A computational Architecture to identify and classify LTR retrotransposons in plant genomes

Simon Orozco-Arias

Universidad de Caldas
Doctoral Program in Engineering
Line of research in Biocomputational models and Bioinformatics
Faculty of Engineering
Manizales, Colombia
2022

A computational Architecture to identify and classify LTR retrotransposons in plant genomes

Simon Orozco-Arias

Thesis for the degree of:
Ph.D. in Engineering

Advisor:
Ph.D. Gustavo Isaza Echeverri

Co-Advisor:
Ph.D. Romain Guyot

Line of research in Biocomputational models and Bioinformatics
Research group: GITIR

Universidad de Caldas
Faculty of engineering
Manizales, Colombia
2022

Dedicatory

It is paradoxical, yet true, to say, that the more we know, the more ignorant we become in the absolute sense, for it is only through enlightenment that we become conscious of our limitations. Precisely one of the most gratifying results of intellectual evolution is the continuous opening up of new and greater prospects.

Nikola Tesla.

To my family, friends, students and professors.

Acknowledgements

I thank my professors, who instilled in me the curiosity of research, my parents who supported me at all times and made me believe that limits only exist in our heads. I thank my advisors, Dr. Gustavo Isaza for believing in me, for guiding me in both academic and personal aspects and Dr. Romain Guyot for supporting my entire training process since before the doctorate and for sharing his vast experience and knowledge. I also thank my students Johan Piña, Nicolás Tobón, Mariana Candamil, Paula Jaimes, Estiven Valencia, Maradey Arias, Luis López, among many others, for trusting me and for allowing me to accompany them in their educational process. I thank my colleague and friend Reinel Tabares-Soto for the long hours of work, discussion and ideas to overcome all the ups and downs of our doctoral processes, which we always graciously said, ”we are doing two doctorates”(his and mine). Finally, I thank the Universidad de Caldas for my entire training process (from undergraduate to Ph.D.), the Ministry of Science, Technology and Innovation of Colombia, Minciencias, for financing my studies, thank you for believing that high level education is an excellent way to build a different country and the Universidad Autónoma de Manizales that believed in this project, financed several of the publications, supported me with undergraduate students and helped me to pay for several scientific events.

Abstract

This PhD thesis focused on the application of machine learning and deep learning techniques for the study of LTR retrotransposons, with the aim of improving the understanding at the genomic level of plants of agro-industrial interest such as rice, maize, coffee and sugar cane, and which could be applied to any other plant genome or other organisms.

Recent research has demonstrated the impact of transposable elements on the phenotype of crops of interest, such as the colour of maize kernels, the colour and flavour of oranges, the skin colour of potatoes, the size and shape of tomatoes, and the colour and flavour of grapes, which are produced by the insertion of these elements near or into genes. Although bioinformatics techniques and tools exist for the detection and classification of transposable elements, it is not yet possible to obtain reliable results, due to the great diversity of their structures, replication patterns and life cycles. In addition, these genomic components have characteristics that make their study very complex, such as species specificity, high diversity at the nucleotide level (low homology between sequences), long non-coding regions and their repetitive nature. Therefore, new techniques such as machine learning and deep learning could improve performance in terms of both execution time and accuracy of results.

In the development of this research project, the most well-known machine learning algorithms were used, as well as some deep neural network architectures that have become widespread in the scientific community in recent years. Feature extraction and selection methods, pre-processing techniques, algorithms and architectures that have been successfully used on datasets similar to transposable features were extrapolated. Also, this Ph.D. thesis will have a positive impact on the scientific community in the fields of bioinformatics, genomics and agriculture, as the software developed here and its use on other genomes could serve as a basis for future research related to genetic improvement, understanding the evolution of species and the relationship between organisms and the environment. In addition, knowledge was generated on the use of new techniques on genomic data (especially LTR retrotransposons), such as the influence of the nature of the data on the accuracy of the results, better pre-processing techniques (feature selection and extraction, dimensionality reduction, data transformation, among others), better hyper-parameters and metrics that better fit such elements.

Finally, this research proposal led to the creation of a functional bioinformatics software that, thanks to the selected techniques, allows the detection and classification of LTR retrotransposons in plants of interest. This software is available to the scientific community and can be used in the context of several massive genome sequencing and assembly projects, such as the $3,000$ rice genomes project, the sequencing of $10,000$ plant genomes or the 1.5 million eukaryotic species sequencing project. All the codes and scripts developed during this project are available at https://github.com/simonorozcoarias/MLinTEs.

Contents

1. Introduction

1.1. Background

Transposable elements (TEs) are genomic units that have the ability to replicate or move through the chromosomes of virtually all living organisms. These elements make up the majority of the nuclear DNA content of many plant genomes. This is particularly true for large cereal genomes such as wheat, barley and maize, for which up to $85\,\%$ of the sequenced DNA is sorted into repeated sequences [1]. In contrast, compact genomes such as those of *Arabidopsis thaliana* ($10\,\%$) and the carnivorous plant *Utricularia gibba* ($3\,\%$) have a lower content of TEs [2], suggesting that their copy number can vary dramatically and are associated with genome size variation [3]. TEs can be activated across a broad panel of biotic and abiotic stresses [4, 5], suggesting that they may play a significant role in environmental adaptation [6]. In addition, several investigations have demonstrated the profound impact of TEs on their host genomes, especially in plants, such as within-species variability [7], inactivation [8] or over-expression of genes [9], key functions in chromosomal structures [10] and fundamental roles in species evolution [11].

Transposable elements are traditionally classified according to their life cycle or structure into two classes [12]: class I or retrotransposons and class II or DNA transposons. In addition, Class I includes four orders including LTR (Long Terminal Repeats) retrotransposons, non-LTR retrotransposons, Penelope-like elements (PLEs) and Dictyostelium intermediate repeat sequence (DIRS); while Class II contains transposons with Terminal Inverted Repeats (TIRs) and helitrons. The most common TEs in plant genomes are LTR retrotransposons, which can reach up to $75\,\%$ of the maize genome [13], $67\,\%$ of wheat [1], $55\,\%$ of sorghum (*Sorgum bicolor*) [14] and $42\,\%$ of the coffee genome [15].

Because each type of TE can cause different effects on organisms, reliable deep-level classification (into superfamilies and lineages) is of great importance. For example, it has been shown that the centromeres of plants such as coffee and maize are mainly composed of a specific lineage of LTR retrotransposons called Centromeric Retrotransposons [16]. In addition, each subclass of TEs has different distributions within chromosomes, thus having different relationships with genes [17]. For this reason, knowing the superfamily and lineage to which a TE belongs could help to understand what effect it has on the organism under study.

Along the same lines, in recent years the influence of TEs in the variability of the phenotype of

multiple crops has been demonstrated, as can be seen in Figure **1-1**, being influential for example in the different skin colours of grapes [18], pigment instability in maize kernels (from completely yellow, spotted kernels to completely purple), tomato shape and size [19], potato skin colour [20], different orange colours and flavours [21] and presence or absence of peach skin hairs [22]. These mutations can be categorised into six mechanisms: (i) gene inactivation caused by TE insertions in coding regions or introns, (ii) differential gene expression caused by TE insertions in or near regulatory regions, (iii) TE-mediated genomic rearrangements resulting in gene insertion, deletion or duplication, (iv) transduction-based regulation of gene expression, (v) transposase exaptation responsible for modification of transcription factor capacities [23], and (vi) mechanisms based on epigenetics such as siRNA, methylation, among others [24].

In bioinformatics, a science that integrates computational methods for solving biological problems [25], some methods have been developed to annotate TEs. This process consists of some steps such as identification and classification [26]. In the identification or detection step, the main goal is to separate TEs from any other genomic components such as genes, microsatellites, tandem repeats, among others. This process generates the input for the classification step, where using order, superfamily or lineage-specific characteristics, each element is assigned a grouping. Annotation methods are generally classified into four main categories [27, 28, 29]: de novo, structure-based, comparative genomics-based and homology-based. However, annotation in genomes is a well-known problem in genomics. This is due to the repetitive nature of TEs, their high diversity at the nucleotide level even among elements of the same lineage (low sequence homology) and their species-specific nature [30]. The aforementioned attributes make their identification and classification very complex and unreliable [9], which causes problems not only for researchers interested in repetitive sequences, but also for functional studies. These problems include unassembled genome sections, altered gene annotation, inability to deeply understand certain mechanisms of gene activation or silencing, among others.

In recent years, several datasets consisting of thousands of TEs from various species have been created and published, such as Repbase [31], RepetDB [32], PGSB Plants DB [33] and InpactorDB (released as part of this work) [34]. These datasets constitute valuable resources for improving tasks such as TE detection and classification, and have motivated the proposal and evaluation of novel computational techniques to obtain substantial results in terms of accuracy and speed in executing these tasks [26, 35].

Computational approaches such as supercomputing [36], artificial intelligence [37] and data mining [25] are currently widely used in the biological sciences, demonstrating great improvements in both obtaining results and decreasing run times. Machine learning (ML) is defined as the programming of computers to improve and optimise a performance criterion using already processed data or past experience [38]. ML has been applied to solve many bioinformatics problems, such as in genomics [39], in systems biology and evolution [40], and in the identification and classifica-

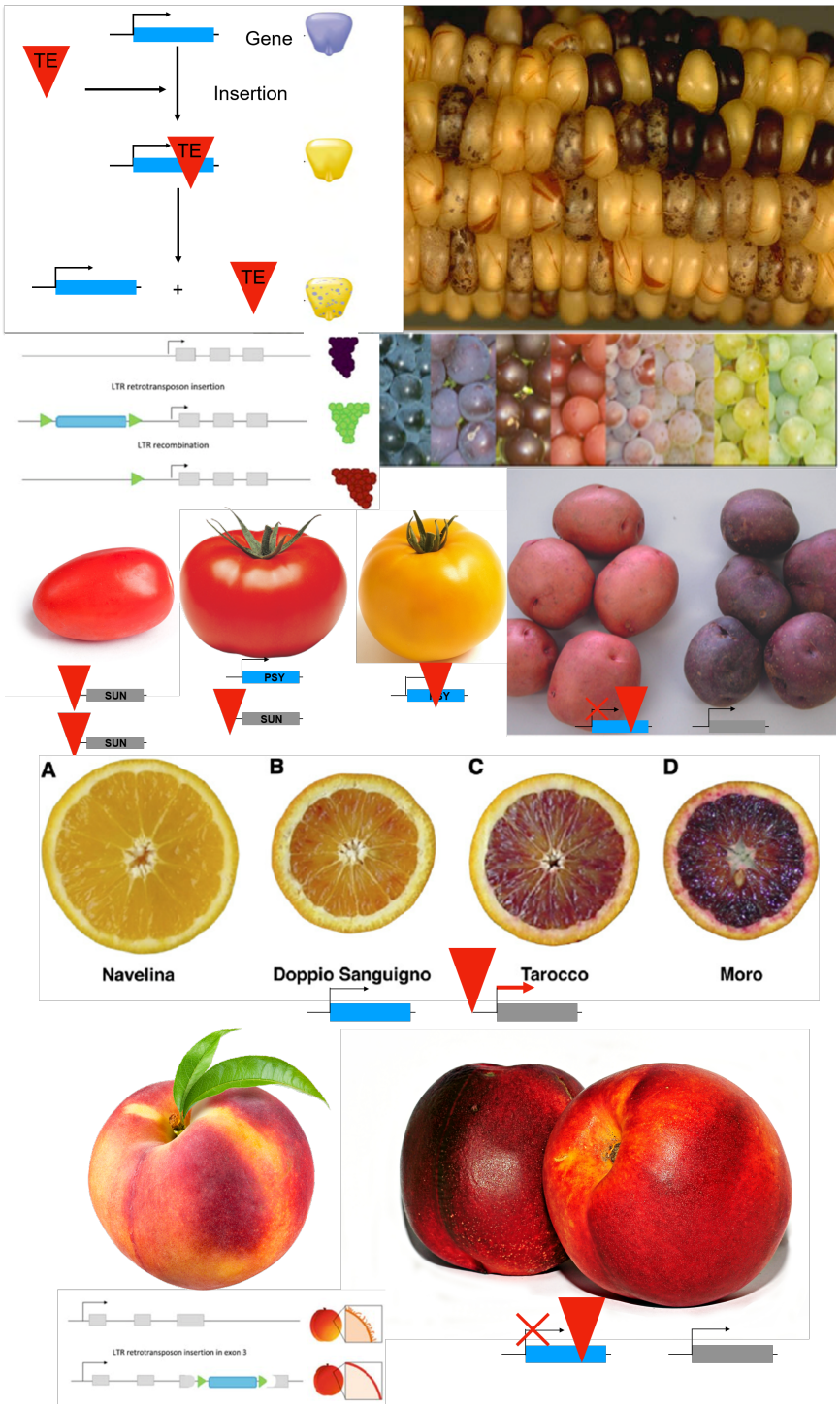

Figure 1-1.: *Crop phenotype variations caused by TEs. Adapted from [18, 19, 20, 21, 22].*

tion of a specific order of retrotransposons [26, 37], demonstrating the benefits of using this type of technology in biological fields. Additionally, the literature reports that the use of deep neural networks (also called deep learning, DL) in the classification of TEs at the superfamily level improves the result dramatically [41, 42, 43]. However, the aforementioned works have been mostly exploratory and have not conducted in-depth studies of various key computational components such as selection of the most informative metrics, various pre-processing techniques (different DNA encoding schemes, dataset augmentation, feature extraction and selection, among others), hyper-parameter tuning processes, selection of different ML algorithms and architectures, and ensemble methods.

1.2. Research problem

According to the United Nations (UN) climate change is one of the biggest challenges today. The effects caused by this phenomenon include rising sea levels, stronger tropical storms, more extreme droughts, more intense rainfall and even threatens to reduce agricultural production. This puts the food security of millions of people around the world at risk. According to the UN, the number of people suffering from malnutrition has been increasing since 2014, reaching 750 million in 2019 and rising up to 840 million in 2030 [44]. The most affected regions are Africa and Latin America, where $21\,\%$ and $5\,\%$ of the population respectively suffer from this problem [45]. Therefore, $21.3\,\%$ (144 million) of children under five years of age were stunted in 2019 due to food security issues [44].

On the other hand, there has been an increase in the number of crops lost in their entirety due to extreme environmental conditions. In 2005 and 2006, most producing countries suffered from conditions that greatly reduced agricultural production, decreasing for example cereal production by $2.1\,\%$ [45]. Not only do high temperatures affect crops, they can also accelerate the metabolism, consumption and growth rate of certain types of insects that infect plants of major food security concern. In addition, the increase in land temperature is considered a risk, as if it increases by an average of $2\,\%$, losses due to pest pressure could be $46\,\%$, $19\,\%$ and $31\,\%$ for wheat, rice and maize, respectively. It is worth mentioning that these crops provide on average $42\,\%$ of the calories that are consumed by humans globally [46].

The loss of crops due to climate change greatly affects the economy of producing countries. Millions of people around the world depend exclusively on agricultural production on their land for them and their family's livelihood, but due to global warming, they have been forced to change practices, crop periods and duration, or even the type of crop grown. Some farmers have even had to change the location of their crops, for example to higher regions, in order to mitigate the effects of climate change [47]. According to the UN, natural disasters have increased to an average of 205 per year, with floods and storms being the most frequent events. However, extreme droughts are the most damaging to crops, which can cause about $80\,\%$ of total damage and losses

in agriculture [45].

To mitigate the effects that climate change is causing in agriculture, the development of varieties or species that are more resistant to severe conditions is required. For the generation of these new varieties that adapt to global warming, a deep knowledge of the plants that are cultivated worldwide such as rice, maize, coffee, sugar cane, among others, is required. This knowledge guarantees higher crop productivity, greater resistance to pests, shorter times between sowing and harvesting, and better adaptation to changes in the environment.

Since the discovery of the structure of DNA, it has been shown that most cell functions are encoded and passed on from generation to generation. This information has caused a major revolution in the understanding of the biology of species and has been of great use in creating varieties that are better adapted to climatic changes. One of the key events driving this revolution has been the development and subsequent improvement of sequencing technologies, which have produced large amounts of DNA sequences of many species of interest for subsequent *in silico* study. Through bioinformatics techniques, it has been possible to process, analyse and subsequently understand many of the functions that are vital in organisms, through a pipeline that has become standard in recent years. The first step in this pipeline is sequencing, where through various methodologies a biological DNA or RNA sample is obtained, cut into multiple portions and duplicated (in certain technologies) to improve the quality, obtaining a computer file with the genomic information contained in the sample. Next, the assembly of the reads obtained in the previous step is carried out; this process attempts to group and re-form the original chromosome structures. Finally, all features of interest such as genes and their functions, coding and non-coding portions, promoter sequences, transposable elements, among others, are annotated. Although many plant species have now been fully sequenced and studied at the molecular level, many of the crops of interest, such as maize and sugar cane, still have many sections of their genomes that are unknown. These gaps are mainly generated by repetitive sequences such as TEs and simple repeats such as microsatellites, because most assemblers have problems with regions that are highly repetitive and highly variable [9]. Furthermore TEs can affect gene annotation [48], so it is recommended to identify and hide them before executing the annotation process.

Although the last five years have seen a great deal of effort in the development of bioinformatics applications using the techniques described above (de novo, structure-based, homology-based and using comparative genomics), novel techniques are required to improve the TE discovery process and thus improve understanding at the molecular level of plants of interest.

ML offers the advantage of optimising tasks using previous experience. Thanks to the large amount of data that has been generated in recent years, many researchers have applied ML techniques in various areas of genomics, particularly in the identification and classification of TEs. For example, Loureiro *et al.* [35] tested various ML algorithms such as neural networks, Bayesian net-

works, Random Forest and decision trees to improve identification or classification results using as input the outputs of various bioinformatics tools. Although this work showed very good results, the researchers used simulated data, which may be impractical if taken to a real application. Abrusan *et al.* [49] developed TEclass to classify transposable elements through support vector machines (SVM), but only down to the order level. Schietgat *et al.* [26] proposed the TE-Learning framework integrating traditional bioinformatics techniques for TE detection and random forest techniques for classification of LTR retrotransposon down to the superfamily level (leaving aside lineage classification), but did not implement this framework in software that interested researchers could use. In addition, the researchers used only the internal sections (which are coding and therefore the most conserved) for classification, which limits the framework to whole elements, leaving out a large portion of TEs that have mutated and removing important information from non-coding portions such as LTRs. DL has emerged as a branch of ML, where neural networks with multiple hidden layers are used to obtain patterns in the training data. Several literature reviews [50, 51, 52, 37, 53] have demonstrated the great utility that this type of algorithms can have on complex and large data such as DNA, however its application in the field of TEs is still very limited and in the available literature there are no algorithms and very few proposed architectures that integrate DL to overcome the problems inherent to TEs [41, 54, 55, 56]. Nevertheless, these architectures only perform classification down to the superfamily level and do not use plant-only data, which could lead to unreliable results on newly released plant genomes of interest.

Although complex genomes such as maize, sugar cane and coffee have already been sequenced, they still have large gaps, especially in regions with a high number of repeats (mainly TEs) such as centromeres, which makes it impossible to deeply understand these species that are considered of high impact in producing countries. In addition, although it has been shown that TEs are activated under certain external or internal stimuli, the exact dynamics of these elements and the effects they may cause on plants are still not known.

The design, implementation and validation using real TE data of an ML-based architecture that improves the identification and classification of TEs at superfamily and lineage levels is therefore required, with the aim of improving knowledge about their diversity, dynamics and impacts on plant genomes. This will lay the foundations for the subsequent improvement of crop varieties of global agronomic interest, such as rice, maize, sugar cane, wheat, barley or coffee, providing solutions to crop losses due to climate change.

1.3. Justification

Transposable elements have key roles in genome, chromosome organisation, in particular in sex chromosomes, participation in rearrangement events [57, 17] (e.g. translocations, fusions or cleavages), and contribution to genome size variations [8]. Also, TEs can have great influence on chromosome structure, especially centromeres, where certain lineage (centromeric retrotranspo-

sons - CRs) of LTR retrotransposons constitute essential components for centromere recognition by kinetochore proteins. CR retrotransposons have been found in centromeres of plants such as rice, coffee, species of the genus *Brachypodium*, wheat, maize, cereals and other grasses [16].

In addition, TEs can cause effects on host organism phenotypes due to interaction with gene activity [58]. These impacts may include imposing intragenomic selection pressures through their effects on gene expression [59], inactivation of gene coding or regulatory regions [8], mutations that change the protein sequence, variation of the expression pattern or alternative splicing [30], alteration of the expression of neighbouring genes by epigenetic effects [17] or through modification of the expression of transcription factors [60], redirection of stress stimuli to contiguous genes [61] and influence on the conservation, rearrangement and deletion of gene pairs [62]. The long-term impact of such variation involves, for example, genetic variation with important effects on species evolution [11]; variation in phenotypes of agricultural interest; genomic diversification and speciation [63]; and modification of organismal health [64] through the production of sense or antisense transcripts of genes [65]. Due to the impacts and importance described above, there is a need for reliable identification and classification of TEs in plants of agricultural interest. This is to understand in depth the mechanisms of adaptation to the environment, species evolution and intra-species variability. In addition, the correct annotation of TEs can improve the accuracy of coding region annotation and also facilitate functional genetic studies [66]. These advances could be based on the development of different identification and classification strategies through machine learning, deep learning and bioinformatics algorithms.

ML techniques such as support vector machines (SVM), Random Forest, Hidden Markov Models (HMM) and neural networks have been successful in the analysis of life science data due to their ability to handle the randomness and uncertainty of data noise and generalisation [67]. Some tools and frameworks have even been developed in recent years to detect and classify TEs. TE-Class [49], TE-Learner [26] and RED [68] are some examples of tools that apply ML (SVM, Random Forest and HMMs respectively) especially on repetitive items. However, these applications still have limitations such as: they only cover a single task (detection, classification or filtering), some are not easily installable and executable tools, others do not use heterogeneous architectures to accelerate software execution (they only use CPU and in some cases without parallel strategies to use multiple cores at the same time). Also, DL architectures have been applied on genomic data achieving better predictive performance over ML methods, including logistic regression, decision trees or Random Forest [50]. Although, the application of these techniques in TEs is still very limited and taking into account the high complexity of the data, its large size and divergence, DL techniques could increase the performance of bioinformatics algorithms.

In supervised problems, the feature extraction or feature selection process is a crucial step to improve the performance of the whole architecture. In ML, variable or feature selection processes must be carried out by a subject matter expert. Deep network architectures allow features

to be extracted in a non-linear and automatic way. The hidden layers of deep neural networks transform these features into complex patterns relevant to the classification problem [50]. In the specific case of TEs, being DNA sequences, feature extraction is often a too complex process due to the large amount of information, their unstructured form and their sequentiality. In this case, deep networks provide new features that are not possible to extract manually. For example, convolutional neural networks have the ability to discover local patterns in sequential data such as pixels in an image [52]. In DNA, these patterns are known as motifs and have important functions in the genome, such as gene promoters. Motifs will be very informative if they are found in the LTR sequences of retrotransposons, as they can be used to identify or classify TEs by taking their locations and frequencies. Although motifs are important for DNA classification problems, it is not enough to find the exact patterns, because DNA can undergo modifications or mutations and certain motifs may function the same as others even if they do not have exactly the same nucleotides.

This thesis allowed the application of novel ML techniques to TEs, as well as the evaluation of different metrics, data encoding forms, parameters, algorithms, databases and architectures, on diverse data with special characteristics. On the other hand, real data (found by research available in the literature on plant genomes and stored in databases such as PGSB [33]) were used to train the algorithms to obtain more reliable and generalisable results.

With the development of this thesis, reliable approaches for the identification and classification (also other task like library curation) of TEs were developed, which will contribute to researchers in areas such as biology, genomics and in general in the life sciences, with the aim of improving the understanding of the genomic structures of plants of agricultural interest. It will also contribute to the understanding of the dynamics of TEs, their relationships with gene activity and their roles within host organisms.

In addition, this doctoral research provided knowledge of the genomes of plants of agricultural interest such as rice, coffee, maize and sugar cane; and model organisms such as *Arabidopsis thaliana*, which will serve as input for genetic improvements that could be made in the future, which could generate plants that are more resistant to climate change, could be more resistant to pests and unfavourable conditions such as excess water or drought and could be more productive, speeding up harvesting time or increasing food production.

In the research both annual and perennial plants were used, with different genome sizes (from small sizes, 135 Mb like *A. thaliana* to plants with very large and complex genomes like maize with 2.3 Gb), different TE compositions and genomes at different annotation levels (*A. thaliana* and rice being the best quality). *Arabidopsis thaliana* is a model organism and was the first plant to be sequenced [69], is an annual plant and has an approximate genome size of 135 Mb, where $10\,\%$ corresponds to TEs. Rice (*Oryza sativa*) has a genome size of 466 Mb [70], also has an an-

nual cycle and $16\,\%$ of its entire genome corresponds to TEs. In addition, robusta coffee (*Coffea canephora*) is a perennial, diploid plant with a genome size of 710 Mb and approximately $50\,\%$ TE content [15]. Maize (*Zea mays*) is a globally important crop, was domesticated in Central America approximately 10,000 years ago and has undergone several genome duplications, reaching a genome size of 2.3 Gb with a TE composition of $85\,\%$ [13].

1.4. Research questions

Based on the above, the following research questions are formulated:

1. What are the metrics needed to measure the correct detection and classification of transposable elements from bioinformatics-based techniques and a computational architecture based on novel techniques such as Machine Learning?

2. What are the Machine Learning techniques that allow a computational architecture to obtain better results in a reasonable time in large-scale analysis in the identification and classification of transposable elements using various forms of DNA representation?

3. How do Machine Learning techniques contribute to the identification and classification of transposable elements in plant genomes?

4. What are the parameters (pre-processing techniques, activation functions, hyper-parameters of each algorithm and the representation of the input data) most appropriate for the design of a Machine Learning-based architecture for the identification and classification of transposable elements in plant genomes?

1.5. Research hypothesis

Therefore, the hypothesis of the following research proposal is the following:

Computational architectures based on machine learning techniques improve the identification and classification of transposable elements in plants that approximate an *in silico* solution for future genetic improvement of crop varieties of agricultural interest.

