Your manuscript has been peer reviewed by two experts in the field. The two reviewers and I agree that your simulation tool and benchmark of TIP detection tools can be a valuable contribution for the TE community. Nevertheless, they have identified room for improvement in the description of the methodology and the integration of your results in the context of previous work. Also, some aspects of the simulation strategy have been criticized. They recommend a number of revisions that can enhance the quality of the work. I will be happy to consider a revised version of the manuscript for recommendation.

Dear Recommender,
We sincerely thank you and the two reviewers for the time and effort dedicated to evaluating our manuscript. We are grateful for their expertise and valuable feedback, which have significantly contributed to enhancing the quality of our work. The updated version provides greater clarity in the description of our methodology and situates our results within the context of prior studies. Below, we provide a detailed point-by-point response to each of the reviewers' comments.

Reviewer 1

General comments:
Verneret et al present a new system for the simulation of TE sequences in eukaryotic genomes, together with a performance analysis of several short-read TE detection systems in Drosophila, Arabidopsis and Bos on simulated data, followed by an analysis of real Bos data based on simulation results. The simulation system generates synthetic genomes with characteristics of real genomes that are present in GenBank. There is no attempt to compare this system to previous simulation systems that have been used to evaluate short-read TE detection software (i.e., simulaTE by Kofler 2018 and the single TE insertion framework by Nelson et al 2017/Chen et al 2023). A full treatment of prior work and the strengths and weaknesses (e.g., limited to reference genomes in Genbank, limited number of TE insertions, generation of completely artificial genomes) of replicaTE vis-à-vis prior studies would benefit the reader greatly.
Answer: We thank the reviewer for their valuable comments, which have helped us improve the reproducibility of our work and better highlight the unique contributions of our study in the context of prior research. We have added information about previous simulation systems to the Introduction and further elaborated on their strengths and limitations in comparison to our approach.
The McClintock framework (Nelson et al. 2017/Chen et al. 2023) uses both real and simulated data from yeast. For simulations, single TEs from four active families are inserted at biologically plausible positions, with several replicates. This method benefits from yeast's small genome size, where all TE insertions are annotated with minimal ambiguity. However, this approach is not easily generalizable to organisms with larger genomes and more complex TE landscapes. Furthermore, it is restricted to active TE families. Same for Rishishwar et al., TE insertions were simulated by randomly inserting human consensus sequences from active families into autosomes. While these methods provide insight into specific active elements, it introduces biases by excluding inactive or older TE families, which also contribute to polymorphism in real genomes.
Our simulation approach addresses these limitations using real TE annotations to include several hundred insertions from diverse families, not necessarily restricted to active ones. This allows the inclusion of TEs with varying divergence profiles, reflecting the true diversity of TE landscapes across species.
Additionally, our tool indeed requires a GenBank format file as an input which is straightforward to generate from BED of the gene and TE annotations and the corresponding FASTA files, annotations which are in any case mandatory prerequisites for using most of the TE detection tools.

The analysis of TE detection system performance is a valuable addition to the field and underscores many themes that have been reported in prior work by Rishiwara et al 2016, Nelson et al 2017, Vendrill-Mir et al 2019 and Chen et al 2023. The manuscript (and readers) would benefit from more effort to synthesize similarities and differences between the current and prior work (e.g., Chen et al. 2023 also report that 50X coverage is recommended for the optimal detection of non-reference insertions).
Answer: In the revised manuscript, we have included additional comparisons in the Discussion section to highlight these aspects. While prior benchmarks, such as those by Rishishwar et al. (2016), Nelson et al. (2017), Vendrell-Mir et al. (2019), and Chen et al. (2023), primarily focused on factors like sequencing coverage and TE types, our study extends beyond these by investigating why certain TE insertions are detected while others are missed. This analysis provides unique insights into the genomic factors influencing TE detection performance, which were not addressed in previous benchmarks.

Limitations of the current evaluation study should also be more thoroughly discussed (i.e., the authors only analyze a single simulation replicate for each species, and thus quantitative differences among TE detection methods may reflect results for only this replicate).
Answer: We appreciate the reviewer's observation regarding the use of a single simulation replicate for each species. While we acknowledge that analyzing multiple replicates could allow for statistical testing, particularly when tools show similar performance, this approach would come with significant computational resource demands due to the large number of insertions considered in our study.

Unlike other studies that focus on a single insertion or a small number of insertions from one or two TE families, our analysis incorporates a broad range of insertions across various genomic contexts. This diversity provides sufficient statistical power to draw robust conclusions from a single replicate.

Lastly, the analysis of Bos data is lacking a direct investigation of sequence characteristics that affect short-read TE detection performance (as is shown for Drosophila and Arabidopsis), as well as key information about access to empirical datasets and methodology to reproduce main findings.
Answer: In our study, *Bos taurus* was used as a case study to illustrate the importance of selecting tools based on the species being analyzed, as the performance results for *Bos* differed from those observed for *Drosophila* and *Arabidopsis*. The primary objective for *Bos taurus* was not to investigate the sequence characteristics affecting TE detection performance, but rather to identify the most suitable tool for analyzing *Bos taurus* real data, specifically for detecting ERV insertions using pseudo-simulated data. This practical application highlights the relevance of species-specific benchmarking to ensure optimal detection performance. We have clarified this in the revised manuscript and added details about the methodology and access to empirical datasets to enhance reproducibility.

