TRANSPOSONS, A RICH SOURCE OF GENETIC VARIABILITY IN CROPS

Carlos de Tomás Gil

Universitat Autònoma de Barcelona

Facultat de Biociències

Departament de Biologia Animal, Biologia Vegetal i Ecologia

Programa de Doctorat en Biologia i Biotecnologia Vegetal

TESI DOCTORAL

TRANSPOSONS, A RICH SOURCE OF GENETIC VARIABILITY IN CROPS

Memòria presentada per Carlos de Tomás Gil per optar al títol de doctor per la Universitat Autònoma de Barcelona.

Treball realitzat en el Programa de Genòmica de Plantes i Animals del Centre de Recerca en Agrigenòmica (CRAG), Campus UAB Bellaterra.

Director de la tesi:	Tutor de la tesi:
Dr. Carlos Mª Vicient Sánchez	Dr. Josep Mª Casacuberta Suñer

Candidat a doctor:

Carlos de Tomás Gil

Cerdanyola del Vallès, 2023

A mi familia,

INDEX

ABSTRACT

Plant genomes contain a large number of dispersed repetitive sequences. Transposable Elements (TEs) and Endogenous Pararetroviruses (EPRVs) are part of this group of repetitive sequences. Transposable Elements are DNA sequences with the ability to move and change their position in the genome or generate copies of themselves in other genomic locations. Endogenous Pararetroviruses are sequences derived from viruses of the family *Caulimoviridae* and have been identified in monocotyledon and dicotyledon genomes, despite their replicative cycle does not necessarily include integration into the host genome.

In this thesis, transposons and Endogenous Pararetroviruses have been studied as sources of genetic variability in crops. Transposons have been specifically studied in the genus *Prunus*: in peach (*Prunus persica* 'Early Gold'), almond (*Prunus dulcis* 'Texas'), and an F1 hybrid (MB1.37) between both species. Meanwhile, the study of Endogenous Pararetroviruses included different plant groups.

Chapter 1 focuses on characterizing the transcription of Transposable Elements and genes in almond, peach, and the F1 hybrid, to determine if there is *genomic shock* resulting in transcriptional activation in the hybrid. Our results have shown that there is no *genomic shock*, as we did not observe significant changes in the hybrid's transcription compared to the parentals. Additionally, the study of genes with differential transcription between the two species has identified a gene with a polymorphic insertion that, due to its genomic location, is a strong candidate for the powdery mildew resistance gene *Vr3*.

Chapter 2 presents a study of transposons in almond, using a new version of its genome (Texas v.3.0). This new genome sequence is characterized by sequencing both phases or haplotypes, which is advantageous for studying a considerably heterozygous species like almond. This study included transposon annotation, improving the previous genome annotation, and an analysis of allele-specific expression in various almond organs, which identified four clusters of co-expressed alleles. Additionally, we performed an analysis of the impact of homozygous and heterozygous transposons near genes, which allowed for the detection of gene transcriptional differences, and a study of genetic variability in 40 almond cultivars.

Lastly, Chapter 3 was dedicated to the identification, classification, and characterization of the Reverse Transcriptase (RT) domain of recently inserted Endogenous Pararetroviruses in 278 plant genomes. This analysis identified 11,527 sequences, classified into 13 genera of the family *Caulimoviridae*. One of these genera is the newly proposed during this thesis genus *Wendovirus*, characterized by the presence of four open reading frames, two of which encode aspartic proteases.

RESUM

Els genomes de les plantes contenen un gran nombre de seqüències repetitives disperses. Els transposons i els Pararetrovirus Endògens (EPRVs) formen part d'aquest grup de seqüències repetitives. Els transposons són seqüències d'ADN amb la capacitat de moure's i canviar la seva posició en el genoma, o bé de generar còpies d'ells mateixos en altres posicions del genoma. Els Pararetrovirus Endògens són seqüències derivades de virus de la família *Caulimoviridae*, que s'han identificat en els genomes de monocotiledònies i dicotiledònies, tot i que el seu cicle replicatiu no inclou necessàriament la integració en el genoma de l'hoste.

En aquesta tesi s'han estudiat els Transposons i els Pararetrovirus Endògens com a font de variabilitat genètica en els cultius. Els transposons s'han estudiat específicament en el gènere *Prunus*: en presseguer (*Prunus persica* 'Early Gold'), ametller (*Prunus dulcis* 'Texas') i un híbrid F1 (MB1.37) entre les dues espècies. Mentre que l'estudi dels Pararetrovirus Endògens incloïa diferents grups de plantes.

El Capítol 1 es centra en la caracterització de la transcripció dels transposons i dels gens d’ametller, presseguer i de l'híbrid F1, per determinar si es produeix un *xoc genòmic* en l'híbrid, que resulti en una activació de la transcripció. Els nostres resultats han mostrat que no es produeix un *xoc genòmic*, ja que no hem observat grans canvis en la transcripció de l'híbrid respecte a la dels parentals. A més, l'estudi de gens amb transcripció diferencial entre les dues espècies ha permès identificar un gen amb una inserció polimòrfica, que per la seva localització en el genoma és un bon candidat a ser el gen de resistència a l'oïdi *Vr3*.

Al Capítol 2 es va realitzar un estudi dels transposons a l'ametller, utilitzant una nova versió del seu genoma (Texas v.3.0). Aquesta nova seqüència del genoma es caracteritza per la seqüenciació de les dues fases o haplotips, la qual cosa és un avantatge per a l'estudi d'una espècie considerablement heterozigota com és l'ametller. Aquest estudi va incloure l'anotació de transposons, millorant l'anotació de la versió anterior del genoma. A més, es va realitzar un estudi de l'expressió al·lèlica específica en diversos òrgans d’ametller i es van identificar 4 agrupacions d'al·lels co-expressats. Finalment, es va dur a terme un estudi de l'impacte dels transposons homozigots i heterozigots propers a gens, que va permetre la detecció de diferències en la transcripció dels gens, i un estudi de la variabilitat genètica en 40 cultivars d'ametller.

Per últim, el Capítol 3 es va basar en la identificació, classificació i caracterització del domini Transcriptasa Reversa (RT) dels Pararetrovirus Endògens recentment inserits a 278 genomes de plantes. Aquest anàlisi va permetre la identificació de 11,527 seqüències, que es van classificar en 13 gèneres de la família *Caulimoviridae*. Un d'aquests gèneres és el gènere *Wendovirus*, proposat durant aquesta tesi i que es caracteritza per la presència de quatre marcs de lectura oberts, amb dos que codifiquen per proteases aspàrtiques.

RESUMEN

Los genomas de plantas contienen un gran número de secuencias repetitivas dispersas. Los Elementos Transponibles (TEs) y los Pararetrovirus Endógenos (EPRVs) forman parte de este grupo de secuencias repetitivas. Los Elementos Transponibles son secuencias de ADN con la capacidad de moverse y cambiar su posición en el genoma, o bien de generar copias de ellos mismos en otras posiciones del genoma. Los Pararetrovirus Endógenos son secuencias derivadas de virus de la familia *Caulimoviridae*, que se han identificado en los genomas de monocotiledóneas y dicotiledóneas, a pesar de que su ciclo replicativo no incluye necesariamente la integración en el genoma del huésped.

En esta tesis se han estudiado los Transposones y los Pararetrovirus Endógenos como fuente de variabilidad genética en cultivos. Los transposones se han estudiado específicamente en el género *Prunus*: en melocotonero (*Prunus persica* 'Early Gold'), almendro (*Prunus dulcis* 'Texas') y un híbrido F1 (MB1.37) entre ambas especies. Mientras que el estudio de los Pararetrovirus Endógenos incluía diferentes grupos de plantas.

El Capítulo 1 se centra en la caracterización de la transcripción de los Elementos Transponibles y de los genes de almendro, melocotonero y del híbrido F1, para determinar si se produce un *shock genómico* en el híbrido que resulte en una activación de la transcripción. Nuestros resultados han mostrado que no se produce un *shock genómico*, ya que no observamos grandes cambios en la transcripción del híbrido respecto a la de los parentales. Además, el estudio de genes con transcripción diferencial entre las dos especies ha permitido identificar un gen con una inserción polimórfica, que por su localización en el genoma es un buen candidato para ser el gen de resistencia a oídio *Vr3*.

En el Capítulo 2 se realizó un estudio de los transposones en almendro, utilizando una nueva versión de su genoma (Texas v.3.0). Esta nueva secuencia del genoma se caracteriza por la secuenciación de las dos fases o haplotipos, lo cual supone una ventaja para el estudio de una especie considerablemente heterocigota como es almendro. Este estudio incluyó la anotación de transposones, mejorando la anotación de la versión anterior del genoma. Además, se realizó un estudio de la expresión alélica específica en diversos órganos de almendro y se identificaron 4 clusters de alelos co-expresados. Por último, se realizó un estudio del impacto de los transposones homocigotos y heterocigotos cerca de genes, que permitió la detección de diferencias en la transcripción de los genes, y un estudio de la variabilidad genética en 40 cultivares de almendro.

Finalmente, el Capítulo 3 se basó en la identificación, clasificación y caracterización del dominio Transcriptasa Reversa (RT) de Pararetrovirus Endógenos recientemente insertados en 278 genomas de plantas. Este análisis permitió la identificación de 11,527 secuencias, que se clasificaron en 13 géneros de la familia *Caulimoviridae*. Uno de estos géneros es el género *Wendovirus*, propuesto en esta tesis y que se caracteriza por la presencia de cuatro marcos de lectura abiertos, con dos que codifican para proteasas aspárticas.

