Article

Guideline for genome transposon annotation derived from evaluation of popular TE identification tools

Haidong Yan [1], Federica Torchiana [2] and Aureliano Bombarely [2,*]

[1] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA 24061 USA
[2] Department of Bioscience, Universita degli Studi di Milano Milan, Italy, 20133
* Correspondence: Aureliano.bombarely@unimi.it;

Received: date; Accepted: date; Published: date

Abstract:

Background: Transposable elements (TEs) constitute the vast majority of all eukaryotic DNA, and display extreme diversity, with thousands of families. Given their abundance and diversity, TEs discovery and annotation becomes challengeable. At present, tools and databases have built libraries to mask TEs in genomes based on *de novo*- and homology-based identification strategies, but no consensus criteria about which tools should be used have been proposed.

Results: In the de novo-based strategy, we compared performances of TE libraries developed by four commonly used tools, including RepeatModeler, LTR_FINDER, LTRharvest, and MITE_Hunter, by using a simulated genome as a standard control. The results showed that the performance of RepeatModeler decreased as it was combined with either LTR_FINDER or LTRharvest. Combination of RepeatModeler and MITE_Hunter showed better performance than RepeatModeler and MITE_Hunter alone. In the homology-based strategy, we evaluated different sources from a taxonomic point of view to build an accurate TE library. When we selected a library from databases to identify TEs for *Arabidopsis thaliana* genome, the library from a genus genetically closer to Arabidopsis achieved better performance than other genera with further genetic distance. Without the Arabidopsis, combination of top three genera closer to Arabidopsis showed better performance than combination of all genera.

Conclusion: This study proposes a series of recommendations to perform an accurate TE annotation: 1) For *de novo*-based strategy, RepeatModeler and MITE_Hunter are suggested to build a TE library; 2) For homology-based strategy, it is recommended to use library of genus genetically close to the species rather than use combined library from all genera.

Keywords: transposable elements, genome annotation, software evaluation

1. Introduction

Transposable elements (Transposons; TEs) constitute a majority of all eukaryotic genomes. They are a major evolutionary force through their capability to produce mutations [1]. TE composition is extreme diverse in all kingdoms. Thousands of TE families have been identified in plants, accounting for 60% or more of the total plant genomic DNA [2, 3]. In comparison, TEs in metazoans (3-45%) and fungi (3-20%) represent a relatively smaller part of their genomes [4, 5]. TEs fall into two general classes: Class I elements are retrotransposons, transposing from one position to another via the 'copy-and-paste' mechanism [6]. The Class I elements can be further divided into long terminal repeat (LTR) retroelements and non-LTR retroelements. The LTR TEs (LTRs) are the predominant order in plants, but are less abundant in animals. Class II elements are DNA transposons following a 'cut-and-paste' mechanism [6]. They can be classified into autonomous and non-autonomous elements according to their ability to move by themselves. Miniatures inverted repeat transposable element (MITE) is a special type of non-autonomous DNA transposons with high copy number and special structural features present in all eukaryotic genomes [6]. Given the diversity and abundance of TEs, their discovery and annotation could be challengeable in the eukaryotic genomes.

As of now, two main strategies have been developed to create consensus TE sequences or TE libraries. The first strategy is a homology-based method in which the target sequences are compared with a repository or catalog of known TE sequences defined as TE library. This library can be obtained directly from databases such as RepBase and Dfam [7]. These two databases are common TE-centric repositories containing consensus repeat sequences for each transposon family and subfamily [8]. The second strategy is the *de-novo*-based strategy in which transposons and other repetitive elements are identified based by their specific domains and number of occurrences. Several tools have been developed using this approach, such as RepeatScout, RECON, and RepeatModeler, which utilize consensus seeds or pairwise similarity to cluster repetitive sequences and build TE libraries [9]. For LTR identification, LTR_STRUC leveraged certain structural features, including presence of flanking terminal repeats, target site duplications, primer-binding sites, and polypurine tract [10]. However, it is unable to identify incomplete LTR TEs and it is limited to windows-only implementation, significantly restricting automated large-scale analysis. LTR_FINDER [11] and LTRharvest [12] were developed based on similar principles as LTR_STRUC, but these tools produce large numbers of false positives [13]. LTR_retriever has been developed to efficiently remove false positives from results in LTR_FINDER and LTRharvest, and generates high-quality LTR libraries from genomic sequences [14]. For MITE TEs (MITEs), MITE_Hunter [15], MITE_Digger [16], detectMITE [17], MiteFinderII [18], and MITE Tracker were developed based on the Terminal Inverted Repeat (TIR)-like structure.

These tools are able to aid non-specialists to easily identify and annotate TEs, but most studies identified TEs in a new genome using different strategies and tools. We collected 58 plant genome sequencing studies in 2019 (Table S1). Thirteen studies only utilized the de novo-based strategy to build the TE libraries, and six studies only relied on the homology-based strategy. RepeatModeler was utilized in most studies (78%; 45/58), of which 23 studies only used RepeatModeler, while the other 22 studies used RepeatModeler combines with other tools, such as LTR_FINDER, LTRharvest, or MITE-Hunter (Table S1 and S2). In the homology-based strategy, nearly half of the studies (48%; 28/58) used all TE libraries, and eight studies used species- or genus-specific libraries from the RepBase (Table S2). Taken together, no consensus criteria were built to develop the TE libraries. A guideline needs to be proposed to generate high-quality TE library to accurately mask TEs in genomes.

