genes

MDPI

Article

InpactorDB: A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning

Simon Orozco-Arias [1,2,*], Paula A. Jaimes [1], Mariana S. Candamil [1], Cristian Felipe Jiménez-Varón [3], Reinel Tabares-Soto [4], Gustavo Isaza [2] and Romain Guyot [4,5,*]

1 Department of Computer Science, Universidad Autónoma de Manizales, 170002 Manizales, Colombia; paula.jaimesb@autonoma.edu.co (P.A.J.); mariana.candamilc@autonoma.edu.co (M.S.C.)
2 Department of Systems and Informatics, Universidad de Caldas, 170002 Manizales, Colombia; gustavo.isaza@ucaldas.edu.co
3 Department of Physics and Mathematics, Universidad Autónoma de Manizales, 170002 Manizales, Colombia; cristian.jimenezv@autonoma.edu.co
4 Department of Electronics and Automation, Universidad Autónoma de Manizales, 170002 Manizales, Colombia; rtabares@autonoma.edu.co
5 Institut de Recherche pour le Développement, CIRAD, University of Montpellier, 34394 Montpellier, France
* Correspondence: simon.orozco.arias@gmail.com (S.O.-A.); romain.guyot@ird.fr (R.G.)

Abstract: Long terminal repeat (LTR) retrotransposons are mobile elements that constitute the major fraction of most plant genomes. The identification and annotation of these elements via bioinformatics approaches represent a major challenge in the era of massive plant genome sequencing. In addition to their involvement in genome size variation, LTR retrotransposons are also associated with the function and structure of different chromosomal regions and can alter the function of coding regions, among others. Several sequence databases of plant LTR retrotransposons are available for public access, such as PGSB and RepetDB, or restricted access such as Repbase. Although these databases are useful to identify LTR-RTs in new genomes by similarity, the elements of these databases are not fully classified to the lineage (also called family) level. Here, we present InpactorDB, a semi-curated dataset composed of 130,439 elements from 195 plant genomes (belonging to 108 plant species) classified to the lineage level. This dataset has been used to train two deep neural networks (i.e., one fully connected and one convolutional) for the rapid classification of these elements. In lineage-level classification approaches, we obtain up to 98% performance, indicated by the F1-score, precision and recall scores.

Keywords: LTR retrotransposons; machine learning; deep neural networks; bioinformatics; plant genomes; genomics; InpactorDB

Citation: Orozco-Arias, S.; Jaimes, P.A.; Candamil, M.S.; Jiménez-Varón, C.F.; Tabares-Soto, R.; Isaza, G.; Guyot, R. InpactorDB: A Classified Lineage-Level Plant LTR Retrotransposon Reference Library for Free-Alignment Methods Based on Machine Learning. *Genes* **2021**, *12*, 190. https://doi.org/10.3390/genes12020190

Academic Editor: Dariusz Grzebelus
Received: 30 December 2020
Accepted: 22 January 2021
Published: 28 January 2021

1. Introduction

Transposable elements (TEs) have key roles in plant genomes. They are major contributors to genomic size [1,2], rearrangement events (such as fissions, fusions, and translocations) [3], chromosome organization and structure (e.g., centromeres) [4], and evolution and adaptation to the environment [5]. These dynamic elements can be activated under several biotic or abiotic stresses, such as pathogens [6,7], defense-associated stresses [8], heat, drought and salt stresses, freezing, polyploidization and hybridization events [9,10], UV light [11], and X-ray irradiation [12]. Transposable elements are also known to participate in reproductive isolation between genotype of the same species (reviewed in [13]) [14] and to shape the genome architecture during the process of plant speciation [15].

TE classification is still a subject of debate, despite the fact that a standard has emerged. TE classification is generally performed hierarchically [16], whereby TEs are first divided into classes according to their replication cycle: Class I or retrotransposons, which follow a

copy-and-paste strategy using an RNA intermediate; and Class II or DNA transposons that use a cut-and-paste mobility mechanism through a DNA molecule [17]. Next, TE levels correspond to orders, superfamilies, lineages (also called families), and sub-families [18]. Among these, long terminal repeat (LTR) retrotransposons (LTR-RTs, an order of retrotransposons) are the most abundant TEs in plant genomes [19,20] and can account for up to 80% of the plant genome size, such as in wheat, barley, or rubber tree [21]. LTR-RTs are characterized by the presence of one or several open reading frames involved in the mobility of the element, flanked by a direct tandem repeat of 100 pb to more than 5000 bp, called LTR. These LTRs are directly involved in the transcription regulation of the element by the host's machinery [22,23].

LTR-RT in plants are classically divided into two major superfamilies: Copia (also called Ty1) and Gypsy (also called Ty3), based on the organization of coding domains in the element [24,25]. Each superfamily is sub-classified into lineages or families according to coding region similarities and phylogenetic relationships of the reverse transcriptase (RT) domains, a combination of several domains or the complete polyprotein of the elements [24,26,27]. Llorens and coworkers [28,29] classified LTR-retrotransposons based on a phylogenetic analysis of 268 non-redundant element and in plants, 5 Copia and 2 Gypsy lineages have been identified, and further sub-classified into clades. With a bigger sampling composed of 5410 Copia and 8453 Gypsy elements from 80 plant genomes and a phylogenetic approach, Neumann and coworkers identified 16 Copia (that is, Ale, Alesia, Angela, Bianca, Bryco, Lyco, Gymco-I,II,III, IV, Ikeros, Ivana, Osser, SIRE, TAR and Tork), and 14 Gypsy lineages, sub-divided into chromovirus and non chromovirus elements (that is, CRM, Chlamyvir, Galadriel, Tcn1, Reina, Tekay, Athila, Tat-I,II,III, Ogre, Retand, Phygy and Selgy) [30]. Coding domains of these classified elements are available as curated libraries (Gypsydb and RexDB) for fine annotation of elements using homology based software such as RepeatMasker [31].

The classification of LTR-retrotransposons as deep as the classification in lineages finds its justification in numerous studies showing the dynamics of amplification of these elements. For example, in some plant genomes, sudden expansion of genome size is the result of the amplification of one or a small number of lineages [32–34]. Different copy number, amplification history and chromosomal distribution of lineages shape the genome architecture of plants [35–37]. A better fine-scale annotation of LTR retrotransposon in plants will likely reveal new lineage-specific mechanisms of genome size variation and divergence. Currently, a challenge in genomics is to reliably annotate TEs. These elements have certain characteristics that make their identification and classification a complex task [38,39], such as repetitiveness, structural and nucleotide diversity, complex mobilization dynamics (including nested insertions), and species specificity [18,40,41]. Although de novo, homology-based, structure-based, and comparative genomics bioinformatics methods (or a combination of several methods) can automatically detect and classify TEs [42] (for a review see [5,26]), all of these approaches have limitations due to the diversity of TE structures, the quality of genome assemblies into others, and the sole use of any of these cannot produce high quality results. Thus, the TE annotation process usually relies on much manual work done by experts [43]. With the recent advances of sequencing technologies, many plant genomes have been sequenced and the automation of TE annotation is needed to process the large amount of DNA sequence data [44]. Recent studies have demonstrated that machine learning (ML) can be applied to automatically annotate or even to both identify and annotate TEs in short times [18,45–48] using publicly available databases such as Repbase [49], PGSB [50], RepetDB [44], among others (for a list see [5]). Despite the available datasets, none of these attains a lineage-level classification and several do not include plant species from certain families, which could affect the generalization performance of the ML-based algorithms.

In this work, we present InpactorDB, a semi-curated dataset comprising more than 130,000 LTR retrotransposons from 195 plant species belonging to 108 families. These elements are classified to the lineage level and are filtered by length, number of coding

domains present, nested insertion of class II TEs, and other retrotransposons. In addition, we removed the redundancy of elements through consensus creation following the same methodology of REPET, obtaining more than 67,000 sequences. This dataset constitutes a valuable resource for homology-based TE annotation, which is the most used approach [39], such as in RepeatMasker. Furthermore, this database also contributes to developing and testing ML-based algorithms for alignment-free and automatic annotation methods. Finally, we tested InpactorDB using two currently available deep neural networks for the classification of LTR retrotransposons to the lineage level.

2. Materials and Methods

2.1. Databases and LTR-RT Classification Processes

We collected information about LTR retrotransposons from three known TE databases: Repbase (v. 20.05, 2017) [49], PGSB [51], and RepetDB [44]. In addition, we detected LTR retrotransposons using different tools that follow a structure-based identification strategy. First, we used LTR_STRUC [52], due to its low level of false positive rates in plants (Romain Guyot, personal communication), on 69 available plant genomes to produce a dataset named here as "LTR_STRUC"; however, LTR_STRUC can be run only under Windows XP and takes a considerable execution time. Therefore, we used EDTA v1.9.3 [53], which uses LTR_Finder [54], LTRharvest [55], and LTR_retriever [56], to detect LTRs in 87 additional species (Supplementary Material S1) and generated a new dataset called "EDTA". For Repbase, we joined the LTR domains and the internal section (concatenating before and after) of each LTR retrotransposon into a single sequence. Plant genomes to be analyzed were selected to target 103 different Angiosperm species families and in priority assemblies with low genome size (Supplementary Material S1).