1.6. Organization of this Document

After this chapter, the thesis document follows the following order: **Chapter 2** contains the objectives of the thesis. **Chapter 3** contains the state of the art about retrotransposons, its diversity,

their impact over plants, and also about machine learning applied to TE tasks. **Chapter 4** elaborates on how to coding DNA sequences and how to measure ML algorithms trained with LTR retrotransposon data. **Chapter 5** shows a reference library of LTR retrotransposons from 195 plant species designed to be used in the training of ML algorithms, and also to be used in homology-based tools. **Chapter 6** demonstrated the utilization of *k*-mers as features in ML algorithms and also shows the feasibility of doing detection and classification of LTR retrotransposons as separated and also as integrated task. **Chapter 7** shows a proposed neural network to automatically curate libraries of LTR retrotransposons from plant genomes in efficient times. **Chapter 8** presents a one-shot tool that implements four neural networks to detect, classify and also annotated LTR retrotransposons in plant genomes as a united pipeline. **Chapter 9** demonstrated how ML based tools can be used to analyze large amount of data (like dozens of plant genomes) to answer a biological question in relative short time, and **Chapter 10** shows the discussions about all the thesis, the conclusions, and contributions derived from this thesis work. At a general level, **Chapters 1** to **3** describe the entire dissertation proposal, **Chapters 4** to **9** show the development and results of the dissertation and **Chapter 10** closes the dissertation with discussions, conclusions, and contributions at a general level.

2. Thesis Objectives

2.1. General Objective

The general objective of this doctoral research is to:

To develop a computational method based on Machine Learning techniques to identify and classify LTR retrotransposons in plant genomes of agro-industrial interest, to improve genomic and evolutionary understanding of the species.

2.2. Specific Objectives

1. To design and to build a scalable Machine Learning-based architecture for the study of LTR retrotransposons in plants.
2. To integrate and to implement the architecture in a Machine Learning-based bioinformatics software for the identification and classification of LTR-RTs.
3. To validate the architecture using genomes of plants of biotechnological and agro-industrial interest that have already been sequenced and released (coffee, rice, maize, and *A. thaliana*).

3. The State of the Art

This chapter is composed by two published articles. The first one was a narrative review submitted at 21 June 2019, accepted at 2 August 2019 and published at 6 August 2019 in the International Journal of Molecular Sciences. DOI: 10.3390/ijms20153837. The second was a systematic review submitted at 5 August 2019, accepted at 28 November 2019 and published at 18 December 2019 in the journal PeerJ. DOI: 10.7717/peerj.8311.

3.1. Context about retrotransposons and their characteristics

Research for the last 10 year have shown the great impact that transposable elements have inside their host genome in many organism, but specially in plants [71, 72, 73]. Thus, there are much literature demonstrating key functions of those repeated sequences in many aspects such as chromosomal structure and organization [74, 75], variations in phenotype [76], adaptation to environmental changes [24, 77], among others. Therefore, an exhaustive bibliography review was performed in the biology of transposable elements focusing in Class I or retrotransposons, because in plant genomes these types of elements are the most abundant.

This initial review allowed to synthesize the huge amount of information trying to answer the following questions:

1. What is the structure, diversity and function of retrotransposons in host genomes?
2. Why is it important to classify retrotransposons (into superfamilies and lineages)?
3. How to identify and classify retrotransposons?

The TE analysis of plant genomes brings challenges due to their complex dynamics, their huge contribution to the genome size and their specie-specific behaviour [78]. Also, thanks to the advance in sequencing technologies, the problem today is not how to get data, but how to process and analyse it in acceptable times [79].

International Journal of
Molecular Sciences

MDPI

Review

Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning

Simon Orozco-Arias [1,2], Gustavo Isaza [2] and Romain Guyot [3,4,*]

1 Department of Computer Science, Universidad Autónoma de Manizales, Manizales 170001, Colombia
2 Department of Systems and Informatics, Universidad de Caldas, Manizales 170001, Colombia
3 Department of Electronics and Automatization, Universidad Autónoma de Manizales, Manizales 170001, Colombia
4 Institut de Recherche pour le Développement, CIRAD, University Montpellier, 34000 Montpellier, France
* Correspondence: romain.guyot@ird.fr

Received: 21 June 2019; Accepted: 2 August 2019; Published: 6 August 2019

Abstract: Transposable elements (TEs) are genomic units able to move within the genome of virtually all organisms. Due to their natural repetitive numbers and their high structural diversity, the identification and classification of TEs remain a challenge in sequenced genomes. Although TEs were initially regarded as "junk DNA", it has been demonstrated that they play key roles in chromosome structures, gene expression, and regulation, as well as adaptation and evolution. A highly reliable annotation of these elements is, therefore, crucial to better understand genome functions and their evolution. To date, much bioinformatics software has been developed to address TE detection and classification processes, but many problematic aspects remain, such as the reliability, precision, and speed of the analyses. Machine learning and deep learning are algorithms that can make automatic predictions and decisions in a wide variety of scientific applications. They have been tested in bioinformatics and, more specifically for TEs, classification with encouraging results. In this review, we will discuss important aspects of TEs, such as their structure, importance in the evolution and architecture of the host, and their current classifications and nomenclatures. We will also address current methods and their limitations in identifying and classifying TEs.

Keywords: transposable elements; retrotransposons; function; structure; detection; classification; bioinformatics; machine learning; deep learning

1. Introduction

Transposable elements (TEs) are genomic units able to move within and among the genomes of virtually all organisms [1]. They are the main contributors to genomic diversity and genome size variation [2], with the exception of polyploidy events. An important issue in genome sequence analyses is to rapidly identify and reliably annotate TEs. There are major obstacles and challenges in the analysis of these elements [3], including their repetitive nature, structural polymorphism, species specificity, and, conversely, their conservation across genera and families, as well as their high divergence rate, even across close relative species [4].

Among eukaryotic genomes, TEs represent the most repetitive sequences [5]. They are able to move in the genomes, generate mutations, and obviously amplify the number of their copies [6]. Usually they are classified according to their coding regions involved in the replication of the element [7]. TEs moving via an RNA molecule called retrotransposons fall into Class I, while elements moving via a DNA molecule, called transposons, are classified into Class II [8]. They represent the vast majority of TEs found in plant genomes due to their mobility mechanisms. Retrotransposons can be further

3.2. Context about machine learning models in TEs

Bioinformatics methodologies to detect and classify TEs have known limitations [80, 81, 82]. Structure-based tools cannot find elements lacking some general features, so are less sensitive to novel structures. On the other hand, to use homology-based approaches, it is required to build a well curated library with elements from closely related species. This process is complex, usually required a lot of time and manual work. *De novo*, detects elements with a high number of copies, restricting the range of TE detectable. Finally, tools based on comparative genomics required assemblies of high quality, which is a difficult task specially with polyploid plants and with elevate number of repeated sequences.

Therefore, other approaches are needed in order to accelerate the analysis and annotation of those sequences in huge datasets, like plant genomes [83]. Machine learning models have been applied to bioinformatics [84, 85], and also in transposable elements [86] showing promising results. This computational approach learns from available data how to do a task automatically [38] and many researchers have taken advantage of this in genomics [52, 50]. Nevertheless, a method to transform nucleotide data (represented as letters) must be applied in order to use ML algorithms and it is not clear which could generate better results. Thus, a systematic literature review approach was applied to answer following questions:

1. What advantages ML approaches have compared to bioinformatics approaches for TE analyses?
2. Which ML techniques are currently used to detect and classify TEs or other genomic data?
3. What are the best parameters and most used metrics in algorithms and architectures to detect and classify TEs?
4. What are the most used DNA coding schemes in ML tasks?

A systematic review of the application of machine learning in the detection and classification of transposable elements

Simon Orozco-Arias[1,2], Gustavo Isaza[2], Romain Guyot[3,4] and Reinel Tabares-Soto[4]

[1] Department of Computer Science, Universidad Autónoma de Manizales, Manizales, Caldas, Colombia
[2] Department of Systems and Informatics, Universidad de Caldas, Manizales, Caldas, Colombia
[3] Institut de Recherche pour le Développement, CIRAD, University of Montpellier, Montpellier, France
[4] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales, Caldas, Colombia

ABSTRACT

Background: Transposable elements (TEs) constitute the most common repeated sequences in eukaryotic genomes. Recent studies demonstrated their deep impact on species diversity, adaptation to the environment and diseases. Although there are many conventional bioinformatics algorithms for detecting and classifying TEs, none have achieved reliable results on different types of TEs. Machine learning (ML) techniques can automatically extract hidden patterns and novel information from labeled or non-labeled data and have been applied to solving several scientific problems.

Methodology: We followed the Systematic Literature Review (SLR) process, applying the six stages of the review protocol from it, but added a previous stage, which aims to detect the need for a review. Then search equations were formulated and executed in several literature databases. Relevant publications were scanned and used to extract evidence to answer research questions.

Results: Several ML approaches have already been tested on other bioinformatics problems with promising results, yet there are few algorithms and architectures available in literature focused specifically on TEs, despite representing the majority of the nuclear DNA of many organisms. Only 35 articles were found and categorized as relevant in TE or related fields.

Conclusions: ML is a powerful tool that can be used to address many problems. Although ML techniques have been used widely in other biological tasks, their utilization in TE analyses is still limited. Following the SLR, it was possible to notice that the use of ML for TE analyses (detection and classification) is an open problem, and this new field of research is growing in interest.

Subjects Bioinformatics, Genomics, Data Mining and Machine Learning
Keywords Transposable elements, Retrotransposons, Detection, Classification, Bioinformatics, Machine learning, Deep learning

Submitted 5 August 2019
Accepted 28 November 2019
Published 18 December 2019

Corresponding author
Simon Orozco-Arias,
simon.orozco.arias@gmail.com

Academic editor
Kenta Nakai

Additional Information and Declarations can be found on page 24

DOI **10.7717/peerj.8311**

How to cite this article Orozco-Arias S, Isaza G, Guyot R, Tabares-Soto R. 2019. A systematic review of the application of machine learning in the detection and classification of transposable elements. **PeerJ 7:e8311 DOI 10.7717/peerj.8311**

3.3. Conclusions and perspectives

The two literature reviews provided relevant information on transposable elements in plant genomes such as how to classify them, their structure and their impact on host organisms. It was observed that there is a growing number of research studies focusing on aspects such as the relationship of TEs with chromosome structure, influence on genome size, variability between species and individuals, specific distribution by superfamilies and lineages in chromosomes, among other topics. This demonstrates a growing interest of the scientific community in these elements. In addition, it was found that advances in sequencing technologies have created a need for tools that can be executed in less time and produce more accurate results.

Although a large number of algorithms, approaches and tools exist to identify and classify TEs, accurate and reliable results cannot yet be obtained. This is because currently used strategies have limitations and therefore the use of novel approaches such as ML is required. However, literature reviews could not observe any tool that could detect, classify and annotate transposable elements through machine learning in a single tool.

Nevertheless, a increasing number of works related to TEs that are analysed with ML were observed. This provides evidence that machine learning-based methods are feasible and can overcome the limitations of current bioinformatics strategies. Since the publication of these reviews in 2019, some works were published focusing in the utilization of DL in DNA data and TEs. For example, a wrapper specialized in genomics of the well know DL framework keras, named by the authors as keras_dna was reported in 2021 [87], and a framework in Python to apply DL on genomics were also released [88]. Also, some new neural network architectures were proposed to classify TEs, such as TERL [54], and DeepTE [55], or to detect TE insertion boundaries like Frontier [89]. Other approaches based on ML algorithms (others than neural networks) were reported since 2019 too, such as ClassifyTE [90] and TransposomeUltimate [91].

The increase in interest and the number of papers on this subject is evident. However, the need has not yet been met. More research on fundamental components of the application of machine learning to the specific type of genomic data is lacking. These data present an additional challenge for ML algorithms due to their non-categorical nature. It is for this reason that the following chapter investigates how to transform this data and how to measure the performance of ML algorithms.

4. DNA coding schemes and measuring metrics

Article submitted at 25 April 2020, accepted at 22 May 2020 and published at 27 May 2020 in the MDPI Processes Journal. DOI: 10.3390/pr8060638.

4.1. Context

The genome of an organism can be represented by a computational file, where each letter corresponds to a nucleotide. This statement is crucial for bioinformatics, and it is thanks to this type of computational file that large-scale *in silico* analyses can be performed.

However, machine learning algorithms need numerical data to perform the training and parameter tuning processes. For this reason, a pre-processing step, feature extraction or conversion through coding schemes is necessary. In the literature review [92], different coding schemes and two strategies for feature extraction were found, but no information was available on which one might be the best for training ML algorithms.

To implement an algorithm based on machine learning, a number of fundamental steps must be followed [93]. One of the most important steps is to select the best way to measure the performance of the algorithms based on the general characteristics of the dataset and the problem to be solved by the algorithm. Although several applications of machine learning and specifically neural networks were observed in the literature, there was no study that evidenced whether the use of one metric or another affected the performance of the algorithms applied on transposable elements. In order to answer these two questions, the most commonly used metrics were compared and an experimental test was carried out using public databases.

processes

MDPI

Article

Measuring Performance Metrics of Machine Learning Algorithms for Detecting and Classifying Transposable Elements

Simon Orozco-Arias [1,2,*], Johan S. Piña [3], Reinel Tabares-Soto [4], Luis F. Castillo-Ossa [2], Romain Guyot [4,5] and Gustavo Isaza [2,*]

[1] Department of Computer Science, Universidad Autónoma de Manizales, Manizales 170001, Colombia
[2] Department of Systems and Informatics, Universidad de Caldas, Manizales 170004, Colombia; luis.castillo@ucaldas.edu.co
[3] Research Group in Software Engineering, Universidad Autónoma de Manizales, Manizales 170001, Colombia; johan.pinad@autonoma.edu.co
[4] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales 170001, Colombia; rtabares@autonoma.edu.co (R.T.-S.); romain.guyot@ird.fr (R.G.)
[5] Institut de Recherche pour le Développement, Univ. Montpellier, UMR DIADE, 34394 Montpellier, France
* Correspondence: simon.orozco.arias@gmail.com (S.O.-A.); gustavo.isaza@ucaldas.edu.co (G.I.)

Received: 25 April 2020; Accepted: 22 May 2020; Published: 27 May 2020

Abstract: Because of the promising results obtained by machine learning (ML) approaches in several fields, every day is more common, the utilization of ML to solve problems in bioinformatics. In genomics, a current issue is to detect and classify transposable elements (TEs) because of the tedious tasks involved in bioinformatics methods. Thus, ML was recently evaluated for TE datasets, demonstrating better results than bioinformatics applications. A crucial step for ML approaches is the selection of metrics that measure the realistic performance of algorithms. Each metric has specific characteristics and measures properties that may be different from the predicted results. Although the most commonly used way to compare measures is by using empirical analysis, a non-result-based methodology has been proposed, called measure invariance properties. These properties are calculated on the basis of whether a given measure changes its value under certain modifications in the confusion matrix, giving comparative parameters independent of the datasets. Measure invariance properties make metrics more or less informative, particularly on unbalanced, monomodal, or multimodal negative class datasets and for real or simulated datasets. Although several studies applied ML to detect and classify TEs, there are no works evaluating performance metrics in TE tasks. Here, we analyzed 26 different metrics utilized in binary, multiclass, and hierarchical classifications, through bibliographic sources, and their invariance properties. Then, we corroborated our findings utilizing freely available TE datasets and commonly used ML algorithms. Based on our analysis, the most suitable metrics for TE tasks must be stable, even using highly unbalanced datasets, multimodal negative class, and training datasets with errors or outliers. Based on these parameters, we conclude that the F1-score and the area under the precision-recall curve are the most informative metrics since they are calculated based on other metrics, providing insight into the development of an ML application.

Keywords: transposable elements; metrics; machine learning; deep learning; detection; classification

4.2. Conclusions and perspectives

In order to adequately measure the performance of ML algorithms, it was important to search the literature for metrics that are regularly used in problems involving transposable elements, and genomic data in general. In this way it became evident that due to the unique characteristics of TE datasets, the most informative metrics are those that are not affected by class imbalance. For this reason, it was shown that using a metric such as accuracy (very commonly used in ML problems) can lead to an over-realistic estimation because it does not take into consideration the performance of under-sampled classes (e.g. Ikeros, Ivana or Tekay lineages). Finally, it became evident that it is necessary to use more than one metric to demonstrate the performance of the model in different aspects, such as sensitivity, precision, false positive rate, among others.

Also, this work made it possible to establish how informative five coding schemes and two forms of feature extraction were. The aforementioned methodologies are found in the literature and have been used mostly for genomics problems. The *k*-mer frequencies proved to be the most informative form for all ML models tested and in both public databases used. It is worth mentioning that other studies [41, 55] had already used *k*-mer frequencies, but this work was the first to apply it to lineage-level classification of LTR retrotransposons and also to compare it with other coding schemes.

Thanks to the datasets that are released every day, it is now possible to think about using machine learning to automate the tasks of detecting and classifying LTR retrotransposons. However, each dataset has its own characteristics, such as different levels of curation of its sequences (from uncurated to manually curated) and different types of sequences (consensus sequences or individual genomic sequences). These large differences between datasets open a question on how to use them to train an ML algorithm and what would be the best way to unify this large amount of information in order to get the most out of the existing data. To solve this question, in the following chapter we statistically analyze the behavior of the algorithms when trained by the different public databases and some made from available genomes. In addition, a dataset designed to be used in the training process of ML algorithms is released.

5. InpactorDB: A reference library to train machine learning models

Article submitted at 30 December 2020, accepted at 22 January 2021 and published at 28 January 2021 in the MDPI genes Journal. DOI: 10.3390/genes12020190.

5.1. Context

After testing public databases in Chapter 4, it was found that the nature of the data affects the performance of ML models. For example, using a curated database composed of consensus will result in more representative samples and therefore better performance. However, these databases have far fewer sequences than non-curated ones. For this reason, three databases available in the literature with different nature of the data were used and supplemented with software based strategies to improve the reprensentability of species families.

Different ML classifiers were then used to test whether there was a significant difference between the databases and finally a new library was created, classified down to the lineage level and that it was the first one to be designed specifically to train ML algorithms.

After obtaining promising results in ML algorithms, it was decided to use this library to train two types of neural networks available in the literature. A fully connected network published by Nakano *et al* in 2018 [41] and a convolutional network published by Yan *et al* in 2020 [55]. These two architectures showed interesting results, even with species that had not been used to train them.

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

5.2. Conclusions and perspectives

This work established that databases composed of consensus sequences and databases composed of curated sequences are the best for training machine learning algorithms. However, they did not show significant differences between them. This analysis showed that it is possible to create databases with consensus sequences (which can be done automatically) that have comparable results to those that are curated (which require much more manual work). Thus, a library of LTR retrotransposons was designed and released, which is composed of different public databases, but adding sequences from more plant species from different families.

This database of more than $100,000$ sequences in the redundant form (independent sequences) and more than $67,000$ consensus sequences in non-redundant form demonstrated a good level of generalisation, as it contains sequences from 195 plant species.

On the other hand, it was shown that when training neural networks available in the literature using the non-redundant form of this library, even better results were obtained than when using other ML techniques, due to the fact that the training time was reduced when using GPUs and overfitting was reduced. This work opened the door for the design of DL-based tools thanks to the good quality of their data and the amount of sequences from different plant families that in deep networks is very useful for the training process.

Nevertheless, extracting *k*-mer frequencies (in this study it was proposed to use $1 <= k <= 6$) is computationally expensive and generates more than five thousand features which makes the training process of the algorithms slower. Therefore, it is necessary to understand whether it is necessary to use all this large amount of information or whether it is possible to reduce the *k*-mer frequencies without considerably reducing the performance of the algorithms. Therefore, in the next chapter we will delve more into how to design a machine learning based workflow that utilizes the features extracted from the *k*-mers count. Additionally, we experiment on training algorithms to perform the detection and classification tasks individually and jointly, as well as reporting information on how important each of the *k*-mer frequencies are and how to use as few as possible without lossing significant performance in the training process.

6. *K*-mer-based machine learning method to detect and classify LTR retrotransposons

Article submitted at 17 February 2021, accepted at 24 April 2021 and published at 19 May 2021 in the PeerJ Journal. DOI: 10.7717/peerj.11456.

6.1. Context

In Chapter 4, it was observed that using counts of the *k*-mers frequencies of LTR retrotransposons as features to train algorithms that classify these elements into lineages obtains promising and better results than other types of DNA encoding. Additionally, it was found in Chapter 5 that this form of feature extraction also obtains good generalization results (prediction on genomes of species that were not present in the training dataset) by neural networks published in the literature. However, the classification problem was only attacked at the lineage level, based on the assumption that the elements were already identified.

Using the library created in the previous section, this paper approaches three different classification problems for LTR retrotransposons. First, a binary classification problem is proposed, where LTR-RTs are differentiated from other genomic sequences (such as different types of RNAs and transposable elements of other orders). Then, the classification of these elements in the different lineages is further explored. Finally, a unified multiclass problem is presented that can both differentiate other types of genomic sequences and classify LTR retrotransposons into lineages. These problems were approached from the most widely used ML models, as well as using assembly classifiers.

K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes

Simon Orozco-Arias[1,2], Mariana S. Candamil-Cortés[1], Paula A. Jaimes[1], Johan S. Piña[1], Reinel Tabares-Soto[3], Romain Guyot[3,4] and Gustavo Isaza[2]

[1] Department of Computer Science, Universidad Autónoma de Manizales, Manizales, Caldas, Colombia
[2] Department of Systems and Informatics, Universidad de Caldas, Manizales, Caldas, Colombia
[3] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales, Caldas, Colombia
[4] Institut de Recherche pour le Développement, CIRAD, Univ. Montpellier, Montpellier, France

ABSTRACT

Every day more plant genomes are available in public databases and additional massive sequencing projects (i.e., that aim to sequence thousands of individuals) are formulated and released. Nevertheless, there are not enough automatic tools to analyze this large amount of genomic information. LTR retrotransposons are the most frequent repetitive sequences in plant genomes; however, their detection and classification are commonly performed using semi-automatic and time-consuming programs. Despite the availability of several bioinformatic tools that follow different approaches to detect and classify them, none of these tools can individually obtain accurate results. Here, we used Machine Learning algorithms based on *k*-mer counts to classify LTR retrotransposons from other genomic sequences and into lineages/families with an F1-Score of 95%, contributing to develop a free-alignment and automatic method to analyze these sequences.

Submitted 17 February 2021
Accepted 24 April 2021
Published 19 May 2021

Corresponding authors
Simon Orozco-Arias,
simon.orozco.arias@gmail.com
Gustavo Isaza,
gustavo.isaza@ucaldas.edu.co

Academic editor
Gerard Lazo

Additional Information and Declarations can be found on page 15

DOI **10.7717/peerj.11456**

Subjects Bioinformatics, Plant Science, Computational Science, Data Mining and Machine Learning, Data Science
Keywords Transposable elements, LTR retrotransposons, Plant genomes, Machine learning, Classification, Free-alignment approach, *k*-mer based method

INTRODUCTION

The availability of large-scale biological data is changing the way researchers must analyze and find solutions to problems in almost every area of biological sciences. Machine Learning (ML) algorithms can use this data to automatically learn the parameters needed to fit a model to a specific problem (*Shastry & Sanjay, 2020*) in order to predict known labels. This process is called supervised learning (*Zou et al., 2018*). Bioinformatics, which is an intersection between computer sciences, biological sciences, and mathematics (*Orozco-Arias et al., 2017*), plays a central role in storing, analyzing, categorizing, and labeling the huge flow of information generated, for example, by next-generation sequencing (NGS) platforms. Advances in these sequencing technologies have provided a new paradigm in the field of post-genomics (*Rigal & Mathieu, 2011*; *Chen et al., 2014*;

6.2. Conclusions and perspectives

This work confirmed that ML algorithms succeed in classifying transposable elements into subgroups as had been shown in other studies (usually in super-families) [41, 94, 27, 35, 26, 25, 43, 54, 55, 90] through features based on *k*-mers frequencies. Performances of up to $97\,\%$ F1-Score were obtained on the multiclass classification problem at the lineage level (*k*-nearest neighbors algorithm). Additionally, this work pioneered the problem of distinguishing LTR retrotransposons from other genomic sequences through an ML model. For this case, yields of up to $98\,\%$ were achieved (with the Multi-layer perceptron algorithm). On the other hand, the unification of both problems was successfully achieved, designing a dataset with more than 100 thousand sequences, where more than 34 thousand corresponded to sequences other than LTR-RTs, about 28 thousand to elements of the Copia superfamily and approximately 39 thousand to the Gypsy superfamily. Using this dataset, up to $96\,\%$ F1-Score was obtained using the ML assembly method. However, in Chapter 5 it is shown that by using an FNN publish by Nakano [41] in the lineage classification problem, an F1-Score of $98\,\%$ is obtained. Tests performed after the publication of this paper show that using the same FNN can give an F1-Score of $98\,\%$ in the binary classification problems and in the binary classification plus multiclass classification problem. In addition, the training times of this network were shorter and predictions were made faster. For this reason, in the following chapters we will focus on the use of deep neural networks.

Another interesting contribution of this article was the analysis of the importance of the different *k*-mer counts. Other studies had shown significant differences in *k*-mer features by modifying the value of *k* [55]. However, in this work it was found that with only $5.2\,\%$ of all the characteristics (286 of 5460) a $98\,\%$ F1-Score performance and an AUC of $97\,\%$ could be obtained.

These results present a promising alternative for the design and implementation of ML algorithms to attack problems related to LTR retrotransposons and even other DNA sequences such as filtering out sequences that do not meet certain criteria. In the next chapter, we will cover for the first time the problem of detecting whether an LTR-RT is intact (which could be used as a reference) or whether it has nested insertions, mutations or is considered a fragment, not suitable for a reference library through a DL approach.

7. Neural Network to curate LTR retrotransposons libraries

Article presented in the 15th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB) at 6 October 2021. Proceedings were published at 28 August 2021 in the Lecture Notes in Networks and Systems book series (LNNS). DOI: 10.1007/978-3-030-86258-9_9.

