UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Escola d'Enginyeria Agroalimentària
i de Biosistemes de Barcelona

CRAG
CENTRE FOR RESEARCH
IN AGRICULTURAL GENOMICS

Analysis of transposon dynamics in *Prunus* species using a pangenome approach

Bachelor's thesis

Biological systems engineering

Author: Daniel Serrano Jorcano

Tutor: Joan Simó Cruanyes

External tutor (CRAG): Raúl Castanera

9 / July / 2021

Abstract

Transposable elements (TEs) are repetitive DNA sequences found in all prokaryotic and eukaryotic organisms. They are able to move from one genome location to another and replicate in the genome, producing mutations and rearrangements. Due to their capacity to generate genetic variability, TE polymorphisms are starting to be studied in crop plants as they can lead to phenotypic variations of agronomic importance. TE detection is performed by computational analysis of genomic data, and the increasing availability of high-quality genome assemblies allows to study their dynamics using pangenome approaches. The *Prunus* genus is one of the most economically important genera which includes many cultivated trees, such as almonds or peaches, and several *Prunus* genome assemblies have been released recently.

We have used this public genomic data to identify TEs "*de novo*" in 7 trees of this genus belonging to 6 different species: *P. armeniaca*, *P. avium*, *P. dulcis*, *P. mira*, *P. persica* and *P. salicina*. We have identified thousands of intact transposons, which encode all the structural hallmarks and enzymatic domains necessary for its transposition. By using a pangenome approach we have built a database of *Prunus* TE families and analyzed their differential presence in the six *Prunus* species. We discovered a recent burst of TE amplification in *P. salicina,* mainly driven by LTR-retrotransposon activity occurred during the last million years. We used our developed *Prunus* database to search for TE polymorphisms between two peach varieties (Lovell and Earlygold), identifying 801 potential TE insertions specific to Lovell variety. Many of these TE insertions potentially affect gene activity, and we validated two of them by molecular approaches.

The present work will enable the study of *Prunus* TE families using a common, structured classification and will contribute to the understanding of the potential of TE polymorphisms in the generation of intraspecific variability linked to agricultural traits of interest, such as the resistance to drought, pests, as well as fruit quality or yield.

Resum

Els elements transponibles (TE) són seqüències d'ADN repetitives que es troben en tots els organismes procariotes i eucariotes. Tenen capacitat de desplaçar-se d'una localització genòmica a una altra i replicar-se, produint mutacions i reordenacions genètiques. A causa de la seva capacitat de generar variabilitat genètica, els polimorfismes dels TE estan començant a estudiar-se en les plantes de cultiu, ja que poden donar lloc a variacions fenotípiques d'importància agronòmica. La detecció dels TE es realitza mitjançant l'anàlisi computacional de les dades genòmiques, i la creixent disponibilitat d'assemblatges genòmics d'alta qualitat permet estudiar la seva dinàmica mitjançant enfocaments pangenòmics. El gènere *Prunus* és un dels gèneres més importants des del punt de vista econòmic que inclou molts arbres cultivables, com l'ametller o el presseguer, i dels quals recentment s'han publicat diversos acoblaments de genomes de diverses espècies d'aquest gènere.

Hem utilitzat aquestes dades genòmiques públiques per identificar TE *"de novo"* en 7 arbres d'aquest gènere que pertanyen a 6 espècies diferents: *P. armeniaca, P. avium, P. dulcis, P. mira, P. persica* i *P. salicina*. Hem identificat milers de transposons intactes, que codifiquen totes les característiques estructurals i els dominis enzimàtics necessaris per a la seva transposició. Utilitzant un enfocament pangenòmic hem construït una base de dades de famílies de TE de *Prunus* i hem analitzat la seva presència diferencial en les sis espècies de *Prunus*. Observem un recent augment d'amplificació de TE en *P. salicina*, principalment impulsat per l'activitat dels LTR-retrotransposons durant l'últim milió d'anys. Utilitzem la nostra base de dades de *Prunus* per buscar polimorfismes creats per TE entre dues varietats de presseguer (Lovell i Earlygold), identificant 801 potencials insercions de TE específics de la varietat Lovell. Moltes d'aquestes insercions de TE afecten potencialment a l'activitat del gen on es troben inserits, i validem dues d'elles mitjançant enfocaments moleculars.

El present treball permetrà l'estudi de les famílies de TE de *Prunus* utilitzant una classificació comuna i estructurada i contribuirà a la comprensió del potencial dels polimorfismes de TE en la generació de variabilitat intraespecífica lligada a trets agrícoles d'interès, com la resistència a la sequera , a les plagues, així com a la qualitat o el rendiment de la fruita.

Resumen

Los elementos transponibles (TE) son secuencias de ADN repetitivas que se encuentran en todos los organismos procariotas y eucariotas. Son capaces de desplazarse de una localización genómica a otra y replicarse en el genoma, produciendo mutaciones y reordenamientos. Debido a su capacidad de generar variabilidad genética, los polimorfismos de los TE están empezando a estudiarse en las plantas de cultivo, ya que pueden dar lugar a variaciones fenotípicas de importancia agronómica. La detección de los TE se realiza mediante el análisis computacional de los datos genómicos, y la creciente disponibilidad de ensamblajes genómicos de alta calidad permite estudiar su dinámica mediante enfoques pangenómicos. El género *Prunus* es uno de los géneros más importantes desde el punto de vista económico que incluye muchos árboles cultivables, como el almendro o el melocotonero, y recientemente se han publicado varios ensamblajes de genomas de múltiples especies de este género.

Hemos utilizado estos datos genómicos públicos para identificar TEs "*de novo"* en 7 árboles de este género pertenecientes a 6 especies diferentes: *P. armeniaca, P. avium, P. dulcis, P. mira, P. persica* y *P. salicina*. Hemos identificado miles de transposones intactos, que codifican todas las características estructurales y los dominios enzimáticos necesarios para su transposición. Utilizando un enfoque pangenómico hemos construido una base de datos de familias de TE de *Prunus* y hemos analizado su presencia diferencial en las seis especies de *Prunus*. Descubrimos un reciente estallido de amplificación de TE en *P. salicina*, principalmente impulsado por la actividad de LTR-retrotransposones ocurrido durante el último millón de años. Utilizamos nuestra base de datos de *Prunus* para buscar polimorfismos creados por TE entre dos variedades de melocotón (Lovell y Earlygold), identificando 801 potenciales inserciones de TE específicos de la variedad Lovell. Muchas de estas inserciones de TE afectan potencialmente a la actividad del gen donde se encuentra insertados, y validamos dos de ellas mediante enfoques moleculares.

El presente trabajo permitirá el estudio de las familias de TE de *Prunus* utilizando una clasificación común y estructurada y contribuirá a la comprensión del potencial de los polimorfismos de TE en la generación de variabilidad intraespecífica ligada a rasgos agrícolas de interés, como la resistencia a la sequía, a las plagas, así como a la calidad o el rendimiento de la fruta.

Summary

Index of Figures

Index of tables

Symbols and acronyms

Ac	Activator
aORF	additional open reading frames
BC	Before Crhist
bp	Base pair long
BUSCO	Benchmarking Universal Single-Copy Orthologies
CRAG	Centre for Research in Agricultural Genomics
Dc	Dissociation
DIRS	Dictyostelium intermediate repeat sequence
DNA	Deoxyribonucleic acid
dNTPs	nucleotides
EDTA	Extenside de-novo TE annotator
ORF	Open reading frames
Gb	Gigabase
HPC	High performance computing cluster
IRTA	Institute of Agrifood Research and Technology
LAI	LTR Assembly Index
LINEs	Long interspersed nuclear elements
LTR	Lng termial repeats
Mb	Megabase
MITEs	Miniature inverted-repeats transposable elements
MM	Molecular marker
NGS	Next-Generation Sequencing
Par	*Prunus armeniaca*
Pav	*Prunus avium* Titeon
PCR	Polymerase Chain Reaction
PdL	*Prunus dulcis* Lauranne
PdT	*Prunus dulcis* Texas
PLEs	Penelope
Pm	*Prunus mira*
Pp	*Prunus persica*

Ps	*Prunus salicina*
RNA	Ribonucleic acid
rpm	revolution per minutes
RT	Room temperature
SINEs	Short interspersed nuclear elements
TE	Transposable elements
TGS	Third-Generation Sequencing
TIRs	Terminal inverted repeats
TSD	Target site duplication
WGS	whole genome shotgun

Acknowledgements

I would like to thank the research group "Structure and evolution of plant genomes" located in the research centre CRAG, for giving me the opportunity to do my TFG with them and also to experience for the first time what it is like to be part of a research group.

Specifically, thanks to Dr. Raúl Castanera for the supervision and tutoring of this work and for his immense help, teaching and motivation I have received during the last four months. For sharing his knowledge and passion he feels for bioinformatics in an easy and understandable way that has allowed me to be comfortable in a field I wasn't familiar with, and for all the corrections he have done these last days to improve the work.

Thanks to Carlos de Tomás for his great help, for teaching me how to move around and become more independent in the laboratory, as well as for teaching me how to handle all the laboratory equipment and techniques necessary for the molecular validation.

Thanks to Joan Simó for being my UPC tutor, for accepting my TFG and helping me through this process.

Finally, I would like to thank my university friends for the mutual support we have been giving each other these last few days in order to be able to finish and to give the maximum of ourselves in this final degree work, as well as to my family for being by my side all the time.

Background

This project has been carried out at the 'Centre for research in agricultural genomics' (CRAG) within the research group 'Structure and evolution of plant genomes', led by Dr. Josep M. Casacuberta and Dr. Carlos Vicient. The general research line of the group explores the movement of transposons as a source of genetic variability in cultivated and model plants, mainly in rice, almond, peach, melon and *Physcomitrella*. One of the main objectives of the group is to determine the activity of transposons and the possible impact on the genome, which could generate new phenotypes of agronomic interest with the aim of finding useful alleles for genetic improvement. For this, the group uses bioinformatics and molecular approaches.

The present project falls in the framework of the determination of transposons on *Prunus* species and builds on the group´s experience on computational analyses of genomes and pangenomes.

This project has been entirely done by myself, with the help and supervision of Dr. Raúl Castanera (external tutor at CRAG) in the bioinformatics analyses and manuscript writing, and the supervision of Carlos de Tomás in the laboratory methods.

1. Introduction

1.1. Rosaceae family and *Prunus* genus

Rosaceae is one of the largest families in the order Rosales, is composed by herbs, shrubs and trees having a worldwide distribution. Rosaceae trees originated in regions extending from west Asia to the Caucasus (N.I Vavilov, 1951), and have a major presence in north-temperate regions (northern hemisphere) (Shulaev et al., 2008). Rosaceae members produce a wide distinctive type of fruits (i.e. drupes, pomes, achenes or drupetums), and some are very important crops, having a big impact on our economy (Ribeiro Serra, 2017; Xiang et al., 2017). Thus, the Rosaceae family has been the subject of numerous taxonomic and evolutionary studies (Potter et al., 2002).

Rosaceae family is comprised by up to 90-100 genera, 16 tribes and 3000 different species (Potter et al., 2002; Shulaev et al., 2008). According to the most recent classification done in 2007 using molecular evidences, Rosaceae members are divided into three subfamilies: Rosoideae, Dryadoideae, and Spiraeoideae. In 2011, based on the International Code of Nomenclature for Algae, Fungi and Plants (McNeill et al., 2012), Spiraeoideae was renamed as Amygdaloideae (Shulaev et al., 2008).

Prunus genus (*Prunus L.*), is found within Amygdaloideae subfamily and is composed by up to 200 species. The infrageneric classification of *Prunus* genus consists of five subgenera: *Amygdalus* (peaches and almonds), *Cerasus* (cherries), *Prunus* (plums), *Laurocerasus* (evergreen laurel-cherries), and *Padus* (deciduous bird-cherries) (Chin et al., 2014).

The most common species are cultivated peach [*Prunus persica* (L.) Batsch] and almond [*P. dulcis* (Mill.) D.A.Webb] as well as several wild relatives such as *P. mira* Koehene, *P. kansuensis* Rehd and *P. davidiana* (Carr.) Franch, all members of the A*mygdalus* subgenera (Badenes & Byrne, 2012). *Prunus* systematic classification is shown in **Figure 1** and *Prunus* phylogeny in **Figure 2**.

Kingdom: *Plantae*

Division: *Magnoliophyta*

Class: *Magnoliopsida*

Order: *Rosales*

Family: *Rosaceae*

Subfamily: *Amygdaloideae*

Tribe: *Amygdaleae*

Genus: *Prunus*

Figure 1. Representation of the systematic classification of *Prunus* genus, own elaboration.

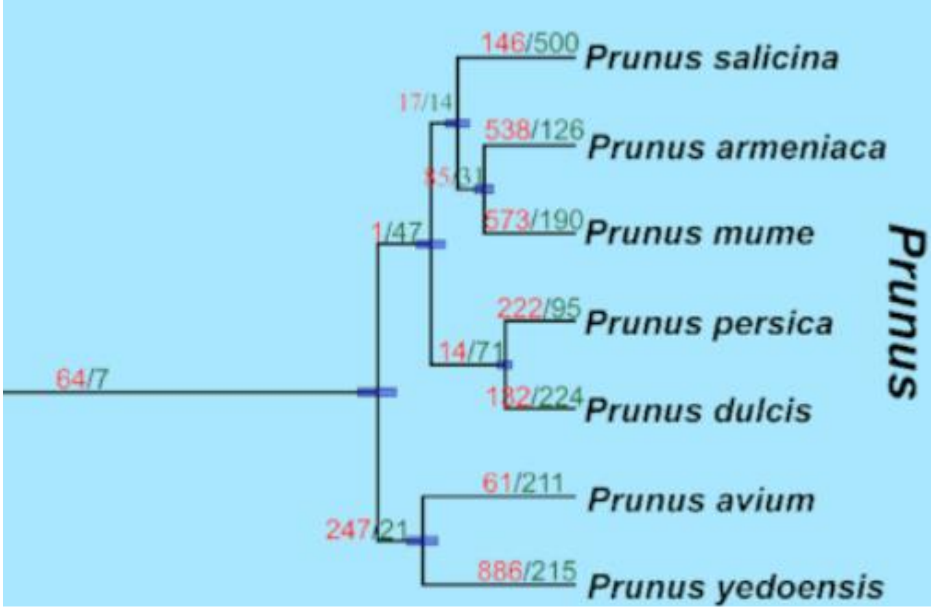


Figure 2: Phylogeny of *Prunus* genus. *Prunus mira*, which lacks to appear, would be near *Prunus persica* as they are close relatives, being both of them peaches. The numbers in red and green indicate the numbers of orthogroups that have expanded and contracted along particular branches, respectively. Extracted from *Liu et. al.* 2021.

1.1.1. *Prunus armeniaca*

P. armeniaca L. (**Figure 3A**), commonly known as apricot , is believed to have originated, and been domesticated in China (Hebei Province) around 2000 BC (Jiang et al., 2019)**.**

Apricots, temperate drupes closely related to peaches and plums are mainly cultivated for fresh market consumption, kernel production and for ornamental use. Mediterranean regions are the most important areas of apricot production (Jiang et al., 2019). The global production of apricots reached 4 million tonnes in 2019, being Turkey the major producer (21 %) of dried and fresh apricots (Asma & Ozturk, 2005) and Spain the 5th with 4 % of the global production (FAO 2019).

Apricot, with a relatively small genome but highly heterozygous, is mainly self-incompatible in China and central Asia, but is self-compatible in the Eastern Europe. *P. armeniaca* whole genome sequencing was obtained in 2019 (Jiang et al., 2019).

1.1.2. *Prunus avium*

Prunus avium (**Figure 3B**), or sweet cherry, was originated in Europe and western Asia, more specifically near the Black Sea and the Caspian Sea and nowadays are mainly cultivated for human consumption as well as for ornamental and for wood use (Tavaud et al., 2004; Wang et al., 2020).

In 2019, 2,6 million tonnes of cherries were produced worldwide, being Turkey the first producer with 25 % of the production (FAO 2019).

P. avium genome Sequencing was completed in 2017 (Shirasawa et al., 2017) but after, a new sequencing in 2020 (Wang et al., 2020) reached a better genome assembly.

1.1.3. *Prunus dulcis*

Prunus dulcis (**Figure 3C**), or most commonly known as Almond tree, is an ancient crop thought to have originated in west Asia as several wild species are found there (Martínez-Gómez et al. 2007). Wild almond tree nuts aren't edible as they are bitter, but through the domestication process, which started 8000 BC, sweet almonds trees have been selected (Ladizinsky, 1990) and are currently cultivated worldwide.

Almond has a diploid genome, is strictly allogamous tree and also self-incompatible (Delplancke et al. 2013). In 2018, the sequencing of whole genome of *P. dulcis* Texas (Alioto et al., 2020) was achieved, and a year later, in 2019, from *P. dulcis* Lauranne (Sánchez-Pérez et al., 2019).

1.1.4. *Prunus mira*

Prunus mira (**Figure 3D**), the Himalayan peach, originally from Qinghai-Tibet plateau (China), is thought to be one of the wild relative species of modern cultivated peach (*P. persica*) (Cao et al., 2020). It grows between 2000 m and 4000 m of altitude, having a strong tolerance to environmental stresses (Tian et al., 2015).

China, the major producer as they recollect the edible peach fruit for human consumption, as well as for Chinese medicine and for ornamental use due to their blossoming in Spring (Bao et al. 2017; Tian et al. 2015). Unfortunately, due to deforestation and other human environmental changes, *P. mira* population has been remarkably reduced and now is classified as an endangered species. *P. mira* has been sequenced for the first time in 2020 by a pangenome analysis (Cao et al., 2020).

1.1.5. *Prunus persica*

Peach botanical name is *Prunus persica (L.)*, **Figure 3E**. In the 19th century, peach origin was thought to be in Persia. Nowadays, the wild ancestor of the peach remains unknown, but due to the high genetic diversity found in China, and the discovery of an endocarp fossil of 2.5 million years old from *P. kunmingensis* (Su et al., 2015), we can assume that the Asian country is the origin of the wild, and also, of the domesticated peach (Zheng et al., 2014).

The cultivation of domesticated peach started in 4000 BC and was dispersed from China to the rest of the world. Firstly around all Asia, then to Europe via the ancient Silk Road through Persia (present day Iran) in the final centuries B.C, and from Europe to the Americas during the 16th century by the Spanish, Portuguese and French explorers (Chin et al., 2014; Yu et al., 2018).

With a global annual production of approximately 26 million tonnes, it is one of the most important trees within the P*runus* genus as it has a huge economic impact worldwide, in 2019 China (57%) was the first producer followed by Spain (6 %) and Italy (5,6 %) (FAO 2019).

Peach, a model fruit species for comparative and functional genomics, has a relatively small genome (230 Mb). It is a diploid specie, distributed in eight pairs of chromosomes and is genetically well known, with a self-compatible mating system (Cao et al. 2014). Peach genome was firstly sequenced in 2012 and further being improved in 2015 (Verde et al., 2013, 2017).

1.1.6. *Prunus salicina*

Prunus salicina (**Figure 3F**), commonly known as Japanese plum or Chinese plum, has a Chinese origin, but was firstly improved in Japan and after, to a much greater extent, in the United States (Liu et al., 2021). In 2020, first sequencing of its genome was carried out (Liu et al., 2021).

P. salicina is the predominant plum tree in the modern crops for commercial production. Taking into account all the plums subgenera, 12,6 million tonnes of plums are produced, being China the first producer with near 7 million tonnes of plums (56 %) (FAO 2019).

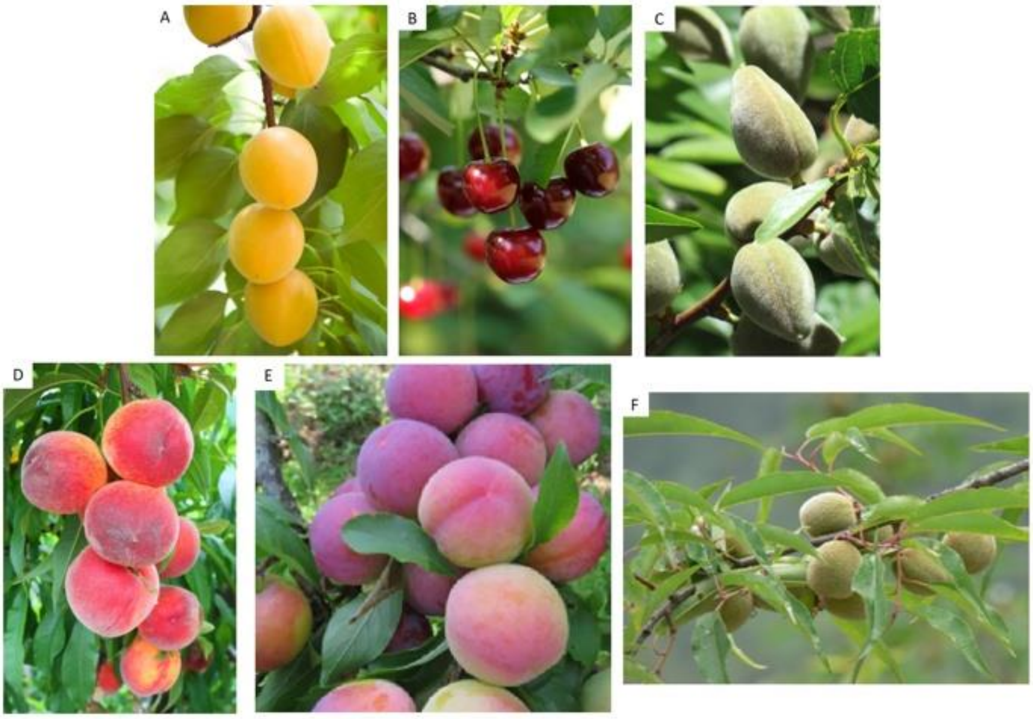


Figure 3. Examples of the *Prunus* fruits of A) *Prunus armeniaca*, B) *Prunus avium*, C) *Prunus dulcis*, D) *Prunus persica*, E) *Prunus salicina* and F) *Prunus mira*. Images obtained from GDR Rosaceae database (https://www.rosaceae.org/).

1.2. Evolution of DNA sequencing: from genes to genomes

The development of DNA sequencing technologies has made possible to determine the nucleotide order sequence within a DNA molecule, allowing us to obtain genomic information from short fragments (i.e. single genes regulatory regions) up to entire genomes (Tipu & Shabbir, 2015).

First DNA sequencing technique started in 1972 by the sequencing of the first complete protein-coding of MS2 bacteriophage and in the next year with the sequencing of the 'lac operator', a small sequence from *Escherichia coli* (Gilbert & Maxam, 1973).

But it was not until 1975 when some revolutionary techniques developed by Maxam and Gilbert (Maxam & Gilbert, 1977) with the chemical cleavage technique and two years later, by Sanger with Sanger's 'chain termination' technique (Sanger, F, 1997) enabled the progress of DNA sequencing.

For Sanger sequencing, four reactions are step up, all containing the four nucleotides (dNTPs) but only one dideoxynucleotide (ddATP, ddGTP, ddCTP, ddTTP) which will be the limiting factor, as it is a chain-terminating nucleotide. Further, it needs the presence of DNA polymerase, primers and DNA chain. As a result, we obtain different size DNA fragments, from 500 up to 1000 bp, that will be separated in PolyAcrylamide Gel Electrophoresis (PAGE) with fluorescence detection (Road, 1979; Zadesenets et al., 2017). In 1979, whole genome shotgun (WGS) appeared by achieving the yielding of Sanger sequencing by overlapping the sequencing fragments with the usage of sophisticated computer programs instead of doing it manually.

From 90s to the 2000s, WGS sequencing was technologically automated and optimized having an extensive use, playing an important role in the determination of the first sequencing eukaryotic genomes, including the human genome (Craig Venter et al., 2001; Lander et al., 2001), and others such as mouse (Waterston et al., 2002) and *Arabidopsis thaliana* (Poczai et al., 2000). But despite the improvement over previous techniques, WGS still required intensive work and also was very expensive and had low productivity (Tipu & Shabbir, 2015; Zadesenets et al., 2017).

