DANTE and DANTE_LTR: computational pipelines implementing lineage-centered annotation of LTR-retrotransposons in plant genomes

Petr Novák, Nina Hoštáková, Pavel Neumann and Jiří Macas*

Biology Centre, Czech Academy of Sciences, Branišovská 31, České Budějovice, Czech Republic

*Corresponding author

e-mail: macas@umbr.cas.cz

phone: +420 387775516

Fax: +420 385310356

http://w3lamc.umbr.cas.cz/lamc/

Abstract

Long terminal repeat (LTR) retrotransposons represent a predominant class of repetitive DNA elements in the genomes of most plant species. As the number of sequenced plant genomes is growing at an accelerating rate, there is a need for computational tools that enable efficient annotation and classification of LTR retrotransposons in plant genome assemblies. Here, we present DANTE, a computational pipeline for Domain-based ANnotation of Transposable Elements, that performs sensitive detection of these elements based on the sequences of their conserved protein domains. The identified protein domains are then used by the DANTE_LTR pipeline to annotate complete element sequences by searching for their structural features, such as long terminal repeats, in the adjacent genomic regions. Moreover, the utilization of domain sequences enables the classification of elements into phylogenetic lineages, offering a more granular annotation compared to conventional coarse classification methods based on superfamilies. The efficiency and accuracy of this strategy is demonstrated by annotating LTR-retrotransposons in 93 plant genomes and comparing the results with several established pipelines. It is shown that DANTE_LTR is able to find significantly more intact LTR retrotransposons. DANTE and DANTE_LTR are provided as user-friendly Galaxy tools that can be accessed via the public server (https://repeatexplorer-elixir.cerit-sc.cz), installed on local Galaxy instances from the Galaxy tool shed or run from the command line.

Key words: LTR-retrotransposons, protein domains, genome annotation, REXdb database

Introduction

Long terminal repeat (LTR) retrotransposons (LTR-RTs) are a dominant type of repetitive DNA in most higher plant genomes, accounting for up to 75% of their nuclear DNA (Baucom et al. 2009). Differential amplification of LTR-RTs is among the main drivers of genome size evolution (Pellicer et al. 2018), leading to extremely large genomes in species that are permissive to retrotransposon accumulation (Kelly et al. 2015; Novák et al. 2020a). Although LTR-RTs have long been considered genomic parasites, they can also be beneficial to their hosts by providing regulatory genetic elements (Hirsch and Springer 2017), driving rapid genomic changes (Dubin et al. 2018) or being an integral part of specific genomic regions such as centromeres (Neumann et al. 2011; Klein and O'Neill 2018). Investigation of these aspects of LTR-RT biology depends on the availability of genome sequencing data from various plant taxa and the computational tools that utilize these data to identify and accurately annotate LTR-RTs. While the former requirement is largely met thanks to recent innovations in sequencing and assembly technologies, there are still significant limitations in the annotation of mobile elements in plant genome assemblies.

The annotation of LTR-RTs consists of two main tasks: the identification of element sequences in the assembly and their classification. Due to the enormous diversity of elements and the low conservation of their nucleotide sequences between plant taxa, most of the currently used tools use structure-based methods to identify LTR-RTs. They typically search for pairs of direct repeats representing LTRs and then screen suitable candidates for the presence of additional features typical of LTR-RTs, such as primer binding site (PBS), polypurine tract (PPT), target site duplication (TSD) and protein coding regions. This strategy is employed by most of the tools currently used to annotate LTR-RTs, including LTR_STRUC (McCarthy and McDonald 2003), LTR_FINDER (Xu and Wang 2007), LTRharvest (Ellinghaus et al. 2008), LTRdigest (Steinbiss et al., 2009), LtrDetector (Valencia and Girgis 2019) and MGEScan3 (Rho et al. 2007). However, despite their common methodology, these programs can provide different results. While some programs are excellent at identifying true LTR-retrotransposons, they can also produce a significant number of false positives, and their performance often depends on the parameters chosen and the organism studied (Lerat 2010). This has led to the development of post-processing tools such as LTR_retriever (Ou and Jiang 2018) and LTRpred (Drost, 2020), which combine the results of multiple tools to increase the sensitivity and accuracy of element annotations. Nevertheless, the structure-based approaches are computationally intensive and generally less efficient in genomes rich in tandem repeats or nested LTR-RT insertions (Ou and Jiang 2018).

Plant LTR-RTs are divided into two superfamilies, Ty1/copia and Ty3/gypsy, which differ structurally by the order of their protein-coding domains. They are eventually further classified into families representing groups of elements that have significant similarities between their nucleotide sequences (the 80-80-80 rule, (Wicker et al. 2007)). The limitation of this system is that the classification into superfamilies is very broad, whereas families consist of very similar elements that are typically only shared by closely related species. To overcome these limitations, we proposed a classification system based on the phylogenetic analysis of conserved domains extracted from LTR-RT polyprotein sequences (Neumann et al. 2019).

Autonomous LTR-RTs encode a polyprotein that has at least five protein domains: GAG, protease (PROT), reverse transcriptase (RT), ribonuclease H (RH) and integrase (INT). The polyprotein is cleaved by the activity of the PROT domain, releasing mature proteins that are essential for the replication and integration of new element copies into the genome. These protein sequences are conserved so that they can be efficiently detected in LTR-RTs from phylogenetically distant plant species based on similarities. However, they exhibit sufficient variation to distinguish specific lineages within the Ty1/copia and Ty3/gypsy superfamilies. The lineages, established primarily on the basis of phylogenetic analysis of the corresponding protein domains, are further supported by lineage-specific sequence and structural features of the elements. This classification system has been implemented in REXdb, a database of more than 75,000 protein sequences extracted from a representative set of green plant genomes. REXdb aims to provide a reference for the comprehensive annotation of LTR-retrotransposons that is consistent across different plant genomes (Neumann et al. 2019). REXdb was originally developed for use in the RepeatExplorer pipeline (Novák et al. 2020b), but later adapted for LTR-RT classification in other tools such as REPET (Quesneville et al. 2005), TEsorter (Zhang et al. 2022) and InpactorDB (Orozco-Arias et al. 2021). However, none of these tools used conserved protein domains as the primary means of identifying LTR-RTs in genome assemblies.

Here we present a novel, protein domain-centered approach for the identification and annotation of LTR-RTs. Instead of primarily searching for structural features of LTR-RTs, we first detect and annotate their conserved protein domains. To this end, we perform a similarity searches against REXdb and evaluate the resulting hits using the DANTE pipeline. Subsequently, the clusters of domains belonging to the same lineage are identified and structural features of LTR-RTs in their neighborhood are detected using the DANTE_LTR tool. We demonstrate that this approach is efficient for the identification and annotation of intact elements while minimizing the number of false detections and allowing the classification of identified elements at the lineage level.

Materials and Methods

The principle of DANTE and DANTE_LTR

The principle of detecting and classifying protein domains implemented in DANTE is shown schematically in Figure 1 A-C. In the first step, the program lastal (Sheetlin et al. 2014) is used to perform a sensitive and frameshift-tolerant DNA-protein similarity search of the query sequence in the REXdb. The regions of the query sequence covered by similarity hits that exceed an initial threshold are then recorded. Due to the redundancy of entries in the database and the high search sensitivity, most retrotransposon coding regions in the query sequence are covered by multiple hits. For each of these regions, the hit with the highest similarity score is identified and used to calculate the threshold for the classification score, which is 80% of the similarity score of the best hit (Figure 1A). The hits with a score above the threshold (“top hits”) are then used in the classification step of DANTE. Using multiple similarity hits in the classification step improves the robustness of the classification as it minimizes the classification

error caused by false similarity hits.

