Review



# Introduction of Plant Transposon Annotation for Beginners

Dongying Gao *

Posted Date: 10 November 2023

doi: 10.20944/preprints202311.0673.v1

Keywords: Transposon annotation; plant; genome; bioinformatics pipeline; database

*Review*

# Introduction of Plant Transposon Annotation for Beginners

**Dongying Gao**

Small Grains and Potato Germplasm Research Unit, USDA-ARS, Aberdeen, ID 83210, USA; dongying.gao@usda.gov

**Simple Summary:** Transposons are the most abundant repeats in plant genomes, and many of them can produce transcripts and encode proteins that may result in overestimating and incorrectly annotating functional genes. Thus, accurate transposon annotation is essential for all plant genome sequencing projects and other research. Although numerous tools have been developed, it is still challenging to annotate plant transposons for most scientists. The aims of this review are to introduce the basic knowledge about plant transposons and to provide a beginner's guide on plant transposon annotation. I accentuate the importance of transposons and summarize the general strategies for transposon annotation. I briefly introduce the unique features of different transposon superfamilies in plants and the related resources for annotating plant transposons. I further present the information on improving the quality of transposon annotation. The challenges and future prospects for plant transposon annotation are also discussed.

**Abstract:** Transposons are mobile DNA sequences that contribute large fractions of many plant genomes. They provide exclusive resources for tracking gene and genome evolution and for developing molecular tools for basic and applied research. Despite extensive efforts, it is still challenging to accurately annotate transposons, especially for the beginners as transposon prediction requires necessary expertise in both transposon biology and bioinformatics. Moreover, the complexity of plant genomes and the dynamic evolution of transposons also bring difficulties for genome-wide transposon discovery. This review summarizes the three major strategies for transposon detections including repeat-based, structure-based, and homology-based annotation, and introduces the transposon superfamilies identified in plants thus far, and some related bioinformatics resources for detecting plant transposons. Furthermore, it describes the transposon classification and explains why the terms of 'autonomous' and 'non-autonomous' cannot be used to classify the superfamilies of transposons. Last, this review also discusses how to identify misannotated transposons and improve the quality of transposon database. This review provides helpful information about plant transposons and a beginner's guide on annotating these repetitive sequences.

**Keywords:** transposon annotation; plant; genome; bioinformatics pipeline; database

## 1. Introduction of plant transposons

Transposons or transposable elements (TEs) are genomic sequences that have the potential to change their positions in host genome. According to their transposition mechanisms, transposons are grouped into two major classes: Class I elements or retrotransposons which move via a copy-and-paste model and class II elements or DNA transposons that transpose via a cut-and-paste model, rolling-circle replication, or other mechanisms [1,2]. Each transposon class can be further divided into different superfamilies based on their sequence structures and the encoded proteins. Thus far, over 30 transposon superfamilies have been identified in prokaryotic and eukaryotic genomes [1–6]. However, only 16 superfamilies were found in the plant kingdom (Figure 1), including two superfamilies of long terminal repeat (LTR) retrotransposons, three superfamilies of long

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2023 doi:10.20944/preprints202311.0673.v1



interspersed nuclear elements (LINEs), one superfamily of short interspersed nuclear elements (SINEs) and nine DNA transposon superfamilies. Additionally, many plant genomes harbor endogenous plant pararetroviruses (EPRVs) which share similar core genes with LTR retroelements for reverse transcription but lack both functional integrase (INT) and LTRs [7,8]. Numerous elements still can be recognized as TEs as they show some typical features of transposons such as LTRs, terminal inverted repeats (TIRs), and the flanking target site duplications (TSDs). However, they do not encode transposase (TPase) proteins and are difficult to be grouped. These elements would usually be considered as unclassified transposons that include large retrotransposon derivatives (LARDs) [9], terminal-repeat retrotransposons in miniatures (TRIMs) [10,11], and miniature inverted-repeat transposable elements (MITEs) [12]. Notably, with the advance of sequencing technologies and related bioinformatic software, some novel types of transposons may be identified in the existing and/or newly sequenced plant genomes.

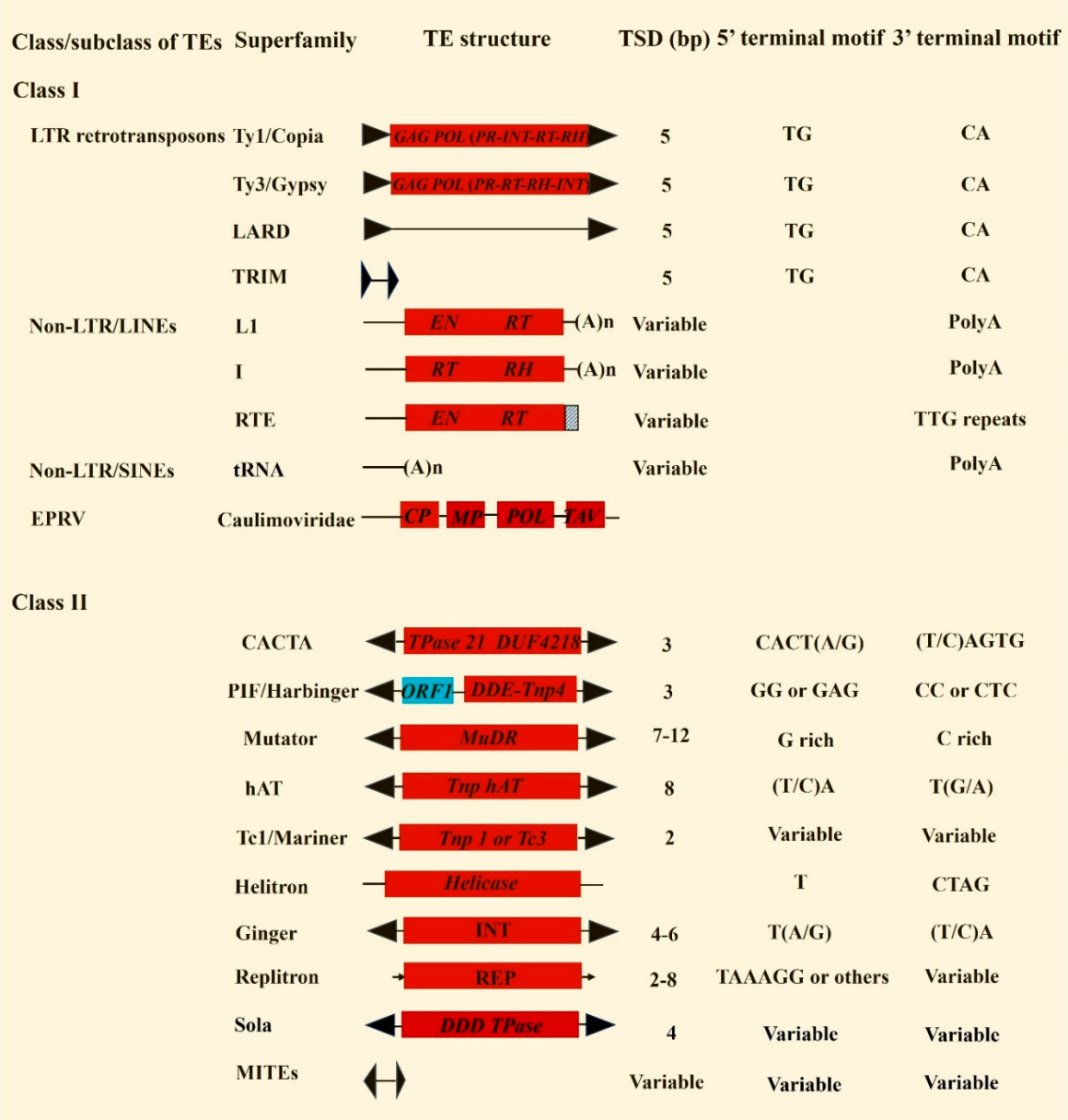


**Figure 1. Structures and typical features of different types of plant transposons.** GAG, group-specific antigen; POL, polyprotein; PR, protease; RT, reverse transcriptase; RH, RNAase-H; INT, integrase; EN, endonuclease; CP, coat protein; MP, movement protein; TAV, transactivation protein; TPase, transposase; REP, replication protein. .

