

RESEARCH

HiTE, an Ensemble Method for High-Precision Transposable Element Annotation

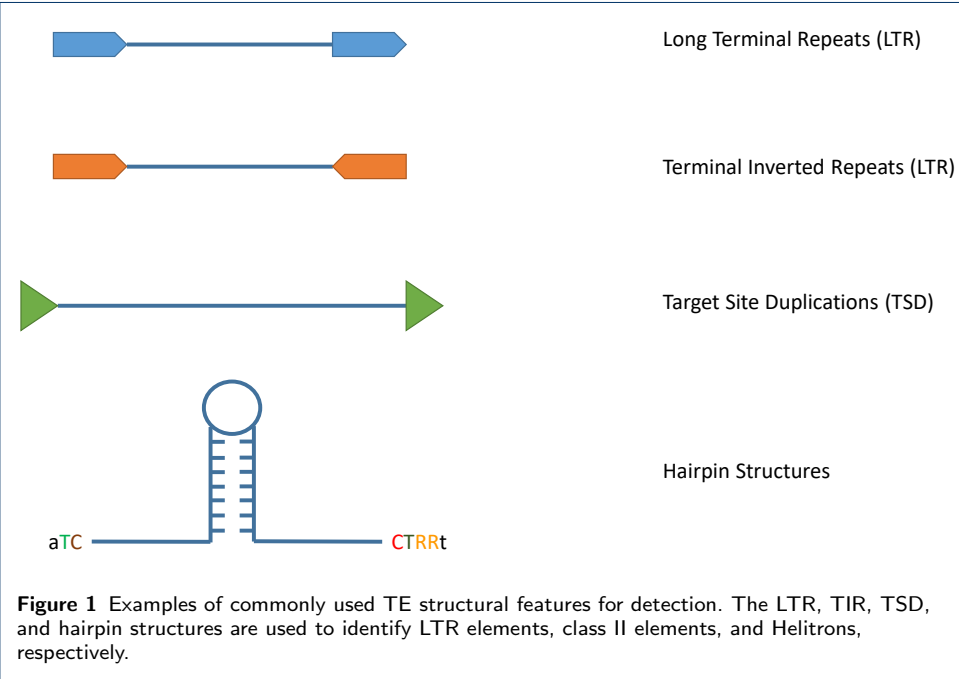
Kang Hu and Jianxin Wang*

*Correspondence:
jxwang@mail.csu.edu.cn
Hunan Provincial Key Lab on
Bioinformatics, School of
Computer Science and
Engineering, Central South
University, Changsha, 410083,
China
Full list of author information is
available at the end of the article

Abstract
Background: Text for this section.
Results: Text for this section.
Conclusions: Text for this section.
Keywords: sample; article; author

Background

Text and results for this section, as per the individual journal’s instructions for authors.



Results

In eukaryotic genomes, transposable elements (TEs) exist widely in the form of both full-length (structurally intact) and fragmented sequences. An ideal library should contain only full-length models of all significantly distinct TEs that have left copies in the genome [1], which are then used to detect fragmented and divergent TE sequences that are hard to recognize using structural features. However, compared

with the limited number of full-length TE sequences, fragmented sequences are more abundant and comprise the majority of TE, creating a challenge for algorithms to find the true ends of the TEs.

The identification methods based on sequence repeatedness tend to produce more fragmented TE sequences, and their sequence boundaries are often approximate, which still need extensive editing before they can be accepted in curated databases such as Dfam or Repbase. For example, the majority of TE models in Dfam come from libraries generated by RepeatModeler, and the great majority of Dfam submissions are currently housed in a non-curated section[2]. Structure-based methods can clearly define the boundaries of the TE structure, but always with a high number of false positives.

To evaluate the performance of different TE identification methods, a high diversity of benchmarking approaches has been proposed, which is a barrier to both the understanding of the true performance of a method and the competitive em of methods. For example, many em methods promote getting the higher copy number of TE, the higher number of models generated, the longer sequences of output, and the higher N50 of the library, which does not take into account the quality of the dataset produced[1].

An ideal method of em should be able to consider both the integrity of TE sequences and the false-positive rate of the TE library. To solve this problem, we take the em methods from the latest study, EDTA[3] and RepeatModeler2[4], which could produce ten metrics including *Sensitivity*, *Specificity*, *Accuracy*, *Precision*, *FDR*, *F1*, *Perfect*, *Good*, *Present*, and *Not_found*. By combining the two complementary em methods, we can accurately evaluate the integrity of TE families and the false-positive rate of the whole TE library.

Selecting benchmarking model species

Despite the universality and importance of TEs in genomes, except for a few model species, the annotation and research of TEs in other species are still poor. In this benchmarking, we mainly focus on 4 typical species: *Oryza sativa*, *Caenorhabditis briggsae*, *Drosophila melanogaster*, and *Danio rerio*, whose TE libraries are well studied and preserved. These four species cover genomes of different sizes as well as different TE landscapes. The smallest genome, the *C. briggsae* genome, is dominated by DNA transposons; the *D. melanogaster* genome is primarily composed of LTR and LINE transposons; the proportion of LTR and DNA transposons on *Oryza sativa* was close, along with a medium-sized genome; and the largest genome, the *D. rerio* genome, comprises the majority of DNA transposons but also some LTR transposons.