We applied the methodology proposed by Inpactor [57] to classify the elements of all all the datasets. Inpactor uses a homology-based strategy with known coding domains belonging to LTR-RTs; specifically, we used the RexDB [27] domain library as the reference. LTR-RTs were classified into superfamilies [for example, Gypsy (RLG) or Copia (RLC)] and sub-classified into lineages according to the similarities of five amino acid reference domains (GAG, AP, RT, RNAseH, and INT domains [58]). In addition, we applied filters to keep only intact elements by removing (1) predicted elements with domains from two superfamilies (that is, Gypsy and Copia, potential chimeric elements), (2) elements with domains belonging to two or more lineages, (3) elements with lengths different than those reported by the Gypsy Database (GyDB) [29], with a tolerance of 20%, (4) incomplete elements with less than three identified domains, and (5) elements with insertions of class II TEs (reported in Repbase). Figure 1 shows a general representation of the classification and filtering process.

The data generated is available at Zenodo under doi:10.5281/zenodo.4453481 and at DataSuds (https://dataverse.ird.fr) under doi:10.23708/QCMOUA.

2.2. Statistical Analysis

The datasets used in this study have different origins and characteristics such as the type of sequences (that is, consensus or individual DNA sequences) and pre-processing (that is, curated or non-curated sequences). We used ML algorithms such as logistic regression (LR), linear discriminant analysis (LDA), K-nearest neighbors (KNN), multi-layer perceptron with one layer (MLP), random forest (RF), decision trees (DT), naïve Bayes network (NB), and support vector machine (SVM) to test the performance of the datasets. We used the F1-score as the performance metric, which is the harmonic mean of precision and sensitivity [39] and we used it as the accuracy indicator; we used k-mer frequencies with $1 \leq k \leq 6$ as features, and we used scaling and dimensional reduction using principal component analysis (PCA) as pre-processing steps, according to [39].

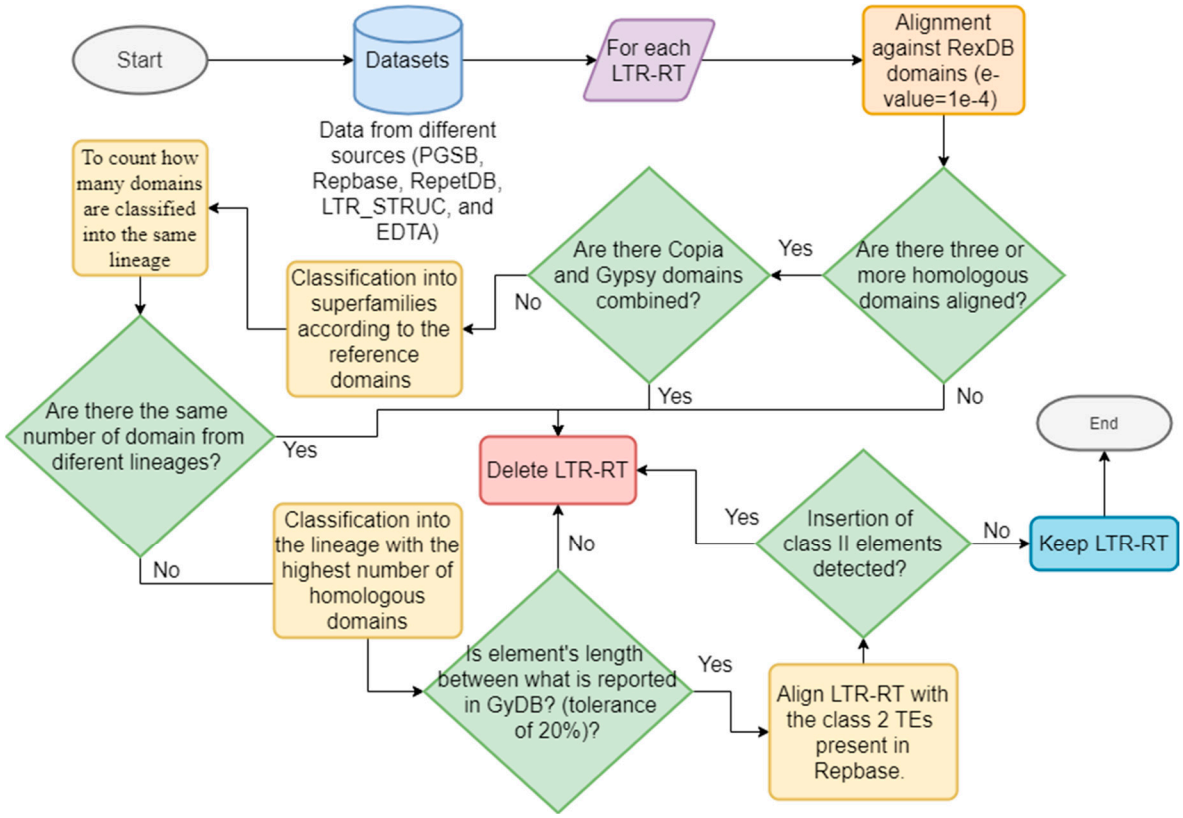

Figure 1. Schematic representation of the classification and filtering process performed for InpactorDB.

We created subsets of the datasets according to their characteristics (Table 1). To avoid bias related to the number of elements in each subset, we randomly selected the same number of LTR retrotransposons of each lineage that was present in the smallest dataset (Repbase; ~2842 elements).

Table 1. Datasets used in the statistical analysis of machine learning (ML) performances.

Name	Observations
Repbase	Curated consensus sequences.
PGSB	Curated individual genomic sequences.
RepetDB	Non-curated consensus sequences.
LTR_STRUC	Non-curated individual genomic sequences.
Consensus	Union between Repbase and RepetDB.
Genomics	Union between PGSB and LTR_STRUC.
Curated	Union between Repbase and PGSB.
Non-curated	Union between RepetDB and LTR_STRUC.
All	Union between Repbase, PGSB, RepetDB, and LTR_STRUC.

A one-way analysis of variance (ANOVA), using subset type as the variable factor, was conducted to determine statistically significant differences between the performances of the algorithms applied to the subsets (Table 1). Additionally, we performed the Shapiro–Wilks normality test on the standardized residuals and a homoscedasticity test through the Bartlett test to determine the need for a non-parametric framework, such as the Kruskal–Wallis test.

Given significant statistical differences, a post hoc test was performed to identify which datasets generated these differences. This test was based on pairwise comparisons

using Bonferroni's method in the non-parametric framework of Duncan's method. The pairwise comparisons were conducted as follows:

$$H0: \mu_i = \mu_j$$
$$H1: \mu_i \neq \mu_j \quad (1)$$

Following the pairwise analysis, we selected a subset of the data that did not display statistically significant differences in the performance of ML algorithms with the other subsets, but taking into account that the average performances are the best among the other subsets.

2.3. Post-Processing and Generalization Tests through Deep Neural Networks

Based on the results of significant differences, we removed the redundancy of LTR retrotransposon sequences in the individual genomic datasets (PGSB, LTR_STRUC, and EDTA). For this, we used the same methodology implemented in REPET. First, we performed a BLASTN v2.4.0 (NCBI-Blast) [59] of all elements against all (separating each dataset) using an evalue = 1×10^{-300} and an identity cutoff $\geq$90. Then, we clustered the sequences using Silix v1.2.11 [60] with a minimum length of 95% and a minimum identity of 90%. Next, we generated a multiple alignment of each group using MAFFT v7.305b [61] and removed the columns in which all but one sequence showed gaps, using trimal v1.2 [62]. Finally, we built consensus sequences based on the majority system using cons (EMBOSS v6.5.7 [63]). This dataset is referred to as the non-redundant version of InpactorDB.

We used the non-redundant version of InpactorDB to explore the automatic alignment-free classification process of LTR retrotransposons to the lineage level. Then, we implemented two deep neural networks (DNN) based on previously published research. First, we tested the hyper-parameter values proposed by Nakano et al. [46] for a fully connected DNN to classify TEs into superfamilies following a hierarchical strategy. The network had three hidden layers with 200 neurons each. The training stage was performed using 200 epochs (times that the entire training set is used to train the network) and mini batches of size 128, stochastic gradient descent (SGD) with a learning rate of 0.01, and Adam as the optimizer with a learning rate $\alpha = 0.001$, $\beta1 = 0.9$, $\beta2 = 0.999$, and $\varepsilon = 10^{-8}$, where $\beta1$ and $\beta2$ are the first and second exponential decay rate of the moment vector, respectively. The loss function used was the mean squared error and the activation function was ReLu (Figure 2).

