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.

Minor comments:
1. Line 60. The number of insertions should be 4.93×10^{-9} per site per generation.

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.

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.