Repbase Update (RU) is a database of representative repeat sequences in eukaryotic genomes that has a long history of TE discovery and annotation since 1992[5]. RU has long been used as a manually curated reference database for nearly all eukaryotic genome sequence analyses. Here, we used TE libraries from RepBase26.05 as the gold standard for all species. The RepBase libraries were then used to annotate the genomes for both structurally intact and fragmented TE sequences, which comprised 47.81% of the *O. sativa* genome, 15.83% of the *C. briggsae* genome, 20.28% of the *D. melanogaster* genome, and 57.36% of the *D. rerio* genome, respectively (Tables 1, 2, 3, and 5).

Table 1 TE content in the *O. sativa* (*Oryza sativa* Japonica Group “assembly IRGSP-1.0”) genome.

	Class	RepBase26.05	Number of elements	Total (%)
LTR	Class I	88.4 Mb	46595	23.61
Non-LTR	Class I	5.7 Mb	13381	1.51
TIR	Class II	67.7 Mb	230281	18.09
Helitron	Class II	17.2 Mb	66469	4.60
Total	-	179.0 Mb	356726	47.81

Table 2 TE content in the *C. briggsae* (*Caenorhabditis briggsae* “assembly CB4”) genome.

	Class	RepBase26.05	Number of elements	Total (%)
LTR	Class I	0.2 Mb	234	0.2
Non-LTR	Class I	0.9 Mb	3085	0.59
TIR	Class II	14.5 Mb	68146	13.41
Helitron	Class II	1.8 Mb	8509	1.63
Total	-	17.4 Mb	79974	15.83

Table 3 TE content in the *D. melanogaster* (*Drosophila melanogaster* “assembly Release 6 plus ISO1 MT”) genome.

	Class	RepBase26.05	Number of elements	Total (%)
LTR	Class I	19.9 Mb	21050	11.78
Non-LTR	Class I	10.8 Mb	15428	6.37
TIR	Class II	2.6 Mb	6204	1.53
Helitron	Class II	1.0 Mb	4822	0.60
Total	-	34.2 Mb	47504	20.28

Table 4 TE content in the *D. rerio* (*Danio rerio* “assembly GRCz11”) genome.

	Class	RepBase26.05	Number of elements	Total (%)
LTR	Class I	119.0 Mb	296556	7.09
Non-LTR	Class I	74.3 Mb	215393	4.42
TIR	Class II	719.4 Mb	3372500	42.84
Helitron	Class II	50.5 Mb	178913	3.01
Total	-	963.2 Mb	4063362	57.36

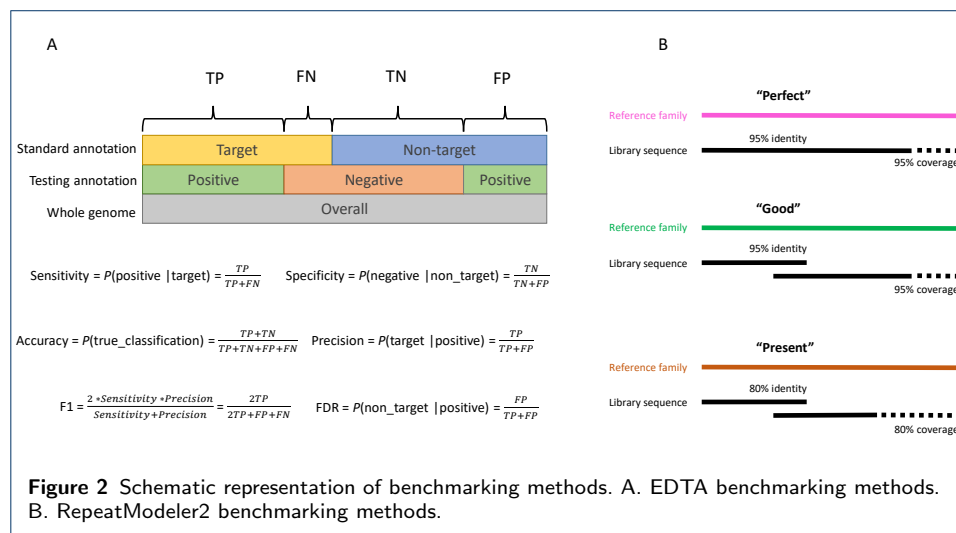
Setting up benchmarking methods for TE library evaluation

To fairly and comprehensively measure the quality of TE libraries generated by different TE identification tools, we use the benchmarking methods from the latest study, EDTA and RepeatModeler2. For convenience, we hereafter refer to the benchmarking methods of EDTA and RepeatModeler2 as BM_EDTA and BM_RM2, respectively.