This evidenced the need of new and improved technologies which led to the advent of second-generation technologies (Next-Generation Sequencing, NGS) for massively parallel DNA sequencing (8). Three NGS platforms are commonly used: Roche/454 FLX (Margulies et al., 2005), Illumina/Solexa Genome Analyzer with a sequencing-by-synthesis approach (Bentley, 2006) and Applied Biosystems SOLiD^{TM} System.

Although NGS have allowed a more in-depth knowledge on many researches fields and enabled the sequencing of new genomes, the techniques used had some problems due to short length of reads produced, normally ranging from 75 to 300 bp. This makes difficult to assembly highly repetitive regions (i.e. centromers) and thus limit the study of the general chromosome architecture (Borgognone et al., 2017; Xiao & Zhou, 2020). Therefore, in the need of technologies that allow the sequencing of long reads, new technologies were recently developed and named Third-Generation Sequencing (TGS) which are capable of producing long reads, from 500 bp up to 2Mb. Two main technologies have been successfully established and are widely used currently, mainly PacBio (Eid et al., 2009) followed by Nanopore DNA sequencing.

Both, second and third generation coexists today and have allowed the sequencing of thousands of species due to their high throughput and low cost, but also thanks to the development of bioinformatics field. The combination of both technologies have allowed researches to study the sequence of complete eukaryotic chromosomes, providing highly a better understanding of the genome fraction occupied by non-coding DNA regions (Costa 2008).

1.3. Pangenomes

The information obtained through a single genome or a single reference when studying large populations is limited as genetic variants and subsequent genotype-phenotype associations have a negative repercussion in the understanding of genomic basis of traits.

Thus, the genetic study using pangenomes (assembly and comparison of multiple individual genomes from the same species) or super-pangenomes (assembly and comparison of multiple individuals from the same species and/or relatives species) (Khan et al., 2020), enables a better mapping accuracy and consequently a higher quality of variants as well as a more accurate gene expression. Furthermore, a pangenome analysis shows the complete diversity within a species, with some of these variations within the genome being potentially responsible of important agricultural phenotypes (Bayer et al., 2020).

By the study of these multiple genomes (pangenome) we can also distinguish conserved genes (*core* genes) which are important for plants development and are present in all genomes, or by contrary, those that are *accessory* or *dispensable* genome genes that are present in some of the individuals. For these reasons, it is expected that in the following years, pangenomes will become the new reference for genomic analyses, instead of using single genomes (Khan et al., 2020).

1.4. Transposable elements

1.4.1. General description

Transposable elements (TEs) are repetitive DNA sequences able to move and replicate within the genome, increasing its copy number, and generating plasticity by producing mutations and genome rearrangements. They constitute the most abundant repetitive fraction of the genome (Wicker et al., 2007). They are found in all eukaryotic organisms constituting a wide, but variable, proportion of their genome and therefore, TEs have large diversity. In maize (*Zea mays)*, TEs occupy about 85% of the genome (Schnable et al., 2009), in humans about 50% (K. Pace II & Feschotte, 2007; Lander et al., 2001), in rice (*Oryza sativa*) around 35% (Matsumoto et al., 2005) and in *Arabidopsis thaliana* about 15% (Poczai et al., 2000). These repetitive sequences have played an important role in the genome evolution and structure enabling them to be adapted to new environments.

1.4.2. History of transposons discovery

Barbara McClintock (1902-1992), was a North American scientist who made great discoveries and advances in the field of genetics, mainly due to her experiments with the maize (*Zea mays*) genome. Her greatest achievement was accomplished in 1949 with the discovery of mobile genetic elements, for which she was awarded the Nobel Prize in physiology or medicine in 1983.

Barbara McClintock, started her studies extending previous work done in drosophila to maize, increasing maize to a model organisms (Ravindran, 2012). Later on, in 1932, McClintock and Harriet Creighton demonstrated genetic recombination or "crossing-over", which involve physical exchange of chromosome segments (Creighton & McClintock, 1931)**.**

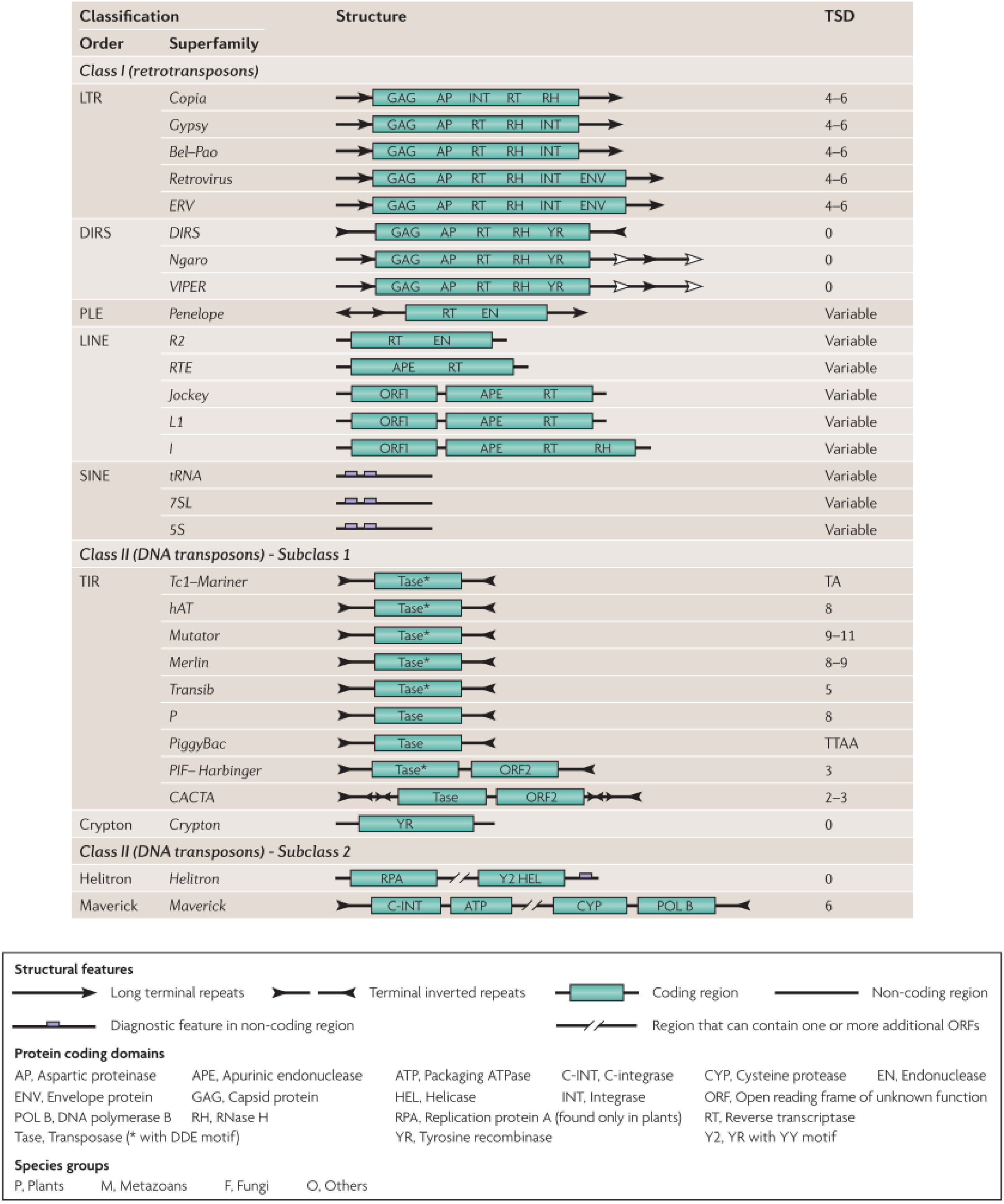
In 1936 she started to study chromosome-breakage by X-irradiation and in further experiments she focused on chromosome-breakage in chromosome 9 of maize (Fedoroff, 1994). In the course of one of her experiments, a phenomenon of rare occurrence appeared with remarkably high frequencies producing drastic structural modifications and genic instability (McClintock, 1950). This chromosome-breaking event always occurred at the same locus and had the ability to change its position within the chromosome and could alter other genes expressions (Pray, 2008).

McClintock described two genetics elements related with this phenomenon (McClintock, 1950). The first element was *Ds* (dissociation), able to create mutations (McClintock, 1945)**,** which could not transpose itself but with the presence of the second element, *Ac* (activator). Both elements, known as *Ac/Ds,* were firstly observed in the aleurone in corn kernels endosperm resulting in unstable phenotypes (McClintock, 1950).

With this discovery she refused the statement that genes are stable entities and proved their instability, redefining the concept about genes and genomes with two revolutionary concepts, by overthrowing the constant genome notion and by demonstrating that unitary genes are not indivisible alternate alleles, but a structure similar to a mosaic with multiple genetic locus (Shapiro, 2015).

1.4.3. History of TEs classification

The earliest classification of TE was made by Finnegan in 1989 depending on their transposition mechanism, whether they use RNA intermediate (Class I or retrotransposons) or, by contrast, use a DNA intermediate (Class II or DNA transposons) (Finnegan, 1989). Few years later, in 2007, due to the emerging data of TEs and their importance in eukaryotic genomes, a hierarchical classification was designed, focusing on their transposition mechanism, structure similarities and relationships, **Figure 4** (Wicker et al., 2007). This classification is the most commonly used, and it divides TEs into classes, subclasses, orders and superfamilies. Classes were distinguished by their transposition mechanism and further divided into subclasses by their integration in the DNA, orders which are TEs that share a common genetic organization as well as a monophyletic origin and finally, in superfamilies that are a closely related group of TEs that can be traced as descendants of a single ancestral unit (Bourque et al., 2018; Wicker et al., 2007). Is worth mentioning that a year later (2008) a new hierarchical classification was submitted in RepBase also focusing on eukaryotic TEs and repetitive sequences, but with a different structure classification (Kapitonov & Jurka, 2008).



Classification		Structure	TSD
Order	Superfamily		
Class I (retrotransposons)			
LTR	*Copia*	GAG AP INT RT RH	4–6
	Gypsy	GAG AP RT RH INT	4–6
	Bel–Pao	GAG AP RT RH INT	4–6
	Retrovirus	GAG AP RT RH INT ENV	4–6
	ERV	GAG AP RT RH INT ENV	4–6
DIRS	*DIRS*	GAG AP RT RH YR	0
	Ngaro	GAG AP RT RH YR	0
	VIPER	GAG AP RT RH YR	0
PLE	*Penelope*	RT EN	Variable
LINE	*R2*	RT EN	Variable
	RTE	APE RT	Variable
	Jockey	ORFI APE RT	Variable
	L1	ORFI APE RT	Variable
	I	ORFI APE RT RH	Variable
SINE	*tRNA*		Variable
	7SL		Variable
	5S		Variable
Class II (DNA transposons) - Subclass 1			
TIR	*Tc1–Mariner*	Tase*	TA
	hAT	Tase*	8
	Mutator	Tase*	9–11
	Merlin	Tase*	8–9
	Transib	Tase*	5
	P	Tase	8
	PiggyBac	Tase	TTAA
	PIF– Harbinger	Tase* ORF2	3
	CACTA	Tase ORF2	2–3
Crypton	*Crypton*	YR	0
Class II (DNA transposons) - Subclass 2			
Helitron	*Helitron*	RPA Y2 HEL	0
Maverick	*Maverick*	C-INT ATP CYP POL B	6

Figure 4. Representation of the main structural and coding features of TEs in a hierarchical classification. TEs are divided into classes, subclasses, orders and superfamilies according to *Wicker et. al.* 2007.

1.4.4. Class I transposons or retroelements

Class I elements (retrotransposons) are the most common class of transposable elements and can make up the bulk of many genomes. They transpose via a 'copy-and-paste' mechanism (**Figure 5**) in which mRNA is transcribed from the element by RNA polymerase II (RNA Pol II), then converted into cDNA by reverse transcription and, finally integrated by an integrase enzyme at a new position in the genome (Lisch, 2013). Each complete transposition cycle produces new TE copies. In consequence, retrotransposons reach high copy numbers and are often the major contributors to the repetitive fraction in large genomes (Wicker et al., 2007). Retrotransposons are further divided into 5 orders: long terminal repeat (LTR) retrotransposons, dictyostelium intermediate repeat sequence (DIRS), penelope (PLE) and non-long terminal repeat (non-LTR) retrotransposons, that includes long interspersed nuclear elements (LINEs) and short interspersed nuclear elements, (SINEs).

LTR retrotransposons (**Figure 4**) are among the most abundant constituents of most eukaryotic genomes, especially relevant in plants, which displays a higher content than animals (Havecker et al., 2004). LTR retrotransposons size ranges from 4 to 31 kb (Orozco-Arias et al., 2019), and are characterised by having long terminal repeats (LTR) sequences of a few hundred up to 5 kb base pairs long in both ends (Finnegan, 2012). These LTR sequences, composed by three domains (R, U5 and U3) are non-coding regions but contain start and stop signals for critical processes to TE replication (Orozco-Arias et al., 2019). Also, they are flanked by target site duplications (TSD) of variable length (Wicker et al., 2007). Between the two LTR sequences there are two open reading frames (ORF), known as *gag* and *pol* (Finnegan, 2012; Orozco-Arias et al., 2019). *Gag* gene encodes structural proteins (i.e. capsid (CA) and nucleocapsid (NC) (Vicient & Casacuberta, 2020)**)**, that assemble virus-like particles. The *pol* gene encodes several enzymatic functions necessary for reverse transcription and integration in the host including aspartic proteinase (AP), reverse transcriptase (RT) and DDE integrase (IN). Depending of the position of the different domains that *pol* encodes, we can differentiate between Ty1-*Copia* and *ty3-Gypsy* superfamilies (**Figure 4**) (Wicker et al., 2007; Wojciech Makałowski, 2019).

Exceptionally, some LTR retrotransposons, found in plants and insects, may present more ORFs, called additionals open reading frames (aORF) frequently positioned between *pol* and 3' LTR in sense or antisense direction (Vicient & Casacuberta, 2020). Some LTR retrotransposons, called endogenous retroviruses, are closely related in evolution with retroviruses due to the presence of a similar coding region that encodes for an envelope-like protein *(env)* (Wicker et al., 2007).

DIRS (**Figure 4**) retroelements are structurally very diverse and are present in almost all organisms, including plants, while PLE retroelements (**Figure 4**) are widely distributed from amoebae and fungi to vertebrates, but not in mammals, and very few are observed in plants (conifers).

Non-LTR retroelements are generally much less abundant in plant genomes than LTR retrotransposons, and lack from LTR sequences. Non-LTR retroelements are usually sub-classified into LINEs and SINEs (**Figure 4**). LINEs can reach several kb long and are flanked by TSD. They are autonomous elements, having 2 ORFs, *gag* and *pol*. By contrast, SINEs, with a length from 80 bp up to 500 bp long and flanked by TSD, are non-autonomous elements that cannot transpose by themselves and depend on LINEs for its transposition. Non-LTR retrotransposons often contain a poly-A tail at 3' end (Carnell & Goodman, 2003; Orozco-Arias et al., 2019; Weiner, 2002)**.**

1.4.5. Class II transposons or DNA transposons

Class II elements transpose via a DNA intermediate. They are present in almost all eukaryotes, generally in lower copy number than retrotransposons. DNA transposons are divided into two subclasses differing on their transposition mechanism (Wicker et al., 2007).

Subclass 1 (**Figure 4**), is characterized by having terminal inverted repeats (TIR). TIRs are sequences of about 9 to 40 base pairs long present on both TE ends that are recognized by the transposase encoded by the element itself. Also, they present TSD of variable length (Wicker et al., 2007). Subclass 1 requires the cleavage of both DNA strands for their transposition via a 'cut-and-paste' mechanism (**Figure 5**), in which the element is physically excised from the chromosome and reintegrated at a new location (Lisch, 2013). This process is not usually replicative, unless the gap caused by excision is repaired using the sister chromatid (Wojciech Makałowski, 2019).

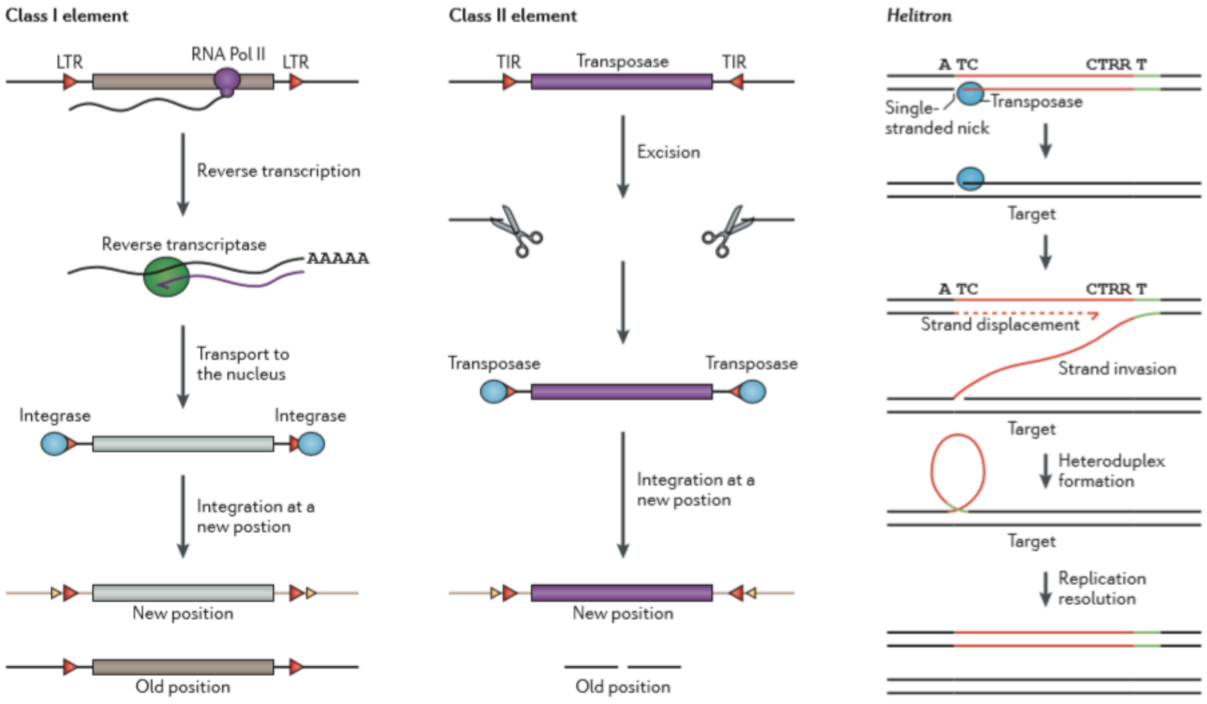


Figure 5. Schematic representation of TE transposition mechanism. Class I element via 'copy-and-paste' mechanism, Class II element via 'cut-and-paste' mechanism and Helitrons (Class II element) via 'rolling circle' mechanism. Extracted from *Lisch et. al.* 2013.

One important group of this subclass are miniature inverted repeat transposable elements (MITEs). They have a relatively small length, between 50-800 bp, but are presented in high copy numbers in plants genomes (Feng, 2003). MITEs are non-autonomous elements, more specifically, are derivative elements from autonomous DNA transposons that lost their capacity to transpose by themselves (Crescente et al., 2018; Lu et al., 2012). MITEs are highly associated with genes (Crescente et al., 2018), being found frequently within or near them.

Subclass 2 (**Figure 4**) is characterized by the lack of TIR sequences and by having a replication process that does not involve a double-strand DNA break, but the cleave of a single DNA strand (Lisch, 2013; Soriano, 2016). Subclass 2 can be further divided into 2 orders, Helitrons and Mavericks. It is worth mentioning that the order Helitrons, firstly discovered in plants, are included in this subclass because of the absence of a RNA intermediary, not because of phylogenetic proximity, further Helitrons lack of TIR and TSD (Wicker et al., 2007) and, transpose via a 'rolling circle' mechanism (**Figure 5**). This process involves nicking at the Helitron terminus, followed by strand invasion, DNA synthesis, strand displacement and resolution of a heteroduplex by DNA replication (Lisch, 2013; Thomas & Pritham, 2015). Mavericks order, also known as polintons, have been observed in few eukaryotes, but not in plants. They are long TE, reaching 10 to 20 kb and are bordered by TIR sequences (Wicker et al., 2007). Mavericks also undergo a replicative transposition cleaving a single DNA strand fragment, which will be replicated extrachromosomically for its integration at a new site (Soriano, 2016).

1.4.6. Autonomous and non-autonomous transposable elements

TEs can be further divided in autonomous and non-autonomous elements. TEs that encodes all the structural hallmarks and enzymatic domains necessary for its transposition (without implying that the element is either functional or active), are referred as autonomous elements. By contrast, non-autonomous TEs lack some, or all coding domains necessary for the transposition. Under specific circumstances, non-autonomous elements can be mobilized by a related autonomous element (i.e. MITEs). First autonomous and non-autonomous elements were discovered by Barbara McClintock, which demonstrated that Ds (non-autonomous TE) needed Ac (autonomous TE) for its transposition in maize kernels in1944.

1.4.7. Impact of TE in genomes

Transposons play an important role in shaping genome and chromosome architecture, since the transposition, activated in response to multiple environmental stresses or as simple by-products of physiological, cellular or genetic stresses (Fambrini et al., 2020), generate permanent genomic modifications such as deletions, inversions, translocations or other types of genomic rearrangements that can affect by being favourable or deleterious for the organism (Schrader & Schmitz, 2019).

Transposition can affect the function and expression of genes, either by inactivating their expression due to modifications of the coding regions caused by the insertion of the TE into the gene, or also if it is inserted relatively close to such genes, disrupting normal gene function and creating new expression patterns by bringing new *cis* regulation (Lisch, 2013; Schrader & Schmitz, 2019).

TE movement can also mediate translocation of DNA segments, such as genes fragments, that are not part of the TE structure and are moved to another genomic region generating new variability and new phenotypes (Schrader & Schmitz, 2019; Wei & Cao, 2016).

But there must be a balance between TEs activity (expression) and TEs repression as a high copy numbers may lead to a disadvantage for the host. Thus, there are some self-regulatory epigenetic mechanisms for transposition repression such as DNA methylation, DNA modification pathways and a variety of small RNA. As a result of this balance, TEs have a higher expression level depending on the tissues and the stage of the life cycle, having a higher activity when is found in germline stage while they are not highly expressed in the somatic stage (Bourque et al., 2018).

So, the genomic divergence created by these mobile elements is a continuous process that plays a very important role in the evolution of plants for adaptation to new environments and has also been used by humans for the selection, unconsciously, of those individuals with best traits for the domestication of today's cultivable plants (Lisch, 2013).

1.4.8. Plant transposable elements

Transposons occupy a large portion of genome plants, but is variable depending on the species, in *Arabidopsis thaliana* it has a content of TEs of 18,5 % of the total genome, in *Malus x domestica* (Apple tree) a 42,4 %, in *Triticum aestivum* (Bread wheat) a 79,8 % and in *Zea mays* (maize) an 80 % (Kim, 2017). Most of these transposons are non-autonomous TEs, which lack some or all the hallmarks and enzymatic domains necessary for its transposition, or old remanent transposons (truncated transposons) that neither can do transposition as they have some structural modifications.

The activity of TEs on plant genomes is highly variable, from gene mutations to genome rearrangements, post-transcriptional silencing, among others (Lisch, 2013). Thus, they constitute an enormous resource of natural genetic variability. Some of the genetic changes that TEs produce can lead to phenotypic variation of agronomic importance and therefore, new techniques are being developed to induce and control transposition for crop improvement by generating new phenotypes with agricultural interests, such as resistant to environmental stresses or pests as well as for food improvement of quality and productivity (Lisch, 2013).