In the classification step, DANTE uses a hierarchical classification system for mobile elements implemented in REXdb (Supplementary Figure S1). If the classification of all top hits points to a single lineage, this classification is reported as final (this case is illustrated by alignment 1 in Figure 1B,C). If the top hits contain sequences from multiple lineages, the next higher classification level (formally the lowest common ancestor) that includes all top hits is reported (alignment 2 in Figure 1B,C). DANTE outputs the positions, classifications and protein sequences of the identified domains as GFF3 and FASTA files.

The algorithm used by DANTE_LTR to identify nucleotide sequences of complete LTR-RTs is shown schematically in Figure 1D-H. First, the output of DANTE is analyzed to find clusters of adjacent protein domains that are likely to belong to the gag-pol coding region of a single element (Figure 1D-F). Such clusters should fulfill the following criteria: (1) all protein domains within the cluster must be classified into the same lineage; (2) they must be in the same orientation (strand); (3) the domains must be in the correct order; and (4) the distances between neighboring domains must not exceed specified threshold. This threshold and the typical number and order of domains were determined for each lineage by analyzing all elements in the REXdb database. All complete LTR-RTs are expected to contain GAG, PROT, RT, RH and INT domains, while some lineages encode additional domains, such as chromodomains (CHD and CHDCR) from chromoviruses (Neumann et al. 2011) or ancestral RNase H (aRH) from Tat elements (Neumann et al. 2019). Clusters with two or more missing domains are excluded from further analysis as they likely represent truncated, rearranged or non-autonomous elements.

In the next step, DANTE_LTR analyzes the nucleotide sequences of the genome regions upstream and downstream of the domain clusters for the presence of LTRs (Figure 1G). The size of the analyzed flanking regions is limited to 2-14 kb, depending on the typical size distribution of LTR lengths of different lineages, increased by 25% to account for possible size variations. If the flanking regions contain additional annotated protein domains indicating the presence of another mobile element, they are truncated to exclude these domains. The nucleotide sequences of the 5' and 3' flanking regions are then compared using BLASTN to identify direct repeats. The direct repeat that is closest to the gag-pol coding region and has at least 80% of the shortest LTR length for a given LTR-RT lineage in the REXdb database is considered the LTR sequence. If present, the TG/CA boundaries typical for long terminal repeats are also reported. Finally, the target site duplication (TSD) is detected as a short (typically 5 bp) direct repeat surrounding the element sequence, and the primer binding site (PBS) of the element is identified based on similarity to a custom database of plant tRNA sequences (the database is available as part of the DANTE_LTR GitHub repository).

Implementation

The DANTE and DANTE_LTR pipelines were implemented in Python and R programming languages and are available as conda packages. The pipelines run under the Linux operating system with a minimum hardware configuration of 8 CPUs and 32 GB RAM.

DANTE is designed to analyze FASTA-formatted nucleotide sequences ranging from a few

hundred nucleotides up to whole genome assemblies. The resulting GFF3 file contains the sequence coordinates of the identified protein domains and lists for each domain its type and classification, the name of the REXdb entry that provided the best similarity hit, and the translation of the annotated region into a protein sequence. It should be noted that DANTE reports all types of protein domains present in REXdb, which in addition to LTR-RTs include protein domains encoded by LINEs, DNA transposons and helitrons (https://doi.org/10.5281/zenodo.10160280). In addition to its full output, DANTE also generates a GFF3 file containing only domains that fulfill strict filter criteria (45% similarity over at least 80% of the reference sequence, with maximum of 3 frameshifts or stop codons per 100 amino acids allowed) and the corresponding protein sequences in FASTA format. While the filtered output is intended for specific downstream analyzes, such as the generation of phylogenetic trees of the domains, the full output of DANTE is used for the annotation of LTR-RTs implemented in DANTE_LTR.

DANTE_LTR should be executed with the same input sequences as DANTE, but requires a GFF3 file with the complete output of DANTE as additional input. It produces a GFF3 output file with detailed annotations of the identified LTR-RTs and a summary table with the numbers of the identified elements. Optionally, a graphical summary can be created that gives a detailed overview of the structure and size distribution of the LTR-RTs and their parts (Supplementary File S1).

Genome assemblies used for testing and benchmarking

DANTE_LTR was tested with the genome assemblies of maize (*Zea mays*) and rice (*Oryza sativa*) which have high-quality repeat annotations that can be used as reliable references. The reference annotation of the maize assembly (acc.no. GCF_000005005.1) was downloaded from MaizeGDB (https://download.maizegdb.org/B73_RefGen_v3/). The reference annotation of the rice genome (acc.no. GCF_001433935.1) was generated as recommended (Ou et al. 2019) using the standard library v7.0.0 obtained from https://github.com/oushujun/EDTA and RepeatMasker version 4.1.1 with the parameters '-e ncbi -pa 36 -q -no_is -norna -nolow -div 40 -cutoff 225'.

Since the repeat annotations of both reference assemblies contained overlapping features resulting in multiple annotations being assigned to a portion of the genomic repeat regions, these conflicting annotations were resolved as follows. The overlaps between "Ty1/copia" and "Ty3/gypsy" were converted to "LTR-RT unspecified", while other types of conflicting annotations (e.g., DNA transposon versus LTR-RT) were converted to "Unknown". If the annotations of overlapping features were identical, they were merged into a single feature to avoid redundancy. These corrections were made using a custom script that ensured that each base in the reference assembly was not annotated more than once (each base was either annotated uniquely or left unannotated). To compare the reference annotation with the annotations generated by the tested methods, each base in the reference genome was assigned to one of the following exclusive categories: "Ty1/copia", "Ty3/gypsy", "LTR-RT unspecified", "Other" (includes all repeat types other than LTR-RTs) and "Not Annotated".

To evaluate the performance of DANTE_LTR in different plant taxa, additional tests were

performed on a set of 91 plant genome assemblies with available LTR-RT annotations selected from the dataset compiled by Zhou et al. (Zhou et al. 2021). Only the GenBank accessions that could be unambiguously matched with the annotations provided by Zhou 2021 were used for testing. The list of analyzed species and the links to their assembly and annotation data can be found in Suppl. Tab. S1.

Performance comparison of DANTE_LTR to Inpactor2 and EDTA

For comparison with other LTR-RT annotation tools, DANTE (ver. 0.1.9) was run with default settings. DANTE_LTR (ver. 0.3.5) was run with the option "-M 1" to tolerate one missing protein domain when identifying intact LTR-RT elements. While DANTE_LTR identifies both intact and partial LTR-RTs elements, when comparing DANTE_LTR with other tools, the partial elements missing LTR sequences were excluded from the analysis. Inpactor2 (commit version 3bd8954) (Orozco-Arias et al. 2022) was run with the options "-t 5 -m 35000 -C 3". The option "-m" was used to set the maximum length of the LTR-RTs to the same value as used in DANTE_LTR. The option -C 3 of Inpactor2 was used to perform the detection process in three cycles. This multi-cycle analysis was designed to mitigate the problem of potential splitting of elements when the analyzed sequences are split into partitions (Orozco-Arias et al. 2022). LTR_retriever was used as implemented in the EDTA pipeline (ver. 2.0.1) (Ou et al. 2019) with the parameters "--species others --overwrite 0 --step all --sensitive 1".

