Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics

Simon Orozco-Arias, Juan Liu, Reinel Tabares-Soto, Diego Ceballos, Douglas Silva Domingues, Andréa Garavito, Ray Ming, Romain Guyot

▶ To cite this version:

Simon Orozco-Arias, Juan Liu, Reinel Tabares-Soto, Diego Ceballos, Douglas Silva Domingues, et al.. Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. Biology, 2018, 7 (2), pp.32. 10.3390/biology7020032 . hal-03472296

HAL Id: hal-03472296
https://hal.umontpellier.fr/hal-03472296v1

Submitted on 9 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

biology

MDPI

Article

Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics

Simon Orozco-Arias [1,†], Juan Liu [2,†], Reinel Tabares-Soto [1], Diego Ceballos [3], Douglas Silva Domingues [4], Andréa Garavito [5], Ray Ming [2,6] and Romain Guyot [1,7,*]

[1] Department of Electronics and Automatization, Universidad Autónoma de Manizales, Manizales 170002, Colombia; simon.orozco.arias@gmail.com (S.O.-A.); rtabares@autonoma.edu.co (R.T.-S.)
[2] FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou 350002, China; relaxljliu@sina.com (J.L.); rming@life.uiuc.edu (R.M.)
[3] Department of Systems and Informatics, Universidad de Caldas, Manizales 170002, Colombia; diego.ceballos@ucaldas.edu.co
[4] Department of Botany, Instituto de Biociências, Universidade Estadual Paulista, UNESP, Rio Claro, SP 13506-900, Brazil; doug@rc.unesp.br
[5] Department of Biological Sciences, Universidad de Caldas, Manizales 170002, Colombia; neagef@gmail.com
[6] Department of Plant Biology, University of Illinois at Urbana-Champaign, Champaign, IL 61801, USA
[7] Institut de Recherche pour le Développement (IRD), CIRAD, Université de Montpellier, Montpellier 34394, France
* Correspondence: romain.guyot@ird.fr
† These authors contributed equally to this work.

Received: 3 May 2018; Accepted: 22 May 2018; Published: 25 May 2018

Abstract: One particular class of Transposable Elements (TEs), called Long Terminal Repeats (LTRs), retrotransposons, comprises the most abundant mobile elements in plant genomes. Their copy number can vary from several hundreds to up to a few million copies per genome, deeply affecting genome organization and function. The detailed classification of LTR retrotransposons is an essential step to precisely understand their effect at the genome level, but remains challenging in large-sized genomes, requiring the use of optimized bioinformatics tools that can take advantage of supercomputers. Here, we propose a new tool: Inpactor, a parallel and scalable pipeline designed to classify LTR retrotransposons, to identify autonomous and non-autonomous elements, to perform RT-based phylogenetic trees and to analyze their insertion times using High Performance Computing (HPC) techniques. Inpactor was tested on the classification and annotation of LTR retrotransposons in pineapple, a recently-sequenced genome. The pineapple genome assembly comprises 44% of transposable elements, of which 23% were classified as LTR retrotransposons. Exceptionally, 16.4% of the pineapple genome assembly corresponded to only one lineage of the *Gypsy* superfamily: *Del*, suggesting that this particular lineage has undergone a significant increase in its copy numbers. As demonstrated for the pineapple genome, Inpactor provides comprehensive data of LTR retrotransposons' classification and dynamics, allowing a fine understanding of their contribution to genome structure and evolution. Inpactor is available at https://github.com/simonorozcoarias/Inpactor.

Keywords: Inpactor; transposable elements; LTR retrotransposons; parallel programming; pineapple; HPC

1. Introduction

Transposable Elements (TEs) constitute the main part of the nuclear DNA content of plant genomes. This is particularly true for large genomes of cereals such as wheat, barley and maize, for which up to 85% of the sequenced DNA is classified into repeated sequences [1]. On the contrary, compact genomes such as those of *Arabidopsis thaliana* (10%) and the carnivorous plant *Utricularia gibba* (3%) show a lesser content of TEs [2], suggesting that their copy numbers may vary drastically and are associated with genome size variation in plant genomes [3]. Occasionally, a rapid increase in copy numbers of a few TE families may lead colossal genome size variations between related species [4]. TEs can be activated through a large panel of biotic and abiotic stresses ([5,6]), suggesting that they could play a significant role in the environmental adaptation of species [7].

Transposable elements are traditionally classified according to their mechanism of transposition [8]: Class I or retrotransposons move through an RNA intermediate via a "copy and paste" mechanism, and Class II or DNA transposons do not use an RNA intermediate and move via a "cut and paste" mechanism. Class I includes LTR (Long Terminal Repeats) retrotransposons and non-LTR retrotransposons, such as LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear Elements), while Class II contains Terminal Inverted Repeat (TIR) DNA transposons and Helitrons. The most common transposable elements in plants genomes are LTR retrotransposons, because they replicate by a "copy and paste" mechanism. They represent 75% of the maize genome [9], 67% of wheat ([1,10]), 55% of *Sorghum bicolor* [11] and 42% of the coffee genome [12]. The sequences of full-length LTR retrotransposons usually carry two coding genes: the GAG gene involved in TE packaging into a virus-like particle and the Pol gene coding for the enzymatic machinery mainly involved in the retro-transcription of the element. LTR retrotransposons are further classified into *Gypsy* and *Copia* super-families according to the position of the integrase domain in the Pol coding region, and they are further separated into lineages and families according to their structural features and domain similarities [8]. Six domains are particularly important for the mobility of the elements. The GAG (Group Specific Antigen) domain is involved in the formation of virus-like particles; the Aspartic Protease (AP) is responsible for processing the polyprotein of the element into smaller proteins; the Reverse Transcriptase domain (RT) is the key enzyme involved in DNA synthesis (using an RNA template); the RNase H domain degrades the RNA template in the DNA-RNA molecule; while the Integrase domain (Int) catalyzes the insertion of the retrotransposon cDNA into the host genome. Occasionally, an Envelope (Env)-like domain is present [8]. In angiosperms, the main *Gypsy* lineages are the closely-related *TAT* and *Athila* lineages, and the *Galadriel*, *Reina*, *CRM* (Centromeric Retrotransposon in Maize) and *Del* lineages [13]. The main *Copia* lineages are classified as *Tork*, *Retrofit*, *Oryco* and *SIRE*. The *Bianca* lineage was also recently described as part of the *Copia* super-family ([14,15]).

Defective elements, lacking several or all of these domains involved in mobility, can be classified as non-autonomous LTR retrotransposons (LTR-RT) elements. They are further sub-classified into TRIM (Terminal-repeat Retrotransposon In Miniature) [16], LARD (Large Retrotransposon Derivative) [17], BARE-2 (Barley RetroElement-2) [18] and TR-GAG (Terminal repeat with Gag domain) [19], according to their internal structures.

Since TEs correspond to the major part of plant genomes, their precise and exhaustive annotation, particularly in large genomes, remains a difficult and extensive work. More efforts in identification and annotation are necessary in partial draft genomes or in the case of new or highly degenerated repeated elements. In the last few years, several tools allowing one to identify and annotate transposable elements based on their structure and/or similarity were developed ([20,21]). REPET, one of these tools, has been developed to identify and classify transposable elements at a whole genome sequence scale. It was recently used to annotate TEs in several plant genomes, such as wheat [1], *Solanum pennellii* (a wild relative of tomato; [22]), *Coffea canephora* [12] and *Capsella rubella* [23].

However, available tools for the classification of TEs, and more particularly LTR-retrotransposons, such as TEclass [24], Repclass [25], Pastec [26], LTRsift [27] and LTRclassifier [28], provide limited information about the identification of super-families, while none of them are capable of classifying elements into lineages, nor identifying non-autonomous elements. Furthermore, optimized bioinformatics tools taking advantage of current supercomputers are now necessary to analyze and classify the large set of genomes and transposable elements available.