Title and abstract
- Does the title clearly reflect the content of the article? I don't know. The title only reflects one of the main results from the paper.
Answer: We chose to focus the title on the added value of our study compared to other benchmark analyses, which lies in the ability to determine the sequence characteristics of missed TE insertions.

- Does the abstract present the main findings of the study? Yes

Introduction
- Are the research questions/hypotheses/predictions clearly presented? Yes
- Does the introduction build on relevant research in the field? No. Prior efforts developing simulation systems to evaluate the performance of TE detection systems are not described (i.e., simulaTE by Kofler 2018 and the single TE insertion framework by Nelson et al 2017/Chen et al 2023).
Answer: In the revised manuscript, references to prior studies have been incorporated into both the Introduction and Discussion sections to provide proper context for our work and highlight how ReplicaTE differs from these efforts.

Materials and methods
- Are the methods and analyses sufficiently detailed to allow replication by other researchers? No.
The version and parameters used to generate the simulated Drosophila, Arabidopsis and Bos genomes using replicaTE are not given.
Answer: We have added the missing version of the genome of *D. melanogaster*. The versions for *B. taurus* (ARS-UCD1.3) and *A. thaliana* (TAIR10) were already mentioned. We have indicated that default parameters were used for the simulation with replicaTE.

The version and parameters used to run the McClintock TE detection system, TEPID and Jitterbug are not given.
Answer: The version and parameters used to run McClintock, TEPID and Jitterbug were added in the text. Default parameters were used for all programs.

The accessions for the data used in the analysis of Bos short and long read sequences are not provided. The version and parameters for analysis of Bos short reads are not provided. The version and parameters for pbsv analysis of Bos long reads are not provided.
Answer: Accession numbers for the data used in *Bos* analysis were added in the Supplementary Table 1. Details about the analysis of *Bos* short and long reads were added in the text.

The software for determining the overlap between TE detectors and simulated data is not described. The software and specific criteria for determining the concordance between short-read and long-read predictions for Bos datasets are not given.
Answer: The overlap between TE detectors results as well as the concordance between short-read and long-read predictions were performed using two homemade scripts. The scripts are now mentioned in the manuscript and available in the Git repository.

- Are the methods and statistical analyses appropriate and well described? No. See above.

Results
- In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? I don't know
- Are the results described and interpreted correctly? No.
It would be helpful for all figures to present Drosophila and Arabidopsis data in the same order. (i.e., make fig 3 and 6 like figs 4, 5, 7 & 8.);
Answer: We have modified the figures according to the reviewer suggestion to present *Drosophila* and *Arabidopsis* in the same order.

It is unclear in Fig 9 which format the TE library is in (LTR and internal combined or separate?).
Answer: In figure 9, the "classical" TE library was used; ie. internal and LTR sequences separately. The figure 9 legend was modified accordingly.

The data in Figure 10 do not support the claim that class I ERVs are "better recognized" than class II ERVs.
Answer: We agree that the difference of TE detection between class I and class II ERVs was not tested statistically so the sentence was deleted from the manuscript.

The lack of access to data and detailed description of methods makes it difficult to evaluate claims about short-read TE detector performance on Bos real data.
Answer: We appreciate the reviewer's feedback regarding the accessibility of data and the detailed description of our methods. In the revised version of the manuscript, we have incorporated additional details and clarified our methodology based on the reviewers' insightful comments. We hope that these improvements address the concerns raised.

The authors do not seem to be aware that some TE detection systems do not attempt to make reference TE predictions (i.e., TEbreak and Retroseq).
Answer: TEbreak and Retroseq are indeed designed exclusively to detect insertions present in the samples but not in the reference genome. To avoid any confusion, we have revised the manuscript to clearly highlight this distinction in comparison to other tools. Additionally, we have updated the corresponding figure legends to explicitly acknowledge that these programs are not intended to detect reference insertions.

Discussion
- Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? No. No comparison to prior TE simulation frameworks is provided. Additionally, unrealistic aspects of the replicaTE system are not discussed (i.e., generation of wholly artificial genomes; generation of TE copy size from an exponential distribution – LTR and TIR elements typically insert with a characteristic size; use of longest TE to represent ancestor – this should be a consensus; access to reference genomes with comprehensive TE annotation as input to simulation).
Answer: We have added a more detailed comparison with the TE simulation approaches used by prior benchmarks in the Introduction and Discussion sections to emphasize how ReplicaTE differs from existing approaches. We have expanded the discussion to address the limitations of our tool, ensuring a balanced presentation of its capabilities and areas for improvement (ie., simulation of only one chromosome, take a genbank file as an input…). While we acknowledge the reviewer's concerns, we respectfully disagree with the characterization of ReplicaTE as unrealistic. A key strength of ReplicaTE lies in its ability to

