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: Long-read sequencing technology can cross the repetitive regions in the genome, which greatly improves the quality of genome assembly and gives a bright future to comprehensive annotation of TEs. While numerous methods exist for the annotation of comprehensive TEs or only for specific classes of TEs, the highly variable TEs make automated TE discovery and annotation challenging and time-consuming. In addition, all automatically generated TE libraries still require extensive manual editing due to their inability to determine the true ends of TEs. Therefore, an automated and high-precision TE annotation tool that can produce structurally intact TE libraries is urgently needed.

Results: We have developed an ensemble method for high-precision transposable element annotation, known as HiTE, which can be used to annotate almost all types of TE. Using the complementary benchmarking methods released in two recent studies, HiTE achieved the highest precision in TE annotation and restored the most number of gold standard sequences on four different model species. Moreover, HiTE can discover novel TEs with low copy numbers that are not included in known libraries.

Conclusions: The pipeline developed here is expected to improve the quality of TE annotation in eukaryotic genomes, which can enhance our understanding of the diversity and evolution of TEs and serve as a valuable addition to the genome annotation toolkit. Since HiTE aims to identify high-precision and structurally intact TEs, it can reduce unnecessary manual repair during the making of curated libraries. HiTE is open-source and freely available: https://github.com/CSU-KangHu/HiTE.

Keywords: High precision; Genome Annotation; Intact; Transposable elements; Ensemble methods

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 size of TSDs[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[13]. Long-read sequencing technologies, especially PacBio HiFi with a lower per-base error rate, can cross highly repetitive regions and improve the quality of genome assemblies[14]. 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[15].

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[16], 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[17], RECON[18], and PILER[19] 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[15].

(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[20], LTR_retriever[21], Generic Repeat Finder[22], EAHelitron[23], HelitronScanner[24], 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[25] and RepeatModeler2[26]. 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[27] and Dfam[28] 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, and classified TE library. As shown in Fig. 1, 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.

Results

HiTE Overview

HiTE is an automated TE annotation pipeline that aims to produce a high-quality and structurally intact 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, in turn, identify TE instances by recognizing features of specific families of TEs, which always suffer from false positives due to the weak structural characteristics of some TEs. A TE

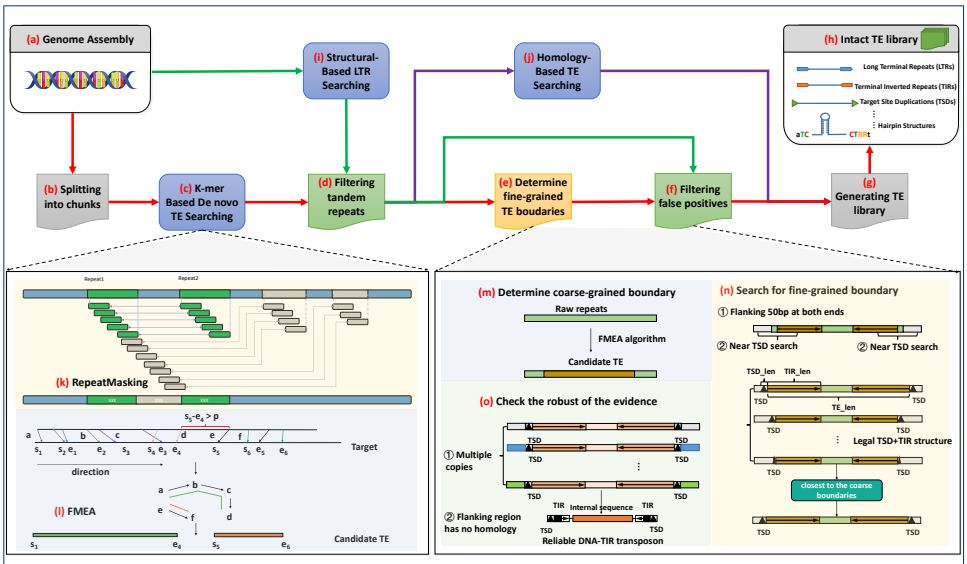

Figure 1 The workflow of HiTE. **a** HiTE takes the genome assembly as input; **b** the genome assembly is split into chunks to reduce the amount of single-round computation; **c, i, and j** three main modules of HiTE are used to identify different types of TEs; **d** the candidates with a large amount of tandem repeats are regarded as false positives and are filtered out; **e** the fine-grained boundaries of TEs are determined using signature-based methods; **f** the candidates with repetitive flanking sequences within copies that are parts of a larger repetitive element are considered false positives and are filtered out; **g** CD-HIT and RepeatClassifier are used to generate TE libraries; **h** the TE libraries generated by HiTE contain intact TE structures; **k** the candidate repeat regions within the genome are masked based on the low-frequency *k*-mer masking method, RepeatMasking; **l** the coarse boundaries of TEs are identified based on the fault-tolerant mapping expansion algorithm, FMEA; and **m, n, and o** the true ends of TEs are discovered by searching for the fine-grained TE boundaries, and the robust evidence of true TEs are checked by non-homology flanking regions within multiple copies.

discovery pipeline can overcome the shortcomings of any one particular approach by integrating a variety of existing tools into a single pipeline. However, using these existing tools directly will introduce their inherent errors (fragmentation or false positives) that will propagate to the whole genome annotation. For the different structural characteristics and distribution of TEs in the genome, we employ three modules, *k*-mer-based *de novo* TE searching, structural-based LTR searching, and homology-based TE searching, 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[21, 29]. 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[15, 30]. To identify reliable LTR-RTs, we use the mature tools LTR_FINDER[31] and LTRharvest[20] to find all candidate LTR candidates, followed by the LTR_retriever[21], the state-of-the-art in LTR identification, as a stringent filtering method for the raw results by identifying LTR-specific signals.

TIR and Helitron elements have weak structure signals, which makes it easy to generate a large number of false positives. TIR elements, for example, may be structured with short terminal inverted repeats (5-27 bp for the hAT superfamily) and target site duplications (TSDs). Helitron elements have a short hairpin structure

lying a few nucleotides before the 3′ end, with clearly defined 5′-TC and CTRR-3′ motifs (where R is a purine). To find the true ends of TIRs and Helitrons, we have developed a high-precision identification method, shown in Fig. 1.

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.

Setting up benchmarking methods for TE library evaluation

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 [15], 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 TEs, creating a challenge for algorithms to find their true ends.

