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

Since being discovered in maize by Barbara McClintock in 1947 [1, 2], transposable elements (TEs), consisting of the major parts of repetitive regions in genomes, have been detected in most eukaryotic species[3, 4]. As mutagens and major contributors to the organization, rearrangement, and regulation of the genome, TEs have been proven to be the major drivers of genome evolution and intraspecific genomic diversity[5, 6, 7].

TEs are generally divided into two classes based on the transposition intermediates (RNA or DNA)[8], and further split into families and subfamilies on the basis of various structural features[9]. Transposing by a TE-encoded reverse transcriptase (RT), Class I TEs are also called as retrotransposons with a "copy-and-paste" transposition mechanism. Based on whether flanked with long terminal repeats or not, they can be further divided into LTRs as well as non-LTRs, which include long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Class II transposable elements are known as DNA transposons with a "cut-and-paste" transposition mechanism, mainly including three major types: (i) TIR elements, which are flanked with terminal inverted repeats (TIRs) of variable length and further divided into nine known superfamilies by the TIR sequences and the TSD size[9]; (ii) Helitrons, which lack TIR features but have conserved 5'-TC and CTRR-3' termini with a short hairpin structure lying a few nucleotides before the 3' end. [10]; (iii) Mavericks, which are large transposons (often 15–40 kbp in size) with long TIRs (several hundred base pairs) and conserved 5'-AG and TC-3' termini [11]. TEs inserted into the integration site of the host genome are usually accompanied by staggered double strand breaks, and the repair of them results in the generation of two short target site duplications (TSDs; usually 2–11 bp)[12].

Over the past decade, high-throughput sequencing technology has made it possible to sequence more large and complex eukaryotic genomes. Long-read sequencing technologies, which can cross highly repetitive regions, are improving the quality of genome assemblies[13]. Faced with the rapid emergence of large quantities of

sequence data as well as the abundance and diversity of TEs, identifying and annotating TEs presents a major challenge, which is driving the need for improved unsupervised annotation of TEs.

Many complications make the identification of TEs not straightforward, including: (i) TEs are degenerating at different speeds since each TE is faced with mutations, which may cause the structural signals of TEs to perish. (ii) The high divergence level between TE instances requires sensitive alignment, making the process impractically slow. (iii) Older TE instances tend to be highly fragmented, which makes it hard to find the true ends of the TE. (iv) The abundance of fragments is much higher than that of full-length TE instances, which hinders the construction of full-length TE models. (v) Regional homology may exist between unrelated TEs, complicating the definition of the true ends of TEs and their classification. (vi) Higher-copy number segmental duplications or large tandem repeats may be falsely regarded as putative TE families[14].

There are a number of tools designed to automate TE identification and/or annotation, which can be divided into three categories:

(i) De novo methods. By identifying exact or closely matching repetitions, de novo methods can identify novel TE instances that do not belong to a known family of TE, which mainly includes a (spaced) k-mer based or self-comparison approach. K-mer-based approaches, such as RepeatScout [15] and P-Clouds[15], are better suited to dealing with young TEs with plenty of copies. For older TEs with large diversity or more complex patterns, such methods tend to generate highly fragmented sequences. Grouper[16], RECON[17], and PILER[18] are examples of self-comparison approaches that require computationally intensive and sensitive alignments with accurate clustering methods to cluster these alignments into "piles" and generate the TE family. Compared with the k-mer based method, these methods can find more sophisticated TE families. However, the high fragmentation and mosaicism present in TE families make accurate clustering of these alignments challenging[14].

(ii) Signature-based methods. Purely de novo methods, which detect TEs by sequence repetition alone, may miss low-copy but well-characterized TEs. At the same time, it is inevitable that they will include non-TE sequences, such as processed pseudogenes and high-copy gene families. Instead, signature-based methods identify TE instances by recognizing features of specific families of TEs, including terminal inverted repeats, direct repeats, conserved terminal motifs, TSDs, etc. For example, LTRharvest[19], LTR_retriever[20], Generic Repeat Finder[21], EAHelitron[22], HelitronScanner[23], and MITEHunter. Unfortunately, signature-based methods always suffer from false positives due to the weak structural characteristics of many TEs.

(iii) TE discovery pipeline. A TE discovery pipeline combines different TE identification tools to comprehensively identify all types of TE within any given genome, such as EDTA[24] and RepeatModeler2[25]. By integrating a variety of tools into a single pipeline, a TE discovery pipeline can overcome the shortcomings of any one particular approach. However, using other tools without any improvements will introduce the inherent defects of these tools. Moreover, the merging of multiple tools requires careful handling of redundant results.

After years of manual curation, Repbase[26] and Dfam[27] are high-quality consensus libraries for a limited set of species, while all automatically generated TE

libraries required extensive manual editing. To generate high-quality TE libraries, we develop an automated TE annotation pipeline called High-precision TE Annotator (HiTE) that produces a high-quality, structurally intact, non-redundant, and classified TE library. It mainly includes four steps: (i) filtering candidate repeat regions within the genome based on the low-frequency k-mer masking method; (ii) identifying the coarse boundaries of TEs based on the fault-tolerant mapping expansion algorithm; (iii) using signature-based methods to accurately determine the boundaries of TEs and filtering out non-intact TE elements, such as fragments, segment duplications, tandem repeats, and nested TEs; (iv) filtering false positive sequences with repetitive flanking sequences within copies, which are parts of a larger repetitive element. HiTE can not only discover novel TE families but also accurately identify structurally intact TE models using highly conservative structural features and copy support. At the same time, the accurate determination of the boundaries reduces a large amount of manual repair. By benchmarking four different kinds of model species, we have proved that HiTE can restore more gold standard TE consensus sequences and produce a higher quality TE library than the existing tools.

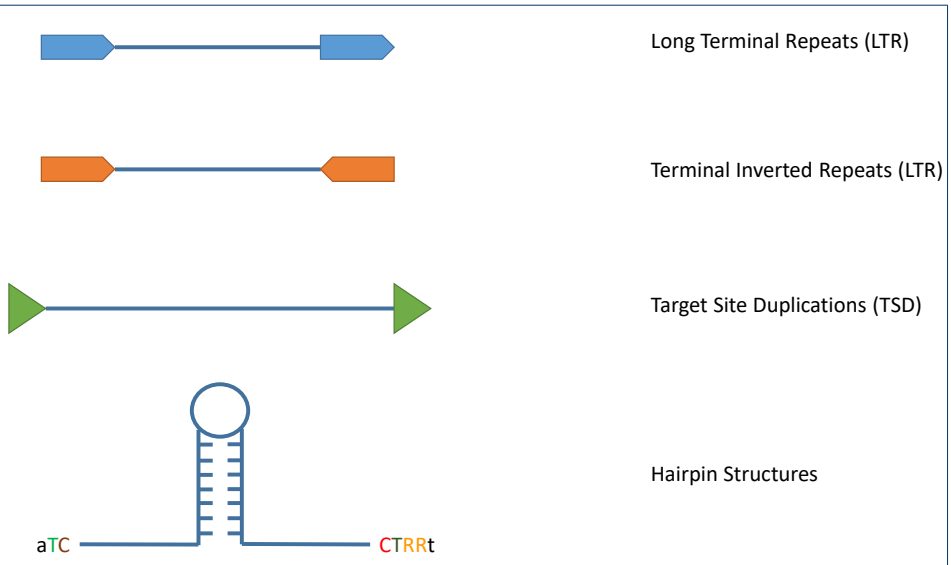

Figure 1 Examples of commonly used TE structural features for detection. The LTR, TIR, TSD, and hairpin structures are used to identify LTR elements, class II elements, and Helitrons, respectively.

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 [14], 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[28]. 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[14].

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[24] and RepeatModeler2[25], 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[26]. 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[24]. 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[24]. 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"[25]. 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.

