



## Article

# DARTS: An Algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies

Mikhail Biryukov  and Kirill Ustyantsev \* 

Sector of Molecular and Genetic Mechanisms of Regeneration, Institute of Cytology and Genetics SB RAS, 630090 Novosibirsk, Russia; biryukov@bionet.nsc.ru

\* Correspondence: ustyantsev@bionet.nsc.ru

**Abstract:** Retrotransposons comprise a substantial fraction of eukaryotic genomes, reaching the highest proportions in plants. Therefore, identification and annotation of retrotransposons is an important task in studying the regulation and evolution of plant genomes. The majority of computational tools for mining transposable elements (TEs) are designed for subsequent genome repeat masking, often leaving aside the element lineage classification and its protein domain composition. Additionally, studies focused on the diversity and evolution of a particular group of retrotransposons often require substantial customization efforts from researchers to adapt existing software to their needs. Here, we developed a computational pipeline to mine sequences of protein-coding retrotransposons based on the sequences of their conserved protein domains—DARTS (Domain-Associated Retrotransposon Search). Using the most abundant group of TEs in plants—long terminal repeat (LTR) retrotransposons (LTR-RTs)—we show that DARTS has radically higher sensitivity for LTR-RT identification compared to the widely accepted tool LTRharvest. DARTS can be easily customized for specific user needs. As a result, DARTS returns a set of structurally annotated nucleotide and amino acid sequences which can be readily used in subsequent comparative and phylogenetic analyses. DARTS may facilitate researchers interested in the discovery and detailed analysis of the diversity and evolution of retrotransposons, LTR-RTs, and other protein-coding TEs.

**Keywords:** LTR retrotransposons; retroelements; domain annotation; software; automatic pipeline



**Citation:** Biryukov, M.; Ustyantsev, K. DARTS: An Algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies. *Genes* **2022**, *13*, 9. <https://doi.org/10.3390/genes13010009>

Academic Editor: Dariusz Grzebelus

Received: 30 November 2021

Accepted: 17 December 2021

Published: 21 December 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Transposable elements (TEs) are important players in the evolution of genomes [1–4]. The activity of TEs drives genetic diversity, contributes to the establishment of new gene regulatory networks and the rewiring of the existing ones, and can result in the origin of new genes sequestered by the host genome for its functioning [5–7]. The long-term existence and evolution of TEs has resulted in a broad diversity of the mechanisms for their transposition and replication, and the origin of a variety of different structural variants [8,9].

Retrotransposons, a group of TEs that move through a reverse transcription mechanism, are the most ubiquitous TEs in eukaryotic genomes. Due to their propensity to increase in copy number, retrotransposons constitute a substantial portion of the host genome, reaching as high as 80% of the total genome size in some plants [10,11]. Thus, studying retrotransposons is an essential part of understanding plant evolution. The majority of retrotransposons in plants are long terminal repeat (LTR) retrotransposons (LTR-RTs), which are structurally and evolutionarily similar to retroviruses of vertebrates [12,13]. Autonomous, i.e., capable of self-replication, LTR-RTs are complex genetic entities consisting of several protein-coding domains and non-coding regulatory sequences (such as LTRs) that mediate transcription, replication, and integration of the TEs [14–17]. Despite similarities in the general replication mechanism, LTR-RTs are structurally diverse and encode for additional protein domains, which are supposed to fine-tune their life cycle [18–20]. Despite structural differences, the central functional domain of all autonomous retrotransposons,

reverse transcriptase (RT), remains conserved through evolution, allowing unbiased phylogenetic delineation and classification of retrotransposon diversity [9]. Evolution of LTR-RTs and other retrotransposons as individual entities attracts attention by itself, being an example of modular evolution [21,22]. In modular evolution, the main driving force is not a random mutational process, but acquisition, reshuffling, and loss of whole structural elements, such as protein domains and transcriptional enhancers. Therefore, the history of a distinct protein domain in a retrotransposon can be different from the evolution of its core RT domain [21,23,24].

The majority of computational tools developed for the annotation of LTR-RTs in genomic sequences initiate their search from identification of LTRs, and not conserved protein domains [25–27]. Alternative approaches, such as RepeatModeler, first look for any repetitive sequences on the nucleotide sequence level and then try to classify them based on the homology information to known TEs [28]. However, in cases when it is important to search for a specific family of TEs, these methods, apart from being too redundant and computationally time-consuming, may end up with a very high rate of false-negative results, since some TE lineages may be present in a very low number of copies, and some LTR-RT copies may lack well-detectable LTRs. On the other hand, homology-based approaches suffer from the incompleteness of the reference databases [29].

