TEsorter: an accurate and fast method to classify LTR-retrotransposons in plant genomes

Ren-Gang Zhang¹,² #, Guang-Yuan Li² #, Xiao-Ling Wang³, Jacques Dainat⁴, Zhao-Xuan Wang⁵, Shujun Ou⁶*, Yongpeng Ma¹*
¹ Yunnan Key Laboratory for Integrative Conservation of Plant Species with Extremely Small Populations, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, 650201, China
² Department of Bioinformatics, Ori (Shandong) Gene Science and Technology Co., Ltd., Weifang, Shandong, 261322, China
³ BGI-Shenzhen, Shenzhen 518083, China
⁴ Department of Medical Biochemistry and Microbiology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden
⁵ Shijiazhuang People’s Medical College, Shijiazhuang, Hebei, 050091, China
⁶ Department of Ecology, Evolution, and Organismal Biology (EEOB), Iowa State University, Ames, IA, 50010, USA

* Authors for contact: zhangrengang@ori-gene.cn (R.G.Z.), oushujun@iastate.edu (S.O.) or mayongpeng@mail.kib.ac.cn (Y.P.M.)

# These authors contributed equally to this work.

© The Author(s) 2022. Published by Oxford University Press. All rights reserved. This is an Open Access article distributed under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/4.0/, which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Dear the editor,

Transposable elements (TEs) constitute the largest portion of repetitive sequences in many eukaryotic genomes, with long terminal repeat retrotransposons (LTR-RTs) being predominant in plant genomes. Various tools have been developed for the identification and classification of TEs, including RepeatModeler\(^1\), REPET\(^2\), LTR_retriever (https://github.com/oushujun/LTR_retriever) and TERL (https://github.com/muriloHoracio/TERL). To our knowledge, most of these can only classify TEs to the superfamily level, in particular the LTR-RT Copia and Gypsy superfamilies in plants, leaving the gap for further classifications. Moreover, although approaches for the automated classification of LTR lineages using amino acid hidden Markov models (HMMs) do exist, but to our knowledge these are used only as script collections, supporting biological studies of plant TEs, and are not curated or specifically designed with the user in mind.

Previous studies have proposed classifications of LTR-RTs on the clade level \(^3\). Particularly, Neumann et al. \(^4\) classified the Copia superfamily into the Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco I–IV, Ikeros, Ivana, Osser, SIRE, TAR, and Tork clades and the Gypsy superfamily into the CRM, Chlamyvir, Galadriel, Ten1, Reina, Tekay, Athila, Tat I–III, Ogre, Retand, Phygy, and Selgy clades. These studies provide protein domain databases for clade-level LTR-RT classifications. Moreover, the update of REXdb \(^4\) also provides classifications for other TEs, such as long interspersed nuclear repeats (LINEs), terminal inverted repeats (TIRs), and Helitrons (http://repeatexplorer.org/?page_id=918). In this study, based on previous classifications of conserved protein domains, we develop an automated, easy-to-use, and accurate classifier, named TEsorter, to perform superfamily-level classification of TEs and to further classify LTR-RTs into detailed clades. The Python code is freely available at https://github.com/zhangrengang/TEsorter.

The TEsorter package is implemented in Python3 and accelerated using multiprocessing. The conda installation approach is supported, which should ease the installation of TEsorter and integration of the package with other workflows. The TEsorter classifier was implemented using HMM profiles obtained from the TE protein domain databases GyDB (http://gydb.org) and REXdb \(^4\). For REXdb, multiple sequence alignments on domains of each clade were performed using MAFFT (https://mafft.cbrc.jp/alignment/software/) and HMM profiles were generated with HMMBuild\(^5\).

To classify the input TE sequences, the TE sequences are first translated in all six frames and
the translated sequences are then searched against one of the two databases using HMMScan. Hits with coverage < 20% or E-value > 1e-3 are discarded. For each domain with multiple hits, only the best hit with the highest score is retained (Figure 1a). The classifications of TE superfamilies (e.g. LTR/Copia, LTR/Gypsy) and clades (e.g. Reina and CRM of Gypsy) are based directly on hits. For the Copia and Gypsy superfamilies, complete elements are identified based on the presence and order of conserved domains including capsid protein (GAG), aspartic proteinase (AP), integrase (INT), reverse transcriptase (RT), and RNase H (RH) as described in Wicker et al.

Mutations like frameshifts and domain losses may obstruct HMM-based classifications. To improve the classification sensitivity, a two-pass strategy was implemented to classify non-autonomous TEs based on their sequence-level similarity to autonomous TEs (Figure 1a). The unclassified TE sequences are searched against the HMM-classified sequences using BLAST and then classified with the 80–80–80 rule (≥ 80bp of alignment, ≥ 80% of sequence identity, and ≥ 80% of sequence coverage). To comply with alignment uncertainties, this step only classified sequences at the superfamily level.