As shown in Fig. 2A, the BM_EDTA evaluates the performance of various tools by annotating the genome with the gold standard TE library and the tested TE library generated by these tools. Based on the total number of genomic DNA bases, six metrics, including sensitivity, specificity, accuracy, precision, FDR, and F1, are used to characterize the annotation performance of the tested library[3]. The BM_EDTA can display detailed metrics, including the rates of false positives, which are common in many TE identification methods. However, it cannot reflect the integrity of the TE models. All general repeat identification programs, even those with many fragments and unclear boundaries, still performed well in benchmarking[3]. For ex-

ample, while a 1 kbp intact TE sequence and ten 100 bp fragments may obtain the same performance, the former is obviously more valuable in terms of TE integrity and biological significance.

As shown in Fig. 2B, the BM_RM2 aligns the tested TE library with the gold standard library and divides the gold standard sequences into four levels: “Perfect”, “Good”, “Present”, and “Not_found”. “Perfect” families are those for which one sequence in the tested library matches with >95% sequence similarity and >95% length coverage to a family consensus in the gold standard library. “Good” families are those in which multiple overlapping sequences in the tested library match with >95% similarity and >95% coverage to the curated consensus. A family is considered “present” if one or multiple library sequences align with >80% similarity and >80% coverage to the reference consensus sequence. Below these thresholds, a family is considered “not found” [4]. The BM_RM2 takes the integrity of the sequence into consideration. Intact TE models usually get a perfect level, while fragments can only get a good, present, or even not found level. However, it cannot display the rate of false positives in the tested TE library. By combining the two complementary benchmarking methods, we can accurately evaluate the integrity of TE models and the rate of false positives in the whole TE library.



Comparison of general-purpose repeat annotators

We compared HiTE with the other three mainstream general-purpose repeat annotators, including RepeatScout, EDTA, and RepeatModeler2. Among them, EDTA is the best annotation tool based on structure, and RepeatModeler2 is the pipeline with the best overall performance. RepeatScout was originally used to identify repetitive sequences, and its algorithm characteristics tend to find highly consistent repetitive regions, such as duplicates or the youngest TE families. For older and more divergent TEs, many fragments are often generated. RepeatModeler, the old version of RepeatModeler2, uses RECON and RepeatScout for de novo TE identification. Although RECON uses the single linkage clustering algorithm to generate TE sequences based on overlapping alignments, accurate clustering of these alignments is

challenging due to the high fragmentation and mosaicism present in TE families[1]. Therefore, the structurally intact TE models generated by RepeatModeler are not satisfactory (a low number of "perfect" models). To solve this problem, RepeatModeler2 adds the LTR_retriever module to the RepeatModeler, which generates structurally intact LTR transposons and greatly increase the number of "Perfect" in the results[4].

The results of BM_RM2 are shown in Fig. 3A. Among these methods, we found that RepeatModeler2 is stable on all datasets; the performance of EDTA is unstable, obtaining a high number of perfect models on the TE-rich genomes (Tables 1 and 5) but a low number on the other genomes. At the same time, the number of presents is significantly higher than that of other tools, indicating that its results contain many fragments. RepeatScout obtained more perfect sequences on *C. briggsae*, indicating that the majority of TE in the *C. briggsae* genome are relatively young, while in other species, it obtained the minimum number of perfect sequences. Since it cannot process more than 1 GB of genomes, it has no results for *D. rerio*. HiTE has the highest number of perfect TE models than other tools on all datasets and a smaller number of good and present TE models, which shows that HiTE can recognize more structurally intact TE models and fewer fragments; we found that all tools have quite a few "Not_Found" TE models, and we discussed in **Section Comparison of TIR annotators**.

The benchmarking results using BM_EDTA are shown in Fig. 3B. We noticed that RepeatScout and RepeatModeler2 both achieved a consistent high performance, which verified that "all general repeat identification programs, which depend on sequence repeats, performed well" as described in the EDTA[3]. The greatest advantage of the BM_EDTA is that it can intuitively describe the false positive rate of the TE library, but it cannot reflect the integrity of the TE models. For example, in *O. sativa* and *D. melanogaster*, RepeatScout has the lowest number of perfect TE models, indicating that there are a large number of fragments, but it has a high BM_EDTA performance. However, we noticed that on all datasets, HiTE shows significantly higher precision performance, including precision, specificity, and accuracy, which indicates that HiTE can identify TE more accurately. Like the structure-based method EDTA, HiTE achieves a similar low sensitivity, which does not mean that HiTE recognizes fewer TE models than RepeatScout and RepeatModeler2. On the contrary, HiTE obtains more perfect TE models and fewer not_found TE models from the BM_RM2 results. Since the BM_EDTA is based on base statistics, some false-positive sequences with short length can be well aligned to the true TEs, resulting in falsely high sensitivity but significantly low precision. Each TE sequence in HiTE has complete structural characteristics and copy verification (at least two full-length copies exist, and the region outside the TE copy boundary has no homology), which leads to high precision and somewhat lower sensitivity. The reason that RepeatScout and RepeatModeler2 can obtain higher sensitivity is that they identify TE based on the repeatedness of the sequence, so many short and incomplete TE models will also be identified, which are filtered out in HiTE.