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[15, 32]. 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[33]. 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 the understanding of the true performance of methods[15]. For example, many benchmarking methods promote getting the higher metrics, including the higher copy number of TE, the higher number of models generated, the longer sequences of output, and even the higher N50, which does not take into account the quality of the TE library produced.

An ideal benchmarking method should consider both the integrity of TE structures and the false-positive rate of the TE library. To address this issue, we use the benchmarking methods released in two recent studies, EDTA[25] and RepeatModeler2[26] (BM_EDTA and BM_RM2 hereafter), which give ten metrics including *Sensitivity*, *Specificity*, *Accuracy*, *Precision*, *FDR*, *F1*, *Perfect*, *Good*, *Present*, and *Not_found* (Additional File 1: Fig. S1).

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 different sizes of genomes and divergent TE landscapes. The C. briggsae genome is the smallest and is dominated

by DNA transposons; the D. melanogaster genome is dominated by LTR and LINE transposons; the Oryza sativa has a medium-sized genome with a similar proportion of LTR and DNA transposons; and the D. rerio has the largest genome, which 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[27]. 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 (Table 1).

Table 1 TE content in the benchmarking genomes.

		Class	RepBase26.05	Number of elements	Total (%)
O. sativa*	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
C. briggsae*	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
D. melanogaster*	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
D. rerio*	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

*Oryza sativa Japonica Group "assembly IRGSP-1.0"
*Caenorhabditis briggsae "assembly CB4"
*Drosophila melanogaster "assembly Release 6 plus ISO1 MT"
*Danio rerio "assembly GRCz11"

HiTE benchmarking for different species

We compared HiTE with the other three mainstream general-purpose TE annotators, including RepeatScout, EDTA, and RepeatModeler2. To date, EDTA is the best annotation tool based on structure signals, 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. RepeatModeler,

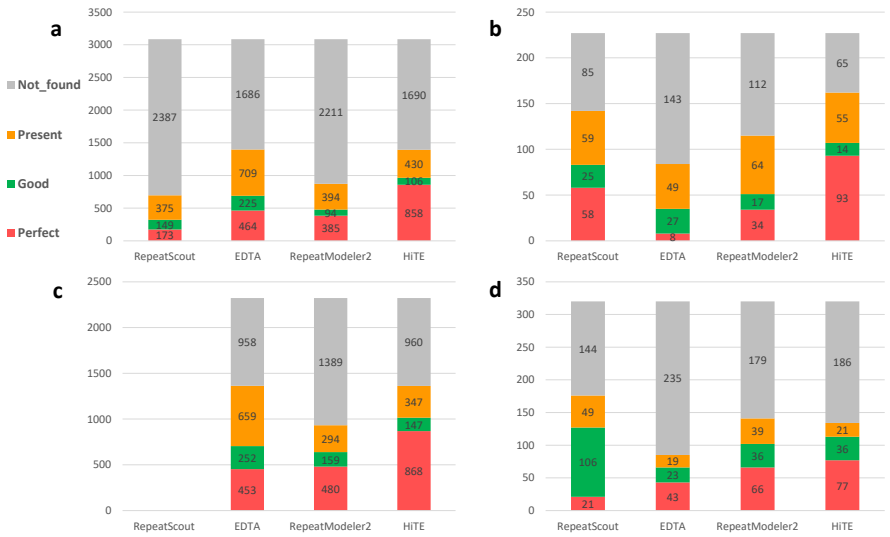

Figure 2 Performance of general-purpose TE annotators based on benckmarking method of RepeatModeler2. **a** Performance based on O. sativa genome. **b** Performance based on C. briggsae genome. **c** Performance based on D. rerio genome. **d** Performance based on D. melanogaster genome.

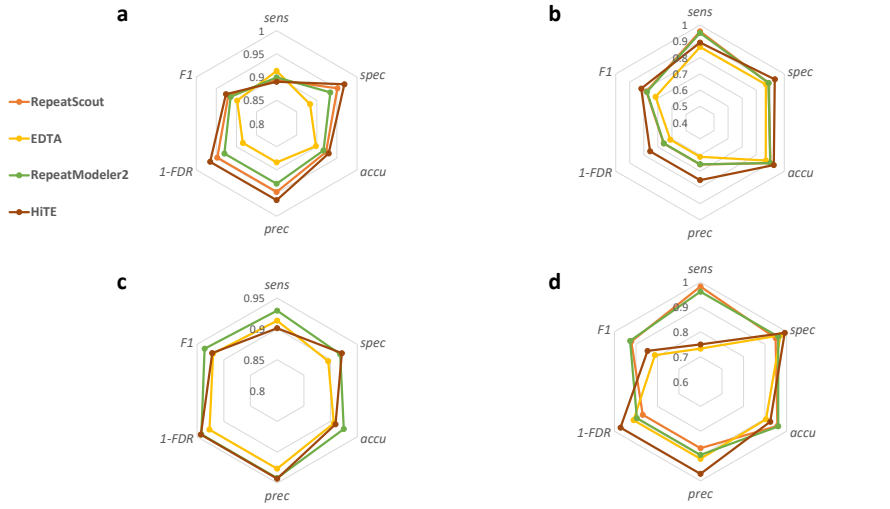

Figure 3 Performance of general-purpose TE annotators based on benckmarking method of EDTA. **a** Performance based on O. sativa genome. **b** Performance based on C. briggsae genome. **c** Performance based on D. rerio genome. **d** Performance based on D. melanogaster genome.

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[15]. Due to the limited ability to generate structurally intact TE models, RepeatModeler2 adds the LTR_retriever module to generate structurally intact LTR transposons, which greatly increases its performance[26].

The benchmarking results using BM_RM2 are shown in Fig. 2 (Additional file 2: Table S1). Among these methods, we found that RepeatModeler2 achieved good performance and was stable on all datasets. The performance of EDTA is unstable instead, producing a high number of *Perfect* models on the TE-rich genomes, such as O. sativa, but a low number on the other genomes. At the same time, the number of *Presents* models generated by EDTA is significantly higher than that of other tools, indicating many fragments may exist. RepeatScout produced more *Perfect* models on C. briggsae, which suggest that the majority of TE in the C. briggsae genome are relatively young. However, when it comes to other species, RepeatScout obtained the minimum number of *Perfect* models. Since RepeatScout cannot process more than 1 GB of genomes, it has no results for D. rerio. Notably, HiTE produced 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 restore more gold standard TE models and fewer fragments.