Computational approaches such as supercomputing, artificial intelligence [29] and data mining [30] are currently used for biological sciences, including sequences comparisons, nucleic acid secondary structure prediction and molecular dynamics [31], demonstrating the importance of speeding-up the analysis processes for large genomes [32]. Message Passing Interface (MPI) is a standard library for parallel programming [33], which is able to take advantage of multi-cores (like servers with many CPUs), many-cores (like GPUs) or heterogeneous (interaction between CPUs and GPUs [34]) architectures. MPI is capable of running in parallel many sub-problems that are previously divided given three focuses: (i) executing independent processes simultaneously; (ii) decomposing the main problem into tasks and resolving them in parallel; and (iii) introducing parallelism at instruction levels [35].

Pineapple (*Ananas comosus* L. 2n = 2x = 50) is a species indigenous to South America, belonging to the family Bromeliaceae (order Poales). Pineapple is the second most important tropical fruit crop after mango (FAO, http://www.fao.org/) and the most economically important species in the family Bromeliaceae. It is also the most economically important crop that assimilates carbon using Crassulacean Acid Metabolism (CAM) and is consequently a model to study the CAM photosynthesis pathway [36]. Over many years, genetic and genomic resources have been developed and reported for pineapple, including genetic maps with F1 and F2 populations [37] and expressed Sequence Tags (EST) and transcriptomes [38]. Only one previous study reported the presence of LTR-retrotransposon *Del* elements [39] in the pineapple genome. Now, the release of the pineapple genome sequence [40], with a draft that covers 72% of the estimated 526-Mb genome, offers the possibility of a large-scale analysis of the TE content.

In this study, we report the development of Inpactor, a parallel and scalable pipeline, able to classify LTR retrotransposons, to identify autonomous and non-autonomous elements, to perform RT-based phylogenetic trees and to analyze their insertion times using High Performance Computing (HPC) techniques. Inpactor was tested through a comprehensive analysis based on the identification and annotation of transposable elements in the pineapple genome. The pineapple genome assembly is comprised of 44% of transposable elements, of which 23% were classified as LTR retrotransposons and 9% as non-autonomous LTR retrotransposons. Only one lineage of the *Gypsy* superfamily: *Del*, corresponds to 16.4% of the pineapple genome assembly, suggesting that this lineage has undergone a significant increase in its copy numbers. Most full-length LTR retrotransposons were recently inserted (<2 Mya), reinforcing the hypothesis that they represent one of the most dynamic fractions of the pineapple genome. Inpactor provides comprehensive data of LTR retrotransposons' classification and dynamics at the lineage level, allowing a fine understanding of their contribution to genome structure and evolution.

2. Materials and Methods

2.1. Implementation of Inpactor

Inpactor is composed of four modules (Figure 1) and was developed using the Message Passing Interface (MPI) standard, in C language. It requires input parameters to be declared in a configuration file (Supplementary S1), where it is possible to define general information such as input file types (LTR_STRUC [41] output, REPET's TEdenovo output in FASTA format or a genome FASTA file), result directory, verbose mode and clean mode at the end of the execution. In addition, each module requires that different parameters be indicated in the configuration file.

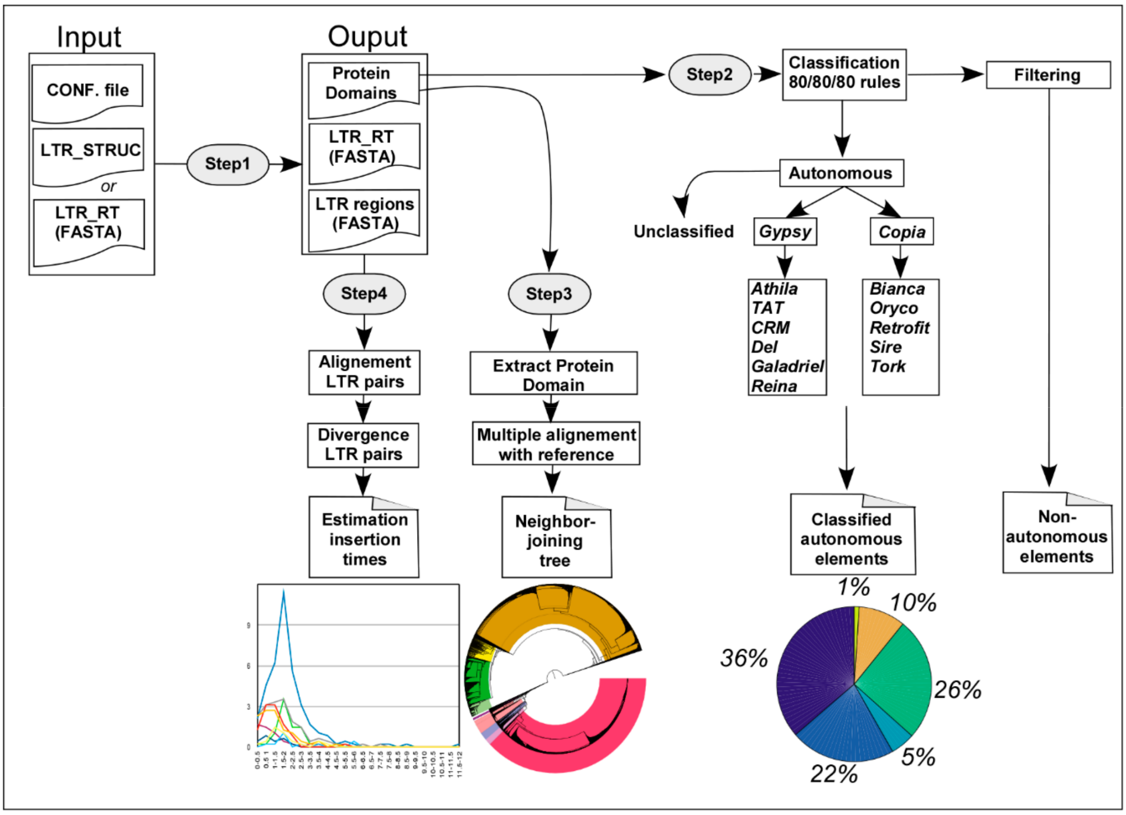

Figure 1. Representation of the different steps of the Inpactor pipeline.

Step 1, preprocessing module. The objective is to sort information and features from the LTR_STRUC output or the FASTA files submitted to Inpactor, such as the full element sequence, LTR identity, length and sequences using tools from EMBOSS [42] and BLASTX results against six references domains (GAG, RT, INT, RNAse H, AP and ENV) available at the Gypsy Database Project [15].

Step 2, classification module. It performs a classification using results from Step 1 as follows: (i) if the element carries at least one of the following domains: RT, INT, and RNAse H, with keywords RLC or RLG, the LTR-RT is classified as a putative autonomous-family element (*Copia* or *Gypsy*); (ii) if the element does not carry any domain, it is classified as a non-autonomous element (TRIM if the LTR-RT length is lower than 2000 bases and LARD if the LTR-RT length is greater than or equal to 2000 bases); (iii) if the element has only a GAG domain or a GAG and an AP domain, the element falls into the non-autonomous TR-GAG elements. Non-autonomous elements are not reclassified into super-families or lineages with autonomous elements. Elements with domains from both super-families (*Copia* and *Gypsy*) are considered as unclassified (possible chimeric elements). In addition, Inpactor creates an extra text file, which contains all LTR-RTs that are unclassified and thus named "no-class elements". Finally, the complete sequence of each classified and unclassified LTR retrotransposon (Figure 1) is extracted. A re-classification is performed following the 80-80-80 rule from Wicker et al. [8] for unclassified elements (Step 2, classification module parameters). Unclassified LTR-RT elements are re-analyzed with previously classified elements by similarity using Censor [43]. If the alignment covers a minimum of 80% of the unclassified element, with a minimum of 80% of nucleotide identity, and a minimum of 80 bases aligned (the 80/80/80 rule), the unclassified element is re-classified into the reference element [8].