LIST OF FIGURES

GENERAL INTRODUCTION AND OBJECTIVES

CHAPTER 1: TRANSCRIPTOME CHARACTERIZATION OF TRANSPOSABLE ELEMENTS AND GENES IN ALMOND, PEACH AND THEIR INTERSPECIFIC CROSS

CHAPTER 2: STUDY OF TRANSPOSABLE ELEMENTS OF A NEW PHASED VERSION OF ALMOND GENOME 'TEXAS'

CHAPTER 3. IDENTIFICATION, CLASSIFICATION, AND CHARACTERIZATION OF RECENTLY INSERTED ENDOGENOUS PARARETROVIRUSES, INCLUDING THE DISCOVERY OF THE NEW PUTATIVE GENUS *WENDOVIRUS*

LIST OF TABLES

GENERAL INTRODUCTION AND OBJECTIVES

CHAPTER 1: TRANSCRIPTOME CHARACTERIZATION OF TRANSPOSABLE ELEMENTS AND GENES IN ALMOND, PEACH AND THEIR INTERSPECIFIC CROSS

CHAPTER 2: STUDY OF TRANSPOSABLE ELEMENTS OF A NEW PHASED VERSION OF ALMOND GENOME ‘TEXAS’

CHAPTER 3. IDENTIFICATION, CLASSIFICATION, AND CHARACTERIZATION OF RECENTLY INSERTED ENDOGENOUS PARARETROVIRUSES, INCLUDING THE DISCOVERY OF THE NEW PUTATIVE GENUS *WENDOVIRUS*

ABBREVIATIONS

Activator locus: Ac

Additional open reading frames: aORF

Adenine-Thymine: AT

Allele-specific expression: ASE

Almond: A

Amino acid: aa

Annealing temperature: Tm

Aspartic proteinase / aspartyl proteinase: AP

Backcross: BC

Base Pairs: bp

Capsid protein: CP

Cauliflower mosaic virus: CaMV

Center for Research in Agricultural Genomics: CRAG

Centromeric Retrotransposons of Maize: CRM

Chromosome: Chr

Citrus tristeza virus: CTV

Cytoplasmic Male Sterility: CMS

Deoxyribonucleic acid: DNA

Dissociation locus: Ds

Double-stranded DNA: dsDNA

Early Gold: EG

Endogenous banana streak viruses: eBSV

Endogenous Non-Retroviral Elements: ENREs

Endogenous Pararetroviruses: EPRVs

Endogenous petunia vein-clearing virus: ePVCV

Endogenous Retroviruses: ERVs

Endogenous tobacco vein-clearing virus: eTVCV

Endogenous Viral Elements: EVEs

European Union: EU

Extensive de-novo TE Annotator: EDTA

False Discovery Rate: FDR

Genome Database for Rosaceae: GDR

Hectares: ha

High-throughput chromatin conformation capture: Hi-C

Hybrid: H

Institute of Agrifood Research: IRTA

Integrase: IN

Integrative Genomics Viewer: IGV

International Commite on Taxonomy of Viruses: ICTV

Kilobases pairs: Kbp

Long Interspersed Nuclear Elements: LINEs

Long Terminal Repeats: LTR

LTR Assembly Index: LAI

Maker-Assisted Selection: MAS

Maximum likelihood: ML

Megabase: Mb

Megabase pairs: Mbps

Million years ago: MYA

Million Years: MY

Miniature Inverted-Repeat Transposable Elements: MITEs

Multipurpose virion-associated protein: VAP

Near-isogenic lines: NILs

Nucleotide binding domain: NBD

Number: N

Open reading frames: ORFs

Operational taxonomic units: OTUs

Oxford Nanopore Technologies: ONT

Peach: P

Phase 0: P0

Phase 1: P1

Picograms: pg

Plant pararetroviruses: PRVs

Powdery Mildew resistance: PPM

Pregenomic 35S RNA: pgRNA

Presence-absence variations: PAVs

Principal Component Analysis: PCA

***Prunus dulcis*:** Pd or Prudul

***Prunus persica*:** Pp or Prupe

Quantitative trait loci: QTLs

Regularized log: rlog

Replication Protein A: RPA

Reverse transcriptase: RT

Rice tungro bacilliform virus: RTBV

Rice tungro spherical virus: RTSV

Ribonucleic acid: RNA

RNA Integrity Number: RIN

RNA interference: RNAi

RNA Polymerase II: RPII

RNAse H1 enzyme: RH1

Satellite DNA: satDNA

Short Interspersed Nuclear Elements: SINEs

Simple Sequence Repeats: SSRs

Simple Tandem Repeats: STRs

Single Nucleotide Polymorphism: SNP

Single-molecule real-time sequencing: SMRT

Small interfering RNAs: siRNAs

Structural variations: SV

Target Site Duplication: TSD

TE insertion Polymorphism: TIP

Terminal Inverted Repeats: TIR

Texas: TEX or T

Tons: t

Translation Elongation Factor 2: TEF2

Translation transactivator: TA

Transmembrane domain: TMD

Transposable Elements: TEs

United States of America: USA

Viral movement protein: VMP

Virus-like particles: VLPs

GENERAL INTRODUCTION

AND OBJECTIVES

PART A. Almond and Peach

1. Taxonomy

Peach [*Prunus persica* (L.) Batsch] and almond [*Prunus dulcis* (Mill.) D.A. Webb] are two fruit-bearing trees belonging to the genus *Prunus* in the family *Rosaceae*. Their systematic classification is as follows:

Kingdom: *Plantae*

Phylum: *Tracheophyta*

Division: *Magnoliophyta*

Class: *Magnoliopsida*

Order: *Rosales*

Family: *Rosaceae*

Subfamily: *Amygdaloideae*

Tribe: *Amygdaleae*

Genus: *Prunus*

Subgenus: *Amygdalus*

The family *Rosaceae* encompasses around 3,000 angiosperm eudicot species, many of which have significant economic importance (Potter *et al.*, 2007). For example, *Rosaceae* species such as apples (*Malus domestica* Borkh.), pears (*Pyrus communis* L.) and woodland strawberries (*Fragaria vesca* L.) are cultivated as food crops. Other *Rosaceae* species like roses (*Rosa* L.) and japanese apricots (*Prunus mume* Siebold & Zucc.) have ornamental value (Xiang *et al.*, 2016).

This family is divided into three subfamilies (Figure 1) in the most recent classification (Xiang *et al.*, 2016): *Rosoideae* (around 2,000 species of shrubs, herbs and fruit plants), *Amygdaloideae* (around 1,000 species that include the major fruit trees) and *Dryadoiledae* (less than 30 actinorhizal shrubs).

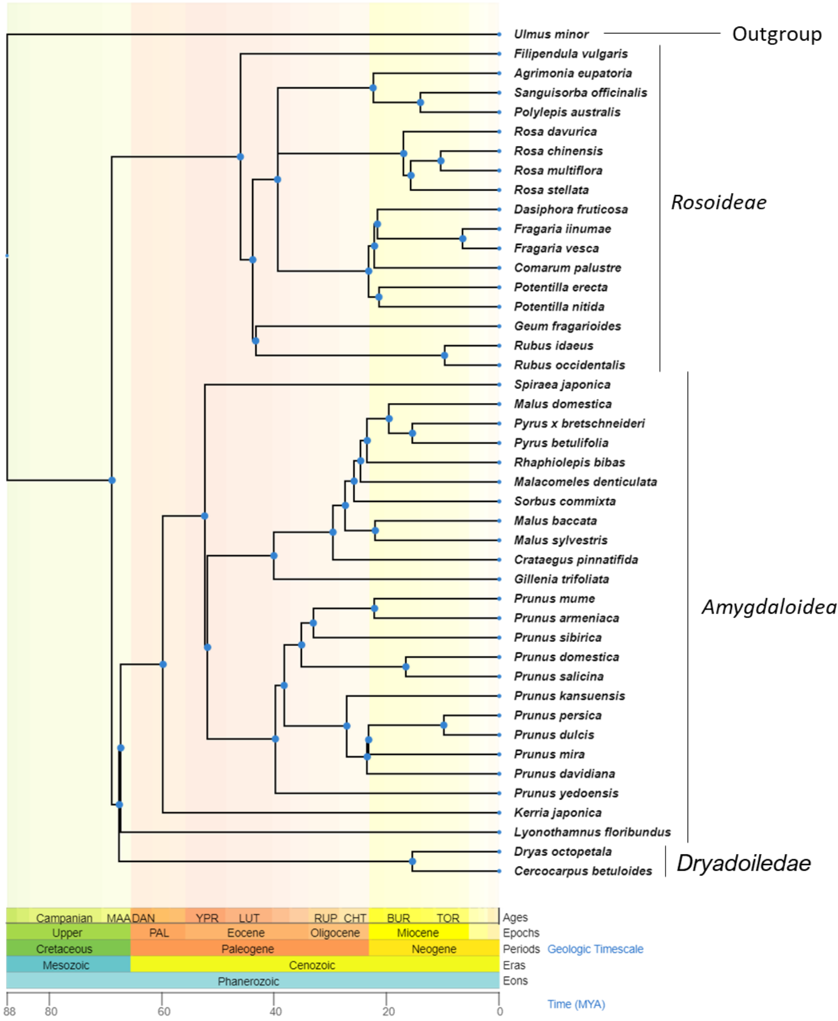


Figure 1. Phylogenetic tree of *Rosaceae* plants with a time tree at a scale based on phylogenetic analysis and fossil data. *Ulmus* minor (*Rosales* order) was used as an outgroup. The three subfamilies of *Rosaceae* are indicated. Tree developed using the timetree database (Kumar *et al.*, 2022).