As shown in Fig. 3 (Additional file 2: Table S1), we noticed that RepeatScout and RepeatModeler2 both achieved a consistently high performance using BM_EDTA, which is also described in the EDTA[25]. The greatest advantage of the BM_EDTA is that it can intuitively describe the false positive rate of the tested TE library, but it cannot reflect the integrity of the TE models. For example, in O. sativa and D. melanogaster, RepeatScout obtained the lowest number of *Perfect* TE models, indicating that there are a large number of fragments, but it still achieved a high BM_EDTA performance. Instead, HiTE showed somewhat lower sensitivity but significantly higher precision performance, including *Precision*, *Specificity*, and *Accuracy*, which indicates that HiTE can identify more accurate TE. At the same time, 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. Like the structure-based method EDTA, HiTE produces a similar lower sensitivity, which does not neccessary mean that HiTE or EDTA discover fewer TE models than RepeatScout and RepeatModeler2. On the contrary, HiTE produces more *Perfect* and fewer *Not_found* TE models using the BM_RM2, which indicates more gold standard TE models are restored.

TEs discovered by HiTE are consistent with previous studies

To identify sequence motifs of typical TEs, we aligned both terminals and flanking sequences of LTR, Helitron, and EnSpm identified by HiTE and produced sequence logos that represent the nucleotide usage at each position in both terminals and their flanking regions (Fig. 4). As previously documented[35], LTRs are typically flanked by 2-bp palindromic motifs, commonly 5′-TG...CA-3′, which is shown in Fig. 4a, b. Helitrons are always inserted into 5′-AT-3′ target sites and do not have terminal inverted repeats. As shown in Fig. 4c, d, we clearly discovered the canonical terminal structure 5′-TC...CTRR-3′ (with 5′-TC...CTAG-3′ dominating) in Helitrons that are inserted into 5′-AT-3′ target sites. We also noticed a significantly higher AT content in the 5′ terminal and enriched CG content at the 3′ terminal, especially at the −29 and −13 positions, which could produce a canonical Helitron feature, a hairpin loop, consistent with previous observations[36, 24]. Furthermore, as shown in Fig. 4e, f, we discovered highly conserved CACT(A/G) motifs in the short terminals of EnSpm elements that had previously been documented [9, 37].

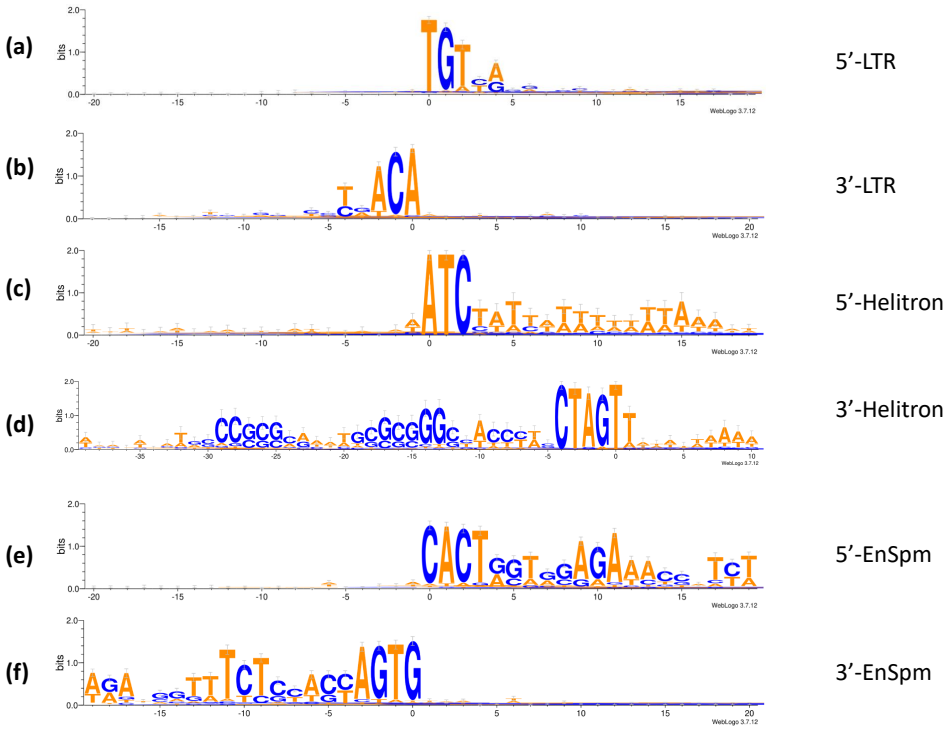

Figure 4 Sequence logos of terminal and flanking sequences of LTR, Helitron, and EnSpm candidates generated by HiTE. **(a)** 5′ terminal of LTR candidates with the starting position labeled as 0; **(b)** 3′ terminal of LTR candidates with the last position labeled as 0; **(c)** 5′ terminal of Helitron candidates with the starting position labeled as 1; **(d)** 3′ terminal of Helitron candidates with the last position labeled as -1; **(e)** 5′ terminal of EnSpm candidates with the starting position labeled as 0; **(f)** 3′ terminal of EnSpm candidates with the last position labeled as 0. Sequence logos are produced by WebLogo[34].

As an additional assessment of the ability for HiTE to discover known TEs in each of these genomes, we run RepeatMasker with each output library generated by different tools and measure the percentage of the genome masked by each major TE subclass. HiTE restores the TE landscapes of these species consistent with the reference libraries (Fig. 5). The genome of O. sativa is known to contain DNA-TIR and LTR elements in close proportions[25], which is recovered by our HiTE library (Fig. 5a). As shown in Fig. 5b, the HiTE library discovered a similar proportion to the reference library, which indicates an abundant percentage of DNA-TIR elements in the genome of C. briggsae[38]. The genome of D. rerio is dominated by class II DNA-TIR transposons, but it also has a diverse composition of LTR retroelements with many distinct families (Fig. 5c)[39]. In addition, our results show that the genome of D. melanogaster is dominated by retrotransposons, especially LTR and LINE retroelements (Fig. 5d)[40].