Some examples that have happened naturally and are well known are: the presence of an LTR (TY1/Copia) in the third exon of the PpeMYB25 gene that caused a loss of expression of this gene generating this new peach phenotype called nectarine (Vendramin et al., 2014), the presence of a MITE within the ZmNAC111 region of the maize genome allowing greater resistance to drought (Mao et al., 2015) or the presence of an LTR in the VvmybA1b allele that causes mutations in grape skin colour (Kobayashi et al., 2004).

1.4.9. Annotation of TE sequences in complete genomes

Current sequencing and genome assembly techniques allow us to obtain high quality, almost complete genomes, and the annotation of the transposons that compose the repetitive fraction implies the identification of their exact place within the genome. For this identification it is currently using methods of genome self-alignment that search for structural elements characteristic of TEs (i.e. LTRs or TIRs) or also the search for coding regions that allow transposon transposition (i.e. integrase, reverse transcriptase, etc.).

In some model organisms, such as maize or drosophila, the existing models have been created manually but due to the interest in the identification of TEs in many other cultivated species, these models have been automated. These automated models are combined in pipelines, which are the joint result of multiple programs (Ou et al., 2019).

During the last few years the appearance of multiple TE annotation programs has increased, but the ones that are currently the most complete are EDTA (**Figure 6**) (Ou et al., 2019), REPET (Flutre et al., 2011) and RepeatModeler2 (Flynn et al., 2020).

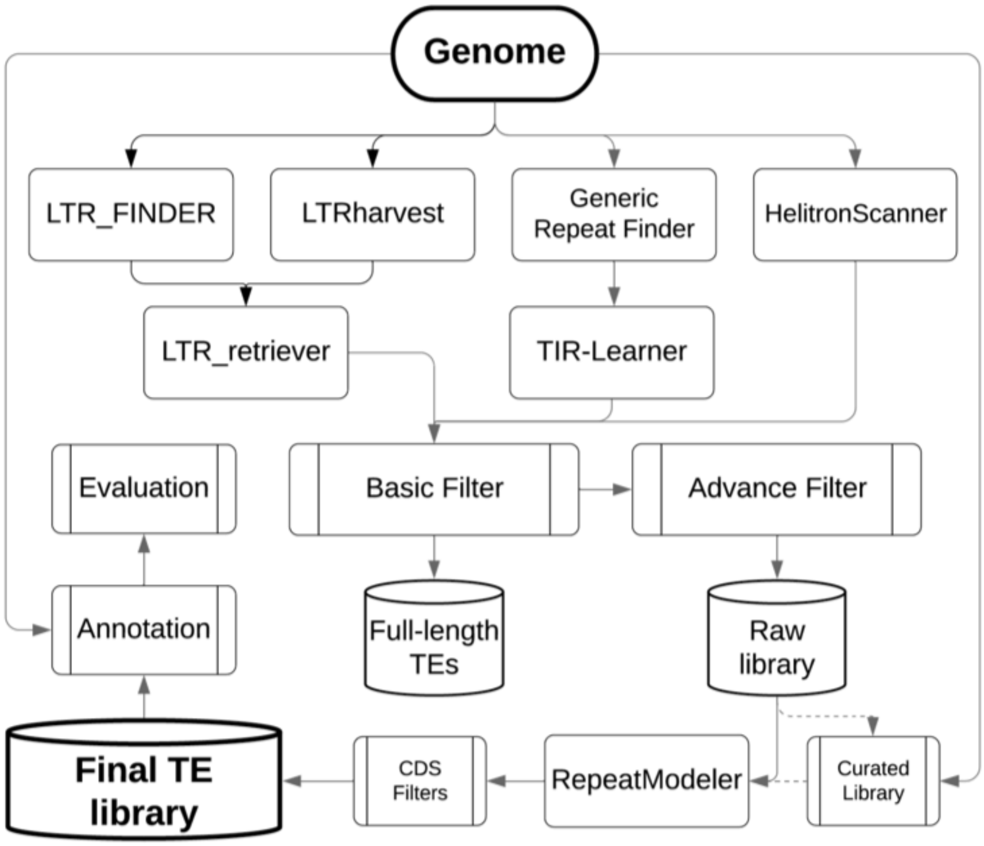


Figure 6. EDTA workflow for LTR retrotransposons, TIR elements, and Helitrons candidates identification from the genome sequence. Extracted from *Ou et. al.* 2019.

2. Objectives

2.1. General objective

The aim of this TFG is the characterization of transposons in the pangenome of P*runus* genus and in cultivated varieties of peach (*Prunus persica*) from their genomic sequence.

2.2. Specific objectives

- Annotation of transposons in seven *Prunus* genomes using bioinformatics tools.

- Construction of a non-redundant database of transposon sequences of the *Prunus* genus.

- Identification of polymorphic transposon insertions in peach varieties (*P. persica*) using the *Prunus* transposons database and publicly available resequencing data.

- Molecular validation of transposon insertions in *P. persica* with potential impact on genes.

3. Materials and methods

3.1. Bioinformatics analysis

BIoinformatics analyses were carried out mainly in a laptop with Linux Operating System, using R and Bash programmes languages. The most computationally intensive data analyses were performed in the high-performance computing cluster (HPC) available at Centre for research in agricultural genomics (CRAG). All the code generated to carry out this work is available in **Annex A, scripts 1 -14**.

3.1.1. Genomic data retrieval

Prunus genomes used were downloaded in FASTA format from the "GDR Rosaceae Database" (https://www.rosaceae.org), queried on March 08, 2021. The following *Prunus* genomes were used; *Prunus armeniaca* (Par) (Jiang et al., 2019), *Prunus avium* Tieton (Pav) (Wang et al., 2020), *Prunus dulcis* Texas (PdT) (Alioto et al., 2020),*Prunus dulcis* Lauranne (PdL) (Sánchez-Pérez et al., 2019), *Prunus mira* (Pm) (Cao et al., 2020), *Prunus persica* (Pp) (Verde et al., 2013, 2017), *Prunus salicina* (Ps) (Liu et al., 2021).

3.1.2. Evaluation of genome assembly quality

We used different metrics and indicators to evaluate the quality of the *Prunus* assemblies. Genome size, contig and scaffold number, scaffold N50, scaffold L50, contig N50 and contig L50 were calculated using bbmap tool (https://sourceforge.net/projects/bbmap/) using **Script 2**. Gap content (unkwnown sequences) was obtained using Seqtk tool (https://github.com/lh3/seqtk) with **Script 3**. LAI (LTR Assembly Index) value was obtained by running LTR-retriever Scripts on RepeatMasker LTR annotation (**Script 4**). The BUSCO score (Benchmarking Universal Single-Copy Orthologs) and genome assembly method of each of the seven *Prunus* genomes were obtained from the original publications.

3.1.3. Transposon annotation

Genomes were analysed with EDTA (Extensive de novo TE annotator) (Ou et al., 2019), a computational pipeline for whole-genome TE annotation that integrates the results of multiple TE annotation tools. EDTA package filters false TE discoveries and creates a high-quality non-redundant TE library which is then used to annotate whole genomes.

EDTA shows a high sensitivity and precision in the detection of LTR retrotransposons, but lower for MITEs, TIRs and Helitrons (Ou et al., 2019). The files obtained from the EDTA analysis (**Script 1**) were the main starting point for all the analyses presented in this work.

3.1.4. Classification, length and insertion time of intact transposons

Superfamilies of transposons were classified at the order and superfamily into 6 different groups by EDTA (LTR/Copia, LTR/Gypsy, LTR/Unknown, MITEs, TIRs, Helitrons). Distribution of elements length (bp) was obtained using a custom R Script (**Script 5**).

LTR retrotransposons insertion time was calculated using LTR retriever, as part of EDTA analysis, and plotted using R and RStudio through **Script 6**, obtaining density and frequency plots.

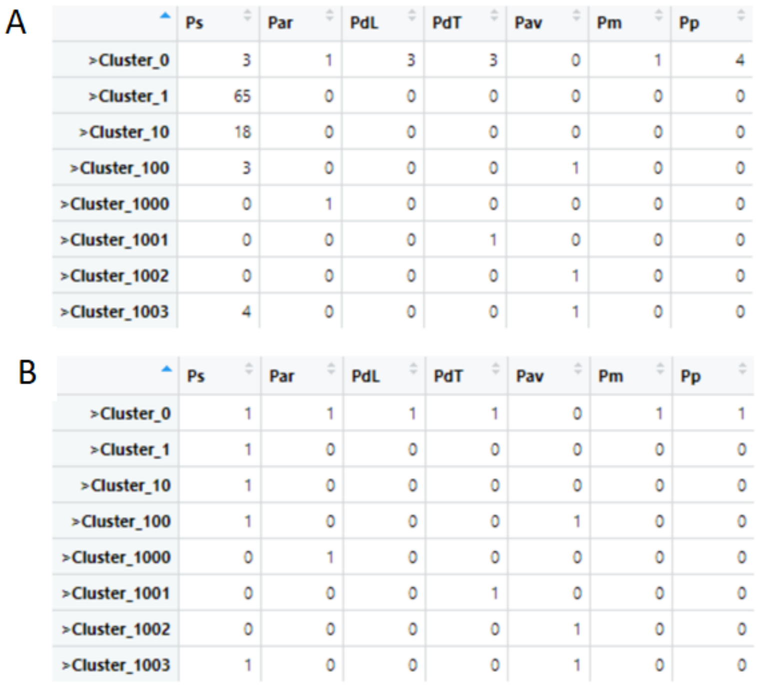
3.1.5. Non-redundant *Prunus* TE library

Individual TE libraries (TE sequences of all autonomous elements in fasta format from each species) were concatenated in order to build a global *Prunus* TE library (**Script 7**). To eliminate redundancy (due to the same TE being detected in two or more species), we used CD-HIT (Li & Godzik, 2006) to cluster sequences of all intact elements by homology using an 80 % identity cutoff (**Script 8**). Each cluster was considered a TE family, as previously described (Wicker et al., 2007). The longest sequence per cluster (centroid sequence) was retained as the representative sequence of this TE family. *Sed* and *cat* commands were used to rename and concatenate sequences when necessary.

3.1.6. TE-based *Prunus* pangenome analysis

A Script was built to parse the CD-HIT results and transform the clustering analyses into abundance matrices (number of TE copies per cluster in each variety) and binary matrices (presence/absence of each cluster in each genome) (**Table 1**) summarizing the TE-based pangenome of the *Prunus* genera. A stringent filter was applied only to MITEs, TIRs and Helitrons to avoid false positives: for MITEs, we retained only families with a cluster size larger than 5 in at least one species. For TIRs and Helitrons, we retained only families with significant homology (BLASTX evalue < 1e-10 , **Script 8**) to TE proteins deposited in Repbase database (Jurka et al., 2005) a blastx was done by **Script 9**. This step was not necessary for LTR retrotransposons due to the high precision of EDTA in the detection of these elements.

The results were further processed to build the corresponding histograms with ggplot2 (**Script 10**).



A

	Ps	Par	PdL	PdT	Pav	Pm	Pp
>Cluster_0	3	1	3	3	0	1	4
>Cluster_1	65	0	0	0	0	0	0
>Cluster_10	18	0	0	0	0	0	0
>Cluster_100	3	0	0	0	1	0	0
>Cluster_1000	0	1	0	0	0	0	0
>Cluster_1001	0	0	0	1	0	0	0
>Cluster_1002	0	0	0	0	1	0	0
>Cluster_1003	4	0	0	0	1	0	0

B

	Ps	Par	PdL	PdT	Pav	Pm	Pp
>Cluster_0	1	1	1	1	0	1	1
>Cluster_1	1	0	0	0	0	0	0
>Cluster_10	1	0	0	0	0	0	0
>Cluster_100	1	0	0	0	1	0	0
>Cluster_1000	0	1	0	0	0	0	0
>Cluster_1001	0	0	0	1	0	0	0
>Cluster_1002	0	0	0	0	1	0	0
>Cluster_1003	1	0	0	0	1	0	0

Table 1. Example of A) Abundance matrix (copy number per TE cluster). B) Binary matrix (presence/absence of each TE cluster in each species) in Copia superfamily.

3.1.7. Hierarchical clustering based on TE abundance and family polymorphisms

Following the R Script discussed in section 2.1.6 (**Script 10),** hierarchical clustering combined with heatmaps were build using heatmap.2 function, taken into account either the presence or absence of TEs or the frequency abundance of each TE (TE matrices, **Table 1**).

3.1.8. Percentage occupied by TEs in each genome

To know the percentage occupied by the TEs in the genomes, after CdHit clustering, we used a RepeatMasker (**Script 11**) and then a ParseRM (**Script 12**). RepeatMasker identifies all the regions in the genome with significant homology to TEs in our global Prunus TE library, and ParseRM is a Script used to summarize the RepeatMasker output.

Using excel, the percentage occupied by TEs was found by dividing the length of each TE superfamily by the total length of the corresponding genome assembly.

3.1.9. TE distribution on chromosomes

In this section only the arrangement of TEs in the chromosomes of *P. salicina* and *P. armeniaca* has been compared. Each chromosome was divided in 50 windows, and the number of TEs per window was calculated using Bedtools software and RepeatMasker results (**Script 13**). For the generation of the corresponding graphs, we followed the **Script 12**. The number of TEs per windows was plotted using ggplot2.

3.1.10. Identification of polymorphic TE insertions between *P. persica* var EarlyGold and Lovell

The library created in this work (PRUNUS_TE_LIBRARY) was used to detect non-reference insertions on *P. persica EarlyGold* (insertions present in this variety that are absent in the reference genome). For this we used the software Jitterbug (Hénaff et al., 2015) with default parameters. This tool allows to identify TE insertion signatures from raw re-sequencing reads mapped to a reference genome. For this, we took advantage of the genome mapping that was already performed in the host laboratory with the resequencing reads of EarlyGold variety using as reference the *P. persica Lovell* genome.

3.2. Molecular validation

3.2.1. Sampling and DNA extraction

Young leaves were sampled from *Prunus persica* EarlyGold tree from Gimenelles IRTA field (Lleida) in 2019 by Carlos de Tomás (research group member), and deposited in the freezer at -80 ºC in the Centre for research in agricultural genomics (CRAG) until its use in June of 2021.

For the total genomic DNA extraction, 100 mg of sample (young leave) was taken. First, the sample was put in a mortar with the addition of liquid nitrogen and ground until the obtention of a powder. The result was put in an eppendorf tube and then followed the Doyle and Doyle protocol (Doyle, J.J., 1990) with small modifications specific for P*runus* DNA extraction.

The quality and concentration of DNA was checked by using a Nanodrop and by an agarose 1 % gel.

3.2.1.1. Doyle protocol for genomic DNA extraction

100 m g of frozen young leaves sample (*Prunus persica* EarlyGold) were powdered in a mortar, 660 µL of CTAB buffer (for 8 mL of CTAB we added 0,16 g de CTAB (2 %), 2,24 mL of NaCl at 5 M, 800 µL of tris HCL (pH = 8) at 1 M, 320 µL of EDTA at 0,5 mM, 16 µL of β-Mercaptoethanol (0,2 %) and 4,624 mL of H_2O) was added and then the mix was incubated in a termoblock between 60-65 ºC for 40 min. After, 660 µL of chloroform-isoamyl alcohol (24:1, v:v) was added and centrifuged at 3000 rpm during 15 min at room temperature (RT). The supernatant (aquose fase) is transferred to another tube and incubated with RNAse at 37 ºC for 20 minutes. A second step of chloroform-isoamyl alcohol (24:1, v:v) was repeated and the supernatant was transferred to another tube. Same volume of isopropanol as the volume of supernatant recovered was added. The tube was inverted two or three times and then centifugated at 3000 rpm during 30 min at 4 ºC. Supernatant was removed and at the same time the same volume of buffer "Rentat" (alcohol 70 %) was added. Then, it was centrifugated at 3000 rpm for 10 min at 4 ºC. Following, the supernatant was removed and the tube was left until it was dry at room temperature, finally, 50 µL of H_2O was added and then stored in the freezer at -20 ºC.

3.2.2. Primer design

Primers were designed in the adjacent regions of a TE insertion detected by Jitterbug. According to the reference genome (without the insertions, empty loci), they were designed to amplify a sequence, of at least 300 to 600 bp. Up to 4 putative TE insertions were selected for validation, and for each sequence a specific pair of primers was designed (**Table 2**). Primers were design to meet the following characteristics: Tm had to be between 56 and 61 ºC, % GC between 40 to 60 % and a length size of 20 bp. All primers were designed using Primer3 (https://primer3.ut.ee/) following the indications above, or by a % GC calculator. Once primers were created, they were purchased to "Integrated DNA Technology" company (https://eu.idtdna.com/pages). The characteristics of the primers used are summarized in **Table 2**.

3.2.3. PCR and Agarose gel

PCR reactions were set in a total volume of 20 µL, containing 7,72 µL of H_2O, 4 µL of LongAmp *Taq* Reaction Buffer of 5X, 2 µL of dNTPs at 2 µM, 2 µL of forward and reverse primers respectively at 2 µM, 0,28 µL of LongAmp *Taq* Reaction Polymerase of 2.500 units/mL and 2 µL of DNA of *Prunus persica* EarlyGold variety with a concentration of 65 ng/µL (total of 130 ng DNA).

The PCR protocol used is shown in **Table 3.**

After, electrophoresis was performed in an agarose gel at 1 %, with one drop of ethidium bromide for each 50 mL at a concentration of 0,7 mg/mL, and was to check the presence or absence of transposons within the selected genes sequences. The molecular marker (MM) used is shown in **Figure 7.**

3.2.4. Purification of PCR clean-up gel extraction and PCR clean-up

Purification of PCR DNA product was done directly from the agarose gel bands following the protocol of the NucleoSpin Gel Clean-up from Macherey-nagel brand (https://www.mn-net.com/bioanalysis/kits/?p=1) in the cases where double band was present. On contrary, if there was a single band, the purification was done directly from PCR product.

Concentrations of DNA were measured by spectrophotometry using a Nanodrop, the concentration and each wavelength measured are resumed in **Table 4.**

Table 2. Primers design used for TE amplification.

Primers	Primers sequences	Primer length (bp)	Chromosome	Start	Finish	Amplification product size (bp)	GC (%)	Tm (ºC)
Amplicon 1	GACTTGCTCACGTGCCACAT	20	Pp08	16558258	16558277	361	55	58.2
	ATAGCTGAAGGTGACCGCAA	20	Pp08	16558619	16558600		50	56.5
Amplicon 2	GCACCTTTTCACGCCATACT	20	Pp04	6366632	6366651	304	50	55.8
	TTTGTCAGCCGCTTCAATCC	20	Pp04	6366936	6366917		50	56.0
Amplicon 3	TTTCTCCCGGCACACTACTT	20	Pp07	15412567	15412586	361	50	56.2
	CACCTGTGCCCAATGATAGC	20	Pp07	15412928	15412909		55	56.2
Amplicon 4	TGGTGTCAACGTGAAGGGAT	20	Pp04	9624357	9624376	338	50	56.5
	CGATGGTGCCCGTAATGTTG	20	Pp04	9624695	9624676		55	56.6

Note. bp means 'base-pair-long' and Tm means 'melting temperature'.

Table 3. *PCR methodology: main step, temperature, time and number of cycles.*

Step	Temperature (ºC)	Time	Cycles
Initial denaturation	94	30 s	
Denaturation	94	20 s	X 30
Annealing	56	20 s	
Extension	65	6:30 min	
Final extension	65	10 min	
Repose	16	∞	

Note 1. "∞" symbol represents the time from when PCR has finished to when we took it from PCR machine.

Note 2. Extension is of 6:30 min as we expect to have an insertion in the samples, thus more time is needed and LongAmp *Taq* Reaction Polymerase is used.

Table 4. *Nanodrop concentrations after DNA purification.*

Plate ID	Sample ID	Concentration (ng/µL)	A260	A280	260/280	260/230
A1	BG1	23,1	0,462	0,259	1,78	0,53
B1	BP1	35,92	0,718	0,392	1,83	0,88
C1	BG3	18,63	0,373	0,204	1,83	0,95
D1	BP3	92,76	1,855	1,085	1,71	0,17
E1	2	29,89	0,598	0,33	1,81	0,68
F1	4	37,03	0,741	0,4	1,85	0,71

Note 1. Samples BG1, BP1 (sample 1) and BG3, BP3 (sample 3) were DNA gel purification while samples 2 and 4 were PCR clean-up purification.

Note 2. Wavelength measured: A260 measures DNA at 260 nm. A280 measures protein at 280 nm. 260/280 absorbance ratio is an indicator of protein contamination, if ≥ 1.8, it is pure DNA sample. 260/230 absorbance ratio smaller than 1.8 indicates contamination caused by organic compounds or chaotropic agents.

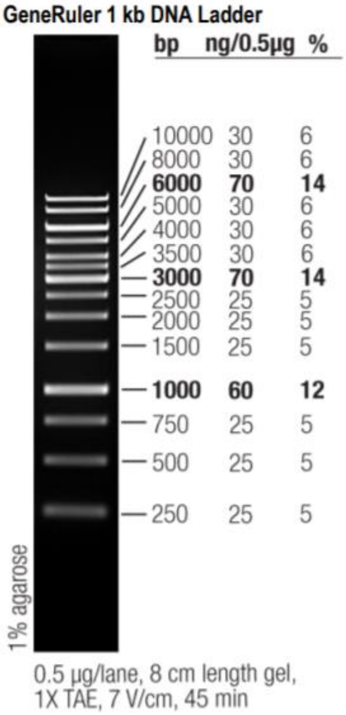


Figure 7. GeneRuler 1 kb DNA Ladder. The ladder is composed of fourteen chromatography-purified individual DNA fragments (in base pairs): 10000, 8000, 6000, 5000, 4000, 3500, 3000, 2500, 2000, 1500, 1000, 750, 500, 250. It contains three reference bands (6000, 3000 and 1000 bp) for an easy orientation.

3.2.5. Sequencing

Purified PCR DNA product (5 µL for samples 3 and 4 and 10 µL for samples 1 and 3, all concentrations are shown in **Table 4**) was taken to the CRAG genomic service to perform Sanger sequencing.

3.2.6. Verification of the presence or absence of the sequenced TE

Once the sequencing of the DNA purification was obtained, a BLASTN (**Script 14**) was performed against *P. persica* and against the *Prunus* TE library (PRUNUS_TE_LIBRARY) to check if the sequences were homologous to a TE present in our library, and thus validate the presence or absence of TE.

4. Results and discussion

4.1. *Prunus* genomic characteristics and assembling quality

All the analyses presented in this work rely on publicly available *Prunus* genome assemblies, most of them carried out by international consortia (Alioto et al., 2020; Verde et al., 2013, 2017). As genome assembly quality critically impacts the detection of transposable elements (Ou et al., 2018), we evaluated the main characteristics of genomes assemblies of the different *Prunus* species used in this work, in comparison to other agronomically important *Rosaceae* species such as *Malus x domestica* (apple tree) belonging to *Malus* genus and *Fragaria vesca* (strawberry plant) belonging to *Fragaria* genus. All data collected for this comparison is shown in **Table 5A** and **Table 5B**.

Genome size within *Prunus* species ranges between 200 Mb and 280 Mb, with *P. salicina* having the largest genome size. This range is similar to the genome size of *Fragaria vesca*, but much smaller than *Malus x domestica*, whose genome size is about 700 Mb. The genome size of most of the known plant genomes ranges between 700 Mb and 2 Gb, and in this context, we can conclude that *Prunus* species have a very compact genome. This small size in comparison to other plants can be explained by the lack of recent whole-genome duplications in the *Prunus* genus and the different dynamics of TE amplifications (Verde et al., 2013; Wendel et al., 2016).