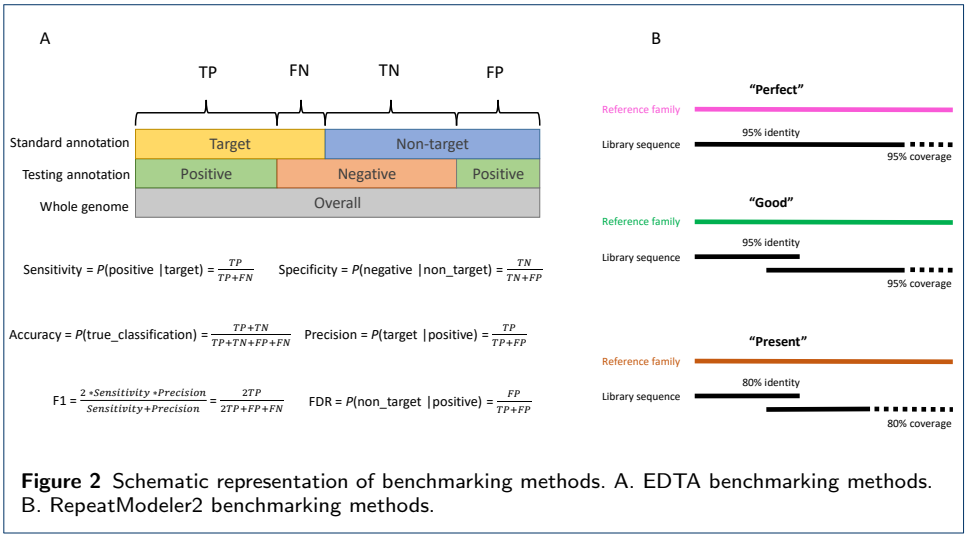

Figure 2 Schematic representation of benchmarking methods. A. EDTA benchmarking methods. B. RepeatModeler2 benchmarking methods.

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[14]. 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[25].

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 Discussion**.

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[24]. 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.

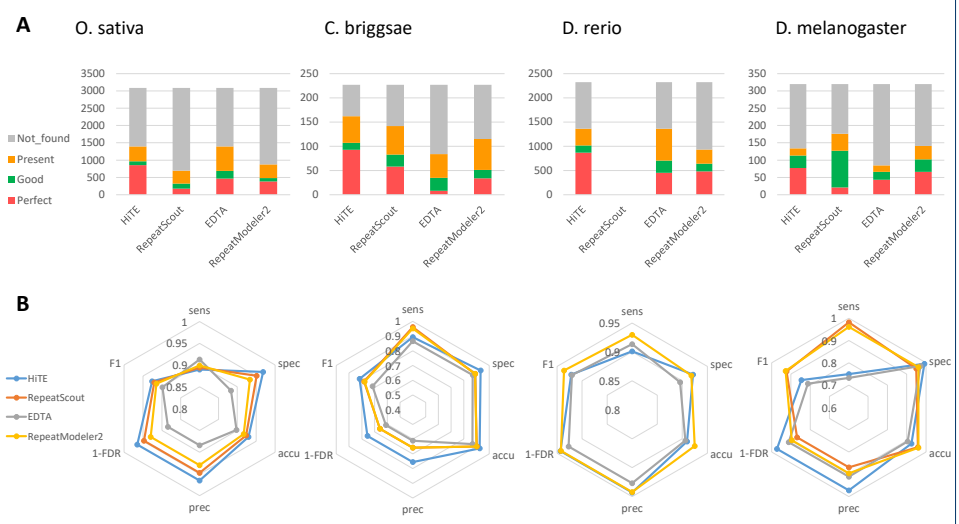

Figure 3 Performance of general-purpose repeat annotators. A the performance based on benckmarking method of RepeatModeler2 and B the performance based on benckmarking method of EDTA.

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[9]. 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[11]. 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[29].

Many tools have been designed for their identification, such as IRF[30], TIRvish[31], TIR-Learner[11], and GRF[21], 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[24]. 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.

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[9]. 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

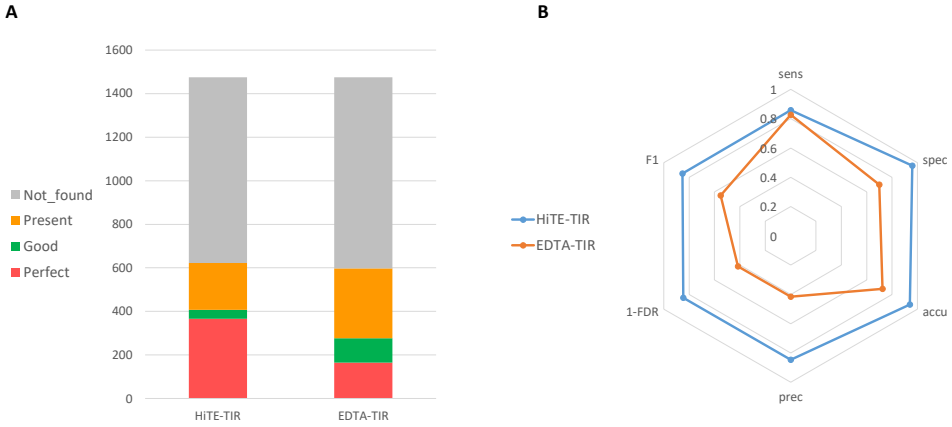

Figure 4 Performance of TIR annotators. A the performance based on benckmarking method of RepeatModeler2 and B the performance based on benckmarking method of EDTA.

Table 10 Details of performance among all types of TE annotators based on O. sativa.

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
Non-LTR	Non_LTR_library	0.7302	0.9879	0.9819	0.5883	0.4117	0.6516	77	3	21	48
	HiTE-Non-LTR	0.6519	0.9999	0.9916	0.9906	0.0094	0.7863	24	0	21	104

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[10]. The weak structural signals of Helitrons make identification of these elements particularly challenging.

To date, there are only two tools, HelitronScanner and EAHelitron, that can produce useful Helitron predictions. HelitronScanner 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

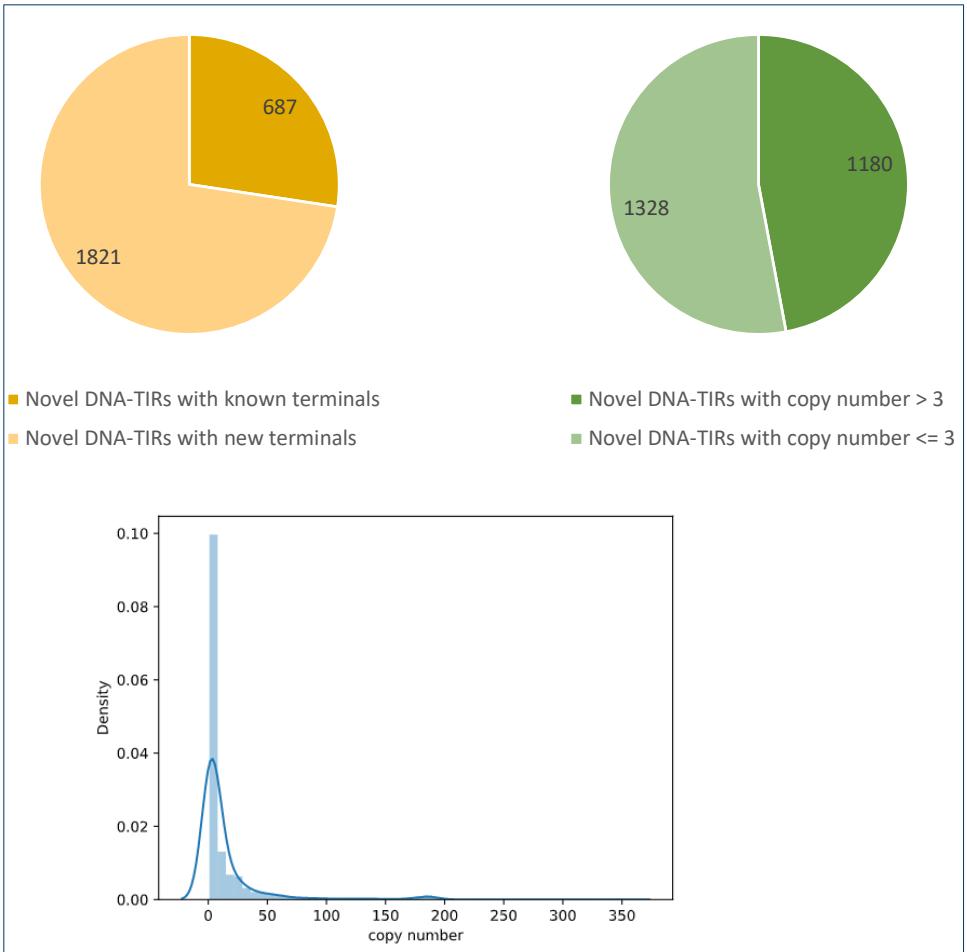

Figure 5 Characteristic of novel TIR elements identified by HiTE. A the number of novel TIR elements with different types of terminals; B the ratio of copy numbers > 3 and <= 3; and C the distribution of copy numbers of TIR elements identified by HiTE.