In the de novo-based strategy, we evaluated performances of four tools, the most frequently used in the collected studies: RepeatModeler, LTR_FINDER, LTRharvest, and MITE-Hunter (Table S2). In order to evaluate the specificity and sensibility of these tools we have developed a simulated genome with randomly inserted TEs. PILER is the fifth most frequent tool used in nine studies (Table S2). It was not included in our evaluation, since its one dependence PALS tool is no longer supported [19]. LTR_retriever was used to generate consensus TE sequences from LTR_FINDER and LTRharvest [14]. Recent studies utilized different sources from databases to build the TE libraries *via* the homology-based strategy (Table S2). We have evaluated different sources from a taxonomic point of view in order to build an accurate TE library for novel genomes using homology-based methods. Taking these results together, we have synthesized a series of recommendations to perform an accurate TE annotation.

2. Materials and Methods

Simulation of genome and TEs

A clean genome without any TE was constructed based on *Arabidopsis thaliana* genome (ver. TAIR10) (https://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Sequences) [20]. TE libraries of *Arabidopsis* from Repbase (ver. 20170127) [21] and Plant Genome and Systems Biology (PGSB) (v9.3p) [22] were used to identify TEs in *Arabidopsis thaliana* genome sequence downloaded from the TAIR database (https://www.arabidopsis.org/download/index-auto.jsp?dir=/download_files/Sequences) *via* RepeatMasker (v. 4.0.7) [7]. The identified TEs were removed from the *A. thaliana* genome. After repeating this filtering process for three times, no TEs were identified in the fourth time RepeatMasker run, considering that the genome was clean of TEs.

A total of 250 LTRs and 250 MITEs were randomly collected from Repbase and PGSB database. Three copy types were generated: 1) Type A: the 250 TEs copied one time. 2) Type B: 25 TEs were randomly extracted from the 250 TEs and copied ten times to form the 250 copies. 3) Type C: ten TEs from the 250 TEs copied 25 times to form 250 copies (Figure 1a). Eight kinds of mutations were separately introduced to the TEs by Simulome tool (v1.2) [23] (Table S3). These mutations contain four major mutation types including single nucleotide changes, nucleotide insertions and deletions, and fusion of inserted TE. For single nucleotide changes, three variation levels were set, including 1%, 5%, and 10% of the total nucleotides for each copy which underwent with random mutations. The deletion and insertion changes were set at the 1%, 5%, and 10% levels, similar to the single nucleotide changes. Different combinations of these mutation types were also generated (Table S3). A total of 3,780 TEs with mutations were generated for each copy type. These TEs were randomly inserted into the clean genome to form a simulated genome (Table S3; Figure 1a). Target Site Duplication (TSD) sequences were introduced to flanking regions of each TE copy, since TSD is one of structures these evaluated tools could detect them. For MITE copy, we set 'TA' as TSD [24]. For LTR copy, we randomly set a short sequence with random nucleotides and 4-6 bp length as TSD [25].

Evaluation of tool performances

The testing tools were used to predict TE locations in the simulated genome (Figure 1a). We used True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) to evaluate the tool performances by the following criteria: 1) if a predicted TE covered the simulated TE, it will be regarded as TP; 2) otherwise, it will be FP; 3) FN was used if no simulated TE was covered; 4) The inter TE region (R) is considered TN, since no FP exists in that region. Four evaluation scores, including Sensitivity (SE), Specificity (SP), Accuracy (AC), and Precision (PR), were calculated to compare performance for each tool (Figure 1b).