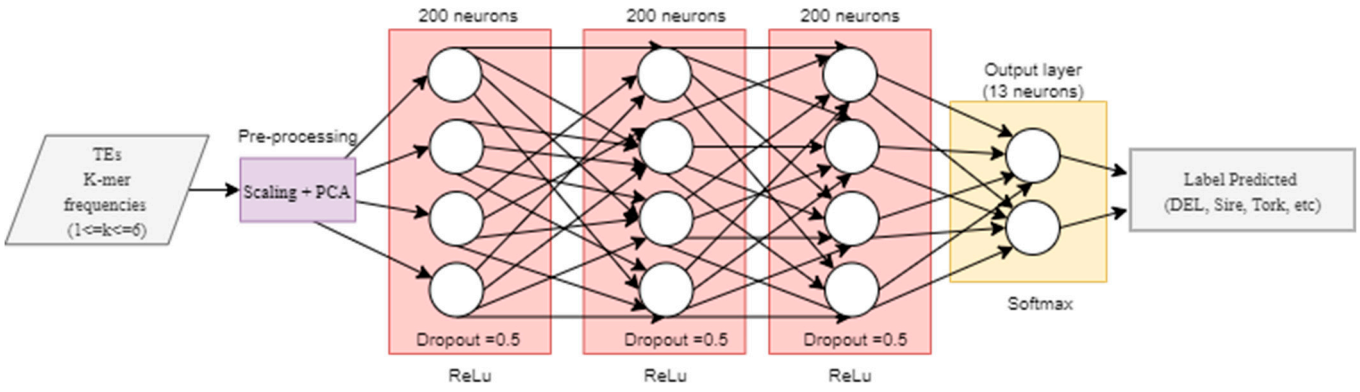

Figure 2. Implementation of the fully connected neural network architecture proposed by Nakano et al. [46].

We also implemented a convolutional neural network (CNN) using the hyper-parameters proposed in DeepTE [48], which was previously used to classify TEs from all classes (that is, retrotransposons and DNA transposons) into superfamilies. This CNN consisted of three layers with 100, 150, and 225 filters with a kernel size of 3. Max pooling was used after

each convolutional layer with a window size of 2. A dropout of 0.5 was used after the last convolutional layer. Finally, a fully connected layer with 128 units was used, and a softmax output layer was set to calculate the probabilities of the predicted classes. ReLu was used as the activation function in the three convolutional layers and the fully connected layer. Furthermore, we used a categorical_crossentropy loss function and an ADAM optimizer with a learning rate of 0.001 (Figure 3). The implementation of these DNN architectures in Python and Tensorflow 2 (Keras) can be consulted in Supplementary Material S2.

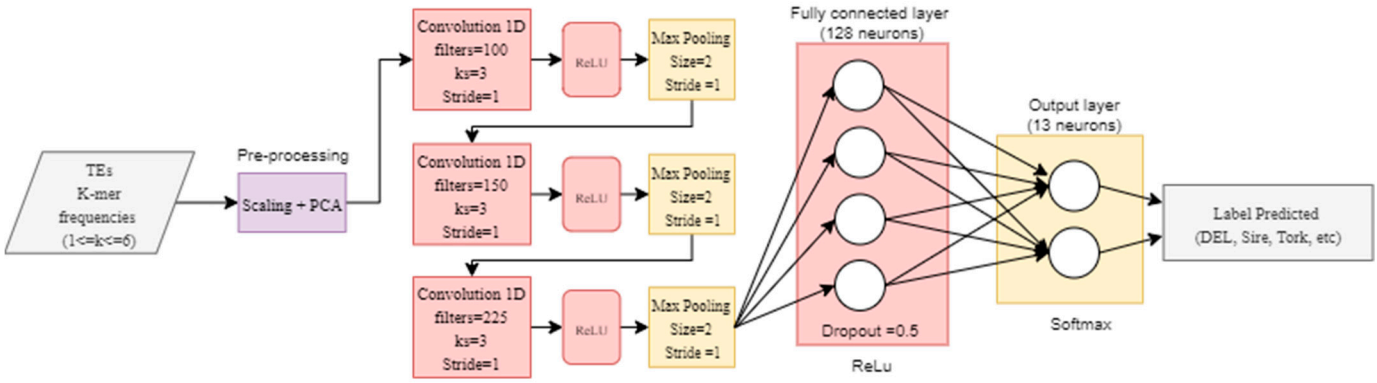

Figure 3. Implementation of the convolutional neural network (CNN) architecture proposed in DeepTE [48].

As features, we used k-mer frequencies with $1 \leq k \leq 6$. We also applied scaling and reduction of dimensionality using principal components analysis (PCA) as suggested by [39]. Each dataset was partitioned into 80% for training, 10% for validation, and 10% for testing. We measured the generalization performance using LTR retrotransposons from the genomes of Gardenia jasminoides [64], Daucus carota [65], Abrus precatorius [66], and Asparagus officinalis [67], which were not included in the training dataset. These LTR retrotransposons were detected using EDTA and were processed with the same pipeline used for the elements in InpactorDB (Figure 1). These genomes were downloaded from NCBI (assemblies: ASM1310374v1, ASM162521v1, Abrus_2018, and Aspof.V1). We used all lineages, except those absent from Angiosperms, like Chlamyvir, Tcn1, Phygy, Selgy, TatI,II,III, Osser, Bryco, Lyco, GymcoI,II,III, and IV. In addition, considering their close relationships, we decided to merge Tar and Tork groups into the Tar/Tork group, Ivana and Oryco into the Ivana/Oryco group, and Ogre and Retand into TAT groups [27]. For a better visibility of the lineage names, we renamed Ale as Ale/Retrofit, Tekay as Del/Tekay and Ivana as Ivana/Oryco.

All the experiments were performed using Python 3.6, Scikit-Learn library 0.22 [68] for pre-processing, data partition, and ML algorithms, and tensorflow 2 (keras) [69] for deep neural networks, installed in an Anaconda environment in a Linux operating system over GPU. We ran our tests using the HPC clusters of the Institut Français de Bioinformatique (https://www.france-bioinformatique.fr), IRD (https://bioinfo.ird.fr/), and Genotoul Bioinformatics platform (http://bioinfo.genotoul.fr/), managed by Slurm, and in the BiRD platform (https://pf-bird.univ-nantes.fr/) and Migale Bioinformatics facility (http://migale.jouy.inra.fr/), managed by Sun Grid Engine (SGE).

3. Results

First, we downloaded 9278, 61,730 and 16,137 plant LTR-RT sequences from Repbase, PGSB, and RepetDB, respectively. Additionally, we identified 49,896 elements using LTR_STRUC and 221,052 elements using EDTA. However, EDTA did not predict full-length LTR-RT in eight plant genomes, namely *Calotropis procera*, *Spergula arvensis*, *Diospyros lotus*, *Magnolia ashei*, *Moringa oleifera*, *Passiflora edulis*, *Rafflesia leonardi*, and *Aristotelia chilensis*. It is likely due to low N50 of some assemblies (below 10kb for *S. arvensis*, *D. lotus*, *M. ashei*, *P. edulis*, *R. leonardi*, and *A. chilensis*) and/or the absence of full-length copies of LTR-RT.

A lineage-level classification process was performed for all identified elements and a filtration process was applied. The final redundant library of InpactorDB comprised 130,439 elements from 195 plant species belonging to 108 Angiosperm families (Figure 4, Supplementary Material S3).

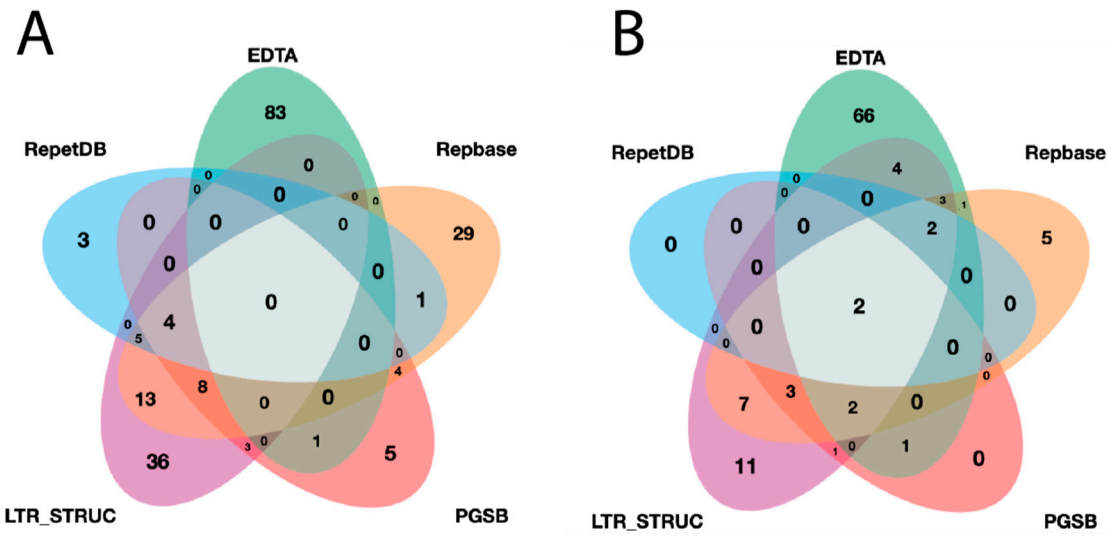

Figure 4. Venn diagrams representing: (**A**) the number of unique and shared plant species between datasets and (**B**) the number of unique and shared plant families between datasets.