The annotations of complete elements generated by DANTE_LTR, Inpactor2, and EDTA were preprocessed to remove overlapping features within each output file and convert them into a unified classification containing only the superfamily-level categories "Ty1/copia", "Ty3/gypsy", and "LTR-RT unspecified". This was necessary to make the annotations comparable, as EDTA, unlike DANTE_LTR and Inpactor2, does not provide a classification of the identified elements at lineage level. The resulting modified annotations were analyzed to determine the total number of elements identified by each tool and the proportion of elements that were simultaneously identified by different tools (defined as annotations that overlap >90% of their length).

To compare the annotations generated by each tool with the reference repeat annotations of rice and maize genomes, each base in the respective genome was assigned to one of four categories, depending on the match between the tested and reference annotation: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). The number of bases assigned to each category was used to calculate the benchmarking metrics of sensitivity, specificity, accuracy, false discovery rate (FDR), precision, and F1 score (see Supplementary Figure S2) using custom R scripts.

To perform the similarity-based annotation of LTR-RTs, reference databases containing nucleotide sequences of elements identified by individual tools in the previous analysis were compiled and used for the RepeatMasker search with the parameters '-e ncbi -pa 36 -q -no_is -norna -nolow -div 40 -cutoff 225'. The resulting annotations were processed and evaluated as described above.

The benchmarking of the execution times of the analyses was carried out on a computer with Intel(R) Xeon(R) CPU E5-2630 v4 @ 2.20GHz, 32 CPU, 64 GB RAM under the operating

system Ubuntu 16.4 Linux. In addition to running the programs as described above, Inpactor2 was also run with the -C 1 option to use the same settings as in (Orozco-Arias et al. 2022). Benchmarking was performed with whole genome assemblies of *O. sativa* and *Z. mays* as well as with the centromere 6 assembly of *Pisum sativum* ((Macas et al. 2023), GCA_947076115.1).

Results

We tested the performance of DANTE and DANTE_LTR on a set of assemblies from 93 species of flowering plants (Magnoliopsida). In the first phase of testing, we analyzed rice and maize genome assemblies to evaluate the sensitivity and accuracy of LTR-RT identification. We selected these assemblies because their detailed and high-quality annotations of all types of repeats provided suitable references for benchmarking. In the second phase of testing, we examined the ability of DANTE_LTR to identify LTR-RTs in a group of 91 species that differ in genome assembly size (42 - 19,000 Mb) and represent different taxonomic groups of flowering plants (Suppl. Tab. S1). The assemblies of these species were previously analyzed for the presence of LTR-RTs by Zhou et al. using a combination of LTRharvest and LTRdigest (Zhou et al. 2021). We also compared the performance of DANTE_LTR with the results of two established pipelines run with the same set of assemblies: Inpactor2 (Orozco-Arias et al. 2022), which uses a machine learning approach, and Extensive *de-novo* TE Annotator (EDTA), which uses a combination of several structure-based approaches for LTR-RT annotation (Ou et al. 2019).

Evaluation of DANTE_LTR using reference annotations of rice and maize genomes

The analysis of the rice and maize genome assemblies by DANTE_LTR led to the identification and annotation of 2771 and 48,276 complete elements, respectively. Elements were considered complete if they contained both LTRs and an internal protein-coding region in which conserved protein domains were recognized by DANTE in the correct order and orientation, with the tolerance that one domain was missing. Primer binding sites and target site duplications were also detected and annotated. Elements were classified into phylogenetic lineages of LTR-RTs defined in REXdb (REF). This revealed that elements from individual lineages occur with different frequencies in the analyzed genomes (Tab. 1 and Supplementary Tab. S2).

DANTE_LTR not only provides the LTR-RT annotation in the form of GFF3 files, but also generates a summary report and a graphical representation of various features of the detected elements. In addition to the number of elements identified (Tab. 1 and Supplementary Tab. S2), it provides an overview of their consensus structures (Figure 2A) and the size and structural variations of elements within individual lineages (Figure 2B-D). Each lineage is also analyzed with respect to the sequence divergence of LTRs in individual elements and the type of primer binding site (Figure 2E-F). The complete summary reports for rice and maize, including data for all identified LTR-RT lineages, can be found in the Supplementary File S1.

The comparison of the LTR-RT annotations generated by DANTE_LTR with the reference annotations showed a high specificity and a low false discovery rate (Figure 3A). Compared to

the results of Inpactor2 and EDTA, DANTE_LTR outperformed both in terms of the number of complete elements identified (Figure 3B). In addition to the elements that were also recognized by the other tools, there were 31.8 % and 39.7 % elements identified exclusively by DANTE_LTR in rice and maize assemblies, respectively (Figure 3C). Comparison with the reference annotations for these species showed that the higher number of elements identified by DANTE_LTR was due to its higher sensitivity compared to other methods, combined with a consistently low false discovery rate (1.4 % and 6.2 % for rice and maize, respectively). In addition, the better F1 score of DANTE_LTR indicates that the algorithm provides a balanced trade-off between precision and sensitivity while achieving higher overall annotation accuracy (Figure 3A, Supplementary Table S3).

Since the tools developed for the identification of complete elements are in principle unable to annotate non-autonomous elements and fragmented LTR-RT sequences, they often include an additional annotation step based on sequence similarity. This approach was tested using RepeatMasker similarity search with reference libraries compiled from the nucleotide sequences of the complete elements identified in the previous analysis (Fig. 4A-B). The sensitivity of LTR-RT sequence detection was slightly lower for the reference libraries generated from the DANTE_LTR output than for the libraries generated with EDTA and Inpactor2. It was 75 %, 96 % and 91 % for rice and 84 %, 85 % and 94 % for the maize genome, respectively. However, the higher sensitivity of EDTA and Inpactor2 was associated with a 2-4-fold higher false discovery rate (Fig. 3C). An important factor affecting the specificity of the similarity-based annotation was the proportion of misclassified sequences in the reference libraries. Although these contaminations were minor, they led to a considerable increase in false positives in the subsequent annotation. This effect was observed in all three pipelines, but was least pronounced in DANTE_LTR, which maintained the lowest false discovery rate (Supplementary Figure S3).

LTR-RT detection in a set of 91 species of flowering plants

Since the performance of DANTE_LTR depends on the recognition and accurate classification of the retrotransposon protein domains, we first verified that the domains reported by DANTE were unambiguously assigned to the specific LTR-RT lineages defined in the REXdb database (see Supplementary Figure S1). The opposite case, in which a large proportion of domains would only be assigned to a general category as a superfamily, would indicate that the corresponding elements from phylogenetically diverse plant species are not represented in REXdb and therefore cannot be efficiently identified. However, this was ruled out by the observation that the majority of the 21,054,179 identified domains could be unambiguously classified for both the Ty1/copia (97.7%) and Ty3/gypsy (97.0%) lineages (Supplementary Figure S4A). The number of identified domains generally increased with increasing assembly size, reflecting a higher proportion of LTR-RTs in species with larger genomes (Supplementary Figure S5).