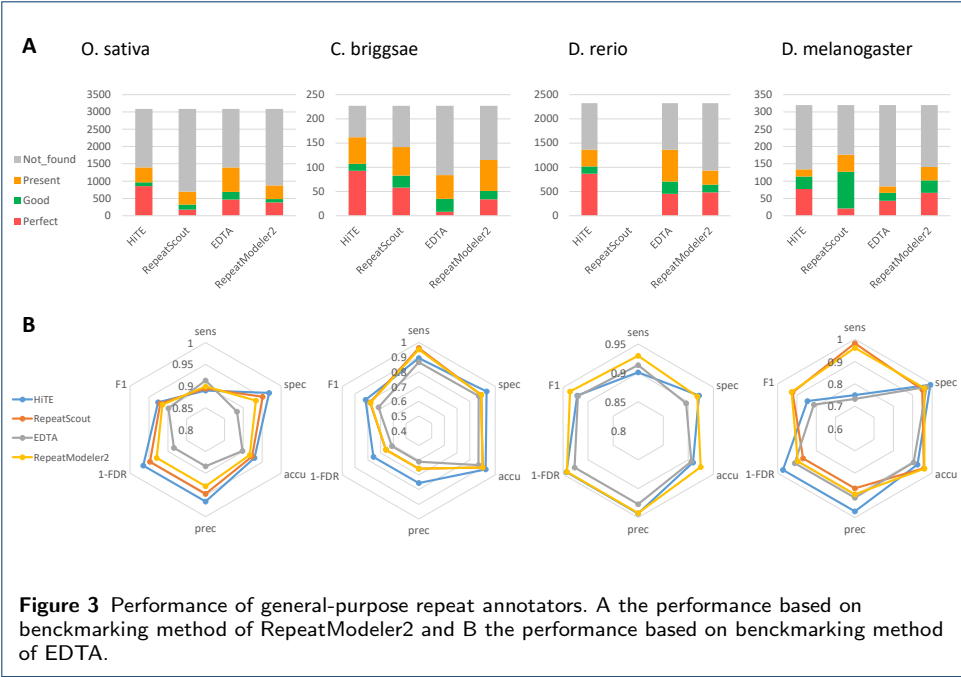


Figure 3 Performance of general-purpose repeat annotators. **A** the performance based on benchmarking method of RepeatModeler2 and **B** the performance based on benchmarking method of EDTA.

Table 5 TE content in the *D. rerio* (Danio rerio “assembly GRCz11”) genome.

	Class	RepBase26.05	Number of elements	Total (%)
LTR	Class I	119.0 Mb	296556	7.09
Non-LTR	Class I	74.3 Mb	215393	4.42
TIR	Class II	719.4 Mb	3372500	42.84
Helitron	Class II	50.5 Mb	178913	3.01
Total	-	963.2 Mb	4063362	57.36

Table 6 Details of performance among general-purpose repeat annotators based on *O. sativa*.

Tools	EDTA evaluation						RepeatModeler2 evaluation			
	Sensitivity	Specificity	Accuracy	Precision	FDR	F1	Perfect	Good	Present	Not_found
HiTE	0.8897	0.9686	0.9295	0.9653	0.0347	0.9260	858	106	430	1690
RepeatScout	0.8940	0.9514	0.9230	0.9476	0.0524	0.9200	173	149	375	2387
EDTA	0.9131	0.8831	0.8979	0.8840	0.1160	0.8983	464	225	709	1686
RepeatModeler2	0.8993	0.9336	0.9166	0.9299	0.0701	0.9144	385	94	394	2211

Table 7 Details of performance among general-purpose repeat annotators based on *C. briggsae*.

Tools	EDTA evaluation						RepeatModeler2 evaluation			
	Sensitivity	Specificity	Accuracy	Precision	FDR	F1	Perfect	Good	Present	Not_found
HiTE	0.8926	0.9323	0.9247	0.7550	0.2450	0.8181	93	14	55	65
RepeatScout	0.9616	0.8875	0.9011	0.6583	0.3417	0.7815	58	25	59	85
EDTA	0.8656	0.8690	0.8683	0.6115	0.3885	0.7167	8	27	49	143
RepeatModeler2	0.9519	0.8876	0.8995	0.6576	0.3424	0.7778	34	17	64	112

Comparison of TIR annotators

TIR TEs, which belong to class II TEs, are ancient TEs found in almost all eukaryotes. They are flanked by characteristic terminal inverted repeat sequences (TIRs), usually presenting in low to moderate numbers[6]. TIR TEs may contribute to

Table 8 Details of performance among general-purpose repeat annotators based on *D. rerio*.

Tools	EDTA evaluation						RepeatModeler2 evaluation			
	Sensitivity	Specificity	Accuracy	Precision	FDR	F1	Perfect	Good	Present	Not_found
HiTE	0.9012	0.9213	0.9094	0.9430	0.0570	0.9216	868	147	347	960
RepeatScout	-	-	-	-	-	-	-	-	-	-
EDTA	0.9134	0.8955	0.9061	0.9268	0.0732	0.9201	453	252	659	958
RepeatModeler2	0.9297	0.9180	0.9249	0.9421	0.0579	0.9359	480	159	294	1389

Table 9 Details of performance among general-purpose repeat annotators based on *D. melanogaster*.