7.1. Context

The conventional process for annotating transposable elements consists of a two main steps workflow [83]. First, transposable elements are detected from the genome through different techniques [35] such as structure-based, de novo, homology or based on comparative genomics. Next, a curation process is performed where elements that are not considered intact and therefore do not fulfill the role of reference elements are removed. The second step consists of using this curated library to screening all the sequences of the copies and fragments of the TEs present in the library, usually using the homology strategy [95].

The curation process is of vital importance to achieve good quality TE masking and annotation because if element fragments (for example soloLTRs) are present in the library, which come from intact LTR-RTs that are also in the library, the annotation could show an over-estimation of the element contribution. Since it would take into account both the whole element and its fragments [96]. Another possible case would be to have an LTR-RT with another element inserted inside it and, in addition, to have the element nestedly inserted in the library. The annotation process would double count the nested element and overestimate its contribution to the genome. However, this task requires a lot of manual work and is very time consuming. For this reason, the alternative of training an ML algorithm to automate this task and thus speed up the analysis of large amounts of data was proposed.

Deep Neural Network to Curate LTR Retrotransposon Libraries from Plant Genomes

Simon Orozco-Arias[1,2( )], Mariana S. Candamil-Cortes[1], Paula A. Jaimes[1], Estiven Valencia-Castrillon[1], Reinel Tabares-Soto[3], Romain Guyot[3,4], and Gustavo Isaza[2]

[1] Department of Computer Science, Universidad Autónoma de Manizales, Manizales, Colombia
simon.orozco.arias@gmail.com, {mariana.candamilc,paula.jaimesb, estiven.valenciac}@autonoma.edu.co
[2] Department of Systems and Informatics, Universidad de Caldas, Manizales, Colombia
gustavo.isaza@ucaldas.edu.co
[3] Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales, Colombia
rtabares@autonoma.edu.co
[4] Institut de Recherche Pour Le Développement, CIRAD, Univ. Montpellier, Montpellier, France
romain.guyot@ird.fr

Abstract. Transposable elements are mobile sequences in all eukaryotic genomes. LTR (Long Terminal Repeat) retrotransposons are the most abundant elements in plant genomes where they play a fundamental role in evolution, gene function and genetic diversity. It is therefore important to develop bioinformatic tools to identify them in sequenced genomes and to classify them, taking into account that over time these elements may undergo deletions, insertions or recombination, generating incomplete and inactive elements, which are no longer considered a valid reference for identification and classification studies. LTR retrotransposons play fundamental roles in evolution and genetic diversity, hence the importance of understanding their function and studying in depth the variations that they may present. With the increase of whole genome sequencing, it is necessary to automate the analysis process and reduce the execution time, and to develop more advanced tools. Here, we propose an automatic curator of plant LTR retrotransposons libraries, based on Deep Learning (DL), in which a percentage F1-score of 91.18% was obtained for the test dataset. Generalization tests using four different genomes were performed, obtaining the best results for *Oryza granulata*, with a performance of 93.6% F1-score, and with an execution time of 22.61 seconds for the prediction by the neural network, using LTR retrotransposons obtained with the LTR_STRUC software. Taking into account that the conventional bioinformatics methods require a time of approximately six hours to curate the same genome, we conclude that our proposed method is efficient and can speed up the curation of libraries of LTR retrotransposons of plants genomes published in massive sequencing projects.

Keywords: LTR retrotransposons · Curation · Nesting insertions · Bioinformatics · Machine learning · Deep neural networks · *k*-mer-based methods

M. Rocha et al. (Eds.): PACBB 2021, LNNS 325, pp. 85–94, 2022.
https://doi.org/10.1007/978-3-030-86258-9_9

7.2. Conclusions and perspectives

This work showed another application of machine learning in tasks related to transposable elements. In this opportunity, it was observed that by using a dataset with semi-curated elements (such as those contained in the library presented in chapter 5) and elements that were found to have some kind of nested insertion, an ML algorithm can learn to classify them to improve the quality of the library.

On the other hand, it was shown that a performance of $91\,\%$ is achieved and that the prediction times are only a few seconds. However, the bottleneck of feature extraction still remains. Although Chapter 6 shows evidence that *k*-mer frequencies are a good source of information for ML classifiers, obtaining them has a high computational cost that in the case of automatic curation can take minutes, reducing the impact of this approach.

For this reason, for a tool to be able to use this filtering method, it must contemplate the use of some computational technique to speed up the counting of *k*-mer frequencies. Nevertheless, this article continues to open the paradigm of automating bioinformatics tasks that are currently performed manually, through artificial intelligence algorithms in order to create new software. In the following chapter, the complex task of integrating the different ML algorithms, activities of pre-processing and input data processing will be addressed to create a software that is easy to use and run and also requires considerably short time to analyze large amounts of genomic data.

8. Inpactor2: A one-shot software based on deep learning

Article submitted at 28 February 2022 in the Oxford Briefings in Bioinformatics Journal. Currently, it is under review.

8.1. Context

The work shown in the previous chapters has demonstrated the feasibility of applying a machine learning-based approach to perform tasks with transposable elements in plant genomes. However, all these tasks had been developed independently.

It had been seen in the literature [92] that very few works addressed both the task of identifying and classifying TEs through artificial intelligence. In addition, very few tools were found that could be easily installed and used by life science users [37].

In this work, the option of creating a one-shot tool that integrates all the necessary activities to detect, classify, filter and annotate LTR-RTs in plant genomes was raised. The goal was to create a tool that was easy to install and use, as well as to perform analysis in short times using heterogeneous architectures (CPUs + GPU).

Journal Title Here, 2022, pp. 1–10

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

Inpactor2: A software based on deep learning to identify and classify LTR-retrotransposons in plant genomes

Simon Orozco-Arias,[1,2,*] Luis Humberto Lopez-Murillo,[1] Mariana S. Candamil-Cortés,[1] Maradey Arias,[1] Paula A. Jaimes,[1] Alexandre Rossi Paschoal,[3,*] Reinel Tabares-Soto,[4] Gustavo Isaza[2,*] and Romain Guyot[4,5,*]

[1]Department of Computer Science, Universidad Autónoma de Manizales, 170001, Caldas, Colombia, [2]Department of Systems and Informatics, Universidad de Caldas, 170004, Caldas, Colombia, [3]Department of Computer Science, Federal University of Technology (UTFPR) - Paraná, 80230-901, Paraná, Brazil, [4]Department of Electronics and Automation, Universidad Autónoma de Manizales, 170001, Caldas, Colombia and [5]Institut de Recherche pour le Développement, CIRAD, Univ. Montpellier, 34000, Montpellier, France
*To whom correspondence should be addressed. Email: simon.orozco.arias@gmail.com, paschoal@utfpr.edu.br, gustavo.isaza@ucaldas.edu.co, romain.guyot@ird.fr
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

LTR-retrotransposons are the most abundant repeat sequences in plant genomes and play an important role in evolution and biodiversity. Their characterization is of great importance to understand their dynamics. However, the identification and classification of these elements remains a challenge today. Moreover, current software can be relatively slow (from hours to days), sometimes involve a lot of manual work and do not reach satisfactory levels in terms of precision and sensitivity. Here we present Inpactor2, an accurate and fast application that creates LTR-retrotransposon reference libraries in a very short time. Inpactor2 takes an assembled genome as input and follows a hybrid approach (deep learning and structure-based) to detect elements, filter partial sequences and finally classify intact sequences into superfamilies and, as very few tools do, into lineages. This tool takes advantage of multi-core and GPU architectures to decrease execution times. Using the rice genome, Inpactor2 showed a run time of five minutes (faster than other tools) and has the best accuracy and F1-Score of the tools tested here, also having the second best accuracy and specificity only surpassed by EDTA, but achieving 28% higher sensitivity. For large genomes, Inpactor2 is up to seven times faster than other available bioinformatics tools.

Key words: Inpactor2, LTR retrotransposons, plant genomes, deep learning, neural networks, detection, classification

Key Messages

- The hybrid approach used by Inpactor2 allows the creation of quality LTR-retrotransposon libraries, maintaining a high level of precision, accuracy, and sensibility and keeping a low number of false positives.
- Inpactor2 can be run using CPUs + GPUs, speeding up the execution time up to 7 times, being the fastest software in the creation of libraries of the tested software. This allows to analyze more genomes in less time, being useful for large scale analysis.
- Inpactor2 is the first freely available tool that integrates in a single software the detection, curation, classification and annotation process that provides lineage-level classification in an easy-to-install and use manner, eliminating the need for manual operations.
- Inpactor2 is the first neural network-based tool to detect LTR-retrotransposons de novo. It can be installed in an anaconda environment and can be run in a single Python command.

8.2. Conclusions and perspectives

The strategy implemented in Inpactor2 proved to be effective in integrating the different tasks required for the annotation of LTR-RTs in plant genomes. Although in Chapter 6 it was shown that the unification of the detection and classification problem could obtain good results, in an application with real data (such as an assembly) it was found to be more complex than having a network that predicts the class of a sequence.

The main difficulty of this integration was due to the large variability in length of the sequences of the different families of LTR retrotransposons. These elements can measure from two thousand bases to more than 20 thousand bases. For this reason, the use of non-overlapping windows of 50 thousand bases was considered as input. Finally, three neural networks trained with thousands of sequences were used to obtain more accurate results in each of the necessary activities such as detection, filtering, and classification of the elements.

This work also covered the bottleneck in *k*-mers frequency counting raised earlier in Chapter 5. Inpactor2 proposes a strategy based on a convolutional neural network with untrainable layers, to simultaneously count the 5460 *k*-mers frequencies necessary for the execution of the neural networks. This calculation is done in seconds, removing the limitation of using all of the *k*-mers frequencies.

Finally, thanks to the use of Python libraries such as Keras and tensorflow, this tool can be executed both on CPU and CPU + GPU, to accelerate the analysis times, being up to seven times faster than other reference software (with the two Gb genome of *Zea mays*). This tool proves to be effective in running complete analyses of LTR-RTs in minutes, which is especially important in large-scale genomic analyses, as a study at the level of a plant genus. In the next chapter, the high speed of Inpactor2 was used to analyze the correlation between the proportion of LTR-RTs and genome size variability in *Coffea* genus species.

9. Application of a DL-based tool to the identification and classification of LTR retrotransposons in the genus *Coffea*

9.1. Abstract

Transposable elements are DNA sequences that can move from one chromosomal position to another and are widely found in virtually all organisms. These elements are activated under certain biotic and abiotic stresses and play important roles within the genome. In plants, LTR retrotransposons are the most abundant and are associated with evolutionary mechanisms and, in particular, with variability in genome size. New applications based on machine learning allow large-scale analysis of these elements in search of answers to evolutionary questions in the area of, for example, a species order. After developing Inpactor2 and comparing it against other software, the possibility of using this software to answer a biological question involving the analysis of a large amount of data was raised. In this study, 46 species of the genus *Coffea* were used to understand the relationship of LTR retrotransposons in the genome size variation shown in this group. Our results demonstrate a high correlation between the Gypsy superfamily and especially the Tekay/Del lineage with genome size within the genus.

9.2. Introduction

Transposable elements (TEs) are short segments of DNA that are found in abundance within plant genomes [97]. These elements can move from one chromosome location to another, causing various rearrangements such as translocations, inversions and duplications [98]. TEs can produce genetic variations that, in the plants evolutionary field, are related with the contribution to the development of skills for the adaptation to their environment [99], thus it has been found that the TEs activation is generally given by stress conditions [100] dispersing through the genome the regulatory sequences found there, managing to increase the network of genes related to stress [101]. Thus, although the number of coding genes is not highly variable in plants [102], genome size does vary drastically, even in species within orders or families. These variations are mainly due to the repetitive regions and especially to TEs [103], due to the ability of some species to control the activity of these elements and to the mechanisms of them to escape regulation in some

other organisms [104, 105].

According to the transposition mechanism, TEs can be classified into Class I or retrotransposons", which transposed via a RNA intermediary, through a process called çopy-and-paste", generating a decrease in the number of copies of TEs; and Class II or "transposons", that moves directly using a DNA intermediary through the çut-and-paste"strategy [12]. LTR retrotransposons (LTR-RTs), belonging to Class I, are the most common in plants [81], contributing up to $80\,\%$ of the genome size [106], together with polyploidization, which is the principal mechanism for the increase in genome size and evolution in plants [107]. The LTR-RTs are flanked by two long terminal repeats (LTRs) at the ends. Between these regions are the structural coding domains and the enzymatic proteins, gag and pol, which are crucial in the transposition process [61, 106]. These TEs present a similar structure to retroviruses, nevertheless, these lack funcional env gene, which is responsible for the formation of infectious particles that can leave the cell and infect other cells [108]. The LTR-RTs are divided into two superfamilies, Ty1-Copia and Ty3-Gypsy, depending on the order of their internal domains [12, 61, 106].

Due to the rise of sequencing data, some methods have been described for the identification and classification of these elements [108, 35, 26], such as those based in structure, homology, de novo and by comparative genomics. However, the identification of these elements presents some difficulties since TEs do not have a universal structure, some families present a specific composition, and some of them acquire mutations over time, which generates fragmented or nested copies [26]. Recent studies have shown that Machine Learning (ML) can be used to propose automatic tools for the detection and classification of TEs [37], based on model training using TEs detected by conventional software [26]. Transfer learning, one of fields of ML, is a process in which the model is first trained with an initial dataset and later, the learned characteristics are transferred to another model to perform the training with the dataset of interest, increasing generalizability [109]. Nonetheless, research on the use of ML for the identification and classification of transposable elements remains scarce.

The genus *Coffea* belongs to the Rubiaceae family, the fourth largest family of Angiosperm. It comprises 124 identified species originating from tropical Africa, Madagascar, Mascarene islands extending to Southern and Southeast Asia and Australasia [110, 111]. *Coffea canephora* and *Coffea arabica* are two important cultivated species to its socio-economic impact in the tropical regions of the world [112]. All species in this genus are diploid, with the notable exception of *C. arabica* (allotetraploid), due to a recent hybridization between *C. eugenoides* and *C. canephora* [113]. Since the publication of the first sequenced genome of *Coffea*, *C. canephora*, in 2014 [15] and thanks to the Next Generation Sequencing (NGS) other sequencing data has been published such as genotyping-by-sequencing data with which gave the first resolved phylogeny of the genus *Coffea* [114], the identification of coffee species that are naturally decaffeinated such as *Coffea humblotiana* [115] and the construction of the evolutionary history of 52 wild coffee species

[116]. Following these sequencing data, it has become possible to identify more complex genomic structures such as transposable elements and their impact on various characteristics of the genus Coffea, as seen in [111] from the partial sequencing (454 technology) of 16 coffee genomes. In that study, it was possible to identify the transposable element composition and its variation in relation with the biogeographic groups of the species [111]. Now, the emergence of new bioinformatic tools and the availability of *Coffea* species sequences allows us to answer new questions about the diversity of the TEs in the genus and their contribution to the genome size variation [115, 117, 118]. In this study, we investigated the relation between the diversity and quantity of LTR-RT elements, their lineage classification and genome sizes in the frame of the phylogenetic relationships of the *Coffea* species and frame of the following phylogeographic groups [116]: West and Central Africa (WCA), North-Eastern Africa (NEA), Asia (ASIA), East Central East Africa (E-CEA), East Africa (EA), Mauritius (MUS) and Madagascar & Comoro (MDG-COM). We used data from analysis of 46 *Coffea* species showing a genome size ranging from 469 Mb for *C. mauritiana Lam* to 899 Mb for *C. humilis*. The LTR-RT identification and classification in the level of the lineages were performed using a DL-based tool named Inpactor2, and finally statistical tests were carried out.

9.3. Materials and methods

9.3.1. *Coffea* sequencing resources available

Illumina read datasets from 41 species of the *Coffea* genus were used in this study (Table 9.3.1) and 10 from species of the *Psilanthus* genus (Table 9.3.1). Additionally, seven coffee species (*C. canephora* Pierre ex A.Froehner - DH200-94, *C. eugenioides* S.Moore - BUA, *C. heterocalyx* Stoff., *C. homollei* J.-F.Leroy. *C. humblotiana* Baill., *C. pseudozanguebariae* Bridson, and *P. ebracteolatus* Hiern) and one of *Rubiaceae* family (*Kraussia floribunda Harv.*) were sequenced with PacBio, and those assemblies were used to construct the LTR-RT library (see section 9.3.3). The illumina reads were assembled using MaSuRCA [119], genomic size information was obtained via flow cytometry, and completeness was calculated from BUSCO [120]. Complete information can be consulted in Appendix A.

9.3.2. Creation of coffee dataset for re-training Inpactor2

Once Inpactor2 was trained using the InpactorDB dataset [34], we proceeded to do a transfer learning process with the aim of creating a specialized version of the tool for *Coffea* data. For this reason, we selected the most representative species: *Coffea arabica*, *Coffea eugenioides*, *Coffea homollei*, *Coffea pseudozanguebariae* and *Coffea ebracteolata* (ex *Psilanthus ebracteolatus*) because of the availability of near complete genome assembly. Later, we identified LTR-RT elements using LTR_STRUC [121], and classified them at the lineage level using Inpactor (V.1) [78]. For the dataset of negative instances, we performed the same process of [122]: we collected other genomic struc-

Species name	Country of origin	Genome size (Mbp)	N50 Illumina Assembly (bp)	Complete BUSCO %
	Coffea			
C. arabica	Ethiopia	1264.1	12530	92.6
C. boiviniana (Baill.) Drake	Madagascar	489	4437	62.6
C. brevipes Hiern.	Cameroon	743.28	4925	71.7
C. canephora Pierre ex A.Froehner	Democratic Republic of Congo	762.84	14559	92
C. canephora Pierre ex A.Froehner	Uganda	762.84	10617	90.3
C. canephora Pierre ex A.Froehner	Ivory Coast	762.84	9776	86.3
C. canephora Pierre ex A.Froehner	Ivory Coast	762.84	10248	71.6
C. canephora Pierre ex A.Froehner	Brazil	762.84	15981	93.3
C. charrieriana	Cameroon	699	25231	94.8
C. congensis A.Froehner	NA	753.06	14144	78.6
C. congensis A.Froehner	Republique of Congo	753.06	7192	85.1
C. dewevrei De Wild. & T.Durand	Central African Republic	704.16	20547	93.9
C. dolichophylla J.-F.Leroy	Madagascar	669	7888	87.3
C. eugenioides S.Moore	Kenya	723.72	10099	78.4
C. eugenioides S.Moore	Uganda	723.72	14110	91.3
C. heterocalyx Stoff.	Cameroon	889.98	15722	93.1
C. homollei J.-F.Leroy	Madagascar	596.58	14554	84.6
C. homollei J.-F.Leroy	Madagascar	596.58	4613	69.7
C. humblotiana Baill.	Comoros	479.22	19876	92.9
C. humilis A.Chev.	Ivory Coast	899.76	6398	88.9
C. kapakata (A.Chev.) Bridson	Angola	645.48	4127	26.8
C. liberica W.Bull. ex Hiern	Ivory Coast	743.28	5378	37.8
C. macrocarpa A.Rich.	Mauritius	577.02	20136	93
C. mauritiana Lam	Mauritius	469.44	34342	94.2
C. mayombensis A.Chev.	Cameroon	ND	2969	62
C. mufindiensis Hutch. ex Bridson	Tanzania	ND	6784	84
C. myrtifolia (A.Rich. ex DC.) J.-F.Leroy	Mauritius	528.12	7020	75.5
C. myrtifolia (A.Rich. ex DC.) J.-F.Leroy	Mauritius	528.12	99851	96.7
C. sp. â€˜nkolbisoniiâ€™	Cameroon	ND	4482	76.5
C. perrieri Drake ex Jum. & H.Perrier	Madagascar	625.92	9180	87.8
C. pervilleana Drake	Madagascar	547.68	4246	75.6
C. pseudozanguebariae Bridson	Kenya	557.46	15425	89.2
C. racemosa Lour.	Mozambique	508.56	16275	91.3
C. rhamnifolia (Chiov.) Bridson	Somalia	ND	42001	94.9
C. salvatrix Swynn. & Philipson	ND	596.58	22411	80.5
C. sessiliflora Bridson	Tanzania	537.9	32437	89.8
C. sessiliflora Bridson	Tanzania	537.9	4119	76.2
C. sp 3	Cameroon	ND	8072	87.9
C. sp. Congo	Congo	665	6010	88.3
C. stenophylla G.Don.	Ivory Coast	625	13676	92
C. tetragona Jum. & H.Perrier	Madagascar	528	16591	92.3

Table 9-1.: *Coffea* species used in this study. ND means data is not available. The complete information can be found in Appendix A

Species name	Country of origin	Genome size (Mbp)	N50 Illumina Assembly (bp)	Complete BUSCO %
		ex. Psilanthus		
P. benghalensis var. bababudanii (Sivar., Biju & P.Mathew) A.P.Davis	India	709	17218	91.4
P. benghalensis (Heyne ex J.A.Schult.) J.-F. Leroy	India	709	15991	91.6
P. brassii (J.-F.Leroy) A.P.Davis	Australia	ND	1173	6.4
P. ebracteolatus Hiern	Ivory Coast	586	10655	90.6 %
P. horsfieldianus (Miq.) J.-F.Leroy	Indonesia	ND	9059	77.6
P. leroyi Bridson	Sudan	ND	45234	92.3
P. mannii Hook.f.	Cameroon	ND	7320	89.2
P. melanocarpus (Welw. ex Hiern) J.-F.Leroy	Angola	ND	6687	88.5
P. travancorensis (Wight & Arn.) J.-F.Leroy	India	636	3682	54.1
P. wightianus (Wall. ex Wight & Arn.) J.-F.Leroy	India	631	40952	95.2

Table 9-2.: *Psilanthus* Species used in this study. ND means data is not available. The complete information can be found in Appendix A

tures such as CDS, RNAs and other types of TEs already mentioned. Then, the Inpactor2_Detect and Inpactor2_Class networks were retrained using both positive (LTR-RT elements) and negative (other genomic features than LTR-RTs) instances. The selection of the best models of both networks was performed following the loss and F1-score for each epoch, selecting the one that presented less loss and higher F1-score. Finally, the best retrained models of both NNs were kept and used for the execution of Inpactor2 in the rest of the analysis of this study.

9.3.3. Library of LTR-RTs in *Coffea* genus and its annotation

For the construction of the *Coffea* library, we used eight available assemblies of the species *C. canephora* Pierre ex A.Froehner, *C. eugenioides* S.Moore, C. *heterocalyx* Stoff., *C. homollei* J.-F.Leroy, *C. humblotiana* Baill., *C. pseudozanguebariae* Bridson, *C. ebracteolata* Hiern and *Kraussia floribunda* Harv (*Rubiaceae* species, outside the *coffea* genus), since these came from sequencing reads using the PacBio technology. Then, Inpactor2 (using the retrained models for Inpactor2_Detect and Inpactor2_Class) was executed to detect and classify until the lineage level the LTR-RTs. Finally, all the libraries created by Inpactor2 of each species were concatenated to create a unique library used to annotate genomic sequences using RepeatMasker [123].

In this study, we selected 46 assemblies of the species of *Coffea* (including ex *Psilanthus*) genus (Table 9.3.1), belonging to seven geographic groups as follow [116]: West and Central Africa (WCA), North-Eastern Africa (NEA), Asia (ASIA), East Central East Africa (E-CEA), East Africa (EA), Mauritius (MUS) and Madagascar & Comoro (MDG-COM). The annotation of the LTR-RTs was carried out with RepeatMasker, using the -lib parameter to append previously created libraries as a reference. Also, to obtain a complete summary of the annotation for each LTR-RTs lineages, we used the script "buildSummary.pl"which is part of RepeatMasker.

9.3.4. Data analysis and visualization

To test whether there is a relationship between the size of species genomes and the proportion of annotated LTR-RTs, R [124] is used to perform a correlation analysis. In the first instance, a pairwise plot of the proportion of LTR-RTs and assembled size is performed. Subsequently, a model fitting is performed to obtain more information about the relationship of the variables with respect to the genomic size. Finally, we checked the assumptions for a regression model, among them: residual analysis, multicollinearity and correlation.

In order to visualize the results more interactively, iTOL [125] was used to annotate the phylogenetic tree of the 46 species of the genus *Coffea*, following the tree constructed in [114]. In the same way, through leafletR library (http://cran.r-project.org/package=leafletR), a geographical map was constructed in order to locate the species in their respective biogeographical groups and to detail other physiological aspects of these, such as the approximate genome size taken from flow cytometry, caffeine ratio, as well as the ratio of LTR-RTs Copia and Gypsy and of some important lineages such as Athila, TAT and Tekay/DEL.

9.3.5. Raw Illumina reads mapping results

Raw Illumina reads were mapped against their respective library constructed in section 9.3.3 to estimate their redundancy and compare it with the estimate from assemblies. We used Bowtie2 [126] following the following parameters: bowtie2 –time –local –very-sensitive-local, and taking into account only forward reads.

9.4. Results

9.4.1. Re-training of the model for the *Coffea* genus

During the re-training process, five species of the aforementioned *Coffea* genus were used, which were selected due to their diversity in terms of their biogeographic group of origin. A quick test was performed to observe the performance of the model with the species *Coffea humblotiana* (see Table 9.4.1), and adequate accuracy was observed after retraining.

9.4.2. Construction of a LTR-RT library for the *Coffea* genus

A library of $34,063$ sequences of intact LTR-RTs extracted from eight high quality genomes including seven species from *Coffea* and a closely related species from the *Kraussia* genus was obtained (see Table 9.4.2), of which $9,205$ fall into Copia and $24,858$ into Gypsy superfamilies. Figure **9-1** shows the copy numbers for each of the lineages belonging to the mentioned superfamilies.