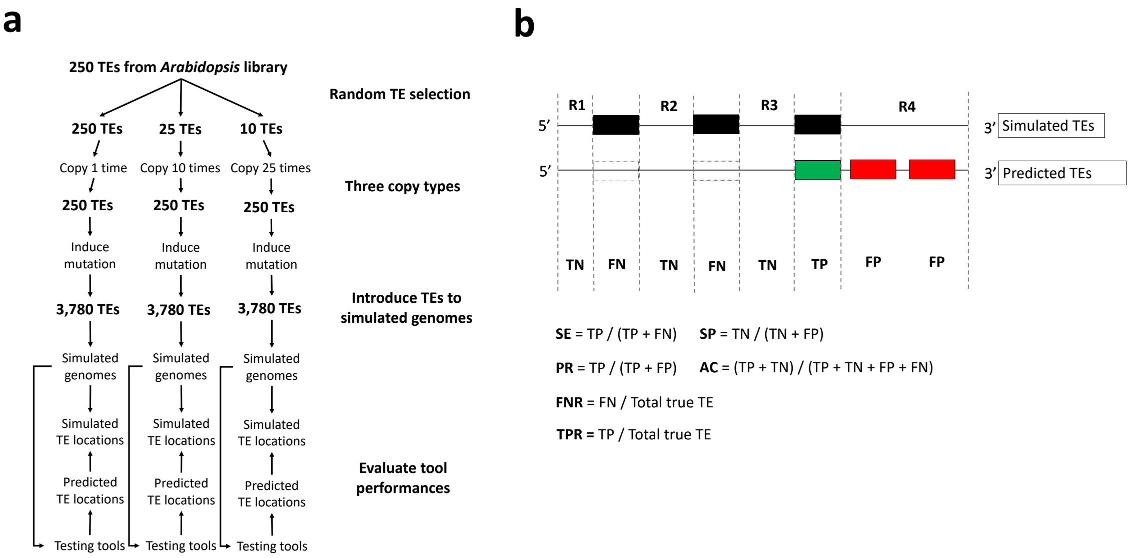

Figure 1 Simulation pipeline. **a,** 250 TEs were collected from *Arabidopsis* TE library from Repbase and PGSB database. A total of 250, 25, and 10 TEs were randomly extracted from the collected TEs to generate 250 TEs by copying one, 10, and 25 times, respectively. Mutations were introduced to the 250 TE copies and 3,780 TEs were generated for each copy type. These TEs were randomly inserted into a clean genome without any TE insertion to form a simulated genome and simulated TE locations. Predicted TE locations were generated by using the testing tools towards the simulated genome. The simulated TE locations were compared with the predicted TE locations to evaluate performances of the testing tools. **b,** Evaluation method. The black box indicates simulated TEs. The dark line suggests sequences between two TEs. True Positive (TP) is defined once the predicted TEs cover with the simulated TEs, otherwise, this TE will be False Positive (FP). False Negative (FN) is defined if no simulated TEs are covered. The inter TE region (R) is defined as True Negative (TN) while no FP exiting in that region. Six evaluation scores including Sensitivity (SE), Specificity (SP), Accuracy (AC), Precision (PR), False Negative Rate (FNR), and True Positive Rate (TPR) are calculated on basis of TP, SE, SP, PR.

Evaluation of performances to identify TEs for each of 22 plant genera

A total of 22 TE libraries from plant genera were collected by combining the Repbase and PGSB database (Table S4). *A. thaliana* transposon location file was downloaded from database (ver. TAIR10) (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_transposable_elements) [20]. TEs from four families including LTR, LINE, SINE, and DNA were extracted from this location file and generated reference TE locations. Each of these 22 libraries was used to mask TEs towards *A. thaliana* genome by RepeatMasker [7] to obtain predicted TE location information. The TE locations from each library were compared with the reference TE locations to evaluate performance of each genus library using a homemade script. Given the current version of the *A. thaliana* reference, TE locations may not contain all true TE information, and it is hard to detect accurate FP, and TN. Therefore, the performance of each genus was evaluated only based on False Negative Rate (FNR) and True Positive Rate (TPR) and SE (Figure 1b).

Pearson correlation analysis between library performance scores for all genera except for *Arabidopsis* and their estimated divergence time (Million Years Ago; MYA) was conducted using R (ver. 3.6.0) [26]. The divergence time between the different generat and *Arabidopsis* was obtained from TimeTree (http://www.timetree.org/) [27]. For Pearson analysis, the evaluation of correlation was based on the following criteria (www.statstutor.ac.uk/resources/uploaded/spearmans.pdf): (1) rho > 0 equated to a positive correlation, (2) rho < 0 signified a negative correlation, (3) very weak or no correlation was indicated by the absolute value of rho < 0.2, (4) weak was indicated by 0.2 ≤ the absolute value of rho < 0.4, (5) moderate was determined by 0.4 ≤ the absolute value of rho < 0.6, (6) strong was determined by 0.6 ≤ the absolute value of rho < 0.8, and (7) very strong was indicated by 0.8 ≤ the absolute value of rho < 1.0.

All scripts for the simulation and evaluation of different methods can be found at Github (https://github.com/yanhaidong1/TEeval).

3. Results

3.1. Comparison of tool performance for identification of TEs with different copy types

In eukaryotic genomes, each TE is duplicated hundreds or thousands of times. To understand performances of the TE identification tools on a genome with different TE copy number, we set three copy types in a simulated genome to compare these tools (Figure 1a). For LTR identification, RepeatModeler, LTR_FINDER, and LTRharvest and their combinations were evaluated under these three copy types (Figure 2 and S2). For accuracy and specificity, combinations of RepeatModeler and LTR_FINDER and of all three tools had significant higher evaluation scores from one to ten TE copy times (Figure 2a, d). No significant difference ($p > 0.05$) was detected in sensitivity (Figure 2b). For precision and specificity, combination of RepeatModeler and LTR_FINDER achieved significant ($p < 0.05$) higher scores in ten than in one copy times (Figure 2c, d). Combination of all three tools achieved increased ($p < 0.05$) specificity from one to ten copy times (Figure 2d). However, from ten to 25 copy times, all tools and combinations did not have a significant change ($p > 0.05$) (Figure 2). Ref-based method was introduced to these tools (Figure S1). The Ref-based method indicates identification of TEs based on a TE library that is from the initial TEs without any duplication. Ref-based method alone had a significant ($p < 0.05$) increasing precision and specificity from one to ten copy times. When combined with RepeatModeler and LTRharvest, there was a significant ($p < 0.05$) decrease in precision and specificity from one to ten copy times (Figure S1c, d).

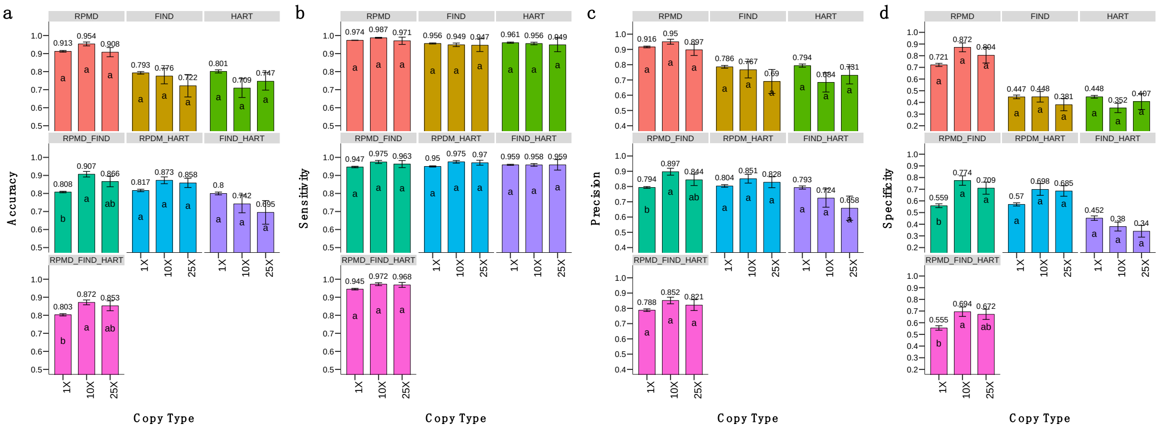

Figure 2 Comparison of tool performances for three TE copy types in LTR TEs. RPMD: RepeatModeler. FIND: LTR_FINDER. HART: LTRharvest. '_' indicates tool combination. 1X: 1 copy time. 10X: 10 copy times. 25X: 25 copy times. The color of bars indicates different tools and their combinations.