3.1. Analysis of Significant Differences

In order to reduce the number of sequences in InpactorDB without losing representativeness and to increase data quality, we used datasets with different characteristics (that is, consensus versus individual genomic sequences and curated versus non-curated sequences) to train the ML algorithms. Using the sequences retained after filtering, we performed an analysis of significant differences to determine if the dataset characteristics (Table 1) affected the performance of eight ML algorithms (LR, LDA, KNN, MLP, RF, DT, NB, and SVC) using k-cross validation with k = 10 (Supplementary Material S4). We could not assume normality of the data or homogeneity of variances; therefore, a non-parametric Kruskal–Wallis test was conducted.

The Kruskal–Wallis test showed a *p*-value lower than 2.2×10^{-16}. Due to the non-normal distribution of the data, a non-parametric pairwise comparison test was applied using Bonferroni's method through a Duncan's range test (Table S1). Table 2 shows the results from the subsets with the best performances. We found no differences between the curated, consensus, PGSB, and RepetDB subsets regarding the performance of the ML algorithms. Since the curation process is more complex than building consensus sequences, we concluded that it is better to remove redundancy through consensus.

Table 2. *p*-values obtained using pairwise comparison through Bonferroni's method.

	Curated	Consensus	PGSB	RepetDB
Curated				
Consensus	1			
PGSB	1	0.188		
RepetDB	1	0.162	1	

3.2. Post-Processing and Classification Using Deep Neural Networks

The methodology used by REPET to build consensus sequences from TEs was applied to our full dataset. Since two datasets are already consensus sequences, we only applied this process to PGSB, LTR_STRUC, and EDTA datasets of InpactorDB. After consensus

creation, we reduced the number of elements to 9608 (6X reduction), 22,530 (6X reduction), and 26,915 (8X reduction) for PGSB, LTR_STRUC, and EDTA subsets, respectively.

The final non-redundant version of InpactorDB consists of 67,241 LTR retrotransposons. Both redundant and non-redundant versions of InpactorDB are available in FASTA format, in which the sequence identifiers have the following general identification code:

>Superfamily-Lineage-plant_family-specie-source-length-ID,

where Superfamily is either RLC (for Copia) or RLG (for Gypsy), Lineage (or family) following the RexDB nomenclature, source (Repbase, RepetDB, PGSB, LTR_STRUC or EDTA datasets), length, and ID, a unique number that identifies each element in InpactorDB. All those fields are separated by a dash character and composed names (as in species) are separated by an underscore character.

The number of consensus sequences for each lineage is unbalanced, probably reflecting the diversity of subfamilies for several lineages (Table 3). Ivana, and Ikeros are the most frequently found lineages in InpactorDB, while TAT, Retrofit and DEL are lineages with the greatest number of elements.

Table 3. Number of elements for each lineage in the non-redundant version of InpactorDB.

Superfamilies	Lineages	Number of Sequences (Redundant)	Number Sequences (Non-Redundant)
Copia	ALE/RETROFIT	19,888	12,026
Copia	ANGELA	6889	1458
Copia	BIANCA	2872	1827
Copia	IKEROS	149	84
Copia	IVANA	88	68
Copia	ORYCO	6135	3468
Copia	SIRE	10,892	3130
Copia	TORK/TAR	11,460	6161
	Total Copia	**58,373**	**28,222**
Gypsy	ATHILA	6611	3499
Gypsy	CRM	4811	2134
Gypsy	DEL/TEKAY	18,330	10,383
Gypsy	GALADRIEL	1715	549
Gypsy	REINA	6387	4531
Gypsy	TAT	34,212	17,923
	Total Gypsy	**72,066**	**39,019**

Using the non-redundant version of InpactorDB, we tested two published deep neural networks, which were used to classified TEs from all orders into superfamilies. First, we used an FNN architecture that applied a hierarchical approach to classify TEs [46]. We also tested a CNN published by [48]. We implemented these architectures using Python 3, Tensorflow 2, and Keras (Supplementary Material S2) with hyper-parameters published by their authors. Both architectures require only 25 epochs to achieve high performance.

Figure 5 shows the training curves for the FNN and Figure 6 for the CNN. Using the FNN, we obtained 98% accuracy, F1-Score, recall, and precision with the validation and test datasets. On the other hand, the CNN had a performance of 97% for the same metrics using the validation and test datasets.

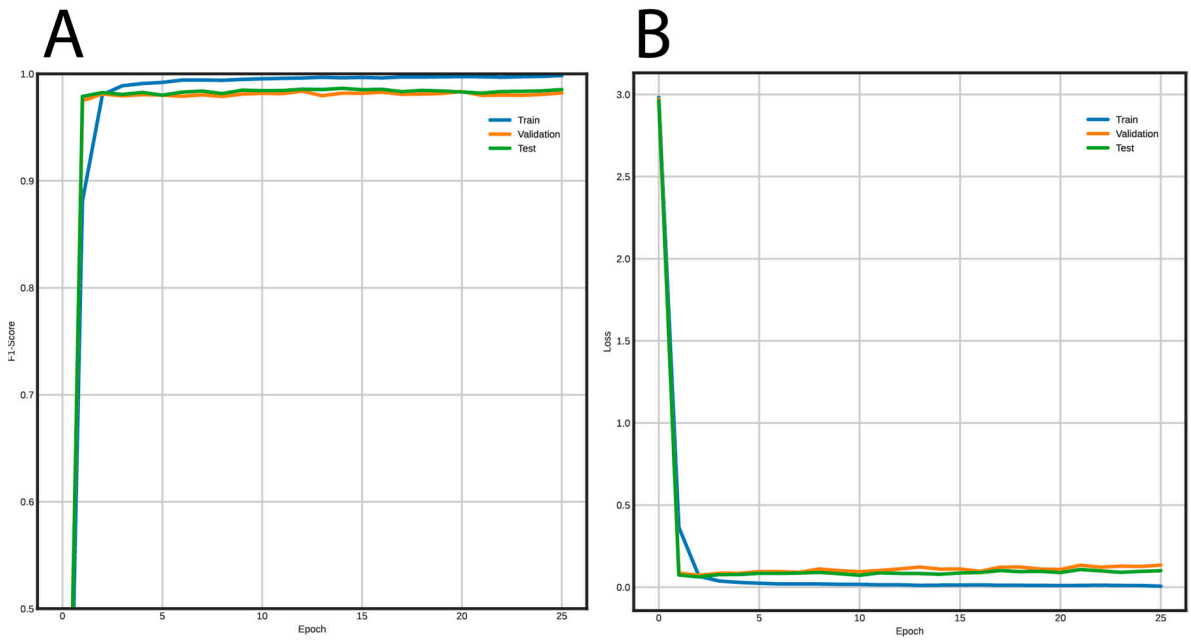

Figure 5. Learning curves using the FNN architecture to classify long terminal repeat (LTR) retrotransposons into lineages. (**A**) F1-Score vs. epochs, and (**B**) loss vs epochs.