Initial model		
No. of errors	***Error percentage***	***Precision percentage***
145	11.2928 %	88.7072 %
Retrained model		
No. of errors	**Error percentage**	**Precision percentage**
65	5,0663 %	94,9337 %

Table 9-3.: Performance test results for initial and re-trained model

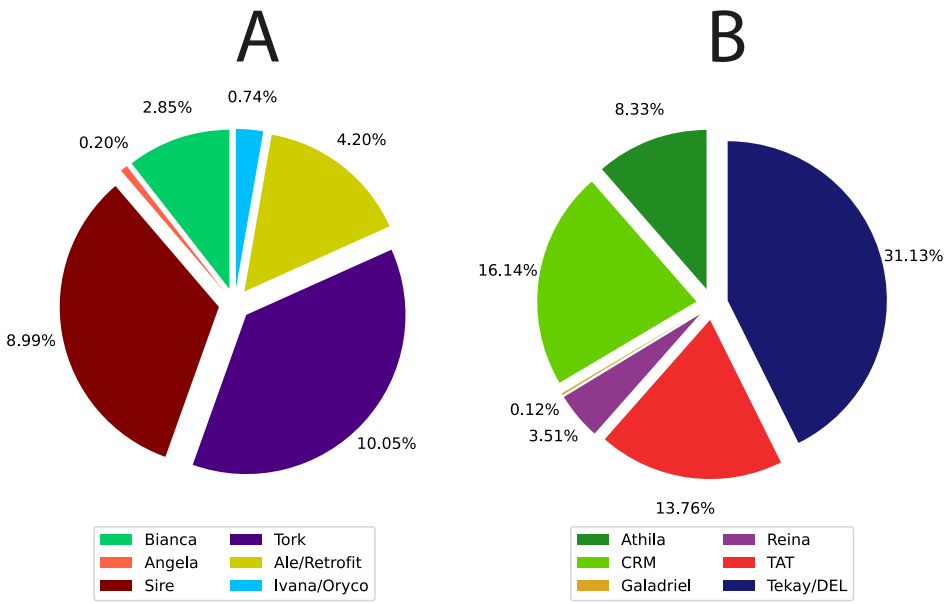

Figure 9-1.: *LTR retrotransposon length distribution among the library. A) Elements that correspond to Copia superfamily, while B) belonging to Gypsy superfamily.*

Species name	PacBio Assembly size (Mbp)	N50 (Mbp)	Total LTR-RTs
C. canephora Pierre ex A.Froehner - DH200-94	672.38	50.12	6,155
C. eugenioides S.Moore - BUA	645.42	54.74	4,731
C. heterocalyx Stoff.	760.30	5.25	5,933
C. homollei J.-F.Leroy	585.00	41.51	4,337
C. humblotiana Baill.	420.72	29.63	1,932
C. pseudozanguebariae Bridson	618.15	41.95	4,426
P. ebracteolatus Hiern	786.80	1.93	1,090
Kraussia floribunda Harv.	212.98	2.69	5,459

Table 9-4.: Species used for the construction of the Coffea library

9.4.3. Utilization of a *Coffea* LTR-RT library for the annotation of assemblies in the *Coffea* genus

Forty-six *Coffea* assembled genomes from different biogeographic groups (see Table 9.3.1 and Appendix A) were used for the annotation of LTR-RT by RepeatMasker. Using the previously created library as a reference, we obtained $23,730,563$ LTR-RTs, classified into superfamilies and lineages (Figure **9-2**). Figure **9-3** shows the nuclear phylogenetic tree corresponding to the species used in this study obtained by [114], and a comparison between the proportion of lineages belonging to Copia and Gypsy found with RepeatMasker in assemblies, according to the assembly size. In addition, the proportion of the superfamilies and lineages are also obtained using a mapping strategy of the raw reads against the LTR-RT library of the species. Our result showed Gypsy and more particularly Del and TAT lineages to be widespread and the most predominant LTR-RT elements in *Coffea*, whatever the genome size and the method to count their proportions. Beside these elements, CRM (Gypsy) and Tork and Sire (Copia) showed significant proportions in the assemblies. It should be noted that the two highest genome sizes present in this study (*C. heterocalyx* and *C. humilis*) showed high abundance of Del and TAT lineages compared to small genome sizes of *Coffea*, suggesting that these lineages might be involved in the genome size of these species. *C. heterocalyx* (890 Mb) is closely related to *C. eugenioides*, a medium genome size species (723 Mb). The difference between these species is clearly on the proportion of Del and TAT lineages (highest for *C. heterocalyx*), suggesting rapid evolutionary changes such as dramatic accumulation of LTR-RT copies. The same observation can be done between *C. humilis* (900 Mb) and *C. dewevrei* (704 Mb). Interestingly the distributions of SIRE appear discontinuous according to phylogeographic groups, with a presence in WCA, ECA groups, low proportion in EA groups and almost absent in Madagascar (MDG-COM) and mascarene (MUS) groups (Figure **9-3**). Altogether, our annotation of LTR-RT and their proportion of diverse *Coffea* species, suggest that some lineages might be correlated with the variation of the genome sizes.

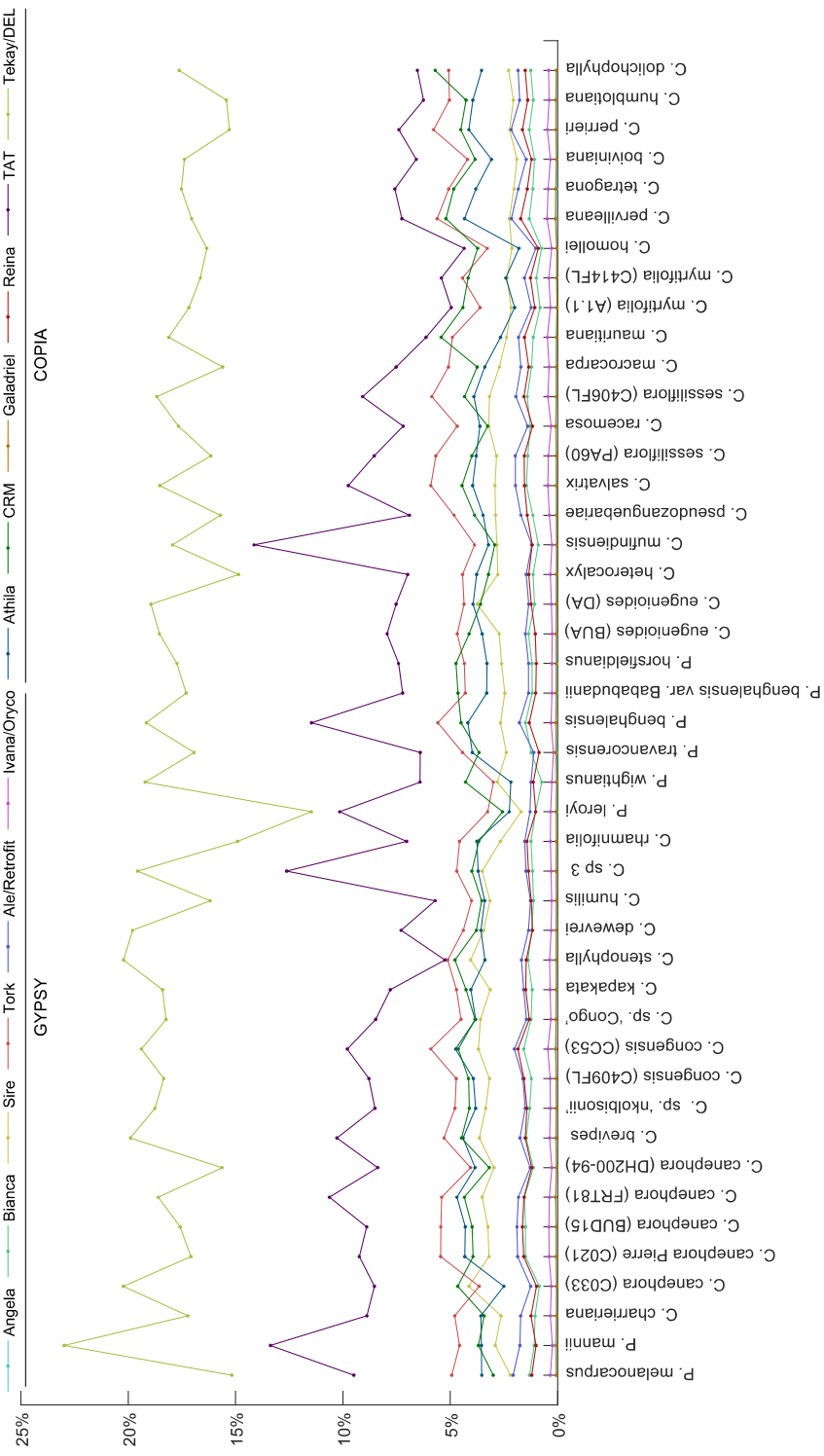

Figure 9-2.: Proportion of LTR-RTs at the lineage level for the 46 *Coffea* species according to their biogeographic groups of origin.

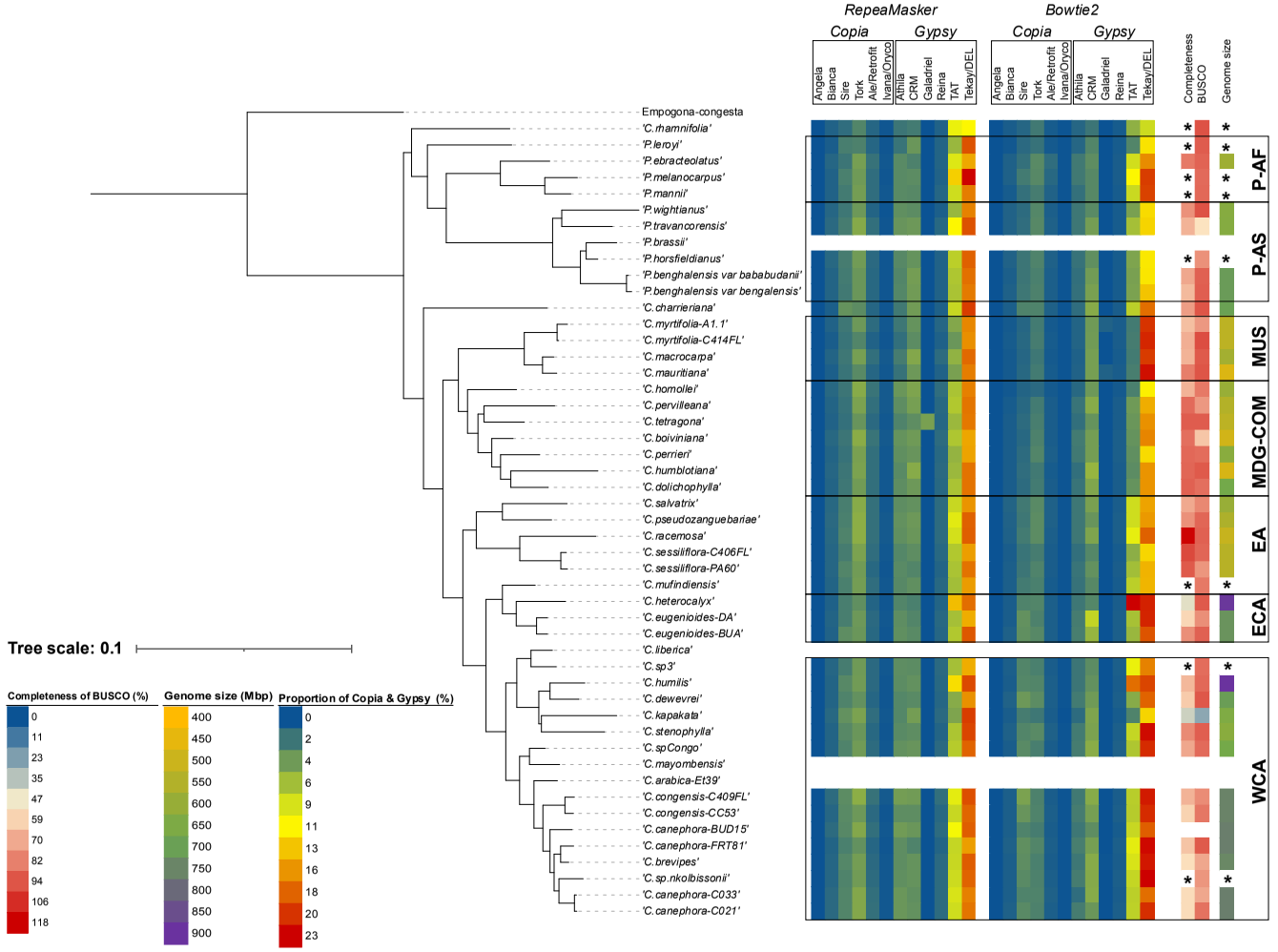

Figure 9-3.: Proportions of LTR-RT superfamilies and lineages of 46 *Coffea* species organized according to the nuclear phylogeny proposed by [114]. LTR-RT proportions are represented as heatmaps. An estimation of the completeness of the assembly is indicated when the genome size is available as well as the BUSCO analysis. Stars indicate missing values in the study.

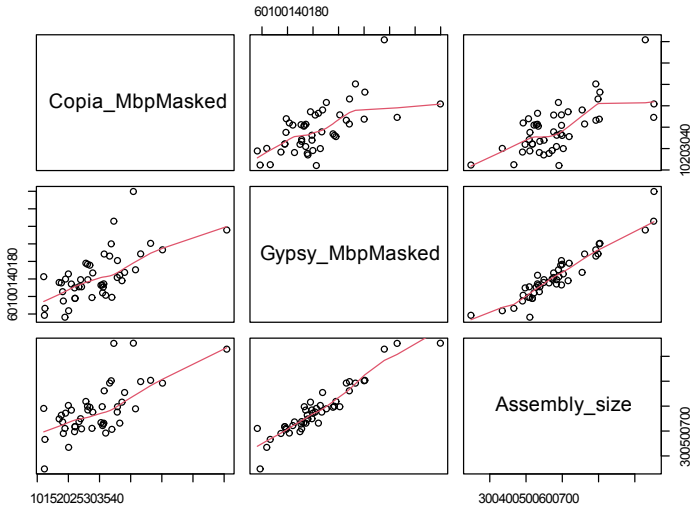

Figure 9-4.: Pairwise plot of the proportion (in percentages) of LTR-RTs of the Copia and Gypsy superfamilies and the genome assembly size (in Mbp). This graph is used to visualize the trend of the variables.

9.4.4. Relationship between the LTR-RT proportion and the genome size assembly

Relationships between Copia and Gypsy superfamilies and the assembly size.

To understand the relationships between the amount of LTR-RT and the size of the genome assemblies, we conducted different statistical analyses. Firstly, a pairwise plot was performed to observe the behavior of the proportions of LTR-RTs of Copia and Gypsy and the size of the genome assembly (Figure **9-4**). Finally, based on the plots, we conducted a multiple linear regression model using the 46 assemblies (Table 9.3.1), obtaining an Adjusted R square of approximately $90\,\%$ (Appendix B).

By checking the assumptions of the proposed regression model, it was found a variance inflationary factor (VIF) for both covariates (proportions of Copia and Gypsy) of 1.74187 and 1.74187 respectively, indicating that there is no problem of multicollinearity. On the other hand, Figure **9-5** shows diagnostics plots for the analysis of residuals and normality of the model (for more information see Appendix B), observing the fulfillment of the assumptions generated for this model in the residuals, being normality and homoscedasticity, thus concluding that the contribution of the proportion of LTR-RTs from both superfamilies is significant in the size of the assemblies.

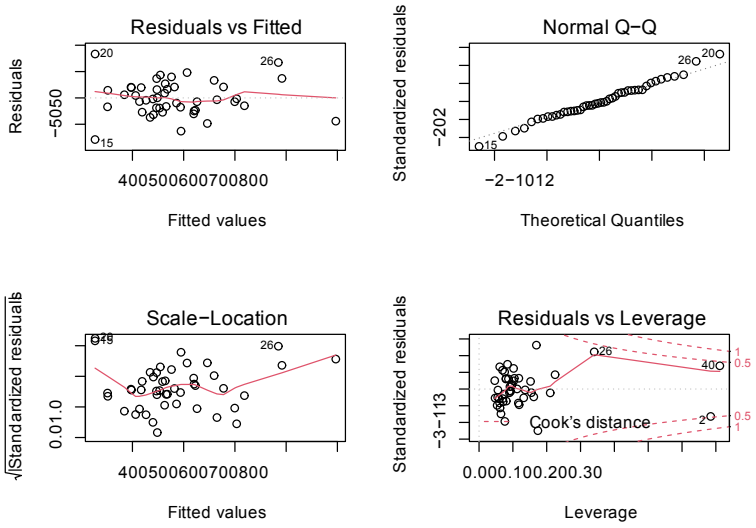

Figure 9-5.: Residual analysis of the model proposed for the proportion of LTR-RTs for the Copia and Gypsy superfamilies and the size of the assemblies.

Relation between Gypsy lineages and assembly size.

As it was done for the superfamilies, a pairwise plot of the six lineages belonging to the Gypsy superfamily was carried out (Figure **9-6**). Thereafter, a variable selection analysis is performed in order to discard the variables that do not have a significant relationship, this is carried out through an information criterion, particularly, the Bayesian information criterion (BIC). The multiple linear regression model is selected with the Tekay/DEL and Galadriel lineages, since these are the variables with the lowest AIC (116.4), obtaining an Adjusted-R square of approximately $90\,\%$ (See Appendix B).

Relation between genome size and proportion of LTR-RTs.

To study the correlation, we carried out a simple linear regression model between the genome size of eight *Coffea* species used to construct the library (Table 9.4.2), and the total proportion of LTR-RTs found in section 9.4.2 (See Figure **9-7**). We can observe that there is a tendency towards linearity despite the R square is not too high (0.76), and this could suggest that the behavior of the LTR-RTs copies in the genome can explain the increase in genome sizes.

9.5. Discussion

The objective of this study was to understand the relationships of the amount of LTR-RTs on the genome sizes of *Coffea* species. In previous studies, using partial sequencing, a contribution

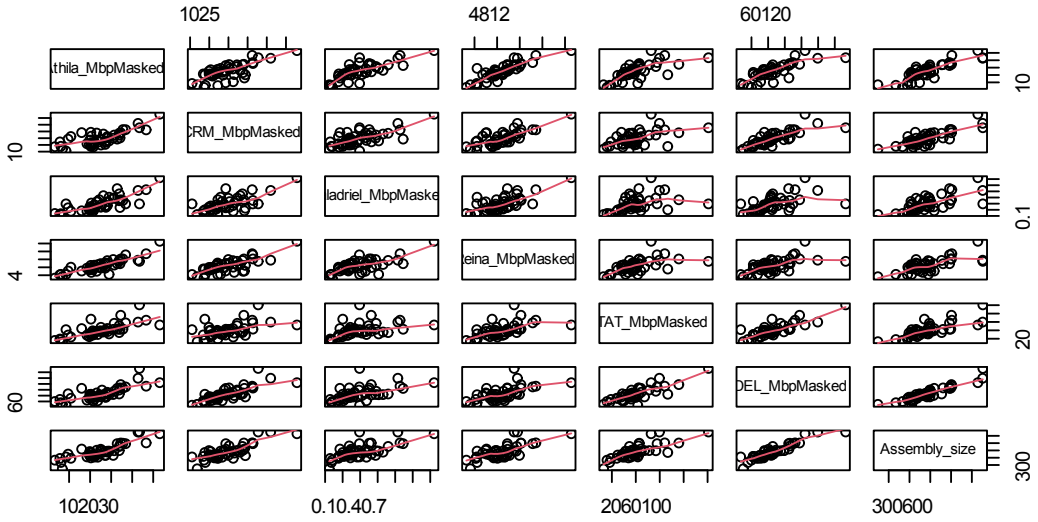

Figure 9-6.: Pairwise plot of the proportion (in percentages) of Gypsy lineages and genome assembly size (in Mbp).

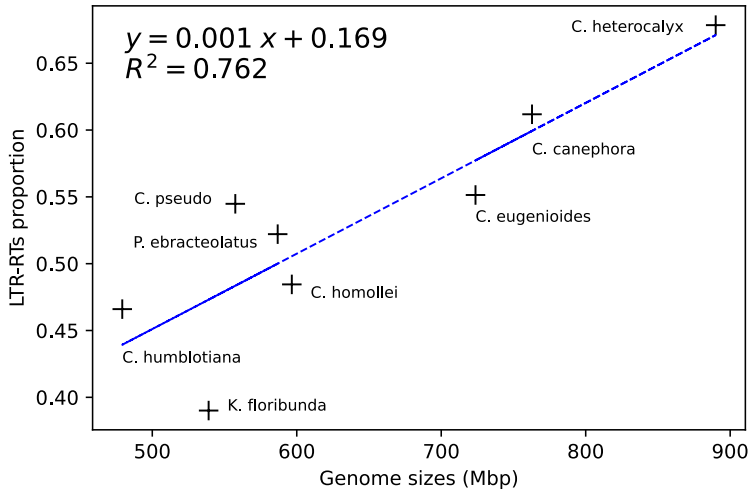

Figure 9-7.: Correlation between LTR-RTs proportions and Genome sizes. This analysis was performed using the genomes of *C. canephora* Pierre ex A.Froehner - DH200-94, *C. eugenioides* S.Moore - BUA, *C. heterocalyx* Stoff., *C. homollei* J.-F.Leroy. *C. humblotiana* Baill., *C. pseudozanguebariae* Bridson, *P. ebracteolatus* Hiern, and *Kraussia floribunda* Harv.

of these elements to the size of various coffee genomes was observed, with a potential increase in lineages such as DEL and Sire [111]. Here using complete sequenced genomes, it was found that, there is a significant contribution of LTR-RTs to up to $50\,\%$ of the genome assembly length, with a higher contribution of elements belonging to the Gypsy superfamily, ranging from $30\,\%$ to $40\,\%$. The results obtained from the statistical analysis of significant relationships confirm the correlation between Assembly length and the Gypsy superfamily, with an Adjusted-R of around $90\,\%$ and normality in the data. Similarly, it is suggested that the DEL lineage contributes to a greater extent to the size of the genomes with a proportion ranging from $11\,\%$ to $23\,\%$, and an Adjusted-R of approximately $90\,\%$ according to the relationship analysis. Surprisingly, TAT lineage did not show statistical correlation with genome size.

Coffee genome sizes vary according to geographic location, with the smallest genomes found in East Africa, Comoros and Indian Ocean Islands, while the largest genomes come from West Africa and Southeast Madagascar, suggesting a gradient in genome size, with growth from East to West in Africa and North to Southeast in Madagascar [127]. This is consistent with the proportion of LTR-RTs, where a higher amount of transposable elements is found in biogeographic groups such as WCA ($48\,\%$) and a decrease in this proportion in groups such as NEA ($38\,\%$) and MUS ($39\,\%$) (Figure **9-4**). An interesting behavior is the change between the contribution of LTR-RTs, which varies between closely related species, an example of this is between *C. deweveri* ($40\,\%$) and *C. congenesis* ($54\,\%$), presenting a difference of up to more than $10\,\%$. This is possibly due to either genome-purging mechanisms, via intra-element recombinations and deletions [128] or recent amplification of some LTR-RTs families.

We tested the robustness of our methodology obtained by RepeatMasker on genome assembly, with a methodology based on raw read mapping on a LTR-RT sequence library. We observed that there is a higher proportion found by mapping, due probably to information that may be lost during the assembly process. The loss of information during assembly can be estimated with assembly metrics (e.g. N50) or BUSCO values (see Appendix A). However, It can be noticed that the trends on the composition of LTR-RTs in the genomes are very similar whatever the method used. Together these data allow us to address a first model of the evolution of the coffee genome sizes linked to LTR-RT contents. The amount of LTR-RTs increases from East Africa to West Africa and remains low in Madagascar and the Mascarene Islands. Considering that it has been established that LTR-RTs are activated via environmental stresses (biotic and abiotic), we will have to investigate the type of stresses that might be involved. One of the future investigations will be the impact of arid and non-arid climates on genome size and TE activation, as demonstrated in palms [129].

9.6. Conclusion

In conclusion, the results obtained here give a clear idea of the contribution of LTR-RTs to the size of the genomes of *Coffea* species, and the distribution of these elements following their biogeographical origins. In the same way, it can be observed that the use of tools built from machine learning algorithms allows obtaining results more efficiently and quickly, unlike other conventional algorithms. This study generates a fundamental resource for research on the proportion of LTR-RTs and their possible implications in evolutionary and genetic processes, however, it is still planned to study whether the distributions of transposable elements in the genus *Coffea* are related to proliferation due to stress events or in response to changes in their environment.

Appendices

A. Appendix A

List of species of the *Coffea* genus and *Psilanthus* used in this study. The table includes detailed information on country of origin, % BUSCO (for completeness of the assembly), estimated genome size, size of the assembly, among other values.