Here, based on our experience in retrotransposon identification [21,23], we developed a new computational tool that takes advantage of the conserved nature of protein domains encoded by retrotransposons—DARTS (an algorithm for Domain-Associated Retrotransposon Search in Genome Assemblies). DARTS uses an open and actively supported database of conserved protein domain sequence profiles instead of relying on databases of representative reference elements. By selecting a certain set of sequence profiles, DARTS can be easily customized for the identification of virtually any group of TEs with a known conserved protein domain sequence, a model of which is present in the database. Additionally, DARTS performs structural annotation and extraction of sequences of the corresponding protein domains. The extracted sequences can be readily used in subsequent comparative and phylogenetic analyses to study the evolution of a particular group of TEs in more detail.

## 2. Materials and Methods

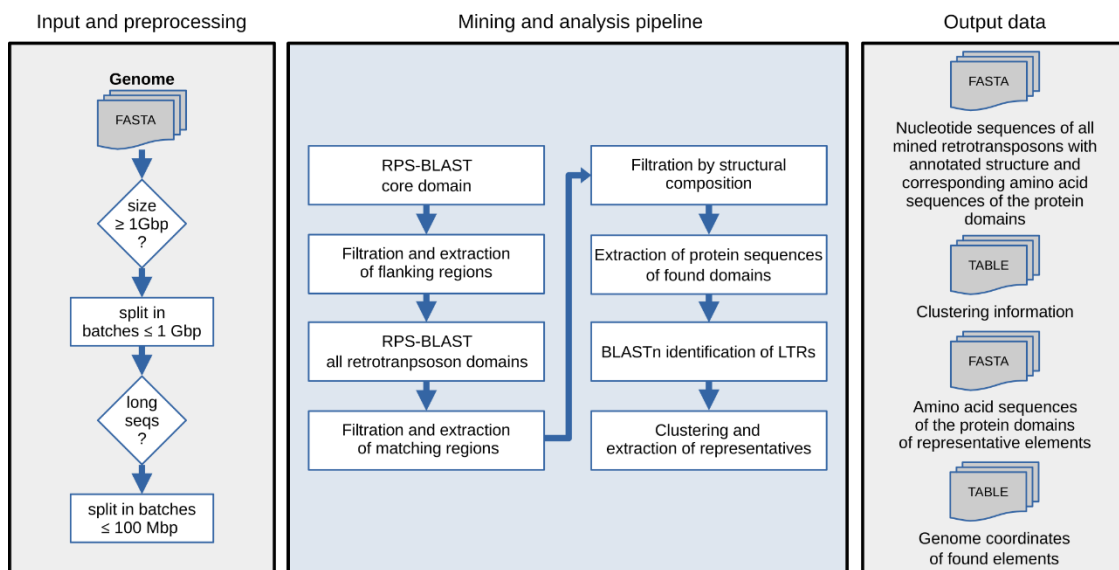
### 2.1. Data Collection

For the analysis, we downloaded genome reference assemblies from the NCBI Genome database (<https://www.ncbi.nlm.nih.gov/genome/>, accessed on 23 November 2021) of four model plant species: *Arabidopsis thaliana* (TAIR10.1, 120 Mbp), *Nicotiana tabacum* (Ntab-TN90, 3736 Mbp), *Selaginella moellendorffii* (GCF\_000143415.4, 212 Mbp), and *Zea mays* (Zm-B73-REFERENCE-NAM-5.0, 2192 Mbp).

### 2.2. Description of the DARTS Pipeline

The DARTS pipeline consists of several scripts written in Python (v 3.6) and Bash programming languages. The scripts, installation, and detailed usage manuals are available on GitHub: [https://github.com/Mikkey-the-turtle/DARTS\\_v0.1](https://github.com/Mikkey-the-turtle/DARTS_v0.1), accessed on 23 November 2021. The general pipeline scheme is shown in Figure 1. A user may choose which parts of the pipeline to execute and can customize every filtering and threshold value presented in the default version, which was originally adapted for identification of LTR-RTs.

Before the analysis, DARTS checks the total genome assembly size, and, if it exceeds 1 Gbp, splits the assembly into several smaller batches to allow for a fast search. If the split is required, the program will attempt to divide the file into individual chromosomes (scaffolds or contigs) without disrupting the original sequences. However, if the genome assembly contains long sequences, such as fully-assembled chromosomes that exceed 1 Gbp in size, the sequences are divided into batches below 100 Mbp creating no more than N-1 breakpoints, where N equals to the number of batches formed from the chromosome. At the same time, chromosomes below 1 Gbp are left intact and put into separate corresponding batches.