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2023 doi:10.20944/preprints202311.0673.v1



Transposons are important contributors for plant gene and genome evolution as their movements may alter gene expression and regulatory networks and result in the chromosome rearrangements [13]. They can create novel genetic and morphological variations that are beneficial for host's fitness under disadvantageous environmental conditions. Transposons are key components of functional centromeres and play pivotal roles in the rapid divergence of centromeres between close related plants [14]. TEs are frequently related to hybrid defects that might cause reproductive isolation across diverse species [15]. Transposons represent the most abundant repetitive sequences in plant genomes, and they were found in all sequenced plants including the model plant Arabidopsis and many major crops in the world such as rice (*Oryza sativa*), wheat (*Triticum aestivum*), barley (*Hordeum vulgare*), maize (*Zea mays*) and soybean (*Glycine max*) (Figure 2). Unlike some repeats in plant genomes such as telomeric tandem repeats and ribosomal DNA (rDNA) repeats, most transposons are poorly conserved and can be quickly replaced in relative short period [16]. In many plants with larger genome sizes, transposons make up large fractions of the genomes. For example, about 85% of the maize genome is composed of transposons [17]. Therefore, transposon annotation is one of the most important and fundamental tasks for genome sequencing projects as it represents the precondition for many genomic analyses and the first step in the computation phase of genome annotation [18,19].

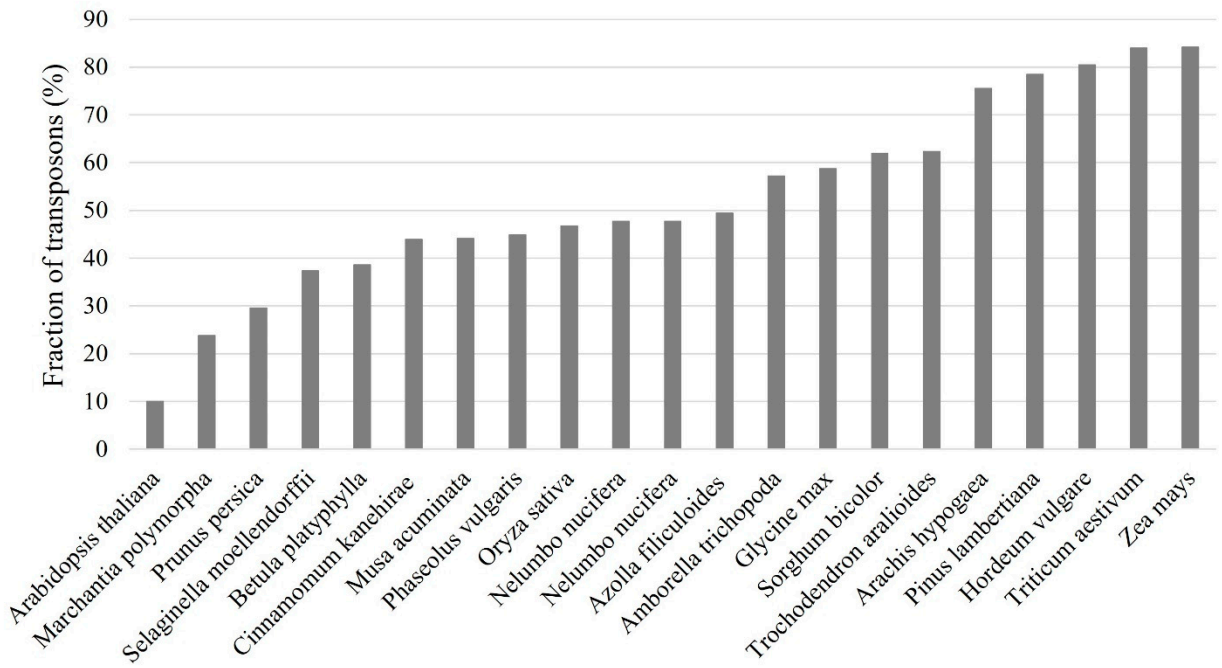


**Figure 2.** Transposon fractions of 21 sequenced plant genomes.

## 2. Strategies of transposon discovery

One of the most, or may be the most, important goal of all plant genome sequencing projects is to accurately identify and annotate functional genes in the genomes. However, it is important to define and understand genes before the gene annotation is started. Generally, a gene is a transcribed unit of heredity that may or may not translate to a protein. It is well known that many TEs are expressed and can generate proteins. Thus, these transposon sequences can be treated as genes. However, except limited TEs called domesticated transposons which have been co-opted by the host genome and confer some biological traits [20], the only function of the transposon-encoding proteins is for catalyzing transpositions. Therefore, TEs are frequently called 'jumping genes' or 'selfish genes', and they were excluded by many gene annotation projects which mostly focused on annotating the genes that encode non-transposase proteins and associate with biological and molecular functions

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2023 doi:10.20944/preprints202311.0673.v1



and phenotypic traits, such as the genes controlling yield, biotic and abiotic tolerance, and other phenotypic traits in crops.

Like genes and other genomic components, transposons are DNA sequences and are composed of four nucleotides including adenine (A), cytosine (C), guanine (G), and thymine (T), it is impossible to distinguish transposons from genes based on the composition of nucleotides. However, transposons are usually repetitive and exhibit unique structures, some of them may encode transposase proteins or reverse transcriptase (RT), which can be used for transposon prediction. Thus far, many resources including bioinformatics software and TE database have been developed. Here, I just mention some of these TE-related software and databases as most of the resources were well documented in the previous publications [21–24]. Generally, three major strategies are widely used for transposon annotations.

### *2.1. Repeat-based annotation*

Transposons are dispersed repeats, and many TE families are present in plant genomes in multiple copies, thus these features can be used to develop bioinformatics tools for the *de novo* transposon identification. One of the most popular programs is RECON which identifies and groups TEs based on the pairwise alignments between genomic sequences [25]. To conduct genome-wide transposon annotations with RECON, the assembled genome sequences are usually split into smaller fragments such as 10-20 Kb, and then used as the queries to search again the whole genome sequences for identifying repetitive sequences and grouping them into different families. The represented element or consensus sequence of each repeat family are extracted, the boundaries of the repeats can be defined by computational analysis and/or manual inspections (Figure 3A). For large plant genomes such as wheat (~17 Gb), it is extremely challenging to conduct all-against-all comparisons. One practical option is to use a certain fraction (10-20% or more) of the genome to search against the whole genome sequences. However, this could miss some repeat families. It should be noted that not all transposon families are highly repetitive, some single-copy or low-copy transposons may not be detected by RECON. Additionally, many genes including gene families and duplicate genes are also scattered throughout plant genomes, and these repeats should be removed during transposon annotations.

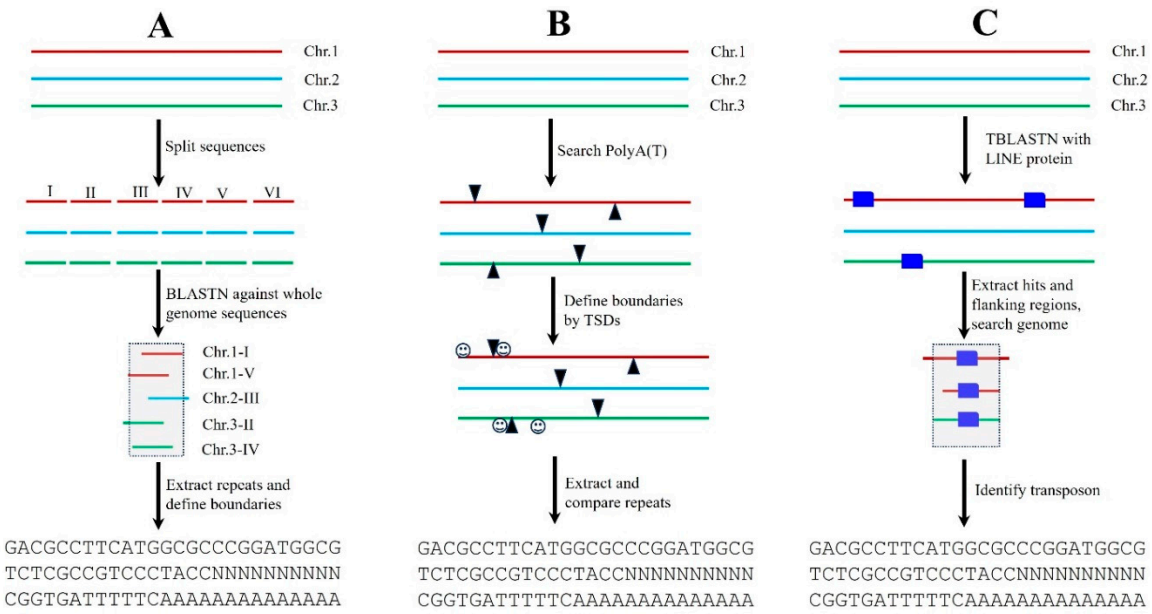


**Figure 3. Three major strategies for annotating transposons (LINEs as the example)**. A. Repeat-based transposon annotation. The plant genome sequences (assembled chromosomes or large scaffolds) were split into smaller fragments, and then used to search against all plant genome sequences. All repetitive sequences were extracted and grouped into different repeat families based

 doi:10.20944/preprints202311.0673.v1



on sequence identity. The boundaries of transposons can be defined by computational comparisons. B. Structure-based transposon annotation. The poly(A) motif of LINEs was recognized by developed software and the boundaries of LINEs were defined based on TSDs (smiley faces). Then the identified LINE sequences were extracted and grouped into different families. C. Homology-based transposon annotation. The proteins of represented LINEs were used to search against the plant genomes and to identify homologous sequences (blue rectangles). Then, the homologous hits and the flanking regions (usually 5-10 Kb for each direction) were extracted together and used to conduct sequence comparisons for defining boundaries of LINEs and grouping them into families.