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[24]. 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 predefined 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 haripin 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 candi-

dates 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.

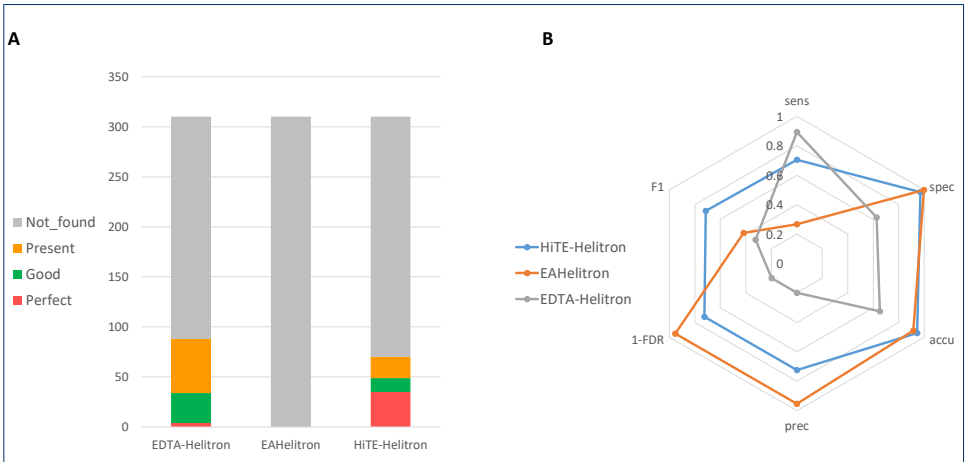

Figure 6 Performance of Helitron annotators. A the performance based on benckmarking method of RepeatModeler2 and B the performance based on benckmarking method of EDTA.

Comparison of LTR annotators

Long terminal repeat retrotransposons (LTR-RTs) (Fig. 7) have a well-conserved structure and are prevalent in plant genomes. There are many tools dedicated to the de novo identification of LTR-RTs, including MGEScan3[32], GRF, LTR_STRUC[33], LTR_FINDER[34], LTRharvest[19], LtrDetector[35], and LTR_retriever[20]. It is worth noting that LTR_retriever was designed as a stringent filtering method for raw results from other LTR tools and does not have its own search engine. We benchmarked the three best existing LTR de novo identification tools, LTR_FINDER, LTRharvest and LTR_retriever (using the output of LTR_FINDER and LTRharvest as input), and found that LTR_FINDER and

LTRharvest achieve higher sensitivity but lower precision, whereas LTR_retriever significantly improves the precision while maintaining the same sensitivity. The LTR_retriever was integrated into a variety of TE detection pipelines, including EDTA and RepeatModeler2, and greatly improved the accuracy of their LTR identification. Although LTR_retriever loses some perfect models, it is still the best LTR identification method at present. Therefore, we integrated LTR_retriever into HiTE.

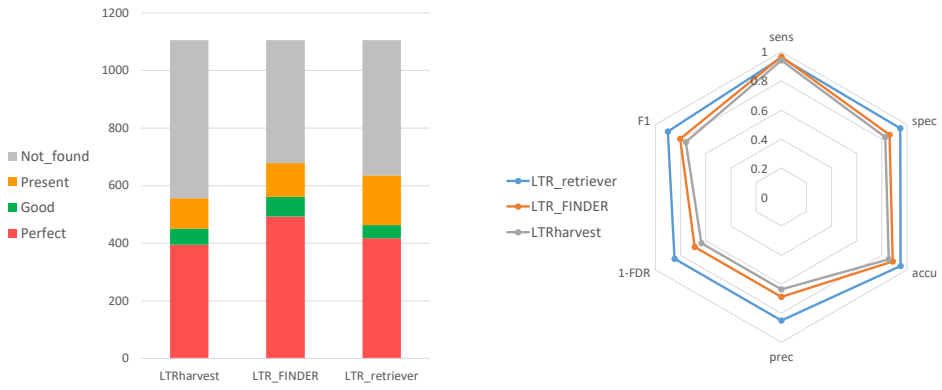

Figure 7 Performance of LTR annotators. A the performance based on benckmarking method of RepeatModeler2 and B the performance based on benckmarking method of EDTA.

Comparison of Non-LTR annotators

Non-LTR retrotransposons include two types of TE: LINEs and SINEs[36]. LINEs, which lack LTRs flanking both ends, can reach several kilobases in length. Although the presence of RT and nuclease in the pol ORF of LINEs seems to provide a confident basis for their identification, there is not a database dedicated to their curation. Worsely, the truncated 5' ends, resulting from the premature termination of reverse transcription, make them difficult to discover. SINEs, on the other hand, are much shorter (80–500 bp)[9]. They do not encode any reverse transcriptase protein and rely on other TEs to transition, especially LINEs[37]. The weak signals of non-LTR retrotransposons make them quite challenging to identify[38].

To accurately identify non-LTR retrotransposons, we have developed a homology-based TE searching module, named HiTE-Non-LTR. HiTE-Non-LTR extracts LINEs and SINES consensus sequences from the Dfam library to form a non-LTR library, which is then used to search for confident candidate sequences based on the coast boundary TE candidates output by the FMEA algorithm. To benchmark the performance of the homology-based TE searching module, we use the non-LTR library to search confident candidates in the assembly based on the same parameter as the competing evaluation, called Assembly-Non-LTR. Although HiTE-Non-LTR sacrificed a little sensitivity, it achieved nearly 100% precision.

Influence of parameter changes on results

To understand how the parameters in HiTE affect the results, we selected the four most important parameters for testing: k_num, freq_threshold, chunk_size, and flank-

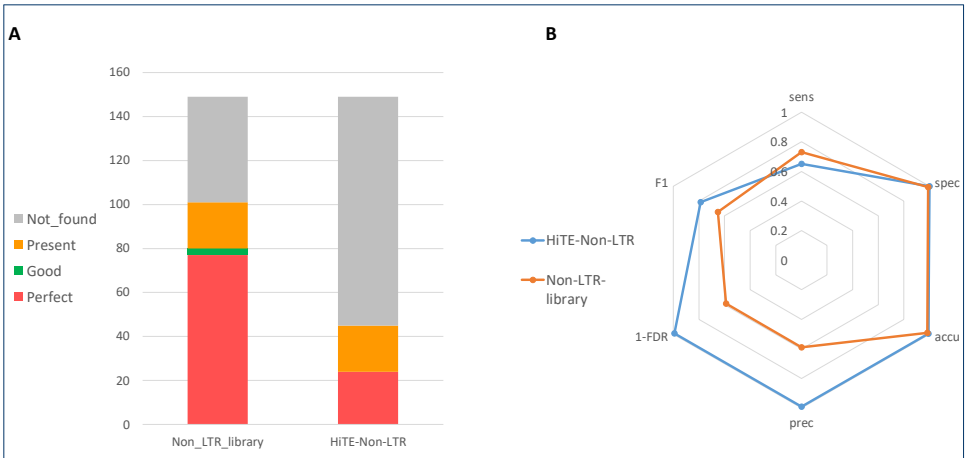

Figure 8 Performance of Non-LTR annotators. A the performance based on benckmarking method of RepeatModeler2 and B the performance based on benckmarking method of EDTA.

ing_len. The k_num is the size of k-mer, the freq_threshold refers to the frequency threshold of k-mer, the chunk_size refers to cutting the genome into blocks of the same size, and the flanking_len is used to extend the candidate TEs identified by FMEA to search the valid TSD. These parameters have no effect on the results of LTR elements, which are discovered by LTR_retriever. Therefore, we chose C. briggsae as the test species, whose genome only contains a small number of LTR elements.