Regarding the genome assembly, the number of contigs is very variable among the different genomes. A high-quality assembly must be arranged in a low number of contigs (ideally one per chromosome), and a low percentage of gaps (unknown nucleotides, represented by Ns). In this sense, two contrasting examples are *P. salicina,* which is assembled into 272 contigs, with 0,007 % of gaps and *P. dulcis* Texas assembled into 4395 contigs with 1,72 % of gaps.

The assembly of contigs into scaffolds and pseudomolecules is also important, although if the number of contigs is high and are arranged directly to a low number of scaffolds (as it can be observed in *P. avium* Tieton and *Malus x domestica*, **Table 5A**), the genome assembly will contain many gaps. In example, in *P. avium* Tieton we can see that the number of contigs is 2488, which are assembled into 8 scaffolds making a higher percentage of gaps in the genome (19.23 %). Similarly, in *Malus x domestica* 3772 contigs are assembled into 18 scaffolds with a gap fraction of 11.94 %, both of them having the highest gap content compared to the other assemblies.

The number of pseudomolecules refers to the number of chromosomes of each species. *Prunus* genomes are arranged in 8 chromosomes while *Malus x domestica* in 17 chromosomes and *Fragaria vesca* in 7. Having genome assemblies organized into pseudomolecules allows to directly compare whole chromosomes from different species.

Focusing on **Table 5B**, two relevant values can be observed: Benchmarking Universal Single-Copy Orthologies (BUSCO) and LTR Assembly Index (LAI index), which are important indicators of genome assembly quality at the gene and transposon level, respectively.

BUSCO allows to analyse genome quality by the gene content based on evolutionary principles. It represents the presence of intact orthologs of universal genes (based on their presence in 90% of the genome of all organisms in a certain lineage). A high value indicates a high completeness of genome assembly at the gene level. In this case, all the genomes used have a BUSCO equal or higher that 95 except for *P. mira* which has a value of 90.3, indicating that they are in general highly complete.

Nevertheless, BUSCO does not account for the quality of the assembly on the non-coding fraction of the genome. In this sense, LAI Index (Ou et al., 2018) indicates the quality of the genome with respect to this fraction, and refers directly to LTR retrotransposons as they are the main constituents of this fraction.

LAI index represents the percentage of intact LTR retrotransposons *versus* total number of elements present in the genome (including truncated and degenerated copies), after correcting for amplification dynamics. A high value indicates a good genome assembly while a low value indicates a poor-quality assembly and therefore affecting directly to the annotation of transposons. A LAI index value around 20 is considered to be a threshold for “gold standard” quality (Qin P et. al, 2021).

The sequencing method (Sequencing technology, **Table 5B**) as well as the assembly method (Assembly method, **Table 5B**) have a direct impact on genome assembly, as reflected by both BUSCO and LAI metrics. In this case, we found that genomes sequenced with PacBio or Sanger as main technologies have the highest value of both BUSCO and LAI index which would imply that these two sequences techniques allow a higher quality assembly than those made primarily by illumina short-read sequencing (i.e. *P. dulcis* Texas). This is expected as long-reads, in contrast to short-reads, allow to resolve the assembly of near-identical repeats such as TE copies by spanning all the element plus the surrounding regions.

Therefore, in a general overview, we can state that *P. persica* and *P. salicina* have the highest quality and metrics of the overall genome while *P. avium* Tieton and *P. dulcis* Texas have the lowest. In terms of TE content, based on the LAI results we conclude that the assemblies range from medium to high quality, being sufficient to accomplish the objectives of this work.

4.2. TE content in *Prunus* genomes

Using EDTA and RepeatMasker we obtained a complete annotation of five TE superfamilies for each genome. Two of these superfamilies, Copia and Gypsy, belong to LTR retrotransposons (Class I), whose structure is explained in section 1.4.4 while the others TE superfamilies, MITEs, TIRs and Helitrons, belong to DNA transposons (Class II), explained in section 1.4.5. These 5 superfamilies are the most representative of their Class, and are the most studied and well known among plant TEs.

Using EDTA we annotated intact copies (which are TEs that have all hallmarks domains and all the structure necessary to transpose and to be active), and by using RepeatMasker with our *Prunus* library we identified all degenerated elements (which lack coding regions or have structural modifications and are not able to transpose).

The percentage occupied by TEs in *Prunus* genomes (including intact and degenerated copies) ranges between 30 and 33 % with some exceptions (**Figure 8, Table 6**). In *P. avium* Tieton TEs represent 21.4 % of the genome, being the lowest while *P. salicina* with the 40,1 % displays the highest TE content. LTR retrotransposons (Copia and Gypsy) occupy a higher genome fraction compared to DNA transposons (MITEs, TIRs, and Helitrons). This variation is striking in *P. salicina,* which shows a very high proportion of Gypsy elements (19 % of the genome) in comparison to other close species such as *P. avium* (5.7 %, **Table 6**).

Table 5A. *Characteristics of genome assembly.*

Species	Genome size (Mb)	Nº of contigs	Nº of scaffolds	Pseudomolecules	Scaffold N50	Scaffold L50 (Mb)	Contig N50	Contig L50 (Mb)	Gap number	Gap (%)
P. armeniaca	221.90	443	182	8	4	25.1	61	1.02	24297	0.011
P. avium Tieton	214.32	2488	8	8	4	25.5	363	0.12	41198918	19.233
P. dulcis Lauranne	200.33	3012	8	8	4	23.1	564	0.10	723082	0.361
P. dulcis Texas	227.60	4395	691	8	4	24.4	511	0.12	3915729	1.72
P. mira	252.38	2605	657	8	4	27.4	283	0.24	3630962	1.439
P. persica	227.41	2525	191	8	4	27.4	250	0.26	2772453	1.219
P. salicina	284.21	272	75	8	4	32.3	45	1.78	19700	0.007
Malus x domestica	709.56	3772	18	17	8	27.6	326	0.60	84710231	11.938
Fragaria vesca	220.36	159	29	7	3	33.9	12	6.97	948312	0.430

Note. Mb refers to 'megabase'.

Table 5B. *Characteristics of genome assembly.*

Species	BUSCO	LAI Index	Sequencing technology	Assembly Method
P. armeniaca	98.0	18.79	PacBio	Canu assembler
P. avium Tieton	97.4	10.80	Illumina sequencing	Supernova assembler (2.0)
P. dulcis Lauranne	95.0	13.31	PacBio	Canua assembler
P. dulcis Texas	96.0	8.15	Illumina and Oxford Nanopore (low coverage)	MaSuRca (v3.2.3)
P. mira	90.3	12.57	Illumina and PacBio	FALCON and ALLPATHS-LG
P. persica	99.0	24.43	Sanger	Arachne
P. salicina	95.7	20.70	PacBio and Illumina	FALCON (v0.3.0)
Malus x domestica	94.9	-	PacBio and Illumina	Bfast, in house developed software
Fragaria vesca	95.0	-	PacBio	Canu assembler

Note. We did not calculate LAI Index in *Malus x domestica* and *Fragaria vesca.*

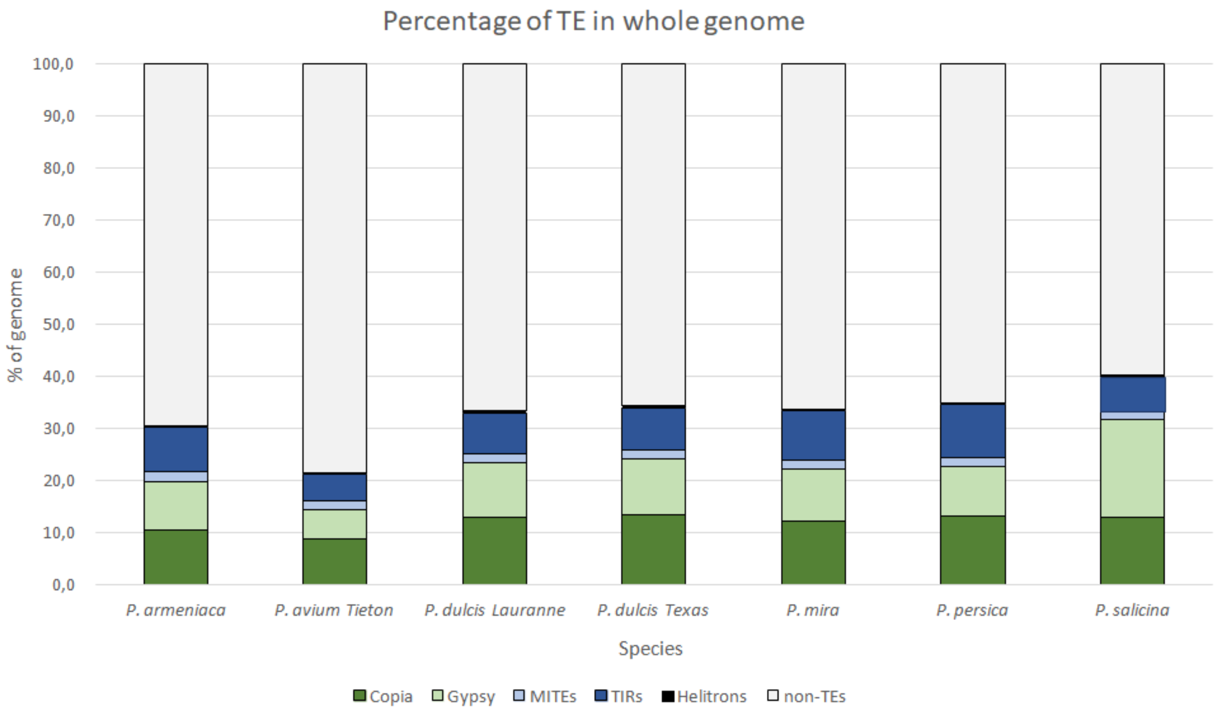


Figure 8. Representation of total TEs in an accumulative histogram.

Table 6. *Percentage of the genome occupied by each TE order and superfamily.*

Species	Copia	Gypsy	MITEs	TIRs	Helitrons	Total TEs	Total non TEs
P. armeniaca	10,4	9,2	1,9	8,4	0,3	30,3	69,7
P.avium Tieton	8,6	5,7	1,6	5,3	0,2	21,4	78,6
P. dulcis Lauranne	12,8	10,4	1,7	7,9	0,3	33,2	66,8
P. dulcis Texas	13,4	10,7	1,7	8,1	0,3	34,3	65,7
P. mira	12,1	10,0	1,7	9,6	0,2	33,6	66,4
P. persica	13,1	9,5	1,7	10,3	0,3	34,8	65,2
P. salicina	12,7	19,0	1,5	6,5	0,3	40,1	59,9

4.3. Intact TE abundance in P*runus* genomes

Intact TEs are the only that retain the potential to transpose and introduce genetic variability, and thus identifying them is of great importance.

In terms of copy number, DNA transposons are almost the doubled RNA transposons, especially due to the high copy numbers of MITEs and TIRs. MITEs are especially abundant in plant genomes such as rice, and polymorphisms of these elements have been recently linked to variability in agronomic traits (Castanera et al., 2021). In the *Prunus* genera, these elements have been described to amplify transcription factor binding sites and are thought to be important regulatory elements for agronomic traits such as stress response, flowering time (Morata et al., 2018). In regard to Helitrons, they have a remarkably low presence compared to the other superfamilies of the same class.

Focusing on each individual species, *P. salicina* stands out from the rest, with a total of 6188 copies of TEs and the highest number of copies in each superfamily (**Table 7**). Remarkably, in the Gypsy superfamily, the increment of copies is up to 5-fold compared to the other close species such as *P. mira* or *P. avium.* Moreover, it is the only species where the number of Gypsy exceeds the number of Copia, when in all the others the presence of Copia is higher.

We analysed the length distribution of intact TEs and found that it was highly variable among all the superfamilies studied. MITEs and TIRs tend to have a smaller size, whereas LTR retrotransposons (Copia and Gypsy) have a bigger size. Intact TEs length are represented as density plot (**Figure 9**) and frequency plot (**Figure 10**) enabling us to better analyse the dynamics of each superfamily in the seven genomes.

TIRs and Helitrons length follow a similar profile in all species. TIRs with the presence of a large peak at 1000 bp and then a small spike between 3500 and 4000 bp while Helitrons have a longer length and a gradually decrease from 1000 bp to 20000 bp. MITEs follow two different dynamics, showing a common peak at 200 bp and then, *P. salicina, P. armeniaca* and *P. avium* Tieton presents the peak at 400 bp while *P. dulcis* Lauranne*, P. dulcis* Texas*, P. persica* and *P. mira* the peak is at 550 bp.

LTR retrotransposons showed contrasting patterns between the two superfamilies. Copia length distribution was practically the same in all species with the presence of a large peak at 5000 bp, except for *P. salicina* which showed a second peak at 10000 bp. This second peak could reflect the amplification of a specific family of long Copia elements only in *P. salicina*. Alternatively, it could also be due to a misclassification by EDTA, as this second peak matches with the main peak observed in Gypsy at 10000 bp for most species. Also, it is remarkable that the number of Copia copies at about 5000 bp is highly variable, being *P. mira* followed by *P. persica* the two species with higher presence with approximately 300 copies while *P. avium* has less than 100 copies (**Figure 10**).

As mentioned above, Gypsy showed a main peak at 10000 bp in most species, but also species-specific peaks at 5000 bp and another at 15000 bp. The peak at 5000 bp also matches with the main peak of the Copia. Following the previous reasoning, we hypothesize that either it was a miss classification of EDTA or that there is a lineage of Gypsy with a length of 5000 bp. Given the high precision in LTR retrotransposon classification reported in a recent benchmark (Ou et al., 2019), the second hypothesis is more plausible. Based on the results shown in the frequency plot (**Figure 10**), *P. salicina* shows a much higher number of Gypsy elements of all reported lengths.

In section 4.2 we stated that LTR retrotransposons occupy a larger genome fraction than DNA transposons. Nevertheless, intact DNA transposons double the number LTR retrotransposons (**Table 7**). As shown in **Figure 8** and **Figure 9**, DNA transposons are smaller than LTR retrotransposons thus, representing a lower percentage within the genome besides its higher copy number. Another reason that explains this apparent contradiction is that the proportion of intact/degenerated elements is much higher for DNA transposons than for LTR retrotransposons.

Table 7. *Classification of intact TEs.*

Species	RNA transposons				DNA transposons				
	Total	**Copia**	**Gypsy**	**LTR/Unknown**	**Total**	**MITEs**	**TIRs**	**Helitrons**	
P. armeniaca	877	393	327	157	2471	787	1684	87	3435
P. avium Tieton	488	283	77	128	2477	891	1586	100	3065
P. dulcis Lauranne	930	496	143	291	1771	763	1008	71	2772
P. dulcis Texas	702	340	126	236	1979	844	1135	95	2681
P. mira	1106	642	281	183	2470	977	1493	83	3659
P. persica	1489	633	311	545	2005	812	1193	169	3663
P. salicina	2605	785	1158	662	3495	1129	2366	88	6188
	8197	3572	2423	2202	16668	6203	10465	693	**Total**

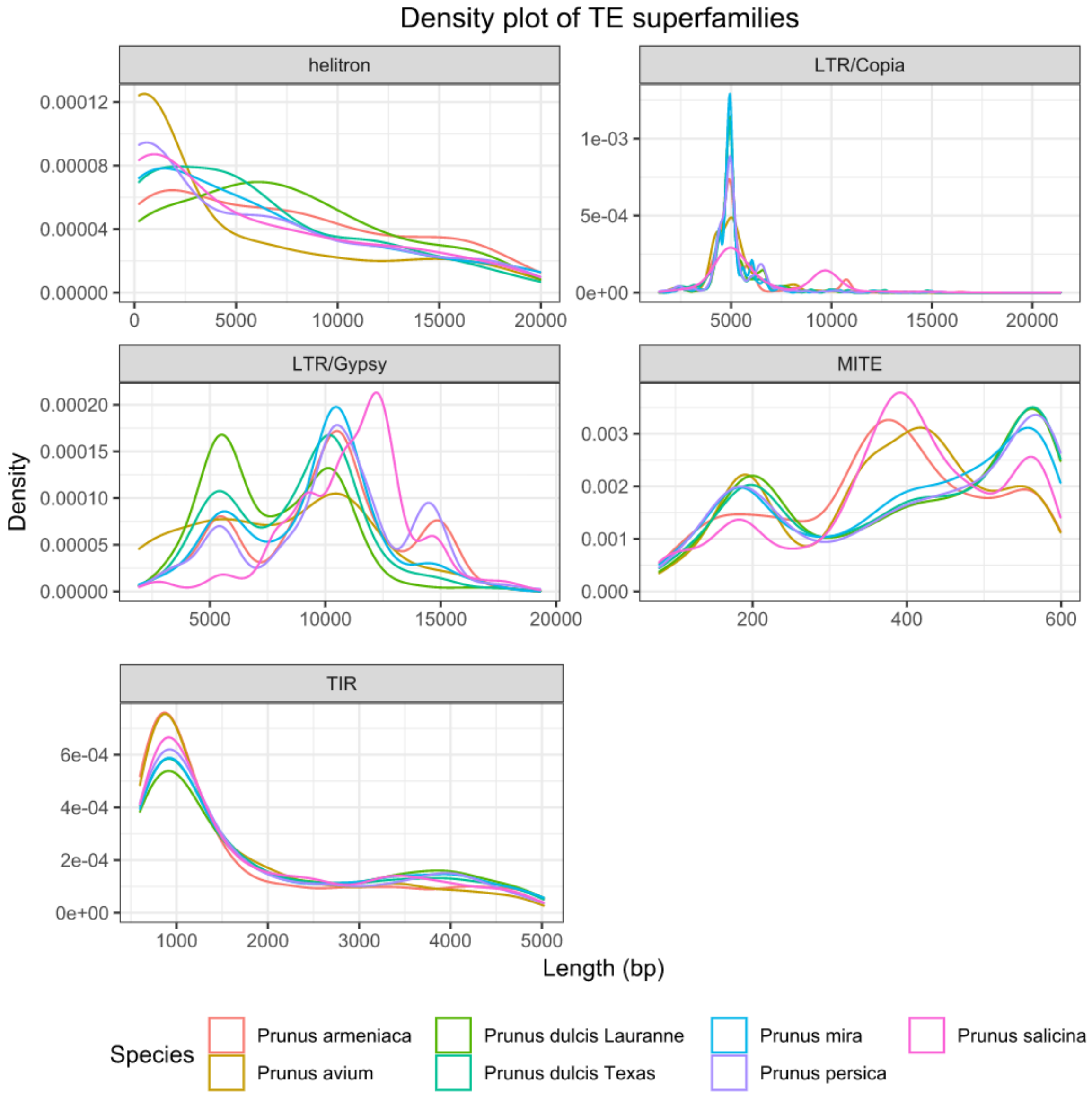


Figure 9. Density plot representation of intact TEs length in base pair long for each superfamily and *Prunus* species.

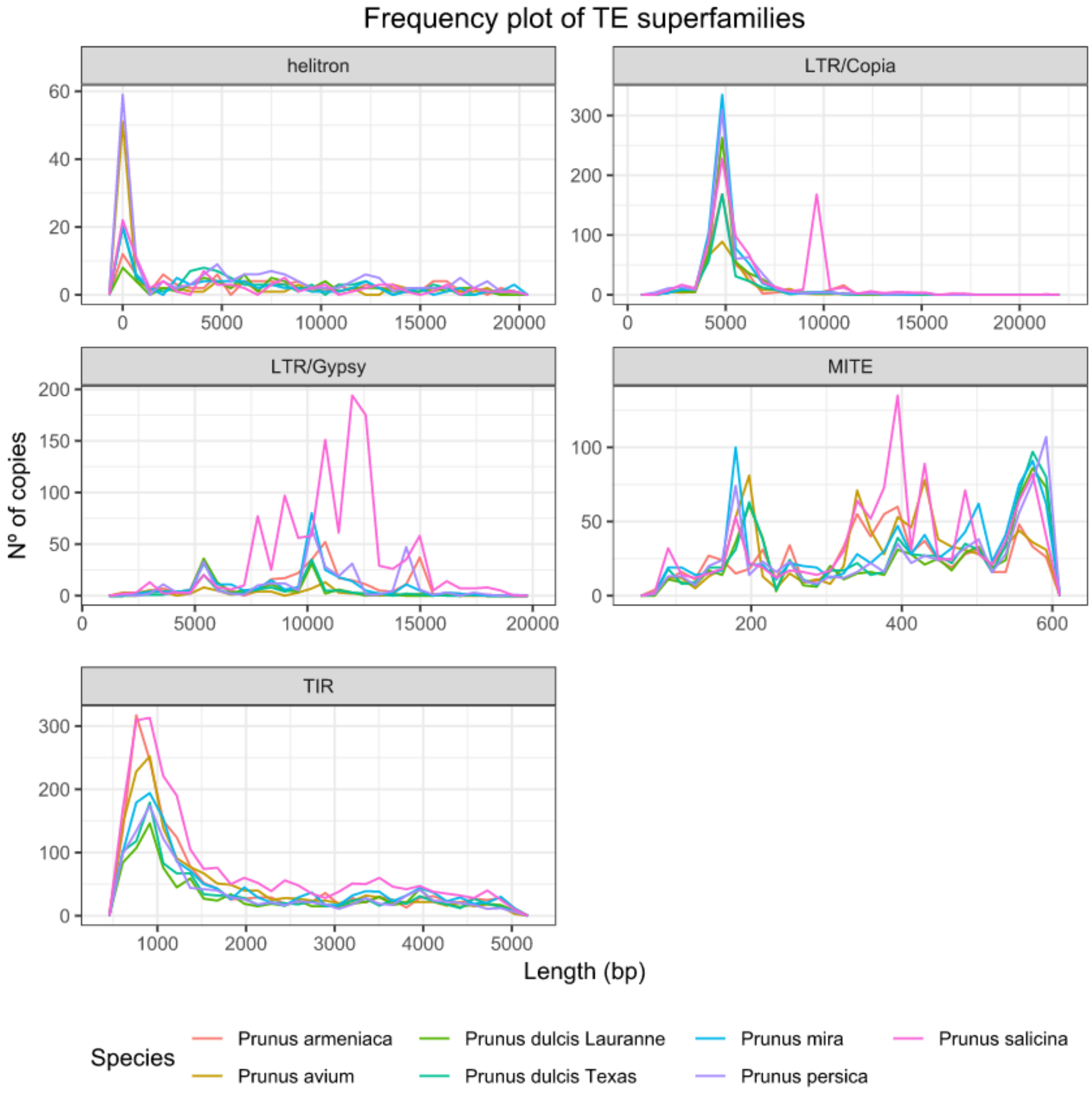


Figure 10. Frequency plot representation of intact TEs length in base pair long for each superfamily and *Prunus* species.

Based on all the comparative analyses presented, we observed that *P. salicina* had a higher genome size, higher content of potentially active TEs, different TE length distribution and, above all, a much higher presence of Gypsy elements compared to the other species. Two hypotheses are proposed to explain these important differences:

i) As mentioned in the previous section 4.1, the genome assembly quality could negatively affect the annotation of these TEs, by increasing false negatives and therefore artificially decreasing the number of Gypsy in the species with lower genome assembly quality. This can apply to some genomes (i.e. *P. dulcis* Texas) but can be ruled out as a general explanation, given that *P. salicina*, *P. armeniaca* or *P. persica* have a high LAI value and there are large differences among them in the Gypsy content.

ii) A recent burst of TE activity may have taken place in *P. salicina*, especially in the Gypsy superfamily, after the split with its closest species analysed (*P. armeniaca*). This is in concordance with its higher genome size and chromosomes length. To test this hypothesis, we calculated the insertion time of LTR-retrotransposons, and explored the distribution of Gypsy along the chromosomes of these two species*,* as Gypsy tend to insert in centromeric and pericentromeric regions, if a recent burst of Gypsy elements has occurred in *P. salicina*, these regions may have recently expanded in comparison with *P. armeniaca*.