incorporate TE characteristics derived from real TE sequences, which reflect the true diversity of TEs. This ensures that our simulations are not entirely artificial but instead grounded in biological reality. The distributions we used to generate the sequence characteristics are fitted on the data. Moreover, we disagree with the fact that using a consensus is a good approach. A consensus cannot be considered as the ancestral sequence. It is only a representative sequence built by the alignments of all the copies present in a genome and thus according to the level of divergence of the different copies, it may be completely different from the original active ancestral insertion. Mapping reads on consensus may be a source of information loss as it has been shown in the case of transcriptomic analyses on TEs. In sum, we recreate a TE insertion landscape with a mix of recent and ancient insertions, as observed in the real genome. Finally, having access to "reference genomes with comprehensive TE annotation as input to simulation" is not unrealistic since this is a prerequisite to be able to perform TE polymorphic analyses using the tested tools. It is not possible to make this type of analysis without a proper annotation of TEs with the current approach and we do not think it is wise to use automatic TE annotation that may be included in some of the tools to do it since it does not include any indispensable curation.

- Are the conclusions adequately supported by the results (without overstating the implications of the findings)? Yes

Reviewer 2

Verneret and colleagues generated a benchmark to evaluate the performance of polymorphic transposon insertion detection tools. Specifically, they considered the effect of TE and genomic characteristics to insertion detection, including copy size, divergence, and GC content. This manuscript didn't give suggestions on which tools should be used in certain conditions, but it highlighted all existing tools are sensitive to these characteristics. This is generally a good idea to take into these features into account. However, my biggest concern is that the authors simulated their benchmark based on real TE features, e.g. sequence divergence and truncation, but the real TEs annotated in the reference genome are typically fixed TEs that inserted into the genome millions of years ago and underwent many mutations. That said, a polymorphic TE, which should be inserted into the genome recently, are different to reference TEs. Polymorphic TEs will have much less divergence and less truncation compared to reference TEs where the simulation based on, and this will lead to strong bias. Thus, I suggest the simulation of features should base on not only the reference genome, but also real biological data that gives us an idea of how many divergences and truncation should a real TE insertion/deletion has.

Answer: We thank the reviewer for its thoughtful comments and for raising concerns about the simulation process and its alignment with real biological data. The different points have been detailed below.

Minor comments:

Line 60. The number of insertions should be 4.93 × 10−9 per site per generation.

Answer: This was this exact formulation in the text. We have modified the text to make sure "-9" is superscript.

Title and abstract

- Does the title clearly reflect the content of the article? Yes.
- Does the abstract present the main findings of the study? Yes.

Introduction

- Are the research questions/hypotheses/predictions clearly presented? Yes.
- Does the introduction build on relevant research in the field? Yes.

Materials and methods

- Are the methods and analyses sufficiently detailed to allow replication by other researchers? Yes.
- Are the methods and statistical analyses appropriate and well described? No.

As I mentioned, the simulation process doesn't reflect real biology, thus will produce bias in benchmarking transposon polymorphic detection tools.

Answer: While using real biological data for benchmarking might seem ideal, this approach has significant limitations. High-quality TE annotations are available for only a limited number of species, and even for these, the annotations are often incomplete or imperfect. Missing annotations can lead to an underestimation of TE insertions and, consequently, a higher rate of false positives during benchmarking. In contrast, our simulation approach allows us to control various parameters systematically, providing a robust framework for evaluating the performance of detection tools. We acknowledge the potential bias in simulations based only on the reference genome. However, by incorporating insertions simulated from real annotation data with a range of parameters regarding sequence divergences, sizes and GC content, we aim to capture the diversity of TE characteristics across different scenarios. This flexibility enables us to model conditions that closely approximate real biological data.

We also disagree with the assumption that polymorphic TEs of interest are exclusively recent insertions. While it is true that some polymorphic TEs represent recent insertions and may be involved in particular biological questions like cancer development for example, others may result from more ancient events depending on the genetic distance between the samples and the reference genome, and may nevertheless be involved in important mechanisms like adaptation to environmental changes. The reference genome serves as a snapshot of an individual or a group of individuals at a given time and thus contains insertions that are not necessarily fixed. Some TEs are indeed present in the reference but may be absent in some or all the

analyzed samples. To avoid any confusion, we changed the term “new” insertions into “non-reference” insertions in the whole manuscript.
Our study emphasizes the necessity of conducting species-specific benchmarks when evaluating tools for TE polymorphism detection. By simulating parameters tailored to a species of interest, we can better understand why certain tools perform more efficiently than others under specific conditions. For instance, our results demonstrated that the best-performing tool for *B. taurus* differed from those for *A. thaliana* and *D. melanogaster*, underlining the value of our approach in selecting appropriate tools based on the species being analyzed.

Results

- In the case of negative results, is there a statistical power analysis (or an adequate Bayesian analysis or equivalence testing)? Yes.
- Are the results described and interpreted correctly? Yes.

Discussion

- Have the authors appropriately emphasized the strengths and limitations of their study/theory/methods/argument? Yes.
- Are the conclusions adequately supported by the results (without overstating the implications of the findings)? Yes