**HiTE, an Ensemble Method for High-Precision Transposable Elements Annotation**

目录

[HiTE方法介绍 2](#_Toc118744268)

[背景 2](#_Toc118744269)

[HiTE的创新点 2](#_Toc118744270)

[TE识别方法 3](#_Toc118744271)

[候选重复区的识别 3](#_Toc118744272)

[容错的比对扩展算法（Coarse-grained boundary mapping） 5](#_Toc118744273)

[Structural-Based De Novo TE Search 7](#_Toc118744274)

[Determine fine-grained boundary 7](#_Toc118744275)

[Check the robust of the evidence 10](#_Toc118744276)

[生成一致性序列 10](#_Toc118744277)

[Filtration and Classification 10](#_Toc118744278)

[HiTE测试结果 11](#_Toc118744279)

[测试结论 17](#_Toc118744280)

[参数变化对结果的影响 17](#_Toc118744281)

[应用案例 20](#_Toc118744282)

[TE影响基因组大小 20](#_Toc118744283)

[TE影响基因表达 21](#_Toc118744284)

[主流的重复识别方法原理图 23](#_Toc118744285)

[RECON方法 23](#_Toc118744286)

[RepeatScout方法 23](#_Toc118744287)

[RepeatModeler2方法 24](#_Toc118744288)

[EDTA方法 24](#_Toc118744289)

[参考文献 25](#_Toc118744290)

## Introduction

Since being discovered in maize by Barbara McClintock in 1947 (McClintock et al., 1947; McClintock, 1950), transposable elements (TEs), consisting of the major parts of repetitive regions in genomes, have been detected in most eukaryotic species (引用：Ten things you should know about transposable elements; A Field Guide to Eukaryotic Transposable Elements). 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（Quesneville, 2020; Kalendar et al., 2021; RepeatModeler2

Kazazian Jr, 2004；）.

TEs are generally divided into two classes based on the transposition intermediates (RNA or DNA) (Finnegan, 1989), and further split into families and subfamilies on the basis of various structural features (Wicker et al., 2007). 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 (Wicker); (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.  (Helitrons on a roll: eukaryotic rolling-circle transposons); (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 (TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome). 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). (The mechanism of Ac/Ds transposition. In Plant Transposons and Genome Dynamics in Evolution).

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 across highly repetitive regions, are improving the quality of genome assemblies(引用Long-read sequencing for identification of insertion sites in large transposon mutant libraries). 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.

There are many complications exist that leads to the identification of TEs is not straightforward, including but not limited to[引用Methodologies for the De novo Discovery of Transposable Element Families]: (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.

Detection of new

transposable element families in Drosophila melanogaster and Anopheles

gambiae genomes.

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[Identification of repeat structure in large genomes using repeat probability clouds], 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[Detection of new transposable element families in Drosophila melanogaster and Anopheles gambiae genomes], RECON[Automated de novo identification of repeat sequence families in sequenced genomes], and PILER[PILER: identification and classification of genomic repeats] are examples of self-comparison approaches, which 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 (Methodologies for the De novo Discovery of Transposable Element Families). (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, LTR\_harvest, LTR\_retriever, Generic Repeat Finder, EAHelitron, HelitronScanner, 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 and RepeatModeler2. 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 [Repbase Update, a database of repetitive elements in eukaryotic genomes] and Dfam [The Dfam database of repetitive DNA families] are high-quality consensus libraries for a limited set of species. Until now, all automatically generated TE libraries required extensive manual editing before being accepted. [Methodologies for the De novo Discovery of Transposable Element Families；Repbase Update, a database of repetitive elements in eukaryotic genomes；The Dfam community resource of transposable element families, sequence models, and genome annotations；]. 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 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 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 non full length 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. Like the traditional de novo method, our pipeline can discover novel TE families. More importantly, it can accurately identify the structurally intact TE families by using highly conservative structural features and copy sequence support. At the same time, the accurate definition of the boundary 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.

转座子 (TE)，也称为移动遗传元件，已被证明是物种进化和多样性的重要驱动因素，因为它们具有移动和诱导基因组突变的能力。 TE 构成了许多物种基因组的主要部分，例如在人类基因组中至少占 45%。然而，基因组中的 TE 衍生序列会随着时间的推移积累变异，这使得它们的发现和表征非常具有挑战性。

**TE识别的难点主要表现在三个方面**：

1. TE拷贝序列通常发生了大量的突变，很难判断两条序列的同源性。
2. TE的起始和终止位置（边界）难确定。基于相似性搜索或者比对的方法很难识别到精确的TE边界，而精确的边界和TSD特征对于TE的分类至关重要（尤其是非自治的TE）[1]。
3. TE经常插入到其他TE中，形成nested TE，这样的嵌套可能持续多次。

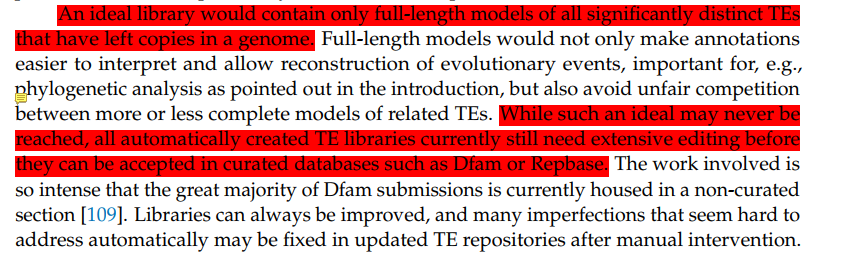
对于绝大多数的de novo识别方法都是基于pairwise比对的相似性来识别重复家族的，例如：

* Single-linkage clustering (Agarwal and States, 1994)
* REPuter (Kurtz et al., 2000)
* RepeatFinder (Volfovsky et al., 2001)
* **RECON (Bao and Eddy, 2002)**
* RepeatGluer (Pevzner et al., 2004)
* PILER (Edgar and Myers, 2005)
* REPET (Flutre et al., 2011)

其中以RECON为代表，作为从头重复家族识别的主要工具，被用于许多基因组项目。

### HiTE的创新点

1. **利用k-mer覆盖的方法减少计算量。**与经典的基于k-mer种子扩展方法识别工具RepeatScout不同的是，HiTE利用低频k-mer来确定候选重复区，能以较快的速度减少后续的计算量。基于比对的识别方法需要将整个基因组进行pairwise比对，而HiTE只需要对候选重复区进行比对，节省了大量的计算资源。
2. **设计了具备容错的比对扩展算法。**经典的基于pairwise比对的识别重复方法（如RECON）根据overlapped subsequence使用Single Linkage Clustering 算法生成TE序列。由于TE序列在演化过程中产生分化或nested TE等因素的影响，从而导致一条完整的TE产生多段比对，这种现象非常常见。常规的基于自比对的方法可能将同一条TE识别成多个没有边相连的piles，生成多个family，从而产生了大量的fragment，这种fragment会严重影响完整TE家族的识别和分类。我们设计了一种能够具备容错性质的比对扩展方法，能够很好地解决序列由于发生insertion、deletion和nested TE（常见于转座子中）等导致的序列不一致产生的影响，尽可能保留完整的TE结构。
3. **定义TE边界。**在综述“*Storer J M, Hubley R, Rosen J, et al. Methodologies for the De novo Discovery of Transposable Element Families[J]. Genes, 2022, 13(4): 709.*”中提到 “These are largely open problems for automated methodologies and still require extensive manual curation to identify and remedy.” 自动识别方法生成的TE库仍需要大量的人工识别和修复，这主要源于这些方法无法准确的定义TE的边界。而我们的方法首先在容错比对扩展阶段利用比对信息确定了TE的**粗**边界，同时我们在粗边界的基础上向两端各延伸了50 bp，并基于LTR、TIR和Helitron不同的终端结构、TSD以及5’-…-3’ motif等特征精确识别TE的边界，能够极大减少后期人工识别和修复的成本。
4. **高可靠的过滤方法。**在许多de novo工具中，假阳性是困扰它们的主要因素。de novo方法不依赖已知的转座子库（Known TE/protein），因此能够识别全新的TE，然后这也造成了假阳性序列的泛滥，尤其对于DNA-TIR类型转座子（较短的终端结构）。为了能过滤掉假阳性序列，我们设计了一种严格的基于拷贝flanking区域的比对方法，能获得高可靠的转座子序列。



## Methods

研究目的、设计、设置

### Calculation of benchmarking metrics

Benchmarking:

**RepeatModeler2:**

Run RepeatMasker v4.1.2 with parameters “-lib ${repbase\_lib} –nolow –pa 48 ${test\_lib}”.

Run get\_family\_summary\_paper.sh script using “${test\_lib}.out” as input.

(download from “https://github.com/jmf422/TE\_annotation/blob/master/get\_family\_summary\_paper.sh”)

**EDTA:**

Run RepeatMasker with parameters “-e ncbi -pa 48 -q -no\_is -norna -nolow -div 40 -lib ${repbase\_lib}/${test\_lib} -cutoff 225 ${genome\_path}”. These parameters identified homologous sequences with up to 40% divergence without detecting bacterial insertion elements, small RNA (pseudo) genes, and low complexity DNA.

Run lib-test.pl in the EDTA package with parameters ”-std ${repbase\_lib\_out} –tst ${test\_lib\_out} –cat Total”

HiTE:

With parameters “-a oryza\_sativa -t 48 -k 31 --chunk\_size 400 --plant 1 --flanking\_len 50”

CD-HIT-EST:

With parameters “-aS 0.95 –aL 0.95 –c 0.8 –G 0 –g 1 –A 80”.

TRF:

Identification of tandem sequences was achieved by Tandem Repeats Finder [74] with parameters “2 7 7 80 10 50 500 -f -d -m”.

LTR\_FINDER：

With parameters “-harvest\_out –threads 48”.

LTRharvest:

With parameters “-seed 20 -minlenltr 100 -maxlenltr 7000 -similar 85 -motif TGCA -motifmis 1 -mintsd 4 -maxtsd 6 -vic 10 -seqids yes”

LTR\_retriever:

With parameters “-inharvest ${input\_path} -noanno –threads 48”

TIR:

With parameters “-i 0.7 -l 7”

EAHelitron:

With parameters “-p 48 -o EAHelitron -u 20000 -T "ATC" -r 3”

Remove nested TE:

coverage=0.95, identity\_threshold=95, 迭代5次。

repbase中的TIRs以及novel candidate TIRs中的终端序列: itrsearch with parameters “-i 0.7 -l 7”

(novel TIR elements and new terminal repeats method: use cd-hit-est, cluster with - aS 0.8 – aL 0.8 – c 0.8, and then analyze that the transposons in the class that do not appear in the curated library are new TIR).