**Figure 1.** Principle scheme of the DARTS (Domain-Associated Retrotransposon Search) workflow. Detailed description of each step is in the text.

To identify target protein domains, DARTS uses standalone Reverse PSI-BLAST (RPS-BLAST) from the BLAST+ package [30] supplemented with corresponding multiple sequence alignment protein models, or profiles, obtained from a local copy of the NCBI Conserved Domain Database (CDD) [31] (<https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml>, accessed on 23 November 2021). Two sets of CDD profiles related to a certain group of TEs must be defined by the user prior to the analysis. The first set is a single CDD profile representing a core domain of the TE group (e.g., reverse transcriptase domain specific to LTR-RTs). The second set contains all other additional CDD profiles expected to be found in TEs of interest as well as the first CDD profile. In the first RPS-BLAST search round, the genomic assembly is scanned by the first CDD profile. This results in a set of matches that are pre-filtered by *e-value* ( $1 \times 10^{-3}$ ) and length of the match. The genomic coordinates of a match are identified, and the corresponding nucleotide sequence with flanking regions of 7500 bp in length is extracted for each match. The second RPS-BLAST search utilizes the second user-defined set of CDD profiles and is applied on the extracted sequence regions instead of the whole assembly. Processing of the second RPS-BLAST run results in the structural annotation of TEs of interest and subsequent filtration of the elements by presence of a user-defined set of protein domains. Importantly, when several core domains are present in the same expanded matching region, DARTS will try to delineate them into separate domain assemblies. When a domain match is interrupted by frameshifts or small insertions, DARTS will assemble its parts in a single unit for annotation (Supplementary Figure S1). Amino acid sequences of each of the identified domains are extracted and stored in separate FASTA-formatted files. For LTR-RTs, using the BLASTn tool from the BLAST+ package [30], DARTS will attempt to identify LTRs flanking the first and the last identified protein domains with more than 80% identity that are more than 100 bp but less than 3000 bp in length. Each element obtains a score (%score) based on the number of identified protein domains, length and quality of the matches, presence of uninterrupted open reading frames (ORFs), and LTR identity for LTR-RTs. Each sequence that passed the filtration stage will have a unique name identifier presented in the following format: “%project\_name\_%batch\_%num\_ID|%structure|%LTR\_information|%score”, where %project\_name is the user-defined name of the DARTS run, %batch is the number of the corresponding genome batch-file, %num\_ID is the numerical identifier in the current genome batch-file, %structure is the generalized protein domain-based structure presented for Ty3/gypsy LTR-RTs (e.g., “GAG.Pro.gRT.gRH.INT”), %LTR\_information is shown as LTR%identity-length (e.g., LTR%99.567-232), and %score is the float number.

For the purpose of subsequent comparative and phylogenetic analyses, DARTS can reduce the redundancy of the dataset through clustering using MMseqs2 [32] and subsequent selection of clusters' representatives based on the %score value and structural composition. Clustering information is stored as a tab-separated values table file and can later be reanalyzed using custom criteria. For LTR-RTs, clustering is performed by default using the core and the most conserved domain, reverse transcriptase (RT), with the following default parameters: "easy-cluster -min-seq-id 0.8 -c 0.8". Nucleotide sequences of the representative elements and amino acid sequences of each of their protein domains are deposited in separate FASTA-formatted files. These sequences can later be directly used for multiple sequence alignment generation for subsequent phylogenetic analysis.