Tools	EDTA evaluation						RepeatModeler2 evaluation			
	Sensitivity	Specificity	Accuracy	Precision	FDR	F1	Perfect	Good	Present	Not_found
HiTE	0.7491	0.9917	0.9243	0.9718	0.0282	0.8461	77	36	21	186
RepeatScout	0.9824	0.9493	0.9577	0.8685	0.1315	0.9220	21	106	49	144
EDTA	0.7322	0.9721	0.9046	0.9113	0.0887	0.8120	43	23	19	235
RepeatModeler2	0.9614	0.9613	0.9613	0.8956	0.1044	0.9273	66	36	39	179

genome evolution by generating allelic diversity, inducing structural variation, and regulating gene expression[7]. TIR TEs are divided into nine known superfamilies by the distinguished TIR sequences and the TSD size (usually 2–11 bp). However, due to the short terminal inverted repeat sequences, TIR TE identification and annotation are quite challenging. For example, members of the hAT superfamily have TSDs of 8 bp and relatively short TIRs of 5–27 bp[8].

Many tools have been designed for their identification, such as IRF[9], TIRvish[10], TIR-Learner[7], and GRF[11], which identify TIR elements by structural signals and are comprehensively evaluated in EDTA. Unfortunately, due to the short structural characteristics of TIR, these methods discover a high number of false positives. For example, the IRF and GRF-TIR produce a large number of candidates, with 4.7 GB and 630 GB (13x–1684x the size of the 374 MB rice genome, respectively) of raw TIR candidate sequences. Among these tools, the TIR module (GRF and TIR-learner) of EDTA has demonstrated great promise for structural annotation and achieved higher performance than other tools[3]. However, it is far from high-precision TIR identification. To solve this problem, we developed a new method to achieve high-precision TIR TE identification (see the “Methods” section).

As shown in Fig. 4, according to the benchmarking results of the BM_RM2, our method can identify more perfect TE models, while the number of good and present models is lower. According to the benchmarking results of the BM_EDTA, our method has higher sensitivity, specificity, accuracy, precision, F1, and a lower FDR than EDTA. These two benchmarking methods both demonstrate that our method can achieve high-precision identification of TIR TEs.

However, we do observe some losses of real TIR TEs, which are mainly caused by the following reasons: (i) Repbase contains a large number of single-copy sequences, even zero-copy sequences. To ensure the high reliability of identified transposons, we filtered single-copy TEs, which require high homology with known transposons or TE proteins to identify. For zero-copy sequences, it is possible that these sequences come from multiple genomes of the same species, such as different types of rice, which we cannot identify based on a single genome, or they are from degraded nested