For DNA MITE identification, RepeatModeler and MITE_Hunter were analyzed (Figure S2). Accuracy and sensitivity scores significantly ($p < 0.05$) increased from one to ten copy times in RepeatModeler, but decreased from ten to 25 copy times for accuracy (Figure S2a, b). In MITE_Hunter, precision and specificity had a significant ($p < 0.05$) increasing from one to ten copy times (Figure S2c, d). When RepeatModeler combined with MITE_Hunter, precision score also improved ($p < 0.05$) from one to ten copy times (Figure S2c). Ref-based method and its combination with MITE_Hunter had a significant increasing ($p < 0.05$) in all evaluation scores except for

doi:10.20944/preprints202008.0275.v1

sensitivity from one to ten copy times (Figure S2d). Inversely, combination of ref-based method and RepeatModeler showed a decreasing precision from ten to 25 copy times (Figure S2c).

3.2. *Evaluation of performances for tools identifying MITE and LTR transposons*

The MITEs and LTRs are two major families of TEs, and most tools were developed for identifying these two families (Table S1). We evaluated performances of RepeatModeler, LTR_FINDER, and LTRharvest for LTR detection, and of RepeatModeler and MITE_Hunter for MITE detection.

For detecting LTRs, RepeatModeler had the highest evaluation scores comparing with LTR_FINDER and LTRharvest under the one and ten TE copy types (Figure 3a, b), while in the 25 TE copy type, these two LTR tools performed same as RepeatModeler ($p > 0.05$), except for their lower specificity ($p < 0.05$) (Figure 3c). Comparing with single LTR_FINDER or LTRharvest, combinations of RepeatModeler with either one of them had significantly ($p < 0.05$) increased specificity in all three copy types (Figure 3). RepeatModeler also achieved better performance than combination of all three tools at one TE copy type (Figure 3a).

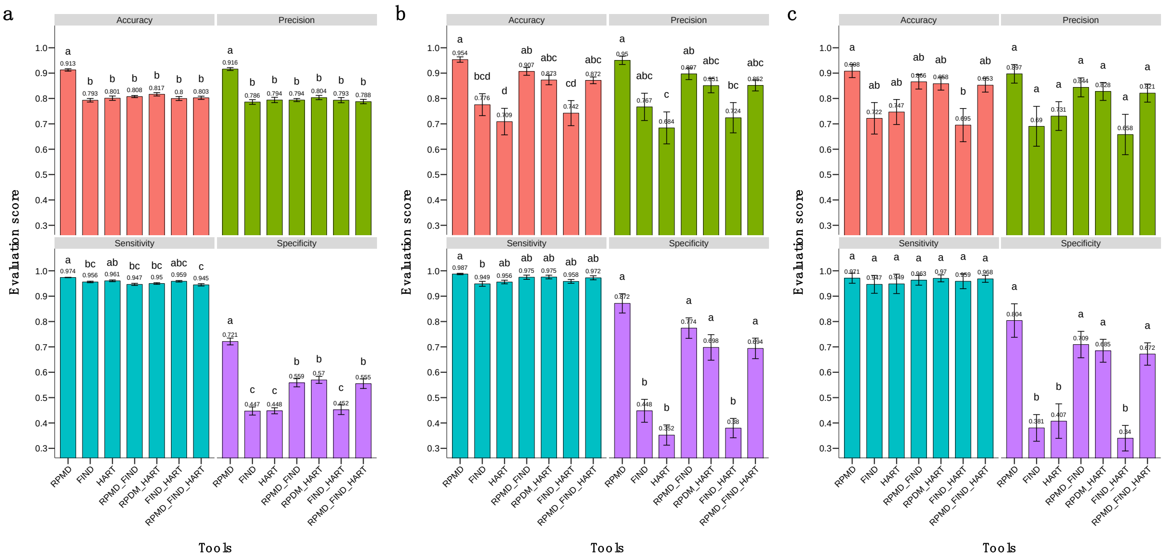

Figure 3 Comparison of performances of tools for LTRs detection. a-c indicate TE copy number is one, ten, and twenty-five, respectively. PRMD: RepeatModeler. FIND: LTR_FINDER. HART: LTRharvest. '_' indicates tool combination. Red bar indicates Accuracy. Green bar indicates Precision. Light blue bar indicates Sensitivity. Purple bar indicates Specificity.

We then compared performances of different tool combinations with ref-based method (e.g. LTRharvest with RepeatMasker). All tools achieved good performances (evaluation scores > 0.8) when they combined with ref-based method (Figure S3). But some of the combinations had significant ($p < 0.05$) lower evaluation scores than ref-based method alone. For example, accuracy and sensitivity of RepeatMasker (ref-based method) declined significantly ($p < 0.05$) when combined with other tools, except when it is combined with RepeatModeler (Figure S3a). Specificity of ref-based method decreased when it was combined with two or more other tools in all three copy types (Figure S3). Performance of ref-based method had overall and significant ($p < 0.05$) decreasing when it was combined with all other three tools (Figure S3), except for sensitivity at the ten and 25 TE copy types (Figure S3b, c). RepeatMasker was superior to RepeatModeler as expected, but performance of their combination did not significantly ($p > 0.05$) decrease relative to ref-based method alone in all three copy types (Figure S3).