The genus *Prunus* includes 200-400 species of shrubs, and deciduous and evergreen trees belonging to the subfamily *Amygdaloideae.* This genus has trees with important agronomic interest as peaches, almonds, apricots (*Prunus armeniaca* L.), sweet cherries [*Prunus avium* (L.) L.], japanese plums (*Prunus japonica* Thunb.) and european plums (*Prunus domestica* L.) (Chin *et al.*, 2014; Hodel *et al.*, 2021). Taxonomists have extensively studied the genus *Prunus* over the past centuries, resulting in various classifications within this genus. The most widely accepted infrageneric classification is based on the proposal by Rehder (1940), which has been revised by Chin *et al.* (2014). *Prunus* is divided into five subgenera: *Amygdalus* (peaches and almonds), *Cerasus* (cherries), *Prunus* (plums and apricots), *Laurocerasus* (evergreen laurel-cherries) and *Padus* (deciduous bird-cherries). Additionally, a sixth subgenus was included by Ingram

(1948) and recognized by Okie (2003) when *Lithocerasus* (flowering sand cherries) was added to the five subgenera of Rehder (1940).

The unity of the genus *Prunus* was verified at the molecular level by the Genome Database of Rosaceae (GDR) (GDR, https://rosaceae.org/; Jung *et al.*, 2019) and by the synteny conservation between the different *Prunus* genomes (Jung *et al.*, 2006; Vilanova *et al.*, 2008).

2. Origin and distribution

Velasco *et al.* (2016) estimates that peach and almond diverged approximately 8 million years ago (MYA). During the second half of the Miocene, there was an active period of uplift in the northeast Tibetan Plateau and Himalayan orogeny, accompanied by subsequent Asian climate change, which may have resulted in the isolation of a part of the population of the ancestor of peach and almond (Chin *et al.*, 2014). So, both species originated from different regions in Asia: peach is native to China, and almond comes from central Asia. Both species had independent processes of domestication that occurred 5,000 years ago (Zohary *et al.*, 2012).

Probably, peach spread from China to Persia (Iran) in the Silk Route before being introduced in Europe. Its scientific name, *Prunus persica,* is derived from its introduction and association with Persia. Subsequently, peach was introduced to America, with European and Chinese varieties serving as the ancestors of modern American cultivars. Almond, on the other hand, were moved from Asia to Europe and North Africa, via ancient trade routes (GDR, https://rosaceae.org/; Jung *et al.*, 2019).

Despite their distinct environmental requirements (warmer and more humid for peaches and colder and xerophytic for almonds) (Watkins, 1976), both can grow under diverse climatic conditions. Nowadays, the peach tree predominantly grows in temperate and subtropical regions (Scorza & Okie, 1991) and the almond tree is cultivated in hot climate regions (García-Tejero *et al.*, 2018), but both have a wide geographical distribution.

3. Botanical characteristics

The peach and almond are deciduous, robust and vigorous trees of medium height. There are differences between the varieties of each species but, in general, they are quite similar, with almond trees being taller and more vigorous than peach trees (Bassi & Monet, 2008; Socias & Gradziel, 2017).

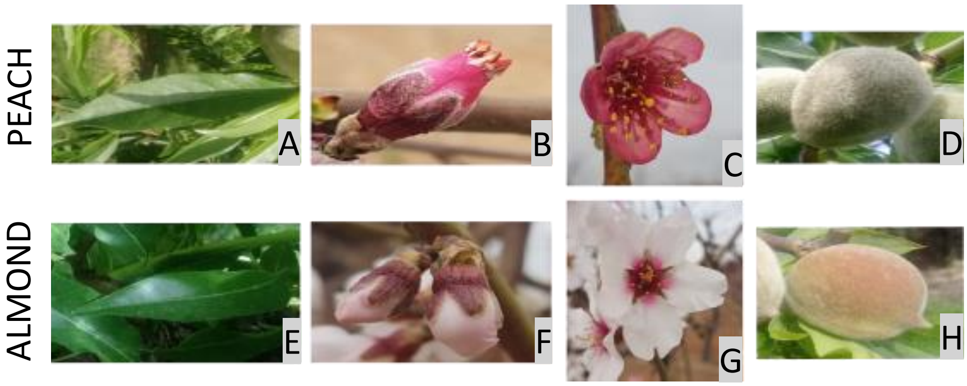


Figure 2. Leaves, flowers (pink and bloom stage) and immature fruits of the peach variety 'Early Gold' and the almond variety 'Texas'. ABCD, peach 'Early Gold'; EFGH, almond 'Texas'; AE, leaves; BF, immature flower (pink stage); CG, mature flower (bloom stage); DH, immature fruit.

The leaves are spaced out and arranged alternately on the branches. The leaves can be considered lanceolate, meaning they have a lance-shaped form and are wider near the petiole than at the apex (Figure 2A and E). They are also glabrous, without any type of hairiness, and slightly serrated on the margins (Bassi & Monet, 2008; Socias & Gradziel, 2017).

The flowers of peach and almond are hermaphrodite, but they have different reproductive strategies. While most peach varieties are auto-fertile, almond varieties are mostly self-incompatible. It means almonds require cross-pollination from pollinizer cultivars and insect pollinators (Gradziel, 2022).

The flowers of both appear before the leaves, in late winter or early spring. They typically have five sepals, five petals, a variable number of stamens, and a unique pistil. Peach flowers are generally pink in color (Figure 2B and C), but they can also have shades of red and white. Almond flowers, on the other hand, are generally white or pink (Figure 2F and G). The shape of the corolla is composed generally of five petals. It allows for the classification of the flowers into two categories: rosette and campanulate. Rosette (rose shaped) or showy flowers are characterized by long petals that hide the anthers and inhibit pollen shed before anthesis (Figure 2C). Campanulate (bell-shaped) or non-showy flowers are characterized by short petals that allow the anthers to be clearly observed among the petals (Figure 2G) (Bassi & Monet, 2008; Chen & Okie, 2015; Socias & Gradziel, 2017).

The fruits of peach and almond trees are very similar during their development (Figure 2D and H), and they begin to differentiate in the final stages. Peaches can have a globular or flattened shape and are classified as drupes. They are characterized by being simple, indehiscent, fleshy fruits with a pit inside, which corresponds to the endocarp (Bassi & Monet, 2008). The developed flesh or mesocarp of peaches can be white, yellow or red, and it can be adhered (clingstone) or separated from the pit. The skin or exocarp can be velvety (peach) or glabrous (nectarine) (Alvarado & Gonzalez, 1999; Bassi & Bonet, 2008). Almonds are also classified as drupes. They have a smaller and dry mesocarp and their endocarp contain their seed or kernel. These kernels represent the commercial part of the fruit (Social & Gradziel, 2017).

4. Fruit production

Peach and almond are highly produced and consumed and have an important economic impact.

The peach tree is the third most cultivated fruit tree in temperate climates, behind the apple and the pear trees. According to FAOSTAT (2021) data, the global production of peaches and nectarines amounts to 24,994,352 tons (t), covering a total harvested area of 1,504.682 hectares (ha). China emerges as the largest producer, accounting for over 60% of the annual production of peaches and nectarines. Following China, the European Union (EU) holds the second-largest share. Spain plays a significant role within the European Union (EU), producing 1,197,840 tons of peaches (accounting for 4.79% of the global total). Italy also contributes significantly with 3,98% of the global production (Figure 3).

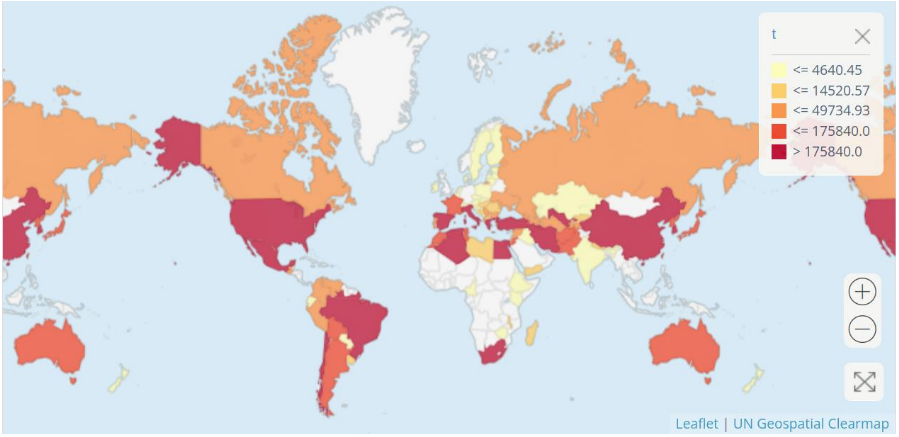


Figure 3. Production of peaches and nectarines around the world (FAOSTAT, 2021).

Almond is the most important and widely planted nut-producing fruit trees in the world (García-Tejero *et al.*, 2018). The FAOSTAT (2021) data shows that the worldwide production of almonds (with shell) amounts to 3,993,998 t, covering a total harvested area of 2,283,414 ha. The United States of America (USA) stands as the major producer, contributing 54.80% of the total production, followed by Spain, that produces 365,210 t (representing 9.14% of the global total) and Australia (7.15%) (Figure 4).

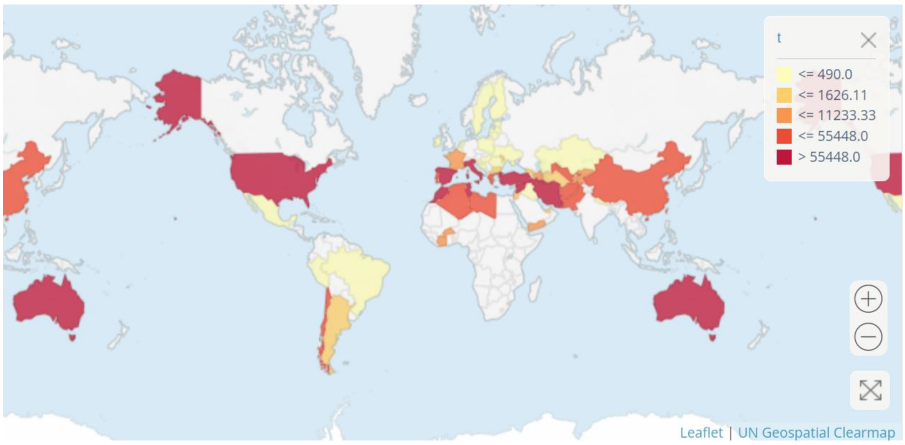


Figure 4. Production of almond in shell around the world (FAOSTAT, 2021).