Subsequent analysis of the identified domains and their adjacent genomic regions by DANTE_LTR resulted in the annotation of 819,175 complete LTR-retrotransposon sequences (67 to 166,487 per species). The complete element sequences were defined by the presence of

LTRs and the gag-pol region. Of the protein domains identified, 38.7% were assigned to complete elements of the Ty1/copia lineage and 34.8% to the Ty3/gypsy lineage, with the remaining domains either representing solo domains or being part of partial elements (Supplementary Figure S4B). A strong positive correlation was found between the number of protein domains and the number of complete elements identified in individual species, consistent across all lineages with an R-squared value between 0.8 (Bianca lineage) and 0.97 (Retand lineage) (Supplementary Figure S6).

Compared to the other tools, DANTE_LTR identified significantly more complete elements than EDTA and Inpactor2 for both, Ty1/copia and Ty3/gyspy (Figure 5A-B). The number of elements identified was comparable to that reported by Zhou et al. (Zhou et al. 2021). However, only 43-44% of elements were identified simultaneously by both, while 17.3-17.7% of elements were reported only by DANTE_LTR and 11.2-15.8% were specific to the Zhou et al. annotation (Figure 5C-D). Similar partial overlap was found in the annotations of the other two tools, such that only 5.1-9.1% of the elements were common to all four annotations (Figure 5C-D).

Execution time

Execution time benchmarking of DANTE_LTR, EDTA and Inpactor2 was performed using rice and maize genome assemblies, with the addition of the 178 Mb assembly of the centromeric region of *Pisum sativum* chromosome 6 (Macas et al. 2006). This region is rich in tandemly organized satellite DNA, which is known to pose a significant challenge for structure-based tools for LTR-RT annotation. DANTE_LTR proved to be faster than EDTA and Inpactor2 in all three assemblies, with one exception where Inpactor2 was run without iterations (option -C 1), resulting in the fastest analysis of the maize assembly (Table 2). However, when Inpactor2 was run with three iterations as recommended (option -C 3), there was a significant increase in execution time, particularly pronounced when analyzing the maize genome (with a 36-fold increase), most likely due to the increased complexity of output processing. However, apart from this exceptional case, the execution times of the three tools remained relatively comparable and fluctuated in a range between 1.3 and 3.7 times. Measured in megabases of analyzed assemblies per minute (Mbp/min), the analysis of rice was significantly faster for all tools compared to the LTR-retrotransposon-rich genomes of maize and *P. sativum*. Remarkably, DANTE_LTR proved to be most efficient in analyzing the satellite DNA-rich centromere of *P. sativum* (Table 2).

Discussion

In this work, we have introduced and validated a new approach to annotate LTR-RTs that is centered around their most conserved protein-coding regions. Besides the high sensitivity of element detection, the sequences of conserved protein domains offer the additional advantage of allowing classification of elements into phylogenetic lineages. This classification is more detailed than the commonly used superfamily classification (Ty1/copia and Ty3/gypsy) and is nevertheless generally applicable to all plant taxa (Neumann et al. 2019). In addition, there is a wealth of information collected during the annotation of individual elements, such as the element

structure and the properties of the LTR, PBS and TSD regions, and this information is summarized for individual LTR-RT lineages. These data not only provide a quick overview of the variability of element populations in the studied species, but can also be used for a number of downstream studies, such as dating their element insertions based on the divergence of LTR sequences. These features make DANTE_LTR unique among similar tools that usually only provide genome annotation files.

There are two potential limitations of the approach used by DANTE_LTR to identify LTR-RTs. First, elements whose protein sequences are too different from those in the REXdb database may not be identified. However, thanks to the relatively high conservation of LTR-RT protein domains, these domains can be efficiently recognized in the elements from related plant taxa with only a limited number of representative reference sequences. The current version of REXdb (ver. 3.0) fully covers flowering plants (angiosperms), so that the elements from these taxa are reliably detected. However, the other groups of green plants are represented with much fewer elements in REXdb, which may limit the sensitivity of DANTE_LTR and will therefore be the subject of the next updates of the database. The second limitation of DANTE_LTR is the inability to recognize non-autonomous elements whose protein-coding domains are missing. However, this problem can be mitigated by performing a similarity search in the second step of the annotation procedure, using a library of the elements identified in the first step as a reference. In this case, only the lineages of elements that do not contain a complete element in the analyzed genome would be overlooked.

Compared to other tools developed for the annotation of LTR-RTs in plant genomes, DANTE_LTR proved to be more sensitive and identified significantly more elements than EDTA and Inpactor2. In addition, DANTE_LTR proved to be more specific thanks to the evaluation of a combination of many patterns, including the type and arrangement of protein domains, LTRs, TSD, PBS and TG/CA boundaries. A surprising finding was that there was relatively little overlap in the LTR-RT annotations generated by the three pipelines compared, as well as those generated by Zhou at al. (Zhou et al. 2021), with most elements identified by a single or few tools, while elements identified by all tools simultaneously were rare. This could be due to the high sequence and structural variation of the elements in plant genomes, combined with the different efficiency of the compared approaches to identify specific variants of the elements. An important consequence of this result is that it emphasizes the importance of using a combination of several tools implementing different recognition principles for the annotation of LTR-RTs in plant assemblies, rather than a single tool.

It should be noted, however, that benchmarking the performance of genome annotation tools, including those for annotating repeats, inherently suffers from the lack of a definitive 'ground truth'. Traditionally, reference annotations are treated as the best available approximation to the ground truth and serve as benchmarks against which new annotations are compared. However, this approach assumes that the reference annotations are complete and accurate — an assumption that is not always true. Benchmarking can then primarily evaluate the relative performance of the tools against each other and against the existing reference annotations. Although this provides valuable insights into the strengths and weaknesses of the various methods, it does not provide an absolute measure of the reliability of the annotations.

Some of the limitations inherent to structure-based approaches for the annotation of LTR-RTs can be overcome by an additional step that utilizes the detection of nucleotide sequence similarities in a reference library of complete elements(Ou et al. 2019; Orozco-Arias et al. 2022) . This approach also proved to be efficient for the DANTE_LTR output and led to a significant increase in the proportion of annotated LTR-RT sequences due to the detection of non-autonomous or fragmented element sequences. A similar effect was observed for the other two pipelines used for comparison, EDTA and Inpactor2 (Figure 4). Therefore, performing this step can generally be recommended to improve the analysis result. However, the reference library with complete elements must be free of errors, such as chimeric or incorrectly annotated elements. Since some degree of such contamination is unavoidable, the identification and removal of these sequences is crucial to avoid further propagation of erroneous annotations, as shown here (Supplementary Figure S3). Although DANTE_LTR appears to be the least affected by this problem of the three pipelines compared, it is still advisable to perform a thorough validation and curation of the reference library. The graphical summary reports provided for each lineage of identified elements can be used for this purpose. For example, they help to identify and remove elements with atypical structure or unexpectedly long non-coding regions that might contain insertions of unrelated elements such as MITEs or other DNA transposons.