Performance comparison for TIR element detection

TIR elements, 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 elements may contribute to genome evolution by generating allelic diversity, inducing structural vari-

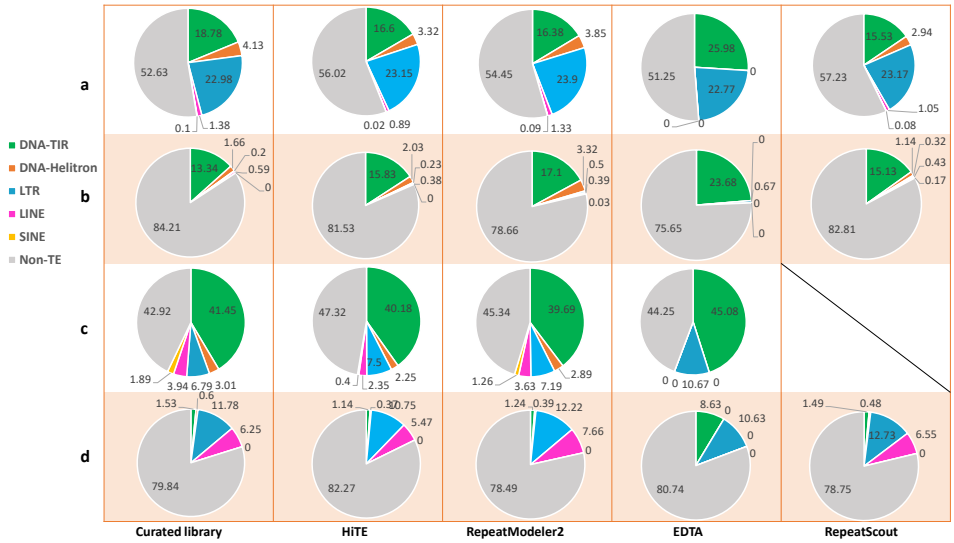

Figure 5 Percent of genome masked by each major subclass based on five types of TE libraries, generated by manually curated, HiTE, RepeatModeler2, EDTA, and RepeatScout, respectively. **a** The percentage of the O. sativa genome masked by TE libraries. **b** The percentage of the C. briggsae genome masked by TE libraries. **c** The percentage of the D. rerio genome masked by TE libraries. **d** The percentage of the D. melanogaster genome masked by TE libraries.

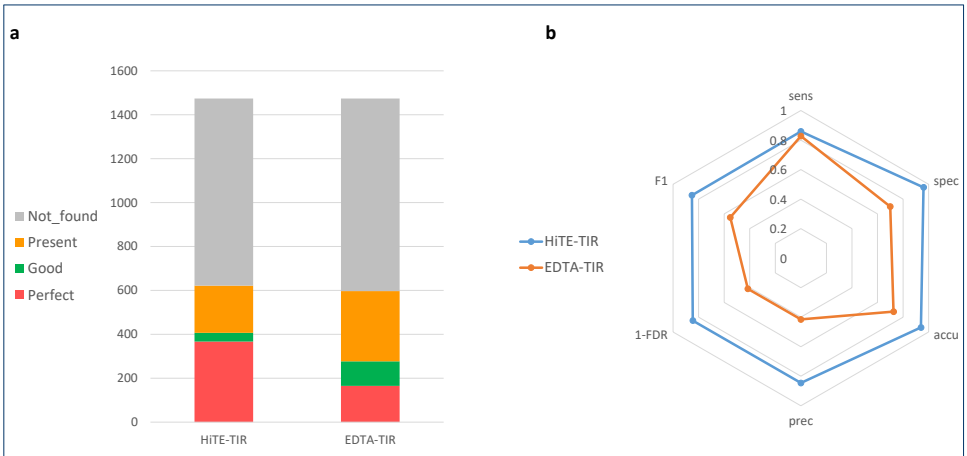

Figure 6 Performance of TIR annotators. **a** the performance based on benckmarking method of RepeatModeler2 and **b** the performance based on benckmarking method of EDTA.

ation, and regulating gene expression[11]. TIR elements are divided into nine known superfamilies by the distinguished TIRs and the size of TSDs (usually 2–11 bp). However, the identification and annotation of TIR elements are quite challenging due to the short TIRs. For example, members of the hAT superfamily have TSDs of 8 bp and relatively short TIRs of 5–27 bp[41].

Many tools have been designed for the identification of TIR elements using structural signals, such as IRF[42], TIRvish[43], TIR-Learner[11], and GRF[22], which are comprehensively evaluated in EDTA, and most of them are proven to discover a high number of false positives[25]. 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 candidates. Although the TIR module (GRF and TIR-learner) of EDTA has demonstrated great promise for structural annotation and achieved higher performance than other tools[25], it is far from satisfactory. To solve this problem, we developed a new method to achieve high-precision identification of TIR elements (see the **"Methods"** section).

As shown in Fig. 6a (Additional file 2: Table S2), according to the benchmarking results of the BM_RM2, our method can identify more TE models labelled as *Perfect* with a lower number of *Good* and *Present*. According to the benchmarking results of the BM_EDTA, our method produces a higher sensitivity, specificity, accuracy, precision, F1, and a lower FDR than EDTA (Fig. 6b; Additional file 2: Table S2). These two benchmarking methods both demonstrate that our method can achieve high-precision identification of TIR elements.

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 complete TIRs and TSDs, and the boundaries between their copies are clear. Notably, most of them have low copy numbers (Fig. 7c). At the same time, nearly half of the TIR elements have more than 3 copies (Fig. 7b), 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 elements have TIRs similar to the known TIRs in Repbase (Fig. 7a), which are likely to be non-autonomous TIR elements.