As shown in Fig. 9A, the smallest k_num (such as 11) will mark the most parts of the genome as repeat regions, which cannot effectively distinguish TE from non-TE, resulting in low sensitivity and precision. Large k_num will lose part of the true TE (lower sensitivity), but the sequences it identifies are more likely to be true TE (higher precision). Moderate k_num (such as 31) achieves a balance between sensitivity and precision, the highest F1 value. When k_num exceeds 41, we observe a significant drop in the number of perfect models. As shown in Fig. 9B, with the freq_threshold increased, all metrics except precision decreased significantly, which indicates that the higher the frequency of k-mer in the sequence, the more likely the sequence is to be a real TE. As shown in Fig. 9C, genome slicing will result in the loss of some low copy and scattered TE, reducing the sensitivity of the results significantly. The smaller the cut, the more TE will be lost. As shown in Fig. 9D, when flanking_len is set to 0, the number of sensitivities and perfect models is very low, which indicates that most of the TE we identify in the FMEA algorithm have coarse boundaries. The real boundary of most TE can already be included when flanking_len is set to 10, indicating that the error between the rough boundary in FMEA and the real boundary is not large. The metrics tend to be stable after flanking_len is set to 40.

Contribution of TEs to genome size

The amplification or contraction of transposable elements, affected by environmental stressors, is closely related to the genome size[39, 40]. LTR retrotransposons, especially the Ty3-gypsy elements, which are the major component in most plants, play an important role in the genome size variation across the Oryza genus[41].

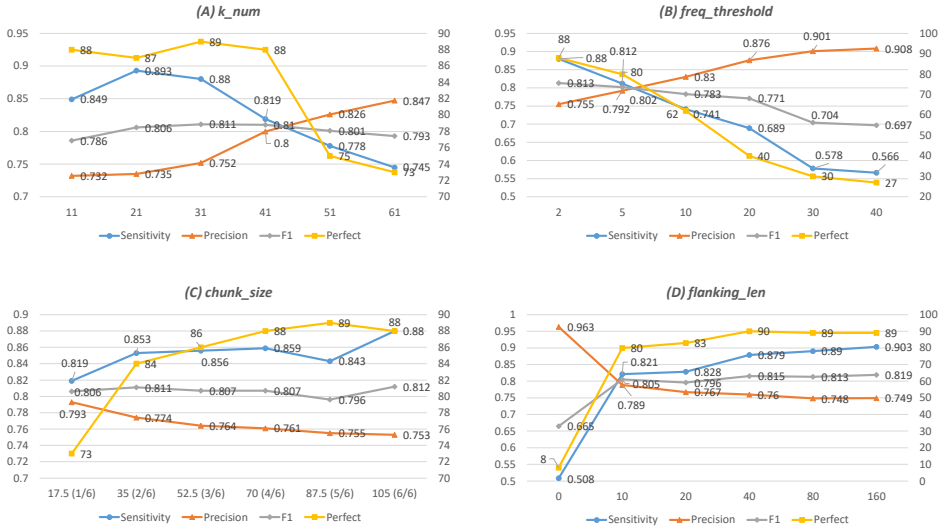

Figure 9 Influence of parameter changes on results. A influence of k_num parameter on results; B influence of freq_threshold parameter on results; C influence of chunk_size parameter on results; and D influence of flanking_len parameter on results.

By applying HiTE to several common rice subspecies, Oryza sativa, Oryza rufipogon, and Oryza glaberrima, we observed that there was significant genome size variation among these rice genus, and the main source of genome size difference is the Gypsy transposon (Fig. 10), as previously documented. The TE libraries of rice genomes are generated by HiTE using the default parameters. RepeatMasker is then used to generate the length coveage based on these TE libraries.

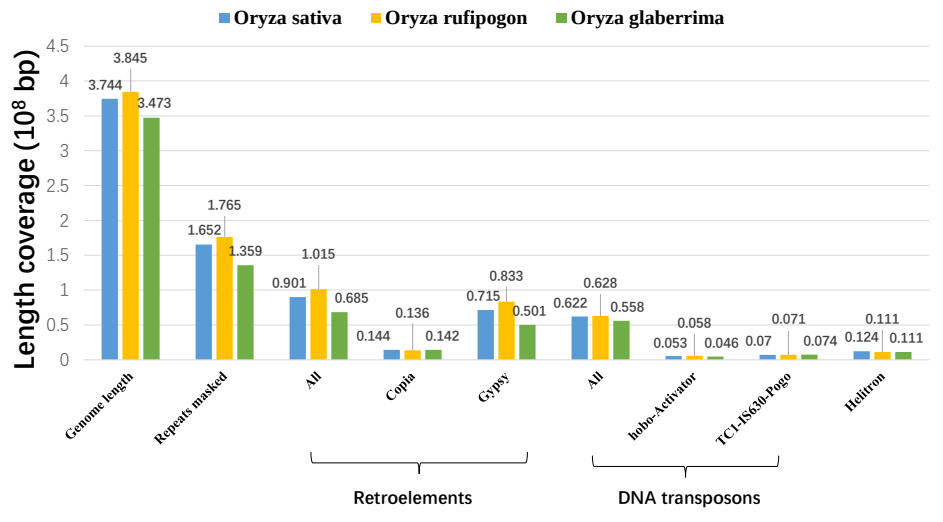

Figure 10 The length coverage distribution of different types of transposons based on the analysis of rice genus.

Discussion

Thanks to decades of manual annotation results, we have obtained a highly reliable TE library for a limited number of species. With the development of third-generation (long-read) sequencing technology, repetitive regions in the genome can be crossed, greatly improving the quality of genome assembly. While quantities of high-quality genome assemblies are being generated, an automated and high-precision TE annotation tool is urgently needed for these newly assembled genomes. To solve this problem, we have developed an ensemble method for high-precision transposable element annotation, known as HiTE, which has performed extensive benchmarking on four model species and achieved higher metrics and restored more perfect gold standard sequences compared with other tools.

The identification of TEs requires intensive and sensitive sequence alignments, which is a computationally demanding task. HiTE uses k-mer coverage to reduce computation. Unlike the traditional k-mer-based seed expansion method, RepeatScout, HiTE uses low-frequency k-mer to determine candidate repeat areas, which reduces the number of sequence alignments and speeds up subsequent computation.

The TE-derived sequences in the genome accumulate variations over time, making their discovery and characterization challenging for the TE annotation methods. As time goes by, TEs are often accompanied by a large number of deletion and insertion variations when replicating and copying themselves. At the same time, their insertion sites on the genome are usually random, leading to complex sequence patterns of TE in the genome, such as nested TE structures, making accurate TE identification and annotation extremely difficult. It is easy for a complete TE sequence to generate multi-segment alignment due to the influence of divergence and nested TE during its evolution. The pairwise alignment-based identification methods, such as RECON, may identify a complete TE model as multiple pieces without edges connected and generate multiple TE models using the single linkage clustering algorithm. We have designed an alignment expansion method with fault tolerance that can easily cross the large gaps caused by insertion, deletion, and nested TE and retain the complete TE structure as much as possible.

Although it is important to accurately identify the structures and boundaries of TEs, repeatedness-based methods, such as RepeatModeler[42], always obtain uncertain boundaries, and intensive manual repairs are required to enable them to be saved in the cured library[14]. HiTE first used the sensitive sequence alignment information to determine the coarse boundaries of TEs based on the fault-tolerant alignment expansion method. Then, the coarse boundaries are flanked to search for valid TSD and terminal motifs. Finally, a reliable false-positive filtering method is developed to get confident TEs with multiple intact copies and clear TE boundaries.

Although HiTE can achieve high-precision TE identification and annotation, 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 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. We discover some lost real TIRs by manually reviewing FMEA results. 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 found that the identification of TEs with weak structural characteristics, such as Helitron and non-LTR elements, is very challenging. Although we have greatly improved the identification performance of Helitron, there is still potential for improvement. For example, a more comprehensive hairpin loop pattern will significantly improve the sensitivity.