4.4. Insertion time of intact TEs

Insertion time of intact LTR-retrotransposons was calculated based on the divergence of their two Long Terminal Repeats, given that, at the time of insertion both are identical (SanMiguel et al., 1998). Using the LTR-retriever module of the EDTA package, we obtained the estimated insertion times (in million years, MY) of LTR-retrotransposons on the seven genomes analysed. Insertion time was represented either as density plots and frequency plots (**Figure 11**).

We found that most insertions of all *Prunus* species occurred in the last 5 million years (**Figure 11**), and especially in *P. salicina* we see a strong increase of both Copia and Gypsy at 1 MY. This increase in TEs can be explained by a recent high TE activity which led to the generation of new insertions increasing the TE copy number which is directly related with genome size, since an increase in copies implies an increase in genome size. As already explained in section 4.1, *P. salicina* genome size is about 280 Mb being the largest genome of all P*runus* species and the number of copies is higher compared to the other species (section 4.2). This profile of LTR retrotransposon activity in *P. salicina* is compatible with a recent amplification burst after the split of *P. salicina* and *P. armeniaca* species, giving strength to our hypothesis.

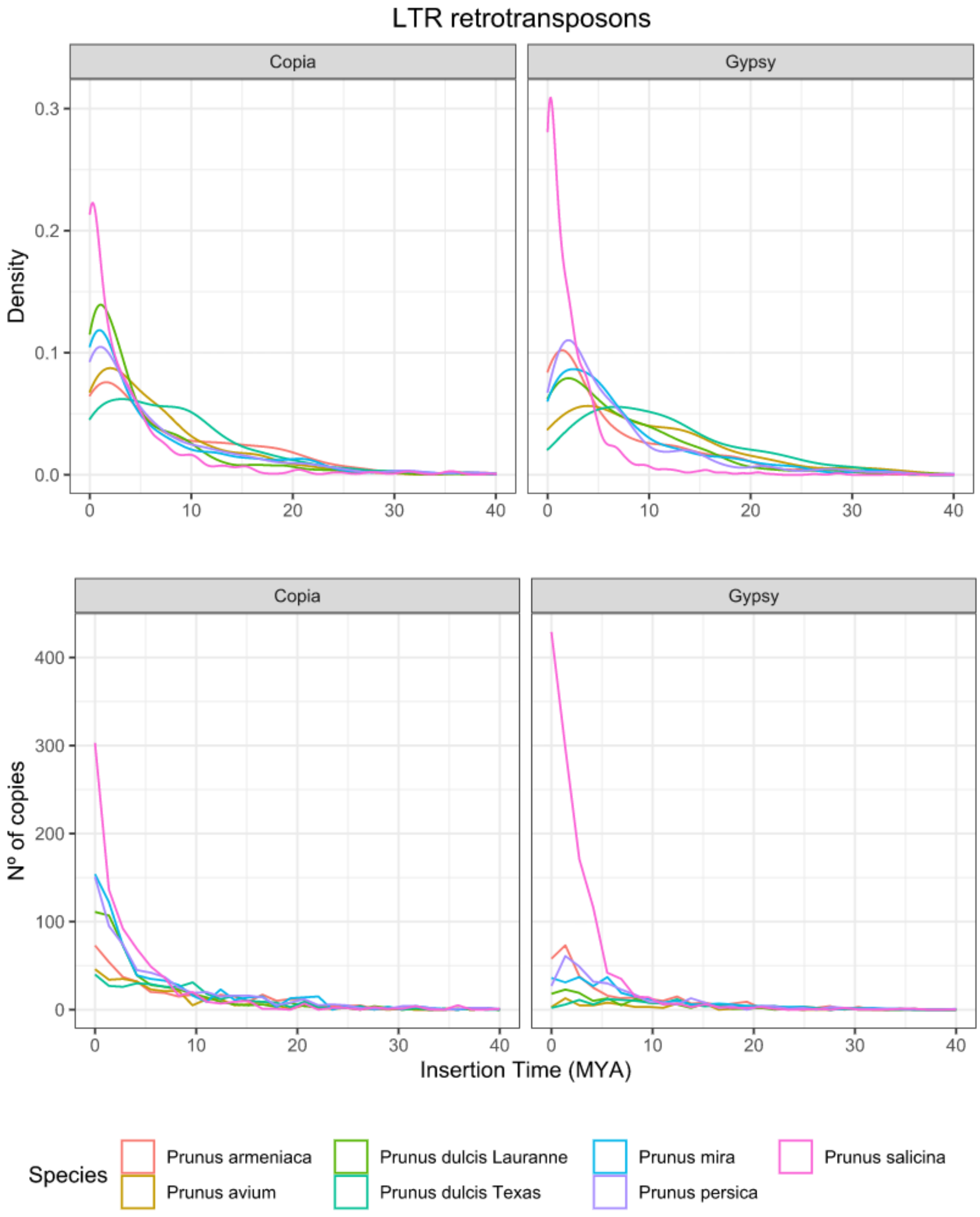


Figure 11. Representation of the estimated insertion time in million years ago (MYA) in the seven *Prunus* species.

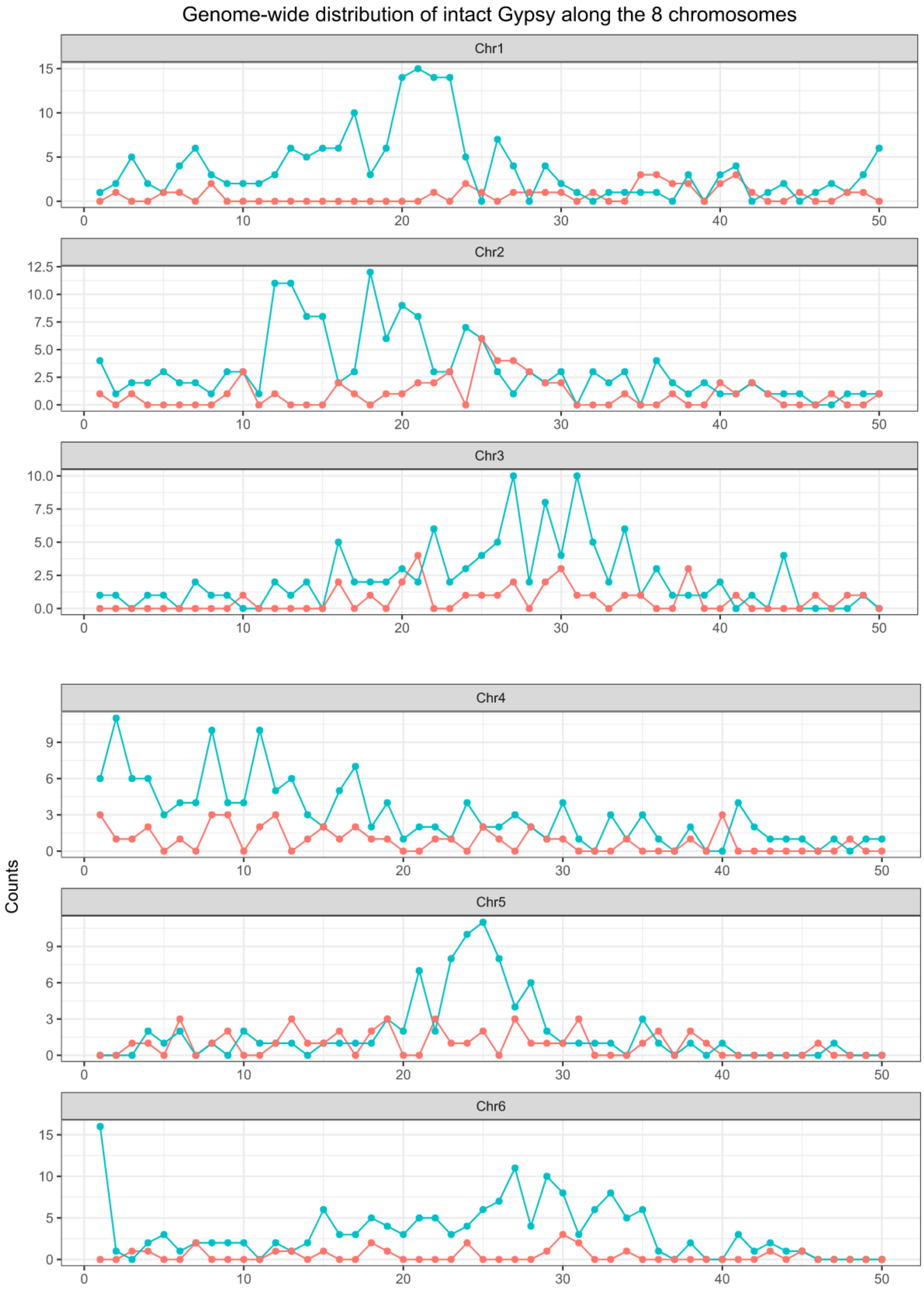
4.5. Distribution of LTR/Gypsy retrotransposons along *P. salicina* and *P. armeniaca* chromosomes

To verify the putative pericentromeric expansion of *P. salicina,* a comparative analysis was made between *P. salicina* and *P. armeniaca* Gypsy superfamily (**Figure 12** and **Figure 13**) as their phylogenetic distance is smaller than to the other *Prunus* species under analysis. In addition, as LAI index is very high in both species any possible technical bias can be ruled out.

P. salicina showed a higher number of Gypsy elements in all chromosomes, especially in the pericentromeric regions, and a lower TE density regions represents the chromosomal arms.

The highest content of Gypsy elements was found in chromosome 1 of *P. salicina* (the largest chromosome, almost twice as large as the others), having a maximum peak in the pericentromeric region of about 500 copies when in the other chromosomes it does not exceed 300 copies (**Figure 13**).

These results confirm that the pericentromeric region of *P. salcina* is more expanded with respect to *P. armeniaca* and, presumably, with respect to the other *Prunus* due to the recent insertion of Gypsy LTR retrotransposons elements. These results are in line to which was described in the melon genome.


Genome-wide distribution of intact Gypsy along the 8 chromosomes
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Counts
15
10
5
0
12.5
10.0
7.5
5.0
2.5
0.0
9
6
3
0
10
20
30
40
50

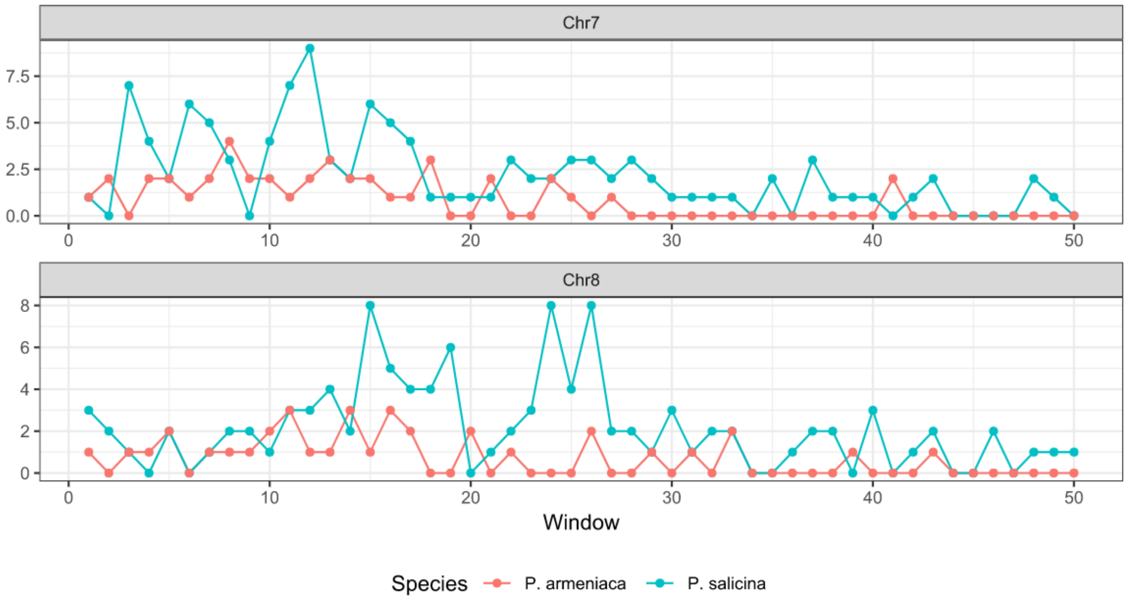
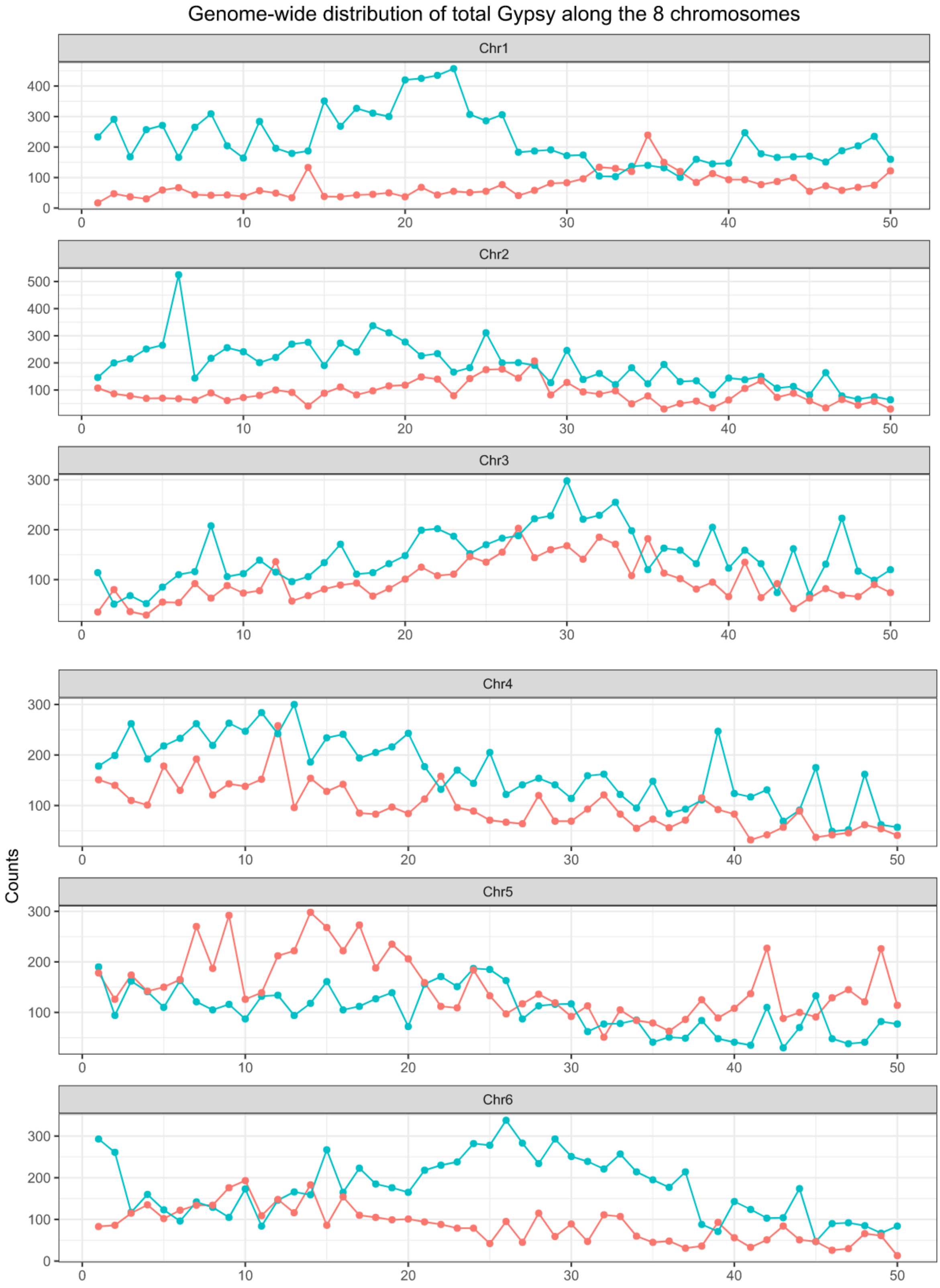


Figure 12. Intact Gypsy distribution along each chromosome in *P. salicina* and *P. armeniaca*.


Genome-wide distribution of total Gypsy along the 8 chromosomes
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Counts
0
100
200
300
400
500
0
10
20
30
40
50

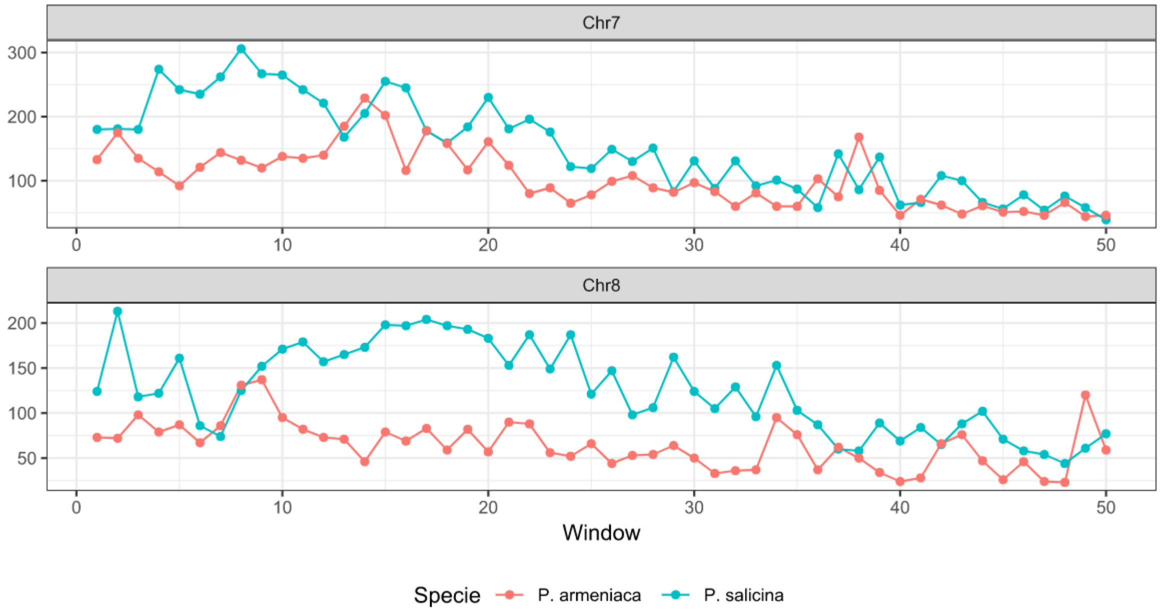


Figure 13. Total Gypsy distribution along each chromosome in *P. salicina* and *P. armeniaca*.

4.6. TE based *Prunus* pangenome

The increasing availability of plant pangenomes is changing crop genomics and improvement, and structural variation is gaining importance as a source of phenotypic variation (Castanera et al., 2021; Domínguez et al., 2020). In this section, our interest is to see if TEs follow the same dynamic as genes in a pangenome. Previous studies have shown an important fraction of genes being present in most genomes (*core* genes, which are necessary for the survival of the organism), whereas a smaller fraction is found only in a few genomes (“*dispensable genes”* which would imply that their function is not essential). In general, the distribution of gene conservation in gene-based pangenome are represented as histograms, and follow a "U-shape" dynamic, with an excess of genes present in most genomes (Cao et al., 2020).

In order to carry out this analysis, we clustered together all similar TEs into common families. For this we used CD-HIT at 80 % identity, meaning that those transposons whose sequence identity is equal or higher than 80% are considered the same transposon family (also referred as “cluster”), while those with a similarity lower than 80% are considered a different family. Then, for each cluster, we determined how many species have copies inside (these copies would be analogous to “homologous genes” in a gene-based pangenome), and how many copies each species had.

For the creation of these histograms, *P. dulcis* Lauranne was excluded as the presence of two varieties of the same species could artificially affect the result enriching the TEs found in *P. dulcis*. These histograms have been plotted with the presence and absence of singletons (clusters that have only 1 copy). In the case of MITEs, due to a previous filter which those clusters with fewer than 3 copies have been eliminate, it was not necessary to remove singletons as they were removed with the filter.

The result of this analysis is presented in **Figure 14**. For this analysis we do not take into account singletons (clusters with only one sequence, **Annex B. Figure S1**), as singletons modify the result by artificially increasing the number of clusters in only one species and have higher chance to be false positives or TEs with multiple structural variations inside making difficult to interpret the results.

We observed strong differences in the dynamics of LTR retrotransposon and MITES. Copia (**Figure 14A**) and Gypsy (**Figure 14B**) showed a higher presence of families present on 1 or 2 species (150 families) while very few were present in 3 or more species. This result can be explained by a strong recent activity of LTR retrotransposons in *Prunus* genus and a very fast turnover. The structural modifications that LTR retrotransposons cause after transposition are often not advantageous or even deleterious for the organism and therefore difficult to fix in a population.

If we look at Copia superfamily, in proportion they are slightly more conserved than LTR/Gypsy as they have a higher prevalence in all the genomes, a result that may be explained by their differences in recent activity (**Figure 11**).

On the other hand, MITEs (**Figure 14C**) are more conserved as their presence in only one species (40 families) is similar to the presence at all the species (approx. 30 families). This result could be partially explained because MITEs activity in the genome is lower compared to LTR retrotransposons activity. Also, due to MITEs are smaller (200-600 bp, **Figure 10**) and their impact in the genome is less deleterious.

In general, Copia and Gypsy profile found in the *Prunus* genus follows a similar dynamic as Copia and Gypsy found in rice populations (*Oryza sativa*) as a result of a recent activity and a high turnover. Focusing on MITEs dynamics, we can see that it follows a "U-shape" again similar to what has been recently found in rice (Castanera et al., 2021). It is worth mention that MITEs are often found close or within genes, sometimes playing a regulatory role. Consequently, if they are close to *core* genes they will persist more over time being present in all species, but if they are found in *dispensable* genes, they will be found in fewer species (Alioto et al., 2020; Schrader & Schmitz, 2019).

As a next step, we used our pangenome matrices to perform hierarchical clustering and observed if the result matched the already know phylogenetic relationships of *Prunus* species. The analyses were performed with the binary matrix (**Figure 15**) and with the abundance matrix (**Annex B, Figure S2**). The results showed a strong congruency with the sequence-based phylogeny of the *Prunus* genus (**Figure 2**).

We observed that, independently of the TE group, the clustering resolved well the phylogenetic relationships of *P. dulcis* (Texas and Lauranne) as well as the two peach trees (P. *mira* and *P. persica*). The topology of the trees slightly varied depending on TE group, being that of MITEs the one that better matched with the phylogeny of the *Prunus* genus (**Figure 2**). Previous studies have proposed that TEs could be used as molecular markers (Ruslan Kalendar and Alan H. Schulman, 2007). Our results suggest that TE polymorphisms at the family level can be used to reconstruct the phylogenetic relationships of close species. This, combined with recent data stressing their potential impact on agronomic traits make TEs a very promising molecular markers for future studies on crop plants.