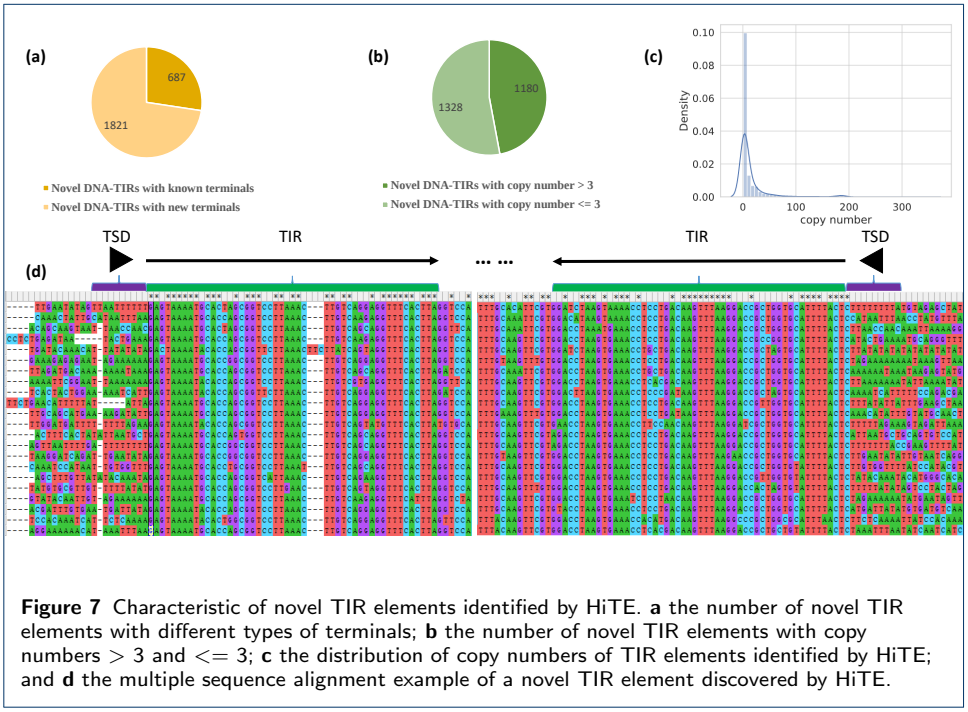

Figure 7 Characteristic of novel TIR elements identified by HiTE. **a** the number of novel TIR elements with different types of terminals; **b** the number of novel TIR elements with copy numbers > 3 and $<= 3$; **c** the distribution of copy numbers of TIR elements identified by HiTE; and **d** the multiple sequence alignment example of a novel TIR element discovered by HiTE.

Performance comparison for Helitron element detection

Helitron elements (Helitrons hereafter) 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

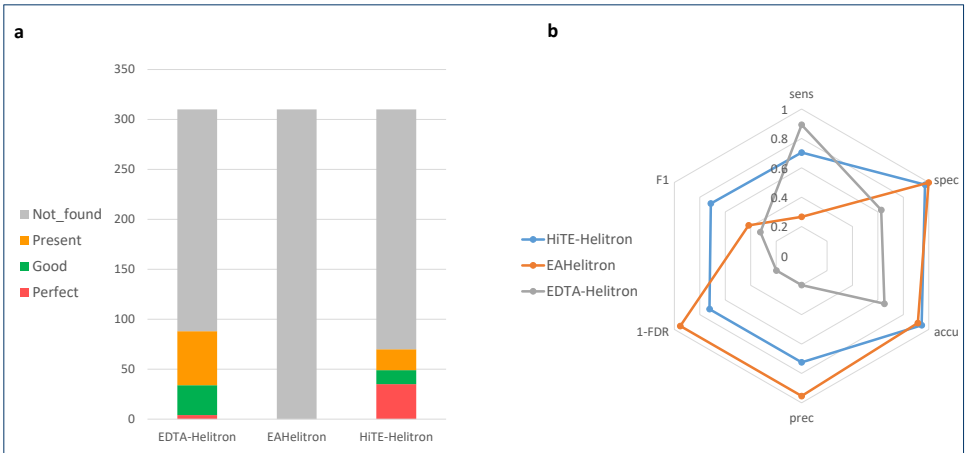

Figure 8 Performance of Helitron annotators. **a** the performance based on benckmarking method of RepeatModeler2 and **b** the performance based on benckmarking method of EDTA.

the other TEs. Helitrons have 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]. However, the weak structural signals of Helitrons make identification of these elements particularly challenging.

To date, only two tools, HelitronScanner and EAHelitron, can produce useful Helitron predictions. HelitronScanner identifies the sequence patterns in Helitrons 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 (∼4%). EDTA filters the results of HelitronScanner, greatly improving its specificity and accuracy without reducing the sensitivity[25]. Nevertheless, the precision of the Helitron identification module in EDTA is still very low (Fig. 8b; Additional file 2: Table S2).

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′ (Additional file 2: Table S4), while it is hard to know the real 5′ end of Helitron. As suggested by the authors, 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. 8b; Additional file 2: Table S2). Moreover, it cannot restore any gold standard models according to the BM_RM2 (Fig. 8a; Additional file 2: Table S2).

To discover the intact Helitrons, we have developed a new Helitron identification method, which is a further usage of the coarse boundaries of TE candidates output

by the FMEA algorithm (see the **"Methods"** section). 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 false positives are filtered out by homology outside the boundaries of copies.

The experimental results show that our Helitron identification method has the highest performance (Fig. 8; Additional file 2: Table S2), 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. 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, and our sensitivity will be further improved once EAHelitron includes a more comprehensive hairpin loop pattern.

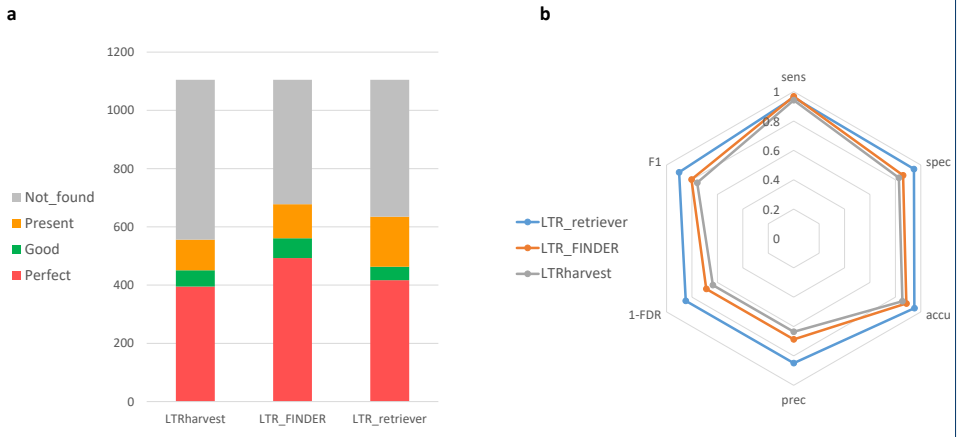

Figure 9 Performance of LTR annotators. **a** the performance based on benckmarking method of RepeatModeler2 and **b** the performance based on benckmarking method of EDTA.