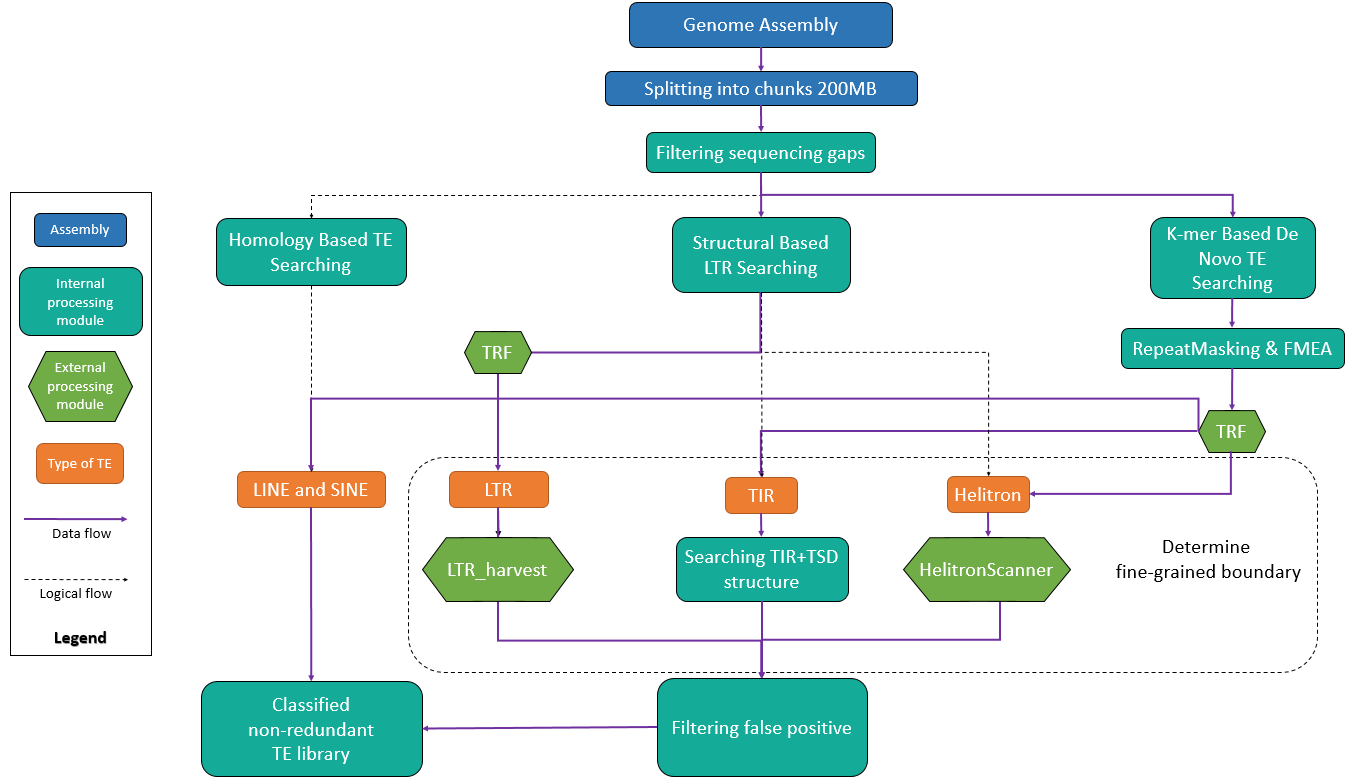
### HiTE Overview

HiTE is an automated TE annotation pipeline that aims to produces 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 are less susceptible to these particular problems. Unfortunately, signature-based methods always suffer from false positives due to the weak structural characteristics of many TEs. Both Purely de novo and Signature based methods have their own defects. By integrating a variety of tools into a single pipeline, a TE discovery pipeline can overcome the shortcomings of any one particular approach. However, the existing tools have different defects, such as fragmentation, false-positive sequences, chimaeras of TEs, and so on. It is difficult for us to deal with the results of the direct use of these tools. HiTE only uses some functions of mature tools to edit the raw boundary of candidate TEs and finally reduces the false-positive rate based on its own reliable filtering steps. Therefore, HiTE will not introduce the inherent errors of other tools.

Because of different transposons' structural characteristics and distribution in the genome, we have employed different strategies. HiTE uses three modules, k-mer-based de novo TE searching, structural-based LTR searching, and homology-based TE searching, to identify almost all common transposons, including LTRs, TIRs, Helitrons, LINEs, and SINES. Although LTR-RTs have significant structural characteristics, they are affected by many factors, such as intraelement recombination and mutations, which lead to intact LTR-RTs contributing only a small fraction of all LTR-RT related sequences in a genome [LTR\_retriever]. Long insertions are also more likely to be selectively disadvantageous to the genome. Full-length LTR elements are often reduced to solo LTRs via LTR–LTR recombination [Methodologies for the De novo Discovery of Transposable Element Families]. Therefore, for LTR-RTs, we use the mature tool LTR Finder [LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons] to identify reliable LTR-RTs by identifying LTR-specific signals (including the TG..CA box, TSD, PBS, PPT, and protein domains). TIR transposons have a TIR and a TSD structure. Helitron ends are defined only by TC or CTRR motifs (where R is a purine) and a short hairpin structure lying a few nucleotides before the 3′ end. However, using only such structures makes it easy to generate a large number of false-positive sequences. Therefore, we have designed a different strategy for these two types of transposons. 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 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 non full length 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. SINE and LINE transposons are very difficult to identify due to their weaker structural features and fuzzy terminals. At present, there is no reliable tool to identify these two types of transposons with high accuracy. In this case, we extract SINE and LINE transposons from the Dfam library to form a reliable library, and then define the true ends of candidate TEs based on this library.

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 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 recognition method requires pairwise alignment of the whole genome, while HiTE only needs to compare candidate repeats, saving a lot of computing resources. (ii) **A Fault-tolerant Mapping Expansion Algorithm is designed to restore an intact TE.** Due to the influence of factors such as differentiation or nested TE during the evolution of TE sequences, highly fragmented sequences are often generated, resulting in multi segment alignment of a complete TE. Traditional pairwise alignment based recognition 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 generate multiple families, resulting in a large number of fragmentation results. This fragmentation will seriously affect the identification and classification of the complete TE family. We have designed an alignment expansion method with fault tolerance, which can well solve the impact of sequence inconsistency caused by insertion, deletion and nested TE, and retain the complete TE structure as much as possible. (iii) **Accurately define the TE boundary.** These are largely open problems for automated methodologies and still require extensive manual curation to identify and remedy[*Methodologies for the De novo Discovery of Transposable Element Families*]. The TE library generated by automatic identification methods still needs a lot of manual identification and repair, mainly because these methods cannot accurately define the boundary of TE. In our method, we first used the self-alignment information to determine the coarse boundary of TE. And then, we extended both ends, and searching for exactly terminal signals to accurately identified the boundary of TE, 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). To reduce the false positive TIR sequences, we first filter out the false positive TSD. Because the real TSD should be different in most copies except for 2bp TA, 3bp TAA/TTA, and 4bp TTAA, we have removed those TSD that appear in more than half of the total copies in different copies. In addition, we have designed a strict alignment method focused on the flanking regions of the copies. We believe that after the specific boundary is determined, the region near the copy of the true transposon 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, our pipeline can discover novel TE families. More importantly, it can accurately identify the structurally intact TE families by using highly conservative structural features and copy sequence support. At the same time, the accurate definition of the boundary reduces a large amount of manual repair.



**Fig. 1.** The workflow of HiTE

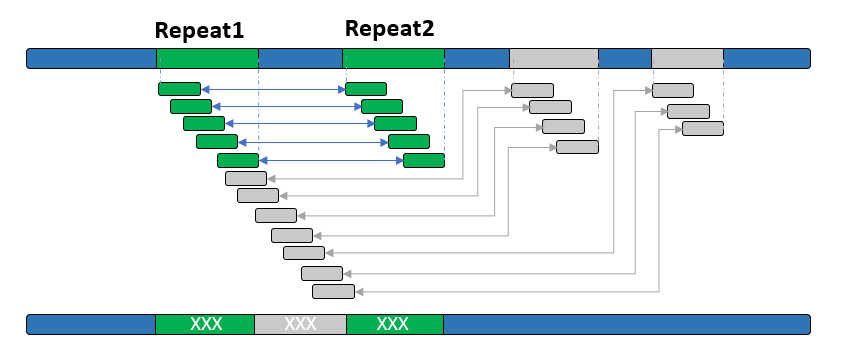
### Kmer-Based De Novo TE Searching

#### RepeatMasking

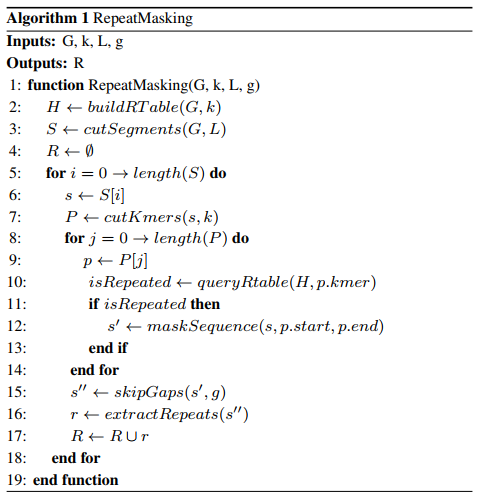
The majority of de novo identification methods, such as RECON, are based on the similarity of pairwise alignment to identify repeats. However, direct pairwise alignment of genomes will consume a lot of computing resources. To solve this problem, we designed the RepeatMasking method, which can mark candidate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

The recognition of the original repeat sequence is based on the following fact: if there are two repeat sequences, regardless of their differences, 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.

As shown in Figure 2, 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. It is worth noting that due to divergence between the two repeat sequences, the originally continuous repeat regions may break because there is no duplicate k-mer, forming a small gap, or because the false positive k-mer connects the originally discrete multiple repeat regions together, so the candidate repeat sequences are not completely true repeat sequences. For the first problem, we use the fault tolerance parameter fault\_tolerant\_bases to span these small gaps and connect the scattered, repeated areas. For the solution to the second problem, the fault-tolerant mapping expansion algorithm in the next section can distinguish different TEs.



**Fig. 2.** The illustration of RepeatMasking



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, which is equivalent to fault\_tolerant\_bases parameter; 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.

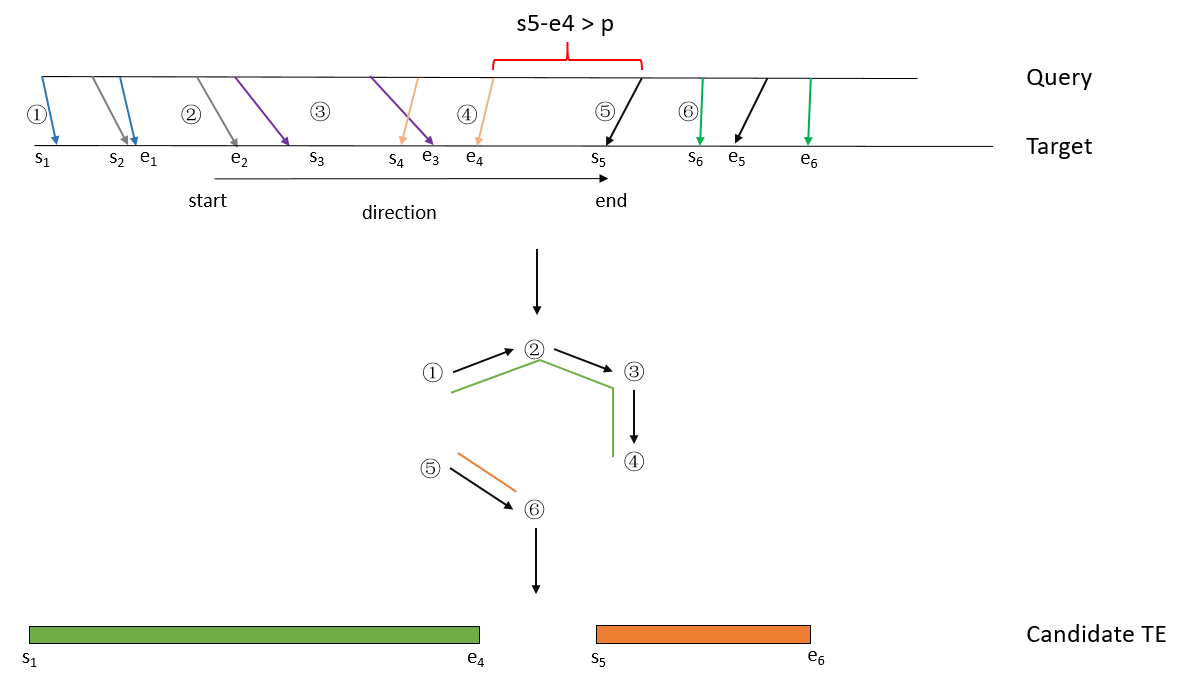
#### #RepeatMasking缩小待比对序列

#实验证明RepeatMasking能将待比对序列缩小至原先的2倍以上，例如在布氏秀丽隐杆线虫基因组上，由105MB缩小至23MB，缩小了**4.6**倍；在果蝇基因组上，待比对序列从原本的165MB缩小至55MB，缩小了**3**倍；在水稻基因组上，由362MB缩小至151MB，缩小了**2.4**倍; 在斑马鱼基因组上，由1.6GB缩小至737MB，缩小了**2.2**倍。

### Fault-tolerant Mapping Expansion Algorithm (FMEA)

The pairwise alignment method can identify more complete and biologically meaningful TE sequences. At the same time, due to the serious differentiation between TE copies and the existence of a large number of indels, 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.

The algorithm first performs self-alignment on the raw RepeatMasking repeats. Adjacent alignments are gathered for each query 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. We need to remove redundancy from the extension sequences. We use a representative sequence to represent the extended sequence with overlap and update the representative sequence iteratively to make it contain the farthest boundary. At the same time, count how many extended sequences this representative sequence supports. Remove the non-full length representative sequence that is not supported by more than two sequences.



**Fig. 3.** The illustration of the fault-tolerant mapping expansion algorithm

Algorithm description:

1. Obtain pairwise local alignments between candidate repeat sequences in **Sn**.
2. Set an extension threshold **p**, and then sort the alignments according to the start and end positions on the alignment target.
3. Construct graph **G (V, E)**, where **V** represents all the subsequences and **E** represents the connectivity between subsequences. Two subsequences **Sk**, **Sk-1** are considered connected if **ek>ek-1** and **sk-ek-1 < p** with the same direction of strand.
4. Define an element **S(sk, ek)**, the longest fragment in the component, as the candidate TE sequence.
5. Obtain a non-overlapped sequence set to represent the candidate repeat sequences of each query.