Step 3, domain extraction module. Reverse Transcriptase (RT) domain sequences are extracted from each autonomous-family element, because this domain is the most conserved and appropriate for phylogenetic analysis (Figure 1). Other domains from the LTR-RT polyprotein might be used alternatively. BLASTX is executed using the FASTA file of all autonomous-family elements as the query and the reference RT domain database (the Gypsy Database Project [15]). Sequences that match with the database are extracted (extractseq, EMBOSS), and the domain is translated into amino acids using Genewise [44] with the option −pep. Only translated sequences larger than 200 amino acids are conserved for further analysis.

Step 4, LTR-RT insertion times' analysis and phylogenetic analysis. Using the FASTA protein file from the RT domain extraction module, a multiple alignment is performed using Mafft [45] with the option −thread to indicate the number of cores. Then, a phylogenetic tree is created based on the maximum likelihood method with Mafft using −retree and −treeout options with the multiple alignment obtained previously (Figure 1). The insertion times of full-length copies, as defined by a minimum of 80% of nucleotide identity over 100% of the reference element length, are dated [19]. Timing of insertion is based on the divergence of the 5′ and 3′-LTR sequences of each copy. The two LTRs are aligned using Stretcher (EMBOSS) and the divergence calculated using the Kimura 2-parameter method implemented in Distmat (EMBOSS) [46]. The insertion dates are estimated using an average base substitution rate of 1.3×10^{-8} as the default parameter [47]. This default parameter can be changed in the configuration file.

Inpactor produces different types of files: a sequence file (FASTA), a global alignment matrix, a phylogenetic tree and tabular files with the insertion time analysis. In addition, the Preprocessing module produces one tabular file, which contains all information from the LTR_STRUC output; the Classification module creates one tabular file and one FASTA file for each LTR-RT type found, including the unclassified. Finally, the domain extraction section generates only one FASTA file with all of the domains found.

Inpactor requires using external bioinformatics software to perform specific functions such as sequences extraction and translation: NCBI-Blast (v.2.5.0, https://www.ncbi.nlm.nih.gov/BLAST/), EMBOSS (v.6.6.0, http://emboss.sourceforge.net), Wise 2 (v.2.4.0, http://www.ebi.ac.uk/~birney/wise2/), OpenMPI (v.1.8.8, https://www.open-mpi.org/), Censor (v.4.2.29, http://www.girinst.org/downloads/software/censor/), Mafft (v.7.305, http://mafft.cbrc.jp/alignment/software/) and LTR_FINDER (v.1.0.5, https://code.google.com/archive/p/ltr-finder/).

2.2. Availability of Inpactor

Inpactor's source code can be found at https://github.com/simonorozcoarias/Inpactor, under the GNU GPLv3 license and is composed of one source code in C language, two bash scripts and an example of the configuration file. All of these need to be in the same folder. Installation instructions and a user manual are available at https://github.com/simonorozcoarias/Inpactor/blob/master/User%20manual%20Inpactor%20V%201.0%20final.pdf. Sample data and results are also available.

2.3. Computational Resources

All executions were done using a server with a 32-core Xeon E5-2670 (with HT enabled), 256 GB of RAM and the Centos 6.7 operating system, managed by Slurm [48]. All software used by Inpactor were installed in a non-standard directory and were loaded using Environmental Modules [49].

2.4. Sequence Data Sources

Inpactor was tested using five plant genomes with different genome sizes. *Arabidopsis thaliana* (117 Mb, http://plants.ensembl.org/Arabidopsis_thaliana/Info/Index) and maize (*Zea mays*, 2048 Mb; http://plants.ensembl.org/Zea_mays/Info/Index) were downloaded from the Ensembl genomes project [49]; rice (*Oryza sativa*, 362 Mb; http://ensembl.gramene.org/Oryza_sativa/Info/Index) was downloaded from the Gramene Project [50]; and Robusta coffee (*Coffea canephora*, 553 Mb; http://coffee-genome.org/) was downloaded from Coffee Genome Hub Project [51]. The pineapple genome sequence (variety "F153") was generated from a combination of Illumina, Moleculo, PacBio, and 9400 Bacterial Artificial Chromosomes (BACs) and released [40]. The genome sequences were deposited at the iPlant

CoGe database, and they can be downloaded at https://genomevolution.org/CoGe/NotebookView.pl?nid=937. The final assembly includes 382 Mb, corresponding to 72.6% of the estimated 526-Mb genome size.

2.5. Identification of Repeated Elements

REPET (TEdenovo package V.2.2-RC) [52] was used to find and classify repeated sequences in the pineapple genome sequences. In total, 3380 scaffolds accounting for 382,063,720 bp were processed. Consensus sequences obtained in REPET were annotated according to the REPBASE database (v.19.6, http://www.girinst.org/repbase/). They were named according to the acronym classification developed by Wicker and coworkers [8] (i.e., DHX (Helitron), DMX (Maverick), DTX (TIR Transposon), DXX (MITE) for Class II elements and RIX (LINE), RLX (LTR retrotransposon), RSX (SINE), RXX (unclassified or non-autonomous retrotransposons), RYX (DIRS) for Class I elements). Consensus sequences were classified as chimeric if they showed characteristics of more than one classification, representing potential nested elements. Additional tools were used to specifically predict full-length LTR retrotransposons (LTR_STRUC, [41]) based on their structure in order to complete the REPET detection.

2.6. Annotation, Phylogenetic Analysis and Insertion Time Analysis of LTR Retrotransposons

Consensus sequences from REPET that were identified, as “complete” (autonomous) or ”incomplete” (non-autonomous) LTR retrotransposons were further classified into lineages and families using Inpactor. At the genome level, putative RT domains were identified using BLASTX [53], with an e-value cut-off of 1×10^{-4}, and translated into amino acid sequences using Genewise [44]. The resulting RT sequences (with a minimum length of 150 residues) and reference RT domains from the Gypsy Database 2.0 were aligned, and a maximum likelihood tree was inferred and edited with Figtree (http://tree.bio.ed.ac.uk/software/figtree/). Insertion time analysis of LTR retrotransposons was performed as in Dupeyron et al., 2017, with the average substitution rate of 1.3×10^{-8} as implemented in Inpactor. LTR retrotransposons were used to annotate pineapple pseudo-molecules using RepeatMasker (-div 20 option; [54]; http://www.repeatmasker.org).

3. Results and Discussion

3.1. Testing Inpactor on Reference Plant Genomes

Genome annotation studies require the annotation and detailed classification of transposable elements, and more particularly LTR retrotransposons, representing the main part of plant genomes. Classification into main classes and lineages, insertion time analysis and phylogenetic analysis [55,56] constitute basic information for understanding the impact, dynamics and evolution of LTR retrotransposons. Inpactor has been developed to combine automatic annotation, classification, insertion time and phylogenetic analyses into a limited time process, taking advantage of supercomputers.

We first tested Inpactor using several numbers of cores (1, 4, 8, 16 and 32), with 10 repetitions for each experiment in order to calculate the speed-up and average run time per module (Table 1 and Supplementary S2).

Table 1. Results of Inpactor on 4 different plant genomes.