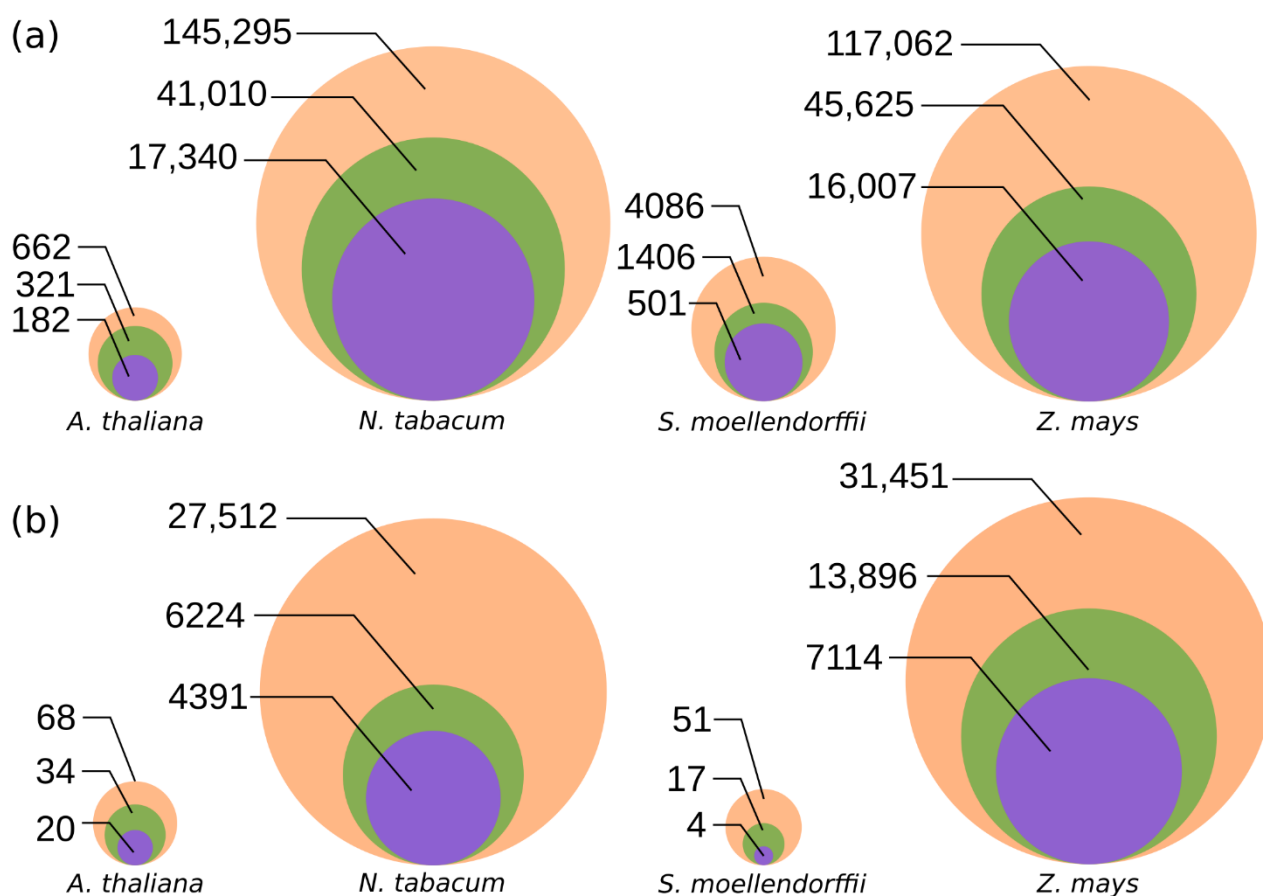
### 2.3. Identification of LTR Retrotransposons Using LTRharvest

To mine LTR-RTs from the selected plant genomes using the de novo LTR-RT prediction tool LTRharvest [26], we ran the program with the following parameters: "-minlenltr 200, -maxlenltr 2000, -mindistltr 3000, -maxdistltr 22000, -similar 85.0, -overlaps no, -mintsd 3, -maxtsd 20". The resulting file with all the hypothetical full-length LTR-RT nucleotide sequences produced by LTRharvest was then processed by DARTS to identify sequences containing the RT domain and to ensure unbiased comparison between both tools. To compare the number of elements uniquely identified by both the DARTS and LTRharvest tools, we performed reciprocal BLASTn searches with a "-max\_target\_seqs 1" parameter.

## 3. Results

Previously, we performed a study on the diversity and evolution of a structurally variable group of Ty3/gypsy plant LTR-RTs—Tat [23,33,34]. Tat LTR-RTs have an additional ribonuclease H domain (aRNH) of the so-called archaeal origin, which is fixed in several positions with regard to other domains in different Tat lineages [23]. In our previous study on Tat [23], we used a conventional tool for the de novo prediction of LTR-RTs—LTRharvest [26]. Later, when doing an independent search using tBLASTn with an aRNH sequence as a query, we found that a substantial fraction of aRNH-containing Tat LTR-RTs were underrepresented in the LTRharvest output. We reasoned that this could be explained by the majority of LTR-RT copies in the studied plant genomes being damaged, fragmented (not intact), and lacking detectable LTRs. The fact that LTRs are used as a starting point for LTR-RT identification in several published software [25–27,35], including LTRharvest (now a part of the most popular de novo repeat identification pipeline RepeatModeler [28]), inspired us to develop a new algorithm that could automatically perform identification of protein-coding TEs and LTR-RTs in particular. We named it Domain-Associated Retrotransposon Search (DARTS), as the initiation of the screen and subsequent structural annotation are based on the prediction of conserved protein domains and not LTR sequences. The basis for DARTS is our experience in semi-automated identification of both protein-coding LTR-RTs and non-LTR retrotransposons [21], as well as conceptually similar approaches performed by other researchers [20,36,37].