To date, due to the truncated 5' ends of LINEs, there is no method to identify LINEs based on the structure method. A few tools designed for identification of SINEs, which suffer from the high false positives and low sensitivity. To achieve high-precision non-LTR element annotation, we developed a homology-based TE searching method, which improves precision by nearly 100%. However, we do lose some true non-LTR elements, and the structure-based identification methods of LINEs are needed, which is also the direction of our future efforts.

Conclusions

The rapid development of sequencing technology enables us to obtain a more reliable genome assembly. The TE library generated by an inaccurate TE identification tool will contain many errors, which will be propagated during the whole-genome annotation process. HiTE makes full use of the strengths and weaknesses of existing methods, including ensemble methods of many types, and can comprehensively and accurately identify and annotate TEs in assembly. By benchmarking on four model species with different TE landscapes, we prove that HiTE can achieve higher accuracy and restore more perfect gold standard TE models, which can be fully applied to any new sequencing genome assembly.

Methods

HiTE Overview

HiTE is an automated TE annotation pipeline that aims to produce a high-quality, structurally intact, non-redundant TE library. Purely de novo methods, which detect TEs by sequence repetition alone, may miss low-copy but well-characterized TEs. At the same time, it is inevitable that they will include non-TE sequences, such as processed pseudogenes and high-copy gene families. Signature-based methods identify TE instances by recognizing features of specific families of TEs, which

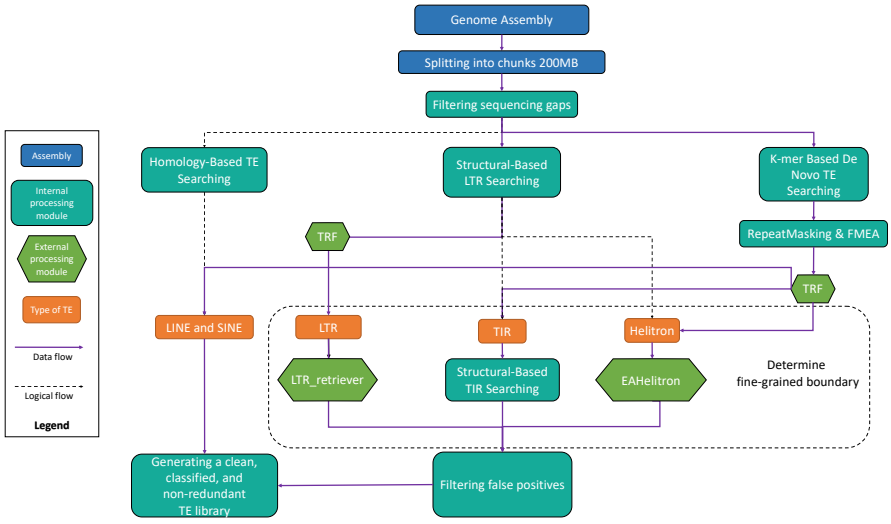

Figure 11 The workflow of HiTE.

are less susceptible to these particular problems. Unfortunately, signature-based methods always suffer from false positives due to the weak structural characteristics of many types of TEs. Both purely de novo and signature-based methods have their own defects. By integrating a variety of existing tools into a single pipeline, a TE discovery pipeline can overcome the shortcomings of any one particular approach. However, these existing tools have different defects, such as fragmentation, false-positive sequences, chimaeras of TEs, and so on. Using these tools without improvement will introduce inherent errors that will propagate to the whole genome annotation.

We employ different strategies to handle the different structural characteristics and distribution of TEs in the genome, and three modules, k-mer-based de novo TE searching, structural-based LTR searching, and homology-based TE searching, are used to identify almost all types of transposons, including LTRs, TIRs, Helitrons, LINEs, and SINES.

Due to intraelement recombination and mutations, intact LTR-RTs contribute only a small fraction of all LTR-related sequences in a genome[20]. In addition, long insertions are also more likely to be selectively disadvantageous to the genome, and full-length LTR elements are often reduced to solo LTRs via LTR–LTR recombination[14]. Therefore, to identify reliable LTR-RTs, we use the mature tools LTR_FINDER[34] and LTRharvest[19] to find all candidate LTR candidates and then use the LTR_retriever[20] as a stringent filtering method for the raw results by identifying LTR-specific signals, which has been proven to be the state of art in LTR identification.

TIR and Helitron elements have weaker structure signals, which makes it easy to generate a large number of false-positive sequences. For example, TIR elements may have a short terminal inverted repeat (5–27 bp of hAT superfamily) and a target site duplication (TSD) structure. Helitrons are defined by 5′-TC and CTRR-3′ motifs (where R is a purine) and a short hairpin structure lying a few nucleotides

before the 3′ end. Therefore, we have developed a reliable identification method of TIR elements, shown in Fig. 11, which mainly includes four steps: (i) filtering candidate repeat regions within the genome based on the low-frequency k-mer masking method; (ii) identifying the coarse boundary of TE based on the fault-tolerant mapping expansion algorithm; (iii) using signature-based methods to accurately define the boundary of TE and filtering out not intact TE elements, such as segment duplication, tandem repeats, and nested TE; (iv) filtering false positive sequences with repetitive flanking sequences, which are parts of a larger repetitive element.

Compared with existing tools, HiTE has the following four innovations: (i) Use repeated k-mer coverage to reduce the amount of computation. Unlike the traditional recognition tool, RepeatScout, based on the k-mer seed expansion method, HiTE uses low-frequency repeated k-mer to determine candidate repeat areas, which can reduce the subsequent calculation amount at a faster speed. The alignment-based identification method requires pairwise alignment of the whole genome, while HiTE only needs to compare candidate repeats, which saves a lot of computing resources. (ii) A fault-tolerant mapping expansion algorithm is designed to restore an intact TE. Highly fragmented sequences are often generated due to the divergences and nested TEs, which result in the multi-segment alignment of a complete TE. Pairwise alignment-based identification methods, such as RECON, use the single linkage clustering algorithm to generate TE sequences based on overlapping subsequences. It is possible to identify the same TE as multiple "piles" without edges connected, and multiple TE models are generated, resulting in a large number of fragments. We have designed an alignment expansion method with fault tolerance that can well solve the impact of sequence inconsistency caused by insertion, deletion, and nested TE while retaining the complete TE structure as much as possible. (iii) Accurately define the TE boundary. The TE library generated by automatic identification methods still needs a lot of manual identification and repair[14], mainly due to the inability to find the true boundaries of TEs. HiTE first used the self-alignment information to determine the coarse boundary of TE. Then, HiTE flanks both ends and searches for TSD to accurately find the boundaries of TEs, which can greatly reduce the cost of manual identification and repair in the later stage. (iv) highly reliable filtration method. Weak structural characteristics of many TEs caused a flood of false positive sequences, especially for DNA-TIR type transposons (short terminal structures). We have designed a strict filtration method based on the following truth: we believe that after the specific boundary is determined, the region near the copy of the true TE should be close to the random sequence. Therefore, more than half of the copies have homology outside the boundary area, indicating that these copies belong to a larger repeat, which is considered a false positive and should be filtered.

Like the traditional de novo method, HiTE can discover novel TE families. More importantly, it can accurately identify the structurally intact TE families by using highly conservative structural features and copy support. At the same time, the accurate definition of the boundaries reduces a large amount of manual repair.

Kmer-Based De Novo TE Searching

The majority of de novo identification methods, such as RECON, are based on the similarity of pairwise alignment to identify repeats. However, direct pairwise align-

ment of genomes will consume a lot of computing resources. To solve this problem, we designed the RepeatMasking method, which can mark candidate repeats on the genome to reduce the search scope.

The discovery of the raw repeat region is based on the following observation: if there are two repeat sequences, regardless of the variations, the k-mers composed of these two repeat sequences are also repeated. Therefore, we can in turn identify candidate repeat sequences by covering the repeated k-mers in the genome, which is shown in Fig. 12.

It is worth noting that due to the variations between the two repeat sequences, the continuous repeat regions may break into small pieces due to a lack of duplicate k-mers. We use the fault tolerance parameter to skip these small gaps and connect the scattered repeat areas. In addition, the fake k-mers may connect multiple repeat regions together and generate a larger repeat region. We use the fault-tolerant mapping expansion algorithm to distinguish different TEs.