Species	Total Average Sequential Runtime in Seconds	Sequential Standard Deviation in Seconds	Total Average Parallel Runtime in Seconds	Parallel Standard Deviation in Seconds	Number of Cores	Speed-Up
Arabidopsis thaliana	995.3	14.42	361.28	5.82	4	2.8
			201.16	12.65	8	4.9
			134.5	9.24	16	7.4
			158.47	11.28	32	6.3
Oryza sativa	3228.7	94.07	1099.67	42.48	4	2.9
			677.25	41.65	8	4.8
			428.02	24.64	16	7.5
			412.75	18.61	32	7.8
Coffea canephora	9569.48	11.91	3292.15	155.64	4	2.9
			2029.39	108.92	8	4.7
			1143.97	23.64	16	8.4
			1015.44	31.71	32	9.4
Zea mays	65,031.07	1143.79	22,186.47	306.43	4	2.9
			11,657.74	582.24	8	5.6
			8452.74	394.94	16	7.7
			7907.58	495.85	32	8.2

Four different plant genomes (*Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays*) were used. Only the 32 cores' outputs were used for classifying predicted LTR-RTs into autonomous elements (*Gypsy* and *Copia* super-families and lineages) and for filtering putative non-autonomous element types (LARD, TRIM and TR_GAG; Figure 2 and Supplementary S3). Autonomous elements were sub-classified into lineages, and a phylogenetic tree per genome was constructed using the output files (Figure 2). Inpactor provided the insertion time analyses, indicating the insertion activity of LTR-RT elements over recent periods of time (Figure 2 and Supplementary S4).

Executions were performed using one server (Supplementary S5–S8), with the 80-80-80 rule option disabled. Each Inpactor module was executed independently in the correct order (i.e., preprocessing, classification, domain extraction, insertion time and phylogenetic tree creation) to calculate the runtime of each module. The total runtime is the sum of each runtime module. Finally, Inpactor was run on the pineapple genome sequence similarly to the four reference genomes used in order to study its LTR retrotransposons diversity (Figure 3 and Supplementary S9).

Inpactor can use different input files such as the LTR_STRUC output files and any FASTA files from other predictors of full-length elements LTR retrotransposons. LTR_STRUC is a relatively slow algorithm (running under a Windows-XP PC), compared to more recent prediction software [57], but it seems to offer a low percentage of false positives in plant genomes and an overall low number of putative elements. In the future, Inpactor will integrate more recent software used to predict LTR retrotransposons, such as LTRharvest [57], LTR-FINDER [58], and LTR_retriever [59]. Additionally, we will also include Hidden Markov Models (HMM) to perform a more sensitive annotation of protein domains. Inpactor uses a Shell script gluing together other programs to construct analysis. To speed up the overall analysis, Inpactor will be implemented as a single C binary.

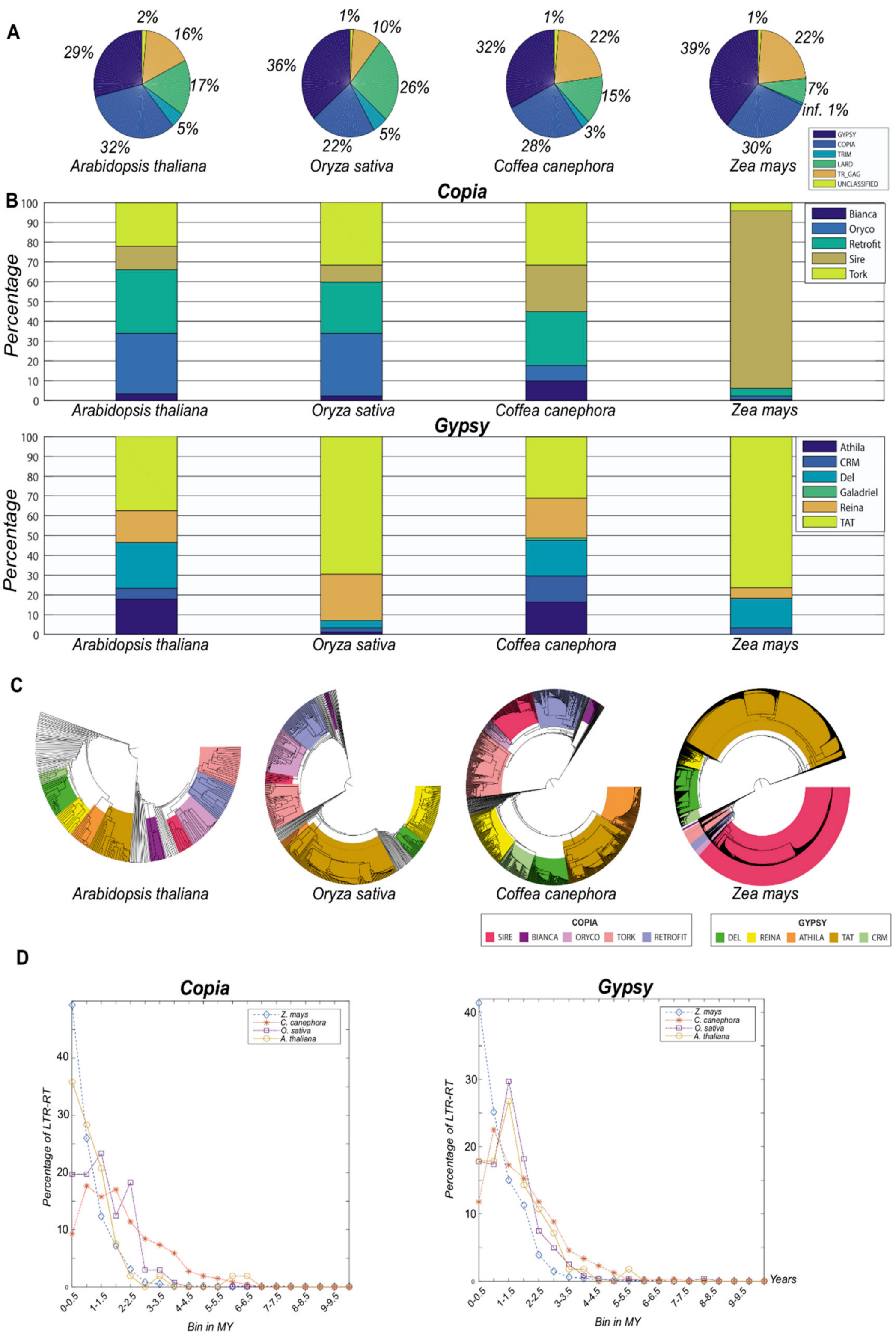

Figure 2. Inpactor results for the four species tested (*Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays*) based on LTR_STRUC detection. (**A**) Initial classification of LTR-RTs into autonomous (*Gypsy* and *Copia*) or non-autonomous (Terminal-repeat Retrotransposon In Miniature (TRIM), Large Retrotransposon Derivative (LARD) or Terminal repeat with Gag domain (TR-GAG)); (**B**) classification of the autonomous elements into lineages showing the variability that can be found in plant genomes; (**C**) phylogenetic trees using the RT domain; (**D**) insertion time analysis using autonomous elements (*Copia* and *Gypsy*).

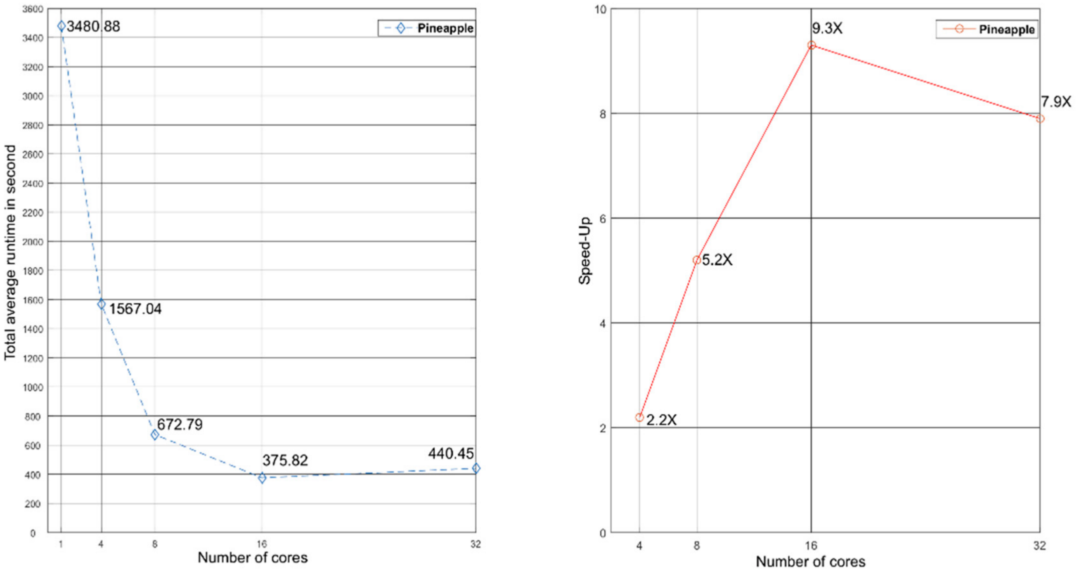

Figure 3. Total average runtime and speed-up of Inpactor using 4, 8, 16 and 32 cores with pineapple data.