In contrast to LTR detection, performance of RepeatModeler was inferior to MITE_Hunter in identification of MITEs in all three copy types (Figure 4). Combination of RepeatModeler and MITE_Hunter had significantly ($p < 0.05$) higher sensitivity but lower specificity than MITE_Hunter alone in all three copy types, but no significant ($p > 0.05$) difference was found in accuracy and precision. The reference based method alone showed better performance than other tools, and the performance did not have significant ($p > 0.05$) changes when it combined with other tools, except for RepeatModeler (Figure 4). Combination of RepeatModeler and ref-based method had lower specificity than ref-based method alone ($p < 0.05$) (Figure 4b, c). When these three approaches combined, performance of this combination did not show significant change ($p > 0.05$) by comparing with ref-based method alone (Figure 4).

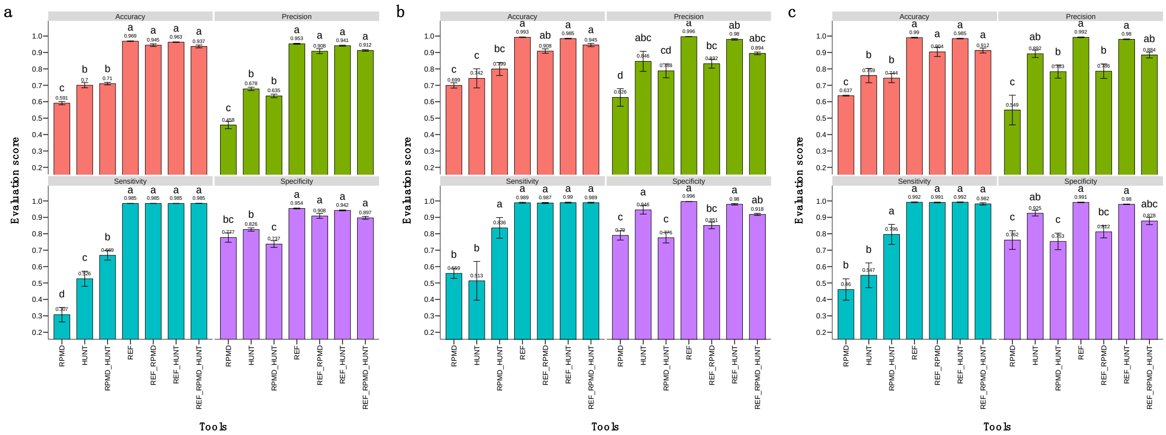

Figure 4 Comparison of performances of tools for MITEs detection. a-c indicate TE copy number is one, ten, and twenty-five, respectively. REF: Ref-based method. PRMD: RepeatModeler. HUNT: MITE-Hunter. '_' indicates tool combination. Red bar indicates Accuracy. Green bar indicates Precision. Light blue bar indicates Sensitivity. Purple bar indicates Specificity.

3.3. *Evaluation of impacts of taxonomic distance on performance of TE identification using the homology-based method*

About half of studies (48%; 28/58) used a combined library from all species to identify TEs, while eight studies used species- or genus- specific libraries in the RepBase (Table S2). To clarify which strategy could achieve better prediction, we evaluated performance of TE library from each of 22 plant genera (Table S4) and used *A. thaliana* as the reference genome.

The library of *Arabidopsis* showed the highest sensitivity and TPR, and the lowest FNR. It outperformed the combined library with all genera (Figure 5a, b, c). *Medicago*, *Triticum*, *Malus*, and *Solanum* performed better than other genera. *Chlamydomonas* displayed the worst performance. *Gossypioides* had lower FNR that nine genera, but it had lower sensitivity and TPR than other 20 genera (Figure 5a, b, c).

To test hypothesis that the closer genetic background to *Arabidopsis* for a genus, the better performance of this genus library when it is used to identify TEs in the *A. thaliana* genome, we collected estimated divergence time from all genera except for *Arabidopsis* from TimeTree (http://www.timetree.org/) [27]. Pearson correlation analysis was conducted between the estimated divergence time from the 21 genera to *Arabidopsis* and their three evaluation scores. A strong negative correlation was detected between the divergence time and the evaluation scores sensitivity (*rho* = -0.72, $p < 0.05$) and TPR (*rho* = -0.70, $p < 0.05$), respectively (Figure 5d, e), while a strong positive correlation (*rho* = 0.71, $p < 0.05$) was found between the divergence time and FNR (Figure 5f).