5. Genetic and genomic characteristics

Peach and almond are diploid species with eight chromosomes (2n = 2x = 16) (Jelenkovic & Harrington, 1972; Baird *et al.*, 1994). Their genomes have been considered relatively small; for example, the 1C genome size of peach is 0.30 picograms (pg), while that of almond is 0.33 pg according to the Plant DNA Cvalues Database (https://cvalues.science.kew.org/; Pellicer & Leitch, 2020). Furthermore, these genomes have been fully sequenced (Verde *et al.*, 2017; Alioto *et al.*, 2020).

The genomic characteristics of the peach tree confer several advantages making it a model species within the *Rosaceae* family. Peach has undergone a limited number of duplications in its evolutionary history (Verde *et al.*, 2013). Moreover, unlike other *Prunus* species, peach lacks gametophytic self-incompatibility mechanisms, which allows the development of F2 hybrid populations through self-fertilization. These populations are highly valuable for genetic mapping projects. Additionally, peach trees have a short juvenile period of 2 to 3 years, which is considerably shorter compared to other woody species with juvenile periods ranging from 5 to 10 years, such as apples, pears and cherries (Arús *et al.*, 2012).

The almond and peach genomes present a high degree of similarity and synteny. As previously explained, the sequences of both genomes have been obtained and genes and TEs have been annotated. On average, their sequences display around 20 nucleotide substitutions per kbp, and they share basically the same genes and TEs. However, almond is seven times more variable than peach because most almond varieties are self-incompatible, which contributes to a higher level of genetic variability compared to peach (Velasco *et al.*, 2016). The TE content is similar in both species with only some few differences in the dynamics and recent activity (Alioto *et al.*, 2020).

Currently, in the GDR, there are available genomes of different *Prunus* species, such as four varieties of peach, including the varieties 'Lovell' (Verde *et al.*, 2017), '124 Pan' (Zhang *et al.*, 2021), 'Chinese Cling' (Cao *et al.*, 2021) and 'Zhongyoutao 14' (Lian *et al.*, 2021). Almond has genomes available for three varieties: 'Lauranne' (Sánchez-Pérez *et al.*, 2019), 'Texas' (Alioto *et al.*, 2020) and 'Nonpareil' (D'Amico-Willman *et al.*, 2022). Different sequencing technologies has been used for the different available genomes (Table 1) (GDR, https://rosaceae.org/; Jung *et al.*, 2019).

Table 1. Available genomes of peach and almond in the GDR (15/07/2023).

SPECIE	GENOME NAME	VARIETY	SEQUENCING TECHNOLOGY	GENOME SIZE (Mbp)
Prunus persica	Prunus persica 124 Pan Genome v1.0	124 Pan	Pacbio and Illumina HiSeq	206.09
	Prunus persica Zhongyoutao 14 Genome v1.0	Zhongyoutao 14	PacBio, Illumina and Hi-C	236.58
	Prunus persica Chinese Cling Genome v1.0	Chinese Cling	PacBio, Illumina and Hi-C	247.33
	Prunus persica Genome v2.0.a1[1]	Lovell	Sanger and Illumina	227.40
Prunus dulcis	Prunus dulcis Lauranne Genome v1.0	Lauranne	PacBio and Illumina	246.12
	Prunus dulcis Texas Genome v2.0	Texas	Illumina and Oxford Nanopore Technologies	227.60
	Prunus dulcis Nonpareil Genome v1.0	Nonpareil	PacBio, Illumina and optical mapping technologies	257.66

[1] This genome is the improved second version of the 'Lovell' genome called Prunus persica Genome v1.0 (Verde *et al.*, 2013).

6. Interspecific crosses

According to studies based on isoenzymes (Byrne, 1990) and microsatellites (Mnejja *et al.*, 2010), the peach tree is the *Prunus* species with the lowest genetic variability, in fact, it is considered to have low genetic diversity. The main reason for this low genetic variability is the previously mentioned high level of self-fertilization by autogamy in this species, largely due to the absence of gametophytic self-incompatibility mechanisms (Aranzana *et al.*, 2012).

Other factors contributing to this situation include the fact that most commercial varieties used in Europe and the United States are derived from a very limited genetic pool that was utilized by breeders in the United States over a century ago. This bottleneck situation has led to increased homozygosity and further limited the genetic variability of the species (Scorza *et al.*, 1985). However, oriental varieties show greater genetic variability than western varieties, mainly due to the high heterozygosity still present in native Chinese varieties, from where the peach tree was originated (Li *et al.*, 2013).

The peach tree is reproductively compatible with other *Prunus* species, such as *P. dulcis*, *P. davidiana*, *P. kansuensis*, *P. ferganensis*, *P. cerasifera*, *P. brigantiaca*, *P. americana*, and *P. spinosa* (Gradziel, 2003; Bouhadida *et al.*, 2007). This reproductive compatibility allows to increase genetic variability by the introgression of new traits. These interspecific crossings could address urgent and current issues in peach cultivation, such as disease resistance, the development of innovative traits for longer life or improved fruit quality, or adaptation to climate change.

Unfortunately, interspecific crossings using traditional hybridization techniques require long-term strategies, which are a challenge for most breeders who rarely produce offspring through these types of crossings (Arús *et al.*, 2015). Other methods, such as Maker-Assisted Selection (MAS) have been developed to save time and resources (Ru *et al.*, 2015). In MAS, the desired phenotype is selected based on the genotype given by a molecular marker, which can be detected much earlier than the manifestation of the interested phenotype (Eduardo *et al.*, 2015).

Interspecific crossings between the peach tree and the almond tree have been extensively studied. In cases where there are offspring resulting from these interspecific crossings, the offspring can be fertile or infertile. There are studies that indicate they can obtain fertile offspring with a probability of 50% (Jáuregui *et al.*, 2001). These interspecific hybridizations can occur through the pollination of peach flowers with almond pollen or through the pollination of almond flowers with peach pollen. In the last case, higher viability of the crossings has been observed.

In fact, post-zygotic reproductive isolation mechanisms have been discovered that prevent genetic flow between these two species. Recently, the presence of male sterility has been discovered in F2 population and in the first backcross (BC1) of hybrids between almond ('Texas') and peach ('Early Gold'). This male sterility is the result of a binary system where nuclear and cytoplasmic genomes interact, and the combination of one or more nuclear restorer genes can interfere with mitochondrial proteins and restore fertility (Schnable & Wise, 1998).

Donoso *et al.* (2015) concluded that the male fertility of F2 and BC1 hybrids is determined by two almond nuclear restorer genes (*Rf1* and *Rf2*), which are transmitted independently. The presence of the almond allele (which is the dominant allele) in at least one of the two nuclear restorer genes results in a fertile phenotype. The combination of these two genes interacts with a mitochondria-encoded protein and provides fertility. When the allele from the almond is absent in both genes, Cytoplasmic Male Sterility (CMS) occurs. This discovery must be taken into account when planning any crossing to introduce new genes into peach varieties. It is necessary to preserve the cytoplasm of the peach tree by using the peach tree as the female parent in at least one crossing or introduce one of the two *Rf* alleles from the almond to avoid unwished and unexpected sterilities (Donoso *et al.*, 2015).

PART B. Repetitive Elements

1. Concept and types

Eukaryotic genomes are characterized by the widespread presence of repetitive sequences (Lower *et al.*, 2019). In plants, these repeats can be highly abundant, constituting up to 90% of the genome in some species. They play crucial roles in various biological processes, such as chromosome movement and pairing, centromeric condensation, chromosome recombination, sister chromatid pairing, chromosome association with the mitotic spindle, chromosome arrangement, interaction of chromatin proteins, histone binding, determination of chromosome structure, karyotypic evolution, regulation of gene expression, and the genome's response to environmental stimuli and physiological changes (Mehrotra & Goyal, 2014).

These repetitive sequences can vary in length, ranging from short to long, and may be arranged either in tandem or interspersed throughout the genome (Lower *et al.*, 2019).

Tandem repeats can be further classified based on the number of nucleotides forming the repeated sequence and the frequency with which the sequence repeats. There are three main tandem repeats elements: microsatellites, minisatellites and satellite DNA (satDNA). Both microsatellites and minisatellites are moderate-repeats, in contrast with high-repeat satDNA, which replicates thousands or millions of times (Rao *et al.*, 2010).

Microsatellites also known as Simple Sequence Repeats (SSRs) or Simple Tandem Repeats (STRs) consist of short repeats of 2 to 5 bps with an array of 10 to 100 repeat units. Minisatellites, on the other hand, range from 6 to 100 bps, typically 15 bps, with an array size of 0.5 to 30 kilobases pairs (Kbp) (Mehrotra & Goyal, 2014). Lastly, satDNA consists of a variable Adenine-Thymine (AT) rich repeat unit that can form arrays up to 100 megabase pairs (Mbps) (Mehrotra & Goyal, 2014).

The interspersed repeats differ from tandem repeats in that they may or may not be proximal, because they do not necessarily need to be in consecutive order; instead, they are dispersed throughout the entire genome (Rao *et al.*, 2010). Transposable Elements (TEs) and Endogenous Viral Elements (EVEs) are part of this type of repetitive sequences and will be developed in more detail in this section.

2. Transposable Elements

Definition and discovery

The most abundant fraction of repetitive sequences is constituted by mobile genetic elements called Transposable Elements (TEs). TEs are DNA sequences with the ability to move and change their position in the genome or to create copies of themselves in other positions in the genome. This process, known as transposition, is an important source of genetic variability and can have a high impact on the size and structure of the genome (Wells & Feschotte, 2020).