Species name	Plant code	Country of origin	Genome size (Mbp)	References	N50 Illumina Assembly (bp)	Illumina Assembly Sum (Mb)	% completeness	Complete BUSCO %	N50 PacBio assembly (Mb)	PacBio assembly Sum (Mb)	Complete BUSCO %
Coffea											
C. arabica	ET39	Ethiopia	1264,1	Cros et al.,1995	12530	649	51,34087493	92.6	32,59	1085	ND
C. boiviniana (Baill.) Drake	A.980	Madagascar	489	Razafinarivo et al.,2012	4437	426	87,11656442	62.6	/	/	/
C. brevipes Hiern.	C417FL	Cameroon	743,28	Noirot et al.,2003	4925	414	55,69906361	71.7	/	/	/
C. canephora Pierre ex A.Froehner	DH200-94	Democratic Republique of Congo	762,84	Razafinarivo et al.,2012	14559	496	65,02018772	92	50,12	672,3	97.1
C. canephora Pierre ex A.Froehner	BUD15	Uganda	762,84	Razafinarivo et al.,2012	10617	620	81,27523465	90.3	/	/	/
C. canephora Pierre ex A.Froehner	C021	Ivory Coast	762,84	Razafinarivo et al.,2012	9776	431	56,49939699	86.3	/	/	/
C. canephora Pierre ex A.Froehner	C033	Ivory Coast	762,84	Razafinarivo et al.,2012	10248	433	56,76157517	71.6	/	/	/
C. canephora Pierre ex A.Froehner	FRT81 (cultivated)	Brazil	762,84	Razafinarivo et al.,2012	15981	482	63,18494049	93.3	/	/	/
C. charrieriana	OA22	Cameroon	699	Hamon, P personnal communication	25231	519	74,24892704	94.8	/	/	/
C. congensis A.Froehner	C409FL	NA	753,06	Razafinarivo et al.,2012	14144	513	68,12206199	78.6	/	/	/
C. congensis A.Froehner	CC53	Republique of Congo	753,06	Razafinarivo et al.,2012	7192	448	59,49061164	85.1	/	/	/
***C. dewevrei* De Wild. & T.Durand**	EB51	Centrafrique Republique	704,16	Noirot et al.,2002	20547	431	61,20768007	93.9	/	/	/
C. dolichophylla J.-F.Leroy	A.206 (P)	Madagascar	669	Hamon, P personnal communication	7888	608	90,8819133	87.3	/	/	/
C. eugenioides S.Moore	DA (P)	Kenya	723,72	Razafinarivo et al.,2012	10099	418	57,75714365	78.4	/	/	/
C. eugenioides S.Moore	BUA	Uganda	723,72	Razafinarivo et al.,2012	14110	566	78,2070414	91.3	54,74	645	96
C. heterocalyx Stoff.	C413FL	Cameroon	889,98	Razafinarivo et al.,2012	15722	394	44,27065777	93.1	5,248	760,3	96.6
C. homollei J.-F.Leroy	SZ	Madagascar	596,58	Razafinarivo et al.,2012	14554	404	67,71933353	84.6	/	/	/
C. homollei J.-F.Leroy	SZ	Madagascar	596,58	Razafinarivo et al.,2012	4613	753	126,2194509	69.7	41,51	585	96.4
C. humblotiana Baill.	BM19/20 (K, MO, TAN)	Comoros	479,22	Razafinarivo et al.,2012	19876	429	89,52047076	92.9	29,63	420,7	87.5
C. humilis A.Chev.	G57 (K)	Ivory Coast	899,76	Razafinarivo et al.,2012	6398	601	66,79558994	88.9	/	/	/
C. kapakata (A.Chev.) Bridson	KAP	Angola	645,48	Noirot et al.,2003	4127	257	39,81533123	26.8	/	/	/
***C. liberica* W.Bull. ex Hiern**	EA61	Ivory Coast	743,28	Noirot et al.,2003	5378	283	38,07448068	37.8	/	/	/
C. macrocarpa A.Rich.	PET (P, K)	Mauritius	577,02	Razafinarivo et al.,2012	20136	393	68,10855776	93	/	/	/
C. mauritiana Lam	BM17-25	Mauritius	469,44	Razafinarivo et al.,2012	34342	389	82,86468984	94.2	/	/	/
C. mayombensis A.Chev.	-	Cameroon	ND	ND	2969	347	ND	62	/	/	/
C. mufindiensis Hutch. ex Bridson	-	Tanzania	ND	ND	6784	432	ND	84	/	/	/
C. myrtifolia (A.Rich. ex DC.) J.-F.Leroy	A1.1	Mauritius	528,12	Razafinarivo et al.,2012	7020	340	64,37930773	75.5	/	/	/
C. myrtifolia (A.Rich. ex DC.) J.-F.Leroy	C414FL	Mauritius	528,12	Razafinarivo et al.,2012	99851	365	69,11308036	96.7	/	/	/
C. sp. 'nkolbisonii'	-	Cameroon	ND	ND	4482	323	ND	76.5	/	/	/
C. perrieri Drake ex Jum. & H.Perrier	A.12	Madagascar	625,92	Razafinarivo et al.,2012	9180	558	89,14877301	87.8	/	/	/
C. pervilleana Drake	A.957	Madagascar	547,68	Razafinarivo et al.,2012	4246	486	88,73794917	75.6	/	/	/
C. pseudozanguebariae Bridson	C407	Kenya	557,46	Noirot et al.,2003	15425	379	67,98694077	89.2	41,95	618,1	96.7
C. racemosa Lour.	IB62 (K)	Mozambique	508,56	Noirot et al.,2003	16275	597	117,3902784	91.3	/	/	/
C. rhamnifolia (Chiov.) Bridson	-	Somalia	ND	ND	42001	412	ND	94.9	/	/	/
C. salvatrix Swynn. & Philipson	C408FL	ND	596,58	Noirot et al.,2003	22411	422	70,73653156	80.5	/	/	/
C. sessiliflora Bridson	C406FL	Tanzania	537,9	Noirot et al.,2003	32437	527	97,97360104	89.8	/	/	/
C. sessiliflora Bridson	PA60	Tanzania	537,9	Noirot et al.,2003	4119	501	93,13998885	76.2	/	/	/
C. sp 3	-	Cameroon	ND	ND	8072	418	ND	87.9	/	/	/
C. sp. 'Congo'	C416FL	Congo	665	Hamon, P personnal communication	6010	477	71,72932331	88.3	/	/	/
C. stenophylla G.Don.	FB55 (K)	Ivory Coast	625	Noirot et al.,2003	13676	500	80	92	/	/	/
C. tetragona Jum. & H.Perrier	A.252 (K, MO, TAN)	Madagascar	528	Razafinarivo et al.,2012	16591	487	92,23484848	92.3	/	/	/
ex. Psilanthus											
P. benghalensis var. bababudanii (Sivar., Biju & P. Mathew) A.P.Davis	PBT1 (CCRI)	India	709	Jingade et al., 2021	17218	508	71,65021157	91.4	/	/	/

P. benghalensis (Heyne ex J.A.Schult.) J.-F. Leroy	PBT5 (CCRI)	India	709	Jingade et al., 2021	15991	467	65,8674189	91.6	/	/	/
P. brassii (J.-F.Leroy) A.P.Davis	-	Australia	ND	ND	1173	278	ND	6.4	/	/	/
P. ebracteolatus Hiern	PSI11 (K, P)	Ivory Coast	586	Cros et al.,1994	10655	495	84,47098976	90.6%	3,35	1122	93.5
P. horsfieldianus (Miq.) J.-F.Leroy	HOR (K)	Indonesia	ND	ND	9059	768	ND	77.6	/	/	/
***P. leroyi* Bridson**	-	Sudan	ND	ND	45234	447	ND	92.3	/	/	/
P. mannii Hook.f.	-	Cameroon	ND	ND	7320	436	ND	89.2	/	/	/
P. melanocarpus (Welw. ex Hiern) J.-F.Leroy	-	Angola	ND	ND	6687	750	ND	88.5	/	/	/
P. travancorensis (Wight & Arn.) J.-F.Leroy	-	India	636	Jingade et al., 2021	3682	431	67,7672956	54.1	/	/	/
P. wightianus (Wall. ex Wight & Arn.) J.-F.Leroy	-	India	631	Jingade et al., 2021	40952	488	77,33755943	95.2	/	/	/

B. Appendix B

Statistical analysis using the R programming language. These analyses demonstrate in more detail the correlation between the Gypsy lineage and, in particular, the Tekay/Del lineage.

Supplementary material 2: Correlation analysis between LTR-RTs proportion and assembly size

1 Correlation between Copia and Gypsy superfamilies and Assembly size

1.1 Exploratory data analysis

First, install and load the corresponding libraries.

```
install.packages('car')
library(car)
```

Then, load the data.

```
dataset1=read.csv('GypsyCopia_AssemblySize.csv',header = T) #Gypsy-copia-Assembly size
head(dataset1)
```

```
  Copia_MbpMasked Gypsy_MbpMasked Assembly_size
1        20.42517         93.13979      497.8163
2        25.43117        199.76897      752.7466
3        16.69567         91.05151      437.9793
4        17.80556        117.93845      518.8235
5        20.22993         92.85315      433.8472
6        20.67914         94.37619      431.4707
```

```
pairs(dataset1, panel=panel.smooth)
```

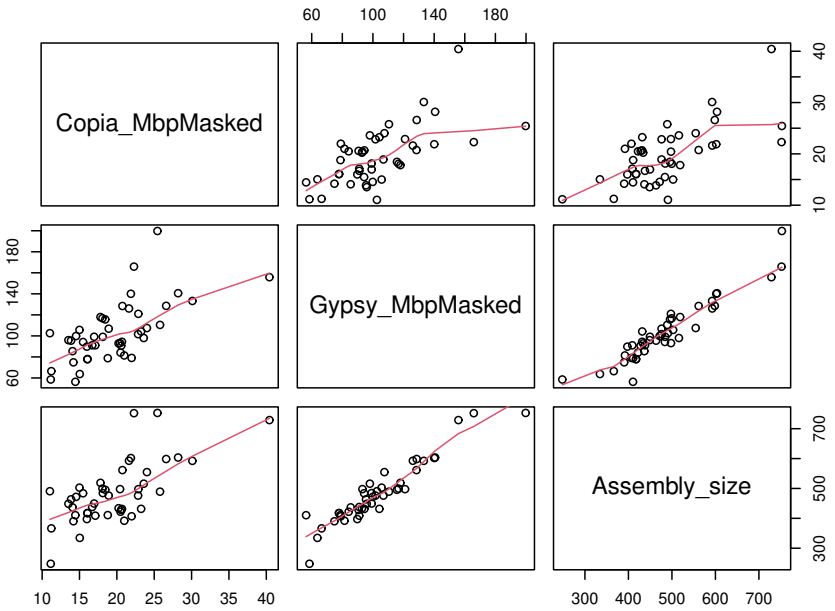

Figure 1: Pairwise plot of Gypsy superfamily, Copia superfamily, and Assembly size

As we can see from the pairwise plot, all regressors, seems to be statistically significant with respect to the assembly size (the response variable), besides this, we can also notice that there is a linear increasing pattern from both regressors with the response. According to this, we propose a multiple linear regresion model, to see whether or not this information is statistically accurate.

1.2 Model fitting

```
#Multiple linear regression model
Y=dataset1$Assembly_size
X_1=dataset1$Copia_MbpMasked
X_2=dataset1$Gypsy_MbpMasked
model<-lm(Y~X_1+X_2, data=dataset1)
summary(model)
```

```
Call:
lm(formula = Y ~ X_1 + X_2, data = dataset1)

Residuals:
     Min       1Q   Median       3Q      Max
-79.025 -19.110   -2.735   20.105   83.316

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 121.2830    20.2641   5.985 3.84e-07 ***
X_1           2.1124     1.1847   1.783   0.0816 .
X_2           3.1129     0.2333  13.343  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 33.15 on 43 degrees of freedom
Multiple R-squared:  0.8958,    Adjusted R-squared:  0.891
F-statistic: 184.9 on 2 and 43 DF,  p-value: < 2.2e-16
```

By fitting a multiple linear regression model, we observed that both regressors, X_1 and X_2 are statistically significant with the response, and the Adjusted R squared is almost 90%, meaning the model is as well statistically significant. Now we check some of the assumptions of the multiple linear regression models.

1.3 Assumptions and multicollinearity analysis

```
#Multicoline
vif(model)
```

```
     X_1      X_2
1.74187 1.74187
```

```
#Correlation. (There is no correlation between the regressors)
with(dataset1, cor(X_1,X_2))
```

```
[1] 0.6526135
```

```
#Residual analysis
par(mfrow=c(2,2))
plot(model)
```

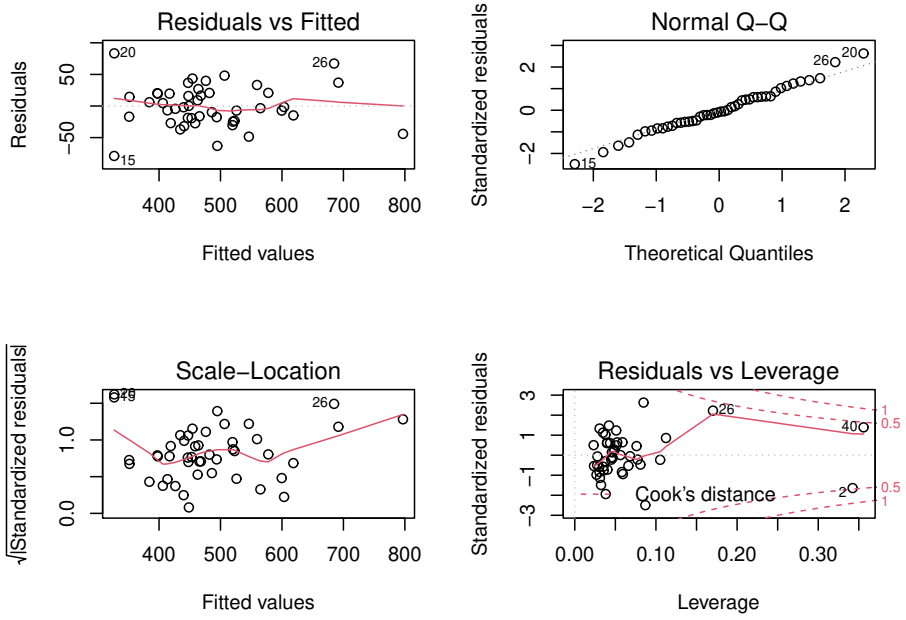

Figure 2: Residual plot

```
marginalModelPlots(model)
```

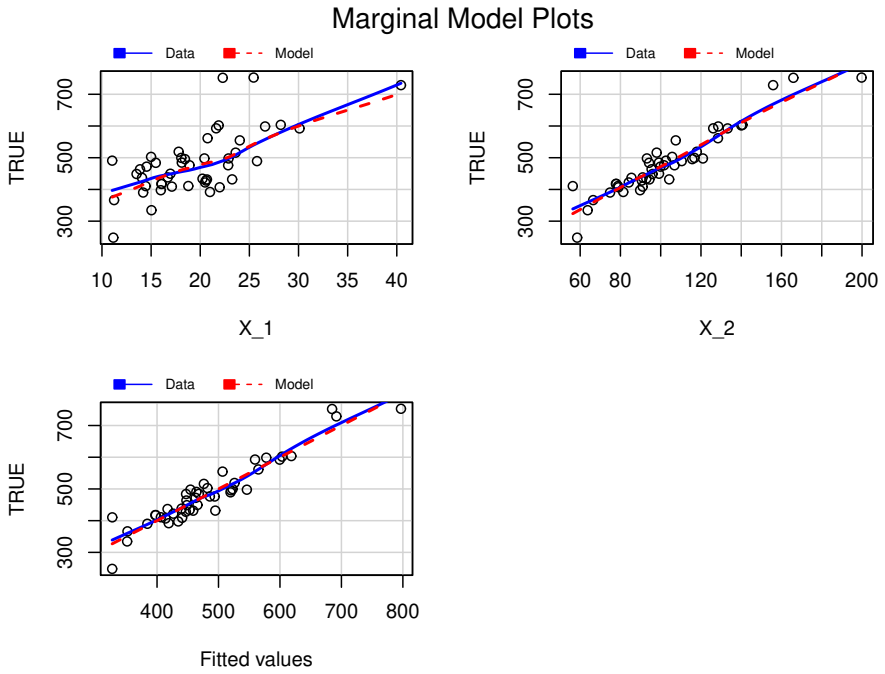

Figure 3: Marginal Model

Firstly, we conduct two initial analysis, to check whether or not our model suffers from multicollinearity and auto-correlation among covariates, for this purpose, we compute the variance inflationary factor (VIF) and correlation respectively. From the VIF we see that in our model we don't have problem with multicollinearity as $VIF < 10$, on the other hand, the correlation for both regressors is of 65%, which is relatively high, however, we ignore this effect into our model analysis.

Secondly, we generate some plots diagnostics for our model that allow us to assume normality and homocedasticity of residuals.

From this analysis and from the marginal model plot, we can conclude that both covariates are significant with respect to the response variable and therefore, there is a relationship between LTR-Retrotransposons and the Assembly size.

2 Analysis for Gypsy and its Lineages

Now that we know that there is a statistically significant relationship between the Gypsy family and the assembly size, in this analysis our goal is to determine whether there is a relationship between the Assembly size and some of the lineages within Gypsy family. To this purpose, we start again with a exploratory data analysis, as follows:

```
dataset2=read.csv('LineagesGypsyMbpMasked.csv',header=T)
head(dataset2)
```

```
  Athila_MbpMasked CRM_MbpMasked Galadriel_MbpMasked Reina_MbpMasked TAT_MbpMasked
1         17.56890       14.91923            0.243854        5.954702       47.24123
2         26.72453       27.75716            0.296558        7.500795      100.74147
3         15.62218       14.96645            0.231496        5.421237       38.90838
4         12.98660       24.15055            0.184256        5.058590       44.23099
5         18.74120       17.09529            0.240447        6.871339       40.10906
6         18.53197       17.18694            0.257677        7.085702       38.36968
  DEL_MbpMasked Assembly_size
1      75.57089       497.8163
2     173.04444       752.7466
3      75.42933       437.9793
4     104.95185       518.8235
5      74.11195       433.8472
6      75.84422       431.4707
```

```
pairs(dataset2, panel=panel.smooth)
```

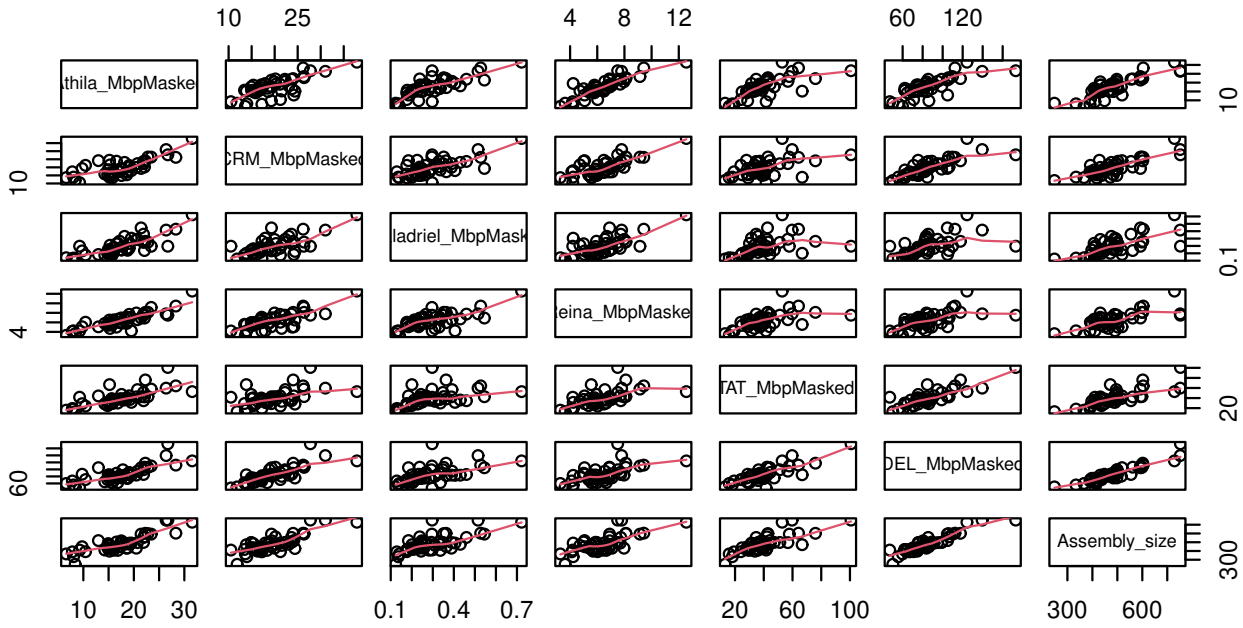

Figure 4: Pairwise plot of Gypsy lineages and Assembly size.

From the pairwise plot, we observed that some of the covariates seems to have an increasing pattern with the response variable, for instance, the covariates W_1, W_3, and W_6 seems to have a stronger relationship with the response, now let's conduct an multiple linear regression model, to see whether this information is accurate.

```
W_1=dataset2$Athila_MbpMasked
W_2=dataset2$CRM_MbpMasked
W_3=dataset2$Galadriel_MbpMasked
W_4=dataset2$Reina_MbpMasked
W_5=dataset2$TAT_MbpMasked
W_6=dataset2$DEL_MbpMasked
W=dataset2$Assembly_size
model2=lm(W~W_1+W_2+W_3+W_4+W_5+W_6,data=dataset2)
summary(model2)
```

```
Call:
lm(formula = W ~ W_1 + W_2 + W_3 + W_4 + W_5 + W_6, data = dataset2)

Residuals:
    Min      1Q  Median      3Q     Max
-74.569 -18.140  -0.259  20.885  61.940

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 136.0976    22.5802   6.027 4.75e-07 ***
W_1           3.7259     2.9195   1.276  0.20943
W_2           1.9174     2.4899   0.770  0.44589
W_3         127.3444    73.3976   1.735  0.09064 .
W_4          -3.3962     7.6662  -0.443  0.66021
W_5           0.4419     0.6533   0.676  0.50281
W_6           2.4610     0.6693   3.677  0.00071 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 32.48 on 39 degrees of freedom
Multiple R-squared:  0.9093,    Adjusted R-squared:  0.8953
F-statistic: 65.14 on 6 and 39 DF,  p-value: < 2.2e-16
```

From the multiple linear regression analysis, we observed that just the covariate W_6 is significant to the response, and the adjusted r-squared is around 90%, however, since the covariates are not significant,it's important to perform a variable selection analysis. To this purpose, we propose using the Bayes information criterion.

```
#Since the covariates W_1 to W_5 are not significant, let's perform a variable selection method.
install.packages("MASS")

library(MASS)
# BIC criterion for variable selection. Bayes Information Criterion.
mod.BIC=stepAIC(model2, direction="both", scope=(~.+W_1+W_2+W_3+W_4+W_5+W_6), k=log(length(W)))
```

```
Start:  AIC=339.43
W ~ W_1 + W_2 + W_3 + W_4 + W_5 + W_6

       Df Sum of Sq   RSS    AIC
- W_4   1     207.1 41352 335.83
- W_5   1     482.6 41628 336.13
- W_2   1     625.6 41771 336.29
- W_1   1    1718.3 42863 337.48
- W_3   1    3175.8 44321 339.02
<none>              41145 339.43
- W_6   1   14263.2 55408 349.29

Step:  AIC=335.83
W ~ W_1 + W_2 + W_3 + W_5 + W_6

       Df Sum of Sq   RSS    AIC
- W_5   1     354.2 41706 332.39
- W_2   1     419.9 41772 332.46
- W_1   1    1925.9 43278 334.09
<none>              41352 335.83
- W_3   1    3652.2 45004 335.89
+ W_4   1     207.1 41145 339.43
- W_6   1   22422.1 63774 351.93

Step:  AIC=332.39
W ~ W_1 + W_2 + W_3 + W_6

       Df Sum of Sq   RSS    AIC
- W_2   1       187 41893 328.77
- W_1   1      2891 44597 331.65
- W_3   1      3506 45212 332.28
<none>              41706 332.39
+ W_5   1       354 41352 335.83
+ W_4   1        79 41628 336.13
- W_6   1     50145 91852 364.88

Step:  AIC=328.77
W ~ W_1 + W_3 + W_6

       Df Sum of Sq    RSS    AIC
- W_1   1      2857  44750 327.98
<none>              41893 328.77
- W_3   1      4980  46873 330.11
+ W_2   1       187  41706 332.39
+ W_5   1       121  41772 332.46
+ W_4   1         0  41893 332.60
- W_6   1     82139 124032 374.87

Step:  AIC=327.98
W ~ W_3 + W_6

       Df Sum of Sq    RSS    AIC
<none>              44750 327.98
+ W_1   1      2857  41893 328.77
+ W_4   1      1292  43458 330.46
+ W_5   1       773  43976 331.00
+ W_2   1       153  44597 331.65
- W_3   1     21225  65975 342.00
- W_6   1    184898 229648 399.38
```

From the BIC analysis we can conclude that just the covariates W_3 and W_6 are statistically significant to the response.

```
model3=lm(W~W_3+W_6,data=dataset2)
marginalModelPlots(model3)
```

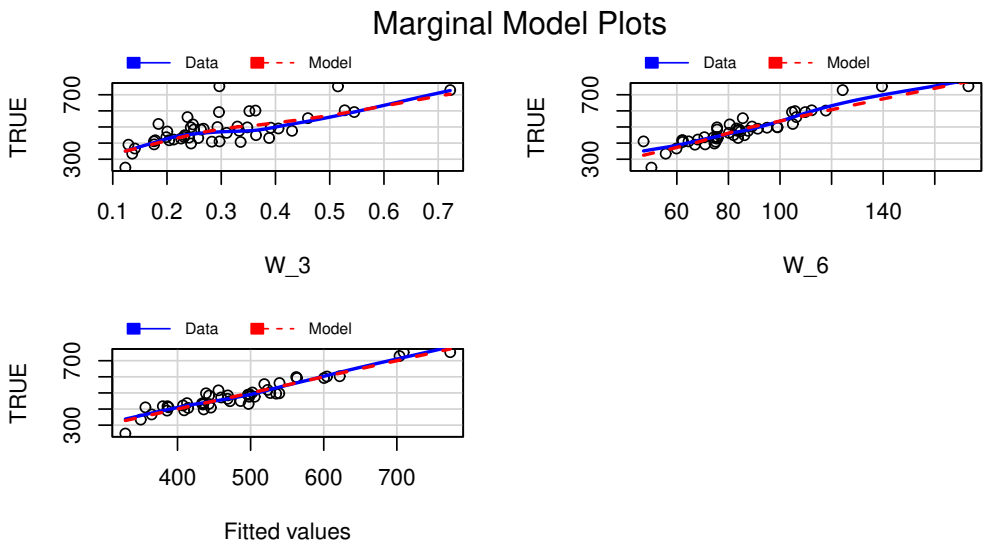

Figure 5: Marginal Model

```
summary(model3)
```

```
Call:
lm(formula = W ~ W_3 + W_6, data = dataset2)

Residuals:
     Min       1Q   Median       3Q      Max
-80.657 -19.742  -3.528  23.542  59.996

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 136.0330    18.2519   7.453 2.86e-09 ***
W_3         216.3516    47.9066   4.516 4.85e-05 ***
W_6           3.3101     0.2483  13.329  < 2e-16 ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1         1

Residual standard error: 32.26 on 43 degrees of freedom
Multiple R-squared:  0.9013,    Adjusted R-squared:  0.8967
F-statistic: 196.4 on 2 and 43 DF,  p-value: < 2.2e-16
```

After conducting the last model just with the covariates W_3 and W_6, we can conclude that these two Gypsy lineages are statistically significant with the Assembly size.