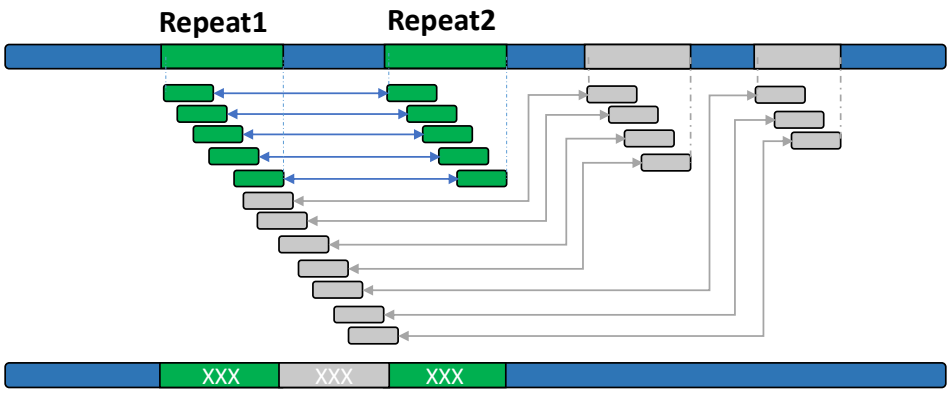

Figure 12 The illustration of RepeatMasking. The long blue bar represents the genome, and the green box on the genome represents two repeat sequences, Repeat1 and Repeat2. The small green box below represents the k-mers of the real repeat sequence (the frequency of occurrence is at least 2 times), and the small grey box represents the confused k-mers, that is, although the frequency is more than 2 times, it does not belong to the real repeat sequence. The area marked with an X below is the candidate repeat area identified by the program.

Algorithm 1 describes the RepeatMasking algorithm, where G is genome assembly, k is the size of the k-mer, L is the length of divided genome segments, and g is the maximum length of the gap between adjacent repeat regions; R is the set of candidate repeat regions; The buildRTable(·) function is used to construct the hash table of repeated k-mers, the cutSegments(·) function is used to divide the whole genome into genome segments, the cutKmers(·) function cuts the genome segments into k-mers, the queryRtable(·) function is used to judge whether k-mers are repeated by querying the repeated k-mers hash table, and the maskSequence(·) is a function to mark repeated sequences, The skipGaps(·) function connects adjacent repeat sequences to skip small gaps, while the extractRepeats(·) function is used to extract candidate repeat sequences from masked sequences.

To analyze the time complexity of RepeatMasking, we noticed that DSK is a highly efficient tool that can process a mammalian genome in a few minutes, so the buildRTable(·) function actually takes very little time. The cutSegments(·) function

divides the whole genome assembly into N/L segments, where N=length(G). Both cutKmers(·) and maskSequence(·) have O (L) time complexity, while queryRtable(·) has O (1) time complexity. The total time complexity of algorithm 1 is O(N·L), which is a function of N and L. In general, L is set as a fixed constant. In addition, since we use multiprocessing technology to accelerate the program, the running time of RepeatMasking can be reduced to t times the original, where t is the number of processes. Therefore, this algorithm has high efficiency in actual application.

Algorithm 1 RepeatMasking

Inputs: G, k, L, g
Outputs: R

function RepeatMasking(G, k, L, g)
  $H \leftarrow buildRTable(G, k)$
  $S \leftarrow cutSegments(G, L)$
  $R \leftarrow \emptyset$
  for $i = 0 \rightarrow length(S)$ **do**
    $s \leftarrow S[i]$
    $P \leftarrow cutKmers(s, k)$
    for $j = 0 \rightarrow length(P)$ **do**
      $p \leftarrow P[j]$
      $isRepeated \leftarrow queryRtable(H, p.kmer)$
      if $isRepeated$ **then**
        $s' \leftarrow maskSequence(s, p.start, p.end)$
      end if
    end for
    $s'' \leftarrow skipGaps(s', g)$
    $r \leftarrow extractRepeats(s'')$
    $R \leftarrow R \cup r$
  end for
end function

Fault-tolerant Mapping Expansion Algorithm

The pairwise alignment method can identify more complete and biologically meaningful TE sequences. At the same time, due to the serious divergence between TE copies and the existence of a large number of insertions and deletions, we must consider fault tolerance when identifying TE. Due to divergence or the existence of nested TE, it is very common for a TE sequence to generate multiple alignments. Traditional self-alignment-based methods may divide a single TE instance into multiple families, resulting in a large number of fragments that negatively affect the identification and classification of complete TE families. Therefore, we designed a fault-tolerant mapping expansion algorithm (FMEA) that can span a large gap.

As shown in Fig. 13, the algorithm first performs self-alignment on the raw RepeatMasking repeats. For each query, adjacent alignments are gathered based on their alignment positions on the subject and then sorted ascending based on query alignment positions. If the next alignment is still in the adjacent area of the previous alignment, expand the previous alignment until it cannot be expanded. Each query will obtain multiple extension sequences, and the redundant sequences are removed. The longest sequence with more than two copies represents an intact repeat. A representative sequence with more than two copies is used to obtain the intact repeat for as long as possible.

Example description

Due to the existence of insertion, deletion, and multiple TE sequences, multiple subsequence alignments will be generated in the candidate repeat area, as shown in

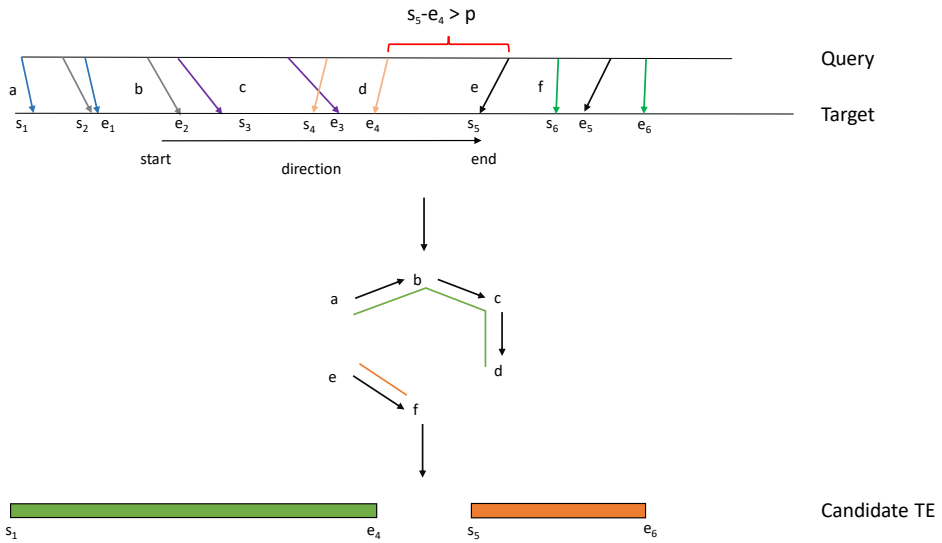

Figure 13 The illustration of the fault-tolerant mapping expansion algorithm.

Fig. 13: a, b, c, d, e, f. The above algorithm can be simply described as the following process:

1. We start by setting an extended threshold value p, then sort the alignments by starting and ending positions.

2. For each alignment, judge whether its adjacent alignment can expand the sequence length. For example, the first is the alignment of subsequence a, whose starting and ending positions are s1 and e1; the starting and ending positions of subsequence b are s2 and e2. Since e2¿e1 and s2-e1¡p, it means that adding b can expand the length of the current subsequence, so we connect the subsequences a and b. Similarly, connect subsequences c and d. However, since s5-e4¿p, it indicates that the subsequence e is too far from sequence d to cross the gap in the middle, which should belong to two different TE instances. The TE instances in the above example are TE sequence 1 (starting s1, terminating e4) composed of subsequences a, b, c and d and TE sequence 2 (starting s5, terminating e6), corresponding to subsequences e and f.