As expected, Inpactor's results showed a different composition of LTR retrotransposon lineages in the reference genomes based on the detection of LTR_STRUC's full-length elements. Our results illustrate considerable variation in the classification of elements, despite that the quantitative detection of elements may be biased by the quality of the genome sequence and assembly. Insertion time and phylogenetic tree modules also provided evidence of different insertional activity of LTR retrotransposons during similar periods of time.

Inpactor surpasses other classification tools for LTR retrotransposons such as TEclass, Pastec and LTR classifier. None of them are able to give detailed information about the LTR retrotransposons' lineages, to identify non-autonomous elements and to estimate insertion times. As a consequence, it was not possible to compare the performance of Inpactor with these tools. Similarly, it was not possible to compare the classification of Inpactor with those from the published genomes of *A. thaliana*, rice, coffee and maize due to the lack of detailed information.

3.2. Using Inpactor on the Pineapple Genome

To use Inpactor on the pineapple genome, we first identified repeated sequences with the REPET TEdenovo package. After clustering and cleaning, 2860 consensus sequences were obtained from the genomic scaffolds and classified according to their structural features and similarities with the REPBASE protein database [60]. As a result, 75% of them were classified into Class I elements (retrotransposons) and 11% into Class II (DNA transposons), following the hierarchical classification proposed by Wicker and coworkers [8] (Supplementary S10). The remaining 14% of repeats were not identified at this step (Figure 4). Furthermore, 1402 LTR retrotransposon consensus sequences (RLX) were identified via TEdenovo, but 1148 sequences were classified as incomplete elements due to missing structural features detected by REPET [52]. Among the 1402 LTR retrotransposon consensus, 939 RLX consensus sequences were annotated and classified into lineages using Inpactor. Most of them (714, 76%) fell into the *Gypsy* superfamily, and more particularly into the *Del* lineage (590, 63%, Figure 5A). Most consensus elements that were classified into the *Del* lineage were closely related to the *Peabody* family based on their RT domains. We did not identify any reverse transcriptase domains from the *Athila* and *Bianca* lineages ([61,62]). Only 225 consensus sequences (23%) belonged to the *Copia* super-family. The remaining LTR retrotransposon consensus sequences that did not carry any recognizable RT domain were classified by Inpactor as TR-GAG (353) or other non-autonomous elements (RXX, 97); probably built from deletion derivative elements. In total, Inpactor did not classify 13 consensus sequences. Finally, Inpactor recovered RT domains for each consensus and released a maximum likelihood phylogenetic tree (Figure 5B), confirming the classification and the overrepresentation of consensus sequences from the *Del* lineage.

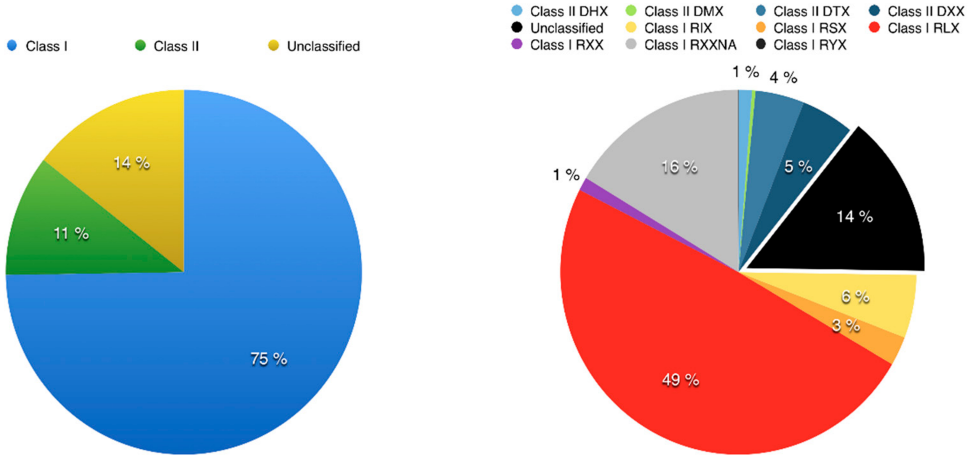

Figure 4. Transposable element abundance found in the pineapple genome. Classification of Transposable Elements (TEs) (left) and a detailed composition using the acronym classification developed by Wicker [8] (right) are presented: DHX (Helitron), DMX (Maverick), DTX (TIR Transposon), DXX (MITE), RIX (LINE), RLX (LTR retrotransposon), RSX (SINE), RXX (unclassified or non-autonomous retrotransposons), RYX (DIRS).

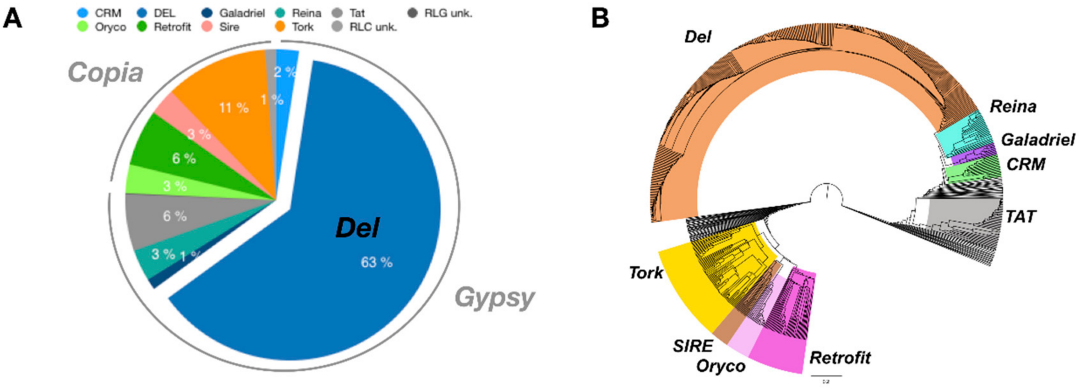

Figure 5. LTR retrotransposon lineages identified in consensus sequences identified by Inpactor. (**A**) Proportion of the different lineages in consensus sequences. The *Del* lineage represented 64% of all LTR retrotransposon annotated consensus sequences. (**B**) Phylogenetic analysis of annotated LTR retrotransposon consensus sequences.

RT domains were also directly recovered from the pineapple genome sequence, and 6379 aligned amino acid sequences were used to construct a phylogenetic tree to classify *Gypsy* and *Copia* super-families and lineages at the genome level (Figure 6). Similarly to the consensus analysis, the phylogenetic tree indicates the overrepresentation of RT domains from the *Del* lineage at the genome level. Most of the branches were closely linked to the *Peabody* family RT domain, confirming previous observations at the molecular level [40].