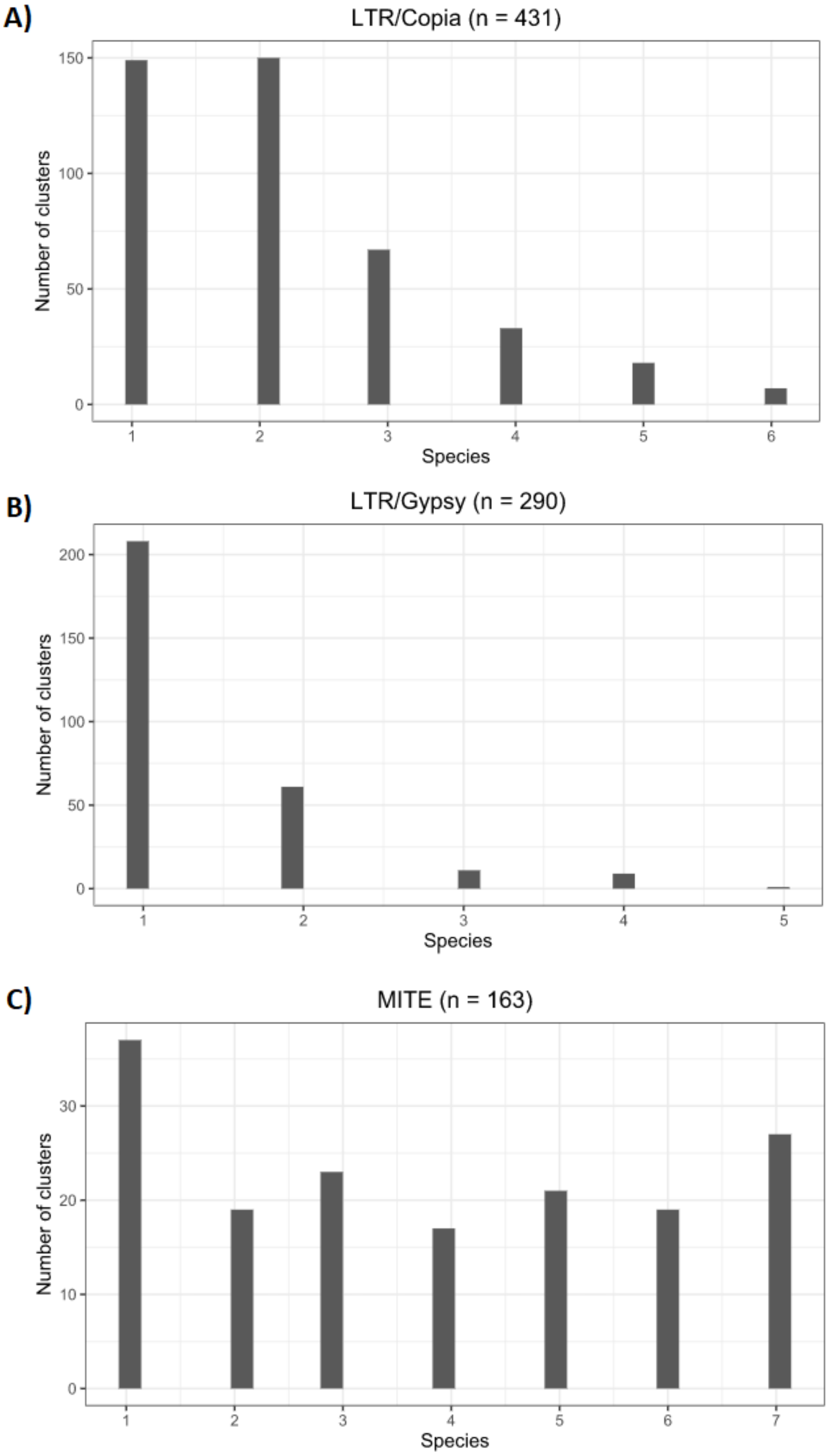


Figure 14. Number of species found in each cluster (without singletons) represented in histograms. A) Copia, B) Gypsy and C) MITEs.

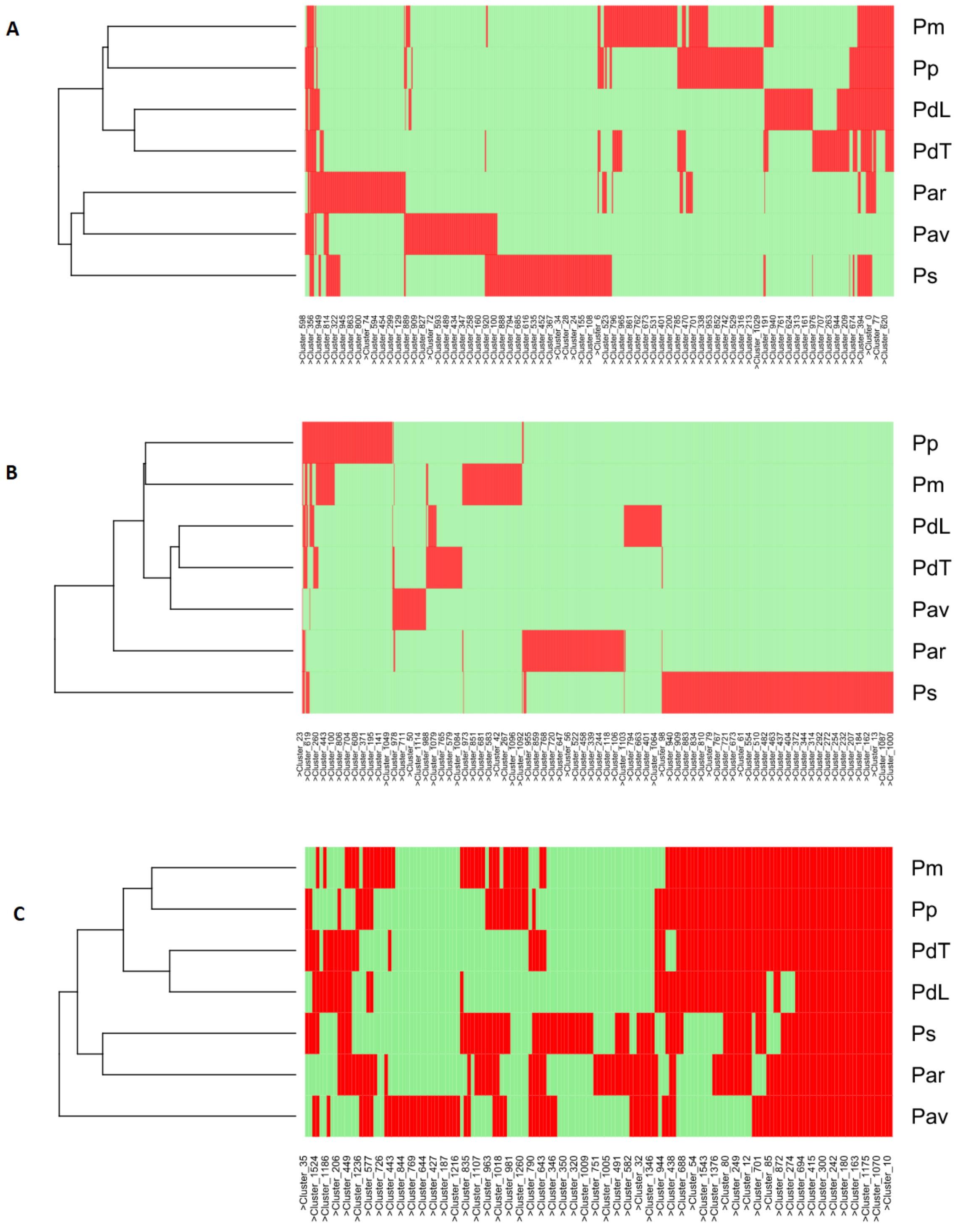


Figure 15. Phylogeny of *Prunus* genus based on TEs presence from binary table, red represents the presence of TEs and green the absence and A) Copia B) Gypsy and C) MITEs.

4.7. Detection of polymorphic TE insertions in peach varieties

Using the *Prunus* TE library (created in this work) and the re-sequencing data of *P. persica* EarlyGold, we tried to find TE insertions specific to this variety by using Jitterbug (Hénaff et al., 2015), a program developed by the group in which this study has been carried out (CRAG - "Structure and evolution of plant genomes"). The objective of this is to understand if a *Prunus* generic TE library (interspecific analysis) can be suitable for this kind of analyses (intraspecific analysis), and to obtain a first idea of the level of TE polymorphisms among cultivated peaches. We found 801 TE insertions that were specific to this variety (absent in the reference *P. dulcis* Texas genome). Among them, 297 were inserted inside genes (including introns or exons). These insertions potentially can have an impact on the target gene affecting its expression. For the validation of this results, we selected 4 possible TEs insertions that were predicted by Jitterbug to be inserted into genes. PCR primers were designed in the flanking regions of the TE insertion.

After the PCR and the agarose gel electrophoresis (**Figure 16**) 2 of our predicted TEs (sample 1 and sample 2, **Figure 16**) insertions were validated as there is a presence of a double band, which means that the insertion was heterozygote. The band with a bigger size (BG) which points to the insertion, is named in sample 1 as BG1 with a size ranging from 1000 to 1500 bp and in sample 3 as BG3 with a size of approx. 500 bp while the smaller band which matches the size of the DNA fragment if there were no insertion (empty site, BP) is named in sample 1 as BP1 and in sample 3 as BP, both bands of both samples ranging from 350 to 450 bp. The result for the other two samples with a single band (samples 2 and 4, **Figure 16**) imply that there may not be a transposon, or that we could not be amplified by using the primers and PCR conditions used due to a very large insertion.

To ensure that the amplified bands were specific, and not false positives, we purified the PCR products from the gel and conducted Sanger sequencing (**Annex C, sequencing dataset**). We used the sequences obtained as input for a homology search (BLASTN, **Script 14**) to the genome assembly and to the Prunus TE library. As expected, the sequences from the bands at lower molecular weight matched the region were the primers were designed.

Finally, the BLASTN performed with the sequences with possible insertions, samples 1 and 3, as they have the possible TEs insertion and as both bands (BG1 and BG3, **Figure 16**) unequivocally identified regions of a TIR and an Copia with a significant e-value (**Table 8**) confirming the presence of a TEs insertion.

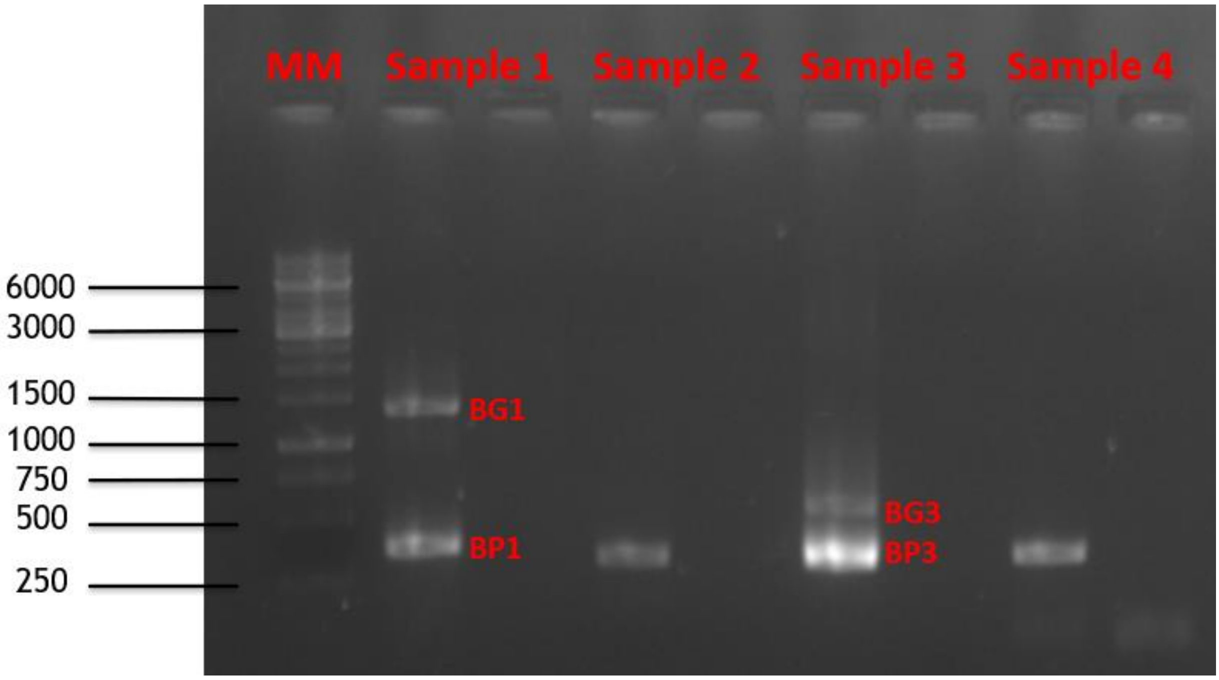


Figure 16. Agarose gel electrophoresis of the 4 possible TEs insertion within genes. MM refers to molecular marker (**Figure 7**), BG ("Banda Grande") refers to possible TEs insertion and BP ("Banda pequeña") refers to absence of TE insertion in those samples. Negative controls were run in the adjacent wells for each sample.

Table 8. *Results of Blast against Prunus_TE_library and P. persica genome.*

	Band ID	Query	Start	End	E-value	Identity
		Blast *VS* Prunus_TE_library				
Sample 1	BG1	Pav_230#TIR	272	380	7,22e^{-20}	82,9
	BP1	-	-	-	-	-
Sample 2	Sample 2	-	-	-	-	-
Sample 3	BG3	Pm_247#LTR/Copia	112	176	1,42e^{-20}	93,9
	BP3	-	-	-	-	-
Sample 4	Sample 4	-	-	-	-	-
		Blast *VS* Peach genome				
Sample 1	BG1	Pp08	16558294	16558399	1,02e^{-37}	94,3
	BP1	Pp08	16558289	16558620	1,49e^{-167}	98,8
Sample 2	Sample 2	Pp04	6366670	6366936	3.50e^{-133}	98,9
Sample 3	BG3	Pp07	15412604	15412714	3.29e^{-46}	97,3
	BP3	Pp07	15403525	15403791	1,73e^{-52}	81,0
Sample 4	Sample 4	Pp04	9624392	9624696	5.09e^{-147}	97,7

5. Conclusions

The research carried out during these months has allowed us to draw the following conclusions:

- The sequencing method affects the quality of the genome assembly, which directly affects the TE annotation. PacBio and Sanger sequencing are the ones that allow to obtain the best assembly quality.

- *P. salicina* genome has been recently expanded due to the activity of LTR retrotransposons, especially in the Gypsy superfamily. These insertions are mostly inserted into pericentromeric regions of chromosomes 1, 2 and 6.

- Copia and Gypsy LTR retrotransposons are less conserved than DNA transposons across the *Prunus* species. By contrast, MITEs tend to be conserved, with many families found present in all the species.

- TE polymorphisms at the family level reflect the phylogenetic relationships of close species.

- The *Prunus* TE library build in this work is useful for the detection of TE polymorphisms in any *Prunus* species using re-sequencing data and a unified classification system based on common TE families.

- We were able to validate the presence of two heterozygous TE insertions potentially affecting genes in peach varieties.

6. Bibliography

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., Dhingra, A., Duval, H., Fernández i Martí, Á., *et al.* (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *Plant Journal*, *101*(2), 455–472. doi: 10.1111/tpj.14538

Arús, P., Verde, I., Sosinski, B., Zhebentyayeva, T., & Abbott, A. G. (2012). The peach genome. *Tree Genetics and Genomes*, *8*(3), 531–547. doi: 10.1007/s11295-012-0493-8

Asma, B. M., & Ozturk, K. (2005). Analysis of morphological, pomological and yield characteristics of some apricot germplasm in Turkey. *Genetic Resources and Crop Evolution*, *52*(3), 305–313. https://doi.org/10.1007/s10722-003-1384-5

Badenes, M. L., & Byrne, D. H. (2012). Fruit breeding. [Online] Springer-Verlag New-York, 2012. ISBN: 978-1-4419-0763-9. [Consulted on 10th of June, 2021]. doi: 10.1007/978-1-4419-0763-9

Bao, W., Wuyun, T., Li, T., Liu, H., Jiang, Z., Zhu, X., Du, H., & Bai, Y. e. (2017). Genetic diversity and population structure of Prunus mira (Koehne) from the Tibet plateau in China and recommended conservation strategies. *PLoS ONE*, *12*(11), 1–19. doi: 10.1371/journal.pone.0188685

Bayer, P. E., Golicz, A. A., Scheben, A., Batley, J., & Edwards, D. (2020). Plant pan-genomes are the new reference. *Nature Plants*, *6*(8), 914–920. doi: 10.1038/s41477-020-0733-0

Bentley, D. R. (2006). Whole-genome re-sequencing. *Current Opinion in Genetics and Development*, *16*(6), 545–552. doi: 10.1016/j.gde.2006.10.009

Borgognone, A. (2017). *Characterization of transposon activity and genome-wide epigenetic regulation throughout the life cycle of Pleurotus ostreatus*. Doctoral Thesis, UPNA, Agricultural Prodution Department, 2017 [Consulted on 1st of June, 2021]. Available at: https://hdl.handle.net/2454/32180

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., *et al.* (2018). Ten things you should know about transposable elements.

Genome Biology, *19*(1), 1–12. doi: 10.1186/s13059-018-1577-z

Cao, K., Peng, Z., Zhao, X., Li, Y., Liu, K., Arus, P., Zhu, G., Deng, S., Fang, W., *et al.* (2020). Pan-genome analyses of peach and its wild relatives provide insights into the genetics of disease resistance and species adaptation. *BioRxiv* doi: 10.1101/2020.07.13.200204

Cao, K., Zheng, Z., Wang, L., Liu, X., Zhu, G., Fang, W., Cheng, S., Zeng, P., Chen, C., *et al.* (2014). Comparative population genomics reveals the domestication history of the peach, Prunus persica, and human influences on perennial fruit crops. *Genome Biology*, *15*(7), 1–15. doi: 10.1186/s13059-014-0415-1

Carnell, A. N., & Goodman, J. I. (2003). The long (LINEs) and the short (SINEs) of it: Altered methylation as a precursor to toxicity. *Toxicological Sciences*, *75*(2), 229–235. doi: 10.1093/toxsci/kfg138

Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M., Panaud, O., & Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *The Plant Journal*, 1–18. doi: 10.1111/tpj.15277

Chin, S. W., Shaw, J., Haberle, R., Wen, J., & Potter, D. (2014). Diversification of almonds, peaches, plums and cherries - Molecular systematics and biogeographic history of Prunus (Rosaceae). *Molecular Phylogenetics and Evolution*, *76*(1), 34–48. doi: 10.1016/j.ympev.2014.02.024

Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene*, *410*(1), 9–17. doi: 10.1016/j.gene.2007.12.008

Craig Venter, J., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., *et al.* (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351. doi: 10.1126/science.1058040

Creighton, H. B., & McClintock, B. (1931). The Correlation of Cytological and Genetical Crossing-Over in Zea Mays. A Corroboration. *Proceedings of the National Academy of Sciences*, *17*(8), 492–497. doi: 10.1073/pnas.21.3.148

Crescente, J. M., Zavallo, D., Helguera, M., & Vanzetti, L. S. (2018). MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics*, *19*(1), 1–10. doi:10.1186/s12859-018-2376-y

Delplancke, M., Alvarez, N., Benoit, L., Espíndola, A., Joly, H. I., Neuenschwander, S., & Arrigo, N. (2013). Evolutionary history of almond tree domestication in the Mediterranean basin. *Molecular Ecology*, *22*(4), 1092–1104. doi: 10.1111/mec.12129

Doyle, J.J., Doyle, J.L. (1990) Isolation of plant DNA from fresh tissue. *Focus*, (12) 13-15

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J. M., Colot, V., & Quadrana, L. (2020). The impact of transposable elements on tomato diversity. *Nature Communications*, *11*(1). doi: 10.1038/s41467-020-17874-2

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138. doi: 10.1126/science.1162986

Fambrini, M., Usai, G., Vangelisti, A., Mascagni, F., & Pugliesi, C. (2020). The plastic genome: The impact of transposable elements on gene functionality and genomic structural variations. *Genesis*, *58*(12). doi: 10.1002/dvg.23399

Fedoroff, N. V. (1994). Barbara McClintock - June 16, 1902-September 2, 1992. *Biographical Memoirs. National Academy of Sciences (U.S.)*, *136*, 1–10.

Feng, Y. (2003). Plant MITEs: useful tools for plant genetics and genomics. *Genomics, Proteomics & Bioinformatics/Beijing Genomics Institute 1*(2), 90–100 doi: 10.1016/s1672-0229(03)01013-1

Finnegan, D. J. (1989). *Eukaryotic transposable elements and genome evolution*. *5*(4), 1–10. doi: 10.1016/0168-9525(89)90039-5

Finnegan, D. J. (2012). Retrotransposons. *Current Biology*, *22*(11), 432–437. doi: 10.1016/j.cub.2012.04.025

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering transposable element diversification in de novo annotation approaches. *PLoS ONE*, *6*(1). doi: 10.1371/journal.pone.0016526

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(17), 9451–9457. doi: 10.1073/pnas.1921046117

Gilbert, W., & Maxam, A. (1973). The nucleotide sequence of the lac operator. *Proceedings of the National Academy of Sciences of the United States of America*, *70*(12 (I)), 3581–3584. doi: 10.1073/pnas.70.12.3581

Havecker, E. R., Gao, X., & Voytas, D. F. (2004). The diversity of LTR retrotransposons. *Genome Biology*, *5*(6). doi: 10.1186/gb-2004-5-6-225

Hénaff, E., Zapata, L., Casacuberta, J. M., & Ossowski, S. (2015). Jitterbug: Somatic and germline transposon insertion detection at single-nucleotide resolution. *BMC Genomics*, *16*(1), 1–16. doi: 10.1186/s12864-015-1975-5

Jiang, F., Zhang, J., Wang, S., Yang, L., Luo, Y., Gao, S., Zhang, M., Wu, S., Hu, S., *et al.* (2019). The apricot (Prunus armeniaca L.) genome elucidates Rosaceae evolution and beta-carotenoid synthesis. *Horticulture Research*, *6*(1). doi: 10.1038/s41438-019-0215-6

Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., & Walichiewicz, J. (2005). Repbase Update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research*, *110*(1–4), 462–467. doi: 10.1159/000084979

K. Pace II, J., & Feschotte, C. (2007). *The evolutionary history of human DNA transposons: Evidence for intense activity in the primate lineage*. *06*(817), 422–432. doi: 10.1101/gr.5826307.422

Kapitonov, V. V., & Jurka, J. (2008). A universal classification of eukaryotic transposable elements implemented in Repbase. *Nature Reviews Genetics*, *9*(5), 411–412. doi: 10.1038/nrg2165-c1

Khan, A. W., Garg, V., Roorkiwal, M., Golicz, A. A., Edwards, D., & Varshney, R. K. (2020). Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends in Plant Science*, *25*(2), 148–158. doi: 10.1016/j.tplants.2019.10.012

Kim, N. (2017). *The genomes and transposable elements in plants: are they friends or foes? 39*, 359–

370. doi: 10.1007/s13258-017-0522-y

Kobayashi, S., Goto-Yamamoto, N., & Hirochika, H. (2004). Retrotransposon-Induced Mutations in Grape Skin Color. *Science*, *304*(5673), 982. doi: 10.1126/science.1095011

Ladizinsky, G. (1990). *On the origin of almond* (pp. 143–147).

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., et al. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. doi: 10.1038/35057062

Li, W., & Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, *22*(13), 1658–1659. doi: 10.1093/bioinformatics/btl158

Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*(1), 49–61. doi: 10.1038/nrg3374

Liu, C., Feng, C., Peng, W., Hao, J., Wang, J., Pan, J., & He, Y. (2021). Chromosome-level draft genome of a diploid plum (Prunus salicina). *GigaScience*, *9*(12), 1–11. doi: 10.1093/gigascience/giaa130

Lu, C., Chen, J., Zhang, Y., Hu, Q., Su, W., & Kuang, H. (2012). Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. *Molecular Biology and Evolution*, *29*(3), 1005–1017. doi: 10.1093/molbev/msr282

Mao, H., Wang, H., Liu, S., Li, Z., Yang, X., Yan, J., Li, J., Tran, L. S. P., & Qin, F. (2015). A transposable element in a NAC gene is associated with drought tolerance in maize seedlings. *Nature Communications*, *6*, 1–7. doi: 10.1038/ncomms9326

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380. doi: 10.1038/nature03959

Martínez-gómez, P., Sánchez-pérez, R., Dicenta, F., Howad, W., & Arús, P. (2007). Fruits and Nuts.