3. Since the query will be aligned to multiple different targets, we will get a set with overlapped sequences. We think that two sequences in the overlapped set have more than 95% overlap, and they are considered to be copies of each other. A representative sequence is used to represent all copies with overlap, and the boundary of the representative sequence is updated to include all copy sequences. Finally, we get a collection of non-overlapping repeats.

Structural-Based TE Searching

TEs have certain structural characteristics, such as LTR and TIR characteristics at both ends of LTR and TIR elements. In addition, when TE is inserted into the genome, it is usually accompanied by DNA double-strand breaks, whose repair results in the formation of two short target site duplications (TSD; usually 2–11 bp) at the integration site. The size of the TSD can be used as a diagnostic feature

for TE identification and classification. Structural-based TE searching methods can discover structurally intact TEs by identifying these TE superfamily-specific structural features. We describe the structural characteristics of three main types of TE and refer to the review for more TE structural characteristics[9].

LTR-RTs typically have long direct repeat sequences (85 to 5000 bp), 2-bp palindromic motifs, 5'-TG..CA-3' at both ends, and a 4-6 bp TSD flanked. The strong structural features of LTRs allow us to identify them directly based on the genome. At present, there are some mature tools that can accurately identify TSD and LTR boundaries, such as LTR_Finder and LTRharvest. We use LTR_harvest and the parallel version of LTR_Finder[43] to identify candidate sequences with LTR structures in the genome. LTR_Finder uses the default parameters, and LTR_harvest uses the parameter "-seed 20 -minlenltr 100 -maxlenltr 7000 -similar 85 -motif TGCA -mintsd 4 -maxtsd 6 -vic 10". Finally, judge whether the candidate LTR sequence is a true transposon (see Section Filtering false positives).

TIR elements have terminal inverted repeat sequences (usually a few bp to hundreds of bp) and conserved motif characteristics of some specific superfamilies. For example, DTC (CACTA) starts and ends with the conserved sequence 5'-CACTA...TAGTG-3'; DTT and DTH transposons have conserved TSDs of "TA" and "TNN", respectively. However, TIR transposons are challenging to identify due to their short terminal structure. Most TIR identification tools still suffer from a large number of false positives. To discover the intact TIRs while reducing false positives, we first use the RepeatMasking and FEMA algorithms to determine the coarse boundaries of candidate TEs. Then, we flank a certain length of the coarse boundaries and enumerate all legal TSDs, as shown in Fig. 14B. To reduce false positives, we identify the consistent TSDs and 5-bp at the ends of TIRs with at most a 1 bp mismatch, respectively. Next, we use the itrsearch tool included in TE Finder 2.30 (a part of the REPET[44] package) with the parameter "-i 0.7 - l 5" to search TIR structures, and the elements with intact TIR and TSD structures closest to the coarse boundaries are chosen as the candidate TIRs. Finally, we will determine whether the candidate TIRs identified are true transposons (see Section Filtering false positives).

Helitron transposon replicates through the rolling circle mechanism. When replicating, only the single strand of DNA is broken, and no TSD is generated. Helitron transposon has a 5'-TC...-CTRR-3' conserved structure (R refers to purine, A or G), and there is a short hairpin structure about 10 bp upstream of the 3' end. All Helitrons previously identified in plants, fungi, words, insights, verticals, and mammals have been characterized by precise transitions between the 5'-A and T-3' into host AT target sites[10]. The weak structural signals of Helitrons make identification of these elements particularly challenging. The identification of Helitrons still based on the candidate TEs with coarse boundaries generated by the RepeatMasking and FEMA algorithms. The difference is that we use EAHelitron[22] to identify candidate sequences with Helitron structure. Finally, the same filtration method with TIR identification is used to filter out false positives.

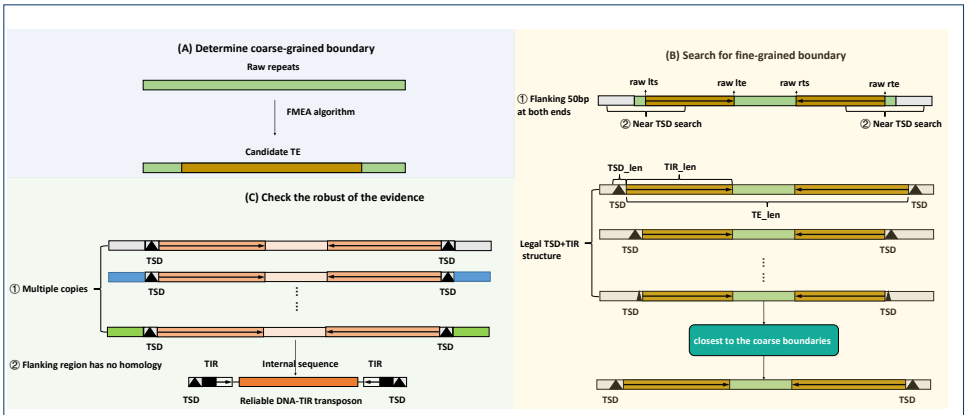

Figure 14 Determine the fine-grained boundary of TIR. A the candidate TE sequences with coarse-grained boundaries are generated by the FMEA algorithm. B the candidate TIR elements with fine-grained boundaries are identified by exhaustively searching TSD; raw lts, lte, rts, and rte refer to the raw left/right terminal start/end positions, respectively. C the elements with multiple copies and non-homology flanking regions outside the boundaries are regarded as true transposons.

Filtering false positive

Sequencing gaps

Gap sequences represent the most uncertainty in a genome assembly and are more likely to be associated with misassembly in a repetitive sequence[24]. Candidate TE sequences that contain continuous gaps longer than 10 bp are excluded.

Tandem repeat

Tandem Repeats Finder (TRF)[45] is used to identify tandem repeats with parameters "2 7 7 80 10 50 500 -f -d -m". Sequences in which tandem repeats account for more than 50% of the whole sequence are filtered out. At the same time, we found that in the candidate LTRs and TIRs, there would be many false positives with tandem repeats at the terminal sequences. Therefore, we take 100 bp and 20 bp of the terminal sequences in the candidate LTRs and TIRs, respectively. If there are more than 50% tandem repeats in their terminal sequences, the candidate sequences are considered false positives and filtered out.

Fake TIRs with LTR terminals

We observed that some of the identified TIRs candidate sequences are actually LTR transposons (LTR terminals or LTR internals). This is mainly because the long LTR terminal structure unexpectedly contains a short TIR terminal structure, and it has legitimate TSDs and more than two full-length copies, which led our TIR recognition module to incorrectly identify it as a legitimate TIR candidate. To filter out such false positive TIRs, align the LTR sequences, identified by the LTR module in HiTE, to the TIR candidate. If the TIR candidate contains 80% of an LTR element, the TIR sequence is considered a false positive and filtered out.

Filtering candidates with homology outside the boundaries based on copies

False-positive sequences are common in the genome, such as accidental terminal structures and TSD features. Our method of filtering false positives is based on

the following principles, as shown in Fig. 14C: (i) A transposon, as a repetitive sequence, appears at least twice in the genome (regardless of the old TEs, whose instances have generated a lot of divergence after a long evolution), and (ii) the boundaries of transposons determine the starting and ending positions of repeats, and the region outside the boundaries should be regarded as random sequences and should not have homology.

Based on the above principles, we flank the copies of candidate TEs and then perform alignment between these flanked copies. If more than half of the copies have homology in the flanking region, the candidate sequence is regarded as a false positive and filtered out. These candidate sequences are not true TEs, but rather long repeat sequences with a TE-like structure. Since it is difficult for LTR-RTs to find their full-length copies, we have removed the limit of at least two occurrences of LTR identification.

Homology-Based TE Searching

The autonomous LINE typically has a polyA tail and at least one RT and nuclease for transposition. The LINEs usually form TSD at the insertion site, but their truncated 5' ends make it hard to determine the true ends. SINEs have a similar structure to LINEs but are much shorter (80–500 bp), which are non-autonomous transposons that cannot transpose themselves and rely on other transposon enzymes to express, such as RT in LINEs.

Non-LTR elements (LINES and SINEs) are particularly challenging to discover due to their variability and undetectable structural signals. To date, fewer methods can give SINEs predictions, and no method can structurally identify LINEs. Most methods, such as SINE-Finder[46] and SINE_Scan[38], produce high rates of false positives but low sensitivity.