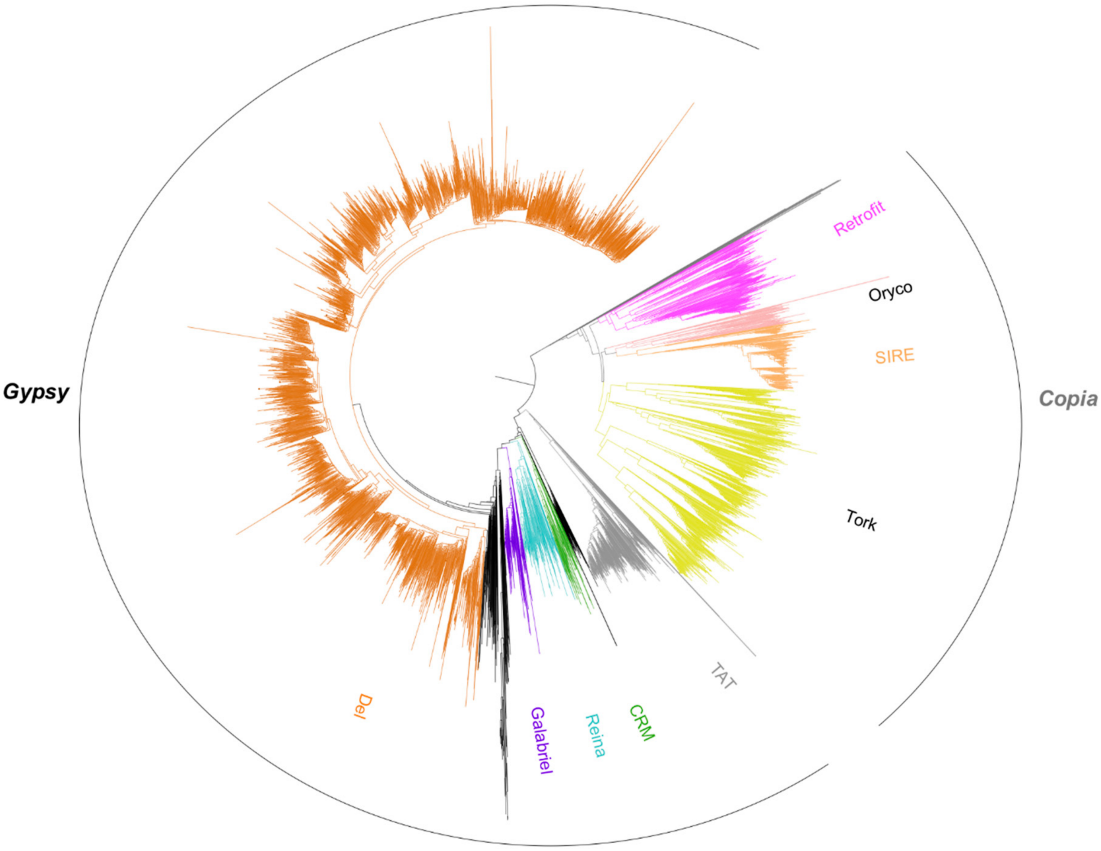

Figure 6. Phylogenetic analysis of 6379 Reverse Transcriptase (RT) domains from the pineapple genome assembly. RT domains were classified into *Gypsy* and *Copia* super-families and lineages using the reference RT domain from the Gypsy Database. The branches from the Gypsy *Del* family are represented in orange.

3.3. Pineapple LTR Retrotransposons Abundance and Dynamics

The final LTR-retrotransposon repertoire annotated by Inpactor, composed of 1389 sequences (for 5,263,860 bp of sequence), was used for pineapple pseudochromosome annotation using RepeatMasker. This repertoire masked 31.08 percent (118,709,778 bp) of the genome sequence. *Copia* and *Gypsy* elements represent 2.7% and 19.3%, respectively, of the genome, while non-autonomous elements represent 1.4% for RXX and 7.7% for TR_GAG. Indeed, the *Del* lineage represents a significant proportion of the genome with 16.4%. Along with pseudo-molecules, LTR retrotransposons range from 22.16 (LG4) to 33.18% (LG24), with the notable exception of the LG25 pseudo-molecule showing an overall percentage of 10.70% (Supplementary S11 and S12). *Del*, the most abundant lineage, ranges from 11.48 (LG17) to 18.65% (LG24), along with pseudo-molecules. Once again, the pseudo-molecule LG25 showed the lowest percentage of *Del* with 4.74%. The very low detection of LTR retrotransposons on LG25 remains intriguing and might be a result of reduced pericentromeric regions. Indeed, this reduction could also originate from difficulties in assembling reads from highly repetitive regions. Beside *Del*, non-autonomous elements (TR-GAG) represent the most significant group with a variation between 5.79% and 7.78%.

The pineapple genome was also processed by LTR_STRUC, and the output was used to estimate the time insertion of full-length LTR retrotransposons by Inpactor (Figure 7). Two different peaks were observed at 1.5–2 Million Years (MY) for *Gypsy* elements and at 1–1.5 MY for *Copia*, suggesting two different rounds of LTR retrotransposon amplification. The time insertion analyzed by lineages confirmed the amplification of *Del* lineages at 1.5–2 MY as the origin of its large copy numbers in the pineapple genome (Figure 7B).

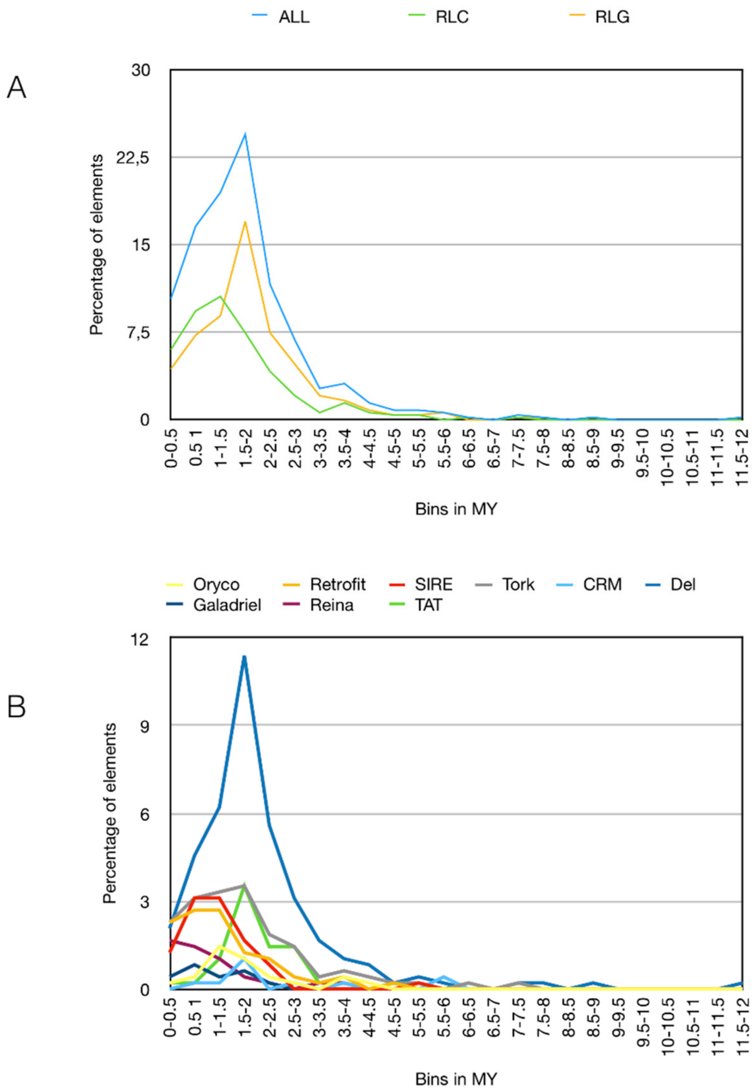

Figure 7. Timing of full-length LTR retrotransposon insertions. (**A**) Blue, yellow and green lines represent respectively the percentage of *Gypsy* and *Copia* full-length LTR retrotransposons per bins of 0.5 Million Years (MY); (**B**) colored lines represent the percentage of full-length LTR retrotransposon lineages per bins of 0.5 MY. Only the full-length LTR retrotransposons found by LTR_STRUC were used here. An average base substitution rate of 1.3 × 10^{-8} was used as the default [47].