### *2.2. Structure-based annotation*

Transposons exhibit some unique structural features such as LTRs, TIRs and Poly(A) tails (Figure 1). Additionally, many transposon superfamilies can generate TSDs when they moved and inserted into new genomic positions. Therefore, these unique characteristics can be used to develop bioinformatics programs for *de novo* transposon annotations. For example, many non-LTR retrotransposons contain 3' poly(A) terminal motif and are flanked by variable TSDs, thus scientists can develop software to recognize the poly(A) motif, and the boundaries of non-LTR retroelements can be defined by TSDs. Then, the annotated sequences are extracted and grouped into different families based on their sequence comparisons (Figure 3B). Compared to RECON, structure-based software can annotate transposons with well-defined boundaries, and many of those programs are able to conduct genome-wide transposon annotations for the plants with larger genome sizes in a relatively short time. Thus far, most of the transposon annotation programs were developed based on the unique structures of distinct transposon superfamilies including those for annotating LTR retrotransposons [26–31], non-LTR retrotransposons [32,33], and DNA transposons [34–41]. Notably, the structure-based annotation software may find full length transposons with both low and high copies, but they may miss some transposons without the typical transposon features.

### *2.3. Homology-based annotation*

This strategy is based on the hypothesis that transposons from related organisms could share common origins and show certain sequence similarities. Thus, researchers can use the known transposons which have already been annotated to find their homologous sequences in new genomes. The best known and most widely used program for homology searches is RepeatMasker (https://www.repeatmasker.org) which applies customer or precompiled repeat libraries to identify homologous repeats. There are several ways to obtain known transposon sequences. The first is to check related transposon databases such as Repbase Update or Repbase, which is a well curated and the most comprehensive database of repetitive elements in eukaryotic genomes [42]. Another way is to use the published transposon database from relative plants. For example, users can use the annotated transposons in common bean (*Phaseolus vulgaris*) [43] to find the homologous sequences in other *Phaseolus* species. The third way is to download deposited transposons from GenBank. However, the efficiency and accuracy of homology-based annotation heavily depend on the genomic similarity between the host plants of reference transposons and the plants under study. As plant transposons are dynamic sequences and many TE families are species-specific, the reference transposons may yield poor annotations in distantly related plants. Also, this approach is not practical for the orphan plant lineages for which no transposon database is available in the relative species. Despite we can use the repeats in Repbase or other databases, in most cases, the homology sequences were short and fragmental in distantly organisms. To get better annotation, the proteins of identified transposons can be used to conduct TBLASTN search against the newly sequenced plant genomes and to identify homologous sequences, then the hits and their flanking regions (5-10 Kb for each side) are extracted and used for BLASTN searches and genome-wide comparisons to identify complete transposons (Figure 3C). As TBLASTN search is very sensitive to detect homologous sequences, technically, it can identify more transposase-encoding elements including both full-length and truncated transposons as well as both single copy and multiple copies of elements.

 doi:10.20944/preprints202311.0673.v1



Except the three major strategies above, other new methods of TE annotations also were developed such as TASR which is based on that transposons are epigenetically silenced by 24 nt-siRNAs, thus this feature can be used to develop bioinformatics pipeline to recognize the regions targeted by the small RNAs and to define transposons [44]. Occasionally, some transposons inserted into genes and caused phenotypic variations. Thus, these transposons can be identified by cloning and comparing the gene sequences of mutants and the wild types. However, this method is not suitable for genome-wide transposon discovery. As plant genomes are very complicated and each of the transposon annotation methods has its own limitations, it is better to combine multiple computational pipelines to obtain high-quality of transposon annotations.

## 3. Steps for transposon annotation

There are three basic steps in genome-wide transposon annotation: preliminary annotation, classification, quality check and data improvement. However, many bioinformatics pipelines for transposon annotation combine the first and second steps and can directly output classified transposons.

### *3.1. Brief introduction of plant transposon superfamilies and their annotations*

#### 3.1.1. LTR retrotransposons

LTR retrotransposons represent the most abundant repeats in many, maybe most, plant genomes such as maize which consists of about 75% LTR retroelements [17]. These retroelements exhibit some structural hallmarks including LTR, TSD, reverse transposase proteins, primer binding site (PBS), polypurine tract (PPT) and the 5'TG…CA3' terminal motifs that can be used for characterizing LTR retrotransposons. Many plant transposon annotations usually start with LTR retrotransposons for three reasons: 1) LTR retrotransposons contribute large fractions of plant genomes; several excellent bioinformatics programs have been developed [26–31] for annotating LTR retrotransposons; and 3) these software may generate lower error rates of data mining than many programs for *de novo* detection of DNA transposons and non-LTR retrotransposons. By following the protocols of the developed software, users can generate sufficient data for preliminary predictions of LTR retrotransposons. There are two superfamilies of LTR retrotransposons, including *Ty1*/*Copia* and *Ty3*/*Gypsy*, are found in plants, they share similar structures but have different order of the reverse transcriptase and integrase (INT) domains (Figure 1). Most complete LTR retrotransposons are large (about 4-10 Kb) and some can be over 20 Kb. However, some type of LTR retroelements called TRIMs is small (about 250 bp to 1500 bp) and with tiny LTRs (< 100 bp) [11]. Therefore, users need to pay attention to the default parameters of the related software. For LTR_Finder, the default minimum size for LTRs and the internal regions is 100 bp and 1,000 bp, respectively [27]. To get better annotations with LTR_ Finder, we usually use the sets: ./ltr_finder -d 30 -D 15000 -l 30 -L 5000 -s ./tRNAdb/Athal-tRNAs.fa where d represents the minimum distance between 5' and 3' of the LTRs, D represents maximum distance between 5′ and 3′ LTRs, l represents minimum size of 5' and 3' LTRs, L represents maximum size of 5′ and 3′ LTRs, and s means the tRNA sequence file we used.

#### 3.1.2. LINEs

LINEs are non-LTR retrotransposons and may be the most ancient types of retrotransposons in plant genomes. The typical LINEs usually contain non-LTR reverse transposases and the poly(A) terminal motif. The general methods for LINEs annotations are presented in Figure 3, and the related software including RECON [25] and MGEScan-Non-LTR [45] can be used for automatic LINE annotations. The boundaries of full-length LINEs can be defined by TSDs. However, most LINE elements in plant genomes are truncated at their 5' end due to unsuccessful reverse transcription or other reasons. It should be noted that the RTE retrotransposons lack a poly (A) tail and instead contain tandem repeats [46,47]. It seemed that the RTEs in flowering plants were horizontally transferred from aphids or other unsequenced animals [47]. As the RTEs from different plants show

 doi:10.20944/preprints202311.0673.v1



high sequence similarity, thus the identified RTEs can be used to find the homologous elements in new plant genomes.

#### 3.1.3. SINEs

SINEs are another type of non-LTR retrotransposons. Unlike LINEs, they are very small (< 500 bp) and lack conserved coding domains. Also, the 5' end of many SINEs are truncated. Thus, SINEs may represent the superfamily that is most difficult to precisely annotate among all plant transposons. Nearly all plant SINEs were derived from tRNAs, they usually contain an internal RNA polymerase III (pol III) promoter consisting of box A and box B motifs, and the 3' poly (A) flanking by TSDs [48]. Thus far, some SINE annotation software has been developed to recognize these structural motifs and conduct genome-wide SINE prediction [32,33]. SINEs were found in a wide range of plants, many of which are present in specific lineages such as p-SINE1 in the *Oryza* genus [49], but some SINE families such as Au are widely distributed in both dicots and monocots [50] that provide good resources for homology-based SINE annotations [51].