To benchmark the classification performance of TEsorter, we firstly selected two curated TE libraries from rice (https://github.com/oushujun/LTR_retriever) and maize with 2, 431 and 1, 546 element sequences, respectively. We then compared TEsorter with five TE classifiers, including the RepeatClassifier module of RepeatModeler, the machine-learning-based classifiers DeepTE (https://github.com/LiLabAlVT/DeepTE), TERL (https://github.com/muriloHoracio/TERL), the annotate_TE module of LTR_retriever (https://github.com/oushujun/LTR_retrieve), and the online-only LTRclassifier (http://LTRclassifier.ird.fr/). We noted here that other software also able to classify TEs has not been included due to the difficulty of direct comparison with TEsorter. For example, Inpactor (https://github.com/simonorozcoarias/Inpactor) required LTR structural features and does not support sequences as sole input. TEClass (http://www.compgen.uni-muenster.de/teclass), REPCLASS and PASTEC only provided confident classifications at order level, not allowing further comparison with the six software in the present study.

Another advantage of TEsorter over the above-mentioned software is that clade assignments to LTR-RT Copia or Gypsy elements can also be performed. To date, however, there is no available
reference (or TE library) able to evaluate the precision of these clade-level assignments by TEsorter. We therefore performed phylogenetic analyses based on the hypothesis that LTR-RT elements that were classified as being in the same clade are likely to have closer phylogenetic relationships. Briefly, protein domain sequences were extracted using TEsorter and aligned with MAFFT (https://mafft.cbrc.jp/alignment/software/), and the phylogenetic trees were reconstructed using IQ-TREE using the JTT matrix distance model and bootstrap values of $\geq 50\%$ after 1000 replicates. To further validate the phylogenetic relationships of the clades assigned by TEsorter, we selected the same domains (i.e., RT, RH, INT and concatenated RT–RH–INT) that were used by Neumann et al. to construct phylogenetic relationships among clades of LTR-RT Copia and Gypsy.

Our results show that TEsorter has the highest precision for classifying LTR-RTs when compared with the other TE classifiers tested (Figure 1b). When classifying the LTR-RT Copia and Gypsy superfamilies, the precision values of TEsorter with REXdb were both 1 in rice, and 0.966 and 1 in maize (Figure 1b). LTR_retriever and LTRclassifier performed the same precision when classifying LTR-RT Copia in rice; however, this dropped significantly in maize, and was also lower for the identification of LTR-RT Gypsy in rice (Figure 1b). We also tentatively tested the powers of these six pieces of software to classify TEs other than LTR-RTs. TEsorter with REXdb also had the highest precision here, with values of 1 in rice and 0.997 in maize (Figure 1b).

Unlike precision values, which were consistently higher in TEsorter, sensitivity varied among these TE classifiers. Specifically, TEsorter with REXdb and DeepTE had the highest value of sensitivity in classifying LTR-RT Copia and LTR-RT Gypsy in rice, respectively. RepeatClassifier had the highest sensitivity in classifying both LTR-RT Copia and LTR-RT Gypsy in maize (Figure 1b). We noted that some TEs have become non-autonomous and could not be classified into superfamilies due to the loss of their characteristic protein domains, which was confirmed by searching against the Pfam database (http://pfam.xfam.org).

For the execution time, the superiority of TEsorter is very evident. RepeatClassifier needed to take more than 10 hours to finish the calculation, while TEsorter needed less than 10 minutes for the same calculation. Furthermore, TEsorter performed better with REXdb than with GyDB in most cases (Figure 1b) due to the systematic collection of plant LTR-RTs by Neumann et al. Overall, all these results suggested that TEsorter is a well-rounded and competitive classifier at the
superfamily level.

In addition to classification at superfamily levels, TEsorter was able to assign 76.8–91.9% of LTR-RT *Copia* or *Gypsy* elements into diverse clades in plants (Figure 1b). Moreover, the clade-level classification of TEsorter was found to be highly consistent (ranging from 99.06% based on the RT domain to 100% on the concatenated RT–RH–INT) with the reconstructed phylogeny (Fig. 1c-fa). Furthermore, phylogenetic relationships among these clades detected by TEsorter were matched well with the clade classification of LTR-RT elements firstly raised by Neumann *et al.* 4. These results revealed that Tesorter classifies with high-confidence classifications at the clade level by TEsorter and suggested its ability to reflect the diversity of and phylogenetic relationships within the classified LTR-RTs.

Overall, TEsorter has demonstrated substantial improvements over the current tools in terms of precision and execution time. Moreover, it demonstrates high-confidence classifications of LTR-RTs at the clade level. Since the development of TEsorter, many people have started using the tool and we believe that more people will benefit from it in the future.

**Author contributions**

R.Z., S.O. and Y.M. designed the study. R.Z., X.W., J.D. and Z.W. performed the experiments and analysed the data. R.Z., S.O. and Y.M. wrote the paper.

**Funding**

This work was supported by the Reserve Talents for Academic and Technical Leaders of Middle-aged and Young People in Yunnan Province (Grant No. 2018HB066), Ten Thousand Talent Program of Yunnan Province (Grant No. YNWR-QNBJ-2018-174) and the Key Basic Research program of Yunnan Province, China (grant no. 202101BC070003).