**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 3: ①, ②, ③, ④, ⑤, ⑥. 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 ①, whose starting and ending positions are s1 and e1; the starting and ending positions of subsequence ② are s2 and e2. Since e2>e1 and s2-e1<p, it means that adding ② can expand the length of the current subsequence, so we connect the subsequences ① and ②. Similarly, connect subsequences ③ and ④. However, since s5-e4>p, it indicates that the subsequence ⑤ is too far from sequence ④ 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 ①, ②, ③ and ④ and TE sequence 2 (starting s5, terminating e6) corresponding to subsequences ⑤ and ⑥.

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

#### Determine fine-grained boundary

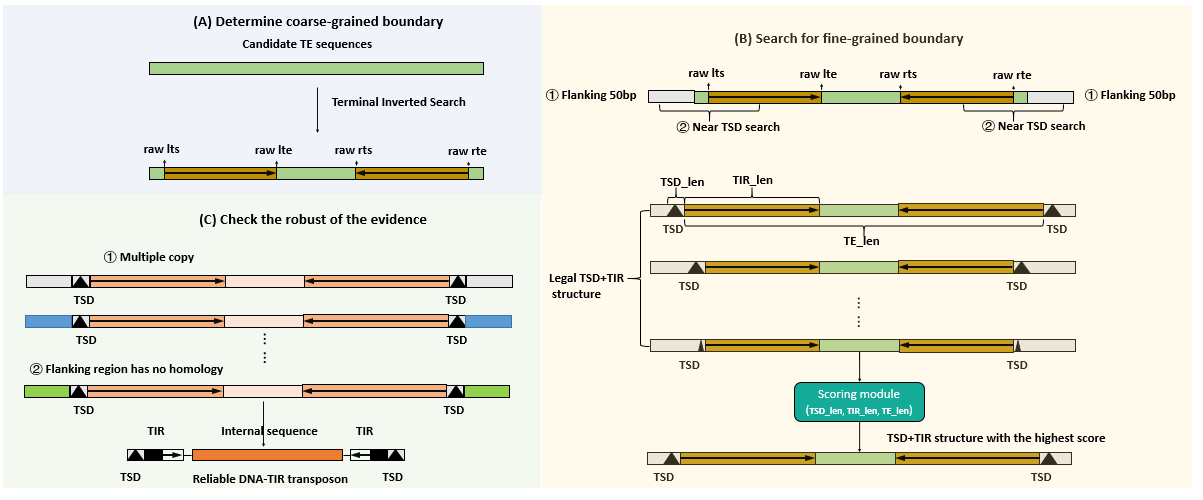
TE usually has certain structural characteristics, for example, LTR-RT has very obvious long terminal repeat (LTR) characteristics at both ends of the sequence, and DNA transposon sequences have terminal reverse repeat (TIR) characteristics at both ends. In addition, when TE is inserted into the genome, it is accompanied by DNA double strand breaks, and its repair results in the formation of two short target site duplications (TSD; usually 2–11 bp) at the integration site (TE flank). The size of target site duplication (TSD) is a feature of most superfamilies and can be used as a diagnostic feature of TE. By identifying these TE superfamily specific structural features, accurate TE sequences can be obtained. We describe the structural characteristics of three main types of TE and refer to the literature for more TE structural characteristics.[2]。

① LTR transposon sequences typically have long direct repeat sequences (85 to 5000 bp) at both ends and 2-bp palindromic motifs, 5 '- TG..CA-3' at both ends. At the same time, its flank has 4-6 bp TSD characteristics[3]. The more prominent structural features of LTR enable us to identify directly based on genome while ensuring a manageable number of false positives. At present, there are some very mature tools that can accurately identify TSD and LTR boundaries, such as LTR\_Finder[4]. We use parallel version of LTR\_Finder with default parameters and LTR\_harvest with the parameter " -seed 20 -minlenltr 100 -maxlenltr 7000 -similar 85 -motif TGCA -mintsd 4 -maxtsd 6 -vic 10" to identify candidate sequences with LTR structure in the genome. Finally, judge whether the candidate LTR sequence is a true transposon. See Filtering false positives for details.

② Both ends of the TIR transposon have short terminal inverted repeat sequences (usually a few bp to hundreds of bp). At the same time, TIRs of some superfamilies have conserved motifs characteristics. 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. Due to their short terminal structure, TIR transposons are not easy to identify. At present, common TIR recognition tools often have a large number of false positive sequences, such as GRF[5]and TIR-learner[6]。To accurately identify TIR transposons while reducing false positives, we first use RepeatMasking and FEMA to determine the repeat sequence's coarse boundary. Secondly, to determine the true boundary of TE, we expand the flank of the coarse boundary by 50 bp and then enumerate the legal TSDs near the coarse boundary, as shown in Figure 4 (B). Because TSD features degenerate in older TE, we identify completely consistent TSDs to avoid introducing too many false positive sequences. Only for TSDs above 8 bp, we allow a 1 bp mismatch. Then, after determining the candidate TSD, we use the itrsearch tool with the parameter "-i 0.7 –l 5" in TE Finder 2.30 (a part of the REPET [7] package) to determine whether the candidate sequence has a TIR structure and retain the candidate sequence with a TIR and TSD structure. TSDs that appear in more than half of the total copies across multiple copies have been removed. It may generate multiple candidate sequences with a TIR and TSD structure for each coarse boundary sequence. We select the candidate sequence with the TIR and TSD structures closest to the coarse boundary as the identified TIR sequence. Finally, we will judge whether the identified TIR sequence is a true transposon. Also see Filtering false positives.

我们不直接使用itrsearch搜索TIRs，因为它会识别到很多类似TIR结构的序列，这样的结构+一个可能的TSD就会被我们识别为一个候选TIR，如果此时它离原始边界比较近，那么真实的TIR+TSD就会被我们忽略。而且据我们统计，在Repbase中，绝大多数的TIR的开始、结束5bp完全一致，906 out of 1474，加上1bp mismatch 有 1151 out of 1474。因此我们首先判断开始、结束5bp是否至多有1bp mismatch，然后再使用itrsearch来判断序列是否具有TIRs结构。

③ 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' nucleus into host AT target sites [Helitrons on a roll: eukaryotic rolling circle transmissions]. Helitron lacks TSD characteristics, and only the conservative structure 5'-TC...CTRR-3' can be used for identification. In ②, we used RepeatMasking and FEMA to determine the coarse boundary of the repeat sequence, which also includes the Helitron transposon. As a result, we added 50 bp to the flank of the repeat sequence with the coarse boundary and then used the tool EAHelitron[7] to find candidate sequences with Helitron structure in the sequence. Finally, determine whether the candidate Helitron sequence is a true transposon. Also see Filtering false positives.



**Fig. 4. Determine the fine-grained boundary of TIR**

### Homology-Based Other TE Searching （LINE and SINE）

For SINE/LINE detection, we found very low sensitivity and very high FDR, which is likely due to variation in these TEs (e.g., most LINEs are truncated upon insertion) and the lack of terminal repeats, making detection very challenging.[EDTA]

The LINE transposon usually forms TSD at the insertion site, but the reverse transcription process of LINE often terminates in advance when reverse transcription generates copies, resulting in frequent truncation at the 5' end. The autonomous LINE element usually has a polyA tail and at least one RT (reverse transcriptase) and nuclease for transposition. SINE is a non-autonomous transposon, which cannot transpose itself and relies on other transposon enzymes to express, such as RT in LINE. They originate from accidental reverse transcription of various polymerase III (Pol III) transcripts and have an internal Pol III promoter. SINE is short (80–500 bp), generates TSD (5–15 bp), and has polyA or polyT tails. The SINE superfamily was defined according to the "head" containing the Pol III promoter, and their origins were revealed: tRNA, 7SL RNA, and 5S RNA. [A unified classification system for eukaryotic transposable elements]

Generally, the method to identify LINE elements is to search for homologous RT domains in candidate sequences, while the determination of 5' usually requires manual judgment. Repbase, the gold standard TE database, retains only a few younger LINE elements. The identification tools of the SINE element are SINE\_Finder[8] and SINE\_Scan[9], where SINE\_Finder searches the A box and B box in TSD, the RNA polymerase III promoter, and the sequence rich in A/T at the 3 ends based on regular expression to identify the similar SINE sequence, while SINE\_Scan is an enhanced polisher of SINE\_Finder, which is used to find matches with tRNA, 7SLRNA, and 5SRNA, locate TSD and poly-A regions, and filter common false positives that match LTR/TIR results and/or do not repeat in the genome.

Since the structural characteristics of LINE and SINE transposons are very weak and the existing SINE element identification tools, such as SINE\_Scan and SINE\_Finder, cannot achieve the desired recognition effect, we identify LINE and SINE transposons based on the method of homology search.

1. We use scripts to extract known LINE and SINE transposons from the Dfam library of RepeatMasker 4.1.2 (the public TE database, freely available under the Creative Commons Zero ("CC0") license) to form the Other\_TE library.

2. Using the Blastn tool, search for and the TE library has repetitive sequences with high coincidence (the part of the alignment should account for more than 90% of the Other\_TE library sequence and the coarse boundary repetitive sequence at the same time).

### Filtering false positive

### Sequencing gaps and tandem repeat filtering

**Sequencing gaps:**

Gap sequences represent the most uncertainty of a genome assembly. Particularly, gaps in a repetitive sequence are more likely associated with misassembly [EDTA]. Sequences contain gaps more than 10 bp are excluded.

**Tandem repeat:**

Tandem Repeats Finder (TRF) [10]is used to identify tandem repeat 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 the extracted sequence, the sequence is considered a false positive 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,

1. we first use ltrsearch with the parameter "-i 0.85" to filter sequences that have long terminal structures. At the same time, the identified sequence with a 4-6 bp TSD and LTR structure is added to the LTR module for further filtering.

2. 将TIR候选比对到LTR cut上，分析比对结果。如果alignment\_len/query\_len>=0.95 and alignment\_len/subject\_len >= 0.95，说明这条TIR序列实际是一条LTR序列。

**A TE Copies-Based Filtering Method for Region Homology Outside the Boundary：**

False positive sequences, such as accidental terminal structures and TSD features on the genome, are easy to introduce using structure-based methods. Our method of filtering false positives is based on the following principles, as shown in Figure 4 (C):

① 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).

② The boundary of transposons determines the starting and ending positions of repeats, and the sequences outside the boundary should be regarded as random sequences. Therefore, the region outside the boundary of different copies of the same transposon should not have homology.

Based on the above principles ① ②, we expand the flanks of the copies of candidate TEs by 50 bp and then perform alignment between the flanked copies. If more than half of the copies have homology in the flanked 50 bp region, the candidate sequence is regarded as a false positive and filtered out. These candidate sequences are not true TEs, but rather a long repeat sequence with a TE-like structure. Since many LTR RTs do not exist or it is difficult to detect their full length copies, we have removed the limit of at least two occurrences of LTR identification.

### Nested TE treatment method:

Nested TE, usually formed by transposons inserted into other transposons, has a complex chimeric structure, so it is difficult for nested TE to form more than two full-length copies. Therefore, for transposons except for LTRs, our 识别的TE至少包含了两条全长拷贝支持, make it difficult to include nested TE.

LTR转座子相对其他转座子而言更长，也更容易被其他转座子插入，为了解开LTR内部序列中包含的nested TE，HiTE对LTR内部序列与生成的所有TE序列进行比对。对于nested TE而言，剔除其中的全长TE（为了保证去除nested TE的可靠性，这里定义全长TE为alignment length>=自身序列的95%，且比对的identity>= 95%），然后连接剩余序列，如果剩余序列长度小于100，则过滤掉这个序列，否则将剩余序列当做新的TE序列对待，这个过程会迭代三轮以解开多层嵌套的TE。

####Although we do not handle nested TE, the abundance of TE we recognize is unaffected. If the TE that forms the nested TE has an intact instance in other parts of the genome, our method can recognize it. In the most extreme case, there is only one intact instance of a TE in the genome, and other instances are inserted by other TEs, resulting in a large difference between the formed nested TE and itself. However, we can identify the copy of TE by the fault-tolerant mapping expansion algorithm. We can skip the gap generated by the insertion of another TE and identify the nested TE instance as another full-length copy of the TE. Since the minimum requirements for two copies have been met at this time, we can identify this TE.

### Generate a classified, non-redundant TE library

Before clustering, we divide the LTR-RT into 5' LTRs, 3' LTRs, and LTR internal regions, and then use the clustering tool CD-HIT[11] with the parameter "-aS 0.95 -aL 0.95  -c 0.8 -G 0 -g 1 -A 80" to generate a non-redundant TE library. At the same time, to determine the classification information of TE, we use the RepeatClassifier module in RepeatModeler2[12] to classify our non-redundant TE database so as to determine the category of each TE sequence.