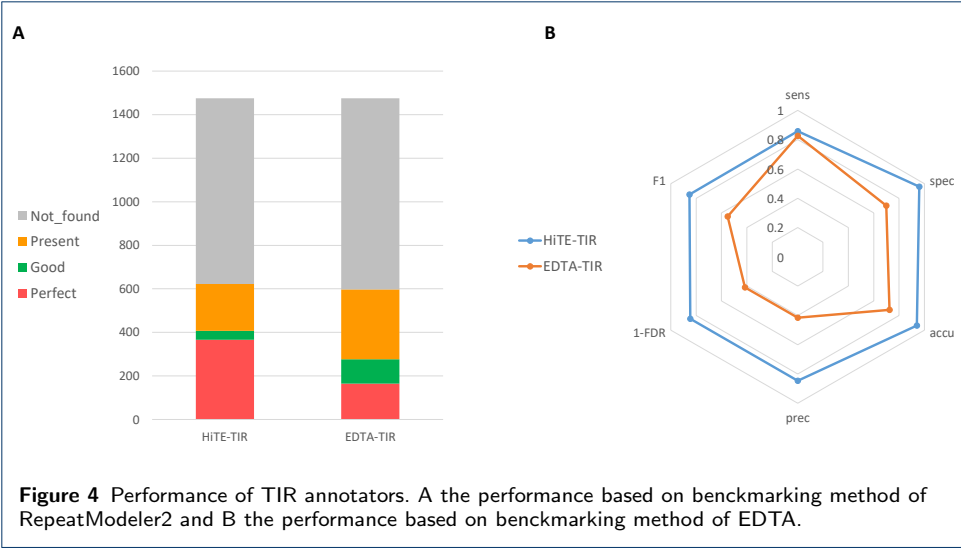


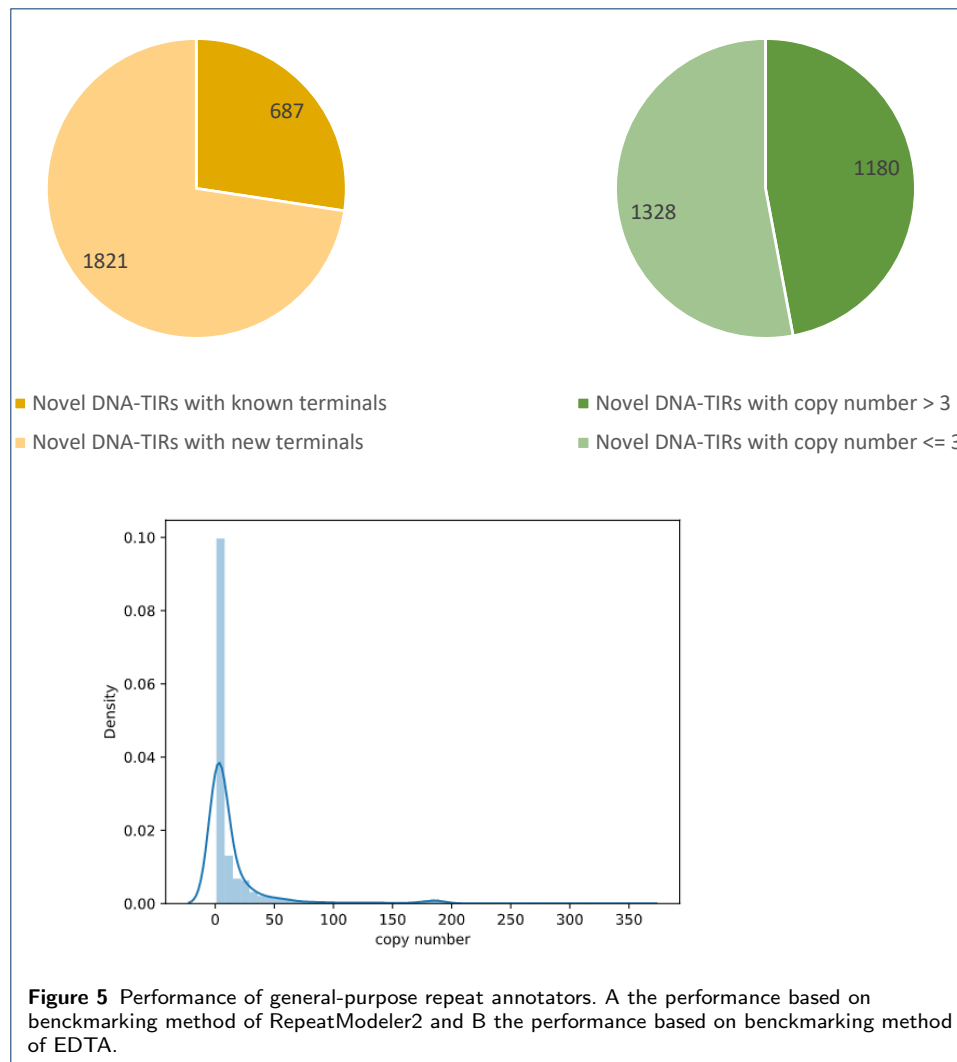
Table 10 Details of performance among general-purpose repeat annotators based on *D. melanogaster*.

TE class	Tools	EDTA evaluation						RepeatModeler2 evaluation			
		Sens	Spec	Accu	Prec	FDR	F1	Perfect	Good	Present	Not_found
TIR	EDTA-TIR	0.8274	0.6989	0.7253	0.4149	0.5851	0.5526	165	112	320	877
	HiTE-TIR	0.8607	0.9591	0.9391	0.8431	0.1569	0.8518	374	39	217	844
Helitron	EAHelitron	0.2666	0.9983	0.9135	0.9536	0.0464	0.4167	0	0	0	310
	EDTA-HelitronScanner	0.8930	0.6259	0.6509	0.1982	0.8018	0.3244	4	30	54	222
	HiTE-Helitron	0.7040	0.9703	0.9438	0.7240	0.2761	0.7138	35	14	21	240
LTR	LTRharvest	0.9420	0.8269	0.8548	0.6356	0.3644	0.7590	395	56	105	549
	LTR_FINDER	0.9676	0.8601	0.8860	0.6874	0.3126	0.8038	493	68	117	427
	LTR_retriever	0.9604	0.9456	0.9491	0.8491	0.1509	0.9013	417	46	172	470

TE, and there are no other full-length copies of these sequences in the genome. Our method needs at least two full-length copies to determine whether a sequence is a true transposon, so we have left out most of the single-copy and zero-copy sequences. (ii) Some transposons do not have consistent TSD or even any TSD. To achieve high-precision identification, we identify LTR and TIR TEs by TSD, so those TEs that do not have consistent TSD are filtered out. Highly divergent terminal inverted sequences (identity less than 0.7) and the candidate TEs with accidental sequence homology outside the boundary, which is similar to many false positive patterns, are also filtered out. By manually checking the results of FMEA, we do find many lost real TIRs. These TIRs are filtered out for various reasons, such as the lack of a consistent TSD and the big divergence in the first 5-bp of the TIRs. This further proves the effectiveness of the FMEA method. At the same time, a more accurate and comprehensive filtering method helps to find more real TIRs.