Data Availability

Source code for DANTE and DANTE_LTR programs are available under GPLv3 licence on https://github.com/kavonrtep/dante and https://github.com/kavonrtep/dante_ltr, respectively Source code DANTE and DANTE_LTR for Galaxy implementation are avaliable on https://toolshed.g2.bx.psu.edu/view/petr-novak/dante and https://toolshed.g2.bx.psu.edu/view/petr-novak/dante_ltr. Custom script for analysis of annotation data is avalable on https://github.com/kavonrtep/dante_ltr_evaluation. Output from DANTE_LTR, Inpactor2 and EDTA annotations of 93 analyzed species is available in Zenodo at https://zenodo.org/doi/10.5281/zenodo.10891048 . DANTE and DANTE_LTR programs can be used on public Galaxy server (https://repeatexplorer-elixir.cerit-sc.cz/) which is part of ELIXIR-CZ project.

Author contributions

JM, PNo and PNe designed the study. PNo and NH programmed the pipelines, PNo analyzed data, JM and PNo wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by the ELIXIR-CZ Research Infrastructure Project (LM2023055).

References

Baucom RS, Estill JC, Chaparro C, et al (2009) Exceptional Diversity, Non-Random Distribution, and Rapid Evolution of Retroelements in the B73 Maize Genome. PLoS Genetics 5:e1000732. https://doi.org/10.1371/journal.pgen.1000732

Drost H-G (2020) LTRpred: _de novo_ annotation of intact retrotransposons. Journal of Open Source Software 5:2170. https://doi.org/10.21105/joss.02170

Dubin MJ, Scheid OM, Becker C (2018) Transposons: a blessing curse. Current Opinion in Plant Biology 42:23–29. https://doi.org/10.1016/J.PBI.2018.01.003

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC bioinformatics 9:18. https://doi.org/10.1186/1471-2105-9-18

Hirsch CD, Springer NM (2017) Transposable element influences on gene expression in plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1860:157–165. https://doi.org/10.1016/J.BBAGRM.2016.05.010

Kelly LJ, Renny-Byfield S, Pellicer J, et al (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytologist 208:596–607. https://doi.org/10.1111/nph.13471

Klein SJ, O'Neill RJ (2018) Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Research 26:5–23. https://doi.org/10.1007/s10577-017-9569-5

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104:520–533. https://doi.org/10.1038/hdy.2009.165

Macas J, Ávila Robledillo L, Kreplak J, et al (2023) Assembly of the 81.6 Mb centromere of pea chromosome 6 elucidates the structure and evolution of metapolycentric chromosomes. PLoS Genet 19:e1010633. https://doi.org/10.1371/journal.pgen.1010633

Macas J, Navrátilová A, Koblízková A (2006) Sequence homogenization and chromosomal localization of VicTR-B satellites differ between closely related Vicia species. Chromosoma 115:437–447

McCarthy EM, McDonald JF (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics 19:362–367. https://doi.org/10.1093/bioinformatics/btf878

Neumann P, Navrátilová A, Koblížková A, et al (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mobile DNA 2:4. https://doi.org/10.1186/1759-8753-2-4

Neumann P, Novák P, Hoštáková N, MacAs J (2019) Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mobile DNA 10:. https://doi.org/10.1186/s13100-018-0144-1

Novák P, Guignard MS, Neumann P, et al (2020a) Repeat-sequence turnover shifts fundamentally in species with large genomes. Nat Plants 6:1325–1329. https://doi.org/10.1038/s41477-020-00785-x

Novák P, Neumann P, Macas J (2020b) Global analysis of repetitive DNA from unassembled sequence reads using RepeatExplorer2. Nature Protocols. https://doi.org/10.1038/s41596-020-0400-y

Orozco-Arias S, Humberto Lopez-Murillo L, Candamil-Cortés MS, et al (2022) Inpactor2: a software based on deep learning to identify and classify LTR-retrotransposons in plant genomes. Briefings in Bioinformatics bbac511. https://doi.org/10.1093/bib/bbac511

Orozco-Arias S, Jaimes PA, Candamil MS, et al (2021) InpactorDB: A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning. Genes 12:190. https://doi.org/10.3390/genes12020190

Ou S, Jiang N (2018) LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. Plant Physiol 176:1410–1422. https://doi.org/10.1104/pp.17.01310

Ou S, Su W, Liao Y, et al (2019) Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biology 20:275. https://doi.org/10.1186/s13059-019-1905-y

Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes 9:88. https://doi.org/10.3390/genes9020088

Quesneville H, Bergman CM, Andrieu O, et al (2005) Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLOS Computational Biology 1:e22. https://doi.org/10.1371/journal.pcbi.0010022

Rho M, Choi J-H, Kim S, et al (2007) De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics 8:90. https://doi.org/10.1186/1471-2164-8-90

Sheetlin SL, Park Y, Frith MC, Spouge JL (2014) Frameshift alignment: Statistics and post-genomic applications. Bioinformatics 30:3575–3582. https://doi.org/10.1093/bioinformatics/btu576

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Research 37:7002–7013. https://doi.org/10.1093/nar/gkp759

Valencia JD, Girgis HZ (2019) LtrDetector: A tool-suite for detecting long terminal repeat retrotransposons de-novo. BMC Genomics 20:450. https://doi.org/10.1186/s12864-019-5796-9

Wicker T, Sabot F, Hua-Van A, et al (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973–982. https://doi.org/10.1038/nrg2165

Xu Z, Wang H (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic acids research 35:W265-8. https://doi.org/10.1093/nar/gkm286

Zhang R-G, Li G-Y, Wang X-L, et al (2022) TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Horticulture Research 9:uhac017.

https://doi.org/10.1093/hr/uhac017

Zhou S-S, Yan X-M, Zhang K-F, et al (2021) A comprehensive annotation dataset of intact LTR retrotransposons of 300 plant genomes. Sci Data 8:174. https://doi.org/10.1038/s41597-021-00968-x

Tables

Table 1. The number of elements identified in *Z. mays* genome by DANTE_LTR

Lineage	Number of elements	Element mean length [bp]	LTR mean length [bp]	Number of elements with PBS & TSD	Number of elements with PBS only	Number of elements with TSD only
Ty1/copia|Ale	675	5363	141	307	184	61
Ty1/copia|Angela	382	7804	1274	216	79	5
Ty1/copia|Bianca	118	7777	198	82	12	5
Ty1/copia|Ikeros	23	7199	559	15	3	1
Ty1/copia|Ivana	314	5829	393	193	55	9
Ty1/copia|SIRE	19283	9198	1233	9822	4079	355
Ty1/copia|TAR	776	6380	928	107	112	249
Ty1/copia|Tork	5	4734	296	2	2	0
Ty3/gypsy|chromovirus|CRM	536	6858	569	129	255	18
Ty3/gypsy|chromovirus|Reina	765	5302	379	397	70	187
Ty3/gypsy|chromovirus|Tekay	9121	10331	1955	1843	3493	231
Ty3/gypsy|non-chromovirus|OTA|Athila	1	11313	1441	0	0	0
Ty3/gypsy|non-chromovirus|OTA|Tat|Ogre	373	9269	687	30	15	204
Ty3/gypsy|non-chromovirus|OTA|Tat|Retand	15904	11992	632	9860	2160	402