10. Discussions, conclusions, and contributions

In the development of the present thesis it was evidenced that the use of ML for automation in the analysis flow of LTR retrotransposons is not only feasible, but it was also demonstrated to be better in different aspects such as execution speed, metrics (such as F1-score) and accuracy thanks to the ability of the models to identify complex patterns in large datasets [130]. However, the application of ML presents significant challenges especially for processing unstructured information such as DNA because algorithms such as CNNs and RNNs were proposed for other types of data [131]. This chapter will discuss these challenges, how they were solved or how future work could consider them for better results.

10.1. Discussions

10.1.1. DNA coding schemes and available datasets

The first step in virtually all ML-based workflows focuses on the data available to train the algorithms. Unlike image processing and like natural language processing, genomic datasets contain categorical information. Therefore, the first challenge corresponds to performing a transformation of these data to a numerical representation or to do a feature extraction. One of the most basic and intuitive representations is to replace each nucleotide by an integer value, called the DAX coding scheme [132]. However, this simple form of representation could generate biases in ML algorithms, because ML could learn that the nucleotide with the highest number (being C=0, T=1, A=2 and G=3) would have more importance than the others. In addition, it does not take into consideration any biological property taking away from the algorithm a lot of information to improve its performance. There are other ways where the base complementary [133], or certain dinucleotide properties [134, 135] could also be considered important. However these coding schemes present important difficulties like the generation of thousands of features that the model must use for the learning process, considering that the input is thousands (even tens of thousands) of considerably long DNA sequences (e.g. transposable elements of between five thousand and 20 thousand bases). This considerably huge load can make the training times very long and the models need to learn many more parameters, which generates models that are too complex, very difficult to train and that do not obtain good results in terms of accuracy, precision, among others.

Another possibility is a two-dimensional representation called a one-hot vector. In this approach, each nucleotide is converted into a vector of four bits where one will have the value of 1, and the rest of 0 depending on the nucleotide to be represented [136]. If the sequence is of n length, then its one-hot representation would be *4 x n*. This encoding scheme has been frequently used especially for convolutional neural networks [54, 136], and even in other applications in genomics other than on transposable elements [137, 138, 139, 140]. This form of representation overcomes the bias due to the numerical value given to each nucleotide presented by other approaches, however, it worsens the size growth problem by multiplying each sequence in the dataset by four times. Nevertheless, it has presented good results thanks to the fact that convolutional filters can find patterns within local sections of the sequence (e.g. certain motifs of a few bases such as TSDs or PPTs) and generate feature maps that can then be processed by other convolutional filters, finding increasingly higher level patterns (e.g. full LTRs, enzymatic domains, among others).

The extraction of certain features could also be considered, in order to provide as much information as possible to the algorithm to perform the task of interest autonomously. Some structural properties such as sequence length, GC (guanine-cytosine) content, presence or absence of motifs, among others are extractable from the sequences without further computational complexity. However, these features alone are not sufficient for the algorithm to learn. For example, to classify LTR retrotransposons into lineages/families, it is not sufficient to use only the length of the sequence and the length of the LTR domains because different elements of the same lineage can have important variations in these characteristics. Even elements from different lineages may have similar lengths. For this reason, it is necessary to obtain other types of features such as the frequencies of DNA sub-sequences of different lengths *k* called *k*-mers. These sub-sequences are widely used in many areas of bioinformatics such as sequence quality control [141], metagenomics [142], de novo assembly of genomes and transcriptomes [143, 144], genome size estimation [145], and de novo detection of transposable elements [146]. In addition, *k*-mers have been used in ML applications in the classification of TEs [41, 55, 147, 49, 148] and in other genomics tasks [149, 150]. In Chapters 4, 5 and 6 of the present document it was shown that using different public databases and one designed in this work, to train different ML algorithms and different encoding schemes, the best results are obtained using *k*-mers. However, *k*-mers have the restriction of losing positional information within the sequence and of being a one-way transformation (*k*-mers frequencies can be obtained from DNA but not vice versa). Keeping positional positions of *k*-mer on DNA should be an interesting way of research in the future. In addition, it should be considered that calculating these features requires a large computational cost, although there are methods in the literature to overcome this bottleneck [151, 152, 153]. In Chapter 8, a novel way to compute these counts from convolution operations was presented, accelerating the obtaining of more than five thousand *k*-mers frequencies using a CNN with untrainable filters and running on GPUs.

On the other hand, currently available datasets present their own challenges. For an ML-based

tool to be useful, it must exhibit outstanding performance on real-world cases and not just on data used for training. This desirable feature is called generalization [154, 155]. However, LTR-RTs are species-specific [68] and have high diversity at their nucleotide level [81]. This characteristic of LTR-RTs makes it necessary to have a sufficiently representative training dataset for the model to generalize adequately. However, the available data do not have TEs from all plant families or all species, which would generate limitations in the predictions made by the tool in genomes of distant species from those used for training. For this reason, Chapter 5 of this thesis presents the design and release of a dataset that integrates three public databases (Repbase, RepetDB and PGSB) and enriches them with sequences of species and plant families that were not found in these, in order to have a greater representativeness and thus improve the generalization of the models. However, this new dataset, called InpactorDB, cannot yet correctly represent all families and orders of the kingdom plantae because they have not yet been sequenced and released. Thanks to massive sequencing projects (10K plants, Earth BioGenome, among others) and new large-scale TE analysis [156], this limitation in the algorithms will be reduced in the future through periodic retraining using these newly released data.

Another challenge presented by the use of the available databases is the heterogeneity of the strategies used to detect, classify and filter LTR retrotransposons. This is because there is no standardized method for the analysis of these elements and each researcher proposes different methodologies to identify TEs within the genome, to classify it into orders, superfamilies and rarely into lineages/families and to consider whether or not it constitutes a reference. According to these decisions, each dataset has its own characteristics, such as different sequence length distributions for each superfamily or lineage (Figure **10-1**), different levels and curation strategies, and their sequences may correspond to consensus or individual genomic sequences. One of the contributions of this thesis was to statistically analyze that training ML algorithms with databases that have the mentioned differences produces significant changes in performance, especially in the F1-Score (Chapter 5). In this way, it was evidenced that using curated data and with consensus sequences produced the best results in the models without significant differences between them. This finding is due to the fact that curing the data eliminates a large amount of false positives, which are frequently encountered by structure-based or de novo algorithms [157, 158, 159]. On the other hand, consensus sequences are a representation of the ancestral sequence state of those that were clustered and aligned to create the consensus [160]. For this reason, both curated and consensus sequences correspond to a good reference for an ML algorithm to learn to recognize and classify them. However, sequence curation requires a great deal of work usually done manually by experts. On the other hand, consensuses can be computed automatically, in much less time and it is for this reason that such a database was also designed and used in this work.

Due to the intrinsic dynamics of LTR-RTs, some lineages are found to a much greater extent than others. For example, in the coffee genus it was found that the Tekay/DEL and TAT lineages are much more frequent contributing on average $17\,\%$ and $7\,\%$ of the genome respectively, than Angela and Galadriel elements that correspond to less than $0.05\,\%$ (Chapter 9). These dy-

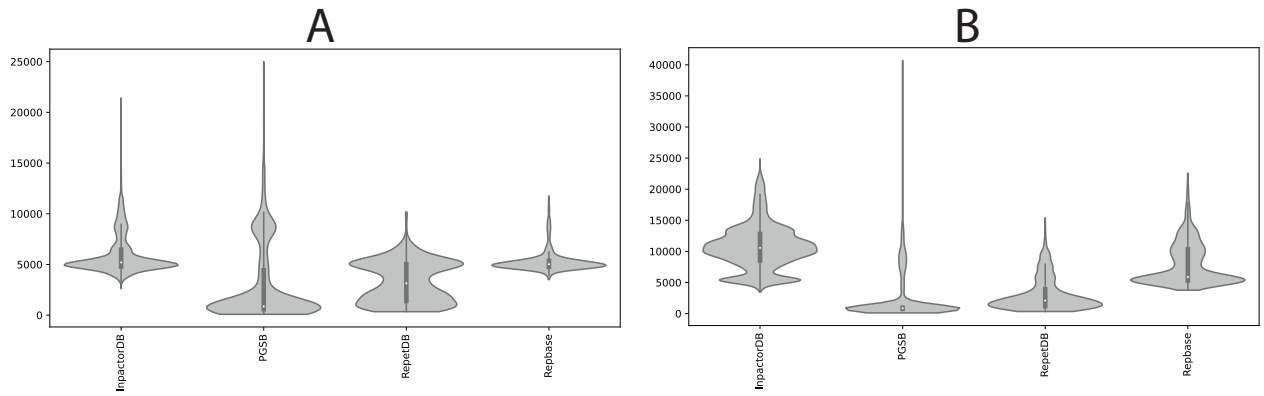

Figure 10-1.: *LTR retrotransposon length distribution among databases. A) Elements that correspond to Copia superfamily, while B) belonging to Gypsy superfamily.*

namics generate very unbalanced datasets, where certain lineages have thousands of sequences while others have only a few records. Even some lineages such as Bryco, Gymco and Lyco [161] are almost non-existent in the available databases, due to low representation of non Angiosperm species in available complete genome sequences. This class imbalance presents an additional challenge for ML models, which achieve better performance on classes with more samples than those with fewer (Chapter 4). Additionally, generating synthetic data from classes with fewer samples is more challenging. Unlike other data types such as images, where more data can be generated with simple operations such as rotations and zooming, DNA sequences are much more complex to generate randomly if it is desirable to retain their biological properties and thus not affect ML models. Although methods exist for balancing underrepresented classes such as SMOTE [162] and ADASYN [163], these would deprecate a great deal of information from the most represented classes, which is necessary for good model performance and generalization. For this reason, a possible line of future work would be the generation of algorithms for data augmentation in genomic datasets, for example through autoencoders or GANs (Generative adversarial networks), but implementing loss functions and metrics adapted to the biological properties of LTR retrotransposons such as total sequence lengths, presence of LTR domains at the beginning and at the end of the element, presence of enzymatic domains, presence of PBS and PPT, among others characteristics.

10.1.2. The detection problem

To implement ML-based software following the approach of this thesis, two essential problems must be considered, the direct detection of full length LTR-RTs from the genome and the classification of these elements into superfamilies and lineages/families. Both problems can be viewed as classification tasks for an ML model. The first one is considered a binary classification, where LTR-RTs are the positive instances and all other genomic sequences are the negative instances. For its

part, the distinction into lineages is taken as a multiclass classification, with each lineage/family being a different class (Chapter 6). Although lineage/family classification is multiclass, it is more trivial than detection. This is because if the premise that the sequences of interest are LTR retrotransposons is met, then it is sufficient to extract numerical features (in this case *k*-mers frequencies), perform some pre-processing steps such as data scaling and dimensionality reduction with PCA (principal component analysis) and train a model to predict to which lineage/family each sequence corresponds. Currently available databases, such as InpactorDB (Chapter 5), can be used for this task.

In contrast, detection requires other considerations. In this task, the input should be genomic sequences assembled as contigs, scaffolds and in the best cases whole chromosomes (called pseudochromosomes). Therefore one must split these long length sequences into shorter sections so that an ML model (which must receive inputs of the same length) predicts which segments correspond to LTR-RTs and which do not. However, this approach presents additional challenges. Consider dividing an input sequence of length n, into segments of length m, with an overlap of l nucleotides. The task of the model would then be to predict which segments are considered to be LTR-RTs. But it cannot be guaranteed that the entire segment corresponds to one element and even, it cannot be considered to be a complete LTR-RT. The only claim that could be made is that the segment contains nucleotides that correspond to an element. This difficulty is due to the fact that LTR retrotransposons have lengths ranging from 4 thousand bases (Ikeros lineage) to 21 thousand bases (TAT lineage) (GyDB; https://gydb.org/). Therefore, if the segment length $m > 4,000$, then predicted segments containing an element (e.g. from the Ikeros lineage) of length less than m could contain other genomic components in addition to the LTR-RT. Conversely elements of lengths greater than m (e.g. from the TAT lineage) would be split into several segments.

A possible solution to this challenge could be setting a segment length m small enough to detect the shortest LTR-RTs and in the end unify the predicted segments as the longest length lineages. Under this perspective, the classification problem for the ML model would be modified to the following: Given a DNA segment of length m, predict whether this corresponds to a fragment of an LTR-RT or not. However, this approach poses more challenges than solutions. First, the ML model would be trained with segments of LTR-RTs (of length m) and not with the whole sequences, taking away from the model the possibility to use important structural features such as LTRs at the beginning and at the end, which are deterministic. In addition, different regions of the elements have important differences. For example, LTRs are non-coding sections that are extremely variable between LTR-RTs of different lineages/families and between plant species, but particularly rich in AT contents. The internal regions, on the other hand, encode enzymatic domains and are much more conserved (Chapter 3). Therefore, using such a diverse dataset could confound the model and not yield good results. An example case would be for the model to predict a segment containing an enzymatic domain as a negative instance (gene) or vice versa. Another problem is the accuracy with which the start and end position of the element within the genome

is detected. The goal of a model that detects elements is that the predicted start and end positions are as close as possible to the actual ones. However, this accuracy depends on the overlapping length l, the smaller it is, the more accurate the predictions will be. Suppose $l = 10$. The algorithm would then give predictions in the range of 10 nucleotides. Naturally, the desired accuracy would be given with $l = 1$. However, such a small value would accommodate the following challenge: the amount of information that must be repeatedly analyzed by the model. Consider that the number of subsequences t that can be extracted from a sequence of length n, each of length m, and with an overlapping of l, is given by the equation 10-1

$$t = \frac{n - m}{l} \tag{10-1}$$

Therefore, by defining a $l = 1$, $n - m$ sequences would be analyzed. If the input is a chromosome of about 40 million bases, and $m = 4,000$ then more than 39 million sequences would have to be analyzed, which would make the algorithm too slow and require too many computational resources.

To overcome the challenges presented by the LTR-RTs detection problem, a hybrid approach is proposed in this thesis. First, the input sequences are divided into 50 thousand base segments without overlapping. Then, these sequences are fed into a convolutional neural network called Inpactor2_Detect (Chapter 8), which predicts whether or not the segment contains LTR-RTs and are stored for further analysis. The remaining segments are discarded to reduce the amount of memory required. Then, a structure-based algorithm (LTR_Finder [164]) is run to detect the contained elements and predict their start and end positions. Finally, these positions are used to extract the elements predicted by the hybrid approach of container segments. This hybrid approach is more efficient than analyzing each segment by the model. It also allows us to filter out the segments that are not of interest through a neural network (which takes seconds) and focus the rest of the study on the segments that contain the elements. Because LTR_Finder runs sequentially, this approach uses a parallelization strategy similar to [165], where the tool is run multiple times, once for each available core, on different segments. At the end, all the results are integrated, storing only the sequences of the predicted elements, in order to free up memory space. Another benefit of this approach is that the use of different approaches contributes to improving the reliability of the results compared to using only one [81]. Although this approach overcomes most of the challenges, there is a problem with dividing the sequences into segments, because some elements could be split and thus the structure-based approach would not be able to detect them. However, in [165] it is proposed that most of these undetected LTR-RTs are represented by complete copies identified in other segments, with a loss of less than $1\,\%$. Additionally, in order to reduce this problem as much as possible, the approach proposed in this thesis can be executed in different cycles (from 1 to 5), where each cycle divides differently the input sequences in order to predict the elements that remain split in any of the partitions. At the end, the result of all the cycles are unified, eliminating those elements detected in more than one cycle.

Despite the progress shown in the task of detecting LTR retrotransposon, this challenge is still open, especially because of the computational time required. A possible solution could come from a similar problem, but applied on completely different data. In computer vision, the object detection problem is quite common [166, 167]. This problem is based on extracting features from images in order to identify a region of interest [168] and even to predict what it corresponds to. One of the most widely used neural networks in these tasks are CNNs and especially an architecture called YOLO [169]. The principle of this network is to make a single analysis of the entire image and thus accelerate the detection of objects within it. Following this approach, one could propose the problem of detecting LTR retrotransposons within genomes as an object detection task as follows: consider a scaffold/chromosome as the equivalent of an image and the LTR-RTs within the sequence as the objects of interest. Therefore, the predictions of the neural network would be the starting position, length and lineage/family of the LTR-RTs within the scaffold/chromosome. Applying the principle proposed by YOLO, in a single analysis cycle, both detection and classification of all the LTR retrotransposons contained within the input sequences could be done, unifying both problems and speeding up execution times, possibly reaching the order of seconds (using GPUs). Although this proposal sounds promising, a great deal of work is required to realize it. First, DNA data differ too much from images, even in the two-dimensional representations used in this work (e.g. one-hot). For this reason, the use of transfer learning would not be possible and a dataset specially designed for this network should be generated that is sufficiently representative to reach a good level of generalization. This dataset should be constructed from genomes with a good level of TE annotation, since it should consider both the elements of interest (LTR-RTs in this case) and all other portions of the genome, which leads to analyzing large amounts of data. Finally, a completely new neural network should be constructed that fits this problem well, because it would not be feasible to use the same one published in the YOLO work. This task brings with it some challenges in terms of hyper-parameter tuning, the size of the training dataset and even the hardware needed to train and tune this network. However, thanks to advances in GPUs, with more and more dedicated memory and more CUDA cores, the availability of large-scale analysis of TEs on hundreds of genomes, and the availability of specialized DL frameworks in genomics, this task is now feasible and would be a line of future work that is likely to get very good results.

10.1.3. Integration of ML models in a one-shot tool

Most existing tools to detect LTR-RTs use as input data assemblies (EDTA [82], LTR_finder [164], LTRharvest [170], LTRdetector [171], LTR annotator [159], LTR_retriever [157], DARTS [172]). However, it is well known that assembly tools have many problems with highly repetitive sections of genomes [173, 174, 175], especially using short sequencing reads, causing most LTR-RTs not to be assembled [176], as well as generating misorderings, deletions, collapsed repeats and other assembly errors [177]. For this reason, it is common for detection software to deal with highly fragmented assemblies and large regions of Ns (unknown nucleotides). This situation po-

ses a challenge for all software based on any methodology, including those based on ML, such as the one proposed in this thesis (Inpactor2, Chapter 8). If the input assembly to the algorithm has an $N50$[1] < 10 or 20 kb, for example, it is possible that very few complete LTR-RTs will be detected due assembly fractionation and to the lengths of these elements (ranging from 4 thousand to 21 thousand). The problem lies in the fact that if this library is used to annotate TEs (e.g. using RepeatMasker), there will be an underestimation of these elements to the genome size. Other issues include the impossibility to map the insertions of LTR-RTs into chromosomes, little information on interaction of the elements with genes, impossibility or unrealistic estimates of insertion times, false estimates on the diversity of LTR-RT lineages/families, among others. A different approach could be considered to eliminate the bias produced by assemblers using sequencing reads directly to assemble LTR-RTs and not detecting them from fragmented assemblies due to these same sequences. In the literature there are repetitive sequence assemblers such as NGSReper [178], TEdna [179], REPdenovo [180], and an assembler based on De Bruijn Graphs [181]. However, the confidentiality of the results obtained by these algorithms is still unclear, especially in LTR retrotransposon, because the methodologies on which they are based are not sufficiently robust, the sequencing technologies may present high levels of noise, and there is no clear benchmarking method to compare the results of the different tools [176]. To overcome this challenge, it would be interesting to benchmark different assembly tools using a methodology similar to [82], and include the best tool found as an additional module of the LTR-RTs detection software. However, long read sequencing technologies such as PacBio or Nanopore could offer a more straightforward solution. Considering that these technologies can generate sequences between 10 to 100 kb, and even generate ultra-long-reads (> 100 kb, up to 2 Mbp) [182, 183, 184], a complete LTR-RT element could be contained in a single read. Therefore, one could run detection tools directly on these reads, especially those with lengths close to or greater than 40 - 50 kb, without the need to use an assembler beforehand. Or even using automatic assemblers in order to have sequences of even greater length. This is because some tools analyze segments of a few thousand base pairs at a time (e.g. Inpactor2 uses 50 kb sections or LTR_Finder_Parallel uses 1Mb segments [165]. For this reason, long read sequencing technologies are expected to influence a new revolution in the field of transposable elements [185].

Another interesting challenge arises when comparing an ML-based tool for detection and classification of transposable elements. Due to the assembler problems discussed above, very few fully assembled and annotated genomes currently exist. In plants, one of the best genomes is that of rice (*Oryza sativa*) [82] and for this reason it was used as a reference for comparisons in Chapter 8. The difficulties in using incomplete or low quality annotation is that one does not have enough information to consider whether a prediction made by a tool and not found in the annotation is a false positive. The same is true for those elements that were not found by the tool and are not found in the annotation, they could not be considered true negatives. Therefore, until more and better assemblies and annotations are obtained, one should consider using another type of

[1] N50 is defined as the sequence length of the shortest contig at 50 % of the total genome length

approaches based on two different results: the libraries created by the programs and the annotations made with these libraries. In the first approach (based on the libraries), one could see which elements are found by all programs and which are specific to a group of tools. However, this approach requires a great deal of manual work to check whether these elements are false positives or not, for example by checking the element structure, as well as confirming whether it contains enzymatic domains and whether it has nested insertions from other TEs, making large-scale analyses unfeasible. In addition, this approach has the problem that it would not provide information on how representative the LTR-RTs contained in the library are, which could be obtained using the annotations. In the second case (based on the annotations), one could compare the percentage contribution of the LTR-RTs in the original annotations and in those made through the tools under review. Another challenge in comparing results is that currently only a few programs classify LTR retrotransposons at the lineage/family level (Inpactor V1 [78], Inpactor2 8, TEsorter [186]) and thus only a few genomes have annotations with elements classified at this level. This makes it difficult to obtain the performance of a new program that classifies LTR-RTs into lineages/families.

Another factor to consider when implementing ML-based software is whether to split the different tasks into multiple models or to pursue a single-model approach. These tasks could be: receiving the input data and transforming them accordingly, detecting LTR-RTs, filtering out those that do not correspond to a good reference, classifying them and finally annotating them. On the other hand, the goal of any tool should be to be easy to install and use, so that a user with little computational knowledge should be able to execute all these tasks and obtain the results without further interaction. Thus, one consideration could be to use a single model (e.g. a neural network) that is designed, trained and integrated into the software to perform all the tasks in a unified way. This case would be ideal because predictions could be generated much faster, errors would not propagate through the other models, and maintenance or retraining would be much simpler. However, this approach is much more complex to achieve (see Section 10.1.2). The other way would be to train an independent model to perform a specific task (such as the one shown in Chapter 6). For example, Chapter 7 presents a neural network, called Inpactor2_Filter, which was designed, tuned and trained with the sole objective of filtering sequences that do not correspond to good references. On the other hand, three other neural networks are presented in Chapter 8 to detect, count *k*-mers and classify elements. This approach, although it tends to be a bit more complex, improves the performances of each model separately, because each task has different specificities requiring different datasets that, in some cases, require different coding schemes.

10.2. Conclusions

The research presented in this paper has the following conclusions:

- In the design of an ML-based application to study LTR retrotransposons, it is crucial to use

an encoding or feature extraction scheme that provides as much information as possible to the model. The performance of the model in terms of precision, accuracy, sensitivity, among others, depends largely on this process. In this sense, for the three tasks presented in the approach proposed by this thesis (detection, filtering and classification), the information provided by the *k*-mers frequencies (with $1 <= k <= 6$) are sufficient to train a model and obtain promising results.

- ML models trained with currently available LTR retrotransposon datasets that differ from each other in their properties (such as curation levels and sequence types, consensus or individual genomic sequences) have significant differences in performance, especially for the F1-Score metric. Curated datasets and those containing consensus sequences are the best. For this reason, it is possible to unify available datasets by creating consensus sequences from those containing individual genomic sequences. This technique increases the amount of usable data in the design and construction of an ML-based tool.

- The best algorithms for the detection problem (treated as a binary problem) were MLP, SVC, and LR with performances between 95 and $97\,\%$ on F1-Score. In the lineage classification task (multiclass problem) the best were KNN, LDA, and SVC with F1-Score between 96 to $97\,\%$. In a mixed problem (detection + classification) the best results were obtained by KNN and LDA with F1-Score scores of 94 to $95\,\%$. On the other hand, using fully connected neural networks, higher F1-Scores were obtained in each of the problems treated ($98\,\%$ in all three cases) with shorter training times and predictions in less time.

- Inpactor2 obtained the highest accuracy and F1-Score of the software compared in this work, with $96.1\,\%$ and $91.9\,\%$. It also obtained the second best values in specificity, precision and FDR, second only to EDTA. However, Inpactor2 obtained $28\,\%$ more sensitivity than EDTA.

- Inpactor2 is up to seven times faster than EDTA. This is especially true for larger genomes (*Zea mays*). In addition, Inpactor2 can be run in minutes on genomes up to 1.2 Gb (the *Coffea arabica* genome was run in 26 minutes). Finally, Inpactor2 can be installed in an anaconda environment and run using only one command line. These results demonstrate a proper integration and implementation of ML models for analyzing LTR-RTs into a usable tool for researchers interested in this topic of study.

- The use of Inpactor2 has made it possible to analyze LTR-RTs in large-scale studies. In particular, it was applied to the study of 46 wild species of the genus *coffea* in order to answer the question of whether these elements influence genome size diversity within the genus. Using this tool, it was possible to obtain the results in less time and in an automatic way, demonstrating that LTR-RTs have influenced the evolution of genomes in the genus *coffea*. In particular, it was shown that they influence the variability in genome size of the different phylogeographic groups, especially those of the Tekay/Del lineage/family.

10.3. Contributions

During the execution of this doctoral thesis, new knowledge products such as scientific articles, conference proceedings, papers and scientific software were produced. These products are listed in table **10-1**.