We have observed that some new TIR elements have been found, which differ significantly from those in Repbase and are distinguished by the 80% principle[6]. Through careful inspection, we found that these new TIR elements have a complete TIR and TSD structure, and the boundaries between their copies are clear. Notably, most of them have low copy numbers (Fig. 5). At the same time, nearly half of the sequences in the new DNA TIRs have more than 3 copies (Fig. 5), suggesting that

these are like real TEs that were not included in the Repbase library due to their low number of copies. In addition, we recognize that some TIR TEs have TIRs similar to the known TIRs in Repbase (Fig. 5), which are likely to be non-autonomous TIR TEs.



Comparison of Helitron annotators

Helitrons are a subclass of DNA transposons, which replicate through the rolling circle mechanism. When replicating themselves, only the single strand of DNA is broken, and no TSD is generated, which is different from the other TEs. The Helitron transposon has a 5'-TC...-CTRR-3' conserved structure, where R refers to purine, A or G, and there is a short hairpin structure about 10 bp upstream of the 3' end. Helitrons mostly transition into host AT target sites, resulting in flanking 5'-A and 3'-T nucleus[12]. The weak structural signals of Helitrons make identification of these elements particularly challenging.

As far as we know, there are only two tools, HelitronScanner and EAHelitron, that can produce useful Helitron predictions. HelitronScanner[13] identifies the sequence

patterns in Helitron transposons using the local combinational variable (LCV) algorithm, which produced a large number of candidate sequences, most of which are false positives. For example, 52 MB of raw candidate sequences cover 13.9% of the rice genome, which obviously exceeds the real coverage. EDTA filters the results of HelitronScanner, greatly improving its specificity and accuracy without reducing its sensitivity[3]. Nevertheless, the precision of the Helitron identification module of EDTA is still very low (Fig), which is far from satisfactory.

We also test the other tool, EAHelitron, which identifies Helitrons based on the conservative structure traits using regular expression (RE), such as the 5' terminal with TC, the 3' terminal with CTAGt, and a GC-rich hairpin loop before 2–10 nt of CTAG. The performance of EAHelitron is primarily determined by the pre-defined patterns of hairpin loop regular expressions. We observed that it lost some of the hairpin loop patterns of real Helitrons. For example, many real Helitrons in *C. briggsae* cannot be discovered until we manually add a new pattern of hairpin loop “[GC]4”. EAHelitron specifies a “-u” parameter to search all possible 5'-TC upstream of CTAGt-3', and it is hard to know the real 5' end of Helitron. We take the first 5'-TC closest to CTAGt-3' as the 5' end of candidate Helitrons, which leads to extremely short sequences with only 87 bp average length and 44 candidates in rice. The short candidate sequences produce the highest precision but the lowest sensitivity (Fig. 6B). Moreover, it cannot identify any gold standard models according to the BM_RM2 (Fig. 6A).

To discover the intact Helitron elements, we have developed a new Helitron identification method, which is a further usage of the coarse boundary TE candidates output by the FMEA algorithm. EAHelitron is used to locate the accurate 3'-CTRR and the hairpin loop structure in candidate TE sequences. The 5'-TC closest to the coarse boundary is selected as the true end. To control the false discovery of the candidate Helitrons, we filtered out the candidates that were not inserted into AT target sites. Finally, the TE copy-based filtering method for region homology outside the boundaries is used to obtain confident candidates (see the “Methods” section).

The experimental results show that our Helitron identification method has the highest performance (Fig. 6), which is superior to EDTA with significantly higher precision, specificity, and accuracy. Compared with the pure EAHelitron method, we have greatly improved the sensitivity and F1 value. We infer that our sensitivity would be greatly improved once EAHelitron can include a more comprehensive hairpin loop pattern. At the same time, we identify more perfect Helitrons in the gold standard dataset. However, we do notice that our results are still affected by false positives, which indicates that our method has potential for improvement.

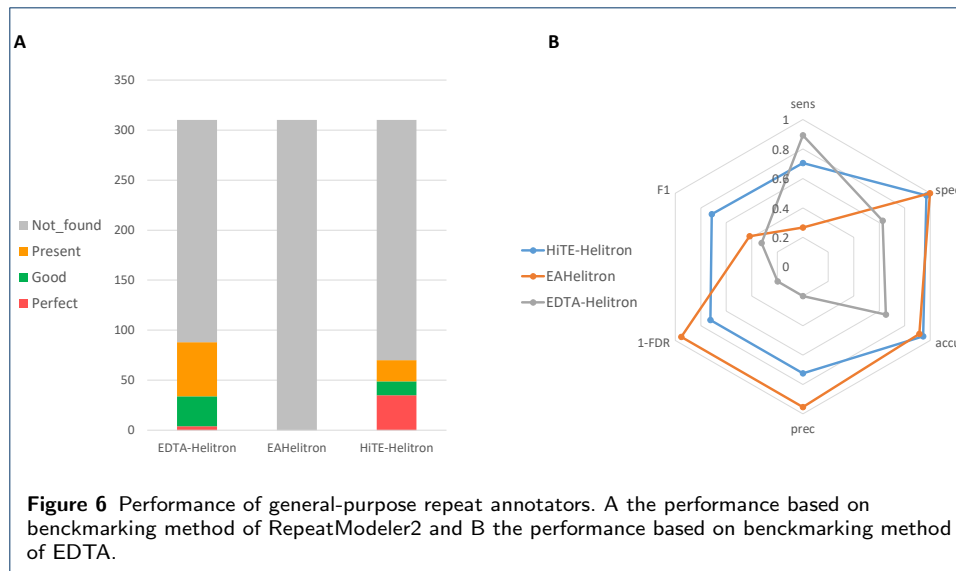
Comparison of LTR annotators

xxx

Sub-sub heading for section

Text for this sub-sub-heading...