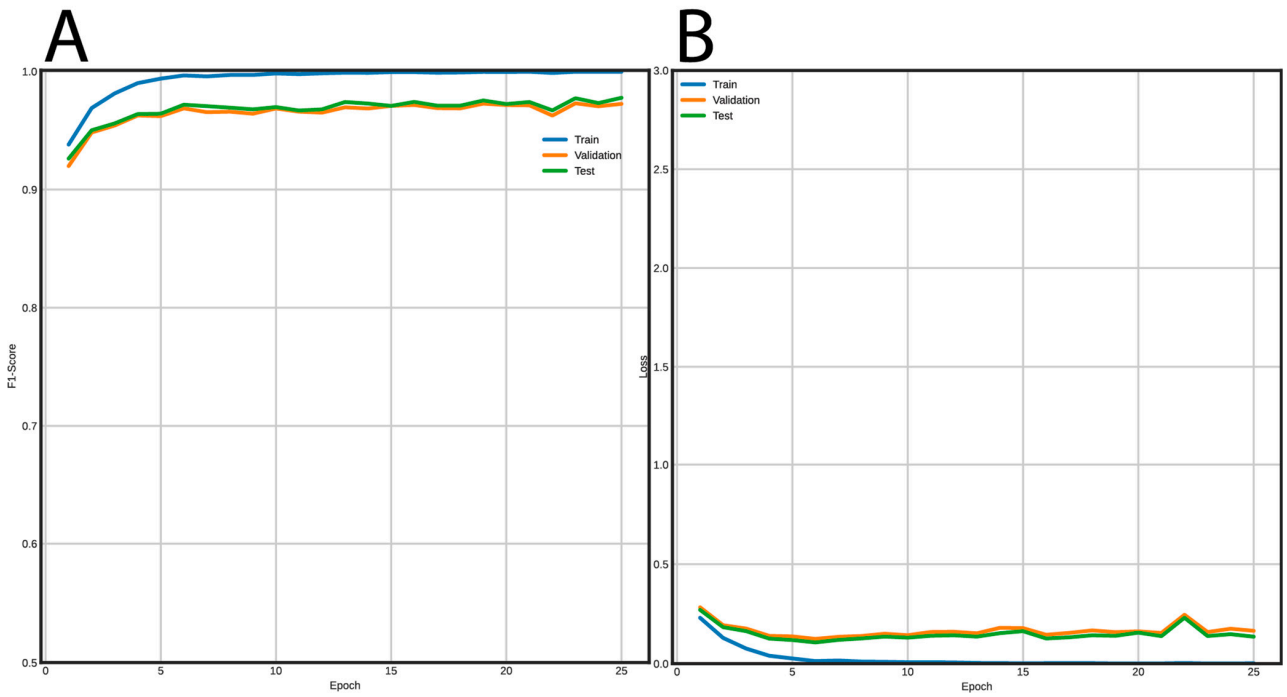

Figure 6. Learning curves using the CNN architecture to classify LTR retrotransposons into lineages. (**A**) F1-Score vs epochs, and (**B**) loss vs epochs.

Most lineages were correctly classified by both DNN architectures, which achieved performances of up to 99–100% (Tables 4 and 5). The lowest F1-Score was found for the Ikeros lineage, likely given the low number of sequences of this lineage in InpactorDB (7).

Table 4. Performance obtained for each lineage using the FNN architecture.

Superfamilies	Lineages/Families	Precision	Recall	F1-Score	Support
Copia	ALE/RETROFIT	0.99	0.99	0.99	1220
Copia	ANGELA	0.96	0.98	0.97	145
Copia	BIANCA	0.99	0.99	0.99	166
Copia	IKEROS	0.67	0.57	0.62	7
Copia	IVANA/ORYCO	0.95	0.97	0.96	319
Copia	TORK/TAR	0.98	0.95	0.96	575
Copia	SIRE	0.99	0.98	0.99	325
Gypsy	CRM	0.98	0.97	0.97	201
Gypsy	GALADRIEL	1.00	0.93	0.96	58
Gypsy	REINA	0.99	1.00	0.99	497
Gypsy	TEKAY/DEL	0.99	0.99	0.99	1059
Gypsy	ATHILA	0.97	0.98	0.97	372
Gypsy	TAT	0.99	0.99	0.99	1787

Table 5. Performance obtained for each lineage using the CNN architecture.

Superfamilies	Lineages/Families	Precision	Recall	F1-Score	Support
Copia	ALE/RETROFIT	0.97	0.99	0.98	1220
Copia	ANGELA	0.96	0.94	0.95	145
Copia	BIANCA	1.00	0.95	0.98	166
Copia	IKEROS	1.00	0.43	0.60	7
Copia	IVANA/ORYCO	0.96	0.93	0.95	319
Copia	TORK/TAR	0.94	0.94	0.94	575
Copia	SIRE	0.99	0.97	0.98	325
Gypsy	CRM	0.97	0.92	0.94	201
Gypsy	GALADRIEL	1.00	0.74	0.85	58
Gypsy	REINA	0.98	0.99	0.98	497
Gypsy	TEKAY/DEL	0.98	0.98	0.98	1059
Gypsy	ATHILA	0.97	0.99	0.98	372
Gypsy	TAT	0.99	1.00	0.99	1787

To test the accuracy of the FNN and CNN under more realistic conditions of LTR-RT classification, we downloaded four plant genomes from species and genera that were not present in our dataset. We selected Gardenia jasminoides, a plant from the Rubiaceae family (Asterids), Daucus carota from Asterids, Abrus precatorius from Rosids, and Asparagus officinalis from monocots. EDTA detected 2648, 1167, 851, and 4692 intact LTR retrotransposons, respectively, and 1010, 628, 579, and 2677 were kept after applying filters, respectively. Using the parameters learned using InpactorDB, the FNN and CNN displayed F1-scores of 97.8% and 86.5% for Gardenia jasminoides, 98.8% and 95.5% for Daucus carota, 99% and 98.1% for Abrus precatorius, and 93.4% and 86.4% for Asparagus officinalis, respectively.

4. Discussion

Given the increasing amount of sequencing data in plants, there is a need to find an automated and rapid way to annotate transposable elements that make up the main part of their genomes. More particularly, LTR retrotransposons constitute the majority of plant DNA (up to 85%) [70] and have crucial roles in genome evolution size, dynamics [71,72] and chromosome organization [3,73]. Furthermore, the high quality detection of TEs improves the accuracy of coding region annotations and functional gene studies [5,74]. Moreover, each lineage of LTR retrotransposons has different dynamics and chromosomal distribution [24,75], and represents different fractions of the genome [5]. For instance, Copia elements are more frequently observed in euchromatin [73,76] and Gypsy retrotransposons are mainly found nested in heterochromatin regions [77,78]. Thus, the classification of

TEs, especially, LTR-RTs, into superfamilies and lineages is crucial to better understanding genome dynamics.

Current computational tools apply several strategies to detect and classify TEs, which can be grouped into homology-based, structure-based, de novo, and those based on comparative genomics [42,46,79,80]. Nevertheless, all these strategies have limitations, such as the dependance on high quality species-specific TE libraries for a homology-based strategy, the incorporation of host multigene families as repeats [26], and the low quality identification for partial or degenerated elements with a structure-based strategy, as well as others [5]. Indeed, the quality of the genome assembly can deeply influence the quality of the detection and the classification.

For detecting LTR-RTs, current tools commonly use structure-based searches in order to take advantage of the well-established features of these kinds of elements. For deep annotation and classification (specifically, classification into linages/families), the homology-based approaches are actually the most frequently used [46]. Since homology-based methods detect TEs based on their similarity to reference TE sequences [81], the quality of the entire process depends on the utilization of a well curated and extensive library or database of TEs. Different plant TEs databases are available (for a list, see [5]), which contain consensus [44,49,82] or genomic [51,83] TE sequences and peptides of coding domains [27,84].

Although there are several databases comprising thousands of TEs, the great structural diversity of these repetitive elements and their species-specificity requires a library with reference LTR retrotransposons from a high number of species from different plant families. For example, PGSB contains 50,000 LTR-RTs from ~60 plant families, RepetDB has ~16,000 from 13 plant species, and Repbase contains ~9700 LTR-RTs from ~70 plant species.

Unlike Repbase, PGSB, and RepetDB, InpactorDB only contains intact full-length LTR-RTs. All elements in this dataset have passed several filters to keep as much as possible LTR retrotransposons that can be used as references. We removed sequences shorter or larger than lengths published by the Gypsy Database (with a tolerance of 20%) to discard incomplete sequences (due to internal deletion for example) and LTR-RTs with nested insertions. Then, we deleted elements with combined domains reported in LTR-RTs from different superfamilies (Copia or Gypsy), suggesting chimeric elements. We also removed ambiguous classified elements with the same number of domains from two or more different lineages (for example, those with two domains from DEL and two domains from REINA lineages). These filters were designed to keep as much as possible LTR-RTs with no nested insertions by other LTR retrotransposons. Finally, we discarded elements with insertions of class II TEs (present in Repbase) to retain putative intact sequences of LTR-RTs.