Title	Product type	Year	Category	Journal	Link
Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning	Review article	2019	Q1	IJMS	Link
A systematic review of the application of machine learning in the detection and classification of transposable elements	Review article	2019	Q1	PeerJ	Link
Measuring performance metrics of machine learning algorithms for detecting and classifying transposable elements	Research article	2020	Q2	Processes	Link
InpactorDB: a classified lineage-level plant LTR retrotransposon reference library for free-alignment methods based on machine learning	Research article	2021	Q1	Genes	Link
K-mer-based machine learning method to classify LTR retrotransposons in plant genomes	Research article	2021	Q1	PeerJ	Link
Inpactor2: A software based on deep learning to identify and classify LTR retrotransposons in plant genomes	Research article	2022	Q1	BIB	-
Automatic curation of LTR retrotransposon libraries from plant genomes through machine learning genomes	Research article	2022	Q2	JIB	-
Deep Neural Network to Curate LTR Retrotransposon Libraries from Plant Genomes	Proceedings	2021	-	LNNS	Link
SENMAP: A Convolutional Neural Network Architecture for Curation of LTR-RT Libraries from Plant Genomes	Proceedings	2021	-	IEEE Xplore	Link
Inpactor2	Software	2022	-	-	Link

Table 10-1.: *Articles, proceedings, and software generated during the doctoral thesis. Articles without link are currently under review.*

In addition, the results obtained in this work were divulged in different scientific events (Table **10-2**).

Title	Year	Event	Link
Identification and Annotation of LTR retrotransposons: A challenge to understand the plant genome structure and evolution	2019	V CCBCOL. Ibagué, Colombia	-
Machine Learning en Identificación y clasificación de retrotransposones en genomas de plantas.	2020	XVIII Jornadas de Ingeniería Universidad de Caldas. Manizales, Colombia	-
A Machine Learning-based approach to identify and classify LTR retrotransposons in plant genomes.	2021	GDR 3546 French meeting of Transposable Elements. Paris, France	-
Deep Neural Network to Curate LTR Retrotransposon Libraries from Plant Genomes	2021	15th PACBB. Salamanca, Spain	Link
Analysis in silico of transposable elements: towards tools based on machine learning.	2021	ALAG 2021. Valdivia, Chile	Link
Inpactor2: A Neural Network-based approach to identify and classify LTR retrotransposons in plant genomes.	2021	1st FLA Workshop on Omics and Bioinformatics Santiago de Chile, Chile	Link

Table 10-2.: *Oral presentations that were done during the doctoral thesis.*

Bibliography

[1] F. Choulet, A. Alberti, S. Theil, N. Glover, V. Barbe, J. Daron, L. Pingault, P. Sourdille, A. Couloux, E. Paux, and Others, "Structural and functional partitioning of bread wheat chromosome 3B," *Science*, vol. 345, no. 6194, p. 1249721, 2014.

[2] E. Ibarra-Laclette and E. Lyons, "Architecture and evolution of a minute plant genome," *Nature*, vol. 498, no. 7452, pp. 1–6, 2013.

[3] M. I. Tenaillon, J. D. Hollister, and B. S. Gaut, "A triptych of the evolution of plant transposable elements," *Trends in Plant Science*, vol. 15, no. 8, pp. 471–478, 2010.

[4] I. Makarevitch, A. J. Waters, P. T. West, M. Stitzer, C. N. Hirsch, J. Ross-Ibarra, and N. M. Springer, "Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress," *PLoS Genetics*, vol. 11, no. 1, 2015.

[5] E. Todorovska, "Retrotransposons and their Role in Plant-Genome Evolution," *Biotechnology & Biotechnological Equipment*, vol. 2818, no. August, pp. 294–305, 2014.

[6] E. Casacuberta and J. González, "The impact of transposable elements in environmental adaptation," *Molecular Ecology*, vol. 22, no. 6, pp. 1503–1517, 2013.

[7] G. Bonchev and C. Parisod, "Transposable elements and microevolutionary changes in natural populations," *MOLECULAR ECOLOGY RESOURCES*, vol. 13, pp. 765–775, sep 2013.

[8] S.-F. Li, T. Su, G.-Q. Cheng, B.-X. Wang, X. Li, C.-L. Deng, and W.-J. Gao, "Chromosome Evolution in Connection with Repetitive Sequences and Epigenetics in Plants," *GENES*, vol. 8, oct 2017.

[9] S. Ou, J. Chen, and N. Jiang, "Assessing genome assembly quality using the LTR Assembly Index (LAI)," *Nucleic Acids Research*, no. August, pp. 1–11, 2018.

[10] D. Hermann, F. Egue, E. Tastard, D.-H. Nguyen, N. Casse, A. Caruso, S. Hiard, J. Marchand, B. Chenais, A. Morant-Manceau, and J. D. Rouault, "An introduction to the vast world of transposable elements - what about the diatoms?," *DIATOM RESEARCH*, vol. 29, pp. 91–104, jan 2014.

[11] F. Mascagni, A. Vangelisti, T. Giordani, A. Cavallini, and L. Natali, "Specific LTR-Retrotransposons Show Copy Number Variations between Wild and Cultivated Sunflowers.," *Genes*, vol. 9, p. 433, aug 2018.

[12] T. Wicker, F. Sabot, A. Hua-Van, J. L. Bennetzen, P. Capy, B. Chalhoub, A. Flavell, P. Leroy, M. Morgante, O. Panaud, E. Paux, P. SanMiguel, and A. H. Schulman, "A unified classification system for eukaryotic transposable elements," *Nature Reviews Genetics*, vol. 8, no. 12, pp. 973–982, 2007.

[13] P. S. Schnable, D. Ware, R. S. Fulton, J. C. Stein, F. Wei, S. Pasternak, C. Liang, J. Zhang, L. Fulton, T. A. Graves, P. Minx, A. D. Reily, L. Courtney, S. S. Kruchowski, C. Tomlinson, C. Strong, K. Delehaunty, C. Fronick, B. Courtney, S. M. Rock, E. Belter, F. Du, K. Kim, R. M. Abbott, M. Cotton, A. Levy, P. Marchetto, K. Ochoa, S. M. Jackson, B. Gillam, W. Chen, L. Yan, J. Higginbotham, M. Cardenas, J. Waligorski, E. Applebaum, L. Phelps, J. Falcone, K. Kanchi, T. Thane, A. Scimone, N. Thane, J. Henke, T. Wang, J. Ruppert, N. Shah, K. Rotter, J. Hodges, E. Ingenthron, M. Cordes, S. Kohlberg, J. Sgro, B. Delgado, K. Mead, A. Chinwalla, S. Leonard, K. Crouse, K. Collura, D. Kudrna, J. Currie, R. He, A. Angelova, S. Rajasekar, T. Mueller, R. Lomeli, G. Scara, A. Ko, K. Delaney, M. Wissotski, G. Lopez, D. Campos, M. Braidotti, E. Ashley, W. Golser, H. Kim, S. Lee, J. Lin, Z. Dujmic, W. Kim, J. Talag, A. Zuccolo, C. Fan, A. Sebastian, M. Kramer, L. Spiegel, L. Nascimento, T. Zutavern, B. Miller, C. Ambroise, S. Muller, W. Spooner, A. Narechania, L. Ren, S. Wei, S. Kumari, B. Faga, M. J. Levy, L. McMahan, P. Van Buren, M. W. Vaughn, K. Ying, C.-T. Yeh, S. J. Emrich, Y. Jia, A. Kalyanaraman, A.-P. Hsia, W. B. Barbazuk, R. S. Baucom, T. P. Brutnell, N. C. Carpita, C. Chaparro, J.-M. Chia, J.-M. Deragon, J. C. Estill, Y. Fu, J. A. Jeddeloh, Y. Han, H. Lee, P. Li, D. R. Lisch, S. Liu, Z. Liu, D. H. Nagel, M. C. McCann, P. SanMiguel, A. M. Myers, D. Nettleton, J. Nguyen, B. W. Penning, L. Ponnala, K. L. Schneider, D. C. Schwartz, A. Sharma, C. Soderlund, N. M. Springer, Q. Sun, H. Wang, M. Waterman, R. Westerman, T. K. Wolfgruber, L. Yang, Y. Yu, L. Zhang, S. Zhou, Q. Zhu, J. L. Bennetzen, R. K. Dawe, J. Jiang, N. Jiang, G. G. Presting, S. R. Wessler, S. Aluru, R. A. Martienssen, S. W. Clifton, W. R. McCombie, R. A. Wing, and R. K. Wilson, "The B73 Maize Genome: Complexity, Diversity, and Dynamics," *Science*, vol. 326, no. 5956, pp. 1112–1115, 2009.

[14] A. H. Paterson, J. E. Bowers, R. Bruggmann, I. Dubchak, J. Grimwood, H. Gundlach, G. Haberer, U. Hellsten, T. Mitros, A. Poliakov, J. Schmutz, M. Spannagl, H. Tang, X. Wang, T. Wicker, A. K. Bharti, J. Chapman, F. A. Feltus, U. Gowik, I. V. Grigoriev, E. Lyons, C. a. Maher, M. Martis, A. Narechania, R. P. Otillar, B. W. Penning, A. a. Salamov, Y. Wang, L. Zhang, N. C. Carpita, M. Freeling, A. R. Gingle, C. T. Hash, B. Keller, P. Klein, S. Kresovich, M. C. McCann, R. Ming, D. G. Peterson, M. ur Rahman, D. Ware, P. Westhoff, K. F. X. Mayer, J. Messing, and D. S. Rokhsar, "The Sorghum bicolor genome and the diversification of grasses.," *Nature*, vol. 457, no. 7229, pp. 551–556, 2009.

[15] F. Denoeud, L. Carretero-Paulet, A. Dereeper, G. Droc, R. Guyot, M. Pietrella, C. Zheng, A. Alberti, F. Anthony, G. Aprea, J.-M. Aury, P. Bento, M. Bernard, S. Bocs, C. Campa, A. Cenci, M.-C. Combes, D. Crouzillat, C. Da Silva, L. Daddiego, F. De Bellis, S. Dussert, O. Garsmeur, T. Gayraud, V. Guignon, K. Jahn, V. Jamilloux, T. Joët, K. Labadie, T. Lan, J. Le-

clercq, M. Lepelley, T. Leroy, L.-T. Li, P. Librado, L. Lopez, A. Muñoz, B. Noel, A. Pallavicini, G. Perrotta, V. Poncet, D. Pot, Priyono, M. Rigoreau, M. Rouard, J. Rozas, C. Tranchant-Dubreuil, R. VanBuren, Q. Zhang, A. C. Andrade, X. Argout, B. Bertrand, A. de Kochko, G. Graziosi, R. J. Henry, Jayarama, R. Ming, C. Nagai, S. Rounsley, D. Sankoff, G. Giuliano, V. a. Albert, P. Wincker, P. Lashermes, and Others, "The coffee genome provides insight into the convergent evolution of caffeine biosynthesis," *science*, vol. 345, no. 6201, pp. 1181–4, 2014.

[16] R. de Castro Nunes, S. Orozco-Arias, D. Crouzillat, L. A. Mueller, S. R. Strickler, P. Descombes, C. Fournier, D. Moine, A. de Kochko, P. M. Yuyama, A. L. L. Vanzela, and R. Guyot, "Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions," *Frontiers in Plant Science*, 2018.

[17] C. M. Vicient and J. M. Casacuberta, "Impact of transposable elements on polyploid plant genomes," *ANNALS OF BOTANY*, vol. 120, pp. 195–207, aug 2017.

[18] P. This, T. Lacombe, M. Cadle-Davidson, and C. L. Owens, "Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1," *Theoretical and Applied Genetics*, vol. 114, no. 4, pp. 723–730, 2007.

[19] H. Xiao, N. Jiang, E. Schaffner, E. J. Stockinger, and E. Van Der Knaap, "A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit," *science*, vol. 319, no. 5869, pp. 1527–1530, 2008.

[20] M. Momose, Y. Abe, and Y. Ozeki, "Miniature inverted-repeat transposable elements of stowaway are active in potato," *Genetics*, vol. 186, no. 1, pp. 59–66, 2010.

[21] E. Butelli, C. Licciardello, Y. Zhang, J. Liu, S. Mackay, P. Bailey, G. Reforgiato-Recupero, and C. Martin, "Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges.," *The Plant cell*, vol. 24, pp. 1242–55, mar 2012.

[22] L. Wei and X. Cao, "The effect of transposable elements on phenotypic variation: insights from plants to humans," *Science China Life Sciences*, vol. 59, pp. 24–37, jan 2016.

[23] C. Vitte, M.-A. Fustier, K. Alix, and M. I. Tenaillon, "The bright side of transposons in crop evolution," *Briefings in Functional Genomics*, vol. 13, no. 4, pp. 276–295, 2014.

[24] P. Baduel and V. Colot, "The epiallelic potential of transposable elements and its evolutionary significance in plants," *Philosophical Transactions of the Royal Society B*, vol. 376, no. 1826, p. 20200123, 2021.

[25] J. Arango-López, S. Orozco-Arias, J. A. Salazar, and R. Guyot, "Application of Data Mining Algorithms to Classify Biological Data: The Coffea canephora Genome Case," in *Advances in Computing*, vol. 735, pp. 156–170, Springer, 2017.

[26] L. Schietgat, C. Vens, R. Cerri, C. N. Fischer, E. Costa, J. Ramon, C. M. A. Carareto, and H. Blockeel, "A machine learning based framework to identify and classify long terminal repeat retrotransposons.," *PLoS computational biology*, vol. 14, p. e1006097, apr 2018.

[27] T. Loureiro, N. Fonseca, and R. Camacho, *Application of Machine Learning techniques on the Discovery and annotation of Transposons in genomes.* Ms.c., Ms.C. Thesis FACULDADE DE ENGENHARIA, UNIVERSIDADE DO PORTO, 2012.

[28] M. Dupeyron, *Dynamique et évolution de deux lignées remarquables de rétrotransposons à LTR dans le genre Coffea (famille des Rubiacées).* PhD thesis, Montpellier, 2017.

[29] K. Rawal and R. Ramaswamy, "Genome-wide analysis of mobile genetic element insertion sites," *Nucleic Acids Research*, vol. 39, no. 16, pp. 6864–6878, 2011.

[30] R. N. Mustafin and E. K. Khusnutdinova, "The Role of Transposons in Epigenetic Regulation of Ontogenesis," *Russian Journal of Developmental Biology*, vol. 49, pp. 61–78, mar 2018.

[31] W. Bao, K. K. Kojima, and O. Kohany, "Repbase Update, a database of repetitive elements in eukaryotic genomes," *Mobile DNA*, vol. 6, no. 1, pp. 4–9, 2015.

[32] J. Amselem, G. Cornut, N. Choisne, M. Alaux, F. Alfama-Depauw, V. Jamilloux, F. Maumus, T. Letellier, I. Luyten, C. Pommier, A. F. Adam-Blondon, and H. Quesneville, "RepetDB: A unified resource for transposable element references," *Mobile DNA*, vol. 10, no. 1, pp. 1–9, 2019.

[33] M. Spannagl, T. Nussbaumer, K. C. Bader, M. M. Martis, M. Seidel, K. G. Kugler, H. Gundlach, and K. F. Mayer, "PGSB plantsDB: Updates to the database framework for comparative plant genome research," *Nucleic Acids Research*, vol. 44, no. D1, pp. D1141–D1147, 2016.

[34] S. Orozco-Arias, P. A. Jaimes, M. S. Candamil, C. F. Jiménez-Varón, R. Tabares-Soto, G. Isaza, and R. Guyot, "InpactorDB: A classified lineage-level plant LTR retrotransposon reference library for free-alignment methods based on machine learning," *Genes*, vol. 12, no. 2, pp. 1–17, 2021.

[35] T. Loureiro, R. Camacho, J. Vieira, and N. A. Fonseca, "Improving the performance of Transposable Elements detection tools.," *Journal of integrative bioinformatics*, vol. 10, no. 3, p. 231, 2013.

[36] S. Orozco-Arias, R. Tabares-Soto, D. Ceballos, and R. Guyot, "Parallel Programming in Biological Sciences, Taking Advantage of Supercomputing in Genomics," in *Advances in Computing* (A. Solano and H. Ordoñez, eds.), vol. 735, pp. 627–643, Zurich: Springer, 2017.

[37] S. Orozco-Arias, G. Isaza, R. Guyot, and R. Tabares-Soto, "A systematic review of the application of machine learning in the detection and classification of transposable elements," *PeerJ*, vol. 7, p. e8311, 2019.

[38] R. Tabares-Soto, S. Orozco-Arias, V. Romero-Cano, V. S. Bucheli, J. L. Rodríguez-Sotelo, and C. F. Jiménez-Varón, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data," *PeerJ Computer Science*, vol. 6, p. e270, 2020.

[39] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321–332, 2015.

[40] P. Larrañaga, B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles, "Machine learning in bioinformatics," *Briefings in Bioinformatics*, vol. 7, no. 1, pp. 86–112, 2006.

[41] F. K. Nakano, S. M. Mastelini, S. Barbon, and R. Cerri, "Improving Hierarchical Classification of Transposable Elements using Deep Neural Networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 8-13 July, (Rio de Janeiro), IEEE, 2018.

[42] O. A. Montesinos-López, A. Montesinos-López, P. Pérez-Rodríguez, J. A. Barrón-López, J. W. Martini, S. B. Fajardo-Flores, L. S. Gaytan-Lugo, P. C. Santana-Mancilla, and J. Crossa, "A review of deep learning applications for genomic selection," *BMC Genomics*, vol. 22, no. 1, pp. 1–23, 2021.

[43] M. H. P. da Cruz, P. T. M. Saito, A. R. Paschoal, and P. H. Bugatti, "Classification of Transposable Elements by Convolutional Neural Networks," in *Lecture Notes in Computer Science*, vol. 11509, pp. 157–168, Springer International Publishing, 2019.

[44] FAO, FIDA, OMS, PMA, and UNICEF, "LA SEGURIDAD ALIMENTARIA Y LA NUTRICION EN EL MUNDO," tech. rep., ONU, Roma, 2020.

[45] ONU, "Alimentación," 2018.

[46] C. A. Deutsch, J. J. Tewksbury, M. Tigchelaar, D. S. Battisti, S. C. Merrill, R. B. Huey, and R. L. Naylor, "Increase in crop losses to insect pests in a warming climate," *Science*, vol. 361, no. 6405, pp. 916–919, 2018.

[47] R. Tito, H. L. Vasconcelos, and K. J. Feeley, "Global Climate Change Increases Risk of Crop Yield Losses and Food Insecurity in the Tropical Andes," *Global Change Biology*, vol. 24, no. 2, 2017.

[48] N. Jiang, "Overview of Repeat Annotation and De Novo Repeat Identification," in *Plant Transposable Elements*, pp. 275–287, Springer, 2013.

[49] G. Abrusán, N. Grundmann, L. Demester, and W. Makalowski, "TEclass - A tool for automated classification of unknown eukaryotic transposable elements," *Bioinformatics*, vol. 25, no. 10, pp. 1329–1330, 2009.

[50] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, "Deep learning: new computational modelling techniques for genomics," *Nature Reviews Genetics*, vol. 20, no. 7, pp. 389–403, 2019.

[51] T. Yue and H. Wang, "Deep learning for genomics: A concise overview," *arXiv preprint arXiv:1802.00810*, 2018.

[52] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, "A primer on deep learning in genomics," *Nature Genetics*, vol. 51, no. 1, pp. 12–18, 2019.

[53] L. Koumakis, "Deep learning models in genomics; are we there yet?," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 1466–1473, 2020.

[54] M. H. P. da Cruz, D. S. Domingues, P. T. M. Saito, A. R. Paschoal, and P. H. Bugatti, "TERL: classification of transposable elements by convolutional neural networks," *Briefings in Bioinformatics*, vol. 22, may 2021.

[55] H. Yan, A. Bombarely, and S. Li, "DeepTE: a computational method for de novo classification of transposons with convolutional neural network.," *Bioinformatics (Oxford, England)*, 2020.

[56] S. Orozco-Arias, M. S. Candamil-Cortes, P. A. Jaimes, E. Valencia-Castrillon, R. Tabares-Soto, R. Guyot, and G. Isaza, "Deep neural network to curate ltr retrotransposon libraries from plant genomes," in *International Conference on Practical Applications of Computational Biology & Bioinformatics*, pp. 85–94, Springer, 2021.

[57] N.-S. Kim, "The genomes and transposable elements in plants: are they friends or foes?," *GENES & GENOMICS*, vol. 39, pp. 359–370, apr 2017.

[58] G. Usai, F. Mascagni, L. Natali, T. Giordani, and A. Cavallini, "Comparative genome-wide analysis of repetitive DNA in the genus Populus L.," *Tree Genetics & Genomes*, vol. 13, p. 96, oct 2017.

[59] C. R. L. Huang, K. H. Burns, and J. D. Boeke, "Active transposition in genomes.," *Annual review of genetics*, vol. 46, pp. 651–75, dec 2012.

[60] A. Testori, L. Caizzi, S. Cutrupi, O. Friard, M. De Bortoli, D. Cora, and M. Caselle, "The role of transposable elements in shaping the combinatorial interaction of transcription factors," *BMC genomics*, vol. 13, no. 1, pp. 1–16, 2012.

[61] M.-A. A. Grandbastien, "LTR retrotransposons, handy hitchhikers of plant regulation and stress response," *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, vol. 1849, pp. 403–416, apr 2015.

[62] N. Krom and W. Ramakrishna, "Retrotransposon insertions in rice gene pairs associated with reduced conservation of gene pairs in grass genomes.," *Genomics*, vol. 99, pp. 308–14, may 2012.

[63] J. Lee, N. E. Waminal, H.-I. Choi, S. Perumal, S.-C. Lee, V. B. Nguyen, W. Jang, N.-H. Kim, L.-Z. Gao, and T.-J. Yang, "Rapid amplification of four retrotransposon families promoted speciation and genome size expansion in the genus Panax.," *Scientific reports*, vol. 7, p. 9045, aug 2017.

[64] M. Elbaidouri and O. Panaud, "Genome-Wide Analysis of Transposition Using Next Generation Sequencing Technologies," in *Plant Transposable Elements*, pp. 59–70, Springer, 2012.

[65] L. Wang, Y. He, H. Qiu, J. Guo, M. Han, J. Zhou, Q. Sun, and J. Sun, "Mdoryco1-1, a bidirectionally transcriptional Ty1-copia retrotransposon from Malus x domestica," *SCIENTIA HORTICULTURAE*, vol. 220, pp. 283–290, jun 2017.

[66] R. C. Paz, M. E. Kozaczek, H. G. Rosli, N. P. Andino, and M. V. Sanchez-Puerta, "Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in Solanum lycopersicum.," *Genetica*, vol. 145, pp. 417–430, oct 2017.

[67] M. Iquebal, S. Jaiswal, C. Mukhopadhyay, C. Sarkar, A. Rai, and D. Kumar, "Applications of bioinformatics in plant and agriculture," in *PlantOmics: The Omics of Plant Science*, pp. 755–789, Springer, 2015.

[68] H. Z. Girgis, "Red: An intelligent, rapid, accurate tool for detecting repeats de-novo on the genomic scale," *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–19, 2015.

[69] G. I. Arabidopsis, S. Kaul, H. L. Koo, J. Jenkins, M. Rizzo, T. Rooney, L. J. Tallon, T. Feldblyum, W. Nierman, M. I. Benito, X. Lin, and Others, "Analysis of the genome sequence of the flowering plant Arabidopsis thaliana," *Nature*, vol. 408, no. December, pp. 796–815, 2000.

[70] J. Yu, S. Hu, J. Wang, G. K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, M. Cao, J. Liu, J. Sun, J. Tang, Y. Chen, X. Huang, W. Lin, C. Ye, W. Tong, L. Cong, J. Geng, Y. Han, L. Li, W. Li, G. Hu, J. Li, Z. Liu, Q. Qi, T. Li, X. Wang, H. Lu, T. Wu, M. Zhu, P. Ni, H. Han, W. Dong, X. Ren, X. Feng, P. Cui, X. Li, H. Wang, X. Xu, W. Zhai, Z. Xu, J. Zhang, S. He, J. Xu, K. Zhang, X. Zheng, J. Dong, W. Zeng, L. Tao, J. Ye, J. Tan, X. Chen, J. He, D. Liu, W. Tian, C. Tian, H. Xia, Q. Bao, G. Li, H. Gao, T. Cao, W. Zhao, P. Li, W. Chen, Y. Zhang, J. Hu, S. Liu, J. Yang, G. Zhang, Y. Xiong, Z. Li, L. Mao, C. Zhou, Z. Zhu, R. Chen, B. Hao, W. Zheng, S. Chen, W. Guo, M. Tao, L. Zhu, L. Yuan, and H. Yang, "A draft sequence of the rice genome (Oryza sativa L. ssp. indica)," *Science*, vol. 296, no. 5565, pp. 79–92, 2002.

[71] R. Akakpo, M.-C. Carpentier, Y. Ie Hsing, and O. Panaud, "The impact of transposable elements on the structure, evolution and function of the rice genome," *New Phytologist*, vol. 226, no. 1, pp. 44–49, 2020.

[72] M. Domínguez, E. Dugas, M. Benchouaia, B. Leduque, J. M. Jiménez-Gómez, V. Colot, and L. Quadrana, "The impact of transposable elements on tomato diversity," *Nature communications*, vol. 11, no. 1, pp. 1–11, 2020.

[73] D. Almojil, Y. Bourgeois, M. Falis, I. Hariyani, J. Wilcox, and S. Boissinot, "The structural, functional and evolutionary impact of transposable elements in eukaryotes," *Genes*, vol. 12, no. 6, p. 918, 2021.

[74] L. Sun, Y. Jing, X. Liu, Q. Li, Z. Xue, Z. Cheng, D. Wang, H. He, and W. Qian, "Heat stress-induced transposon activation correlates with 3d chromatin organization rearrangement in arabidopsis," *Nature communications*, vol. 11, no. 1, pp. 1–13, 2020.