Performance comparison for LTR element detection

Long terminal repeat retrotransposons (LTR-RTs) 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[44], GRF, LTR_STRUC[45], LTR_FINDER[31], LTRharvest[20], LtrDetector[46], and LTR_retriever[21]. 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 (Fig. 9b; Additional file 2: Table S2). 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 (Fig. 9a; Additional file 2: Table S2), it is still the best LTR identification method at present. Therefore, we integrated LTR_retriever into HiTE.

Performance comparison for Non-LTR element detection

Non-LTR retrotransposons include two types of TE, LINEs and SINEs[47]. 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[48]. The weak signals of non-LTR retrotransposons make them quite challenging to identify[49].

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 coarse boundaries of TE candidates output by the FMEA algorithm (see the **"Methods"** section). To benchmark the performance of the homology-based TE searching module, the non-LTR library is used to directly search candidates in the assembly genome as the competing evaluation, called Assembly-Non-LTR. Although HiTE-Non-LTR sacrificed a little sensitivity, it achieved nearly 100% precision (Fig. 10; Additional file 2: Table S2), which demonstrates the reliability of HiTE-Non-LTR.

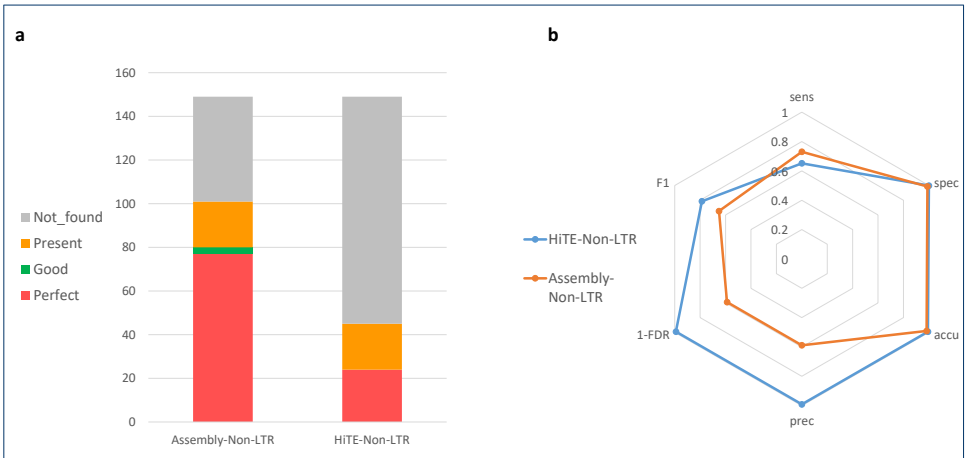

Figure 10 Performance of Non-LTR annotators. **a** the performance based on benckmarking method of RepeatModeler2 and **b** the performance based on benckmarking method of EDTA.

Influence of parameter changes in HiTE

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 flanking_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 for valid TSDs. Since these parameters have no effect on the results of LTR elements, which are discovered by LTR_retriever, we chose C. briggsae as the test species, whose genome only contains a small number of LTR elements.

As shown in Fig. 11a, the smallest k_num (such as 11) will mark 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 identified 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. 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 (Fig. 11b). Genome slicing will result in the loss of some low-copy and scattered TEs, reducing the sensitivity of the results significantly. The smaller the cut, the more TEs will be lost (Fig. 11c). As shown in Fig. 11d, when flanking_len is set to 0, the number of sensitivity and *Perfect* models is very low, which indicates that most of the TEs identified in the FMEA algorithm have coarse boundaries, while the real boundaries of most TEs can be included when flanking_len is set to 10, suggesting that the error between the rough boundaries and the real boundaries is not significant. The metrics tend to be stable after flanking_len is set to 40.

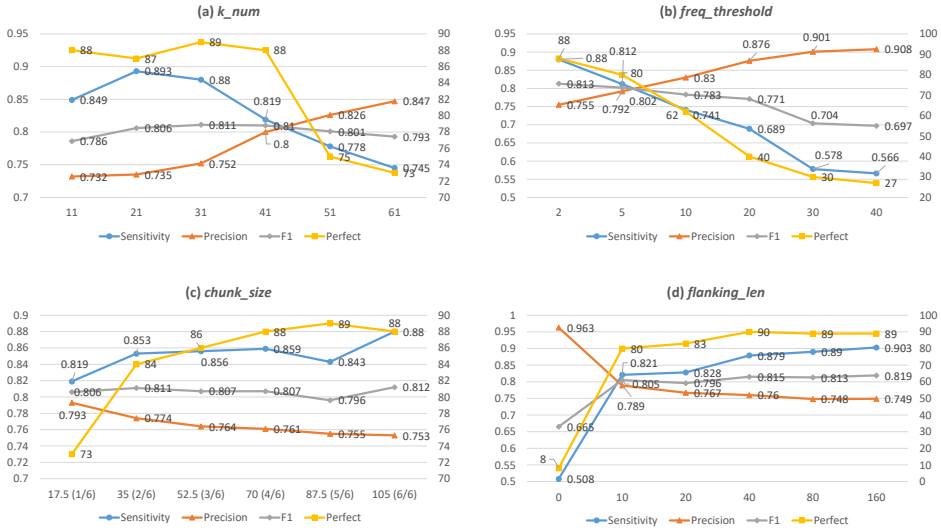

Figure 11 Influence of parameter changes on results. **a** influence of k_num parameter on results; **b** influence of freq_threshold parameter on results; **c** influence of chunk_size parameter on results; and **d** influence of flanking_len parameter on results.

Discussion

With the development of third-generation (long-read) sequencing technology, repetitive regions in the genome can be crossed, which greatly improves 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 TE annotation, known as HiTE, which has undergone extensive benchmarking on four model species. Compared with competing tools, HiTE achieved higher metrics and restored more gold standard sequences.