The currently available databases are valuable resources to annotate LTR-RTs in plant genomes; however, they constitute a small fraction of all sequenced plant species that are representative of plant families. In the data-driven science era, the creation and release of datasets are considered crucial tasks due to their importance in the performance of ML algorithms. A dataset containing repetitive elements from a high diversity of plant species and families is required for tasks regarding LTR retrotransposons, given their natural properties. Thus, more extensive datasets, for training a ML model, could improve its performance and, especially, its generalization because using a data set with more samples for each class (family/lineage) gives the algorithm more information about how the sequences of the LTR-RTs are regardless of the species they come from, reducing the probability of overfitting the model to a certain set of species. Therefore, the algorithm will be more likely to make accurate predictions in genomes that it has never seen before. In an automatic-annotation tool of LTR-RTs, generalization is required since it will be trained using currently sequenced genomes, but it will be used to annotate elements from newly sequenced plant species, which were potentially absent in the training set. Consequently, datasets that include species from a higher number of plant families and genera could improve the probability of the ML algorithm to predict LTR-RTs from new

species. Given this, our main aim was to create a dataset comprising LTR-RTs from different plant species and families; accordingly, InpactorDB contains LTR retrotransposons from 25, 13, and 64 plant species from PGSB, RepetDB, and Repbase, respectively, with additional elements from 69 and 84 plant species using LTR_STRUC and EDTA, respectively. By joining all of these libraries, our dataset consists of more than 130,000 LTR retrotransposons in the redundant version. We observed that consensus and curated databases have the best performance on average for training ML algorithms, with no significant differences between the two. Thus, we applied the process for consensus creation implemented by REPET to build a non-redundant version of InpactorDB with 67,241 elements.

Currently, there is no definitive consensus regarding the LTR-RT classification and naming systems for RT-LTRs. Although many studies use the hierarchical classification of Wicker et al. [16] there is still debate and disagreement on the taxonomy and naming of LTR-RT at different levels of classification [27,29]. Common initiatives are needed for a single classification and naming system at the international level. Our study is based on a robust phylogenetic approach using protein domains for classification [27]. However, as that study used only 56 different plant genomes for the phylogeny of LTR-RT, it cannot be excluded that the diversity of LTR-RT is more complex in plants. An incomplete classification could probably impact automatic classification approaches. Recently, Neumann et al. [27] proposed the separation of several lineages of LTR-RTs based on phylogenetic analyses done with 80 plant genomes. Some of the new lineages appear unevenly distributed among plants families. Clamyvir and Osser are specific of Chlorophyta (a taxon of green algae), while Phygy and Bryco, are specific to Bryophyta (non-vascular land plants), Selgy TatI, Lyco are specific to Lycopodiophyta and finally Gymco is specific to Acrogymnospermae. These lineages are not present in Angiosperm species and as a consequence were not studied here. Interestingly, other lineages show a low number of predicted copies in angiosperm species like Ikeros (84), in the non-redundant version of InpactorDB (Table 3).

Ivana (68) demonstrated very few copy numbers in plant species in the non-redundant version of InpactorDB. In contrast, ALE/Retrofit, TAT, and DEL/Tekay accounted for a large number of samples in the dataset with 12,026, 17,923, and 10,383, respectively. This large unbalance is inadequate for ML algorithms since the model will learn how to classify LTR-RTs from the most frequent lineages, but the performance will reduce in less frequent lineages (Tables 4 and 5). Lineages present in non-angiosperm species were not considered in our DNN tests due to the few number of genomes available in the databases (69 until 2019, where 55 corresponded to green algae [85]), compared with 323 angiosperm genomes found in the databases until 2019 [85]. In the future, it will be essential to include more lineages from non-angiosperms when more genomes will be available for exhaustive classification in plants.

In the current (post) genomic era [86,87], there is a need for automating TE annotation [39,44] to quickly analyze the huge amount of genomic data. Machine learning algorithms have become popular in bioinformatics because they provide promising results in complex tasks and given the availability of large databases. InpactorDB is designed to be a useful tool in the detection and annotation of LTR retrotransposons using both homology-based software (such as RepeatMasker, Supplementary Material S5) and novel free-alignment algorithms based on ML. Using InpactorDB to train two DNN architectures, we obtained up to 98% F1-Score, precision and recall in the problem of classifying LTR retrotransposons into lineages. Additionally, we highlight that only 25 epochs were needed to achieve a good training performance and hyper-parameter tuning was not required. Using more than 4000 LTR-RTs from four plant species that were not included in InpactorDB, we achieved up to a 99% F1-Score using the FNN model, demonstrating good generalization performance. As future work, we propose the use of InpactorDB to generate a new DNN architecture that can improve the performance of the classification of LTR retrotransposons in plant genomes.

5. Conclusions

InpactorDB is a semi-curated dataset of LTR retrotransposons from 195 plant species representing 108 plant families. It comprises more than 130,000 (redundant) and over 60,000 (non-redundant) elements that are classified to the lineage level. InpactorDB was designed to be a tool to annotate LTR-RTs in plant genomes using homology-based algorithms, such as RepeatMasker, and to support automatic, ML-based, and alignment-free software, which is needed to process the large amount of genomic data produced by massive sequencing projects in the current post-genomic era. Given the high diversity of plant species and families contained in the dataset and the filters applied to the LTR-RT sequences, InpactorDB can be used as a basis to train ML algorithms or DNN architectures towards the implementation of automatic TE annotators in plant genomes.

Supplementary Materials: The following are available online at https://www.mdpi.com/2073-4425/12/2/190/s1, Figure S1. Boxplot of F1-Score performance of all algorithms used different subsets, Table S1. Values obtained by pairwise comparisons using Bonferroni's method. The colored cells have a p-value < 0.05, Supplementary Material S1: Assemblies used to find LTR-RTs with LTR_STRUC and EDTA, Supplementary Material S2: Jupyter Notebook of the implementation of both deep neural networks, Supplementary Material S3: Plant species and families of each dataset. Supplementary Material S4: F1-Scores of ML algorithms trained with each subset used in the statistical analysis. Supplementary Material S5: Repeat Masker annotation results of *Coffea canephora*, *Ananas comosus*, and *Oryza sativa* using as custom library InpactorDB, Repbase, and RepetDB.

Author Contributions: Conceptualization, S.O.-A., G.I. and R.G.; methodology, S.O.-A., P.A.J., M.S.C., C.F.J.-V. and R.T.-S.; software, S.O.-A. and R.T.-S.; data curation, P.A.J. and M.S.C.; writing—original draft preparation, S.O.-A., P.A.J., M.S.C., C.F.J.-V., R.T.-S., G.I. and R.G.; writing—review and editing, S.O.-A., P.A.J., M.S.C., C.F.J.-V., R.T.-S., G.I. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: Simon Orozco-Arias is supported by a Ph.D. grant from the Ministry of Science, Technology and Innovation (Minciencias) of Colombia, Grant Call 785/2017. The authors and publication fees were supported by Universidad Autónoma de Manizales, Manizales, Colombia under project 589-089, and Romain Guyot was supported by the LMI BIO-INCA. The funders had no role in the study design, data collection and analysis, the decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: InpactorDB is available at Zenodo under the doi:10.5281/zenodo.4386317 and at Datasud (https://dataverse.ird.fr) under the doi:10.23708/QCMOUA.

Acknowledgments: The authors acknowledge the IFB Core Cluster that is part of the National Network of Compute Resources (NNCR) of the Institut Français de Bioinformatique (https://www.france-bioinformatique.fr), the Genotoul Bioinformatics platform (http://bioinfo.genotoul.fr/), and the IRD itrop (https://bioinfo.ird.fr/) at IRD Montpellier for providing HPC resources that have contributed to the research results reported in this paper. The authors thank STICAMSUD 21-STIC-13" TELearning" for support.

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

References

1. Li, S.-F.; Su, T.; Cheng, G.-Q.; Wang, B.-X.; Li, X.; Deng, C.-L.; Gao, W.-J. Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants. *Genes* **2017**, *8*, 290. [CrossRef] [PubMed]
2. Keidar, D.; Doron, C.; Kashkush, K. Genome-wide analysis of a recently active retrotransposon, Au SINE, in wheat: Content, distribution within subgenomes and chromosomes, and gene associations. *Plant Cell Rep.* **2018**, *37*, 193–208. [CrossRef] [PubMed]
3. Kim, N.-S. The genomes and transposable elements in plants: Are they friends or foes? *Genes Genom.* **2017**, *39*, 359–370. [CrossRef]
4. De Castro Nunes, R.; Orozco-Arias, S.; Crouzillat, D.; Mueller, L.A.; Strickler, S.R.; Descombes, P.; Fournier, C.; Moine, D.; de Kochko, A.; Yuyama, P.M.; et al. Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions. *Front. Plant Sci.* **2018**, *9*. [CrossRef]
5. Orozco-Arias, S.; Isaza, G.; Guyot, R. Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *Int. J. Mol. Sci.* **2019**, *20*, 3837. [CrossRef]