Barbara McClintock was the pioneering scientist who elucidated the concept of Transposable Elements (TEs) during the 1940s and 1950s through experimentation in maize (*Zea mays*). Specifically, her focus was the relationship between the variegated color pattern of maize kernels and chromosome breakage. She detailed the chromosome breaks occurring in a specific locus of chromosome 9, which she named dissociation (Ds) locus. This locus demonstrated the ability to “jump” and change its position in the genome when an additional dominant locus, referred to as Activator (Ac), was present (McClintock, 1950).

This Ac/Ds system, consisting of these Ac and Ds elements, was proposed as the earliest described TEs. Later, in the 1980s, these Ac and Ds elements were cloned and characterized, revealing that Ac functions as an autonomous element, while Ds behaves as a non-autonomous element derivates of Ac elements (Fedoroff, 1989). Additionally, while Barbara McClintock’s groundbreaking work on TEs was indeed recognized by the scientific community, she was awarded the Nobel Prize in Physiology or Medicine in 1983 for her discovery, becoming the first woman to win this prize in solitary (https://www.nobelprize.org/).

Classification

The current classification of TEs that has been followed in this thesis is the one introduced by Finnegan (1989) and later refined by Wicker *et al.*, 2007. In this classification, TEs are divided into two main groups: class 1 (or retrotransposons) and class 2 (or DNA transposons). This classification is based on whether TEs require reverse transcription to transpose (Figure 5).

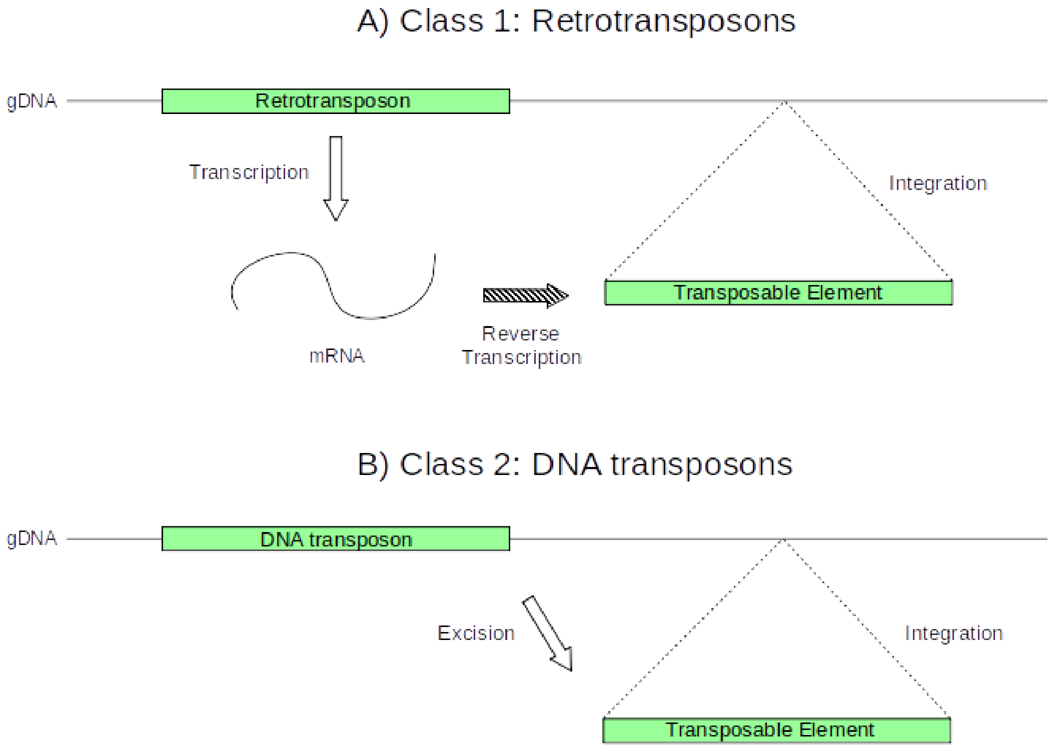


Figure 5. Mechanism of transposition of the two main classes of Transposable Elements: A) Class 1 or Retrotransposons and B) Class 2 or DNA transposons.

Retrotransposons are transposed using an RNA intermediate that is reverse transcribed and integrated into a different position of the genome, retaining their original copy. For this reason, it is known as a 'copy-and-paste' mechanism. In the case of DNA transposons, they do not require an RNA intermediate to transpose. Instead, they are excised from their original position in the genome and integrated into a new place. This mechanism is called 'cut-and-paste'. One exception of this second class is the Helitrons, which have a different type of transposition mechanism that was described as very similar to the "rolling circle" replication of the prokaryotic plasmids (Kapitonov & Jurka, 2007).

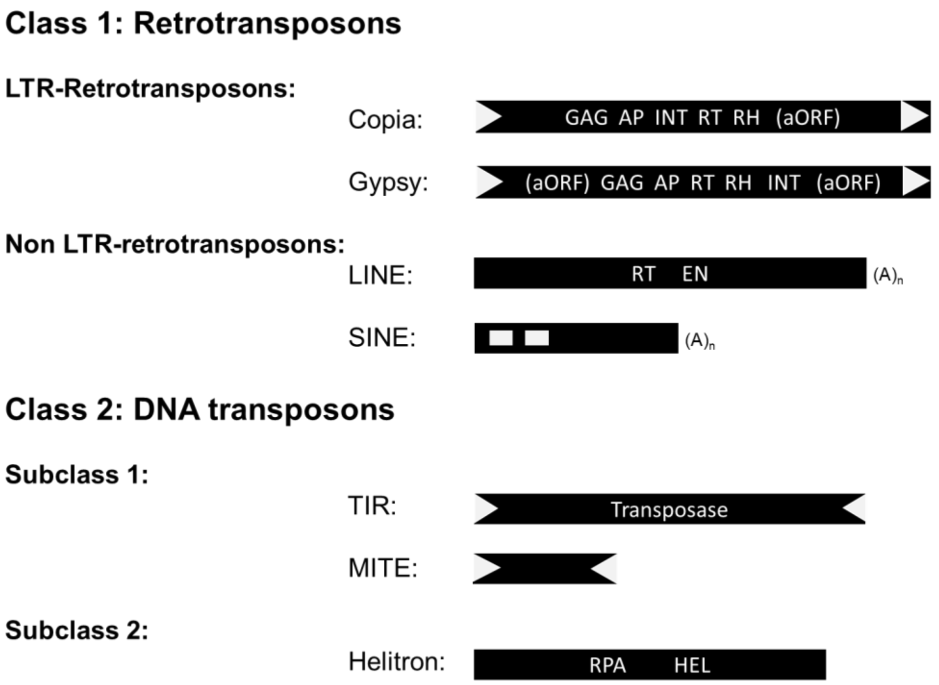


Figure 6. Main classes and orders of Transposable Elements based on Wicker *et al.*, 2007. Black boxes indicate the sequence of the TE. Grey boxes indicate diagnostic features in the coding region. The arrows in the same direction indicate the presence of Long Terminal Repeats, and in the opposite direction, the presence of Tandem Inverted Repeats. GAG indicates the protein Gag, AP Aspartic Protease, INT Integrase, RT Reverse Transcriptase, RH RNAse H, aORF additional Open Reading Frame, EN Endonuclease, (A)n indicates a repetitive sequence at the end of 3', RPA Replication protein A (only found in plants), and HEL Helicase.

Likewise, retrotransposons are further classified based on the presence of repeated sequences at each extreme end, known as Long Terminal Repeats (LTRs) (Figure 6). If they do not have LTRs, they are known as non-LTR retrotransposons, with Long Interspersed Nuclear Elements (LINEs) and Short Interspersed Nuclear Elements (SINEs) belonging to this group (Wicker, 2007).

LTR retrotransposons are closely related to the retroviruses and they are included in the same category of the Wicker Classification. They have a common organization with a 5' LTR followed by *gag* (encodes a structural protein for virus-like particles) and *pol* genes (encodes an aspartic protease, reverse transcriptase, RNAse H, and DDE Integrase), terminating with the 3' LTR (Wicker *et al.*, 2007). Despite the *gag* and *pol* genes being enough for LTR retrotransposons transposition, some elements contain additional open reading frames (aORF) in sense or antisense, such as those encoding for ENV-like proteins, found between the *pol* gene and the 3' LTR. The function of these aORFs is not yet clear, but there is evidence suggesting they may play a regulatory role in retrotransposition (Vicient & Casacuberta, 2017).

LTR retrotransposons are formed mainly by two superfamilies: Copia and Gypsy. Both of them differ in the arrangement of the encoded proteins, in particular in the order of RT and IN in the *pol* gene (Wicker, 2007).

Non-LTR retrotransposons are characterized by the presence of a repetitive sequence at the 3' end. As I mentioned before, there are two main groups: LINEs and SINEs, with LINEs being the most common in plants. LINEs have gag and pol coding regions, including an endonuclease and a RT, while SINEs do not encode any protein. They lack the ability to self-replicate and need the proteins from LINEs for transposition (Orozco-Arias *et al.*, 2019).

The second class of TEs, DNA transposons, is divided into two subclasses (Figure 6). The first subclass uses the 'cut-and-paste' transposition mechanism, while the second subclass replicates through a rolling circle mechanism, as mentioned before, and includes the Helitrons, that have Replication Protein A (RPA) only found in plants (Kapitonov & Jurka, 2007). The first subclass mainly comprises Terminal Inverted Repeats (TIR) TEs, which are characterized by these repeats flanking the internal coding sequence and the presence of an encoded transposase. TIR TEs can be classified based on their transposase motif and the size of their Target Site Duplication (TSD) sequence, which is generated when they integrate into the genome at a new position. There are nine known superfamilies, including Tc1/Mariner TEs, hAT, CACTA, and PIF/Harbinger (Wicker *et al.*, 2007).