Chittaranjan Kole (July 2014). doi: 10.1007/978-3-540-34533-6. ISBN:9783540345336

Matsumoto, T., Wu, J., Kanamori, H., Katayose, Y., Fujisawa, M., Namiki, N., Mizuno, H., Yamamoto, K., Antonio, B. A., … Burr, B. (2005). The map-based sequence of the rice genome. *Nature*, *436*(7052), 793–800. doi: 10.1038/nature03895

Maxam, A. M., & Gilbert, W. (1977). A new method for sequencing DNA. *Biotechnology (Reading, Mass.)*, *74*(2), 560–564. doi: 10.1073/pnas.74.2.560

McClintock, B. (1945). *Cytogenic studies of maize and neurospora*. *44*, 108–112.

McClintock, B. (1950). The origin and behaviour of mutable loci in Maize. *Proceedings of the National Academy of Sciences of the United States of America*, *36*(6), 344–355. doi: 10.1073/pnas.36.6.337

McNeill, J., Barrie, F. R., Buck, W. R., Demoulin, V., Greuter, W., Hawksworths, D. L., Herendeen, P. S., Knapp, S., Marhold, K., *et al.* (2012). International Code of Nomenclature for algae, fungi and plants (Melbourne Code) adopted by the Eighteenth International Botanical Congress Melbourne, Australia, July 2011. Hardback, ISBN: 9783874294256

Morata, J., Marín, F., Payet, J., & Casacuberta, J. M. (2018). Plant lineage-specific amplification of transcription factor binding motifs by miniature inverted-repeat transposable elements (MITEs). *Genome Biology and Evolution*, *10*(5), 1210–1220. doi: 10.1093/gbe/evy073

N.I Vavilov. (1951). *The Origin, Variation, Immunity, and Breeding of Cultivated Plants*. 115(2990):433-434. doi: 10.1126/science.115.2990.433-a

Orozco-Arias, S., Isaza, G., & Guyot, R. (2019). Retrotransposons in plant genomes: Structure, identification, and classification through bioinformatics and machine learning. *International Journal of Molecular Sciences*, *20*(15). doi: 10.3390/ijms20153837

Ou, S., Chen, J., & Jiang, N. (2018). Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Research*, *46*(21), e126. doi: 10.1093/nar/gky730

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., Lugo, C. S. B., *et al.* (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biology*, *20*(1), 1–18. doi: 10.1186/s13059-019-1905-y

Poczai, P., Cernák, I., Varga, I., & Hyvönen, J. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Genetic Resources and Crop Evolution*, *408*(1), 796–815. doi: 10.1134/S1022795411020074

Potter, D., Gao, F., Bortiri, P. E., Oh, S. H., & Baggett, S. (2002). Phylogenetic relationships in Rosaceae inferred from chloroplast matK and trnL-trnF nucleotide sequence data. *Plant Systematics and Evolution*, *231*(1–4), 77–89. doi: 10.1007/s006060200012

Pray, L. (2008). Transposons: The jumping genes. *Nature Education*, *1*(1), 204. http://www.nature.com/scitable/topicpage/transposons-the-jumping-genes-518

Qin P et. al. (2021). Pan-genome analysis of 33 genetically diverse rice accessions reveals hidden genomic variations. *Cell* 184(13), 3542-3558, e16, doi: 10.1016/j.cell.2021.04.046

Ravindran, S. (2012). Barbara McClintock and the discovery of jumping genes. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(50), 20198–20199. doi: 10.1073/pnas.1219372109

Ribeiro Serra, O. M. (2017). Towards increasing genetic variability and improving fruit quality in peach using genomic and bioinformatic tools. [Online] Doctoral Thesis, UAB, Departament of animal biology, vegetal biology and ecology [consulted in 4th of June of 2021] available at: handle/10803/460882

Road, H. (1979). A strategy of DNA sequencing employing computer programs. *Nucleic Acids Research*, *6*(7), 2601–2610. doi: 10.1093/nar/6.7.2601

Ruslan Kalendar and Alan H. Schulman. (2007). *Transposon-Based Tagging: IRAP, REMAP, and iPBS*. *1905*(page 114), 46–48. doi: 10.1007/978-1-62703-767-9

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., Ricciardi, F., Lotti, C., Ricciardi, L., *et al.* (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science*, *364*(6445), 1095–1098. doi: 10.1126/science.aav8197

Sanger, F., Nicklein, S., and Coulsan, A. R. (1997). DNA sequencing with chain-terminating inhibitors.

Obstetrics and Gynecology, *74*(12), 5463–5467. doi: 10.1097/00006250-199004001-00013

SanMiguel, P., Gaut, B. S., Tikhonov, A., Nakajima, Y., & Bennetzen, J. L. (1998). The paleontology of intergene retrotransposons of maize. *Nature Genetics*, *20*(1), 43–45. doi: 10.1038/1695

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., Liang, C., Zhang, J., Fulton, L., *et al.* (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science*, *326*(5956), 1112–1115. doi: 10.1126/science.1178534

Schrader, L., & Schmitz, J. (2019). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *28*(6), 1537–1549. doi: 10.1111/mec.14794

Shapiro, J. (1995). The discovery and significance of mobile genetic elements. *Mobile Genetic Elements - Frontiers in Molecular Biology*

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H., & Isobe, S. (2017). The genome sequence of sweet cherry (Prunus avium) for use in genomics-assisted breeding. *DNA Research*, *24*(5), 499–508. doi: 10.1093/dnares/dsx020

Shulaev, V., Korban, S. S., Sosinski, B., Abbott, A. G., Aldwinckle, H. S., Folta, K. M., Iezzoni, A., Main, D., Arús, P., … Veilleux, R. E. (2008). Multiple models for Rosaceae genomics. *Plant Physiology*, *147*(3), 985–1003. doi: 10.1104/pp.107.115618

Soriano, V. R. (2016). *Transposable element misregulation in Drosophila buzzatii – Drosophila koepferae interspecific hybrids* [Online], Doctoral Thesis, UAB, Department of genetics and microbiology [consulted in 21st of May pf 2021], available at: hdl.handle.net/10803/393906.

Su, T., Wilf, P., Huang, Y., Zhang, S., & Zhou, Z. (2015). Peaches Preceded Humans: Fossil Evidence from SW China. *Scientific Reports*, *5*, 1–7. doi: 10.1038/srep16794

Tavaud, M., Zanetto, A., David, J. L., Laigret, F., & Dirlewanger, E. (2004). Genetic relationships between diploid and allotetraploid cherry species (Prunus avium, Prunus x gondouinii and Prunus cerasus). *Heredity*, *93*(6), 631–638. doi: 10.1038/sj.hdy.6800589

Thomas, J., & Pritham, E. J. (2015). Helitrons, the Eukaryotic Rolling-circle Transposable Elements. *Mobile DNA III*, 891–924. doi: 10.1128/9781555819217.ch40

Tian, Y., Xing, C., Cao, Y., Wang, C., Guan, F., Li, R., & Meng, F. (2015). Evaluation of genetic diversity on Prunus mira Koehne by using ISSR and RAPD markers. *Biotechnology and Biotechnological Equipment*, *29*(6), 1053–1061. doi: 10.1080/13102818.2015.1064780

Tipu, H. N., & Shabbir, A. (2015). Evolution of DNA sequencing. *Journal of the College of Physicians and Surgeons Pakistan*, *25*(3), 210–215. doi 03.2015/JCPSP.210215

Vendramin, E., Pea, G., Dondini, L., Pacheco, I., Dettori, M. T., Gazza, L., Scalabrin, S., Strozzi, F., Tartarini, S., … Rossini, L. (2014). A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach. *PLoS ONE*, *9*(3). doi: 10.1371/journal.pone.0090574

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., Zhebentyayeva, T., Dettori, M. T., Grimwood, J., *et al.* (2013). The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nature Genetics*, *45*(5), 487–494. doi: 10.1038/ng.2586

Verde, I., Jenkins, J., Dondini, L., Micali, S., Pagliarani, G., Vendramin, E., Paris, R., Aramini, V., Gazza, L., *et al.* (2017). The Peach v2.0 release: High-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, *18*(1), 1–18. doi: 10.1186/s12864-017-3606-9

Vicient, C. M., & Casacuberta, J. M. (2020). Additional ORFs in Plant LTR-Retrotransposons. *Frontiers in Plant Science*, *11*(May), 1–5. doi: 10.3389/fpls.2020.00555

Wang, J., Liu, W., Zhu, D., Zhou, X., Hong, P., Zhao, H., Tan, Y., Chen, X., Zong, X., … Liu, Q. (2020). A de novo assembly of the sweet cherry (Prunus avium cv. Tieton) genome using linked-read sequencing technology. *PeerJ*, *2020*(6), 1–18. doi: 10.7717/peerj.9114

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., … Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. doi: 10.1038/nature01262

Wei, L., & Cao, X. (2016). The effect of transposable elements on phenotypic variation: insights from plants to humans. *Science China Life Sciences*, *59*(1), 24–37. doi: 10.1007/s11427-015-4993-2

Weiner, A. M. (2002). SINEs and LINEs: The art of biting the hand that feeds you. *Contemporary Music Review*, *21*(1), 71–79. doi: 10.1080/07494460216644

Wendel, J. F., Jackson, S. A., Meyers, B. C., & Wing, R. A. (2016). Evolution of plant genome architecture. *Genome Biology*, *17*(1), 1–14. doi: 10.1186/s13059-016-0908-1

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., *et al.* (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. doi: 10.1038/nrg2165

Wojciech Makałowski, V. G. P. M. (2019). *Transposable Elements - Classification, Identification, and Their Use As a Tool For Comparative Genomics* (Vol. 1910).

Xiang, Y., Huang, C. H., Hu, Y., Wen, J., Li, S., Yi, T., Chen, H., Xiang, J., & Ma, H. (2017). Evolution of Rosaceae Fruit Types Based on Nuclear Phylogeny in the Context of Geological Times and Genome Duplication. *Molecular Biology and Evolution*, *34*(2), 262–281. doi: 10.1093/molbev/msw242

Xiao, T., & Zhou, W. (2020). The third generation sequencing: The advanced approach to genetic diseases. *Translational Pediatrics*, *9*(2), 163–173. doi: 10.21037/TP.2020.03.06

Yu, Y., Fu, J., Xu, Y., Zhang, J., Ren, F., Zhao, H., Tian, S., Guo, W., Tu, X., … Xie, H. (2018). Genome re-sequencing reveals the evolutionary history of peach fruit edibility. *Nature Communications*, *9*(1), 1–13. doi: 10.1038/s41467-018-07744-3

Zadesenets, K. S., Ershov, N. I., & Rubtsov, N. B. (2017). Whole-genome sequencing of eukaryotes: From sequencing of DNA fragments to a genome assembly. *Russian Journal of Genetics*, *53*(6), 631–639. https://doi.org/10.1134/S102279541705012X

Zheng, Y., Crawford, G. W., & Chen, X. (2014). Archaeological evidence for peach (Prunus persica) cultivation and domestication in China. *PLoS ONE*, *9*(9), 1–9. doi: 10.1371/journal.pone.0106595

Annex A: Scripts used

Script 1

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --cpus-per-task=12
#SBATCH --mem-per-cpu=2G
#SBATCH --output=out1.txt
#SBATCH --job-name="EDTA_1"
#SBATCH --partition=all

module load conda
source activate EDTA


perl /home/rcastanera/bin/EDTA/EDTA.pl --genome $1 --sensitive 0 --anno
0 --threads 12
```

Script 2

```
Stats.sh -Xmx4g in=apricot.genome.fa.gz
```

Script 3

```
Seqtk comp apricot.genome.fa.gz | awk ‘{x+=$9}END{print x}’
```

Script 4

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --cpus-per-task=12
#SBATCH --mem-per-cpu=2G
#SBATCH --output=out1.txt
#SBATCH --partition=all

module load conda
source activate EDTA
RepeatMasker -s -nolow -norna -nois -pa 8 -e rmblast -gff -lib
TElib.fa genome.mod

LAI -genome $1 -intact $2 -all $3
```

Script 5

```
library(ggplot2)
library(ggpubr)
library(ggeasy)
setwd('/Users/danie/Desktop/Data/Data/Length/Todo')

# Copia
data_Texas <-
read.csv("Pdulcis_Copia_Texas_LTR_retrotransposon_length_modificado.tx
t", header = FALSE,sep = "\t")
data_Texas$Species <- "Prunus dulcis Texas"
data_Lauranne <-
read.csv("Pdulcis_Lauranne_Copia_LTR_retrotransposon_length_modificado
.txt", header = FALSE,sep = "\t")
data_Lauranne$Species <- "Prunus dulcis Lauranne"
data_Mira <-
read.csv("Pmira_Copia_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Mira$Species <- "Prunus mira"
data_Avium <-
read.csv("Pavium_Copia_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Avium$Species <- "Prunus avium"
data_Armeniaca <-
read.csv("Parmeniaca_Copia_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Armeniaca$Species <- "Prunus armeniaca"
data_Persica <-
read.csv("Ppersica_Copia_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Persica$Species <- "Prunus persica"
data_Salicina <-
read.csv("Psalicina_Copia_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Salicina$Species <- "Prunus salicina"
data_COPIA <-
rbind(data_Salicina,data_Persica,data_Armeniaca,data_Avium,data_Mira,d
ata_Texas,data_Lauranne)

# Helitrons
data_Texas_helitrons <- read.csv("Pdulcis_Texas_helitron_length.txt",
header = FALSE,sep = "\t")
data_Texas_helitrons$Species <- "Prunus dulcis Texas"
data_Lauranne_helitrons <-
read.csv("Pdulcis_Lauranne_helitron_length.txt", header = FALSE,sep =
"\t")
data_Lauranne_helitrons$Species <- "Prunus dulcis Lauranne"
data_Mira_helitrons <- read.csv("Pmira_helitron_length.txt", header =
FALSE,sep = "\t")
data_Mira_helitrons$Species <- "Prunus mira"
data_Avium_helitrons <- read.csv("Pavium_helitron_length.txt", header
= FALSE,sep = "\t")
data_Avium_helitrons$Species <- "Prunus avium"
```

```
data_Armeniaca_helitrons <- read.csv("Parmeniaca_helitron_length.txt",
header = FALSE,sep = "\t")
data_Armeniaca_helitrons$Species <- "Prunus armeniaca"
data_Persica_helitrons <- read.csv("Ppersica_helitron_length.txt",
header = FALSE,sep = "\t")
data_Persica_helitrons$Species <- "Prunus persica"
data_Salicina_helitrons <- read.csv("Psalicina_helitron_length.txt",
header = FALSE,sep = "\t")
data_Salicina_helitrons$Species <- "Prunus salicina"
data_HELITRONS <-
rbind(data_Salicina_helitrons,data_Persica_helitrons,data_Armeniaca_he
litrons,data_Avium_helitrons,data_Mira_helitrons,data_Lauranne_helitro
ns,data_Texas_helitrons)

# Gypsy
data_Texas_Gypsy <-
read.csv("Pdulcis_Texas_Gypsy_LTR_retrotransposon_length_modificado.tx
t", header = FALSE,sep = "\t")
data_Texas_Gypsy$Species <-  "Prunus dulcis Texas"
data_Lauranne_Gypsy <-
read.csv("Pdulcis_Lauranne_Gypsy_LTR_retrotransposon_length_modificado
.txt", header = FALSE,sep = "\t")
data_Lauranne_Gypsy$Species <- "Prunus dulcis Lauranne"
data_Armeniaca_Gypsy <-
read.csv("Parmeniaca_Gypsy_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Armeniaca_Gypsy$Species <- "Prunus armeniaca"
data_Avium_Gypsy <-
read.csv("Pavium_tieton_Gypsy_LTR_retrotransposon_length_modificado.tx
t", header = FALSE,sep = "\t")
data_Avium_Gypsy$Species <- "Prunus avium"
data_Mira_Gypsy <-
read.csv("Pmira_Gypsy_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Mira_Gypsy$Species <- "Prunus mira"
data_Persica_Gypsy <-
read.csv("Ppersica_Gypsy_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Persica_Gypsy$Species <- "Prunus persica"
data_Salicina_Gypsy <-
read.csv("Psalicina_Gypsy_LTR_retrotransposon_length_modificado.txt",
header = FALSE,sep = "\t")
data_Salicina_Gypsy$Species <- "Prunus salicina"
data_Gypsy <-
rbind(data_Salicina_Gypsy,data_Persica_Gypsy,data_Mira_Gypsy,data_Aviu
m_Gypsy,data_Armeniaca_Gypsy,data_Lauranne_Gypsy,data_Texas_Gypsy)

# MITEs
data_Texas_mite <-
read.csv("Pdulcis_Texas_TIR_transposons_MITEs_length_modificado.txt",
header = FALSE,sep = "\t")
data_Texas_mite$Species <- "Prunus dulcis Texas"
data_Lauranne_mite <-
read.csv("Pdulcis_Lauranne_TIR_transposons_MITEs_length_modificado.txt
", header = FALSE,sep = "\t")
data_Lauranne_mite$Species <- "Prunus dulcis Lauranne"
```

```
data_Mira_mite <-
read.csv("Pmira_TIR_transposons_MITEs_length_modificado.txt", header =
FALSE,sep = "\t")
data_Mira_mite$Species <- "Prunus mira"
data_Avium_mite <-
read.csv("Pavium_TIR_transposons_MITEs_length_modificado.txt", header
= FALSE,sep = "\t")
data_Avium_mite$Species <- "Prunus avium"
data_Armeniaca_mite <-
read.csv("Parmeniaca_TIR_transposons_MITEs_length_modificado.txt",
header = FALSE,sep = "\t")
data_Armeniaca_mite$Species <- "Prunus armeniaca"
data_Persica_mite <-
read.csv("Ppersica_TIR_transposons_MITEs_length_modificado.txt",
header = FALSE,sep = "\t")
data_Persica_mite$Species <- "Prunus persica"
data_Salicina_mite <-
read.csv("Psalicina_TIR_transposon_MITE_length_modificado.txt", header
= FALSE,sep = "\t")
data_Salicina_mite$Species <- "Prunus salicina"
data_MITE <-
rbind(data_Salicina_mite,data_Persica_mite,data_Armeniaca_mite,data_Av
ium_mite,data_Mira_mite,data_Lauranne_mite,data_Texas_mite)

# TIRs
data_Texas_no_mite <-
read.csv("Pdulcis_Texas_TIR_transposon_NO_MITE_length_modificado.txt",
header = FALSE,sep = "\t")
data_Texas_no_mite$Species <- "Prunus dulcis Texas"
data_Lauranne_no_mite <-
read.csv("Pdulcis_Lauranne_TIR_transposon_NO_MITE_length_modificado.tx
t", header = FALSE,sep = "\t")
data_Lauranne_no_mite$Species <- "Prunus dulcis Lauranne"
data_Mira_no_mite <-
read.csv("Pmira_TIR_transposon_NO_MITE_length_modificado.txt", header
= FALSE,sep = "\t")
data_Mira_no_mite$Species <- "Prunus mira"
data_Avium_no_mite <-
read.csv("Pavium_TIR_transposon_NO_MITE_length_modificado.txt", header
= FALSE,sep = "\t")
data_Avium_no_mite$Species <- "Prunus avium"
data_Armeniaca_no_mite <-
read.csv("Parmeniaca_TIR_transposon_NO_MITE_length_modificado.txt",
header = FALSE,sep = "\t")
data_Armeniaca_no_mite$Species <- "Prunus armeniaca"
data_Persica_no_mite <-
read.csv("Ppersica_TIR_transposon_NO_MITE_length_modificado.txt",
header = FALSE,sep = "\t")
data_Persica_no_mite$Species <- "Prunus persica"
data_Salicina_no_mite <-
read.csv("Psalicina_TIR_transposon_NO_MITE_length_modificado.txt",
header = FALSE,sep = "\t")
data_Salicina_no_mite$Species <- "Prunus salicina"
data_NO_MITE <-
rbind(data_Salicina_no_mite,data_Persica_no_mite,data_Armeniaca_no_mit
e,data_Avium_no_mite,data_Mira_no_mite,data_Lauranne_no_mite,data_Texa
s_no_mite)

# Grafica
DATA <-
rbind(data_COPIA,data_Gypsy,data_NO_MITE,data_HELITRONS,data_MITE)

# PLOT density
p2<-ggplot(DATA, aes(x=V5,color=Species)) +
  geom_density()+
  theme_bw()+
  facet_wrap(~V2, scales ="free")+
  labs(x="Length (bp)", y = "Density")
p2 + theme(legend.position="bottom")

# PLOT freqpoly
p1<-ggplot(DATA, aes(x=V5,colour=Species)) +
  geom_freqpoly()+
  theme_bw()+
  ggeasy::easy_center_title()+
  facet_wrap(~V2, scales ="free")+
  labs(x="Length (bp)", y ="N° of copies")
p1 + theme(legend.position="bottom")
```

Script 6

```
library(ggplot2)
library(ggpubr)
library(ggeasy)
setwd('/Users/danie/Desktop/Data/Data/Insertion time/ALL')

# COPIA files
data_Texas_Copia_insertiontime <-
read.csv("Pdulcis_Texas_Copia_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Texas_Copia_insertiontime$Species <- "Prunus dulcis Texas"
data_Lauranne_Copia_insertiontime <-
read.csv("Pdulcis_Lauranne_Copia_InsertionTime.txt", header =
FALSE,sep = "\t")
data_Lauranne_Copia_insertiontime$Species <- "Prunus dulcis Lauranne"
data_Mira_Copia_insertiontime <-
read.csv("Pmira_Copia_InsertionTime.txt", header = FALSE,sep = "\t")
data_Mira_Copia_insertiontime$Species <- "Prunus mira"
data_Avium_Copia_insertiontime <-
read.csv("Pavium_Copia_InsertionTime.txt", header = FALSE,sep = "\t")
data_Avium_Copia_insertiontime$Species <- "Prunus avium"
data_Armeniaca_Copia_insertiontime <-
read.csv("Parmeniaca_Copia_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Armeniaca_Copia_insertiontime$Species <- "Prunus armeniaca"
data_Persica_Copia_insertiontime <-
read.csv("Ppersica_Copia_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Persica_Copia_insertiontime$Species <- "Prunus persica"
```

```
data_Salicina_Copia_insertiontime <-
read.csv("PSalicina_Copia_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Salicina_Copia_insertiontime$Species <- "Prunus salicina"

# GYPSY flies
data_Texas_Gypsy_insertiontime <-
read.csv("Pdulcis_Texas_Gypsy_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Texas_Gypsy_insertiontime$Species <- "Prunus dulcis Texas"
data_Lauranne_Gypsy_insertiontime <-
read.csv("Pdulcis_Lauranne_Gypsy_InsertionTime.txt", header =
FALSE,sep = "\t")
data_Lauranne_Gypsy_insertiontime$Species <- "Prunus dulcis Lauranne"
data_Mira_Gypsy_insertiontime <-
read.csv("Pmira_Gypsy_InsertionTime.txt", header = FALSE,sep = "\t")
data_Mira_Gypsy_insertiontime$Species <- "Prunus mira"
data_Avium_Gypsy_insertiontime <-
read.csv("Pavium_Gypsy_InsertionTime.txt", header = FALSE,sep = "\t")
data_Avium_Gypsy_insertiontime$Species <- "Prunus avium"
data_Armeniaca_Gypsy_insertiontime <-
read.csv("Parmeniaca_Gypsy_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Armeniaca_Gypsy_insertiontime$Species <- "Prunus armeniaca"
data_Persica_Gypsy_insertiontime <-
read.csv("Ppersica_Gypsy_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Persica_Gypsy_insertiontime$Species <- "Prunus persica"
data_Salicina_Gypsy_insertiontime <-
read.csv("PSalicina_Gypsy_InsertionTime.txt", header = FALSE,sep =
"\t")
data_Salicina_Gypsy_insertiontime$Species <- "Prunus salicina"
data_insertiontime <-
rbind(data_Texas_Gypsy_insertiontime,data_Salicina_Gypsy_insertiontime
,data_Persica_Gypsy_insertiontime,data_Armeniaca_Gypsy_insertiontime,d
ata_Avium_Gypsy_insertiontime,data_Mira_Gypsy_insertiontime,data_Laura
nne_Gypsy_insertiontime,data_Salicina_Copia_insertiontime,data_Persica
_Copia_insertiontime,data_Armeniaca_Copia_insertiontime,data_Avium_Cop
ia_insertiontime,data_Mira_Copia_insertiontime,data_Lauranne_Copia_ins
ertiontime,data_Texas_Copia_insertiontime)

# PLOT density
p<-ggplot(data_insertiontime, aes(x=(V4/1000000)*10,colour=Species)) +
  geom_density()+
  theme_bw()+
  ggeasy::easy_center_title()+
  facet_wrap(~V3)+
  xlim(c(0,40))+
  labs(title="LTR",x="Insertion Time (MYA)", y ="Density")
p  + theme(legend.position="bottom")

# PLOT freqpoly
p1<-ggplot(data_insertiontime, aes(x=(V4/1000000)*10,colour=Species))
+
  geom_freqpoly()+
  theme_bw()+
  ggeasy::easy_center_title()+
  facet_wrap(~V3)+
  xlim(c(0,40))+
  labs(title="LTR",x="Insertion Time (MYA)", y ="N° of copies")
p1  + theme(legend.position="bottom")


ggarrange(p, p1, ncol = 1, nrow = 2, common.legend = TRUE)
```