To achieve high-precision non-LTR elements annotation, we identify LINE and SINE transposons based on the method of homology search, which is shown in Fig. 11. Dfam is a public TE database, freely available under the Creative Commons Zero ("CC0") license. We extract known LINEs and SINEs from the Dfam library of RepeatMasker 4.1.2 to generate a non-LTR library. The non-LTR library is aligned to candidate TEs with coarse boundaries generated by the FMEA algorithm, and high-confidence copies are extracted.

Generating the TE library

Disjuncting nested TEs

Nested TE, usually formed by transposons inserted into other transposons, has a complex chimeric structure. HiTE implements a method to disjunct the nested TEs by (i) removing the full-length TEs contained in other sequences with more than 95% coverage and 95% identity and connecting the remaining sequences; (ii) filtering out the sequence if the length is less than 100; otherwise, treating the remaining sequence as a new TE sequence; and (iii) iterating several times to disjunct heavily nested TEs.

Generating classified and consensus models

HiTE generates TE consensus models using the clustering tool CD-HIT[47] with the parameter "-aS 0.95 -aL 0.95 -c 0.8 -G 0 -g 1 -A 80". Note that we divide the

LTR-RTs into 5' LTRs, 3' LTRs, and LTR internal regions before clustering. To determine the classification information of TE, we use the RepeatClassifier module in RepeatModeler2[25] to classify the TE consensus library.

Supplementary information

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 . . .

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

References

1. McClintock, B., et al.: Mutable loci in maize. Mutable loci in maize. (1947)
2. McClintock, B.: The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences **36**(6), 344–355 (1950)
3. Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., *et al.*: Ten things you should know about transposable elements. Genome biology **19**(1), 1–12 (2018)
4. Wells, J.N., Feschotte, C.: A field guide to eukaryotic transposable elements. Annual review of genetics **54**, 539 (2020)
5. Quesneville, H.: Twenty years of transposable element analysis in the arabidopsis thaliana genome. Mobile DNA **11**(1), 1–13 (2020)
6. Kalendar, R., Sabot, F., Rodriguez, F., Karlov, G.I., Natali, L., Alix, K.: mobile elements and plant genome evolution, comparative analyzes and computational tools. Frontiers in plant science **12** (2021)
7. Kazazian Jr, H.H.: Mobile elements: drivers of genome evolution. science **303**(5664), 1626–1632 (2004)
8. Finnegan, D.J.: Eukaryotic transposable elements and genome evolution. Trends in genetics **5**, 103–107 (1989)
9. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, 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)
10. Kapitonov, V.V., Jurka, J.: Helitrons on a roll: eukaryotic rolling-circle transposons. TRENDS in Genetics **23**(10), 521–529 (2007)
11. 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)
12. Peterson, T., Zhang, J.: The mechanism of ac/ds transposition. Plant Transposons and Genome Dynamics in Evolution, 41–59 (2013)
13. Yasir, M., Turner, A.K., Lott, M., Rudder, S., Baker, D., Bastkowski, S., Page, A.J., Webber, M.A., Charles, I.G.: Long-read sequencing for identification of insertion sites in large transposon mutant libraries. Scientific reports **12**(1), 1–9 (2022)
14. 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)
15. Gu, W., Castoe, T.A., Hedges, D.J., Batzer, M.A., Pollock, D.D.: Identification of repeat structure in large genomes using repeat probability clouds. Analytical biochemistry **380**(1), 77–83 (2008)
16. Quesneville, H., Nouaud, D., Anxolabéhère, D.: Detection of new transposable element families in drosophila melanogaster and anopheles gambiae genomes. Journal of molecular evolution **57**(1), 50–59 (2003)

17. Bao, Z., Eddy, S.R.: Automated de novo identification of repeat sequence families in sequenced genomes. Genome research **12**(8), 1269–1276 (2002)
18. Edgar, R.C., Myers, E.W.: Piler: identification and classification of genomic repeats. Bioinformatics **21**(suppl_1), 152–158 (2005)
19. Ellinghaus, D., Kurtz, S., Willhoeft, U.: Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons. BMC bioinformatics **9**(1), 1–14 (2008)
20. Ou, S., Jiang, N.: Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant physiology **176**(2), 1410–1422 (2018)
21. 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)
22. Hu, K., Xu, K., Wen, J., Yi, B., Shen, J., Ma, C., Fu, T., Ouyang, Y., Tu, J.: Helitron distribution in brassicaceae and whole genome helitron density as a character for distinguishing plant species. BMC bioinformatics **20**(1), 1–20 (2019)
23. 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)
24. 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)
25. 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)
26. Bao, W., Kojima, K.K., Kohany, O.: Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile Dna **6**(1), 1–6 (2015)
27. Hubley, R., Finn, R.D., Clements, J., Eddy, S.R., Jones, T.A., Bao, W., Smit, A.F., Wheeler, T.J.: The dfam database of repetitive dna families. Nucleic acids research **44**(D1), 81–89 (2016)
28. 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)
29. Kempken, F., Windhofer, F.: The hat family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma **110**(1), 1–9 (2001)
30. 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)
31. 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)
32. Lee, H., Lee, M., Mohammed Ismail, W., Rho, M., Fox, G.C., Oh, S., Tang, H.: Mgescan: a galaxy-based system for identifying retrotransposons in genomes. Bioinformatics **32**(16), 2502–2504 (2016)
33. McCarthy, E.M., McDonald, J.F.: Ltr_struc: a novel search and identification program for ltr retrotransposons. Bioinformatics **19**(3), 362–367 (2003)
34. Xu, Z., Wang, H.: Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons. Nucleic acids research **35**(suppl_2), 265–268 (2007)
35. Valencia, J.D., Girgis, H.Z.: Ltrdetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. BMC genomics **20**(1), 1–14 (2019)
36. Zhao, D., Ferguson, A.A., Jiang, N.: What makes up plant genomes: The vanishing line between transposable elements and genes. Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms **1859**(2), 366–380 (2016)
37. Dewannieux, M., Esnault, C., Heidmann, T.: Line-mediated retrotransposition of marked alu sequences. Nature genetics **35**(1), 41–48 (2003)
38. Mao, H., Wang, H.: Sine_scan: an efficient tool to discover short interspersed nuclear elements (sines) in large-scale genomic datasets. Bioinformatics **33**(5), 743–745 (2017)
39. Canapa, A., Barucca, M., Biscotti, M.A., Forconi, M., Olmo, E.: Transposons, genome size, and evolutionary insights in animals. Cytogenetic and genome research **147**(4), 217–239 (2015)
40. Zhang, X., Qi, Y.: The landscape of copia and gypsy retrotransposon during maize domestication and improvement. Frontiers in plant science **10**, 1533 (2019)
41. Zuccolo, A., Sebastian, A., Talag, J., Yu, Y., Kim, H., Collura, K., Kudrna, D., Wing, R.A.: Transposable element distribution, abundance and role in genome size variation in the genus oryza. BMC Evolutionary Biology **7**(1), 1–15 (2007)
42. Smit, A., Hubley, R.: Repeatmodeler open-1.0 (2008-2010)
43. Ou, S., Jiang, N.: Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons. Mobile DNA **10**(1), 1–3 (2019)
44. Quesneville, H., Flutre, T., Inizan, O., Hoede, C., Duprat, E., Arnoux, S., Faroux, G., Alfama-Depauw, F., Autard, D., Bely, B., et al.: Repet (2010)
45. Benson, G.: Tandem repeats finder: a program to analyze dna sequences. Nucleic acids research **27**(2), 573–580 (1999)
46. Wenke, T., Döbel, T., Sörensen, T.R., Junghans, H., Weisshaar, B., Schmidt, T.: Targeted identification of short interspersed nuclear element families shows their widespread existence and extreme heterogeneity in plant genomes. The Plant Cell **23**(9), 3117–3128 (2011)
47. Li, W., Godzik, A.: Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics **22**(13), 1658–1659 (2006)

Figure 15 Sample figure title

Figure 16 Sample figure title

Figures
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.