[75] S. A. Montgomery, Y. Tanizawa, B. Galik, N. Wang, T. Ito, T. Mochizuki, S. Akimcheva, J. L. Bowman, V. Cognat, L. Maréchal-Drouard, *et al.*, "Chromatin organization in early land plants reveals an ancestral association between h3k27me3, transposons, and constitutive heterochromatin," *Current Biology*, vol. 30, no. 4, pp. 573–588, 2020.

[76] S. Alseekh, F. Scossa, and A. R. Fernie, "Mobile transposable elements shape plant genome diversity," *Trends in Plant Science*, vol. 25, no. 11, pp. 1062–1064, 2020.

[77] S. Pimpinelli and L. Piacentini, "Environmental change and the evolution of genomes: Transposable elements as translators of phenotypic plasticity into genotypic variability," *Functional Ecology*, vol. 34, no. 2, pp. 428–441, 2020.

[78] S. Orozco-arias, J. Liu, R. T.-s. Id, D. Ceballos, D. Silva, D. Id, R. Ming, and R. Guyot, "Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics," *Biology*, 2018.

[79] L. van Dorp, C. J. Houldcroft, D. Richard, and F. Balloux, "Covid-19, the first pandemic in the post-genomic era," *Current Opinion in Virology*, 2021.

[80] T. Flutre, E. Duprat, C. Feuillet, and H. Quesneville, "Considering transposable element diversification in de novo annotation approaches," *PloS one*, vol. 6, no. 1, p. e16526, 2011.

[81] S. Orozco-Arias, G. Isaza, and R. Guyot, "Retrotransposons in plant genomes: structure, identification, and classification through bioinformatics and machine learning," *International journal of molecular sciences*, vol. 20, no. 15, p. 3837, 2019.

[82] S. Ou, W. Su, Y. Liao, K. Chougule, J. R. Agda, A. J. Hellinga, C. S. B. Lugo, T. A. Elliott, D. Ware, T. Peterson, N. Jiang, C. N. Hirsch, and M. B. Hufford, "Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline," *Genome Biology*, vol. 20, no. 1, pp. 1–18, 2019.

[83] D. R. Hoen, G. Hickey, G. Bourque, J. Casacuberta, R. Cordaux, C. Feschotte, A.-S. Fiston-Lavier, A. Hua-Van, R. Hubley, A. Kapusta, *et al.*, "A call for benchmarking transposable element annotation methods," *Mobile DNA*, vol. 6, no. 1, pp. 1–9, 2015.

[84] K. A. Shastry and H. Sanjay, "Machine learning for bioinformatics," in *Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications*, pp. 25–39, Springer, 2020.

[85] E. Naresh, B. V. Kumar, S. P. Shankar, *et al.*, "Impact of machine learning in bioinformatics research," in *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, pp. 41–62, Springer, 2020.

[86] I.-C. Giassa and P. Alexiou, "Bioinformatics and machine learning approaches to understand the regulation of mobile genetic elements," *Biology*, vol. 10, no. 9, p. 896, 2021.

[87] E. Routhier, A. Bin Kamruddin, and J. Mozziconacci, "keras_dna: a wrapper for fast implementation of deep learning models in genomics," *Bioinformatics*, vol. 37, no. 11, pp. 1593–1594, 2021.

[88] W. Kopp, R. Monti, A. Tamburrini, U. Ohler, and A. Akalin, "Deep learning for genomics using janggu," *Nature communications*, vol. 11, no. 1, pp. 1–7, 2020.

[89] A. Kashfeen and L. McMillan, "Frontier: finding the boundaries of novel transposable element insertions in genomes," in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–10, 2021.

[90] M. Panta, A. Mishra, M. T. Hoque, and J. Atallah, "Classifyte: a stacking-based prediction of hierarchical classification of transposable elements," *Bioinformatics*, 2021.

[91] K. Riehl, C. Riccio, E. A. Miska, and M. Hemberg, "Transposonultimate: software for transposon classification, annotation and detection," *bioRxiv*, 2021.

[92] S. Orozco-Arias, G. Isaza, R. Guyot, and R. Tabares-soto, "A systematic review of the application of machine learning in the detection and classification of transposable elements," *Peerj*, vol. 7, p. 18311, 2019.

[93] C. Ma, H. H. Zhang, and X. Wang, "Machine learning for Big Data analytics in plants," *Trends in Plant Science*, vol. 19, no. 12, pp. 798–808, 2014.

[94] F. K. Nakano, W. J. Pinto, G. L. Pappa, and R. Cerri, "Top-down strategies for hierarchical classification of transposable elements with neural networks," in *Proceedings of the International Joint Conference on Neural Networks*, vol. 2017-May, (Anchorage, AK, United states), pp. 2539–2546, 2017.

[95] E. A. Bell, C. L. Butler, C. Oliveira, S. Marburger, L. Yant, and M. I. Taylor, "Transposable element annotation in non-model species: The benefits of species-specific repeat libraries using semi-automated edta and deepte de novo pipelines," *Molecular Ecology Resources*, 2021.

[96] T. Flutre, E. Permal, and H. Quesneville, "Transposable element annotation in completely sequenced eukaryote genomes," in *Plant Transposable Elements*, pp. 17–39, Springer, 2012.

[97] C. Feschotte, N. Jiang, and S. R. Wessler, "Plant transposable elements: Where genetics meets genomics," *Nature Reviews Genetics*, vol. 3, pp. 329–341, may 2002.

[98] J. F. Pereira and P. R. Ryan, "The role of transposable elements in the evolution of aluminium resistance in plants," *Journal of Experimental Botany*, vol. 70, pp. 41–54, 10 2018.

[99] M. Sahebi, M. M. Hanafi, A. J. van Wijnen, D. Rice, M. Y. Rafii, P. Azizi, M. Osman, S. Taheri, M. F. A. Bakar, M. N. M. Isa, and Others, "Contribution of transposable elements in the plant's genome," *Gene*, vol. 665, pp. 155–166, 2018.

[100] B. McClintock, "The Significance of Responses of the Genome to Challenge," *Science*, vol. 226, no. 4676, pp. 792–801, 1984.

[101] V. Horváth, M. Merenciano, and J. González, "Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response," *Trends in Genetics*, vol. 33, no. 11, pp. 832–841, 2017.

[102] C. A. Thomas, "THE GENETIC ORGANIZATION OF CHROMOSOMES," *Annual Review of Genetics*, vol. 5, no. 1, pp. 237–256, 1971.

[103] T. P. Michael, "Plant genome size variation: bloating and purging DNA," *Briefings in Functional Genomics*, vol. 13, pp. 308–317, 03 2014.

[104] X. Dai, H. Wang, H. Zhou, L. Wang, J. DvoÅ™Ã¡k, J. L. Bennetzen, and H.-G. MÃ¼ller, "Birth and Death of LTR-Retrotransposons in Aegilops tauschii," *Genetics*, vol. 210, pp. 1039–1051, 08 2018.

[105] S.-I. Lee and N.-S. Kim, "Transposable Elements and Genome Size Variations in Plants," *Genomics & Informatics*, vol. 12, no. 3, p. 87, 2014.

[106] E. R. Havecker, X. Gao, and D. F. Voytas, "The diversity of LTR retrotransposons," *Genome biology*, vol. 5, no. 6, p. 225, 2004.

[107] J. M. Casacuberta, S. Jackson, O. Panaud, M. Purugganan, and J. Wendel, "Evolution of Plant Phenotypes, from Genomes to Traits," *G3 Genes—Genomes—Genetics*, vol. 6, pp. 775–778, 04 2016.

[108] C. M. Bergman and H. Quesneville, "Discovering and detecting transposable elements in genome sequences," *Briefings in Bioinformatics*, vol. 8, no. 6, pp. 382–392, 2007.

[109] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Transfer learning for time series classification," in *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1367–1376, 2018.

[110] J.-C. Charr, A. Garavito, C. Guyeux, D. Crouzillat, P. Descombes, C. Fournier, S. N. Ly, E. N. Raharimalala, J.-J. Rakotomalala, P. Stoffelen, S. Janssens, P. Hamon, and R. Guyot, "Complex evolutionary history of coffees revealed by full plastid genomes and 28,800 nuclear SNP analyses, with particular emphasis on Coffea canephora (Robusta coffee)," *Molecular Phylogenetics and Evolution*, vol. 151, p. 106906, 2020.

[111] R. Guyot, T. Darré, M. Dupeyron, A. de Kochko, S. Hamon, E. Couturon, D. Crouzillat, M. Rigoreau, J.-J. Rakotomalala, N. E. Raharimalala, S. D. Akaffou, and P. Hamon, "Partial sequencing reveals the transposable element composition of Coffea genomes and provides evidence for distinct evolutionary stories.," *Molecular genetics and genomics : MGG*, vol. 291, pp. 1979–90, oct 2016.

[112] R. Guyot, P. Hamon, E. Couturon, N. Raharimalala, J.-J. Rakotomalala, S. Lakkanna, S. Sabatier, A. Affouard, and P. Bonnet, "WCSdb: a database of wild Coffea species," *Database*, vol. 2020, 11 2020. baaa069.

[113] P. Lashermes, V. Paczek, P. Trouslot, M. Combes, E. Couturon, and A. Charrier, "Brief communication. Single-locus inheritance in the allotetraploid Coffea arabica L. and interspecific Hybrid C. arabica X C. canephora," *Journal of Heredity*, vol. 91, pp. 81–85, 01 2000.

[114] P. Hamon, C. E. Grover, A. P. Davis, J.-J. Rakotomalala, N. E. Raharimalala, V. A. Albert, H. L. Sreenath, P. Stoffelen, S. E. Mitchell, E. Couturon, S. Hamon, A. de Kochko, D. Crouzillat, M. Rigoreau, U. Sumirat, S. Akaffou, and R. Guyot, "Genotyping-by-sequencing provides the first well-resolved phylogeny for coffee (Coffea) and insights into the evolution of caffeine content in its species: GBS coffee phylogeny and the evolution of caffeine content," *Molecular Phylogenetics and Evolution*, vol. 109, pp. 351–361, 2017.

[115] N. Raharimalala, S. Rombauts, A. McCarthy, A. Garavito, S. Orozco-Arias, L. Bellanger, A. Y. Morales-Correa, S. Froger, S. Michaux, V. Berry, S. Metairon, C. Fournier, M. Lepelley, L. Mueller, E. Couturon, P. Hamon, J.-J. Rakotomalala, P. Descombes, R. Guyot, and D. Crouzillat, "The absence of the caffeine synthase gene is involved in the naturally decaffeinated status of Coffea humblotiana, a wild species from Comoro archipelago," *Scientific Reports*, vol. 11, no. 1, pp. 1–14, 2021.

[116] J.-C. Charr, A. Garavito, C. Guyeux, D. Crouzillat, P. Descombes, C. Fournier, S. N. Ly, E. N. Raharimalala, J.-J. Rakotomalala, P. Stoffelen, *et al.*, "Complex evolutionary history

of coffees revealed by full plastid genomes and 28,800 nuclear snp analyses, with particular emphasis on coffea canephora (robusta coffee)," *Molecular Phylogenetics and Evolution*, vol. 151, p. 106906, 2020.

[117] P. Hamon, P. O. Duroy, C. Dubreuil-Tranchant, P. Mafra D'Almeida Costa, C. Duret, N. J. Razafinarivo, E. Couturon, S. Hamon, A. De Kochko, V. Poncet, and R. Guyot, "Two novel Ty1-copia retrotransposons isolated from coffee trees can effectively reveal evolutionary relationships in the Coffea genus (Rubiaceae)," *Molecular Genetics and Genomics*, vol. 285, no. 6, pp. 447–460, 2011.

[118] M. Dupeyron, R. F. de Souza, P. Hamon, A. de Kochko, D. Crouzillat, E. Couturon, D. S. Domingues, and R. Guyot, "Distribution of Divo in Coffea genomes, a poorly described family of angiosperm LTR-Retrotransposons," *Molecular Genetics and Genomics*, vol. 292, pp. 741–754, aug 2017.

[119] A. V. Zimin, G. MarÃ§ais, D. Puiu, M. Roberts, S. L. Salzberg, and J. A. Yorke, "The MaSuRCA genome assembler," *Bioinformatics*, vol. 29, pp. 2669–2677, 08 2013.

[120] M. Seppey, M. Manni, and E. M. Zdobnov, "BUSCO: Assessing genome assembly and annotation completeness," in *Methods in Molecular Biology*, vol. 1962, pp. 227–245, Humana Press Inc., 2019.

[121] E. M. McCarthy and J. F. McDonald, "LTR STRUC: A novel search and identification program for LTR retrotransposons," *Bioinformatics*, vol. 19, no. 3, pp. 362–367, 2003.

[122] S. Orozco-Arias, M. S. Candamil-Cortés, P. A. Jaimes, J. S. Piña, R. Tabares-Soto, R. Guyot, and G. Isaza, "K-mer-based machine learning method to classify LTR-retrotransposons in plant genomes," *PeerJ*, vol. 9, p. e11456, may 2021.

[123] N. Chen, "Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences," *Current Protocols in Bioinformatics*, vol. 5, no. 1, pp. 4.10.1–4.10.14, 2004.

[124] R. C. Team, "R: a language and environment for statistical 688 computing," *Vienna: R Foundation*, 2016.

[125] I. Letunic and P. Bork, "Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation," *Nucleic Acids Research*, vol. 49, pp. W293–W296, 04 2021.

[126] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature methods*, vol. 9, no. 4, p. 357, 2012.

[127] N. J. Razafinarivo, J. J. Rakotomalala, S. C. Brown, M. Bourge, S. Hamon, A. de Kochko, V. Poncet, C. Dubreuil-Tranchant, E. Couturon, R. Guyot, and P. Hamon, "Geographical gradients in the genome size variation of wild coffee trees (Coffea) native to Africa and Indian Ocean islands," *Tree Genetics and Genomes*, vol. 8, no. 6, pp. 1345–1358, 2012.

[128] C. E. Grover and J. F. Wendel, "Recent Insights into Mechanisms of Genome Size Change in Plants," *Journal of Botany*, vol. 2010, pp. 1–8, 2010.

[129] R. J. Schley, J. Pellicer, X.-J. Ge, C. Barrett, S. Bellot, M. S. Guignard, P. Novák, J. Suda, D. Fraser, W. J. Baker, S. Dodsworth, J. ríMacas, A. R. Leitch, and I. J. Leitch, "The Ecology of Palm Genomes: Repeat-associated genome size expansion is constrained by aridity," *bioRxiv*, 2021.

[130] K. Y. Yip, C. Cheng, and M. Gerstein, "Machine learning and genome annotation: a match meant to be?," *Genome biology*, vol. 14, no. 5, pp. 1–10, 2013.

[131] C. Xu and S. A. Jackson, "Machine learning and complex biological data," 2019.

[132] N. Yu, X. Guo, F. Gu, and Y. Pan, "Dna as x: An information-coding-based model to improve the sensitivity in comparative gene analysis," in *International Symposium on Bioinformatics Research and Applications*, pp. 366–377, Springer, 2015.

[133] M. Akhtar, J. Epps, and E. Ambikairajah, "Signal processing in sequence analysis: advances in eukaryotic gene prediction," *IEEE journal of selected topics in signal processing*, vol. 2, no. 3, pp. 310–321, 2008.

[134] G. Kauer and H. Blöcker, "Applying signal theory to the analysis of biomolecules," *Bioinformatics*, vol. 19, no. 16, pp. 2016–2021, 2003.

[135] G. L. Rosen, *Signal processing for biologically-inspired gradient source localization and DNA sequence analysis*. Georgia Institute of Technology, 2006.

[136] A. C. H. Choong and N. K. Lee, "Evaluation of convolutionary neural networks modeling of dna sequences using ordinal versus one-hot encoding method," in *2017 International Conference on Computer and Drone Applications (IConDA)*, pp. 60–65, IEEE, 2017.

[137] D. Ceballos, D. López-Álvarez, G. Isaza, R. Tabares-Soto, S. Orozco-Arias, and C. D. Ferrin, "A machine learning-based pipeline for the classification of ctx-m in metagenomics samples," *Processes*, vol. 7, no. 4, p. 235, 2019.

[138] Z. Lv, H. Ding, L. Wang, and Q. Zou, "A convolutional neural network using dinucleotide one-hot encoder for identifying dna n6-methyladenine sites in the rice genome," *Neurocomputing*, vol. 422, pp. 214–221, 2021.

[139] F. Wang, P. Chainani, T. White, J. Yang, Y. Liu, and B. Soibam, "Deep learning identifies genome-wide dna binding sites of long noncoding rnas," *RNA biology*, vol. 15, no. 12, pp. 1468–1476, 2018.

[140] D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, and J. Snoek, "Sequential regulatory activity prediction across chromosomes with convolutional neural networks," *Genome research*, vol. 28, no. 5, pp. 739–750, 2018.

[141] D. Mapleson, G. Garcia Accinelli, G. Kettleborough, J. Wright, and B. J. Clavijo, "Kat: a k-mer analysis toolkit to quality control ngs datasets and genome assemblies," *Bioinformatics*, vol. 33, no. 4, pp. 574–576, 2017.

[142] F. P. Breitwieser, D. Baker, and S. L. Salzberg, "Krakenuniq: confident and fast metagenomics classification using unique k-mer counts," *Genome biology*, vol. 19, no. 1, pp. 1–10, 2018.

[143] D. R. Zerbino and E. Birney, "Velvet: algorithms for de novo short read assembly using de bruijn graphs," *Genome research*, vol. 18, no. 5, pp. 821–829, 2008.

[144] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol, "Abyss: a parallel assembler for short read sequence data," *Genome research*, vol. 19, no. 6, pp. 1117–1123, 2009.

[145] H. Sun, J. Ding, M. Piednoël, and K. Schneeberger, "findgse: estimating genome size variation within human and arabidopsis using k-mer frequencies," *Bioinformatics*, vol. 34, no. 4, pp. 550–557, 2018.

[146] A. L. Price, N. C. Jones, and P. A. Pevzner, "De novo identification of repeat families in large genomes," *Bioinformatics*, vol. 21, no. suppl_1, pp. i351–i358, 2005.

[147] B. Z. Santos, G. T. Pereira, F. K. Nakano, and R. Cerri, "Strategies for selection of positive and negative instances in the hierarchical classification of transposable elements," in *2018 7th Brazilian Conference on Intelligent Systems (BRACIS)*, pp. 420–425, IEEE, 2018.

[148] W. Ashlock and S. Datta, "Distinguishing endogenous retroviral ltrs from sine elements using features extracted from evolved side effect machines," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 9, no. 6, pp. 1676–1689, 2012.

[149] F. Liu, H. Li, C. Ren, X. Bo, and W. Shu, "Pedla: predicting enhancers with a deep learning-based algorithmic framework," *Scientific reports*, vol. 6, no. 1, pp. 1–14, 2016.

[150] J. T. Cuperus, B. Groves, A. Kuchina, A. B. Rosenberg, N. Jojic, S. Fields, and G. Seelig, "Deep learning of the regulatory grammar of yeast 5' untranslated regions from $500,000$ random sequences," *Genome research*, vol. 27, no. 12, pp. 2015–2024, 2017.

[151] R. S. Roy, D. Bhattacharya, and A. Schliep, "Turtle: Identifying frequent k-mers with cache-efficient algorithms," *Bioinformatics*, vol. 30, no. 14, pp. 1950–1957, 2014.

[152] L. Pellegrina, C. Pizzi, and F. Vandin, "Fast approximation of frequent k-mers and applications to metagenomics," *Journal of Computational Biology*, vol. 27, no. 4, pp. 534–549, 2020.

[153] P. Melsted and J. K. Pritchard, "Efficient counting of k-mers in dna sequences using a bloom filter," *BMC bioinformatics*, vol. 12, no. 1, pp. 1–7, 2011.

[154] F. Doshi-Velez and B. Kim, "Considerations for evaluation and generalization in interpretable machine learning," in *Explainable and interpretable models in computer vision and machine learning*, pp. 3–17, Springer, 2018.

[155] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[156] S.-S. Zhou, X.-M. Yan, K.-F. Zhang, H. Liu, J. Xu, S. Nie, K.-H. Jia, S.-Q. Jiao, W. Zhao, Y.-J. Zhao, *et al.*, "A comprehensive annotation dataset of intact ltr retrotransposons of 300 plant genomes," *Scientific Data*, vol. 8, no. 1, pp. 1–9, 2021.

[157] S. Ou and N. Jiang, "Ltr_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons," *Plant physiology*, vol. 176, no. 2, pp. 1410–1422, 2018.

[158] E. Lerat, "Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs," *Heredity*, vol. 104, no. 6, pp. 520–533, 2010.

[159] F. M. You, S. Cloutier, Y. Shan, and R. Ragupathy, "Ltr annotator: automated identification and annotation of ltr retrotransposons in plant genomes," *International Journal of Bioscience, Biochemistry and Bioinformatics*, vol. 5, no. 3, p. 165, 2015.

[160] A. C. Wacholder, C. Cox, T. J. Meyer, R. P. Ruggiero, V. Vemulapalli, A. Damert, L. Carbone, and D. D. Pollock, "Inference of transposable element ancestry," *PLoS genetics*, vol. 10, no. 8, p. e1004482, 2014.

[161] P. Neumann, P. Novák, N. Hoštáková, and J. Macas, "Systematic survey of plant ltr-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification," *Mobile DNA*, vol. 10, no. 1, pp. 1–17, 2019.

[162] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[163] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pp. 1322–1328, IEEE, 2008.

[164] Z. Xu and H. Wang, "Ltr_finder: an efficient tool for the prediction of full-length ltr retrotransposons," *Nucleic acids research*, vol. 35, no. suppl_2, pp. W265–W268, 2007.

[165] S. Ou and N. Jiang, "Ltr_finder_parallel: parallelization of ltr_finder enabling rapid identification of long terminal repeat retrotransposons," *Mobile DNA*, vol. 10, no. 1, pp. 1–3, 2019.

[166] G. Chandan, A. Jain, H. Jain, *et al.*, "Real time object detection and tracking using deep learning and opencv," in *2018 International Conference on inventive research in computing applications (ICIRCA)*, pp. 1305–1308, IEEE, 2018.

[167] A. E. Wahabi, I. H. Baraka, S. Hamdoune, and K. E. Mokhtari, "Detection and control system for automotive products applications by artificial vision using deep learning," in *International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 224–241, Springer, 2019.

[168] A. Raghunandan, P. Raghav, H. R. Aradhya, *et al.*, "Object detection algorithms for video surveillance applications," in *2018 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0563–0568, IEEE, 2018.

[169] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

[170] D. Ellinghaus, S. Kurtz, and U. Willhoeft, "Ltrharvest, an efficient and flexible software for de novo detection of ltr retrotransposons," *BMC bioinformatics*, vol. 9, no. 1, pp. 1–14, 2008.

[171] J. D. Valencia and H. Z. Girgis, "Ltrdetector: a tool-suite for detecting long terminal repeat retrotransposons de-novo," *BMC genomics*, vol. 20, no. 1, pp. 1–14, 2019.

[172] M. Biryukov and K. Ustyantsev, "Darts: An algorithm for domain-associated retrotransposon search in genome assemblies," *Genes*, vol. 13, no. 1, 2022.

[173] H. Jung, M.-S. Jeon, M. Hodgett, P. Waterhouse, and S.-i. Eyun, "Comparative evaluation of genome assemblers from long-read sequencing for plants and crops," *Journal of Agricultural and Food Chemistry*, vol. 68, no. 29, pp. 7670–7677, 2020. PMID: 32530283.

[174] Y. Chernyavskaya, X. Zhang, J. Liu, and J. Blackburn, "Long-read sequencing of the zebrafish genome reorganizes genomic architecture," *BMC Genomics*, vol. 23, no. 1, pp. 1–13, 2022.

[175] Y. Suzuki and S. Morishita, "The time is ripe to investigate human centromeres by long-read sequencing," *DNA Research*, vol. 28, 10 2021. dsab021.

[176] Y. Jiang, *Repetitive DNA sequence assembly.* PhD thesis, Deakin University, 2017.

[177] T. J. Treangen and S. L. Salzberg, "Repetitive DNA and next-generation sequencing: Computational challenges and solutions," *Nature Reviews Genetics*, vol. 13, no. 1, pp. 36–46, 2012.

[178] S. Lian, Y. Tu, Y. Wang, X. Chen, and L. Wang, "A repetitive sequence assembler based on next-generation sequencing," *Genetics and Molecular Research*, vol. 15, no. 3, pp. 1–13, 2016.

[179] M. Zytnicki, E. Akhunov, and H. Quesneville, " Tedna: a transposable element de novo assembler ," *Bioinformatics*, vol. 30, pp. 2656–2658, 06 2014.

[180] C. Chu, R. Nielsen, and Y. Wu, "REPdenovo: Inferring De Novo Repeat Motifs from Short Sequence Reads," *PLOS ONE*, vol. 11, no. 3, pp. 1–17, 2016.

[181] R. M. Nowak, "Genome Assembler for Repetitive Sequences," in *Information Technologies in Biomedicine* (E. Piketka and J. Kawa, eds.), (Berlin, Heidelberg), pp. 422–429, Springer Berlin Heidelberg, 2012.

[182] E. Bao, F. Xie, C. Song, and D. Song, "FLAS: fast and high-throughput algorithm for PacBio long-read self-correction," *Bioinformatics*, vol. 35, pp. 3953–3960, 03 2019.

[183] E. L. van Dijk, Y. Jaszczyszyn, D. Naquin, and C. Thermes, "The Third Revolution in Sequencing Technology," *Trends in Genetics*, vol. 34, no. 9, pp. 666–681, 2018.

[184] H. Jung, C. Winefield, A. Bombarely, P. Prentis, and P. Waterhouse, "Tools and Strategies for Long-Read Sequencing and De Novo Assembly of Plant Genomes," *Trends in Plant Science*, vol. 24, no. 8, pp. 700–724, 2019.

[185] S. Shahid and R. K. Slotkin, "The current revolution in transposable element biology enabled by long reads," *Current Opinion in Plant Biology*, vol. 54, pp. 49–56, 2020.

[186] R.-G. Zhang, Z.-X. Wang, S. Ou, and G.-Y. Li, "TEsorter: lineage-level classification of transposable elements using conserved protein domains," *bioRxiv*, 2019.