For the analysis, we selected reference genome assemblies of four widely used model plant species varying in genome size and TE content (see Section 2.1) and applied DARTS and LTRharvest to identify LTR-RTs. The DARTS search was initiated with the most conserved reverse transcriptase (RT) domain, while the LTRharvest attempts to identify regions flanked by direct repeat sequences of hypothetical LTRs [26]. Using DARTS, we mined 267,105 LTR-RT elements (88,389 with LTRs) in the four studied genomes, while only 34,030 sequences predicted by LTRharvest contained the RT domain sequence after filtration of the 55,658 elements originally predicted (Figure 2A). Importantly, all the 34,030 LTRharvest elements were also predicted by DARTS (Supplementary Table S1), suggesting almost eight times higher sensitivity of the latter.



**Figure 2.** Sensitivity of LTR retrotransposon (LTR-RT) identification by the DARTS and LTRharvest pipelines. Orange circles—number of elements found by DARTS; green circles—number of LTR-RTs found by DARTS with predicted LTRs; purple circles—number of elements found by LTRharvest. Sizes of the circles are proportional to the number of elements with relation to the minimum and the maximum values. Exact numbers of elements are indicated to the left of the circles. Parameters of the LTRharvest search were the same for both the approaches (a,b). (a) Prediction of LTR-RTs by DARTS when the search was initiated from the RT domain; LTRharvest elements were retained if the RT domain was present. (b) Prediction of LTR-RTs by DARTS when the search was initiated from the aRNH domain; DARTS and LTRharvest elements were retained if both the RT and aRNH domains were present.

To exemplify DARTS performance when search is initiated from a different protein domain, we screened for *Tat* LTR-RTs in the same genome assemblies using the aRNH CDD profile in the first round of RPS-BLAST. DARTS found 59,082 elements (20,171 with LTRs), while only 11,529 LTR-RTs were identified by LTRharvest (Figure 2B). This suggests that the overall abundance of *Tat* LTR-RTs in our previous study using LTRharvest [23] was largely underestimated.

It must be noted that potential false-positive matches can be present when only the initial target domain is found. However, the chances of this are low since, during the second step of the RPS-BLAST search, all the domains are re-annotated again, which results in an increase of *e-value* since the size of the database is decreased to a single sequence region. Nevertheless, the false-positive hits can be filtered out on the way to phylogenetic analysis, standing as outliers during clustering and multiple sequence alignment compared to true-positive representatives. Alternatively, whenever it is possible, we would suggest filtering the results of DARTS by the presence of one or two additional domains or regulatory sequences, such as LTRs, to completely avoid the problem. In this study, we found that the number of RT-only containing matches in the RT domain search initiated by DARTS

equaled  $5.8\% \pm 2.5\%$  (mean  $\pm$  standard deviation of the mean). Therefore, this range can be considered as a theoretical stringent upper boundary for the false-positive TEs detected by DARTS.

#### 4. Discussion

Although we have shown examples of DARTS usage for general LTR-RT identification and targeted Tat LTR-RT mining in plants, our primary object of interest, the software can be easily customized for search of other TEs with conserved protein-coding domains in other eukaryotic genomes. For example, *Penelope*-like retroelements can be targeted by search for their specific RT and endonuclease domains [38,39] and DIRS-like retrotransposons by their RT and tyrosine recombinase domains [40,41]. Apart from their specific RT domain, various non-LTR retrotransposon groups have two types of endonucleases and two types of RNH domains [23,37,42]. Cut-and-paste DNA transposons can be identified by the transposase domains, e.g., DNA helicases can be found in *Helitrons* and DNA polymerases in *Mavericks* [43–45].

For non-LTR retrotransposons and DNA transposons, the DARTS initial identification approach is similar to the methods implemented in previously published software, such as MGEScan-non-LTR and TransposonPSI [36,46]. However, DARTS is more advantageous since it can also perform automatic structural annotation and clustering, and its algorithm relies on the actively supported RPS-BLAST tool and the CDD database. Therefore, more sensitive profiles can be used to provide a detailed and targeted annotation of TEs of interest. Additionally, compared to TransposonPSI, DARTS returns amino acid sequences of each of the identified domains without the need for additional parsing, allowing direct transition to phylogenetic analysis.