4. Conclusions

In conclusion, Inpactor is a unique tool providing an exhaustive classification and analysis of LTR retrotransposons. It performs the classification of elements into super-families and lineages and efficiently filters non-autonomous elements. An additional benefit of Inpactor is the availability of an RT-based phylogenetic tree for supporting classification into lineages and a lineage-based time insertion analysis for analyzing elements' dynamics. Finally, the analysis of the pineapple genome with Inpactor provided fast and interesting information about the abundance and dynamics of LTR retrotransposons. It also suggests the good complementarity of REPET and Inpactor for the efficient and rapid classification and analysis of LTR retrotransposons.

Supplementary Materials: The following are available online at http://www.mdpi.com/2079-7737/7/2/32/s1, Supplementary S1: Example of the configuration file necessary to run Inpactor, Supplementary S2: Runtimes, standard deviations and break-down runtimes for *Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays* genomes, Supplementary S3: Classification of LTR-RTs found by Inpactor in *Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays*, Supplementary S4: Insertion time analysis done by Inpactor of autonomous elements (*Copia* and *Gypsy*) using *Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays*, Supplementary S5: Total average runtime of Inpactor using *Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays*, Supplementary S6: Comparison between break-down runtimes for each Inpactor module using one and 32 cores with the *Zea mays* genome, Supplementary S7: Speed-up of Inpactor using 4, 8, 16 and 32 cores and the *Arabidopsis thaliana*, *Oryza sativa*, *Coffea canephora* and *Zea mays* genomes, Supplementary S8: Average runtime of each Inpactor module using the *Zea mays* genome, Supplementary S9: Comparison between break-down runtimes of each Inpactor module using one and 32 cores with the Pineapple genome sequences, Supplementary S10: Number and classification of TE consensus found by TEdenovo in the Pineapple sequence, Supplementary S11 and 12: Percentage/proportion of the different LTR retrotransposon lineages and non-autonomous types for each pineapple pseudochromosome.

Author Contributions: S.O.-A. and R.G. developed the software. D.C. tested and improved the software. R.T.-S., D.C. and S.O.-A. carried out the Inpactor executions. A.G. carried out the REPET executions. S.O.-A., A.G., D.S.D., R.M. and R.T.-S. wrote the manuscript. R.G. directed the investigation. All authors read and approved the final manuscript.

Acknowledgments: We thank Alejandro Soto at BIOS for figure editing. We also thank the Centro de Bioinformática y Biología Computacional de Colombia BIOS for the use of the supercomputer.

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

References

1. Choulet, F.; Alberti, A.; Theil, S.; Glover, N.; Barbe, V.; Daron, J.; Pingault, L.; Sourdille, P.; Couloux, A.; Paux, E.; et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **2014**, *345*, 1249721. [CrossRef] [PubMed]
2. Ibarra-Laclette, E.; Lyons, E. Architecture and evolution of a minute plant genome. *Nature* **2013**, *498*, 1–6. [CrossRef] [PubMed]
3. Tenaillon, M.I.; Hollister, J.D.; Gaut, B.S. A triptych of the evolution of plant transposable elements. *Trends Plant Sci.* **2010**, *15*, 471–478. [CrossRef] [PubMed]
4. Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Saniyal, A.; Kim, H.; Collura, K.; Brar, D.S.; Jackson, S.; Wing, R.A.; et al. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **2006**, *16*, 1262–1269. [CrossRef] [PubMed]
5. Makarevitch, I.; Waters, A.J.; West, P.T.; Stitzer, M.; Hirsch, C.N.; Ross-Ibarra, J.; Springer, N.M. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genet.* **2015**, *11*, e1004915.
6. Todorovska, E. Retrotransposons and their Role in Plant—Genome Evolution Retrotransposons and Their Role in Plant—Genome. *Biotechnol. Biotechnol. Equip.* **2017**, *2818*, 294–305.
7. Casacuberta, E.; González, J. The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **2013**, *22*, 1503–1517. [CrossRef] [PubMed]
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] [PubMed]
9. 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]
10. Paux, E.; Roger, D.; Badaeva, E.; Gay, G.; Bernard, M.; Sourdille, P.; Feuillet, C. Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant J.* **2006**, *48*, 463–474. [CrossRef] [PubMed]
11. Paterson, A.H.; Bowers, J.E.; Bruggmann, R.; Dubchak, I.; Grimwood, J.; Gundlach, H.; Haberer, G.; Hellsten, U.; Mitros, T.; Poliakov, A.; et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature* **2009**, *457*, 551–556. [CrossRef] [PubMed]

12. Denoeud, F.; Carretero-Paulet, L.; Dereeper, A.; Droc, G.; Guyot, R.; Pietrella, M.; Zheng, C.; Alberti, A.; Anthony, F.; Aprea, G.; et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **2014**, *345*, 1181–1184. [CrossRef] [PubMed]
13. Llorens, C.; Muñoz-Pomer, A.; Bernad, L.; Botella, H.; Moya, A. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. *Biol. Direct* **2009**, *4*, 41. [CrossRef] [PubMed]
14. Wicker, T.; Keller, B. Genome-wide comparative analysis of copia retrotransposons in Triticeae, rice, and Arabidopsis reveals conserved ancient evolutionary lineages and distinct dynamics of individual copia families. *Genome Res.* **2007**, *17*, 1072–1081. [CrossRef] [PubMed]
15. 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. [CrossRef] [PubMed]
16. 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. [CrossRef] [PubMed]
17. Kalendar, R.; Vicient, C.M.; Peleg, O.; Anamthawat-Jonsson, K.; Bolshoy, A.; Schulman, A.H. Large retrotransposon derivatives: Abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* **2004**, *166*, 1437–1450. [CrossRef] [PubMed]
18. Tanskanen, J.A.; Sabot, F.; Vicient, C.; Schulman, A.H. Life without GAG: The BARE-2 retrotransposon as a parasite's parasite. *Gene* **2007**, *390*, 166–174. [CrossRef] [PubMed]
19. Chaparro, C.; Gayraud, T.; de Souza, R.F.; Domingues, D.S.; Akaffou, S.; Vanzela, A.L.L.; de Kochko, A.; Rigoreau, M.; Crouzillat, D.; Hamon, S.; et al. Terminal-repeat retrotransposons with gAG domain in plant genomes: A new testimony on the complex world of transposable elements. *Genome Biol. Evol.* **2015**, *7*, 493–504. [CrossRef] [PubMed]
20. Bergman, C.M.; Quesneville, H. Discovering and detecting transposable elements in genome sequences. *Brief. Bioinform.* **2007**, *8*, 382–392. [CrossRef] [PubMed]
21. 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. [CrossRef] [PubMed]
22. Bolger, A.; Scossa, F.; Bolger, M.E.; Lanz, C.; Maumus, F.; Tohge, T.; Quesneville, H.; Alseekh, S.; Sørensen, I.; Lichtenstein, G.; et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat. Genet.* **2014**, *46*, 1034–1038. [CrossRef] [PubMed]
23. Slotte, T.; Hazzouri, K.M.; Ågren, J.A.; Koenig, D.; Maumus, F.; Guo, Y.-L.; Steige, K.; Platts, A.E.; Escobar, J.S.; Newman, L.K.; et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.* **2013**, *45*, 831–835. [CrossRef] [PubMed]
24. Abrusán, G.; Grundmann, N.; Demester, L.; Makalowski, W. Teclass—A tool for automated classification of unknown eukaryotic transposable elements. *Bioinformatics* **2009**, *25*, 1329–1330. [CrossRef] [PubMed]
25. Feschotte, C.; Keswani, U.; Ranganathan, N.; Guibotsy, M.L.; Levine, D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. *Genome Biol. Evol.* **2009**, *1*, 205–220. [CrossRef] [PubMed]
26. Hoede, C.; Arnoux, S.; Moisset, M.; Chaumier, T.; Inizan, O.; Jamilloux, V.; Quesneville, H. PASTEC: An automatic transposable element classification tool. *PLoS ONE* **2014**, *9*, e91929. [CrossRef] [PubMed]
27. Steinbiss, S.; Kastens, S.; Kurtz, S. LTRsift: A graphical user interface for semi-automatic classification and postprocessing of de novo detected LTR retrotransposons. *Mob. DNA* **2012**, *3*, 18. [CrossRef] [PubMed]
28. Monat, C.; Tando, N.; Tranchant-Dubreuil, C.; Sabot, F. LTRclassifier: A website for fast structural LTR retrotransposons classification in plants. *Mob. Genet. Elem.* **2016**, *6*, e1241050. [CrossRef] [PubMed]
29. Orozco, S.; Jeferson, A. Aplicación de la inteligencia artificial en la bioinformática, avances, definiciones y herramientas. *UGCiencia* **2016**, *22*, 159–171. [CrossRef]
30. Arango-López, J.; Orozco-Arias, S.; Salazar, J.A.; Guyot, R. Application of Data Mining Algorithms to Classify Biological Data: The *Coffea canephora* Genome Case. *Adv. Comput.* **2017**, *735*, 156–170.
31. Maizel, J.R. Supercomputing in molecular biology: Applications to sequence analysis. *IEEE Eng. Med. Biol. Mag. Q. Mag. Eng. Med. Biol. Soc.* **1988**, *7*, 27–30. [CrossRef] [PubMed]
32. Orozco-Arias, S.; Tabares-Soto, R.; Ceballos, D.; Guyot, R. Parallel Programming in Biological Sciences, Taking Advantage of Supercomputing in Genomics. *Adv. Comput.* **2017**, *735*, 627–643.