## Results

1. 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 [Methodologies for the De novo Discovery of Transposable Element Families], 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.

3. 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 main data in Dfam comes from libraries generated by RepeatModeler, [The Dfam community resource of transposable element families, sequence models, and genome annotations]，and the great majority of Dfam submissions is currently housed in a non-curated section [109]。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 to the competitive evaluation of methods. For example, many evaluation 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.

4. An ideal method of evaluation should be able to consider both the integrity of TE sequences and the false positive rate of the TE library, which has yet to be developed [review]. To solve this problem, we take the evaluation methods from the latest study, EDTA and RepeatModeler2, which could produce ten metrics including Sensitivity, Specificity, Accuracy, Precision, FDR, F1, Perfect, Good, Present, and Not\_found. By combining the two complementary evaluation methods, we can accurately evaluate the integrity of TE families and the false-positive rate of the whole TE library.

### Selecting benchmarking model species

1.介绍数据集，并介绍每个数据集上的TE比例。

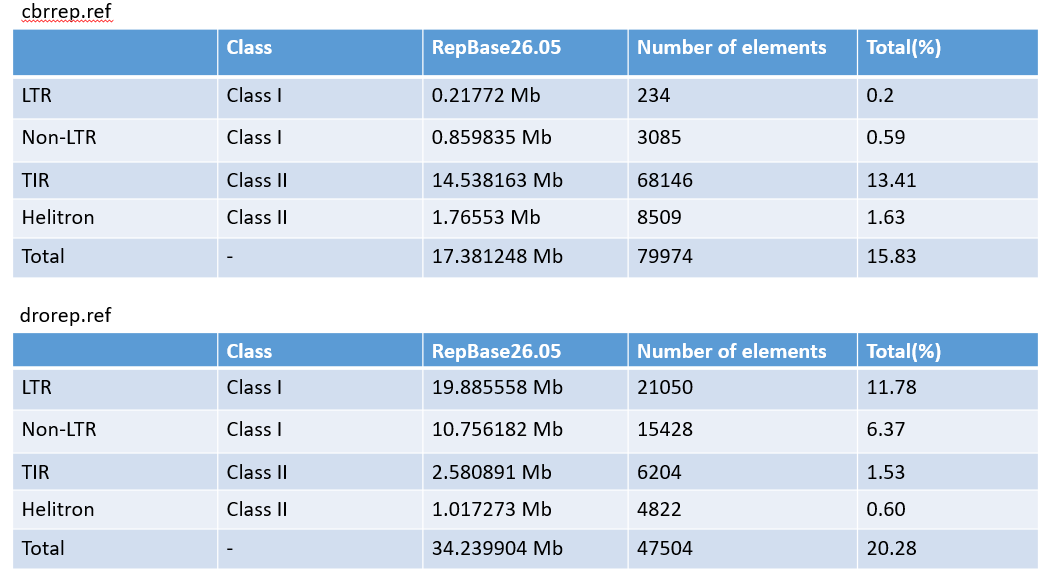
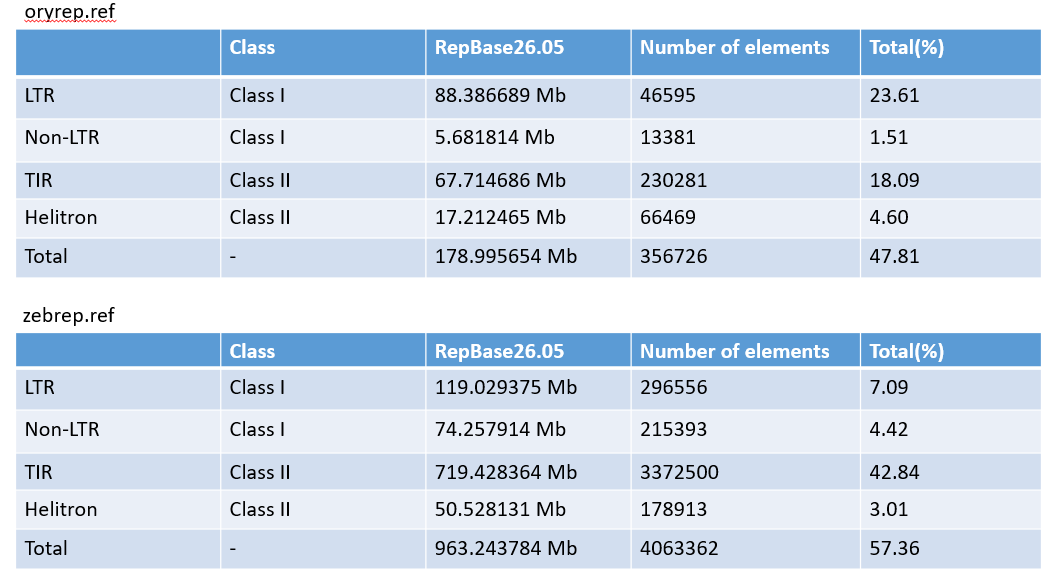
Despite the universality and importance of TEs in genomes, except for a few model species, the annotation and research of TEs in other species are still poor. In this benchmarking, we mainly focus on 4 typical species, Oryza sativa, Caenorhabditis briggsae, Drosophila melanogaster, and Danio rerio, whose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

2.介绍repbase以及每个物种的repbase占基因组的比例(对数据集的描述)

Repbase Update (RU) is a database of representative repeat sequences in eukaryotic genomes, which has a long history of TE discovery and annotation since 1992. RU has been serving as a manually curated reference database fundamental for almost all eukaryotic genome sequence analyses. Thus, we used RepBase26.05 as the gold standard TE libraries for the O. sativa, C. briggsae, D. melanogaster and D. rerio.

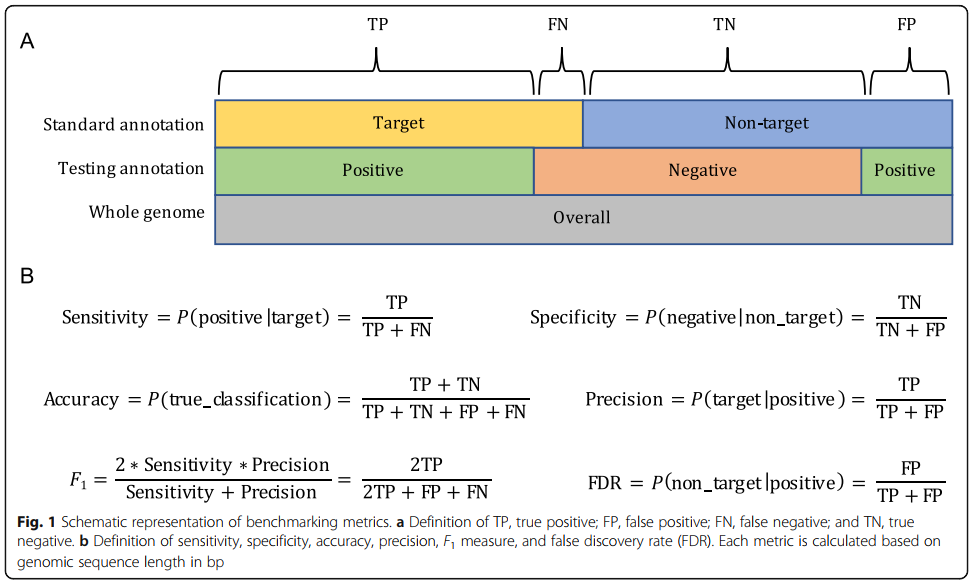
To evaluate each program, we used a high-quality, manually curated library developed for the model species Oryza sativa (rice), which has a long history of TE discovery and annotation

Table 1 TE content in the rice (Oryza sativa ssp. japonica cv. “Nipponbare” v. MSU7) genome



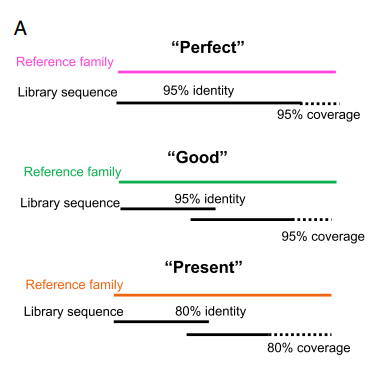
### Setting up benchmarking approaches for TE library evaluation

1. To fairly and comprehensively measure the quality of TE libraries generated by different TE identification tools, we use the evaluation methods from the latest study, EDTA and RepeatModeler2. For convenience, we hereafter refer to the evaluation methods of EDTA and RepeatModeler2 as EM\_EDTA and EM\_RM2, respectively. As shown in Fig. 5A, the EM\_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 [EDTA]. The advantage of this method is that it is clear, and more importantly, it can reflect the false positive rate, which is very common in many TE identification methods. However, the disadvantage is that it cannot reflect the integrity of the TE models. All general repeat identification programs, which depend on sequence repeatedness, performed well [EDTA], even though the sequences identified by these tools are only fragments with unclear boundaries. For example, a 1 kbp intact TE sequence and ten 100 bp fragments may obtain the same performance, but from the perspective of TE integrity and biological significance, the former is obviously more valuable.



**Fig. 5.** The benchmarking method of EDTA

As shown in Fig. 5B, the EM\_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". It 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 reflect the false-positive rate of the tested TE library. By combining the two complementary evaluation methods, we can accurately evaluate the integrity of TE families and the false-positive rate of the whole TE library.



**图6** The benchmarking method of RepeatModeler2

我们结果能识别更多的Perfect数量，但是Sensitivity并没有很高，证明我们能够识别很多低拷贝TE，这是很多工具，如RepeatModeler2所不能做到的。

### Comparison of general-purpose repeat annotators

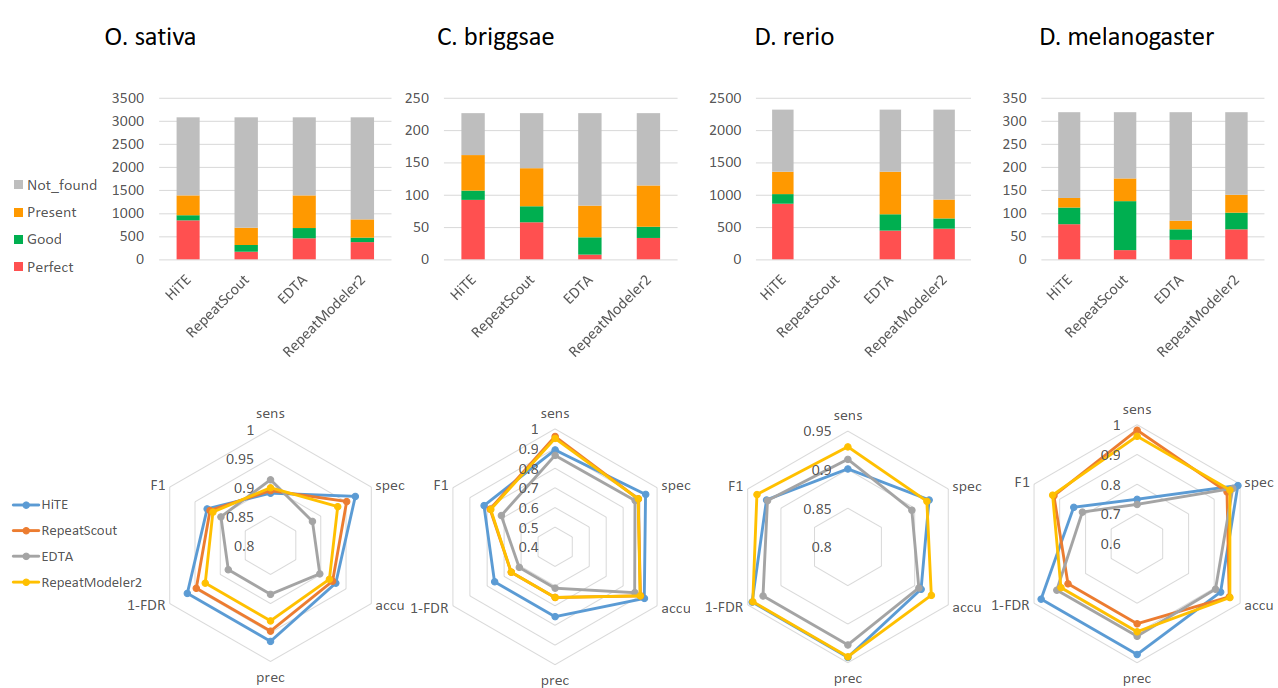
1. We compared HiTE with the other three mainstream general-purpose repeat annotators, including RepeatScout, EDTA, and RepeatModeler2. Among them, EDTA is the best annotation tool based on structure, and RepeatModeler2 is the pipeline with the best overall performance. RepeatScout was originally used 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.[综述] 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 [RepeatModeler2].

