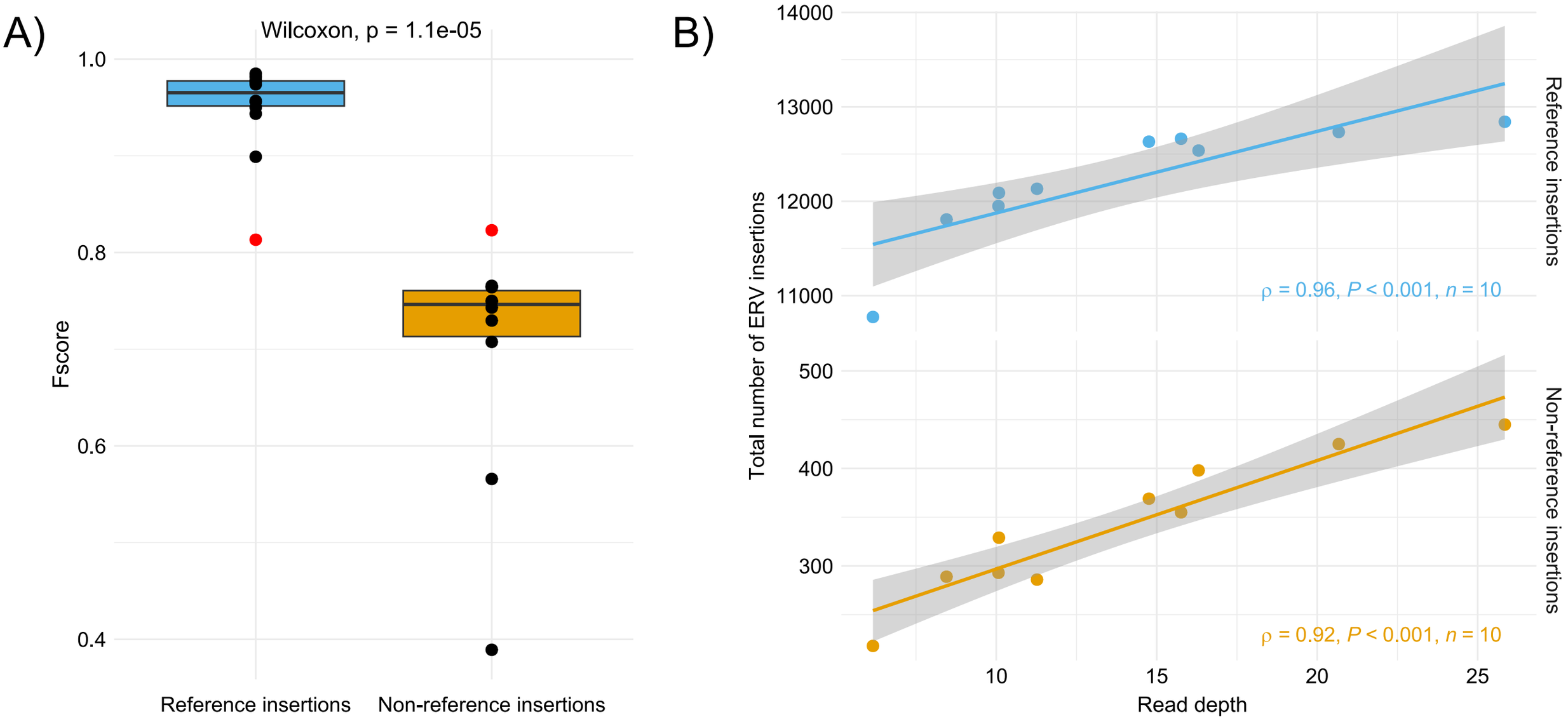


A. thaliana




Reference insertions
Non-reference insertions
Fscore
1
0.8
0.6
0.4
0.2
0
ngs_te_mapper
ngs_te_mapper2
PoPoolationTE
PoPoolationTE2
RelocaTE
RelocaTE2
RetroSeq
TEBreak
TEFLoN
telocate
TEMP
TEMP2
Programs
[0.8;1]
[0.5;0.8[
[0;0.5[


A)
Internal part with
retroviral genes (INT)
LTR
LTR
ERV full-length copy
Consensus sequences used to detect ERV copies
1) LTR
2) INT
3) LTR + INT
4) INT LTR fam
5) LTR + INT+
INT LTR fam
B)
Fscore
1.0
0.8
0.6
0.4
0.2
0.0
LTR
INT
LTR + INT
INT LTR fam
LTR + INT
+ INT LTR fam
Input consensus sequence
C)
Fscore
1.0
0.8
0.6
0.4
0.2
0.0
BtERVF2
ERV1-1
ERV2-2
ERV2-3
Class I
Class II
ERV families
[0.8;1]
[0.5;0.8[
[0;0.5[
Reference insertions
Non-reference insertions


A)
Wilcoxon, p = 1.1e-05
Fscore
1.0
0.8
0.6
0.4
Reference insertions
Non-reference insertions
B)
Total number of ERV insertions
14000
13000
12000
11000
500
400
300
10
15
20
25
Read depth
Reference insertions
Non-reference insertions
ρ = 0.96, P < 0.001, n = 10
ρ = 0.92, P < 0.001, n = 10