33. Gropp, W.; Lusk, E.; Doss, N.; Skjellum, A. A high-performance, portable implementation of the MPI message passing interface standard. *Parallel Comput.* **1996**, *22*, 789–828. [CrossRef]
34. Tabares Soto, R. Programación Paralela Sobre Arquitecturas Heterogéneas. Master's Thesis, Universidad Nacional de Colombia, Manizales, Colombia, 2016.
35. Castro, J.L.A.; Leiss, E. *Introducción a la Computación Paralela*; Editorial Venezolana, Universidad de Los Andes: Mérida, Venezuela, 2004; ISBN 980-12-0752-3.
36. Zhang, J.; Liu, J.; Ming, R. Genomic analyses of the CAM plant pineapple. *J. Exp. Bot.* **2014**, *65*, 3395–3404. [CrossRef] [PubMed]
37. Carlier, J.D.; Sousa, N.H.; Santo, T.E.; d'Eeckenbrugge, G.C.; Leitão, J.M. A genetic map of pineapple (*Ananas comosus* (L.) Merr.) including SCAR, CAPS, SSR and EST-SSR markers. *Mol. Breed.* **2012**, *29*, 245–260. [CrossRef]
38. Ong, W.D.; Voo, C.L.Y.; Kumar, S.V. Development of ESTs and data mining of pineapple EST-SSRs. *Mol. Biol. Rep.* **2012**, *39*, 5889–5896. [CrossRef] [PubMed]
39. Thomson, K.G.; Thomas, J.E.; Dietzgen, R.G. Retrotransposon-like sequences integrated into the genome of pineapple, *Ananas comosus*. *Plant Mol. Biol.* **1998**, *38*, 461–465. [CrossRef] [PubMed]
40. Ming, R.; VanBuren, R.; Wai, C.M.; Tang, H.; Schatz, M.C.; Bowers, J.E.; Lyons, E.; Wang, M.-L.; Chen, J.; Biggers, E.; et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **2015**, *47*, 1435–1442. [CrossRef] [PubMed]
41. McCarthy, E.M.; McDonald, J.F. LTR STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, *19*, 362–367. [CrossRef] [PubMed]
42. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
43. Jurka, J.; Klonowski, P.; Dagman, V.; Pelton, P. CENSOR—A program for identification and elimination of repetitive elements from DNA sequences. *Comput. Chem.* **1996**, *20*, 119–121. [CrossRef]
44. Birney, E.; Durbin, R. Using GeneWise in the. *Genome Res.* **2000**, *10*, 547–548. [CrossRef] [PubMed]
45. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
46. SanMiguel, P.; Gaut, B.S.; Tikhonov, A.; Nakajima, Y.; Bennetzen, J.L. The paleontology of intergene retrotransposons of maize. *Nat. Genet.* **1998**, *20*, 43. [CrossRef] [PubMed]
47. Ma, J.; Bennetzen, J.L. Rapid recent growth and divergence of rice nuclear genomes. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 12404–12410. [CrossRef] [PubMed]
48. Jette, M.; Grondona, M. SLURM: Simple Linux Utility for Resource Management. In *Workshop on Job Scheduling Strategies for Parallel Processing*; Springer: Berlin/Heidelberg, Germany, 2003; pp. 44–60.
49. Furlani, J.L.; Osel, P.W. Abstract Yourself with Modules. In Proceedings of the 10th USENIX Conference on System Administrationm, Chicago, IL, USA, 29 September–4 October 1996; pp. 193–204.
50. Tello-ruiz, M.K.; Stein, J.; Wei, S.; Preece, J.; Olson, A.; Naithani, S.; Amarasinghe, V.; Dharmawardhana, P.; Jiao, Y.; Mulvaney, J.; et al. Gramene 2016: Comparative plant genomics and pathway resources. *Nucleic Acids Res.* **2017**, *44*, 1133–1140. [CrossRef] [PubMed]
51. Dereeper, A.; Bocs, S.; Rouard, M.; Guignon, V.; Ravel, S.; Tranchant-Dubreuil, C.; Poncet, V.; Garsmeur, O.; Lashermes, P.; Droc, G. The coffee genome hub: A resource for coffee genomes. *Nucleic Acids Res.* **2015**, *43*, D1028–D1035. [CrossRef] [PubMed]
52. Duprat, E.; Feuillet, C.; Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *Genome Res.* **2011**, *6*, e16526.
53. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D.J. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402. [CrossRef] [PubMed]
54. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker Open-4.0. 2013–2015. Available online: http://www.repeatmasker.org (accessed on 23 May 2018).
55. Du, J.; Tian, Z.; Hans, C.S.; Laten, H.M.; Cannon, S.B.; Jackson, S.A.; Shoemaker, R.C.; Ma, J. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: Insights from genome-wide analysis and multi-specific comparison. *Plant J.* **2010**, *63*, 584–598. [CrossRef] [PubMed]

56. Dupeyron, M.; de Souza, R.F.; Hamon, P.; de Kochko, A.; Crouzillat, D.; Couturon, E.; Domingues, D.S.; Guyot, R. Distribution of Divo in *Coffea* genomes, a poorly described family of angiosperm LTR-Retrotransposons. *Mol. Genet. Genom.* **2017**, *292*, 741–754. [CrossRef] [PubMed]
57. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **2008**, *14*, 18. [CrossRef] [PubMed]
58. Xu, Z.; Wang, H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **2007**, *35*, 265–268. [CrossRef] [PubMed]
59. 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. [CrossRef] [PubMed]
60. Kohany, O.; Gentles, A.J.; Hankus, L.; Jurka, J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinform.* **2006**, *7*, 474. [CrossRef] [PubMed]
61. Marco, A.; Marín, I. How Athila retrotransposons survive in the Arabidopsis genome. *BMC Genom.* **2008**, *9*, 219. [CrossRef] [PubMed]
62. Pélissier, T.; Tutois, S.; Deragon, J.M.; Tourmente, S.; Genestier, S.; Picard, G. Athila, a new retroelement from *Arabidopsis thaliana*. *Plant Mol. Biol.* **1995**, *29*, 441–452. [CrossRef] [PubMed]

© 2018 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 (http://creativecommons.org/licenses/by/4.0/).