2. The results of BM\_RM2 are shown in the upper part of Fig. 2. RepeatModeler2 is stable on all datasets; the performance of EDTA is unstable, obtaining a high number of perfect models on the TE-rich genomes (O. sativa and D. rerio)[refer to table] but a low number on the other genomes. However, 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. Because it cannot process more than 1 GB of genomes, it has no results for D. rerio. Among these tools, HiTE has a much higher number of perfects than other tools on all datasets and a smaller number of good and present tools. This 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. We made an explanation in Section "Conclusions".

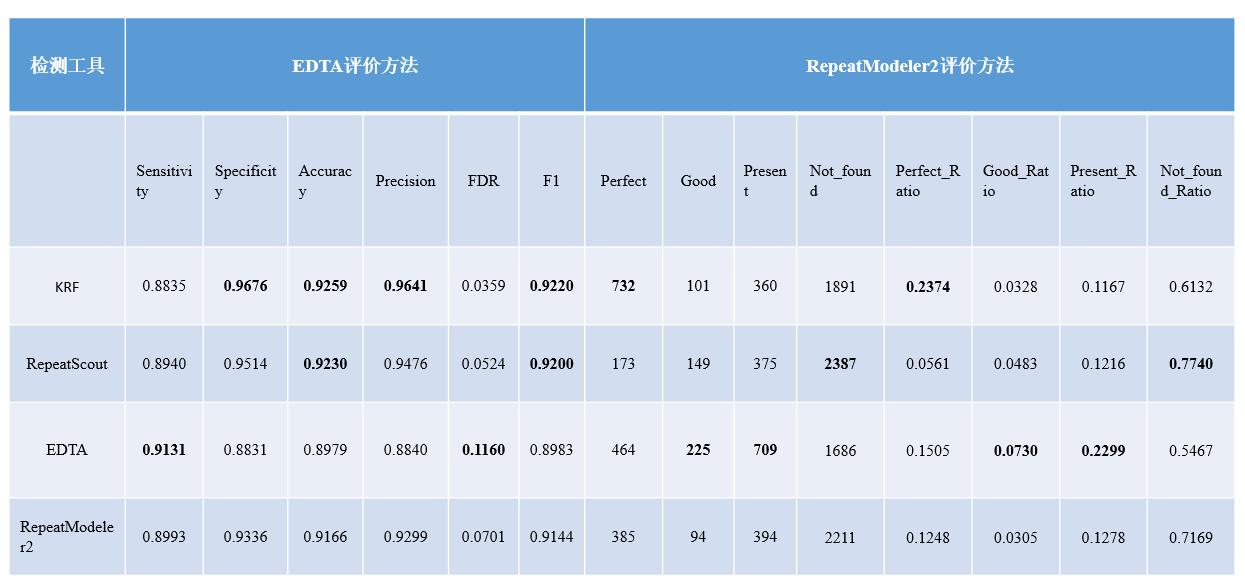
3. The benchmarking results using BM\_EDTA are shown in the lower part of Fig. 2. We noticed that RepeatScout and RepeatModeler2 both achieved relatively consistent high performance, which verified that "all general repeat identification programs, which depend on sequence repeats, performed well" as described in the EDTA article [EDTA]. As we described in the previous section, the greatest advantage of the BM\_EDTA method 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 perfects, indicating that there are a large number of fragments, but it has a very high BM-EDTA performance. However, we noticed that on all data sets, HiTE showed significantly higher precision indicators, 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 evaluation method of 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 repeatability of the sequence, so many short and incomplete TE models will also be identified, and these fragments are filtered out in HiTE because they do not meet our goal of looking for full-length TE.



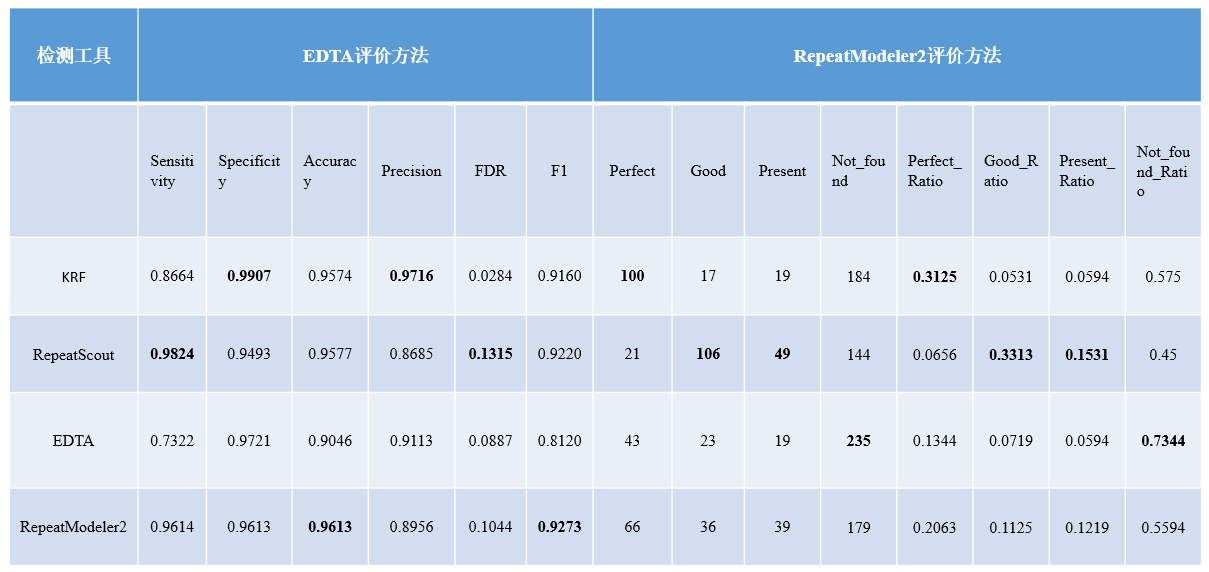
**表3** 基于O. sativa基因组的测试结果



**表1** 基于C. briggsae基因组的测试结果



**表2** 基于D. melanogaster基因组的测试结果



**表4** 基于D. rerio基因组的测试结果



3. 对我们结果的Helitron转座子、LTR转座子、DNA/CMC-EnSpm，弄个Logo；

<https://weblogo.threeplusone.com/create.cgi>

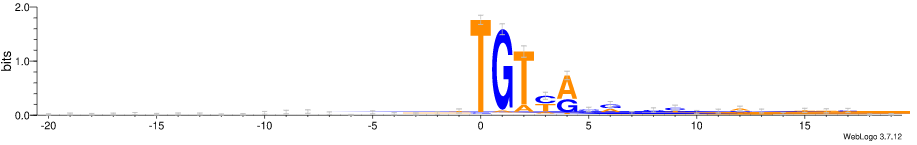
draw\_TE\_log.py 设置成50bp

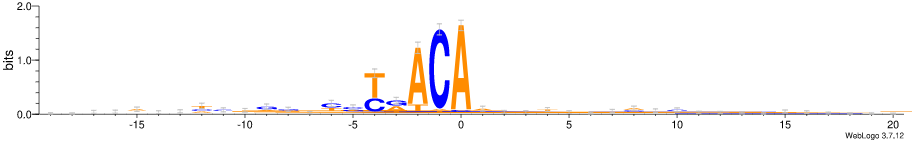
从这个logo图中，我们能看到某类转座子的特性，保守的motif，插入到AT rich区域。

取序列的终端位置的前后flanking区域，形成logo。

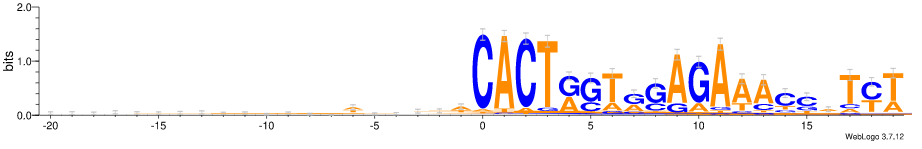
引用[WebLogo: A sequence logo generator]

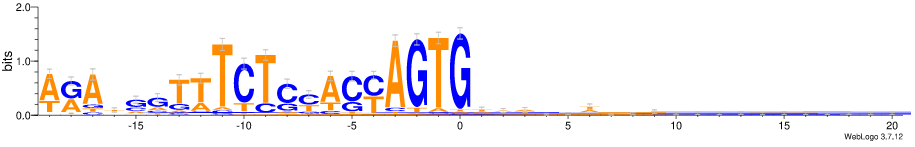
LTR-logos





EnSpm-logos





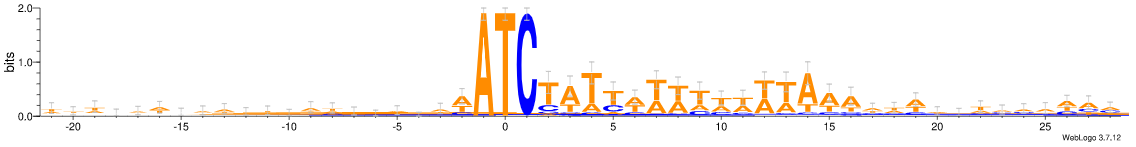
Helitron

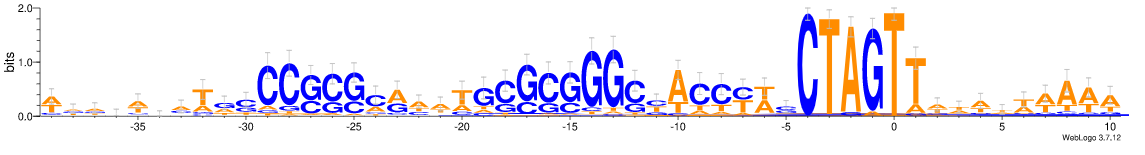
AT靶二核苷酸是Helitron元素高度优选的整合位点，In addition to the central AT dinucleotide, we observe a strong preference for an AT-rich DNA sequence within ∼20 bp around the actual integration site; this preference is the most pronounced towards sequences flanking the 3′-end of the integrated transposon[A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes]

观察到Helitron更倾向插入AT位点，且在3’端末尾AT rich，且在-29 to -13位点之间，是一个hairpin loop。

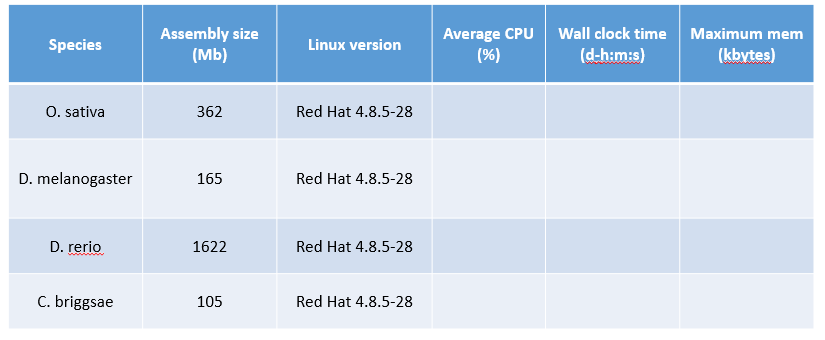
consistent with previous observations by Yang and Bennetzen regarding target site preference [64]. We found enriched CG content at the 3′ terminals especially at the - 13 and - 14 positions, which could produce a hairpin loop, a canonical Helitron feature.

Helitron-logos





4.运行时间、内存



### Comparison of TIR annotators

1.简介TIR的结构、识别难点（TIR+ TSD结构很短，容易有假阳性）。

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 [A unified classification system for eukaryotic transposable elements]. TIR TEs may contribute to genome evolution by generating allelic diversity, inducing structural variation, and regulating gene expression. [TIR-learner].