Another interesting type of DNA transposons is the Miniature Inverted-Repeat Transposable Elements (MITEs). They are short non-autonomous TEs that result from deletion of class 2 transposons. They can contain TIRs and are mobilized by their encoded transposases. In contrast to other DNA transposons, they are highly abundant in the plant genomes and can have an impact on the trait variability of species (Feschotte *et al.*, 2002; Castanera *et al.*, 2021).

Distribution and predominance on plant genomes

TEs are not distributed homogeneously in the genomes. Generally, in angiosperms, TEs tend to accumulate in heterochromatic regions, including telomeres and centromeres, but they can also integrate into euchromatic regions. For example, MITEs are usually found close to genic regions (Casacuberta & Santiago, 2003), and Copia LTR retrotransposons are also found in proximity to genic regions. On the other hand, Gypsy LTR retrotransposons are more commonly found in pericentromeric regions (Alioto *et al.*, 2020). However, there are exceptions, and the integration and elimination of TEs depend on species-specific factors, such as selective pressure, environmental constraints, reproductive systems, population sizes, genome size, and organization (including ploidy level and interspecific crossing contribution), as well as the evolution of coexisting TE families (Pulido & Casacuberta, 2023).

TEs are highly abundant in some genomes, comprising up to 47.6% of the human genome (Hoyt *et al.*, 2022), and more than 80% in the maize genome (Schnable *et al.*, 2009). Specifically, LTR retrotransposons constitute a significant portion of plant genomes (Casacuberta & Santiago, 2003). Due to their 'copy-and-paste' mechanism, LTR retrotransposons are expected to increase their proportion in the genome and contribute to genome size expansion. On the other hand, MITEs are also highly abundant in number in plant genomes (Casacuberta & Santiago, 2003) but due to their small size they do not represent a high proportion of the genomes. The mechanisms used by MITEs to amply are still to be described.

Impact in crop genomes

As previously mentioned, TEs can cause a significant impact on genome structure, as exemplified by their role in genome size expansion. Furthermore, TEs represent a source of genetic variability, influencing processes such as domestication and plant breeding.

Recent studies have highlighted this essential role of TEs in domestication. For example, in rice (*Oryza sativa*), TE insertions have been observed to correlate with alterations in the gene expression of target genes crucial to rice domestication and breeding. Castanera *et al.* (2023) have shown that indica and japonica rice populations encompass distinct selected expression variants of genes associated with signal transduction. Also, Jiménez-Ruiz *et al.* (2020) proposed a significant and current TE activation in domesticated olives (*Olea europaea*), leading to insertions near genes potentially associated to agronomic traits such as oil production and seed development.

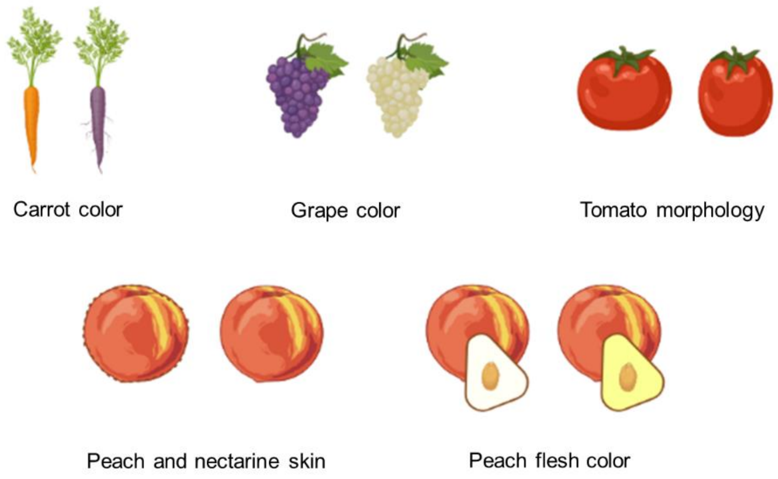


Figure 7. Examples of traits of crops determined by Transposable Elements insertions.

There are many instances where specific traits are influenced by TE insertions, as certain alleles have been selected through domestication (Figure 7). For example, the color of carrots (*Daucus carota*) is a trait that can be attributed to TE insertions in the promoter region of the *DcMYB7* gene, a gene that intervenes in the purple pigmentation of the roots. These insertions lead to the gene's transcriptional inactivation, resulting in nonpurple carrots that fail to accumulate anthocyanins in their roots (Xu *et al.*, 2019). The skin color of the grapes is also determined by a retrotransposon-induced mutation. In this case, this mutation is in *VvmybA1* gene that regulates the anthocyanins biosynthesis, and it causes the loss of pigmentation in the white cultivars of *Vitis vinifera* (Kobayashi *et al.*, 2004). In tomato (*Solanum lycopersicum*), the duplication of a *SUN* gene mediated by an LTR retrotransposon increases the *SUN* expression and produces an elongated fruit shape (Xiao *et al.*, 2008).

In the genus *Prunus*, we encounter the case of nectarines, which play a significant role in the peach industry. The presence of smooth skin in peaches (exhibiting the recessive nectarine phenotype) is caused by the insertion of a Copia LTR-retrotransposons in the third exon of the *PpeMYB25* gene, which regulates trichome formation in peaches (Vendramin *et al.*, 2014). Additionally, the flesh color of peaches (white or yellow) can be determined by three distinct mutational mechanisms: a microsatellite, a SNP and a Copia LTR retrotransposon. These three types of mutations can result in the loss of function of the *PpCCD4* gene, which encodes a carotenoid cleavage dioxygenase, leading to a recessive yellow phenotype. The LTR retrotransposon is integrated into the intron of this gene and may affect the transcription stability (Falchi *et al.*, 2013).

3. Endogenous Viral Elements

Definition and discovery

In the eukaryotes, viral integrations have been found dispersed throughout the entire genomes thanks to the sequencing of whole genomes. These sequences integrated, called Endogenous Viral Elements (EVEs), are sequences derived from viruses that have integrated into the nuclear chromosomes, allowing for vertical transmission and their fixation within the host (Feschotte & Gilbert, 2012).

EVEs have been found in animals, plants, and fungi. Among them, Endogenous Retroviruses (ERVs) that are derived from retroviruses, are very common in the genomes of jawed vertebrates, corresponding to up to 5-8% of the human genome. These elements comprise an internal region with three genes (*gag*, *pol* and *env*) along with two flanking noncoding LTRs, which are identical in the moment of integration. The integration of these sequences into the nuclear genome of the host cells is an integral part of the replicative cycle of retroviruses (Belshaw *et al.*, 2004).

In plants, no known viruses are recognized to integrate into the genome as an integral part of their reproductive cycle. Nevertheless, viral sequences have been identified within plant genomes, particularly Endogenous Pararetroviruses

(EPRVs), which are the most abundant (Diop *et al.*, 2018). Also, though still insufficiently studied, Endogenous Non-Retroviral Elements (ENREs) originating from dsRNA, ssRNA, and ssDNA viruses have been described. For example, ENRE from diverse plant viruses, such as *Partitiviridae*, *Betaflexiviridae* (Chiba *et al.*, 2011), *Chrysoviridae*, and *Geminiviridae* (Bejarano *et al.*, 1996; Chu *et al.*, 2014) have been identified in plants. The first discovery of an EVE occurred in 1996, and it was from a virus of the *Geminiviridae* family integrated in a unique locus of the tobacco (*Nicotiana tabacum*) genome (Bejarano *et al.*, 1996).

If EVEs are integrated into or near host genes, this will be generally detrimental, and they will be removed from host population by purifying selection. In the rare cases that the integration of an EVE is beneficial, it will be fixed in the host population by positive selection, in the same way that occurs with other types of genomic elements like transposons (Catlin and Josephs, 2022). However, most of the EVEs are neutral and will become degraded due to the accumulation of disruptive mutations, insertions or deletions. Due to the random nature of these mutations, it is possible to reconstruct the sequences of the infectious viruses based on the EVEs sequences, particularly for high copy number EVEs (Aiewsakun and Katzourakis, 2015).

Ancient EVEs are considered genomic “fossils” and are studied in Paleovirology as remnants of past viral infections (Etienne, 2017). Despite their age, recent research suggests they can cause advantage and might impact pathogenicity or resistance and could significantly influence viral latency/persistence dynamics (Takahashi *et al.*, 2019).

3.1. Plant pararetroviruses

Definition and classification

Plant pararetroviruses (PRVs) (family *Caulimoviridae*) are a family of double-stranded DNA (dsDNA) reverse-transcribing viruses infecting plants. They replicate by transcription in the nucleus followed by reverse transcription in the cytoplasm (Hohn & Rothnie, 2013). These characteristics allow them to be classified as group VII of the Baltimore Classification, alongside the family of animal viruses known as *Hepadnaviridae* (Koonin *et al.*, 2021).

Plant PRVs contain non-covalently closed circular genomes of 7.1 to 9.8 kbp and do not require integration within the host genome for their replication (Staginnus & Richert-Pöggeler, 2006). They comprise a single family of non-enveloped viruses known as *Caulimoviridae* that infect both monocots and dicots plants encompassing a broad range of plants. It is classified as follows:

Realm: *Riboviria*

Kingdom: *Pararnavirae*

Phylum: *Artverviricota*

Class: *Revtraviricetes*

Order: *Ortervirales*

Family: *Caulimoviridae*

This family consists of 11 genera that are called: *Badnavirus*, *Caulimovirus*, *Cavemovirus*, *Dioscovirus*, *Petuvirus*, *Rosadnavirus*, *Ruflodivirus*, *Solendovirus*, *Soymovirus*, *Tungrovirus* and *Vaccinivirus.* These genera are distinguished by their genome organization, including the number of open reading frames (ORFs) and the arrangement of protein domains within them. The genomes of these viruses have between one and eight ORFs, which encode various proteins such as a viral movement protein (VMP), a capsid protein (CP), a multipurpose virion-associated protein (VAP), an aspartic proteanase (AP), and a reverse transcriptase (RT) with tethered RNAse H1 enzyme (RH1) (Figure 8). Additionally, the morphology of their virus particles can be either isometric (in the case of *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Ruflodivirus*, *Solendovirus*, *Soymovirus*) or bacilliform (in the case of *Badnavirus* and *Tungrovirus*). There is currently no information for the morphology of particles of *Vaccinivirus* (Teycheney *et al.*, (2020); ICTV, https://ictv.global/).