#### 3.1.4. Endogenous plant pararetrovirus (EPRVs)

Pararetroviruses are double-stranded DNA viruses that replicate through an RNA intermediate. They were named by Temin (1985) to distinguish hepadnaviruses in animals from retroviruses that are RNA viruses and integrate their DNA copies into host genomes [52]. Except very few members such as *Petunia vein-clearing virus* (PVCV) [53], the vast majority of plant pararetroviruses do not have the integrase domain. All plant pararetroviruses belong to the family *Caulimoviridae*, they contain one to eight open reading frames (ORFs) and their genome sizes range from 7.1 to 9.8 Kb [54]. Endogenous pararetroviruses (EPRVs) are present in the genomes of a wide range of plants [55,56], they encode polyproteins that share high sequence similarity with Ty3/Gypsy LTR retrotransposons, thus EPRVs were frequently misannotated as Ty3/Gypsy LTR retrotransposons or were ignored. All EPRVs identified thus far lack LTRs [7,8] that bring difficulties to annotate these pararetroviral sequences and accurately define their boundaries. Recently, the bioinformatics pipeline CAULIFINDER has been developed for automatic annotation and classification of EPRVs in plant genomes [7].

#### 3.1.5. DNA transposons with TIRs

Among the nine superfamilies of DNA transposons identified in plant genomes, seven superfamilies have TIRs including Mutator, CACTA, hAT, PIF/Harbinger, Tc1/Mariner, Sola, and Ginger.

##### 3.1.5.1. Mutator transposons

Mutator transposons or mutator-like transposable elements (MULEs) were originally discovered in maize [57] and have since been identified in other plants, animals, and fungi [58]. The transposons of this superfamily usually carry relative long TIRs (can be over 300 bp) and produce TSDs of 7–12 bp [59]. MULEs are abundant in plant genomes and may represent the most mutagenic plant transposons identified thus far [60]. They are usually less than 5 Kb in size and contain one ORF encoding mutator transposase. However, MULEs can be over 8 Kb and may have multiple ORFs including one encoding transposase protein and other(s) which may help transposon movement, or their functions are not very clear [60,61]. Some MULEs called Pack-MULEs have captured functional genes or fragments of expressed genes and play important roles in plant gene evolution [59,62].

##### 3.1.5.2. CACTA transposons

CACTA transposons were first found in maize and named *Enhancer* (*En*) and *Suppressor - mutator* (*Spm*) [63,64]. The typical features of this superfamily are the terminal motifs starting with CACTA or CACTG and ending with TAGTG or CAGTG. The TIRs of CACTA elements are short (mostly < 50 bp) and they generate a 3-bp TSDs. The complete CACTA transposons are usually large (> 10 Kb), some of them can be over 20 kb in size [65]. Due to their large sizes and high copy numbers, CACTA

 doi:10.20944/preprints202311.0673.v1



transposons contribute larger fractions of many plant genomes than other DNA transposons. Like Pack-MULEs, some CACTA transposons have been found to carry host gene sequences [66].

3.1.5.3. hAT transposons

hAT transposons were first identified by Barbara McClintock as the Activator or Ac element [67], but the name of this superfamily is an acronym of its three members: hobo from Drosophila, Ac from maize, and Tam3 from snapdragon [68]. hAT elements are widely distributed in plants and other eukaryotes, they are typically less than 5 kb in length, and usually contain short TIRs (5- 27 bp) flanked by 8-bp TSDs [69,70]. In addition, the transposase of many hAT elements contains the conserved domain of 50 amino acids at the C terminus which may be involved in dimerization or other functions [71].

3.1.5.4. PIF/Harbinger transposons

PIF/Harbinger transposon superfamily gets its name from the two founding members: P instability factor (PIF) from maize [72] and Harbinger from *Arabidopsis thaliana* [73]. They usually contain short TIRs (about 14-50 bp) and generated 3-bp TSDs (TAA or TTA). Unlike other DNA transposons, the intact PIF/Harbinger elements have two main open reading frames (ORFs), one ORF produces catalytic transposase containing a conserved DDE motif and another ORF encodes a Myb-like protein [74]. PIF/Harbinger transposons have been found in the genomes of many plants and some of them were co-opted or domesticated to serve as new molecular functions associated with yield and other traits in plants [75,76].

3.1.5.5. Tc1/Mariner transposons

Tc1 and Mariner transposons were first discovered Caenorhabditis elegans and Drosophila mauritiana in 1980s [77,78], and then were found in a wide range of organisms including both prokaryotes and eukaryotes [58]. The Tc1/Mariner transposons in plants mostly range in size from 1.5 to 6 Kb and contain relative short TIRs (~30 bp) [79,80]. In contrast to other TEs that create variable TSDs, Tc1/Mariner superfamily always generates 2-bp TSDs of TA. They are not very abundant in many plant genomes and were missed by some plant genome annotations.

3.1.5.6. Sola transposons

Sola transposons were first reported in 2009 and they are distributed in a wide range of organisms including bacteria and metazoans [3]. The transposons of Sola superfamily encode DDD-TPase and are flanked by 4-bp TSD. Most Sola elements range in size from 2 Kb to 9 Kb, but some can be over 15 Kb. They also have very variable terminal sequences and the TIRs of the reported Sola transposons ranged from 11 bp to 1,124 bp [3]. Despite Sola transposons were found in some plants such as the moss *Physcomitrella patens*, it seems that they are not common in flowering plants.

3.1.5.7. Ginger transposons

Gypsy INteGrasE Related (Ginger) transposons are unusual DNA transposons as they contain TIRs and a protein which shares high sequence similarity to the integrase encoded by Gypsy LTR retrotransposons [4]. DNA transposons containing a Gypsy-like integrase were first found in *Dictyostelium discoideum* [81,82], and subsequently identified in animals [4] and plants [5]. The TIRs of Ginger transposons are relative long (40-270 bp) and contain the 5′-TG…CA-3′ terminal dinucleotides, and the insertions of Ginger transposons generate TSDs of 4-6 bp [4,5].

3.1.5.8. MITEs

MITEs were first identified in maize in 1992 [12], they are small DNA transposons (mostly < 500 bp) with TIRs surrounded by variable TSDs. In contract to many previously reported DNA transposons that mobile with the ‘cut and paste’ model which usually does not increase new copies

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2023 doi:10.20944/preprints202311.0673.v1



and are present in low copy numbers, MITEs were frequently found in high copy number, can be over 10,000 copies for some families [83]. Additionally, dramatic copy number differences for a same MITE family can be detected between closely related genomes [74] suggesting that MITEs can rapidly increase theirs copy numbers in short period. Therefore, MITEs were considered as a special type of DNA transposons or even Class III transposons [84]. However, more studies revealed that MITEs were likely derived from the internal deletions of large DNA elements as the TIRs of MITEs showed high sequence identity to that of several reported DNA transposon superfamilies such as PIF/Harbinger and hAT [74,85–87].

#### 3.1.5.9. TIR DNA transposon annotation

Several programs have been developed to annotate TIR DNA transposons such as TIRvish [34], TIR-Learner [35], Generic Repeat Finder [36], and Inverted Repeat Finder [88]. The software provides good resources to detect transposons solely based on the recognition of TIRs or the combination of structure and homology transposon annotations. In addition, MITE-Hunter [37], detectMITE [38], MiteFinderII [39], MITE Tracker [40] and other related bioinformatics pipelines can be used to find small TIR-DNA transposons that lack transposases. As the TIRs of DNA transposons are usual short and less conserved, it is still challenging to accurately detect these DNA transposons and define their boundaries.

### 3.1.6. Helitron transposons

Helitron transposons were first identified in Arabidopsis, rice, and nematode (*Caenorhabditis elegans*) in 2001 [89]. They have conservative termini (5'-TC … CTRR (mostly CTAG)-3') and contain the hairpins (16-20 bp) separated by 10–12 nucleotides from the 3' end. However, all Helitron elements do not have terminal inverted repeats and do not generate TSDs, they transpose precisely between 5′AT3'. Helitron like sequences are widely present in plant genomes, some of them are large and can capture gene fragments [90]. Helitron transposons are difficult to annotate and precisely define their boundaries as they lack both TSDs and TIRs. They can be detected using the unique terminal motifs and the helicase sequences or with the related software such as HelitronScanner [41].