6. Todorovska, E. Retrotransposons and their role in plant—Genome evolution. *Biotechnol. Biotechnol. Equip.* **2014**, *21*, 294–305. [CrossRef]
7. Wessler, S.R.; Bureau, T.E.; White, S.E. LTR-retrotransposons and MITEs: Important players in the evolution of plant genomes. *Curr. Opin. Genet. Dev.* **1995**, *5*, 814–821. [CrossRef]
8. Casacuberta, J.M.; Santiago, N. Plant LTR-retrotransposons and MITEs: Control of transposition and impact on the evolution of plant genes and genomes. *Gene* **2003**, *311*, 1–11. [CrossRef]
9. Galindo-González, L.; Mhiri, C.; Deyholos, M.K.; Grandbastien, M.-A. LTR-retrotransposons in plants: Engines of evolution. *Gene* **2017**, *626*, 14–25. [CrossRef]
10. Fan, F.; Wen, X.; Ding, G.; Cui, B. Isolation, identification, and characterization of genomic LTR retrotransposon sequences from masson pine (Pinus massoniana). *Tree Genet. Genomes* **2013**, *9*, 1237–1246. [CrossRef]
11. Schulman, A.H. Hitching a Ride: Nonautonomous Retrotransposons and Parasitism as a Lifestyle. In *Plant Transposable Elements*; Grandbastien, M.-A., Casacuberta, J.M., Eds.; Springer: Berlin/Heidelberg, Germany, 2012; pp. 71–88.
12. Alzohairy, A.M.; Sabir, J.S.M.; Gyulai, G.; Younis, R.A.A.; Jansen, R.K.; Bahieldin, A. Environmental stress activation of plant long-terminal repeat retrotransposons. *Funct. Plant Biol.* **2014**, *41*, 557–567. [CrossRef]
13. Serrato-Capuchina, A.; Matute, D.R. The role of transposable elements in speciation. *Genes* **2018**, *9*, 254. [CrossRef] [PubMed]
14. Kidwell, M.G.; Kidwell, J.F.; Sved, J.A. Hybrid dysgenesis in Drosophila melanogaster: A syndrome of aberrant traits including mutation, sterility and male recombination. *Genetics* **1977**, *86*, 813–833.
15. Zhang, Q.-J.; Gao, L.-Z. Rapid and Recent Evolution of LTR Retrotransposons Drives Rice Genome Evolution During the Speciation of AA- Genome Oryza Species. *G3 Genes Genomes Genet.* **2017**, *7*, 1875–1885. [CrossRef] [PubMed]
16. 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]
17. Chaparro, C.; Gayraud, T.; de Souza, R.F.; Domingues, D.S.; Akaffou, S.; Laforga Vanzela, A.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]
18. Orozco-Arias, S.; Isaza, G.; Guyot, R.; Tabares-Soto, R. A systematic review of the application of machine learning in the detection and classi fi cation of transposable elements. *PeerJ* **2019**, *7*, 18311. [CrossRef]
19. Grandbastien, M.-A.A. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. *Biochim. Biophys. Acta Gene Regul. Mech.* **2015**, *1849*, 403–416. [CrossRef]
20. Gao, D.; Jimenez-Lopez, J.C.; Iwata, A.; Gill, N.; Jackson, S.A. Functional and structural divergence of an unusual LTR retrotransposon family in plants. *PLoS ONE* **2012**, *7*, e48595. [CrossRef]
21. Rahman, A.Y.A.; Usharraj, A.O.; Misra, B.B.; Thottathil, G.P.; Jayasekaran, K.; Feng, Y.; Hou, S.; Ong, S.Y.; Ng, F.L.; Lee, L.S.; et al. Draft genome sequence of the rubber tree Hevea brasiliensis. *BMC Genom.* **2013**, *14*, 75. [CrossRef]
22. Kumar, A.; Bennetzen, J.L. Plant retrotransposons. *Annu. Rev. Genet.* **1999**, *33*, 479–532. [CrossRef] [PubMed]
23. Servant, G.; Deininger, P.L. Insertion of retrotransposons at chromosome ends: Adaptive response to chromosome maintenance. *Front. Genet.* **2016**, *6*, 358. [CrossRef] [PubMed]
24. Gao, D.; Chen, J.; Chen, M.; Meyers, B.C.; Jackson, S. A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. *PLoS ONE* **2012**, *7*, e32010. [CrossRef] [PubMed]
25. Orozco-Arias, S.; Tabares-Soto, R.; Ceballos, D.; Guyot, R. Parallel Programming in Biological Sciences, Taking Advantage of Supercomputing in Genomics. In *Advances in Computing*; Solano, A., Ordoñez, H., Eds.; Springer: Zurich, Switzerland, 2017; Volume 735, pp. 627–643.
26. Arkhipova, I.R. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. *Mob. DNA* **2017**, *8*, 19. [CrossRef]
27. Neumann, P.; Novák, P.; Hoštáková, 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*, 1–17. [CrossRef]
28. 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]
29. 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*, D70–D74. [CrossRef]
30. Palazzo, A.; Lorusso, P.; Miskey, C.; Walisko, O.; Gerbino, A.; Marobbio, C.M.T.; Ivics, Z.; Marsano, R.M. Transcriptionally promiscuous "blurry" promoters in Tc1/mariner transposons allow transcription in distantly related genomes. *Mob. DNA* **2019**, *10*, 13. [CrossRef]
31. Smit, A.F.A.; Hubley, R.; Green, P. RepeatMasker. 1996. Available online: http://www.repeatmasker.org/ (accessed on 25 January 2021).
32. Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Sanyal, 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**, *21*, 1262–1269. [CrossRef]

33. Ammiraju, J.S.S.; Zuccolo, A.; Yu, Y.; Song, X.; Piegu, B.; Chevalier, F.; Walling, J.G.; Ma, J.; Talag, J.; Brar, D.S.; et al. Evolutionary dynamics of an ancient retrotransposon family provides insights into evolution of genome size in the genus Oryza. *Plant J.* **2007**, *52*, 342–351. [CrossRef]
34. 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]
35. Stritt, C.; Wyler, M.; Gimmi, E.L.; Pippel, M.; Roulin, A.C. Diversity, dynamics and effects of long terminal repeat retrotransposons in the model grass Brachypodium distachyon. *New Phytol.* **2020**, *227*, 1736–1748. [CrossRef] [PubMed]
36. Ma, B.; Kuang, L.; Xin, Y.; He, N. New Insights into Long Terminal Repeat Retrotransposons in Mulberry Species. *Genes* **2019**, *10*, 285. [CrossRef] [PubMed]
37. Domingues, D.S.; Cruz, G.M.Q.; Metcalfe, C.J.; Nogueira, F.T.S.; Vicentini, R.; Alves, C.; Van Sluys, M.-A. Analysis of plant LTR-retrotransposons at the fine-scale family level reveals individual molecular patterns. *BMC Genom.* **2012**, *13*, 137. [CrossRef] [PubMed]
38. Ou, S.; Chen, J.; Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **2018**, *46*, 1–11. [CrossRef]
39. Orozco-Arias, S.; Piña, J.S.; Tabares-Soto, R.; Castillo-Ossa, L.F. Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements. *Processes* **2020**, *8*, 638. [CrossRef]
40. Mustafin, R.N.; Khusnutdinova, E.K. The Role of Transposons in Epigenetic Regulation of Ontogenesis. *Russ. J. Dev. Biol.* **2018**, *49*, 61–78. [CrossRef]
41. Loureiro, T.; Camacho, R.; Vieira, J.; Fonseca, N.A. Boosting the Detection of Transposable Elements Using Machine Learning. In *7th International Conference on Practical Applications of Computational Biology & Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 85–91.
42. Loureiro, T.; Camacho, R.; Vieira, J.; Fonseca, N.A. Improving the performance of Transposable Elements detection tools. *J. Integr. Bioinform.* **2013**, *10*, 231. [CrossRef]
43. Santos, B.Z.; Cerri, R.; Lu, R.W. A New Machine Learning Dataset for Hierarchical Classification of Transposable Elements. In Proceedings of the XIII Encontro Nacional de Inteligência Artificial, Recife, Brazil, 9–12 October 2016; pp. 9–12.
44. Cornut, G.; Choisne, N.; Alaux, M.; Alfama-Depauw, F.; Jamilloux, V.; Maumus, F.; Letellier, T.; Luyten, I.; Pommier, C.; Adam-Blondon, A.-F.; et al. RepetDB: A unified resource for transposable element references. *Mob. DNA* **2019**, *10*, 6.
45. Schietgat, L.; Vens, C.; Cerri, R.; Fischer, C.N.; Costa, E.; Ramon, J.; Cararetо, C.M.A.; Blockeel, H. A machine learning based framework to identify and classify long terminal repeat retrotransposons. *PLoS Comput. Biol.* **2018**, *14*, e1006097. [CrossRef]
46. Nakano, F.K.; Mastelini, S.M.; Barbon, S.; Cerri, R. Improving Hierarchical Classification of Transposable Elements using Deep Neural Networks. In Proceedings of the International Joint Conference on Neural Networks, Rio de Janeiro, Brazil, 8–13 July 2018.
47. Da Cruz, M.H.P.; Domingues, D.S.; Saito, P.T.M.; Paschoal, A.R.; Bugatti, P.H. TERL: Classification of Transposable Elements by Convolutional Neural Networks. *bioRxiv* **2020**. [CrossRef] [PubMed]
48. Yan, H.; Bombarely, A.; Li, S. DeepTE: A computational method for de novo classification of transposons with convolutional neural network. *Bioinformatics* **2020**. [CrossRef] [PubMed]
49. Jurka, J.; Kapitonov, V.V.; Pavlicek, A.; Klonowski, P.; Kohany, O.; Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **2005**, *110*, 462–467. [CrossRef]
50. Spannagl, M.; Bader, K.; Pfeifer, M.; Nussbaumer, T.; Mayer, K.F.X. PGSB/MIPS Plant Genome Information Resources and Concepts for the Analysis of Complex Grass Genomes. In *Plant Bioinformatics*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 165–186.
51. Spannagl, M.; Nussbaumer, T.; Bader, K.C.; Martis, M.M.; Seidel, M.; Kugler, K.G.; Gundlach, H.; Mayer, K.F.X. PGSB PlantsDB: Updates to the database framework for comparative plant genome research. *Nucleic Acids Res.* **2015**, *44*, D1141–D1147. [CrossRef] [PubMed]
52. McCarthy, E.M.; McDonald, J.F. LTR STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, *19*, 362–367. [CrossRef]
53. Ou, S.; Su, W.; Liao, Y.; Chougule, K.; Agda, J.R.A.; Hellinga, A.J.; Lugo, C.S.B.; Elliott, T.A.; Ware, D.; Peterson, T.; et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **2019**, *20*, 275. [CrossRef]
54. 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]
55. Ellinghaus, D.; Kurtz, S.; Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform.* **2008**, *14*. [CrossRef]
56. Ou, S.; Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal-repeat retrotransposons. *Plant Physiol.* **2017**, *176*. [CrossRef]
57. Orozco-Arias, S.; Liu, J.; Id, R.T.; Ceballos, D.; Silva, D.; Id, D.; Ming, R.; Guyot, R. Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology* **2018**, *7*, 32. [CrossRef]