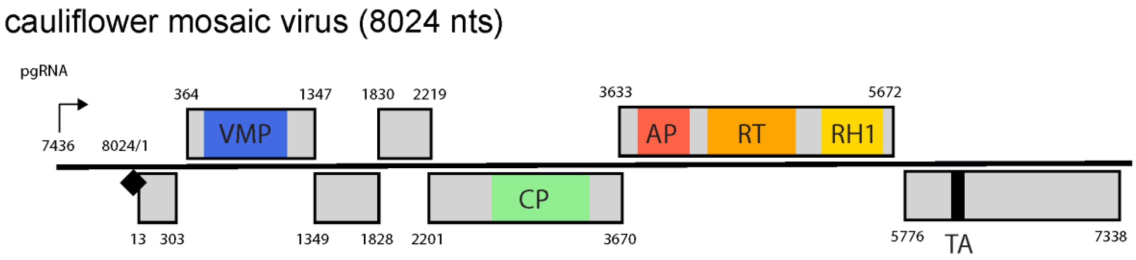


Figure 8. Genome organization of Cauliflower mosaic virus. The linearized map begins at the pregenomic 35S RNA (pgRNA) transcription start site (black arrow). The numbering begins from the first nucleotide of the Met-tRNA primer binding site (black diamond). Light grey boxes mark the distinct ORFs. Conserved protein domains are colored: blue is the viral movement protein (VMP), red is the aspartic proteanase (AP), orange is the reverse transcriptase (RT) and yellow is the RNase H1 (RH1). The conserved C-terminus of the coat protein (CP) is marked green. The conserved translation transactivator (TA) domain is shown in black. Extracted from: ICTV, https://ictv.global/.

Pathology and distribution

These viruses can induce variable symptoms. *Caulimovirus*, *Cavemovirus*, *Petuvirus*, *Rosadnavirus*, *Solendovirus*, and *Soymovirus* primarily cause mottling and mosaic patterns on leaves. *Tungrovirus* and *Badnavirus* can result in a range of symptoms, including chlorotic leaf streaks, leaf mottling, and growth deformations (Hull, 2007; ICTV, https://ictv.global/).

According to Hohn (2013), the transmission of plant PRVs occurs by three insect vectors: aphids for the isometric viruses (Martinière *et al.*, 2013), mealy bugs for *badnaviruses* (Geering *et al.*, 2005) and green leaf hoppers for Rice tungro bacilliform virus (RTBV) and Rice tungro spherical virus (RTSV) complex (Dahal *et al.*, 1997).

Model example

Cauliflower mosaic virus (CaMV) is one of the most studied and well-characterized Caulimoviruses. CaMV was the first plant virus discovered to contain DNA, and its DNA was the first viral genome to be completely sequenced (Franck *et al.*, 1980). The typical member of this species is the specific strain referred to as cauliflower mosaic virus-Cabb-S (V00141), belonging to the *Caulimovirus* genus (ICTV, https://ictv.global/). CaMV has been widely employed as a biological model in studies (Moreno *et al.*, 2005) and has found diverse applications in biotechnology, research and commercial sectors. For instance, the CaMV 35S promoter, known for its high transcriptional activity, is the most often used promoter in transgenic plants and plant biotechnology (Kiselev *et al.*, 2021).

3.2. Endogenous Pararetroviruses

Despite their non-integrative replication, in recent years, sequences remarkably similar to those of several members of the family *Caulimoviridae* have been identified, which are integrated into monocot and dicot genomes (Geering *et al.*, 2014; Diop *et al.*, 2018). These Endogenous Viral Elements (EVEs) have been called Endogenous Pararetroviruses (EPRVs) and have been included as a new category in some repetitive DNA sequence databases like Repbase (Bao *et al.*, 2015).

In contrast to retroviruses, whose integration is catalyzed by an essential retroviral enzyme called integrase (IN) (Passos *et al.*, 2021), plant paratroviruses do not have an active integration mechanism and they are supposed to integrate into the plant genome through illegitimate recombination during somatic DNA repair or meiotic recombination (Harper *et al.*, 2002; Geering *et al.*, 2014; Richert-Pöggeler *et al.*, 2021). EPRVs are usually located in pericentromeric regions of chromosomes and are close to retrotransposons (Staginnus & Richert-Pöggeler, 2006; Yu *et al.*, 2019).

Diop *et al.* (2018) provide evidence that vascular plants, including clubmosses, ferns, and gymnosperms, contain EPRVs. These sequences correspond to known episomal pararetroviruses such as *Rosadnavirus*, *Caulimovirus*, *Soymovirus*, *Petuvirus*, *Cavemovirus*, *Solendovirus*, *Tungrovirus*, and *Badnavirus*. However, there are sequences that correspond to genera whose episomal form has not yet been described, either because it has not been discovered or because it has become extinct now. Some of these exclusively endogenous genera without episomal described representatives are *Florendovirus*, *Gymnendovirus*, *Xendovirus*, *Yendovirus*, *Zendovirus*.

Florendovirus is a major component of flowering plant genomes and was described for the first time in 2014. This tentative genus represents more than 0.5% of the total genome content for some species, such as *Jatropha curcas*, *Amborella trichopoda*, *Citrus clementina* and *Vitis vinifera*. Notably, in *Ricinus communis*, *Florendovirus* sequences represent more than 1% of the genome content (Geering *et al.*, 2014; Diop *et al.*, 2018).

GENERAL OBJECTIVES

This PhD is included within the objectives of the research group "Structure and evolution of plant genomes" of the Center for Research in Agricultural Genomics (CRAG). Specifically, this thesis aims to study transposons and Endogenous Pararetrovirus as the source of genetic variability in crops. The study of the transposons is specifically in the genus *Prunus* and the study of the Endogenous Pararetroviruses encompasses the different plant groups.

In the different chapters, we will be developing more specific objectives, but the general objectives of each chapter are the following:

Chapter 1: Characterization of Transposable Elements and gene transcription in almond, peach and their interspecific cross to determine if a *genomic shock* is produced in the F1 hybrid.

Chapter 2: Study of Transposable Elements of a new phased version of almond genome cultivar 'Texas'.

Chapter 3: Identification and classification of Reverse Transcriptase domains of Endogenous Pararetroviruses in several plant genomes.

CHAPTER 1:

TRANSCRIPTOME CHARACTERIZATION OF TRANSPOSABLE ELEMENTS AND GENES IN ALMOND, PEACH AND THEIR INTERSPECIFIC CROSS

1.1. Introduction

Interspecific hybridization is a highly relevant process in plant evolution and breeding, as it can result in phenotypic changes and sexual isolation and be at the origin of new species (Mason & Batley, 2015). Hybridization results in the combination of diverged and novel genes, which can have strong consequences on the phenotype (Nieto Feliner *et al.*, 2020).

Hybridization can also induce epigenetic changes, including changes of DNA methylation and in the populations of small RNAs (Nieto Feliner *et al.*, 2020). The genomic changes frequently induced by merging two different genomes can be so wide that they have been frequently referred to as *genomic shock* (Comai *et al.*, 2003). It acts as a postzygotic barrier preventing gene flow between organisms. For example, important changes in gene expression have been observed in interspecific crosses of species of *Senecio* (Hegarty *et al.*, 2006), *Tragopogon* (Buggs *et al.*, 2009) or *Gossypium (Yoo et al., 2013)*. On the other hand, structural genome changes have also been reported and the activation of transposable elements (TEs). The transcriptional activation of TEs has been reported in interspecific crosses of, for example, *Spartina* (Parisod *et al.*, 2009), *Solanum* (Paz *et al.*, 2015*)* or *Nicotiana* (Mhiri *et al.*, 2019). Transpositional activation of different TEs was also reported in rice introgression lines derived from crosses with *Zizania latifolia* (Wang *et al.*, 2010) and increases in TE copy number has been reported in interspecific hybrids of *Helianthus* (Ungerer *et al.*, 2006) and *Aegilops* (Senerchia *et al.*, 2016). The activation and mobilization of TEs after hybridization can induce important genome changes through many mechanisms (Nieto Feliner *et al.*, 2020), in line with Barbara McClintock ideas of TEs as controller elements helping to reorganize the genome to overcome stress situations (McClintock, 1984).

TE activity is tightly controlled by epigenetic mechanisms and DNA methylation is the most obvious and frequent chromatin modification associated to TE silencing (Fultz & Slotkin, 2017). The mutation of different enzymes responsible of DNA methylation and chromatin modification results in a decrease of DNA methylation and induces the activation of plant TEs (Deniz *et al.*, 2019). Similarly, some biotic and abiotic stresses can result in a decrease of the DNA methylation and can activate the TE transcription and mobilization (Gutzat & Mittelsten

Scheid, 2012). The merging of two different genomes in allopolyploids can also induce changes in DNA methylation and gene expression, with important consequences on the phenotype (Ding & Chen, 2018). For example, changes in DNA methylation at the *CONSTANS-LIKE2* gene have an impact on the flowering time in domesticated cotton allotetraploid species (Song *et al.*, 2017), and changes in histone modifications in Arabidopsis hybrids and allopolyploids results in an increased biomass, vigor and in starch accumulation (Ni *et al.*, 2009). Massive changes in DNA methylation in TEs have been observed in newly formed hybrids, as, for example, in wheat allohexaploid (Yaakov & Kashkush, 2011). A decrease in 24-nt small RNAs, which are responsible for DNA methylation, has been shown in F1 allopolyploids between *Arabidopsis thaliana* and *Arabidopsis arenosa* (Ha *et al.*, 2009) and in intraspecific hybrids of *Arabidopsis thaliana* (Groszmann *et al.*, 2011). Therefore, the merging of two genomes can modify the epigenetic silencing of TEs and frequently results in TE activation that can induce further changes in the genome. However, examples where the merging of two different genomes does not result in changes in TE activity and genome structure have also been reported, for example, in crosses between *Arabidopsis thaliana* and *Arabidopsis lyrata* (Göbel *et al.*, 2018). TE proliferation has also been reported to be rare in natural *Helianthus* hybrids, despite their widespread transcriptional activity (Kawakami *et al.*, 2011). The reasons for this unpredictable outcome of the merging of two different genomes, and in particular, on TE activation, are not known but it could be related to the level of genome divergence between the two progenitors (Nieto Feliner *et al.*, 2020; Mhiri *et al.*, 2019).