### 3.1.7. Replitron transposons

Replitron is a new group of DNA transposon reported in 2023 [6]. They lack TIRs but contain short direct terminal repeats ranging from 5 to 11 bp and HUH endonuclease that is distantly related to Helitron transposons. Replitron elements are present in the genomes of green algae, liverworts, mosses, lycophytes and ferns, but absent in hornworts and seed plants [6]. The identified Replitrons are 900 bp to 4 kb, and their insertions can generate TSDs of 2-8 bp. Thus far, no software is developed for annotating Replitron transposons. One practical approach is to conduct homology searches with the endonuclease proteins of known Replitrons and identify new Replitron sequences based on sequence comparisons.

### 3.1.8. Others

Despite the programs used to predict specific types of transposons have been developed, there is no need to annotate each of the transposon superfamilies one by one. Several bioinformatics packages have been created through combining multiple programs and methods to annotate all types of transposons, such as Extensive de-novo TE Annotator (EDTA) [91], TransposonUltimate [92], Earl Grey [93], and other comprehensive bioinformatics pipelines.

## *3.2. Classification of transposons*

Once the potential transposon sequences have been generated, the next step is to group them and define their families. Many bioinformatics pipelines can annotate and automatically generate classified transposons, here I just describe how to handle the transposon sequences that are not grouped.

 doi:10.20944/preprints202311.0673.v1



3.2.1. Definition of transposon superfamilies

The superfamilies of transposons can be defined based on the encoded transposase proteins and other sequence features including TIRs, LTRs and the sizes of TSDs. To classify the transposon superfamilies, the widely used method is to use the generated transposon sequences for conducting BLASTX or BLASTN search against GenBank or other database such as Gypsy Database (GyDB) [94] that provides a valuable resource to define the superfamilies of retroelements and determine the conserved domains of LTR retrotransposons and EPRVs. However, many annotated transposons have accumulated mutations or undergone deletions and encode no transposase or short protein sequences, and the superfamilies for these TEs can be defined based on their terminal motifs and flanked TSDs, or they can be classified as unclassified transposons, such as TRIMs, LARDs, or MITEs. It is important to note that unclassified transposons should be real transposons and non-transposon sequences should be excluded.

3.2.2. Definition of transposon families

In some cases, it is challenging to define transposon families as different scientists applied distinct cutoffs or strategies. Thus far, transposon families are commonly defined using the '80-80-80' rule which means that any members of a same transposon family should be over 80 bp and show more than 80% sequence identity over 80% of their sizes [2]. However, other studies argued that this definition may ignore the consensus sequences [95] and may not be applicable for monophyletic groups [96]. As many transposons may contain highly conserved regions such as the internal RT domain of LTR retrotransposons, transposons of a same family defined by the '80-80-80' rule may be grouped into different phylogenetic clades. The terminal sequences including both LTRs and TIRs are less conserved than the transposase-encoded regions, and elements of the same transposon family usually exhibit extensive sequence identity at their termini. Therefore, the terminal sequences of transposons may be more sensitive to define transposon families. In addition, members from a same transposon family should share close phylogenetic relationships and should be catalyzed by same transposase enzymes. The better strategies to define transposon family should consider sequence identity and phylogenetic origins as well as the molecular interactions.

3.2.3. Autonomous and non-autonomous transposons

Transposons can also be grouped into autonomous and nonautonomous elements. The former are usually complete transposons and encode the entire enzymes required for transposition, whereas the latter may have undergone internal deletions or mutations and lack functional transposases and their movement is catalyzed by their autonomous partners. It should be noted that the terms of 'autonomous' and 'non-autonomous' are defined based on the functional mobility, and they cannot be used to classify superfamilies of transposons. Any superfamily of both Class I and II transposons can be divided into either autonomous or non-autonomous elements. In many cases, autonomous and non-autonomous transposons were defined solely based on in silico gene prediction, if any transposons encode transposase proteins and contain the entire domains for transposition, they can be generally considered as autonomous transposons. However, the accuracy and reliability of computational predictions depend on many factors. If the mobility of a transposon has not been experimentally validated, any autonomous elements defined by computational prediction should be treated as potentially (but not true) autonomous transposons even their gene models are well supported by transcriptional data.

*3.3. Quality control and improvement of transposon annotation*

Transposons are highly repetitive, and many plants have complex genome structures, especially for those with large genome sizes and/or duplicated chromosomes. Despite many bioinformatics software or pipelines have been developed and used for transposon prediction, not all annotated sequences are real transposons and some of them may be misannotated. Additionally, some protein-coding genes can also be predicted as repetitive sequences by some *de novo* annotation tools [19].

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 10 November 2023 doi:10.20944/preprints202311.0673.v1



Therefore, users must carefully evaluate the annotated sequences, discard the incorrect annotation, and improve the quality of transposon annotation when they obtain first version of transposon database for newly sequenced plant genomes. Computational analysis is important, but manual inspection is strongly encouraged for this step. Goubert et al has proposed a guideline for manual curation of transposons [97], here I just want to highlight four things that are helpful for improving the quality of transposon annotation.

### 3.3.1. Identification and exclusion of misannotated sequences

Generally, three major types of genomic sequences were frequently misannotated: tandem repeats (TRs), non-transposon sequences and misclassified transposons. Occasionally, some transposons such as TRIMs were tandemly organized [11]. However, nearly all tandem repeats are the long track repeats such as centromeric tandem repeats, they are widely dispersed and consist of various basic units. These sequences were frequently misannotated as LTR retrotransposons and significantly reduce the accuracy of transposon annotations. TRs can be identified with related programs [98] or pairwise sequence alignments (Figure 4). Non transposon sequences, such as genes, pseudogenes and other genomic sequences, can also be misidentified as transposons, especially for those annotated as fragmental or unclassified transposons. In some cases, some transposon sequences were misclassified. For example, I manually inspected a SINE database annotated by a computational program and found that some annotated SINEs were the fragments of true LTR retrotransposons.

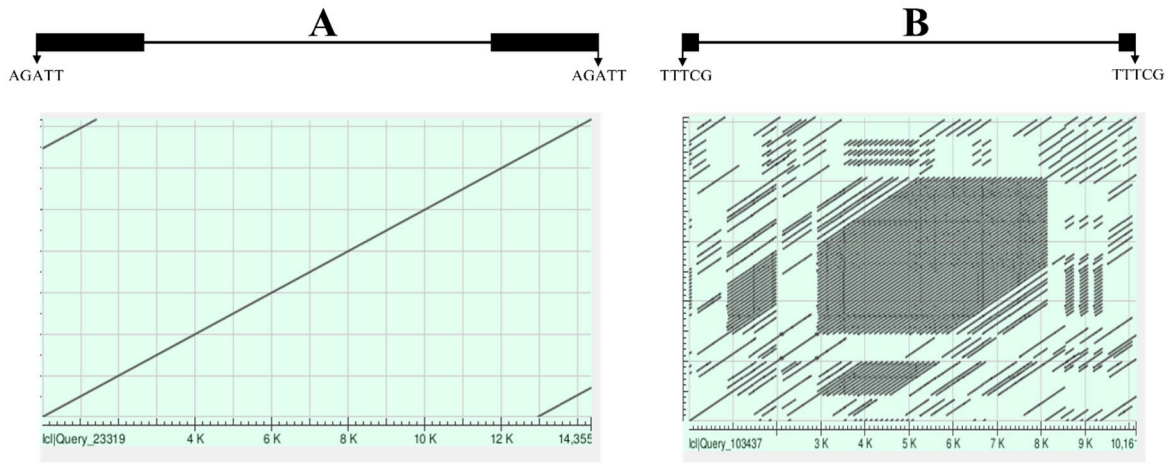


**Figure 4.** Pairwise sequence comparisons of two potential LTR retrotransposons in barley annotated by a bioinformatics software. Both annotated sequences have LTRs (black rectangles) and were flanked by 5-bp TSDs. However, pairwise comparisons revealed that sequence A was real LTR retroelement whereas the sequence B was misannotated as it contained tandem repeats.