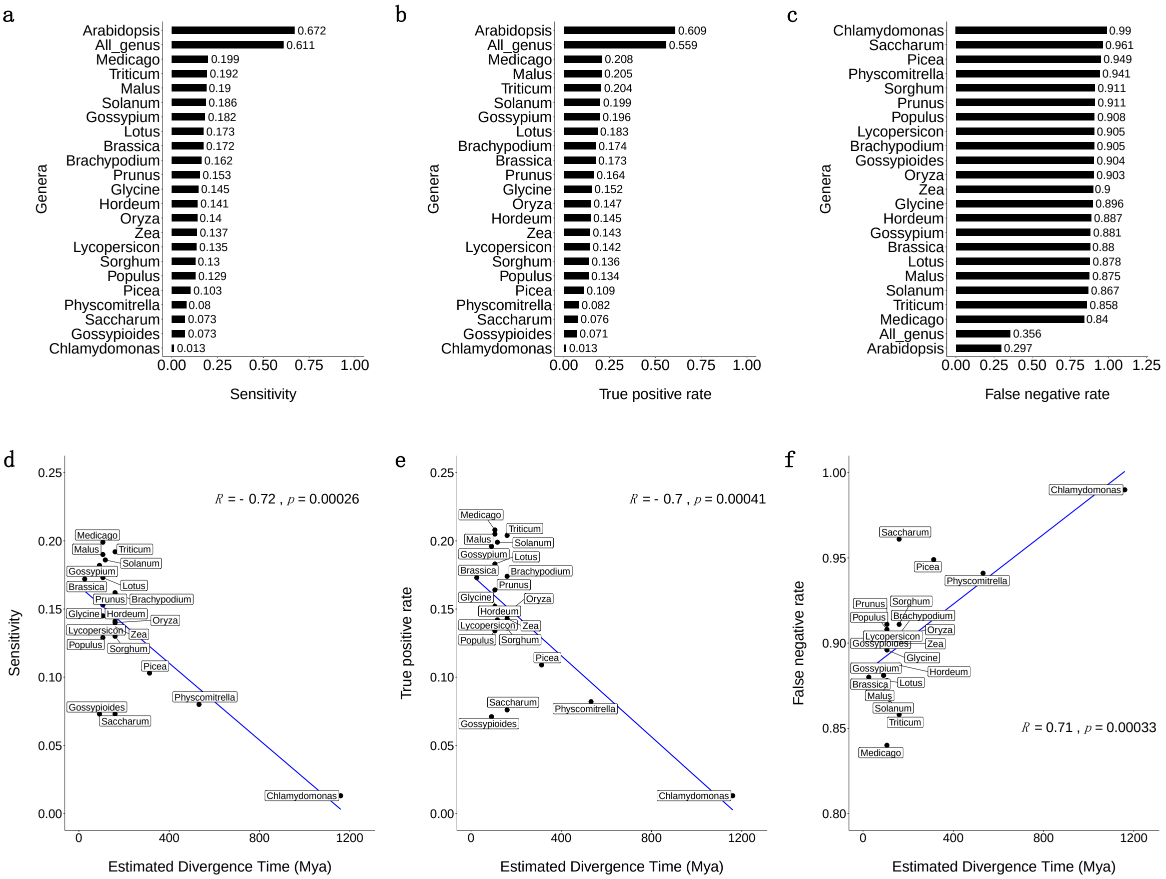

Figure 5 Evaluation of performances from TE libraries of 22 plant genera. a-c indicate ranking order for the 22 genera is based on scores from sensitive, true positive rate, and false negative rate, respectively. d-e indicate Pearson correlation plots between estimated divergence time from 21 genera to *Arabidopsis* and their performance scores (SE, TPR, and FNR).

4. Discussion

TEs are considered to be a main component in genomes [1]. They are highly associated with genome size and chromosomal rearrangements, and can provide regulatory sequences affecting nearby genes [28]. The decrease of cost and improvement of efficiency of new sequencing techniques are promoting sequencing of increasing numbers of genomes [29]. Approximately 300 genome sequencing projects have been completed in plants in recent five years (2010-2019) (https://www.plabipd.de/timeline_view.ep). Most of these studies are from non-model plants lacking annotation of TEs in database. The accurate and efficient annotation of TEs is therefore crucial to our understanding of their influence on genome evolution and gene function. To handle the increasing genomes, we developed guidelines assisting to precisely identify TEs in the genomes *via* building a good TE library.

As novel tools and data types have been developed for biological data, simulations are becoming increasingly essential to bioinformatics researches on developments, testing, and benchmarking [23]. Simulations have been utilized in many cases, such as analyze accuracy of gene expression profiling [30], increase reads mapping quality based on simulated genomes in bacterial strains [31], and correction of read bias in RNA-seq mapping [32]. TE annotation in genomes is challengeable. It is hard to comprehensively identify all TEs in a genome, due to their diversity and abundance [33, 34]. For example, we masked TE consensus sequences of *Arabidopsis* from RepBase [7] to inter-TE regions in the transposon location file downloaded from TAIR database (https://www.arabidopsis.org/download/index-auto.jsp?dir=%2Fdownload_files%2FGenes%2FTAIR10_genome_release%2FTAIR10_transposable_elements). A total of 13,124 TEs could be identified through RepeatMasker [7], suggesting this reference file in *Arabidopsis* does not contain all TE insertions. Thus, it is inadvisable to compare performances of different tools on genomes lacking information of entire TE insertions. This study introduced SNP, deletion, insertion, and fusion events on TEs derived from library in *Arabidopsis*, which simulated evolutionary mutations of TEs in genomes. These TEs were inserted into *Arabidopsis* genome where TEs were manually removed, to generate a reference genome with known TE insertions. Although this reference can be utilized to evaluate performance of any tool, it is still hard to simulate real situation. For example, TEs are not distributed equally in the genome [35] and the simulated mutations on TEs might not match real mutations during genome evolution.

Nearly half of studies (48%; 28/58) used the combined TE library from all species in RepBase (Table S2). This study found the conjunct library achieved worse performance than the library from *Arabidopsis*. Also, the combined library from LTR_FINDER, LTRharvest, and RepeatModeler performed worse than the library only from RepeatModeler (Figure 3). These observations suggest some consensus TEs in the combined libraries are spurious TEs that generate FP when they are used to mask TEs in genomes through local alignment algorithm in RepeatMasker [7]. The quality of the TE library is related to effects of TE detection in genomes.