Table 2. Execution time (min) benchmarks for DANTE_LTR, Inpactor2 and EDTA

Genome		***O. sativa***	***Z. mays***	***P. sativum* (CEN6)**
Accession		GCF_001433935.1	GCF_000005005.1	GCA_947076115.1
Assembly size (Mbp)		374.471	2067.864	177.604
Execution time (min)	**EDTA**	69	687	115
	Inpactor2 (-C 3)	75	15,436	168
	Inpactor2 (-C 1)	25	421	58
	DANTE_LTR	23	529	31
Execution speed (Mbp/min)	**EDTA**	5.44	3.01	1.55
	Inpactor2 (-C 3)	5.03	0.13	1.06
	Inpactor2 (-C 1)	14.98	4.91	3.06
	DANTE_LTR	16.07	3.91	5.67

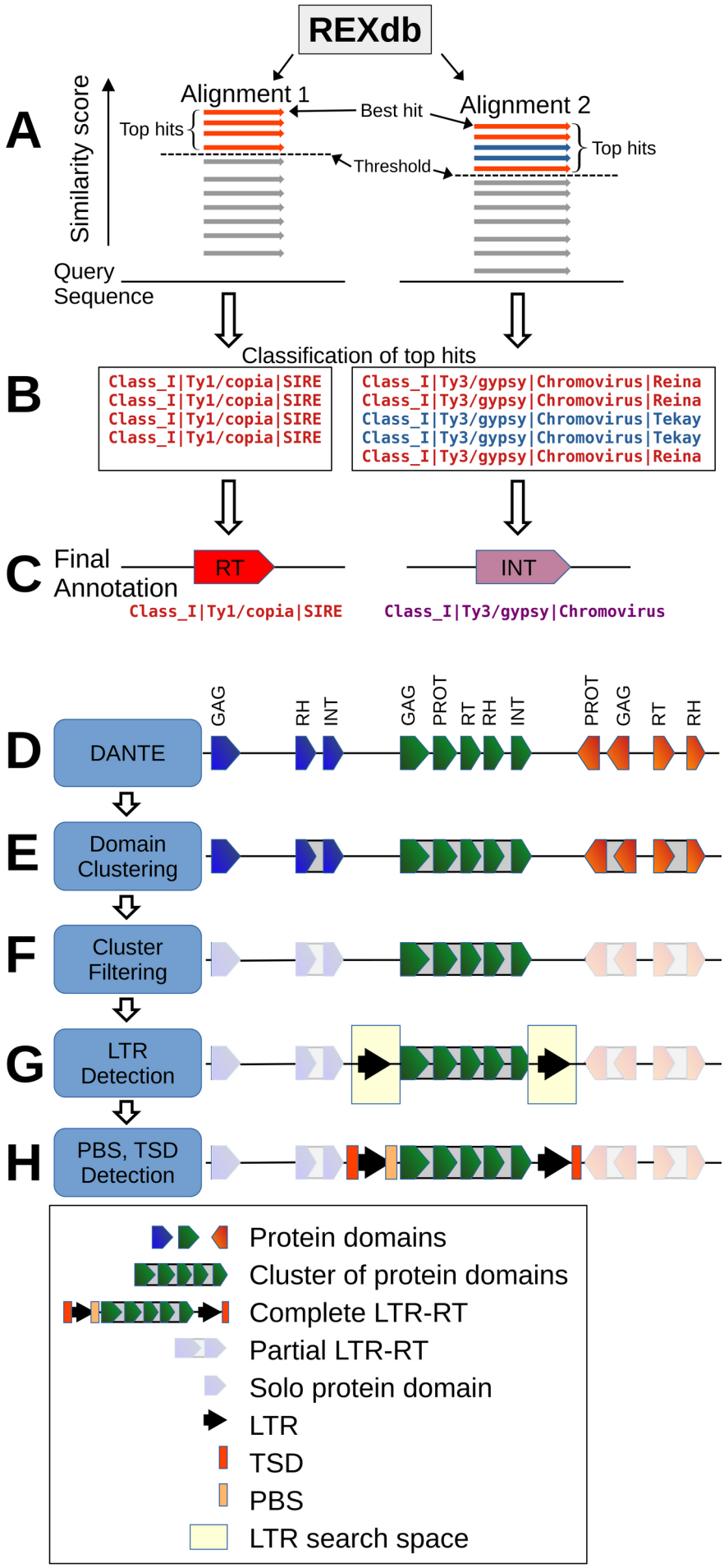

Figure 1. The principle of LTR-retrotransposon annotation by DANTE and DANTE_LTR. See text for details.