### 3.3.2. Elimination of nested transposons

One transposon called bottom transposon may contain other transposon(s) named nested transposon(s). Nested organizations of transposons are very common in many plants with higher repeat contents such as maize [99,100]. They are also frequently found in some specific regions of several plant genomes with lower repeat contents such as the centromeric and pericentromeric regions in rice [14]. In some cases, transposons were organized into multiple layers of nested insertions in which nested elements further served as target transposons for other TEs [14,100]. In plants, LTR retrotransposons frequently acted as the target transposons to harbor other LTR retrotransposons of same or different families, but DNA transposons can also be inserted by LTR retrotransposons and other transposons. Despite nested transposons provided some insights into the evolutionary dynamics of transposons, they may cause problems for transposon classifications. Therefore, nested transposons should be removed from the reference transposons. Nested transposons can be easily identified by comparing different family members of the same transposon

 doi:10.20944/preprints202311.0673.v1



family or orthologous elements between multiple genomes from same or related species. Despite identification of inserted transposons is tedious and time-consuming, this process may find some novel transposon families and many of them may be recently active.

#### 3.3.3. Definition of transposon boundaries

The exact boundaries of transposons can be determined based on the terminal sequences, including LTR and TIRs, and TSDs. However, the movements of some transposons such as Helitrons do not generate TSDs. Additionally, not all transposons possess the typical features of their superfamilies, some of them may have accumulated mutations or undergone deletions, and their termini and TSDs are difficult to recognize. Unfortunately, no software is available for defining the boundaries for fragmental or mutated transposons. One strategy is to extract the flanking sequences of transposons and conduct sequence alignment analysis. It is exhausting and time-consuming but is practical. With the advance of sequencing technologies, multiple genomes from same and closely related species have been sequenced, the exact boundaries of many transposons can also be defined with multiple reference sequence sets [101].

#### 3.3.4. Identification of unannotated transposons

Due to the complexity of plant genomes and the current limitations of available bioinformatics resources for *de novo* TE discovery, some transposons may be missed. There are several ways to identify missing transposons. The first method is to apply more and different annotation approaches and see if any new transposons can be discovered. The second strategy is to check the reported transposons in newly sequenced genomes. Transposons are usually very dynamic, but some transposons are present in a wide range of plants. For example, the An-RTEs (RTE clade of LINEs), Cassandra (TRIMs) and Au-SINEs are present in many dicots and monocots [11,47,102,103]. If they are not found in a new sequenced dicot or monocot, one possibility is that they may be missed by the transposon annotations. The third way is to use the annotated transposons to screen the genome with RepeatMasker and use the transposase proteins of identified superfamilies to conduct BLASTN search against the masked genome. If multiple significant hits were detected, it suggested that some transposons must be missed.

### *3.4. Criteria for good transposon database*

One question is that how to evaluate the quality of transposon database? If two research groups (A and B) independently annotated transposons in a same genome, the A group's transposons can define 75% of the genome whereas the B group's transposons only mask 70% of the genome. Can we say the transposon database of A group was better than that of B group? The answer for this may be yes or no. The coverage is an indicator, but not the sole one to evaluate the quality of transposon annotation. In my opinion, there are four criteria to judge a good transposon database. a. Transposons should be accurately annotated. Two major points for this, 1) the annotated transposons should be real transposons and non-TE sequences should be excluded; 2) all annotated transposons should be correctly classified. For example, a Ty3/Gypsy retrotransposon cannot be classified as a Ty1/Copia or non LTR retrotransposon. b. The boundaries of transposons should be well defined. c. Nearly all transposons should be annotated. It is impossible to identify all transposons in plant genomes, but we shouldn't miss many of them. d. The transposon database should be non-redundant. Transposons are very redundant, it is not necessary to include all or many copies of a transposon family into the database as it will enlarge the size of TE library and impede computational analyses [43], thus it is better to have a well-defined reference transposon for each family.

## 4. Conclusions and Future Directions

Despite extensive research has been conducted, accurate annotation of transposons remains challenging and time-consuming. Thus far, 16 superfamilies of transposons have been found in plants and numerous bioinformatics software and pipelines have been developed to annotate transposons

based on the three major annotation strategies. Although different systems or parameters were proposed for transposon classification and family definition, the terms of 'autonomous' and 'non-autonomous' cannot be used to classify the transposon superfamilies. A good transposon database should be accurately annotated and fully represented, and the boundaries of transposons should be well defined. As transposon annotation heavily depends on the available genomes, it is impossible to have high-quality TE database with poorly sequenced and assembled genomes. With the innovation in software development and sequencing techniques, more complete plant genomes will be well annotated, and new types of transposons may also be discovered in plants. The current bioinformatics programs offer incredible resources for transposon predictions, however, some of them are difficult to install and use for beginners, especially for those with little expertise in bioinformatics and computational biology, and many of these programs also produce numerous false-positive annotations. New methodologies and algorithms are needed to develop next generation annotation tools for significantly improving the accuracy and efficiency of transposon annotation. It is also highly recommended to periodically update the current transposon database and make them public availability.

**Author Contributions:** D.G collected related materials, wrote the manuscript, and takes full responsibility for this publication.

**Funding:** This research was supported by USDA-ARS CRIS Project No. 2050-21000-038-000D**.**

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Not applicable.

**Acknowledgments:** I am grateful for Dr. Ning Jiang for providing the initial training on transposons. I also would like to thank Dr. Scott Jackson for giving the opportunity on plant transposon annotations.

**Conflicts of Interest:** The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Mention of trade names or commercial products in this publication is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the United States Department of Agriculture.

## References

1. Kapitonov, V.V.; Jurka, J. A universal classification of eukaryotic transposable elements implemented in Repbase. *Nat. Rev. Genet*. **2008**, 9, 411–412.
2. 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. *Nat. Rev. Genet.* **2007**, 8, 973–982.
3. Bao, W.; Jurka, M.G.; Kapitonov, V.V.; Jurka, J. New superfamilies of eukaryotic DNA transposons and their internal divisions. *Mol. Biol. Evol.* **2009**, 26, 983–993.
4. Bao, W.; Kapitonov, V.V.; Jurka, J. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mob. DNA.* **2010**, 1, 3.
5. Cerbin, S.; Wai, C.M.; VanBuren, R.; Jiang. N. GingerRoot: A novel DNA transposon encoding integrase-related transposase in plants and animals. *Genome Biol Evol.* **2019**, 11, 3181-3193.
6. Craig, R.J. Replitrons: A major group of eukaryotic transposons encoding HUH endonuclease. *Proc Natl Acad Sci U S A.* **2023**, 120, e2301424120.
7. Vassilieff, H.; Haddad, S.; Jamilloux, V.; Choisne, N.; Sharma, V.; Giraud, D.; Wan, M.; Serfraz, S.; Geering, A.D.W.; Teycheney, P.Y.; et al. CAULIFINDER: a pipeline for the automated detection and annotation of caulimovirid endogenous viral elements in plant genomes. *Mob DNA.* **2022**, 13, 31.
8. Richert-Pöggeler, K.R.; Noreen, F.; Schwarzacher, T.; Harper, G.; Hohn, T. Induction of infectious petunia vein clearing (pararetro) virus from endogenous provirus in petunia. *EMBO J*. **2003**, 22, 4836–4845.
9. Havecker, E.R.; Gao, X.; Voytas, D.F. The diversity of LTR retrotransposons. *Genome Biol.* **2004**, 5, 225.
10. Witte, C.P.; Le, Q.H.; Bureau, T.; Kumar, A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 13778–13783.
11. Gao, D.; Li, Y.; Do Kim, K.; Abernathy, B.; Jackson, S.A. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biol.* **2016**, 17, 7.
12. Bureau, T.E.; Wessler, S.R. Tourist: a large family of small inverted repeat elements frequently associated with maize genes. *Plant Cell*. **1992**, 4, 1283-1294.