LTR_FINDER and LTRharvest performed worse than RepeatModeler, since they generated a large number of FP TEs. These two LTR tools leveraged conserved structure to identify LTRs in genomes [11, 12]. The structural based method may falsely detect TEs in one sequence region with sort of LTR structure, but this sequence is not real LTR. This could explain lower precision and specificity in these two tools. Also, few FN was found in these two tools, suggesting they can identify almost all TP TEs (Figure 3).

Different simulated TE copy types can influence performances of tools for identifying TEs (Figure 2; Figure S1; Figure S2). Some approaches may achieve positive relationship between their performances and TE copy times. When comparing with one TE copy, high copy times (10 and 25) increased performances in ref-based method (Figure S1 and S2), MITE_Hunter (Figure S2), and combination of RepeatModeler and LTR_FINDER (Figure 2). These tools or tool combinations may be efficient to be used in genome with high copy number of TEs. In the contrast, combination of ref-based method, RepeatModeler and LTRharvest (Figure S1) showed a negative relation between the performances and copy times, suggesting this combination could be applied in genome with low copy number of TEs.

Performances of different tool combinations depend on evaluation scores of specific tools inside these combinations. Evaluation scores from one tool may decrease as it combines with another tool with lower scores. For example, performance of ref-based method decreases by combing other tools (Figure 4; Figure S3). Precision and specificity of RepeatModeler decline when it combines with LTR_FINDER and LTRharvest with the lower scores (Figure 3). One exception in the MITE detection is that sensitivity from combination of RepeatModeler and MITE_Hunter is raised relative to RepeatModeler or MITE_Hunter alone (Figure 4).

According to performances of the four common tools to identify TEs, we synthesized two recommendations for the de novo-based strategy:

1) RepeatModeler is recommended to be applied to predict the LTRs, and combination of RepeatModeler and other tools is not recommended;

2) MITE_Hunter and RepeatModeler are recommended to combine to identify the MITEs.

For the homology-based strategy, we propose:

1) if target species has its reference library in a database, it is recommended to use this library rather than combined library from all genera from databases (e.g. RepBase, Dfam, and PGSB), and

2) if the library of target species is not available, it is recommended to build a combined library from its close genera with similar genetic background.

Based on the recommendations, a guideline with the examples of the commands and tools in a Linux environment is proposed to construct a comprehensive TE library and annotate the TE in a genome

genome (Box 1). We will use maize 'B73' genome [36] as an example downloaded from GenBank [37] (https://www.ncbi.nlm.nih.gov/datasets/genomes/?acc=GCF_000005005.2).

Box 1: Instructions to build a TE library and annotate the TE

1- Copy the assembly of interest to a new directory.

```
    CMD1.1: cp genome.fasta my_genome_repeats
CMD1.2: cd my_genome_repeats
```

2- Build a reference database for RepeatModeler and run it.

```
CMD2.1: BuildDatabase -name my_genome genome.fasta
CMD2.2: RepeatModeler -database my_genome -pa 10
CMD2.3: 'cp consensi.fa.classified repeatmodeler.lib
```

3- Run MITE_Hunter.

```
CMD3.1:   MITE_Hunter_manager.pl -i genome.fasta -g mite_hunter -S
12345678
CMD3.2: cp mite_hunter_Step8_singlet.fa mite_hunter.lib
```

4- Combine libraries generated from RepeatModeler and MITE_Hunter.

```
CMD4.1: cat repeatmodeler.lib mite_hunter.lib > denovo.lib
```

5- Build a homology-based TE library using RepeatMasker.

```
CMD5.1: RepeatMasker/util/queryRepeatDatabase.pl -species Zea
CMD5.2: cp Zea.out homology.lib
```

6- Combine the de novo-based and homology-based libraries.

```
CMD6.1: cat denovo.lib homology.lib > combine.lib
```

7- Run RepeatMasker to mask TEs against the maize genome.

```
  CMD7.1: RepeatMasker genome.fasta -lib combine.lib -gff -dir
output_dir
```

5. Conclusions

We evaluated performances of four commonly used tools to identify TEs in genomes including RepeatModeler, LTR_FINDER, LTRharvest, and MITE_Hunter. A simulated sequence randomly inserted by TEs with mutations was constructed to build a reference to evaluate different parameters for these tools such as precision and sensitivity. To build an accurate TE library for novel genomes using homology-based method, we also evaluated different sources from a taxonomic point of view. Based on the evaluation results, we provide a series of recommendations to perform an accurate TE annotation and propose a guideline to develop a comprehensive TE library.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, **Table S1** TE identification tools used in plant genome sequencing studies in 2019; **Table S2** Summary of tools and sources used for TE identification in 2019; **Table S3** TE copy number for each mutation type; **Table S4** Library collection from 22 genera from RepBase and PGSB datasets; **Figure S1** Comparison of tool performances with ref-based method for three TE copy types in LTR TEs. **Figure S2** Comparison of tool performances for three TE copy types in MITE TEs. **Figure S3** Comparison of performances of tools with ref-based method for LTRs detection.

Author Contributions: Conceptualization, A.B. and H.Y; methodology, A.B. and H.Y.; formal analysis, H.Y.; writing—original draft preparation, H.Y.; writing—review and editing, A.B. and F.T.; supervision, A.B.; project administration, A.B.; All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding. H.Y. stipend were funded by the School of Plant and Environmental Sciences, Virginia Tech.