**Availability of data and materials**

The code in Python is freely available at [https://github.com/zhangrengang/TEsorter](https://github.com/zhangrengang/TEsorter).

**Acknowledgments**

We thank Dr. Pavel Neumann for the notification of the release of REXdb and Dr. Jia-Hui Chen for suggestions for the analysis of RT domains in LTR-RTs and LINEs.

**Ethics approval and consent to participate**

Not applicable.

**Competing interests**

...
The authors declare no competing financial interests.

References

1. Flynn, J. M. et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc. Natl. Acad. Sci. U S A. 117, 9451–9457 (2020).
2. Quesneville, H. et al. Combined evidence annotation of transposable elements in genome sequences. PLoS Comput. Biol. 1, e22 (2005).
3. Llorens, C., Munoz-Pomer, A., Berlad, L., Botella, H. & Moya, A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol. Direct. 4, 41 (2009).
4. Neumann, P., Novák, P., Hoštáková, N. & Macas, J. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA. 10, 1 (2019).
5. Eddy, S.R. Profile hidden Markov models. Bioinformatics. 14, 755–763 (1998).
6. Wicker, T. et al. A unified classification system for eukaryotic transposable elements. Nature Review Genetics. 10, 973–982 (2007).
7. Schnable, P.S. et al. The B73 maize genome: complexity, diversity, and dynamics. Science. 326, 1112–1115 (2009).
8. Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L. & Levine, D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol. 1, 205–220 (2009).
9. Hoede, C. et al. PASTEC: an automatic transposable element classification tool. PLoS one. 9, e91929 (2014).
10. Nguyen, L.T., Schmidt, H.A., Haeseler, A.V. & Minh, B.Q. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32, 268–274 (2015).
**Figure 1.** (a) A flowchart illustrating the TEsorter pipeline. (b) Comparison of performance among five previously developed TE classifiers as well as TEsorter. (c-f) High consistency between classifications of TEs assigned by TEsorter, and predicted phylogenetic relationships between these TEs based on the RT (99.06%, c), RH (99.29%, d), INT (99.62%, e), and concatenated RT–RH–INT (100%, f) domains in rice. Conflicts are highlighted with black circles. For detailed information, see https://github.com/zhangrengang/TEsorter/tree/master/example_data. Branches are colored based on TEsorter classifications.

| Library | Classifier          | LTR- and LTR/Gypsy sensitivity (%) | LTR/Gypsy sensitivity (%) | all LTR/RTs sensitivity (%) | other TEs sensitivity (%) | CPU time (hour) |
|---------|---------------------|-----------------------------------|--------------------------|-----------------------------|--------------------------|-----------------|
| Rice    | TEsorter (REXdb)    | 0.931 1.000 85.1% 0.786 1.000 78.6% 0.765 0.994 0.160 1.000 0.09  |
|         | TEsorter (GyDB)     | 0.931 1.000 85.3% 0.768 0.989 78.6% 0.765 0.994 NA NA 0.15  |
|         | RepeatClassifier*   | 0.887 0.922 NA 0.768 0.864 NA 0.773 0.908 0.396 0.881 11.3  |
|         | DeepTE              | 0.874 0.842 NA 0.866 0.713 NA 0.826 0.813 0.671 0.954 0.3  |
|         | TRL_classifier      | 0.818 0.435 NA 0.728 0.698 NA 0.729 0.522 0.186 0.828 0.03  |
|         | LTR_sorter          | 0.868 1.000 NA 0.830 0.979 NA 0.614 0.991 NA NA 0.01  |
|         | LTR_sorter         | 0.824 1.000 NA 0.576 0.679 NA 0.645 0.822 NA NA 1.0  |
| Maize   | TEsorter (REXdb)    | 0.919 0.966 91.9% 0.930 1.000 91.8% 0.793 0.998 0.329 0.997 0.1  |
|         | TEsorter (GyDB)     | 0.914 0.977 89.7% 0.922 0.991 90.6% 0.770 0.998 NA NA 0.12  |
|         | RepeatClassifier*   | 0.908 0.821 NA 0.971 0.707 NA 0.878 0.958 0.365 0.938 12.8  |
|         | DeepTE              | 0.914 0.700 NA 0.963 0.671 NA 0.602 0.925 0.503 0.905 0.01  |
|         | TRL_classifier      | 0.541 0.543 NA 0.791 0.488 NA 0.725 0.710 0.464 0.882 0.02  |
|         | LTR_retriever       | 0.892 0.859 NA 0.918 0.878 NA 0.757 1.000 NA NA 0.01  |
|         | LTR_sorter          | 0.780 0.913 NA 0.664 0.818 NA 0.547 0.916 NA NA 1.2  |

* TEs of rice and maize were excluded from Repeatbase for RepeatClassifier.
† Percentage of elements that were assigned into clades.
Sensitivity = (true positive) / (true positive + false negative) and precision = (true positive) / (true positive + false positive).
NA, not applicable.