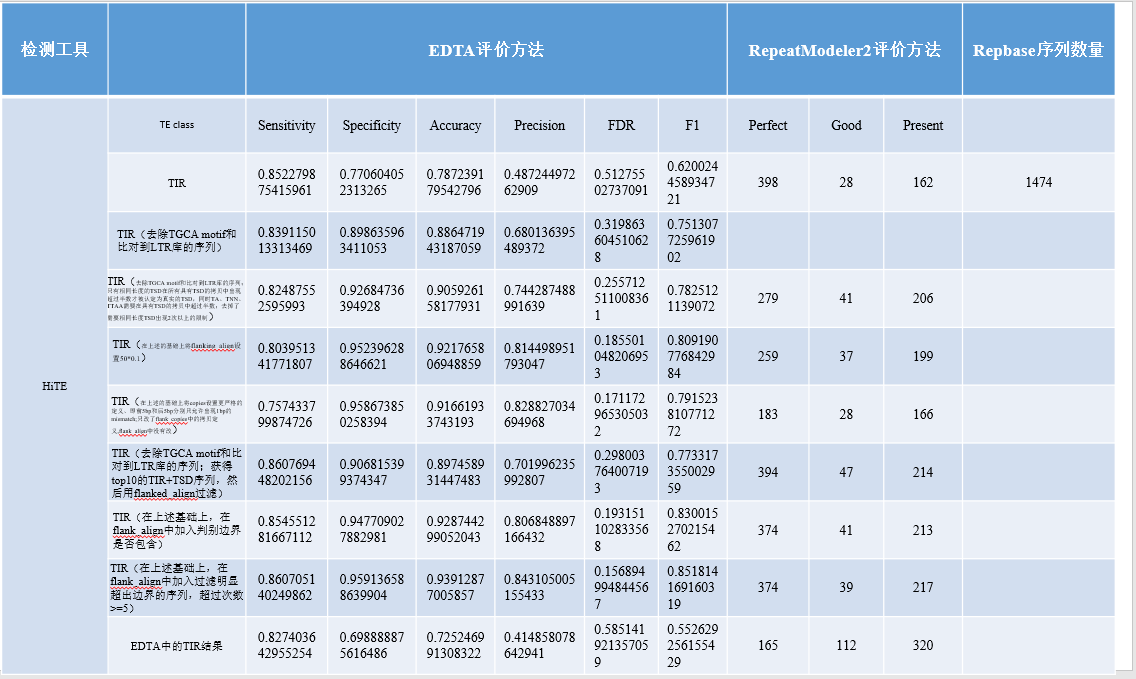
TIR TEs are divided into nine known superfamilies by the distinguished TIR sequences and the TSD size (usually 2–11 bp). However, due to the short terminal inverted repeat sequences, TIR TE identification and annotation are quite challenging. For example, members of the hAT superfamily have TSDs of 8 bp and relatively short TIRs of 5–27 bp [The hAT family: a versatile transmission group common to plants, fungi, animals, and man].

Many tools have been designed for their identification, such as IRF [54], TIRvish [55], TIR Learner [17], and GRF [45], which identify TIR elements by structural signals and are comprehensively evaluated in EDTA. Unfortunately, due to the short structural characteristics of TIR, this method will 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, EDTA's TIR module (GRF+TIRlearner) has demonstrated great promise for structural annotation and achieved higher performance than other tools [EDTA]. 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).

3. As shown in Table 5, according to the benchmarking results of the BM\_EDTA, our method has higher sensitivity, specificity, accuracy, precision, F1, and a lower FDR than EDTA. 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. These two benchmarking methods both demonstrate that our method can achieve high-precision identification of TIR TEs.

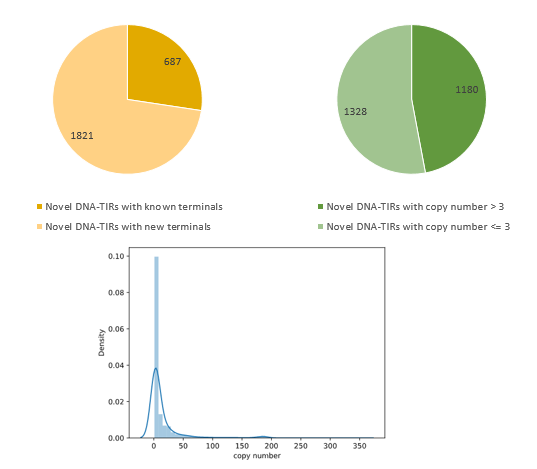
However, we do observe some losses of real TIR TEs, which are mainly caused by the following reasons:

**表5** 基于O. sativa基因组的TIR测试结果



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 [review]. 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 (see the copy distribution diagram). At the same time, nearly half of the sequences in the new DNA TIRs have more than 3 copies (see figure), suggesting that these are like real TEs that were not included in the Repbase library due to their low number of copies.

In addition, we recognize that some TIR TEs have TIRs similar to the known TIRs in Repbase (see figure), which are likely to be non-autonomous TIR TEs.



Verification of novel DNA-TIR candidates identified by HiTE.

### Comparison of Helitron annotators



1.介绍Helitron。

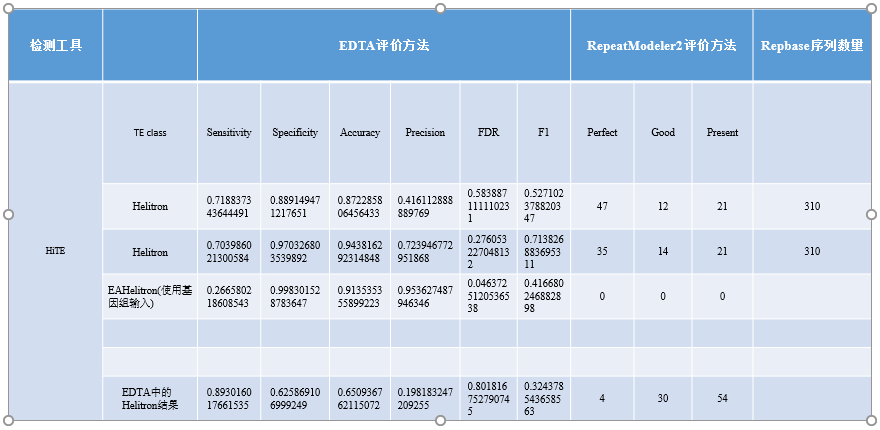
Helitrons are a subclass of DNA transposons, which replicate through the rolling circle mechanism. When replicating themselves, only the single strand of DNA is broken, and no TSD is generated, which is different from the other TEs. The Helitron transposon has a 5'-TC...-CTRR-3' conserved structure, 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 a flanking 5'-A and 3'-T nucleus. [Helitrons on a roll: eukaryotic rolling circle transmissions]. The weak structural signals of Helitrons make identification of these elements particularly challenging.

As far as we know, there are only two tools, HelitronScanner and EAHelitron, that can produce useful Helitron predictions. HelitronScanner [45] identifies the sequence patterns in Helitron transposons using the local combinational variable (LCV) algorithm, which produced a large number of candidate sequences, most of which are false positives. For example, 52 Mb of raw candidate sequences in rice，covering 13.9% of the genome, which is obviously exceeds the real coverage. EDTA filters the results of HelitronScanner, greatly improving its specificity and accuracy without reducing its sensitivity [EDTA]. Nevertheless, the precision of the Helitron identification module of EDTA is still very low (Fig), which is far from satisfactory.

We also test the other tool, EAHelitron, which identifies Helitrons based on the conservative structure traits using regular expression (RE), such as the 5’ terminal with TC, the 3’ terminal with CTAGt, and a GC-rich hairpin loop before 2–10 nt of CTAG. The performance of EAHelitron is primarily determined by the pre-defined patterns of hairpin loop regular expressions. We observed that it lost some of the hairpin loop patterns of real Helitrons. For example, many real Helitrons in C. briggsae cannot be discovered until we manually add a new pattern of 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 starting position of Helitron. We take the first 5'-TC closest to CTAGt-3' as the starting position 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 (Fig. 0.9536) but the lowest sensitivity (Fig. 0.26658). Moreover, it cannot identify any gold standard models according to the BM\_RM2 (Fig).

To discover the intact Helitron elements, we have developed a new Helitron identification method. This method is based on the coarse boundary TE candidates output by the FMEA algorithm, and then EAHelitron is used to locate the accurate 3'-CTRR and the hairpin loop structure. 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 the candidates that were not inserted into AT target sites. Finally, we obtained the confident candidates based on the TE copy-based filtering method for region homology outside the boundaries (see the ``Methods" section).

The experimental results show that our Helitron identification method has the highest performance (Fig). Our Helitron identification method 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 suspect that our sensitivity would be greatly improved once EAHelitron can include a more comprehensive hairpin loop pattern. At the same time, we identified more perfect Helitrons in the gold standard dataset. However, we do notice that our results are still affected by false positives, which indicates that our method has potential for improvement.



### Comparison of LTR annotators

Long terminal repeat retrotransposons (LTR-RTs) (Fig. 1) 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 [49], GRF, LTR\_STRUC [46], LTR\_FINDER [47], LTRharvest [48], LtrDetector [50], and LTR\_retriever [40]. It is worth noting that LTR\_retriever was designed as a stringent filtering method for raw results from other LTR programs 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) [EDTA], 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, so we integrated LTR\_retriever into HiTE.

### Comparison of non-LTR annotators

Non-LTR retrotransposons include two types of TE: LINEs and SINEs [zhao2016makes]. 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) [wicker2007unified]. They do not encode any reverse transcriptase protein and rely on other TEs to transition, especially LINEs [dewannieux2003line]. The weak signals of non-LTR retrotransposons make them quite challenging to identify [mao2017sine\_scan].

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

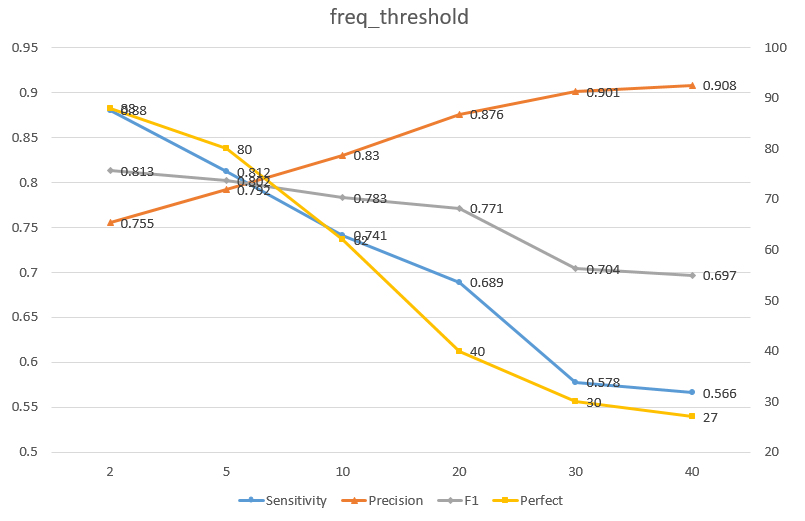
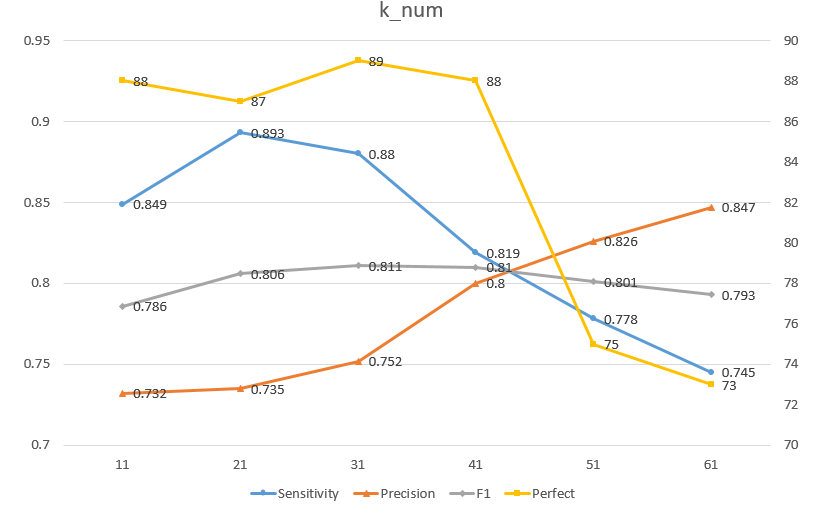
To understand how the parameters in HiTE affect the results, we selected the four most important parameters for testing: k\_num, freq\_threshold, chunk\_size, and flanking\_len. k\_num is the size of k-mer, freq\_threshold refers to the frequency threshold of k-mer, and chunk\_size refers to cutting the genome into blocks of the same size. 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.

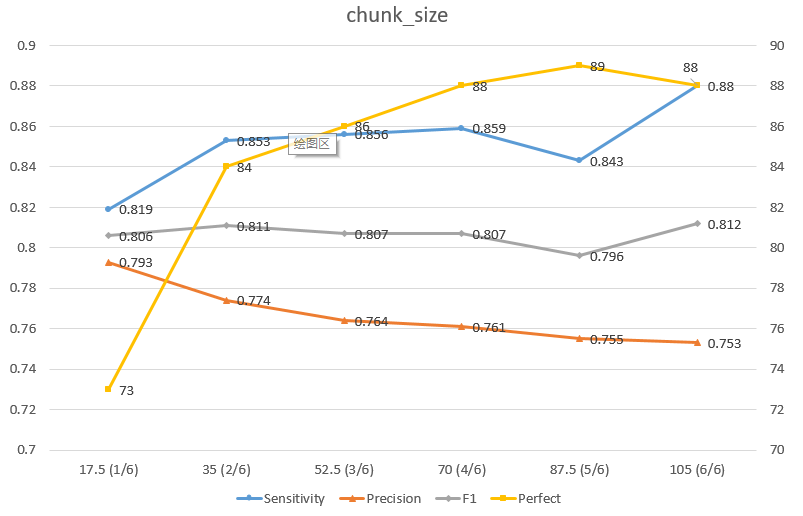
As shown in Figure 1, the smallest k\_num (such as 11) will mark the most parts of the genome as repeat regions, which cannot effectively distinguish TE from non-TE, resulting in low sensitivity and precision. Excessive k\_num will lose part of the true TE (lower sensitivity), but the sequences it identifies 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.