A part of TE analysis that is not covered by DARTS is the annotation of non-protein coding genes and copies lacking the domain of interest that was used to initiate the search. While annotation of such elements as Short Interspersed Nuclear Elements (SINEs) and Miniature Inverted-repeat Transposable Elements (MITEs) indeed requires a substantially different approach for their identification [28,47], severely damaged copies and sequences, such as solo-LTRs, lacking a domain of interest can still be found by applying nucleotide BLAST or RepeatMasker [30,48] using the copies identified by DARTS as queries. Thus, their number can be accounted for in the genome annotation.

#### 5. Conclusions

Here, we developed a new pipeline for automatic search and structural annotation of protein-coding LTR-RTs and other retrotransposons in genomic sequences. DARTS is beneficial when one is interested in analysis of all the structural diversity of a TE group. We showed that DARTS is almost eight times more sensitive in LTR-RT identification than the de novo tool LTRharvest, which is now included in the widely used RepeatModeler version 2 pipeline [28]. The ease of DARTS customization should facilitate many researchers studying the diversity and evolution of different groups of TEs.

**Supplementary Materials:** The following are available online at <https://www.mdpi.com/article/10.3390/genes13010009/s1>, Figure S1: A principle scheme of fragmented protein domain assembly by DARTS, Table S1: BLASTn matches between the LTR-RTs predicted by LTRharvest against the DARTS LTR-RTs matches.

**Author Contributions:** Conceptualization, M.B. and K.U.; formal analysis and software, M.B.; writing—original draft preparation, M.B. and K.U.; writing—review and editing, M.B. and K.U.; supervision, K.U.; funding acquisition, M.B. and K.U. All authors have read and agreed to the published version of the manuscript.

**Funding:** Work of M.B. on the development of the DARTS software was funded by the Russian Foundation for Basic Research grant 20-34-90114. Work of K.U. and access to cluster computing was supported by the Russian state budget project 0259-2021-0013.

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

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The software and datasets produced in this study are openly available on GitHub [https://github.com/Mikkey-the-turtle/DARTS\\_v0.1](https://github.com/Mikkey-the-turtle/DARTS_v0.1).