Acknowledgments: We would like to acknowledge to the Advance Resources Computing at Virginia Tech for the use of the servers.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chuong, E.B.; Elde, N.C., Feschotte, C. Regulatory activities of transposable elements: from conflicts to benefits. *Nature Reviews Genetics* **2017**, *18*, 71.
2. Morgante, M. Plant genome organisation and diversity: the year of the junk! *Current Opinion in Biotechnology* **2006**, *17*, 168-73.
3. Feschotte, C.; Jiang, N., Wessler, S.R. Plant transposable elements: where genetics meets genomics. *Nature Reviews Genetics* **2002**, *3*, 329.
4. Daboussi, M.-J., Capy, P. Transposable elements in filamentous fungi. *Annual Reviews in Microbiology* **2003**, *57*, 275-99.
5. Hua-Van, A.; Le Rouzic, A.; Maisonhaute, C., Capy, P. Abundance, distribution and dynamics of retrotransposable elements and transposons: similarities and differences. *Cytogenetic and genome research* **2005**, *110*, 426-40.
6. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B., et al. A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **2007**, *8*, 973.
7. Smit, A.; Hubley, R., Green, P. RepeatMasker Open-4.0. 2013–2015. 2015.
8. Goerner-Potvin, P., Bourque, G. Computational tools to unmask transposable elements. *Nature Reviews Genetics* **2018**, 1.
9. Smit, A.; Hubley, R., Green, P. RepeatModeler Open-1.0. 2008–2015. *Seattle, USA: Institute for Systems Biology Available from: httpwww repeatmasker org, Last Accessed May* **2015**, *1*, 2018.
10. McCarthy, E.M., McDonald, J.F. LTR_STRUC: a novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, *19*, 362-7.
11. Xu, Z., Wang, H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic acids research* **2007**, *35*, W265-W8.
12. Ellinghaus, D.; Kurtz, S., Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC bioinformatics* **2008**, *9*, 18.
13. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. *Heredity* **2010**, *104*, 520.
14. Ou, S., Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant physiology* **2018**, *176*, 1410-22.
15. Han, Y., Wessler, S.R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research* **2010**, *38*, e199-e.
16. Yang, G. MITE Digger, an efficient and accurate algorithm for genome wide discovery of miniature inverted repeat transposable elements. *BMC bioinformatics* **2013**, *14*, 186.
17. Ye, C.; Ji, G., Liang, C. detectMITE: a novel approach to detect miniature inverted repeat transposable elements in genomes. *Scientific reports* **2016**, *6*, 19688.
18. Hu, J.; Zheng, Y., Shang, X. MiteFinderII: a novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC medical genomics* **2018**, *11*, 101.
19. Edgar, R.C., Myers, E.W. PILER: identification and classification of genomic repeats. *Bioinformatics* **2005**, *21*, i152-i8.
20. Lamesch, P.; Berardini, T.Z.; Li, D.; Swarbreck, D.; Wilks, C.; Sasidharan, R., et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic acids research* **2012**, *40*, D1202-D10.

21. Bao, W.; Kojima, K.K., Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **2015**, *6*, 11.

22. Spannagl, M.; Nussbaumer, T.; Bader, K.C.; Martis, M.M.; Seidel, M.; Kugler, K.G., et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. *Nucleic acids research* **2015**, *44*, D1141-D7.

23. Price, A., Gibas, C. Simulome: a genome sequence and variant simulator. *Bioinformatics* **2017**, *33*, 1876-8.

24. Muñoz-López, M., García-Pérez, J.L. DNA transposons: nature and applications in genomics. *Current genomics* **2010**, *11*, 115-28.

25. Havecker, E.R.; Gao, X., Voytas, D.F. The diversity of LTR retrotransposons. *Genome biology* **2004**, *5*, 225.

26. Allaire, J. RStudio: integrated development environment for R. *Boston, MA* **2012**, *537*, 538.

27. Kumar, S.; Stecher, G.; Suleski, M., Hedges, S.B. TimeTree: a resource for timelines, timetrees, and divergence times. *Molecular biology and evolution* **2017**, *34*, 1812-9.

28. Flutre, T.; Duprat, E.; Feuillet, C., Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PloS one* **2011**, *6*, e16526.

29. Bourque, G. Transposable elements in gene regulation and in the evolution of vertebrate genomes. *Current opinion in genetics & development* **2009**, *19*, 607-12.

30. Hirsch, C.D.; Springer, N.M., Hirsch, C.N. Genomic limitations to RNA sequencing expression profiling. *The Plant Journal* **2015**, *84*, 491-503.

31. Price, A., Gibas, C. The quantitative impact of read mapping to non-native reference genomes in comparative RNA-Seq studies. *PloS one* **2017**, *12*, e0180904.

32. Vijaya Satya, R.; Zavaljevski, N., Reifman, J. A new strategy to reduce allelic bias in RNA-Seq readmapping. *Nucleic acids research* **2012**, *40*, e127-e.

33. Rhoads, A., Au, K.F. PacBio sequencing and its applications. *Genomics, proteomics & bioinformatics* **2015**, *13*, 278-89.

34. Belton, J.-M.; McCord, R.P.; Gibcus, J.H.; Naumova, N.; Zhan, Y., Dekker, J. Hi–C: a comprehensive technique to capture the conformation of genomes. *Methods* **2012**, *58*, 268-76.

35. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M., et al. Ten things you should know about transposable elements. *Genome biology* **2018**, *19*, 199.

36. Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S., et al. The B73 maize genome: complexity, diversity, and dynamics. *science* **2009**, *326*, 1112-5.

37. Benson, D.A.; Karsch-Mizrachi, I.; Lipman, D.J.; Ostell, J., Sayers, E.W. GenBank. *Nucleic acids research* **2011**, *39*, D32.