13. Bennetzen, J.L.; Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu. Rev. Plant Biol.* 2014, 65, 505–530.
14. Gao, D.; Jiang, N.; Wing, R.A.; Jiang, J.; Jackson, S.A. Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Front. Plant Sci.* **2015**, 6, 216.
15. Serrato-Capuchina, A.; Matute, D.R. The role of transposable elements in speciation. *Genes* **2018**, 9, 254.
16. Gao, D.; Gill, N.; Kim, H.; Walling, J.G.; Zhang, W.; Fan, C.; Yu, Y.; Ma, J.; SanMiguel, P.; Jiang, N.; et al. A lineage-specific centromere retrotransposon in Oryza brachyantha. *Plant J.* **2009**, 60, 820–831.
17. Schnable, P.S.; Ware, D.; Fulton, R.S.; Stein, J.C.; Wei, F.; Pasternak, S.; Liang, C.; Zhang, J.; Fulton, L.; Graves, T.A.; et al. The B73 maize genome: Complexity, diversity, and dynamics. *Science* **2009**, 326, 1112–1115.
18. Jamilloux, V.; Daron, J.; Choulet, F.; Quesneville, H. De novo annotation of transposable elements: Tackling the fat genome issue. *Proc. IEEE* **2016**, 105, 474–481.
19. Yandell, M.; Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **2012**, 13, 329
20. Jangam, D.; Feschotte, C.; Betrán, E. Transposable element domestication as an adaptation to evolutionary conflicts. *Trends Genet.* **2017**, 33, 817–831.
21. TE Hub Consortium.; Elliott, T.A.;Heitkam, T.; Hubley, R.; Quesneville, H.; Suh, A.; Wheeler, T.J. TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mob DNA.* **2021**, 12, 16.
22. Lerat, E. Identifying repeats and transposable elements in sequenced genomes: How to find your way through the dense forest of programs. *Heredity* **2010**, 104, 520–533.
23. Storer, J.M.; Hubley, R.; Rosen, J.; Smit, A.F.A. methodologies for the de novo discovery of transposable element families. *Genes* **2022**, 13, 709.
24. Mokhtar, M.M.; Alsamman, A.M.; El, Allali. A. PlantLTRdb: An interactive database for 195 plant species LTR-retrotransposons. *Frontiers in Plant Science* **2023**, 14, 1134627.
25. Bao, Z.R.; Eddy, S.R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **2002**, 12, 1269–1276.
26. McCarthy, E.M.; McDonald, J.F. LTR STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, 19, 362–367.
27. Xu, Z.; Wang, H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res*. **2007**, 35, 265–268.
28. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform*. **2008**, 9, 18.
29. Steinbiss, S.; Willhoeft, U.; Gremme, G.; Kurtz, S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Res.* **2009**, 37, 7002–7013.
30. Ou, S.; Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal-repeat retrotransposons. *Plant Physiol*. **2017**, 176, 1410–1422.
31. Orozco-arias, S.; Liu, J.; Id, R.T.; Ceballos, D.; Silva, D.; Id, D.; Ming, R.; Guyot, R. Inpactor, integrated and parallel analyzer and classifier of LTR retrotransposons and its application for pineapple LTR retrotransposons diversity and dynamics. *Biology* **2018**. *7*, 32
32. 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. *Plant Cell* **2011**, 23, 3117–3128.
33. Li, Y.; Jiang, N.; Sun. Y. AnnoSINE: a short interspersed nuclear elements annotation tool for plant genomes. *Plant Physiol*. **2022**, 188, 955-970.
34. Gremme, G.; Steinbiss, S.; Kurtz, S. GenomeTools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinform*. **2013**, 10, 645–656.
35. 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. *Mol. Plant.* **2019**, 12, 447–460.
36. Shi, J.; Liang, C. Generic Repeat Finder: A High-Sensitivity Tool for Genome-Wide De Novo Repeat Detection. *Plant Physiol.* **2019**, 180, 1803–1815.
37. Han, Y.; Wessler, S.R. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res*. **2010**, 38, e199.
38. Ye, C.; Ji, G.; Liang, C. detectMITE: A novel approach to detect miniature inverted repeat transposable elements in genomes. *Sci. Rep*. **2016**, 6, 19688.
39. Hu, J.; Zheng, Y.; Shang, X. MiteFinderII: A novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC Med. Genom*. **2018**, 11, 101.
40. Crescente, J.M.; Zavallo, D.; Helguera, M.; Vanzetti, L.S. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinform*. **2018**, 19, 348.
41. Xiong, W.; He, L.; Lai, J.; Dooner, H.K.; Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci. USA* **2014**, 111, 10263–10268.

42. Bao, W.; Kojima, K.K.; Kohany. O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob DNA.* **2015**, 6, 11.
43. Gao, D.; Abernathy, B.; Rohksar, D.; Schmutz, J.; Jackson, S.A. Annotation and sequence diversity of transposable elements in common bean (Phaseolus Vulgaris). *Front. Plant Sci*. **2014**, 5, 339.
44. El Baidouri, M.; Kim, K.D.; Abernathy, B.; Arikit, S.; Maumus, F.; Panaud, O.; Meyers, B.C.; Jackson, S.A. A new approach for annotation of transposable elements using small RNA mapping. Nucleic Acids Res. 2015, 43, e84.
45. Rho, M.; Tang, H. MGEScan-Non-LTR: Computational Identification and Classification of Autonomous Non-LTR Retrotransposons in Eukaryotic Genomes. *Nucleic Acids Res*. **2009**, 37, e143.
46. Malik, H.S.; Eickbush, T.H. The RTE class of non-LTR retrotransposons is widely distributed in animals and is the origin of many SINEs. *Mol. Biol. Evol*. **1998**, 15, 1123–1134.
47. Gao, D.; Chu, Y.; Xia, H.; Xu, C.; Heyduk, K.; Abernathy, B.; Ozias-Akins, P.; Leebens-Mack, J.H.; Jackson, S.A. Horizontal Transfer of Non-LTR Retrotransposons from Arthropods to Flowering Plants. *Mol. Biol. Evol*. **2018**, 35, 354–364.
48. Deragon, J.M.; Zhang, X. Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. *Syst Biol.* **2006** 55, 949-956.
49. Umeda, M.; Ohtsubo, H.; Ohtsubo, E. Diversification of the rice Waxy gene by insertion of mobile DNA elements into introns. *Jpn. J. Genet*. **1991**, 66, 569–586.
50. Yasui, Y.; Nasuda, S.; Matsuoka, Y.; Kawahara, T. The Au family, a novel short interspersed element (SINE) from Aegilops umbellulate. *Theor. Appl. Genet*. **2001**, 102, 463–470.
51. Mao, H.; Wang H. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics*. **2017**, 33, 743-745.
52. Temin,H.M. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons and retrotranscripts. *Mol Biol Evol*. **1985**, 2,455-468.
53. Richert-Poggeler, K.R.; Shepherd, R.J. Petunia vein clearing virus: A plant pararetrovirus with the core sequence of an integrase function. *Virology* **1997**, 236, 137–146.
54. Teycheney, P.-Y.; Geering, A.D.W.; Dasgupta, I.; Hull, R.; Kreuze, J.F.; Lockhart, B.; Muller, E.; Olszewski, N.; Pappu, H.; Pooggin, M.; et al. ICTV Virus taxonomy profile: Caulimoviridae. *J. Gen. Virol.* **2020**, 101, 1025–1026.
55. Jakowitsch, J.; Mette, M.F.; van Der Winden, J.; Matzke, M.A.; Matzke, A.J. Integrated pararetroviral sequences define a unique class of dispersed repetitive DNA in plants. *Proc. Natl. Acad. Sci. USA* **1999**, 96, 13241–13246.
56. Geering, A.D.W.; Maumus, F.; Copetti, D.; Choisne, N.; Zwickl, D.J.; Zytnicki, M.; McTaggart, A.R.; Scalabrin, S.; Vezzulli, S.; Wing, R.A.; et al. Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nat. Commun*. **2014**, 5, 5269.
57. Robertson, D. S. Characterization of a Mutator system in maize. *Mutat. Res.* **1978**, 51, 21–28.
58. Feschotte, C.; Pritham, E.J. DNA transposons and the evolution of the eukaryotic genomes. *Annu. Rev. Genet.* **2007**, 41, 331–368
59. Yu, Z.; Wright, S. I.; Bureau, T. E. Mutator-like elements in Arabidopsis thaliana. Structure, diversity and evolution. *Genetics* **2000**, 156, 2019–2031.
60. Lisch, D. Mutator transposons. *Trends Plant Sci.* **2002**, 7, 498–504.
61. Gao, D.; Caspersen, A,M.; Hu, G.; Bockelman, H.E.; Chen, X. A novel mutator-like transposable elements with unusual structure and recent transpositions in barley (*Hordeum vulgare*). *Front Plant Sci*. **2022**, 13, 904619.
62. Jiang, N.; Bao, Z.; Zhang, X.; Eddy, S.R.; Wessler, S.R. Pack-MULE transposable elements mediate gene evolution in plants. Nature 2004, 431, 569–573.
63. Peterson, P. A. A mutable pale green locus in maize. *Genetics* **1953**, 38, 682–683.
64. McClintock, B. Mutations in maize and chromosomal aberrations in Neurospora. *Year B. Carnegie Inst. Wash.* **1954**, 53, 254–261.
65. Zabala, G.; Vodkin, L. A putative autonomous 20.5 kb-CACTA transposon insertion in an *F3'H* allele identifies a new CACTA transposon subfamily in *Glycine max*. *BMC Plant Biol.* **2008**, 8, 124.
66. Kawasaki, S.; Nitasaka, E. Characterization of *Tpn1* family in the Japanese morning glory: *En*/*Spm*-related transposable elements capturing host genes. *Plant Cell Physiol.* **2004**, 45, 933–944.
67. McCLINTOCK, B. The origin and behavior of mutable loci in maize. *Proc Natl Acad Sci U S A*. **1950**, 36, 344-55.
68. Calvi, B.R.; Hong, T.J.; Findley, S.D.; Gelbart, W.M. Evidence for a common evolutionary origin of inverted repeat transposons in Drosophila and plants: hobo, Activator, and Tam3. *Cell*. **1991,** 66, 465-471.
69. Atkinson, P.W. hAT transposable elements. *Microbiol. Spectr*. **2015**, 3, 1–27.
70. Rubin, E.; Lithwick, G.; Levy, A.A. Structure and evolution of the hAT transposon superfamily. *Genetics* **2001**, 158, 949–957.