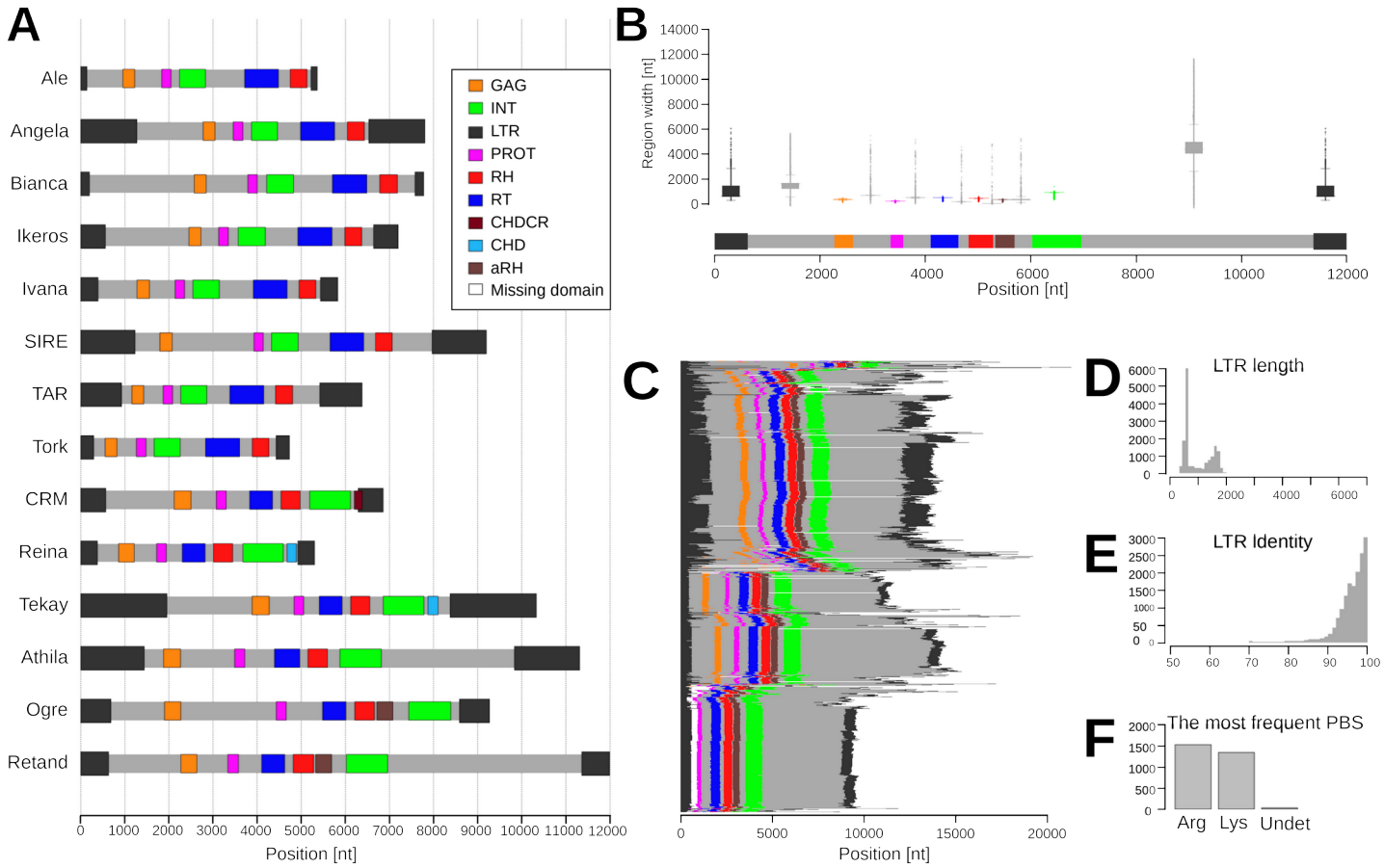

Figure 2. Examples of information contained in the DANTE_LTR summary reports. (**A**) Consensus structure of all LTR-RT lineages identified in the analyzed genome (*Zea mays*). (**B-F**) Detailed information on individual lineages, here represented by the Ty3/Gypsy Retand from the *Z. mays* genome. The report includes the consensus structure and the boxplot of length variations of element features calculated from all identified elements (**B**), whose structures are shown in (**C**) as horizontal lines, with colors indicating the positions of sequence and structural features. The histograms illustrate the size distribution of LTRs (**D**) and their sequence similarities within each element (**E**), as well as the number of occurrences of the most common types of PBSs detected in each lineage (**F**). Complete summary reports are provided as Supplementary File S1.

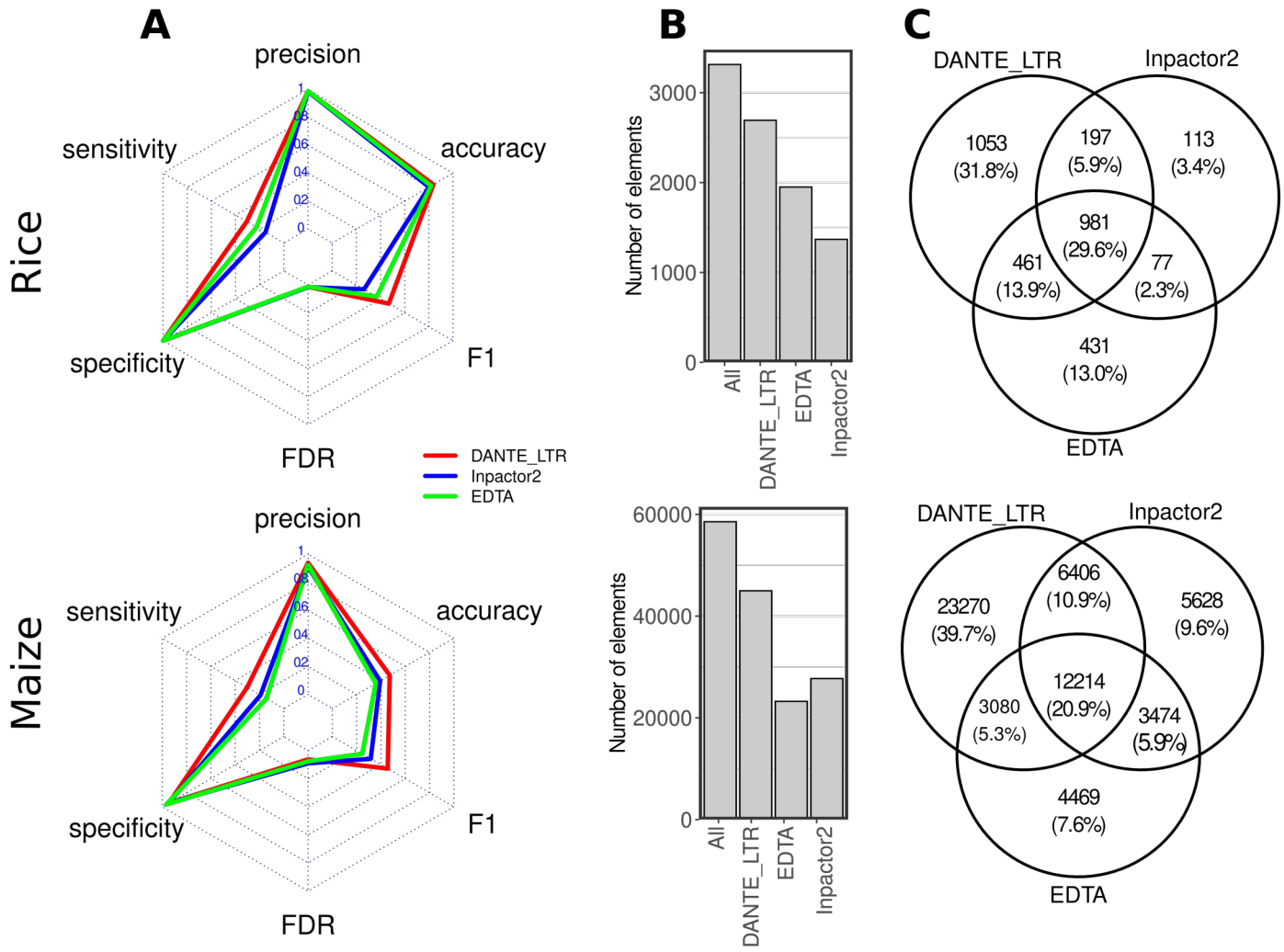

Figure 3. Comparison of LTR-RT annotations generated by DANTE_LTR, Inpactor2 and EDTA for the rice and maize genomes. (**A**) Annotation performance metrics of the three tested pipelines, calculated by comparing their results with the reference annotations. (**B**) The number of identified elements. The bar labeled 'All' corresponds to the set of elements identified by at least one method. **(C)** The number (and percentage) of unique elements and elements identified simultaneously by two or all pipelines. See Supplementary Figure S2 for a detailed explanation of the metrics used in this plot.