**Acknowledgments:** We thank Eugene Berezikov for his valuable comments on the manuscript draft.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Kazazian, H.H. Mobile DNA Transposition in Somatic Cells. *BMC Biol.* **2011**, *9*, 62. [[CrossRef](#)]
2. Solyom, S.; Kazazian, H.H. Mobile Elements in the Human Genome: Implications for Disease. *Genome Med.* **2012**, *4*, 12. [[CrossRef](#)] [[PubMed](#)]
3. Britten, R.J. Transposable Element Insertions Have Strongly Affected Human Evolution. *Proc. Natl. Acad. Sci. USA* **2010**, *107*, 19945–19948. [[CrossRef](#)]
4. Arkhipova, I.R.; Batzer, M.A.; Brosius, J.; Feschotte, C.; Moran, J.V.; Schmitz, J.; Jurka, J. Genomic Impact of Eukaryotic Transposable Elements. *Mob. DNA* **2012**, *3*, 19. [[CrossRef](#)]
5. Deininger, P.L.; Batzer, M.A. Alu Repeats and Human Disease. *Mol. Genet. Metab.* **1999**, *67*, 183–193. [[CrossRef](#)]
6. Volff, J.-N. Turning Junk into Gold: Domestication of Transposable Elements and the Creation of New Genes in Eukaryotes. *BioEssays* **2006**, *28*, 913–922. [[CrossRef](#)]
7. Bennetzen, J.L.; Ma, J.; Devos, K.M. Mechanisms of Recent Genome Size Variation in Flowering Plants. *Ann. Bot.* **2005**, *95*, 127–132. [[CrossRef](#)]
8. 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. [[CrossRef](#)]
9. Eickbush, T.H.; Jamburuthugoda, V.K. The Diversity of Retrotransposons and the Properties of Their Reverse Transcriptases. *Virus Res.* **2008**, *134*, 221–234. [[CrossRef](#)] [[PubMed](#)]
10. 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. [[CrossRef](#)] [[PubMed](#)]
11. Chénais, B.; Caruso, A.; Hiard, S.; Casse, N. The Impact of Transposable Elements on Eukaryotic Genomes: From Genome Size Increase to Genetic Adaptation to Stressful Environments. *Gene* **2012**, *509*, 7–15. [[CrossRef](#)] [[PubMed](#)]
12. Malik, H.S.; Eickbush, T.H. Phylogenetic Analysis of Ribonuclease H Domains Suggests a Late, Chimeric Origin of LTR Retrotransposable Elements and Retroviruses. *Genome Res.* **2001**, *11*, 1187–1197. [[CrossRef](#)] [[PubMed](#)]
13. Hizi, A.; Herschhorn, A. Retroviral Reverse Transcriptases (Other than Those of HIV-1 and Murine Leukemia Virus): A Comparison of Their Molecular and Biochemical Properties. *Virus Res.* **2008**, *134*, 203–220. [[CrossRef](#)]
14. Menéndez-Arias, L.; Sebastián-Martín, A.; Álvarez, M. Viral Reverse Transcriptases. *Virus Res.* **2017**, *234*, 153–176. [[CrossRef](#)]
15. Figiel, M.; Krepl, M.; Park, S.; Poznański, J.; Skowronek, K.; Gołab, A.; Ha, T.; Šponer, J.; Nowotny, M. Mechanism of Polypurine Tract Primer Generation by HIV-1 Reverse Transcriptase. *J. Biol. Chem.* **2018**, *293*, 191–202. [[CrossRef](#)]
16. Schulman, A.H. Hitching a Ride: Nonautonomous Retrotransposons and Parasitism as a Lifestyle. In *Plant Transposable Elements: Impact on Genome Structure and Function*; Grandbastien, M.-A., Casacuberta, J.M., Eds.; Topics in Current Genetics; Springer: Berlin/Heidelberg, Germany, 2012; pp. 71–88, ISBN 978-3-642-31842-9.
17. Sabot, F.; Schulman, A.H. Parasitism and the Retrotransposon Life Cycle in Plants: A Hitchhiker’s Guide to the Genome. *Heredity* **2006**, *97*, 381–388. [[CrossRef](#)] [[PubMed](#)]
18. Malik, H.S.; Eickbush, T.H. Modular Evolution of the Integrase Domain in the Ty3/Gypsy Class of LTR Retrotransposons. *J. Virol.* **1999**, *73*, 5186–5190. [[CrossRef](#)]
19. Rausch, J.W.; Miller, J.T.; Le Grice, S.F.J. Reverse Transcription in the Saccharomyces Cerevisiae Long-Terminal Repeat Retrotransposon Ty3. *Viruses* **2017**, *9*, 44. [[CrossRef](#)] [[PubMed](#)]
20. Novikova, O.; Mayorov, V.; Smyshlyayev, G.; Fursov, M.; Adkison, L.; Pisarenko, O.; Blinov, A. Novel Clades of Chromodomain-Containing Gypsy LTR Retrotransposons from Mosses (Bryophyta). *Plant J.* **2008**, *56*, 562–574. [[CrossRef](#)]
21. Ustyantsev, K.; Blinov, A.; Smyshlyayev, G. Convergence of Retrotransposons in Oomycetes and Plants. *Mob. DNA* **2017**, *8*, 4. [[CrossRef](#)]
22. Lerat, E.; Brunet, F.; Bazin, C.; Capy, P. Is the Evolution of Transposable Elements Modular. *Genetica* **1999**, *107*, 15–25. [[CrossRef](#)]
23. Ustyantsev, K.; Novikova, O.; Blinov, A.; Smyshlyayev, G. Convergent Evolution of Ribonuclease H in LTR Retrotransposons and Retroviruses. *Mol. Biol. Evol.* **2015**, *32*, 1197. [[CrossRef](#)] [[PubMed](#)]
24. Novikov, A.; Smyshlyayev, G.; Novikova, O. Evolutionary History of LTR Retrotransposon Chromodomains in Plants. *Int. J. Plant Genom.* **2012**, *2012*, 874743. [[CrossRef](#)] [[PubMed](#)]
25. Xu, Z.; Wang, H. LTR\_FINDER: An Efficient Tool for the Prediction of Full-Length LTR Retrotransposons. *Nucleic Acids Res.* **2007**, *35*, W265–W268. [[CrossRef](#)]
26. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an Efficient and Flexible Software for de Novo Detection of LTR Retrotransposons. *BMC Bioinform.* **2008**, *9*, 18. [[CrossRef](#)]