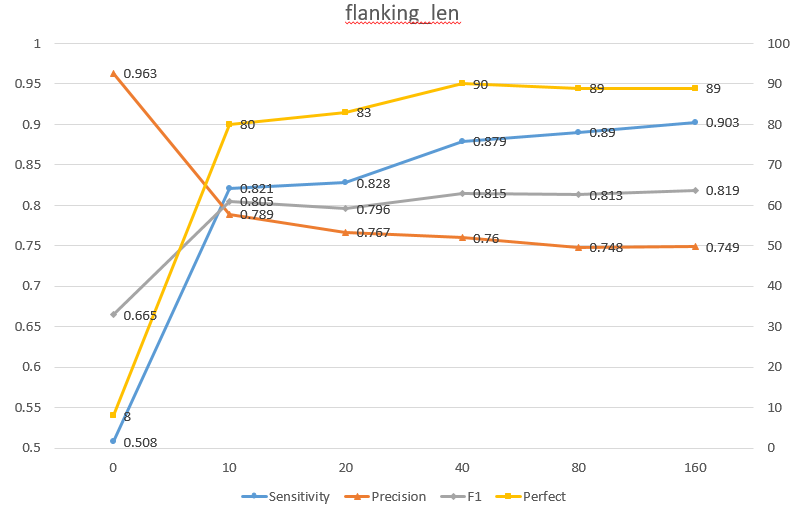
As shown in Figure 2, with the freq\_threshold increased, all metrics except precision decreased significantly, which indicates that the higher the frequency of k-mer in the sequence, the more likely the sequence is to be a real TE.

As shown in Figure 3, 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 Figure 4, 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 flanging\_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.







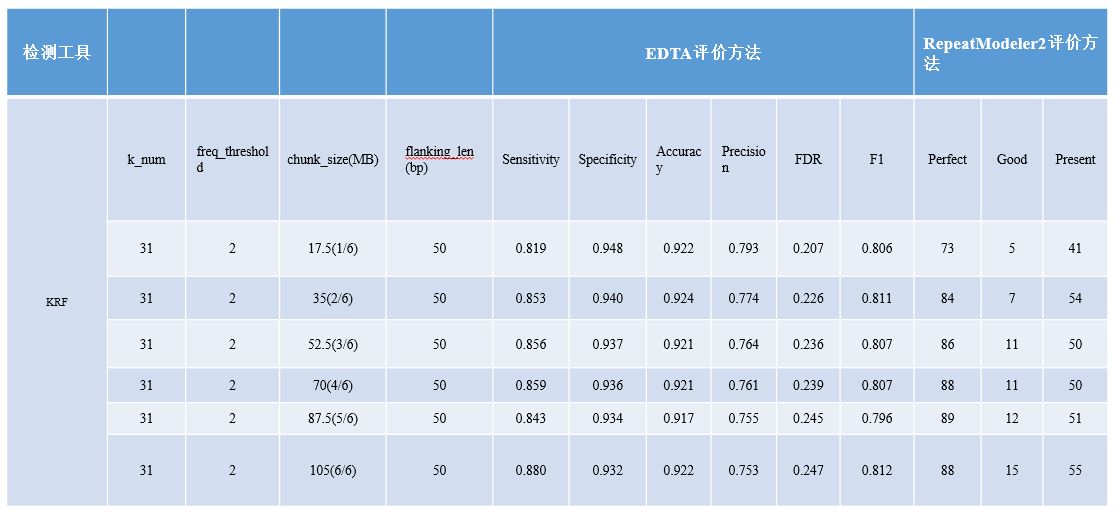
**Table 5** Influence of **k\_num** parameter on results



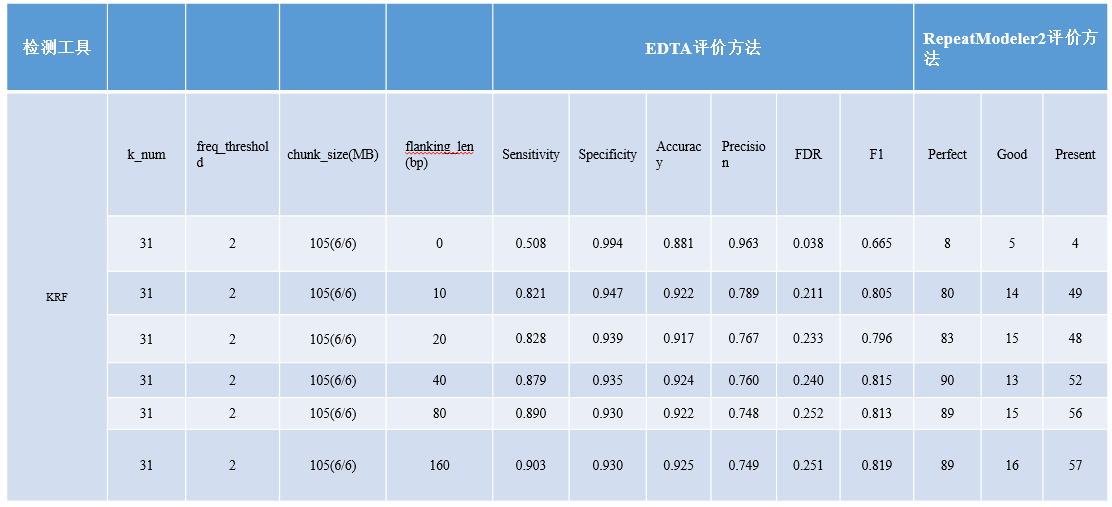
**Table 6** Influence of **freq\_threshold** parameter on results



**Table 7** Influence of **chunk\_size** parameter on results



**Table 8** Influence of **flanking\_len** parameter on results



# 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\cite{Smit2008repeatmodeler}, always obtain uncertain boundaries, and intensive manual repairs are required to enable them to be saved in the cured library\cite{storer2022methodologies}. 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.

Based on the above test results, we can draw the following conclusions:

1. The HiTE has obtained the highest specificity, precision, and lowest FDR on all datasets, indicating that the TE identified by the HiTE is very accurate. But at the same time, because HiTE has filtered out many false positive sequences, its sensitivity can only remain at a moderate level.

2. On all datasets, the HiTE obtained the most perfect sequences while keeping the lower good and present sequences, indicating that the HiTE can recover the gold standard TE sequences and the fragmentation of the results is very low.

3.

3.1 RepeatModeler2 is a program that obtains the best results except for HiTE and can maintain a relatively stable result in most species;

3.2 RepeatScout output is highly fragmented, frequently with low precision, and has the least perfect sequence;

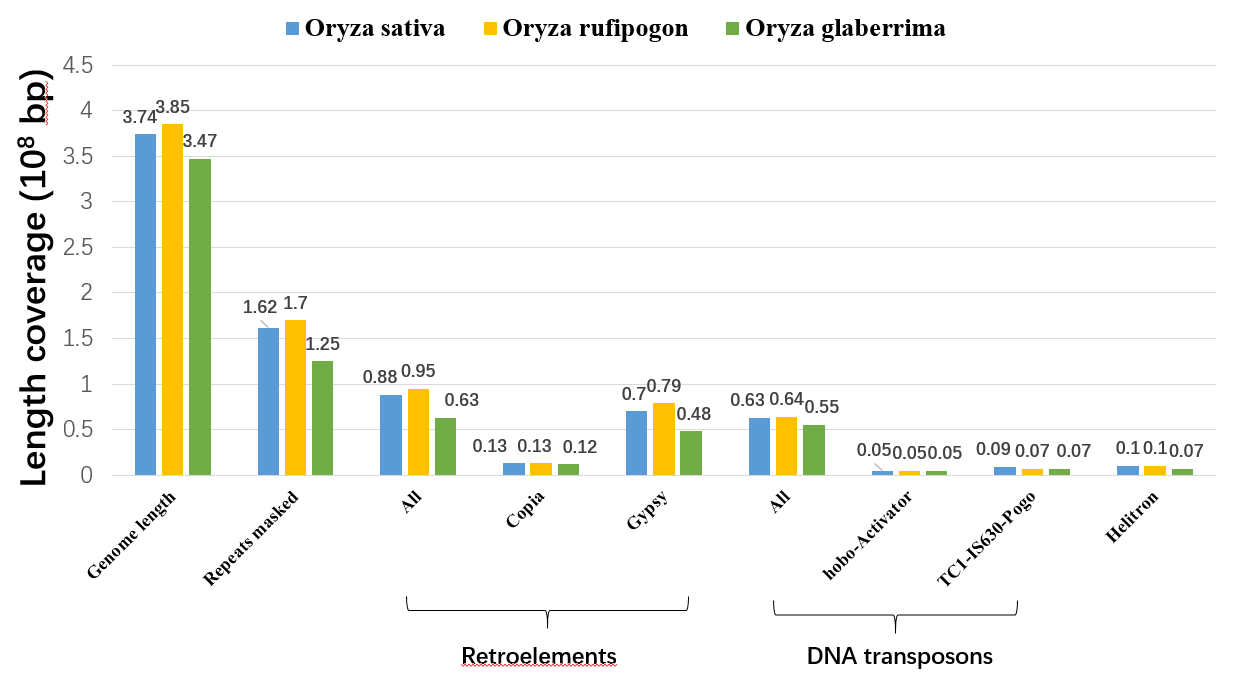
3.3 EDTA can not recognize SINE, LINE, and other types of transposons, so its sensitivity in Drosophila species is low. EDTA obtained the highest sensitivity in rice (almost no SINE and LINE), but also the lowest precision, indicating that there is a high rate of false positives. It was observed that EDTA obtained a significantly higher number of presents on rice and zebrafish, indicating that there were a large number of fragmented sequences.

# 应用案例

## Contribution of TEs to genome size.

“*Long Terminal Repeat retrotransposons are the major component of the Oryza genomes analyzed and, along with polyploidization, are the most important contributors to the genome size variation across the Oryza genus. Two families of Ty3-gypsy elements (RIRE2 and Atlantys) account for a significant portion of the genome size variations present in the Oryza genus.*” 即转座子**Ty3-gypsy**是造成水稻属**基因组大小变化**的主要因素，这一结论引用自文献*Zuccolo, Andrea, et al. "Transposable element distribution, abundance and role in genome size variation in the genus Oryza." BMC Evolutionary Biology 7.1 (2007): 1-15.*