HiTE has the following four innovations compared to the existing tools: (i) Using repeated k-mer coverage to reduce the amount of computation. The traditional tool, RepeatScout, uses the k-mer seed expansion method to identify repeats, while HiTE uses low-frequency repeated k-mer to discover candidate repeat areas, which can reduce the amount of subsequent calculation. For example, the alignment-based identification methods take the whole genome as input to obtain pairwise alignments, while HiTE reduces the whole genome into candidate repeats, which saves a lot of computing resources (Additional file 2: Table S3). (ii) Designing a fault-tolerant mapping expansion algorithm to restore intact TEs. Highly fragmented sequences are often generated due to divergent TE models or nested TEs, which result in the multi-segment alignment of a complete TE. Alignment-based identification methods, such as RECON, use the single linkage clustering algorithm to generate TE models based on overlapping subsequences, which may identify the same TE as multiple "piles" without edges connected, resulting in multiple TE models and fragments. We have designed an alignment expansion method with fault tolerance that can cross the large gaps caused by insertion, deletion, and nested TE while retaining the complete TE structures. (iii) Defining the boundaries of TEs accurately. The TE library generated by automated identification methods still needs a lot of manual identification and repair[15], mainly due to their inability to find the true boundaries of TEs. HiTE first used the self-alignment information to determine the coarse boundaries of TEs. Then, to accurately find the boundaries of TEs, both terminals of candidates are flanked to search for valid TSDs, which can greatly reduce the cost of manual identification and repair in the later period. (iv) Implementing a highly reliable filtering method. Weak structural characteristics of many TEs caused a flood of false positives, especially for DNA-TIR and Helitron transposons. We have designed a strict filtering method based on the following truth: once the boundaries are determined, the regions outside of the true TE instances should be close to the random sequences. Therefore, more than half of the candidate copies have homology outside the boundaries, indicating that these copies belong to a larger repeat, which is considered a false positive and should be filtered out.

While HiTE can achieve high-precision TE identification and annotation, we do observe some losses of real TIR elements in Repbase, which are mainly caused by the following reasons: (i) Repbase contains many single-copy sequences, even zero-copy sequences. 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 TEs without other full-length copies in the genome. At the same time, to ensure the high reliability of identified transposons, single-copy TEs, which require high homology with known transposons or TE proteins to identify, are also filtered out. Our method, such as the TIR identification module, needs at least two full-length

copies to determine a true transposon, and most of the single-copy and zero-copy sequences are left out. (ii) Some transposons have highly divergent terminals or TSDs. Highly divergent TIRs (identity less than 0.7) and the TEs with accidental sequence homology outside the boundaries, which is similar to many false positive patterns, are filtered out. We discover some lost real TIR elements by manually inspecting the results of FMEA, mainly for highly divergent TIRs and TSDs, which further proves the effectiveness of the FMEA algorithm. At the same time, a more accurate and comprehensive filtering method helps to find more real TIR elements.

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 Helitrons, there is still potential for improvement. For example, a more comprehensive hairpin loop pattern will significantly improve the sensitivity. To date, there is no structure-based method to identify LINEs due to the truncated $5'$ ends of LINEs. A few tools designed for the 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%, while the structure-based identification methods of LINEs are still needed, which is also the direction of our future efforts.

Conclusions

The rapid development of sequencing technologies is producing more reliable genome assemblies, which gives a bright future to comprehensive annotation of TEs. However, the TE libraries generated by inaccurate TE identification tools 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, which can comprehensively and accurately identify and annotate TEs in the genome assembly. By benchmarking on four model species with different TE landscapes, we prove that HiTE can achieve higher accuracy and restore more gold standard TE models, which can be fully applied to any new sequencing genome assembly.

Methods

Kmer-Based De Novo TE Searching

The majority of *de novo* identification methods, such as RECON, are based on pairwise alignments to identify repeats. However, pairwise alignments of genomes will consume a lot of computing resources. To solve this problem, we designed the RepeatMasking method, which can reduce the search scope of the whole genome into marked candidate repeats.

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 (Fig. 1k; Additional file 1: AlgorithmS1). 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 try to skip these small gaps and connect the scattered repeat areas. Furthermore, the fake k-mers may falsely

connect multiple repeat regions together to form a larger repeat region, which can be degraded by the fault-tolerant mapping expansion algorithm.

Fault-tolerant Mapping Expansion Algorithm

The alignment-based method can identify more complete and biologically meaningful TE sequences. At the same time, due to the serious divergence between TE instances and the existence of a large number of insertions and deletions, we must consider fault tolerance when identifying TEs. Traditional alignment-based methods may divide a single TE instance into multiple 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.

The algorithm first performs self-alignment on the raw repeats masked by RepeatMasking (Fig. 1l; Additional File 1). For each query sequence, 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.

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 (TSDs; usually 2–11 bp) at the integration site. The size of the TSDs 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.

LTR-RTs typically have long direct repeat sequences (85 to 5000 bp), 2-bp palindromic motifs, 5′-TG..CA-3′ at both ends, and 4-6 bp TSDs flanked. The strong structural features of LTRs allow us to identify them directly based on the genome assembly. At present, there are some mature tools that can accurately identify TSDs and LTRs, such as LTR_Finder, LTRharvest, and LTR_retriever. We use LTR_harvest and the parallel version of LTR_Finder[50, 31] to identify candidate sequences with LTR structures in the genome, and LTR_retriever is used as a stringent filtering method for the raw results from the other two tools. 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" (Additional file 2: Table S4). Finally, false positives are filtered out by homology outside the boundaries of copies (see Section **Filtering false positives**).

TIR elements have terminal inverted repeat sequences (TIRs, 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 elements are challenging to identify due to

their short terminals. Most TIR identification tools still suffer from a large number of false positives. To discover the intact TIR elements, we first use the RepeatMasking and FEMA algorithms to determine the coarse boundaries of candidate TEs. Then, the coarse boundaries are extended by a certain length to search for all legal TSDs (Fig. 1n). We identify the identical TSDs at most a 1 bp mismatch to reduce false positives. Following that, we use the itrsearch tool included in TE Finder 2.30 (a part of the REPET[51] package) with the parameter "-i 0.7 - l 5" to search TIRs (Additional file 2: Table S4), and the sequences with TIRs and TSDs closest to the coarse boundaries are chosen as the candidate TIRs. Finally, we will determine whether the candidate TIR elements identified are true transposons using the filtering method.

Helitron element replicates through the rolling circle mechanism. When replicating, only the single strand of DNA is broken, and no TSD is generated. Helitron element 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, and mammals have been characterized by precise transitions into host AT target sites[10]. The weak structural signals of Helitrons make the identification of these elements particularly challenging. The identification of Helitrons in HiTE is still based on the candidate TEs with coarse boundaries generated by the RepeatMasking and FEMA algorithms. EAHelitron[23] is then used to identify candidate sequences with Helitron structure. Finally, the filtering method is used to filter out false positives.