71. Essers, L.; Adolphs, R.H.; Kunze, R. A highly conserved domain of the maize activator transposase is involved in dimerization. *Plant Cell* **2000**, 12, 211–224.
72. Zhang, X.; Feschotte, C.; Zhang, Q.; Jiang, N.; Eggleston, W.B.; Wessler, S.R. P Instability Factor: An Active Maize Transposon System Associated with the Amplification of Tourist-like MITEs and a New Superfamily of Transposases. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 12572–12577.
73. Kapitonov, V.V.; Jurka, J. Molecular paleontology of transposable elements from Arabidopsis thaliana. *Genetica* **1999**, 107, 27–37.
74. Jiang, N.; Bao, Z.; Zhang, X.; Hirochika, H.; Eddy, S.; McCouch, S.R.; Wessler, S.R. An active DNA transposon family in rice. *Nature* **2003**, 421, 163–167.
75. Velanis, C.N.; Perera, P.; Thomson, B.; de Leau, E.; Liang, S.C.; Hartwig, B.; Förderer, A.; Thornton, H.; Arede, P.; Chen, J.; et al. The domesticated transposase ALP2 mediates formation of a novel Polycomb protein complex by direct interaction with MSI1, a core subunit of Polycomb Repressive Complex 2 (PRC2). *PLoS Genet*. **2020**, 16, e1008681.
76. Mao, D.; Tao, S.; Li, X.; Gao, D.; Tang, M.; Liu, C.; Wu, D.; Bai, L.; He, Z.; Wang, X.; et al. The Harbinger transposon-derived gene PANDA epigenetically coordinates panicle number and grain size in rice. *Plant Biotechnol. J*. **2022**, 20, 1154–1166.
77. Emmons, S.W.; Yesner, L.; Ruan, K.; Katzenberg, D. Evidence for a transposon in Caenorhabditis elegans. *Cell*. **1983**, 32, 55–65.
78. Jacobson, J.W.; Medhora, M.M.; Hartl, D.L. Molecular structure of a somatically unstable transposable element in Drosophila. *Proc. Natl. Acad. Sci. USA*. **1986**, 83, 8684–8688.
79. Dupeyron, M.; Baril, T.; Bass, C.; Hayward, A. Phylogenetic analysis of the Tc1/mariner superfamily reveals the unexplored diversity of pogo-like elements. *Mob. DNA* **2020**, 11, 21.
80. Liu, Y.; Yang G. Tc1-like transposable elements in plant genomes. *Mob DNA*. **2014**, 5, 17.
81. Wells, D.J. Tdd-4, a DNA transposon of Dictyostelium that encodes proteins similar to LTR retroelement integrases. *Nucleic Acids Res*. **1999**, 27, 2408-2415.
82. Glockner, G.; Szafranski, K.; Winckler, T.; Dingermann, T.; Quail, M.A.; Cox, E.; Eichinger, L.; Noegel, A.A.; Rosenthal, A. The complex repeats of Dictyostelium discoideum. *Genome Res*. **2001**, 11, 585–594.
83. Gao, D.; Zhao, D.; Abernathy, B.; Iwata-Otsubo, A.; Herrera-Estrella, A.; Jiang, N.; Jackson, S.A. Dynamics of a novel highly repetitive CACTA family in common bean (Phaseolus vulgaris). *G3 (Bethesda)*. **2016**, 6, 2091-2101.
84. Boutanaev, A.M.; Osbourn, A.E. Multigenome analysis implicates miniature inverted-repeat transposable elements (MITEs) in metabolic diversification in eudicots. *Proc Natl Acad Sci U S A*. **2018**, 115, E6650-E6658.
85. Jurka, J.; Kapitonov, V.V. PIFs Meet Tourists and Harbingers: A Superfamily Reunion. *Proc. Natl. Acad. Sci. USA* **2001**, 98, 12315–12316.
86. Grzebelus, D.; Lasota, S.; Gambin, T.; Kucherov, G.; Gambin, A. Diversity and structure of PIF/Harbinger-like elements in the genome of Medicago truncatula. *BMC Genomics*. **2007**, 8, 409.
87. Deprá, M.; Ludwig, A.; Valente, V.L.; Loreto, E.L. Mar, a MITE family of hAT transposons in Drosophila. *Mob. DNA* **2012**, 3, 13.
88. 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 Res*. **2004**, 14, 1861–1869.
89. Kapitonov, V.V.; Jurka, J. Rolling-circle transposons in eukaryotes. *Proc. Natl. Acad. Sci. USA 2001*, 98, 8714–8719.
90. Du, C.; Fefelova, N.; Caronna, J.; He, L.; Dooner, H.K. The Polychromatic Helitron Landscape of the Maize Genome. *Proc. Natl. Acad. Sci. USA* **2009**, 106, 19916–19921.
91. Ou, S.; Su, W.; Liao, Y.; Chougule, K.; Agda, J.R.A.; 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 Biol*. 2019, 20, 275.
92. Riehl, K.; Riccio, C.; Miska, E.A.; Hemberg, M. TransposonUltimate: software for transposon classification, annotation and detection. *Nucleic Acids Res.* **2022**, 50, e64.
93. Baril, T.; Imrie, R.M.; Hayward, A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *Biorxiv* **2022**.06.30.498289.
94. Llorens, C.; Futami, R.; Covelli, L.; Domínguez-Escribá, L.; Viu, J.M.; Tamarit, D.; Aguilar-Rodríguez, J.; Vicente-Ripolles, M.; Fuster, G.; Bernet, G.P.; et al. The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Res.* **2011**, 39, 70–74.
95. Flutre, T.; Duprat, E.; Feuillet, C.; Quesneville, H. Considering transposable element diversification in de novo annotation approaches. *PLoS ONE* **2011**, 6, 0016526.
96. Hayward, A.; Gilbert. C. Transposable elements. *Curr. Biol*. **2022**, 32, R904-R909
97. Goubert, C.; Craig, R.J.; Bilat, A.F.; Peona, V.; Vogan, A.A.; Protasio, A.V. A beginner's guide to manual curation of transposable elements. *Mob DNA*. **2022**, 13, 7.

98. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **1999**, 27, 573–580.
99. SanMiguel, P.; Tikhonov, A.; Jin, Y.K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.S.; Edwards, K.J.; Lee, M.; Avramova, Z.; Bennetzen, J.L. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **1996**, 274, 765–768.
100. Kronmiller, B.A.; Wise, R.P. TEnest: Automated chronological annotation and visualization of nested plant transposable elements. *Plant Physiol*. **2008**, 146, 45–59.
101. Buisine, N.; Quesneville, H.; Colot, V. Improved detection and annotation of transposable elements in sequenced genomes using multiple reference sequence sets. *Genomics*. **2008**, 91, 467-75.
102. Kalendar, R.; Tanskanen, J.; Chang, W.; Antonius, K.; Sela, H.; Peleg, O.; Schulman, A.H. Cassandra retrotransposons carry independently transcribed 5S RNA. *Proc. Natl. Acad. Sci. USA* **2008**, 105, 5833–5838.
103. Yasui, Y.; Nasuda, S.; Matsuoka, Y.; Kawahara, T. The Au family, a novel short interspersed element (SINE) from Aegilops umbellulata. *Theor. Appl. Genet.* **2001**, 102, 463–470.