Script 7

```
for f in *.fa;
do grep "LTR/Copia" $f | sed 's/>//g' > $f".copia.txt";
   grep "LTR/Gypsy" $f | sed 's/>//g' > $f".gypsy.txt";
   grep "MITE" $f | sed 's/>//g' > $f".MITE.txt";
   grep "DNA" $f | grep -v Helitron | sed 's/>//g' > $f".TIR.txt";
   grep "Helitron" $f | sed 's/>//g' > $f".Helitron.txt";
   seqkit grep -f $f".copia.txt" $f > $f".copia.fa";
   seqkit grep -f $f".gypsy.txt" $f > $f".gypsy.fa";
   seqkit grep -f $f".MITE.txt" $f > $f".MITE.fa";
   seqkit grep -f $f".TIR.txt" $f > $f".TIR.fa";
   seqkit grep -f $f".Helitron.txt" $f > $f".Helitron.fa";
   mkdir $f"_folder";
   mv *.fa $f"_folder";
   mv *.txt $f"_folder";
done

# Ejemplo para prunus salicina (Ps)

for f in *.fa;
do sed 's/>/Ps_/g' $f > $f".Ps.fa";
done
```

Script 8

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --cpus-per-task=12
#SBATCH --mem-per-cpu=2G
#SBATCH --output=out.txt
#SBATCH --job-name="clust"

module load conda
source activate bioperl


cd-hit-est -i $1 -o $1"_cons.fa" -c 0.8 -T 12 -d 0 -M 240000
```

Script 9

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --cpus-per-task=8
#SBATCH --mem-per-cpu=2G
#SBATCH --time=12:00:00
#SBATCH --job-name="blastx2"

module load blast

blastx -query /scratch/075-melo-
TEmovement/RAUL/PRUNUS/CLUSTERING/TIR/TIR_TE.fa_cons.fa  -db
/scratch/075-melo-
TEmovement/RAUL/Almond/HiconfTEannot/repbase_aa/testdb  -evalue 1e-10
-out blastx.out

 -outfmt 6 -max_target_seqs 1 -num_threads 8
```

Script 10

```
library(dplyr)
library(stringr)
library(ggplot2)
library(reshape2)
library(ggeasy)

setwd("/Users/danie/Desktop/unix/Pre-
clustering/RStudio_results/Copia")

clstr <- read.csv("/Users/danie/Desktop/unix/Pre-
clustering/COPIA_TE.fa_cons.fa.clstr", sep = "\t", row.names = NULL,
header = FALSE, stringsAsFactors = FALSE)

## Loop to replace col0 numbers by corresponding cluster
## "\\D" non-digit characters
## grepl returns TRUE if the character is found

clstr2 <- clstr
n = nrow(clstr)
x = 0
numbers_only <- function(x) !grepl("\\D", x)
for (row in c(1:n)) {
  if (numbers_only(clstr2[row,1]) == TRUE) {
    clstr2[row,1] <- x}
  else {NULL}
  x <- clstr2[row,1]
}

# Get rid of empty rows

clstr4 <- clstr2[-which(clstr2$V2 == ""), ]
```

```
# write output

write.table(clstr4,file = "COPIA_TE.fa_cons.fa.clstr_parsed.txt",
row.names = FALSE, col.names = FALSE,quote = FALSE, sep="\t")

# Run bash comands inside R:

system("cut -f1 COPIA_TE.fa_cons.fa.clstr_parsed.txt | tr ' ' '_' |
sort | uniq  > clusters.txt")
system("cat COPIA_TE.fa_cons.fa.clstr_parsed.txt | tr 'Cluster '
'Cluster_' > COPIA_TE.fa_cons.fa.clstr_parsed_2.txt")
system("grep '>Ps_' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1 |
sort | uniq -c | awk '{print $1,$2}' > Ps.txt")
system("grep '>Par_' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1
| sort | uniq -c | awk '{print $1,$2}' > Par.txt")
system("grep '>PdT' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1 |
sort | uniq -c | awk '{print $1,$2}' > PdT.txt")
system("grep '>Pav' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1 |
sort | uniq -c | awk '{print $1,$2}' > Pav.txt")
system("grep '>Pm' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1 |
sort | uniq -c | awk '{print $1,$2}' > Pm.txt")
system("grep '>Pp' COPIA_TE.fa_cons.fa.clstr_parsed_2.txt | cut -f1 |
sort | uniq -c | awk '{print $1,$2}' > Pp.txt")


# then join tables in R using "left_join" (dplyr)

clusters <- read.table("clusters.txt", header =FALSE)
ps <- read.table("Ps.txt", header =FALSE)
par <- read.table("Par.txt", header =FALSE)
PdT <- read.table("PdT.txt", header =FALSE)
Pav <- read.table("Pav.txt", header =FALSE)
Pm <- read.table("Pm.txt", header =FALSE)
Pp <- read.table("Pp.txt", header =FALSE)

names(clusters)  <- c("cluster")
names(ps) <- c("Ps","cluster")
names(par) <- c("Par","cluster")
names(PdT) <- c("PdT","cluster")
names(Pav) <- c("Pav","cluster")
names(Pm) <- c("Pm","cluster")
names(Pp) <- c("Pp","cluster")

# Join tables
test <- left_join(clusters,ps, by='cluster') %>%
  left_join(., par, by='cluster')  %>% left_join(., PdT, by='cluster')
%>% left_join(., Pav, by='cluster') %>% left_join(., Pm, by='cluster')
%>% left_join(., Pp, by='cluster')
row.names(test) <- test$cluster
test$cluster <- NULL

write.table(test, "Copia_clusters_heatmap.txt", row.names = TRUE, sep
= "\t", quote = FALSE)

# full table
```

```
full.table <- test
full.table[is.na(full.table)] <- 0

# binary table for build histogram

test[!is.na(test)] <- 1
test[is.na(test)] <- 0
suma <- as.data.frame(rowSums(test))
names(suma) <- "Species"

# Histogram with ggplot2

ggplot(suma, aes(x=Species, fill=Species)) +
  geom_histogram()+
  theme_bw()+
  scale_x_continuous(breaks=1:7)+
  ggtitle("LTR/Copia (n = 1034)")+
  ylab("Number of clusters")+
  ggeasy::easy_center_title()


# remove NAs and build heatmap

rownames(test) <- test$cluster
test$cluster <- NULL
test <- as.matrix(test)
test <- as.data.frame(test)
test2 <- as.matrix(test)
heatmap(test2)

For excluding singletons we added

data_zero <- suma[apply(suma, 1, function(row) all(row !=0 )), ]
suma2 <- as.data.frame(data_zero)

names(suma2) <- "Species"
```

Script 11

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --cpus-per-task=12
#SBATCH --mem-per-cpu=2G
#SBATCH --time=24:00:00
#SBATCH --output=RMout1.txt
#SBATCH --job-name="RepeatMasker"


# $1 = TE library
# $2 = Assembly

module load wublast
module load hmmer
module load trf
module load repeatmasker

RepeatMasker -s -nolow -norna -no_is -pa 12 -e wublast -gff -lib $1 $2
```

Script 12

```
#!/bin/bash -l

#SBATCH --nodes=1
#SBATCH --ntasks=8
#SBATCH --mem=24G
#SBATCH --job-name="parseRM"


# $1 = RM.out
# $2 = Genome.fa


perl /scratch/075-melo-TEmovement/RAUL/PRUNUS/parseRM/parseRM.pl -i $1
-p -f $2 -n -r  /scratch/075-melo-
TEmovement/RAUL/PRUNUS/CLUSTERING/Prunus_lib_FINAL.fa
```

Script 13

```
# Programs to use:

samtools, bedtools, R (ggplot2)

# Build genome index

samtools faidx apricot.genome.fa.mod
samtools faidx psalicina_v2.0.fasta.mod

# grep only chromosomes

grep Chr psalicina_v2.0.fasta.mod.fai | awk '{print $1"\t"$2}' >
psalicina_chromosomes.txt
grep LG apricot.genome.fa.mod.fai | awk '{print $1"\t"$2}' | sed
's/LG/Chr/g'  > parmeniaca_chromosomes.txt

# extract LTR coordinates

grep LTR psalicina_v2.0.fasta.mod.out.gff | awk '{print $1,$4,$5,$10}'
| sed 's/"//g' | tr ' ' '\t' > psalicina_LTR.bed
grep LTR apricot.genome.fa.mod.out.gff | awk '{print $1,$4,$5,$10}' |
sed 's/"//g' | tr ' ' '\t' | sed 's/LG/Chr/g' > parmeniaca_LTR.bed

# Divide the chromosomes of each genome in 25 windows

bedtools makewindows -g psalicina_chromosomes.txt -n 50 >
psalicina_windows.bed
```

```
bedtools makewindows -g parmeniaca_chromosomes.txt -n 50 >
parmeniaca_windows.bed


# Bedtools to calculate the number of LTR insertions per window in
each genome

bedtools intersect -a psalicina_windows.bed -b psalicina_LTR.bed -wo |
awk '{print $1"_"$2"_"$3}' | sort | uniq -c | awk '{print $2"\t"$1}' |
tr '_' '\t' | sort -k1,1 -k2,2n > psalicina_windows_coutnts.txt

bedtools intersect -a parmeniaca_windows.bed -b parmeniaca_LTR.bed -wo
| awk '{print $1"_"$2"_"$3}' | sort | uniq -c | awk '{print $2"\t"$1}'
| tr '_' '\t' | sort -k1,1 -k2,2n > parmeniaca_windows_coutnts.txt

# create a table to import in R (plot.csv)


# then use "plot.R" to plot the chromosome
```

Script 14

```
# BLAST sequences to prunus lib
# Decrease -word_size to increase sensitivity

for f in *.seq; do mv $f $f".fa";done

for f in *.fa; do echo $f; blastn -query $f -subject
/home/raul/Documents/Science/Docencia/TFG_Daniel_2021/results/clusteri
ng/prunuslibFINAL.fa -word_size 6 -evalue 0.00001 -outfmt 6 -
max_target_seqs 1; done

# BLAST sequences to P. persica genome

for f in *.fa; do echo $f;blastn -query $f -subject
/home/raul/Documents/Science/Almond/db/Prunus_persica_V2-pseudo.fa -
evalue 0.00001 -outfmt 6 -max_target_seqs 5; done


# BLAST vs Ppersica genes


for f in *.fa; do echo $f;blastn -query $f -subject
/home/raul/Documents/Science/Almond/db/Prunus_persica_v2.0.a1.primaryT
rs.cds.fa -evalue 0.00001 -outfmt 6 -max_target_seqs 5; done
```

Annex B: Supplementary graphics

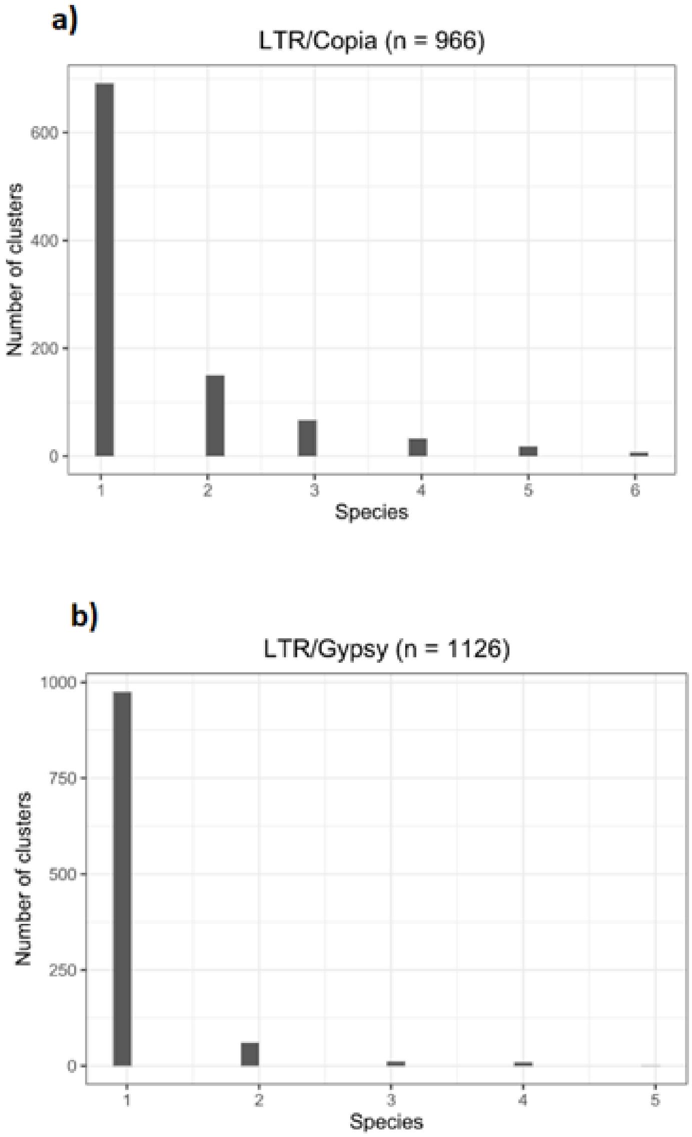


Figure S1. Number of species found in each cluster (with singletons) represented in histograms. a) Copia and b) Gypsy

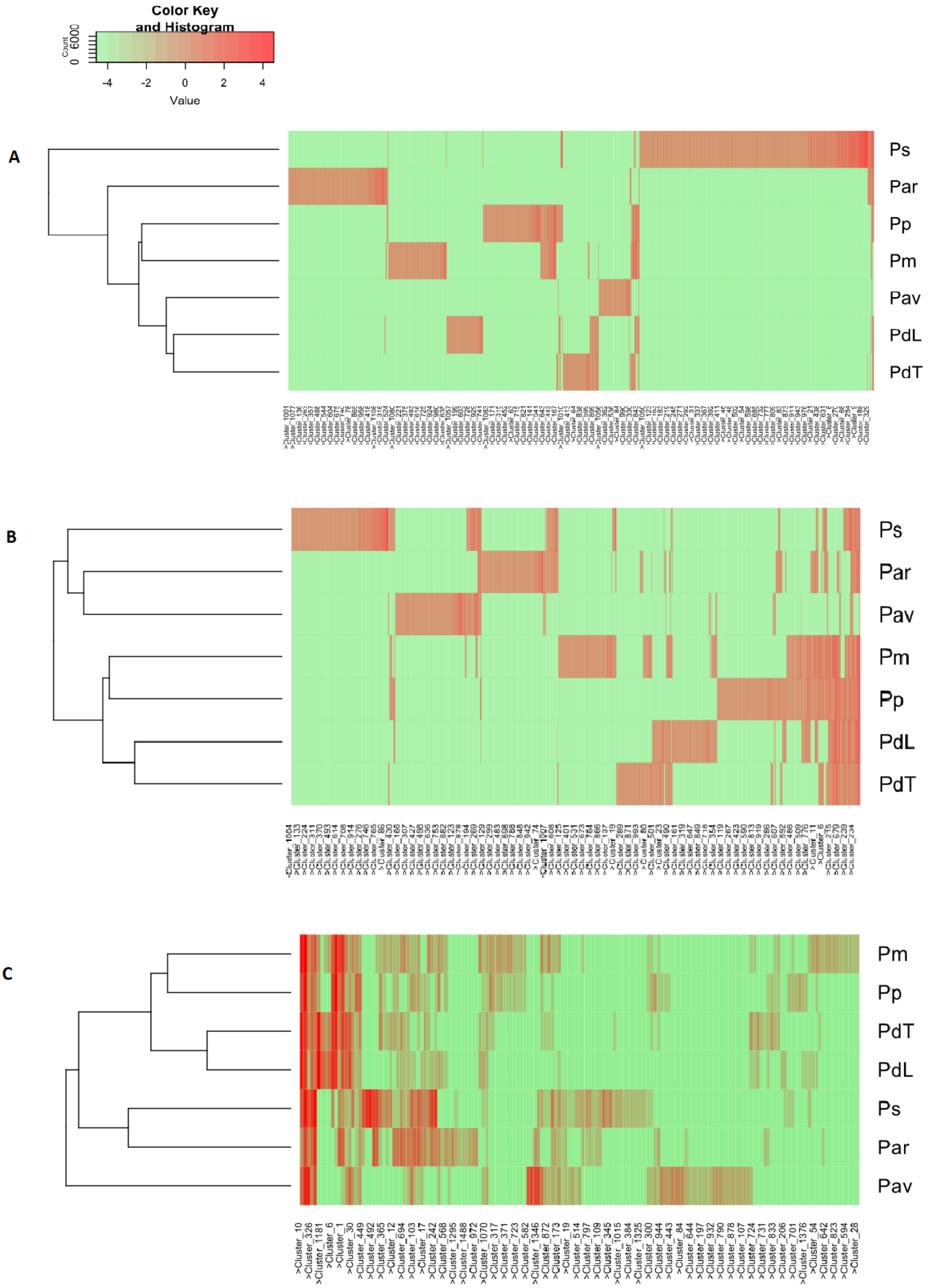


Figure S2. Phylogeny of *Prunus* genus based on TEs presence from abundance table. A) Copia B) Gypsy and C) MITEs.

Annex C: Sequencing dataset (Sanger sequencing)

Sanger sequencing of the 4 selected TEs amplifications which are within genomes for the verification of TEs insertions.

Sample 1; heterozygote, BG1 (possible insertion) BP1 (no insertion).

BG1 (possible insertion):

```
GGCCAAGTGTCTGCTGGAGTTTGGGCATGGGAGGGATGAAGCACCAGAGACGTGAGACAAAGAAG
TACAAAACACAAAGCCATGTGTTTAATTTTTCTTGTACCAAATGACAGAGCCGGTCTTGGGTAAT
CTGAGGCCCAGGGCAAAAATTTAAAGTGTGCCCAAAATAATGTAAACATTAAAATTATAATTTTT
TATCAAGTAATTAATTTTATTAAAATTATAGCATTCATTGTTTTTACAACATAAGCCTTCCCGTT
TGTTATTTTTTGATGATAATAATTTGAGGAAAAATGTCATCGTGTTCATCCATAGGGAATTTGCA
GTTAGTTTCTCCCACATGAAACTTTTCTAAAAGTAAATCTGCCATTTTTGCATCATAAACTTTCA
ATTCATTTTCATTTTCTGCTTATTCTTCCAAATGAATATGAACAATTCATTTGGTAAATCTTCAG
GTAAATTTTCATGTTTAAACCAAGTAATTCATTACAAAACTCACGTTCTCCCTGCAAAACTGATT
AAGACCCAAGAAAAATCATCTTTGTTGTTTATCTGGGTCTGCACCTGAAACCCATCTCAATTCTC
AATGAATAAGTTCTCAATTGGCGATCTTGAATCTGCTAGAATGCCTATATTAAGTGATTATTTGA
AAAAT
```

BP1 (no insertion):

```
GGCTAGGTGTCTGCTGGGAGTTTGGGCATGGGAGGGATGAAGCACCAGAGACGTGAGGCAAAGAA
GTACAAAACACAAAGCCATGTGTTTGATTTTTCTGTGTACCAAATGATTATGTAACTTATGAGGA
TTGTACTGTCAATGTGACATTCTTAGGCTATTTATAGGAAGAAGTACAATTAAATTACAGATTAT
TTATGGACCGTTAAGGGAATATAATATGAAAAAGTATTTTATCAAAAGGATACATGTCACCATTT
AAAGCAACCAAAGTCATCAAAAACTTCCCCCTTGGTCTTGCATCACAAAACTTTGCGGTCACCTT
TCAGCTATG
```

Sample 2; homozygote, no insertion:

```
GTGGGATGTGACGAGCTATGATCATCACAAGATGGAAGTGCACCTTATTTGTAAGGTTGGACTTG
CAACATGTTTAGTGAATAGAGAGTTGAATGTCAAAAATTGCAACATTTCACCCACACTATTTTGA
AATGTACATAAGGAAGCTTATCAGGCCTATCTGTACTCTATTCTCCTCGACTCATCAAATTTATG
AAGGTCATCCATCTAGTGCTGAGCTTTCAATGGCCGAAATTTATCAGCTTTAGTTTGGGGATTGA
AGGGCTGACAAAA
```

Sample 3; heterozygote, BG3 (possible insertion) BG3 (no insertion).

BG3 (possible insertion):

```
GGATTCCGGGGAGATCGCACTTTTAATCTTGTTGATGGTGGGGTTGCCGCTAACAATCCTGTAAG
TAGTTTTGGTAATGTTAACTTGTCAGATAATATCTAACATGTTATGTTTAATTAAATTTAGGTTT
TAGCCATAAAAAAACTTTCACTTTTGAAAAAAGCCAAAGCTTTTTCGTCCCATACAAACTACTTT
CTCGACTTTCCAATAAGTCTATGCAAAGTAAATACATTCCTTAAGATAAGCAACATAAACAAAGA
CAAGTTGAACCACATTGACTACCCGAAAAATGCGGCGCTGTTGTTGGGGCTATCCGTGGGCACCC
GTGATGATTATTCAGCATACCGATATAGAAACATAACATACCAAATTATCAGAGCGTAAAATGTC
TTTCATGTCCCATAAAATGCACTAAGGAATTAATTAATTTTTTCTACAATCATGTACAGACAATG
ATGGTCATAAGCCA
```

BP3 (no insertion):

```
GATTAAGGTAGGACCACTACACTTGATCTGTTGTTGATGGGGGTGCTGCCGCCAACCTCCTAGAA
GTCTGTATGCTTTCTTGCCTCCTATCTAACATGTCTGCCTCAATGCCTGTTATGCAATCTAATAT
GTCTGACTGTTTGACAAACCGACTAAGAATTAAGTCAGACAACGATTCTGTTCAGACCATGATGG
CCATAACCCACATAAGCGCAGACATATTGAAGCACAATTCGGAGCCGATGGATGCTATCATATTG
CACTTGCGTCATTGGGGCACAGGTGAGGCATCGGGCGCTATCGTTGGGCACAGGTGAGGGCACAG
GTGAT
```

Sample 4; homozygote, no insertion:

```
CCTTTCCTACTACCACACTTTGATGGACGAGGGATTCACTGTTCGATTTGGTACCAATTATGTGG
ATTACAATAATGGGCTGAAAAGACTACCAAAACTCTCAACTCGGTGGTTCAAAAGTTTCCTAGGA
AGTAATGAAGAAGCTTATTATTCATAATATAAACTGTTAAGTAGATCATTTCCCTCTTATCATAT
TTGTAATCTTGCATTAATGGATTTTATATATATTTTGCTATGTATAGTATAGATATAATAAGGAT
GACTACAATTATGTACTATATTTAAATAGCAACATTACGACACCATCGGA
```