Sub-sub-sub heading for section Text for this sub-sub-sub-heading...



Discussion

Text and results for this section, as per the individual journal's instructions for authors.

Conclusions

Text and results for this section, as per the individual journal's instructions for authors.

Methods

Text and results for this section, as per the individual journal's instructions for authors.

In this section we examine the growth rate of the mean of Z_0 , Z_1 and Z_2 . In addition, we examine a common modeling assumption and note the importance of considering the tails of the extinction time T_x in studies of escape dynamics. We will first consider the expected resistant population at vT_x for some $v > 0$, (and temporarily assume $\alpha = 0$)

$$E[Z_1(vT_x)] = \int_0^{v \wedge 1} Z_0(uT_x) \exp(\lambda_1) du.$$

If we assume that sensitive cells follow a deterministic decay $Z_0(t) = xe^{\lambda_0 t}$ and approximate their extinction time as $T_x \approx -\frac{1}{\lambda_0} \log x$, then we can heuristically estimate the expected value as

$$\begin{aligned} E[Z_1(vT_x)] \\ = \frac{\mu}{r} \log x \int_0^{v \wedge 1} x^{1-u} x^{(\lambda_1/r)(v-u)} du. \end{aligned} \quad (1)$$

Supplementary information

Supplementary information accompanies this paper at

Acknowledgements

This work was carried out in part using computing resources at the High Performance Computing Center of Central South University.

Funding

This work has been supported by the National Natural Science Foundation of China under Grant: No.61772557 and No.62002388, Hunan Provincial Science and technology Program (No.2018wk4001), 111 Project (No.B18059), Fundamental Research Funds for the Central Universities of Central South University (2021zzts0208).

Abbreviations

Text for this section. . .

Availability of data and materials

Text for this section. . .

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Authors' contributions

Text for this section . . .

Authors' information

Text for this section. . .

Author details

Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha, 410083, China.

References

1. Storer, J.M., Hubley, R., Rosen, J., Smit, A.F.: Methodologies for the de novo discovery of transposable element families. *Genes* **13**(4), 709 (2022)
2. Storer, J., Hubley, R., Rosen, J., Wheeler, T.J., Smit, A.F.: The dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* **12**(1), 1–14 (2021)
3. Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J.R., Hellinga, A.J., Lugo, C.S.B., Elliott, T.A., Ware, D., Peterson, T., *et al.*: Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome biology* **20**(1), 1–18 (2019)
4. Flynn, J.M., Hubley, R., Goubert, C., Rosen, J., Clark, A.G., Feschotte, C., Smit, A.F.: Repeatmodeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences* **117**(17), 9451–9457 (2020)
5. Bao, W., Kojima, K.K., Kohany, O.: Repbase update, a database of repetitive elements in eukaryotic genomes. *Mobile Dna* **6**(1), 1–6 (2015)
6. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capi, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., *et al.*: A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics* **8**(12), 973–982 (2007)
7. Su, W., Gu, X., Peterson, T.: Tir-learner, a new ensemble method for tir transposable element annotation, provides evidence for abundant new transposable elements in the maize genome. *Molecular plant* **12**(3), 447–460 (2019)
8. Kempken, F., Windhofer, F.: The hat family: a versatile transposon group common to plants, fungi, animals, and man. *Chromosoma* **110**(1), 1–9 (2001)
9. Warburton, P.E., Giordano, J., Cheung, F., Gelfand, Y., Benson, G.: Inverted repeat structure of the human genome: the x-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome research* **14**(10a), 1861–1869 (2004)
10. Gremme, G., Steinbiss, S., Kurtz, S.: Genometools: a comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM transactions on computational biology and bioinformatics* **10**(3), 645–656 (2013)
11. Shi, J., Liang, C.: Generic repeat finder: a high-sensitivity tool for genome-wide de novo repeat detection. *Plant physiology* **180**(4), 1803–1815 (2019)
12. Kapitonov, V.V., Jurka, J.: Helitrons on a roll: eukaryotic rolling-circle transposons. *TRENDS in Genetics* **23**(10), 521–529 (2007)
13. Xiong, W., He, L., Lai, J., Dooner, H.K., Du, C.: Helitronscanner uncovers a large overlooked cache of helitron transposons in many plant genomes. *Proceedings of the National Academy of Sciences* **111**(28), 10263–10268 (2014)

Figures

Figure 7 Sample figure title

Figure 8 Sample figure title

Tables

Table 11 Sample table title. This is where the description of the table should go

	B1	B2	B3
A1	0.1	0.2	0.3
A2
A3

Additional Files

Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.