27. Lee, H.; Lee, M.; Mohammed Ismail, W.; Rho, M.; Fox, G.C.; Oh, S.; Tang, H. MGEScan: A Galaxy-Based System for Identifying Retrotransposons in Genomes. *Bioinformatics* **2016**, *32*, 2502–2504. [[CrossRef](#)]
28. Flynn, J.M.; Hubley, R.; Goubert, C.; Rosen, J.; Clark, A.G.; Feschotte, C.; Smit, A.F. RepeatModeler2 for Automated Genomic Discovery of Transposable Element Families. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9451–9457. [[CrossRef](#)]
29. Kennedy, R.C.; Unger, M.F.; Christley, S.; Collins, F.H.; Madey, G.R. An Automated Homology-Based Approach for Identifying Transposable Elements. *BMC Bioinform.* **2011**, *12*, 130. [[CrossRef](#)]
30. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and Applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)] [[PubMed](#)]
31. Lu, S.; Wang, J.; Chitsaz, F.; Derbyshire, M.K.; Geer, R.C.; Gonzales, N.R.; Gwadz, M.; Hurwitz, D.I.; Marchler, G.H.; Song, J.S.; et al. CDD/SPARCLE: The Conserved Domain Database in 2020. *Nucleic Acids Res.* **2020**, *48*, D265–D268. [[CrossRef](#)]
32. Steinegger, M.; Söding, J. MMseqs2 Enables Sensitive Protein Sequence Searching for the Analysis of Massive Data Sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [[CrossRef](#)] [[PubMed](#)]
33. Steinbauerová, V.; Neumann, P.; Novák, P.; Macas, J. A Widespread Occurrence of Extra Open Reading Frames in Plant Ty3/Gypsy Retrotransposons. *Genetica* **2011**, *139*, 1543–1555. [[CrossRef](#)]
34. Neumann, P.; Novák, P.; Hošťáková, N.; Macas, J. Systematic Survey of Plant LTR-Retrotransposons Elucidates Phylogenetic Relationships of Their Polyprotein Domains and Provides a Reference for Element Classification. *Mob. DNA* **2019**, *10*, 1. [[CrossRef](#)] [[PubMed](#)]
35. Ou, S.; Jiang, N. LTR\_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiol.* **2018**, *176*, 1410–1422. [[CrossRef](#)] [[PubMed](#)]
36. 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. [[CrossRef](#)] [[PubMed](#)]
37. Smyshlyaev, G.; Voigt, F.; Blinov, A.; Barabas, O.; Novikova, O. Acquisition of an Archaea-like Ribonuclease H Domain by Plant L1 Retrotransposons Supports Modular Evolution. *Proc. Natl. Acad. Sci. USA* **2013**, *110*, 20140–20145. [[CrossRef](#)] [[PubMed](#)]
38. Evgen'ev, M.B.; Arkhipova, I.R. Penelope-like Elements a New Class of Retroelements: Distribution, Function and Possible Evolutionary Significance. *Cytogenet. Genome Res.* **2005**, *110*, 510–521. [[CrossRef](#)]
39. Craig, R.J.; Yushenova, I.A.; Rodriguez, F.; Arkhipova, I.R. An Ancient Clade of Penelope-like Retroelements with Permuted Domains Is Present in the Green Lineage and Protists, and Dominates Many Invertebrate Genomes. *bioRxiv* **2021**. [[CrossRef](#)]
40. Poulter, R.T.M.; Goodwin, T.J.D. DIRS-1 and the Other Tyrosine Recombinase Retrotransposons. *Cytogenet. Genome Res.* **2005**, *110*, 575–588. [[CrossRef](#)]
41. Poulter, R.; Butler, M. Tyrosine Recombinase Retrotransposons and Transposons. *Microbiol. Spectr.* **2015**, *3*. [[CrossRef](#)]
42. Novikova, O.; Fet, V.; Blinov, A. Non-LTR Retrotransposons in Fungi. *Funct. Integr. Genom.* **2009**, *9*, 27–42. [[CrossRef](#)] [[PubMed](#)]
43. Kapitonov, V.V.; Jurka, J. Helitrons on a Roll: Eukaryotic Rolling-Circle Transposons. *Trends Genet.* **2007**, *23*, 521–529. [[CrossRef](#)] [[PubMed](#)]
44. Pritham, E.J.; Putliwala, T.; Feschotte, C. Mavericks, a Novel Class of Giant Transposable Elements Widespread in Eukaryotes and Related to DNA Viruses. *Gene* **2007**, *390*, 3–17. [[CrossRef](#)] [[PubMed](#)]
45. Munoz-Lopez, M.; Garcia-Perez, J. DNA Transposons: Nature and Applications in Genomics. *Curr. Genom.* **2010**, *11*, 115–128. [[CrossRef](#)] [[PubMed](#)]
46. Haas, B. *TransposonPSI: An Application of PSI-Blast to Mine (Retro-) Transposon ORF Homologies*; Broad Institute: Cambridge, MA, USA, 2007.
47. 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. [[CrossRef](#)] [[PubMed](#)]
48. Smit, A.; Hubley, R.; Green, P. RepeatMasker Open-4.0, 2013–2015. Available online: <http://www.repeatmasker.org> (accessed on 23 November 2021).