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[25]. Candidate TE sequences that contain continuous gaps longer than 10 bp are excluded.

Tandem repeat

Tandem Repeats Finder (TRF)[52] is used to identify tandem repeats with parameters "2 7 7 80 10 50 500 -f -d -m" (Additional file 2: Table S4). Sequences in which tandem repeats account for more than 50% of the whole sequence are filtered out. At the same time, there are many false positives with tandem repeats at the terminal sequences of candidate LTRs and TIRs. 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 TIR elements with LTR terminals

We observed that some of the identified TIR candidates are actually LTR transposons (LTR terminals or LTR internals). The long LTR terminal may unexpectedly contain a short TIR terminal with legal TSDs and more than two full-length copies, which leads our TIR identification module to falsely identify it as a real TIR candidate. To filter out such false positives, the TIR candidates are aligned to the LTR sequences produced by the LTR module. The TIR candidates, which contain more than 80% of an LTR element, are considered false positives and filtered out.

Filtering out false positives by homology outside the boundaries of copies

False positives with TE structures are common in the genome, such as accidental terminal structures and TSD-like features. Our method of filtering false positives is based on the following principles, as shown in Fig. 1o: (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 candidates are not real TEs, but rather long repeat sequences with TE-like structures. 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 TSDs 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, no structure-based methods can identify LINEs, and fewer methods can produce SINEs predictions. Most methods, such as SINE-Finder[53] and SINE_Scan[49], produce high rates of false positives and low sensitivity.

To achieve high-precision non-LTR element annotation, we identify LINE and SINE transposons based on the method of homology search (Fig. 1j). 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 used to discover confident non-LTR elements by searching through the candidate TEs generated by the FMEA algorithm.

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[54] with the parameter "-aS 0.95 -aL 0.95 -c 0.8 -G 0 -g 1 -A 80" (Additional file 2: Table S4). Notably, 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[26] 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).

Availability of data and materials
All data, code, and scripts are freely available at https://github.com/CSU-KangHu/HiTE.

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
KH and JW conceived the study. JW supervised the project. KH conducted the analyses. KH developed the HiTE package. KH and JW wrote the manuscript. All authors read and approved the final manuscript.

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. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., *et al.*: The complete sequence of a human genome. Science **376**(6588), 44–53 (2022)

14. 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)
15. 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)
16. 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)
17. 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)
18. Bao, Z., Eddy, S.R.: Automated de novo identification of repeat sequence families in sequenced genomes. Genome research **12**(8), 1269–1276 (2002)
19. Edgar, R.C., Myers, E.W.: Piler: identification and classification of genomic repeats. Bioinformatics **21**(suppl_1), 152–158 (2005)
20. 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)
21. 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)
22. 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)
23. 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)
24. 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)
25. 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)
26. 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)
27. Bao, W., Kojima, K.K., Kohany, O.: Repbase update, a database of repetitive elements in eukaryotic genomes. Mobile Dna **6**(1), 1–6 (2015)
28. 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)
29. Ou, S., Chen, J., Jiang, N.: Assessing genome assembly quality using the ltr assembly index (lai). Nucleic acids research **46**(21), 126–126 (2018)
30. Jedlicka, P., Lexa, M., Kejnovsky, E.: What can long terminal repeats tell us about the age of ltr retrotransposons, gene conversion and ectopic recombination? Frontiers in plant science **11**, 644 (2020)
31. 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)
32. Bell, E.A., Butler, C.L., Oliveira, C., Marburger, S., Yant, L., Taylor, M.I.: Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated edta and deepte de novo pipelines. Molecular Ecology Resources **22**(2), 823–833 (2022)
33. 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)
34. Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E.: Weblogo: a sequence logo generator. Genome research **14**(6), 1188–1190 (2004)
35. Zhou, S.-S., Yan, X.-M., Zhang, K.-F., Liu, H., Xu, J., Nie, S., Jia, K.-H., Jiao, S.-Q., Zhao, W., Zhao, Y.-J., *et al.*: A comprehensive annotation dataset of intact ltr retrotransposons of 300 plant genomes. Scientific Data **8**(1), 1–9 (2021)
36. Yang, L., Bennetzen, J.L.: Structure-based discovery and description of plant and animal helitrons. Proceedings of the National Academy of Sciences **106**(31), 12832–12837 (2009)
37. DeMarco, R., Venancio, T.M., Verjovski-Almeida, S.: Smtrc1, a novel schistosoma mansoni dna transposon, discloses new families of animal and fungi transposons belonging to the cacta superfamily. BMC Evolutionary Biology **6**(1), 1–13 (2006)
38. Harris, L.J., Baillie, D., Rose, A.: Sequence identity between an inverted repeat family of transposable elements in drosophila and caenorhabditis (1988)
39. Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., *et al.*: The zebrafish reference genome sequence and its relationship to the human genome. Nature **496**(7446), 498–503 (2013)
40. Lerat, E., Rizzon, C., Biémont, C.: Sequence divergence within transposable element families in the drosophila melanogaster genome. Genome research **13**(8), 1889–1896 (2003)
41. Kempken, F., Windhofer, F.: The hat family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma **110**(1), 1–9 (2001)
42. 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)
43. 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)
44. 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)
45. McCarthy, E.M., McDonald, J.F.: Ltr_struc: a novel search and identification program for ltr retrotransposons. Bioinformatics **19**(3), 362–367 (2003)
46. 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)
47. 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)
48. Dewannieux, M., Esnault, C., Heidmann, T.: Line-mediated retrotransposition of marked alu sequences. Nature genetics **35**(1), 41–48 (2003)
49. 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)
50. 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)
51. 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)
52. Benson, G.: Tandem repeats finder: a program to analyze dna sequences. Nucleic acids research **27**(2), 573–580 (1999)
53. 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)
54. 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)

Figures
Tables
Additional Files
Additional file 1 — Supplementary Materials
Additional file 2 — Supplementary Data
Table S1. Details of performance among general-purpose repeat annotators. **Table S2.** Details of performance among all types of TE annotators based on O. sativa. **Table S3.** Time and resource consumption. **Table S4.** Details of tools and parameters used in benchmarking and HiTE.