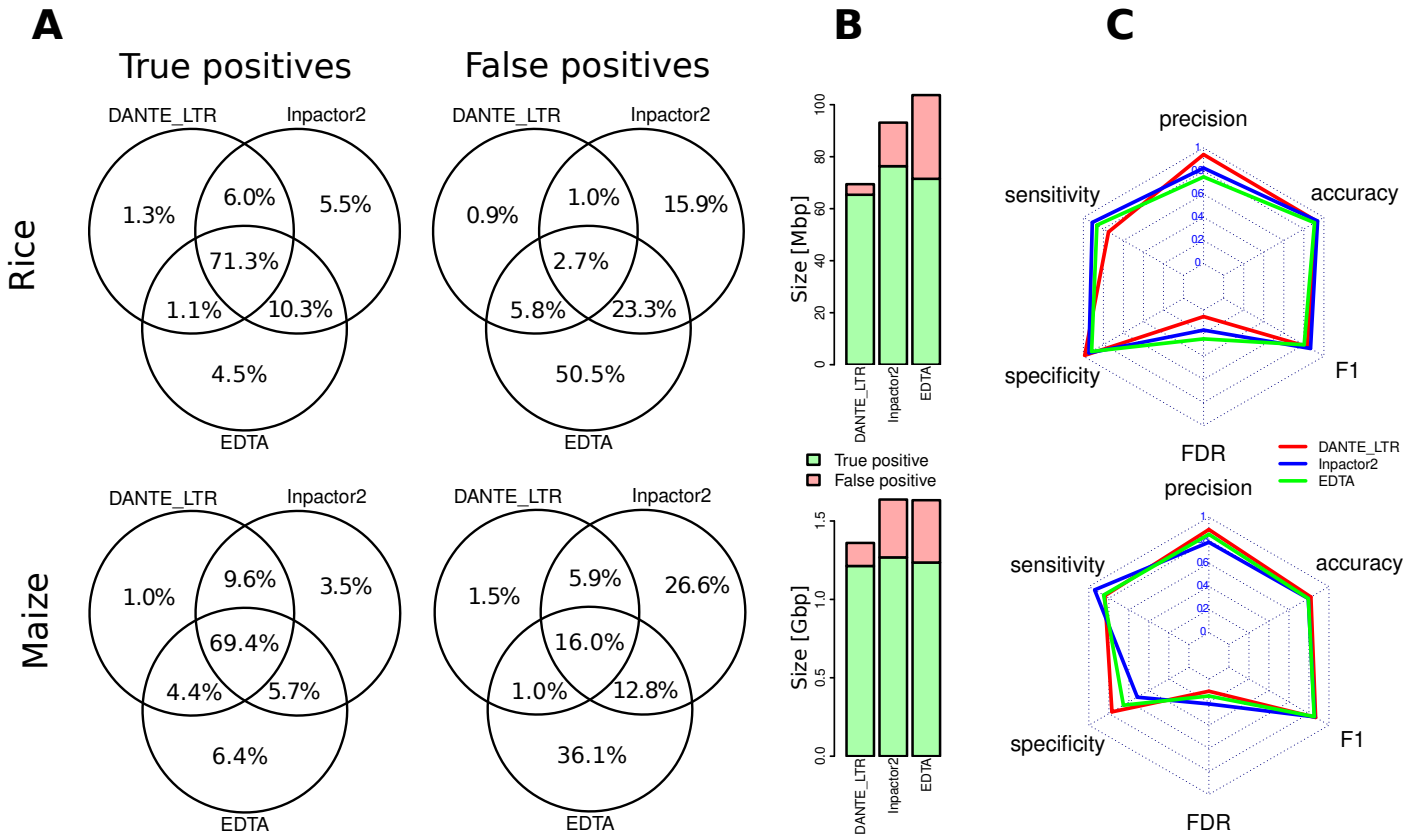

Figure 4. Similarity-based LTR-RT annotations using libraries of elements identified in rice and maize by DANTE_LTR, Inpactor2 and EDTA. The resulting annotations were scored against reference annotations of the rice and maize genomes. (**A**) Overlaps of true- and false-positive annotations generated by the compared pipelines. (**B**) Total length of annotated regions. (**C**) Annotation performance metrics of the three tested pipelines, calculated by comparing their results with the reference annotations. See Supplementary Figure S2 for a detailed explanation of the metrics used in this plot.

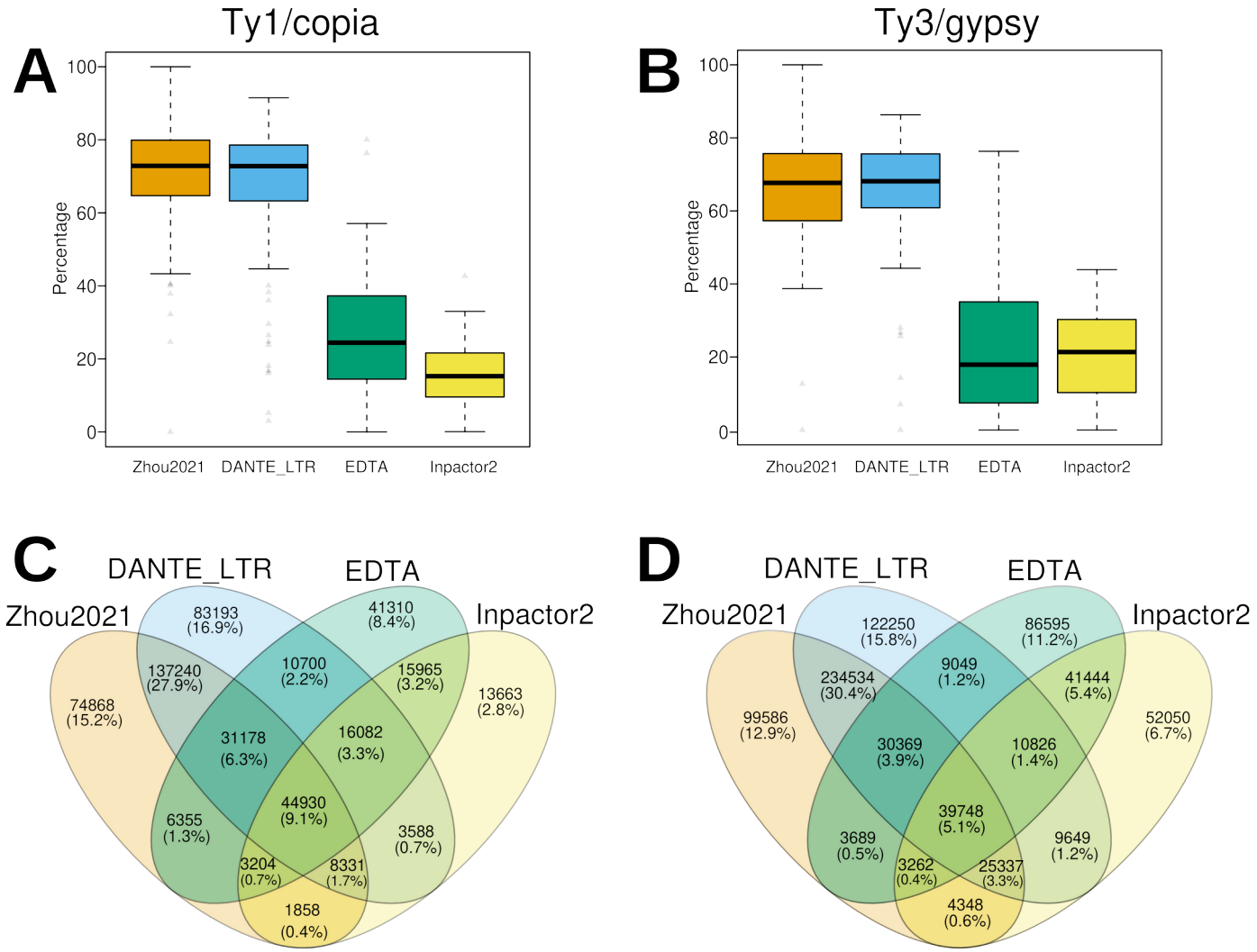

Figure 5. LTR-RT annotation in 91 plant species. Proportions of Ty1/copia (**A**) and Ty3/gypsy (**B**) elements annotated in the analyzed assemblies by the tested pipelines and by Zhou et al.(Zhou et al. 2021) (referred to as 'Zhou2021'). Proportions were calculated as percentages of the combined number of elements identified by all methods. (**C-D**) The number (and percentage) of unique elements and elements identified simultaneously by two or more pipelines calculated separately for Ty1/copia (**C**) and Ty3/gypsy (**D**) elements.