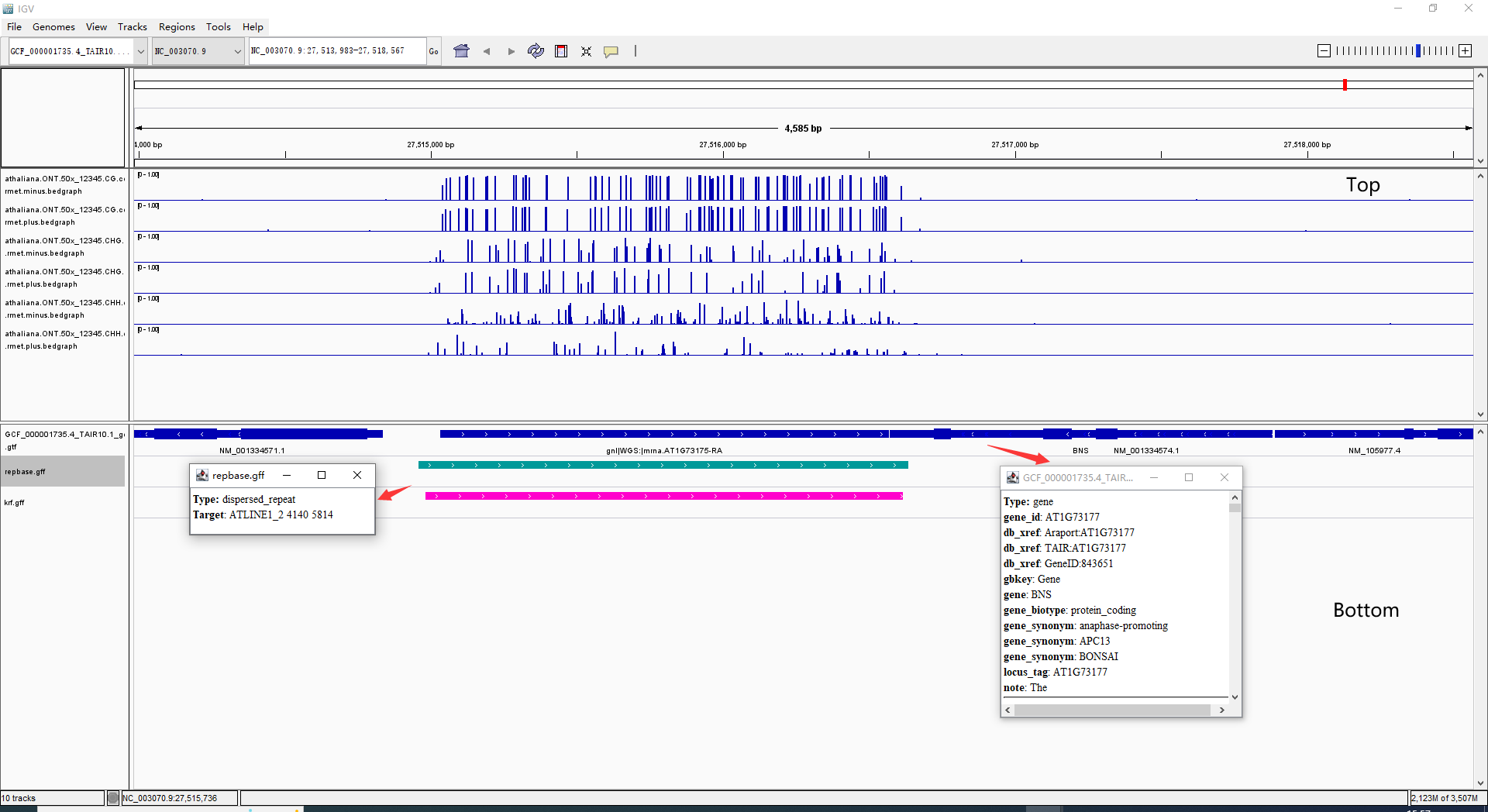
通过将HiTE应用到几种常见的水稻亚种，Oryza sativa、Oryza rufipogon、Oryza glaberrima上，我们验证了这个结论。首先我们从NCBI上下载得到三个水稻亚种的基因组，然后将其输入到HiTE中，使用默认参数，生成相应的TE library，并得到TE在基因组上的覆盖情况。如图10所示，三个不同的水稻亚种的基因组大小有明显的差异，而这种差异主要来自于Gypsy，这也说明了以Gypsy为代表的LTR转座子是水稻基因组产生变化的主要因素。



**图10** 三个水稻亚种上的TE覆盖基因组长度

## TE影响基因表达

有许多研究表明，转座子可以通过插入到基因内部或附近，从而改变基因的表达。例如 “*Hypermethylation of a LINE element can lead to silencing of a nearby Anaphase-Promoting Complex (APC) 13 gene in* Arabidopsis”，引用自文献*Saze, Hidetoshi, and Tetsuji Kakutani. "Heritable epigenetic mutation of a transposon‐flanked Arabidopsis gene due to lack of the chromatin‐remodeling factor DDM1." The EMBO journal 26.15 (2007): 3641-3652.*

经测试，HiTE与RepBase一样能被用来验证这一结论。首先通过查询NCBI，我们知道APC 13 gene在拟南芥基因组的Locus tag为AT1G73177，随后我们将基因组文件、RepBase library、HiTE TE library以及甲基化数据输入到IGV中，观察到如图11的现象，其中Top表示三种类型（CG、CHG、CHH）的甲基化数据；Bottom表示基因与TE注释数据，蓝色是基因注释文件，绿色和粉红色分别是RepBase与HiTE对应的TE注释数据。从图11中我们可以看到，RepBase与HiTE均注释出了基因APC 13附近的LINE-L1元素，且该元素被高度甲基化。

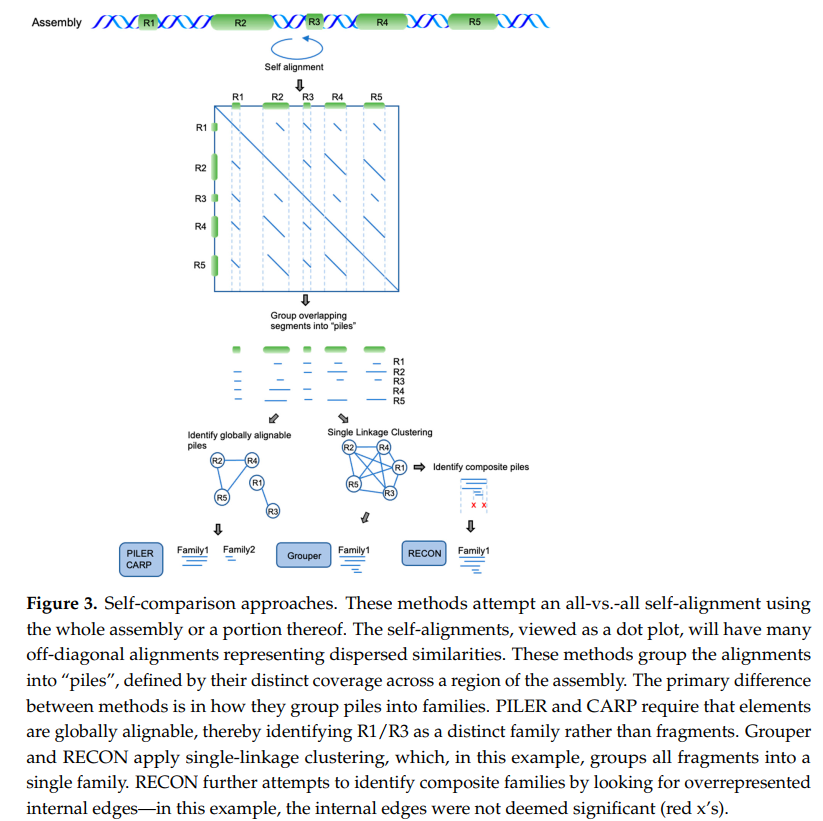
**图11** A hypermethylation LINE element nearby the Anaphase-Promoting Complex (APC) 13 gene in Arabidopsis

The methylation data in this example is obtained from wild type Arabidopsis thaliana (L.) Heynh. Columbia-0 (Col-0) sample processed by deepsignal-plant.

*Ni, Peng, et al. "Genome-wide detection of cytosine methylations in plant from Nanopore data using deep learning." Nature communications 12.1 (2021): 1-11.*

# 主流的重复识别方法原理图

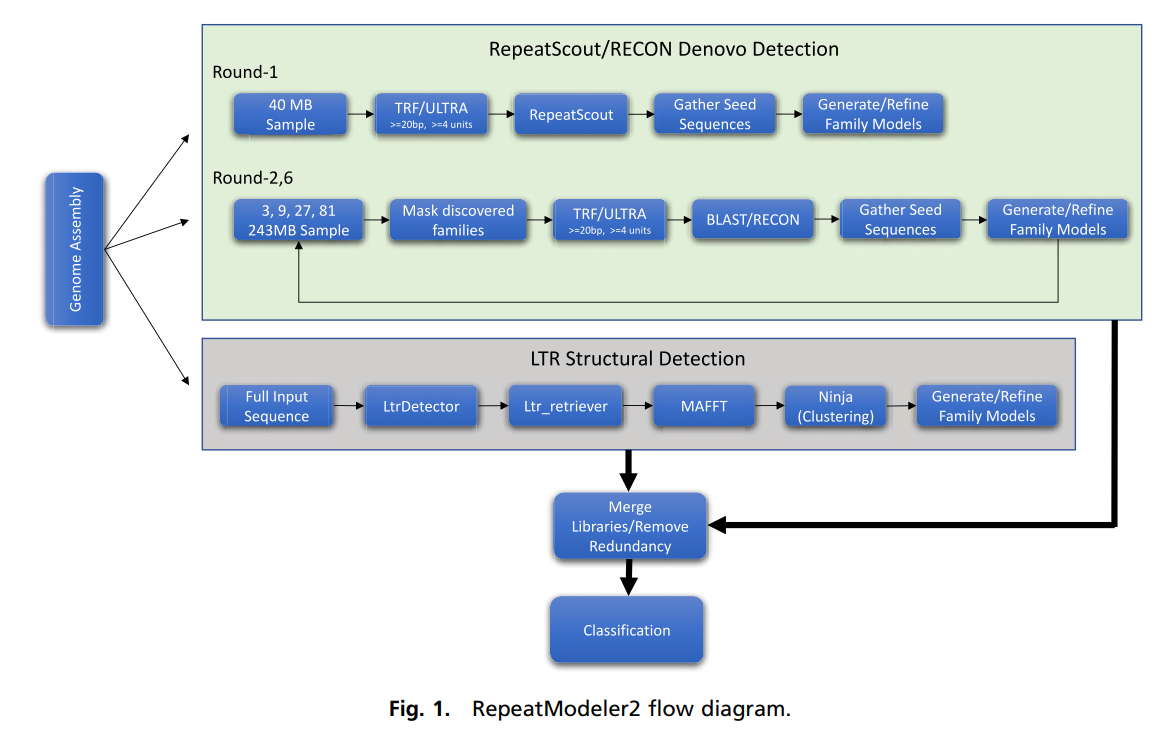
## 基于自比对的方法（RECON）



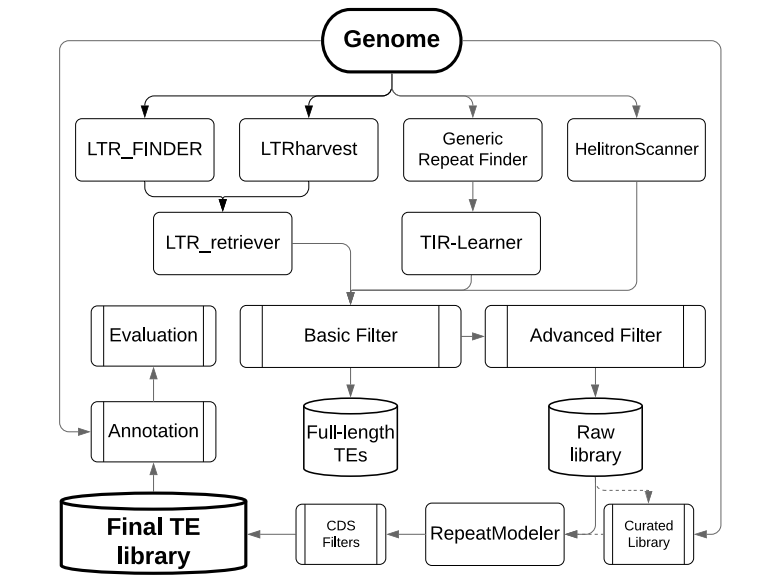
## RepeatScout方法

Price et al., 2005提出了RepeatScout，它以一系列高频的kmer为一致性种子，然后围绕一致性种子开展多次对齐的迭代扩展，直至生成一致性序列。它的优势是速度快，不需要进行pairwise比对。缺点是RepeatScout只适合找到low-divergence的重复家族，因此它会找到许多duplication和TE fragments。

## RepeatModeler2方法



## EDTA方法



# 参考文献

[1] Bao W D, Kojima K K, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes [J]. Mobile DNA-Uk, 2015, 6:

[2] Wicker T, Sabot F, Hua-Van A, et al. A unified classification system for eukaryotic transposable elements [J]. Nat Rev Genet, 2007, 8(12): 973-982.

[3] Ou S, Jiang N. LTR\_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons [J]. Plant Physiol, 2018, 176(2): 1410-1422.

[4] Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons [J]. BMC Bioinformatics, 2008, 9: 18.

[5] Shi J, Liang C. Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection [J]. Plant Physiol, 2019, 180(4): 1803-1815.

[6] 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 [J]. Mol Plant, 2019, 12(3): 447-460.

[7] Xiong W, He L, Lai J, et al. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes [J]. Proc Natl Acad Sci U S A, 2014, 111(28): 10263-10268.

[8] Wenke T, Döbel T, Sörensen T R, et al. Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes [J]. The Plant Cell, 2011, 23(9): 3117-3128.

[9] Mao H, Wang H. SINE\_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets [J]. Bioinformatics, 2017, 33(5): 743-745.

[10] Benson G. Tandem repeats finder: a program to analyze DNA sequences [J]. Nucleic Acids Res, 1999, 27(2): 573-580.

[11] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences [J]. Bioinformatics, 2006, 22(13): 1658-1659.

[12] Flynn J M, Hubley R, Goubert C, et al. RepeatModeler2 for automated genomic discovery of transposable element families [J]. Proc Natl Acad Sci U S A, 2020, 117(17): 9451-9457.

[13] Ou S, Su W, Liao Y, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline [J]. Genome Biol, 2019, 20(1): 275.