58. 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. In *Advances in Computing*; Springer: Berlin/Heidelberg, Germany, 2017; Volume 735, pp. 156–170.
59. Altschup, S.F.; Gish, W.; Pennsylvania, T.; Park, U. Basic Local Alignment Search Tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
60. Miele, V.; Penel, S.; Duret, L. Ultra-fast sequence clustering from similarity networks with SiLiX. *BMC Bioinform.* **2011**, *12*, 116. [CrossRef] [PubMed]
61. 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]
62. Capella-Gutiérrez, S.; Silla-Martínez, J.M.; Gabaldón, T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **2009**, *25*, 1972–1973. [CrossRef] [PubMed]
63. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European molecular biology open software suite. *TIG* **2000**, *16*, 276–277. [CrossRef]
64. Xu, Z.; Pu, X.; Gao, R.; Demurtas, O.C.; Fleck, S.J.; Richter, M.; He, C.; Ji, A.; Sun, W.; Kong, J.; et al. Tandem gene duplications drive divergent evolution of caffeine and crocin biosynthetic pathways in plants. *BMC Biol.* **2020**, *18*, 1–14. [CrossRef]
65. Iorizzo, M.; Ellison, S.; Senalik, D.; Zeng, P.; Satapoomin, P.; Huang, J.; Bowman, M.; Iovene, M.; Sanseverino, W.; Cavagnaro, P.; et al. A high-quality carrot genome assembly provides new insights into carotenoid accumulation and asterid genome evolution. *Nat. Genet.* **2016**, *48*, 657–666. [CrossRef]
66. Zhang, R.; Wang, Y.-H.; Jin, J.-J.; Stull, G.W.; Bruneau, A.; Cardoso, D.; De Queiroz, L.P.; Moore, M.J.; Zhang, S.-D.; Chen, S.-Y.; et al. Exploration of plastid phylogenomic conflict yields new insights into the deep relationships of Leguminosae. *Syst. Biol.* **2020**, *69*, 613–622. [CrossRef]
67. Li, Q.; Zhang, N.; Zhang, L.; Ma, H. Differential evolution of members of the rhomboid gene family with conservative and divergent patterns. *New Phytol.* **2015**, *206*, 368–380. [CrossRef]
68. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
69. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; et al. Tensorflow: A System for Large-Scale Machine Learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
70. Bonchev, G.N. Useful parasites: The evolutionary biology and biotechnology applications of transposable elements. *J. Genet.* **2016**, *95*, 1039–1052. [CrossRef] [PubMed]
71. Cossu, R.M.; Buti, M.; Giordani, T.; Natali, L.; Cavallini, A. A computational study of the dynamics of LTR retrotransposons in the Populus trichocarpa genome. *Tree Genet. Genomes* **2012**, *8*, 61–75. [CrossRef]
72. Bento, M.; Tomás, D.; Viegas, W.; Silva, M. Retrotransposons represent the most labile fraction for genomic rearrangements in polyploid plant species. *Cytogenet. Genome Res.* **2013**, *140*, 286–294. [CrossRef] [PubMed]
73. Vicient, C.M.; Casacuberta, J.M. Impact of transposable elements on polyploid plant genomes. *Ann. Bot.* **2017**, *120*, 195–207. [CrossRef] [PubMed]
74. Paz, R.C.; Kozaczek, M.E.; Rosli, H.G.; Andino, N.P.; Sanchez-Puerta, M.V.; Cristina Paz, R.; Eliana Kozaczek, M.; Guillermo Rosli, H.; Pilar Andino, N.; Virginia Sanchez-Puerta, M. Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in Solanum lycopersicum. *Genetica* **2017**, *145*, 417–430. [CrossRef]
75. Gao, D.; Li, Y.; Kim, K.D.; Abernathy, B.; Jackson, S.A. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. *Genome Biol.* **2016**, *17*, 7. [CrossRef]
76. Tang, X.; Datema, E.; Guzman, M.O.; de Boer, J.M.; van Eck, H.J.; Bachem, C.W.B.; Visser, R.G.F.; de Jong, H. Chromosomal organizations of major repeat families on potato (Solanum tuberosum) and further exploring in its sequenced genome. *Mol. Genet. Genom.* **2014**, *289*, 1307–1319. [CrossRef]
77. Gao, D.; Abernathy, B.; Rohksar, D.; Schmutz, J.; Jackson, S.A. Annotation and sequence diversity of transposable elements in common bean (Phaseolus vulgaris). *Front. Plant Sci.* **2014**, *5*, 339. [CrossRef]
78. Gao, D.; Jiang, N.; Wing, R.A.; Jiang, J.; Jackson, S.A. Transposons play an important role in the evolution and diversification of centromeres among closely related species. *Front. Plant Sci.* **2015**, *6*, 216. [CrossRef]
79. Jiang, S.-Y.; Ramachandran, S. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. *PLoS ONE* **2013**, *8*, e71118. [CrossRef]
80. Rawal, K.; Ramaswamy, R. Genome-wide analysis of mobile genetic element insertion sites. *Nucleic Acids Res.* **2011**, *39*, 6864–6878. [CrossRef] [PubMed]
81. Hermann, D.; Egue, F.; Tastard, E.; Nguyen, D.-H.; Casse, N.; Caruso, A.; Hiard, S.; Marchand, J.; Chenais, B.; Morant-Manceau, A.; et al. An introduction to the vast world of transposable elements—What about the diatoms? *Diatom Res.* **2014**, *29*, 91–104. [CrossRef]
82. Wicker, T.; Matthews, D.E.; Keller, B. TREP: A database for Triticeae repetitive elements. *Trends Plant Sci.* **2002**, *7*, 561. [CrossRef]
83. Du, J.; Grant, D.; Tian, Z.; Nelson, R.T.; Zhu, L.; Shoemaker, R.C.; Ma, J. SoyTEdb: A comprehensive database of transposable elements in the soybean genome. *BMC Genom.* **2010**, *11*, 113. [CrossRef]
84. Arensburger, P.; Piégu, B.; Bigot, Y. The future of transposable element annotation and their classification in the light of functional genomics—What we can learn from the fables of Jean de la Fontaine? *Mob. Genet. Elements* **2016**, *6*, e1256852. [CrossRef]
85. Kersey, P.J. Plant genome sequences: Past, present, future. *Curr. Opin. Plant Biol.* **2019**, *48*, 1–8. [CrossRef]

86. Rishishwar, L.; Wang, L.; Clayton, E.A.; Mariño-Ramírez, L.; McDonald, J.F.; Jordan, I.K. Population and clinical genetics of human transposable elements in the (post) genomic era. *Mob. Genet. Elements* **2017**, *7*, 1–20. [CrossRef]
87. Chen, W.; Feng, P.M.; Deng, E.Z.; Lin, H.; Chou, K.C. iTIS-PseTNC: A sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Anal. Biochem.* **2014**, *462*, 76–83. [CrossRef]