Peach (*Prunus persica*) is one of the best-characterized species among the family *Rosaceae* and an important stone fruit crop (Arús *et al.*, 2012). Peach does not have a functional gametophytic self-incompatibility system and mainly behaves as self-pollinating, and consequently, it shows low levels of genetic diversity (Donoso *et al.*, 2015). For this reason, breeders have started to explore the possibility to use other *Prunus* species as an additional source of variability (Donoso *et al.*, 2016). Almond (*Prunus dulcis*) is one of the closest species to peach, both belonging to the subgenus *Amygdalus* (Hodel *et al.*, 2021). Peach and almond are diploid (2n = 2x = 16) and have relatively small genomes (about

300 Mbp) which has been sequenced (Verde *et al.*, 2017; Alioto *et al.*, 2020). The two genomes show a high level of similarity and are mainly syntenic (Alioto *et al.*, 2020). Most almond cultivars are self-incompatible and the almond genome is seven times more variable than peach (Velasco *et al.*, 2016). Peach and almond can be crossed to produce hybrids that are frequently fertile (Jáuregui *et al.*, 2001). In consequence, almond has been considered as an interesting source of novel alleles for peach breeding (Donoso *et al.*, 2016).

Plant pathogenic microorganisms such as fungi, bacteria or viruses attack different parts of the plants, causing a reduction in the growth rate, a loss of development and a reduction in the yield of the plant, and consequently, a lower production of the crops (Gafni *et al.*, 2015). Powdery mildew is a disease that affects a large number of plants from different families and is caused by different species of fungi of the order *Erysiphales*. These ascomycete fungi are obligate parasites (biotrophs) capable of inhabiting branches, leaves and fruits of more than 10,000 species of higher plants, including some members of the family *Rosaceae*. Powdery mildew appears in environments with elevated humidity and moderate temperatures. The mycelium germination usually occurs during the spring when optimal conditions prevail. During winter, the fungus remains in a state of latency within infected buds (Glawe, 2008).

The fungus *Podosphaera pannosa* (previously known as *Sphaerotheca pannosa*) var. *persicae* is the causal agent of powdery mildew in peach trees. The characteristic symptom of this infection is the appearance of circular whitish spots or white mold on leaves, young branches, and fruits. Consequently, the disease can induce premature leaf senescence and deformation of the fruit, along with a retardation in crop development (Pascal *et al.*, 2010).

Powdery mildew is one of the prevalent pathogens of peach crops in Europe. The majority of peach cultivars are susceptible to powdery mildew, necessitating fungicide applications from the pre-flowering until post-harvest stages (Pascal *et al.*, 2010). An alternative to the use of fungicides is the cultivation of crops varieties resistant to powdery mildew. The major genes associated with Peach Powdery Mildew resistance (PPM) are *Vr1*, *Vr2* and *Vr3*. *Vr1* and *Vr2* genes are located in linkage group G8 of the 'Malo Konare' peach cultivar and the 'Pamirskij 5' peach rootstock, respectively (Lambert, 2018; Pascal *et al.*, 2017). *Vr3* gene is located in the linkage group G2 of the ‘Texas’ almond cultivar (Donoso *et al.*, 2016). Furthermore, quantitative trait loci (QTLs) associated with PPM tolerance have been described in peach (Pacheco *et al.*, 2009).

The presence of monogenic resistance to powdery mildew from *Vr3* gene was described in the F2 and BC1 hybrid populations between 'Texas' and 'Early Gold' (Donoso *et al.*, 2016). Next, Marimon *et al.* (2020) published a fine mapping of this gene, in which I participated during my bachelor’s degree final project (de Tomás, 2016). *Vr3* gene was located between the markers Indel16912 and SNP_17184692. Their positions correspond to a 272 Kb region spanning physical positions 16,912,811 and 17,184,692, encompassing 27 annotated genes in the ‘Lovell’ peach genome (Verde *et al.*, 2017). In the genome of ‘Texas’ (Alioto *et al.*, 2020), *Vr3* is located between the positions 12,907,187 and 13,129,481 of the chromosome 2. This region measures 222 Kb and has 23 annotated genes.

Marimon *et al.* (2020) analyzed the polymorphisms (SNPs and InDels) of the resequencing data of both parents and they employed near-isogenic lines (NILs) for expression analysis of the candidate genes in symptomatic and asymptomatic leaves. Among the differentially expressed genes between resistant and susceptible individuals, the Disease Resistance Protein RGA2 (*Prupe2G111700*) and an Eceriferum 1 protein associated in epicuticular wax biosynthesis (*Prupe2G112800*) were annotated. Only *Prupe2G111700* gene had a variant predicted to disrupt the encoded protein.

In this chapter, we will investigate to what extend the crosses of peach and almond result in the activation of TEs that could lead to a *genomic shock*. Our focus will be the study and comparison of the transcription activity of transposable elements and genes in peach and almond, as well as in their F1 hybrid. This comparison was performed for three organs: leaves, flowers and fruits. Furthermore, we will deepen in the impact of the polymorphic insertions between the two parental lines on their gene expression. Specifically, we aim to characterize a polymorphic gene between peach and almond located within the *Vr3* region.

The transcription results on leaves presented in this chapter, jointly with the methylation analysis performed by Dr. Amélie Bardil, was published in a scientific paper. The paper was titled: Absence of major epigenetic and transcriptomic changes accompanying an interspecific cross between peach and almond. It was authored by de Tomás, Bardil, Castanera, Vicient & Casacuberta (2022) and published in the journal Horticulture Research (see Annexes).

1.2. Objectives

The objectives of this study are as follows:

- Characterize the transposable elements and gene transcription on different organs of almond, peach and their interspecific cross to determine if a *genomic shock* is produced in the F1 hybrid.

- Analyze the impact of the polymorphic transposable elements between peach and almond on their gene transcription.

- Study a polymorphic gene between peach and almond located within the *Vr3* region.

1.3. Material and methods

Plant material and growth conditions

Leaves, flowers, and fruits from the *Prunus dulcis* ‘Texas’, *Prunus persica* ‘Early Gold’ and one interspecific F1 hybrid ‘MB 1.37’ tree were collected from the Experimental Station of Lleida located in Gimenells (Catalonia, Spain), kindly provided by the Institute of Agrifood Research (IRTA). The ‘Texas’ almond tree was cultivated in the field and regularly watered at the same time of the day.

Fully expanded leaves were collected at the end of September, flowers in the pink stage were collected on February and immature fruits with approximately 2cm in diameter, were collected during the initial week of May. For each type of sample, a composite pool of 7 leaves, 10 flowers and 4 fruits was generated. The samples of each organ were harvested from three replicates per genotype of separate branches of the same tree.

In order to maintain sample integrity during transportation, dry ice and nitrogen liquid were employed. Finally, the sampled were stored in a -80º freezer.

RNA and DNA isolation

0.15 grams of each sample were ground using liquid nitrogen and a mortar, until it was transformed into a fine powder. High molecular weight genomic DNA was isolated using a sorbitol pre-wash (Inglis *et al.*, 2019) followed by an adapted CTAB method (Doyle & Doyle, 1990). Total RNA was extracted using the Maxwell RSC Plant RNA Kit and the Maxwell RSC instrument (Promega Corporation, Madison, WI, USA). Complete DNA removal was obtained using the DNA-free DNA Removal Kit (Invitrogen™, Carlsbad, CA, USA).

To assess the quality and purity of the DNA and RNA samples, the density ratios (260/280 and 260/230) were evaluated using NANODROP ND-1000 spectrophotometer (Thermo Fischer Scientific). The ranges of these density ratios are approximately 1.8 and 2, indicating desirable levels of DNA and RNA purity. The DNA integrity was tested using a 1% (w/v) agarose gel. The RNA Integrity Number (RIN) was calculated using Agilent 2100 Bioanalyzer. The determined RIN value for each RNA sample should be higher than 7.

Genome, gene and TE reference datasets

In this Chapter, we have used the genomes “Prunus persica Genome v.2.0.a1” (Verde *et al.*, 2017) and “Prunus dulcis Texas Genome v.2.0” (Alioto *et al.*, 2020). Gene annotations for each genome were obtained from Genome Database for *Rosaceae* (GDR; https://www.rosaceae.org/). TE library described in Alioto *et al.* (2020) was previously curated by Dr. Raúl Castanera (post-doc of CRAG) to retain only high-confidence TE annotations based on the presence of structural features, coding domains or homology to known TEs. It resulted in a more stringent annotation of TEs of almond (Supplementary Table S5 of de Tomás *et al.*, 2022) and peach (Supplementary Table S6 of de Tomás *et al.*, 2022), that we have used during this Chapter (Table 1).

Table 1. Annotation statistics of TEs in almond and peach genomes, including the major classes of TEs: LTR retrotransposons, LINEs, TIRs and MITEs.

CLASS	PEACH (COPIES)	ALMOND (COPIES)	PEACH (bp)	ALMOND (bp)	PEACH (% GENOME)	ALMOND (% GENOME)
LTR retrotransposon	18.807	18.751	40.018.216	35.804.381	17,73	16,01
LINE	1.179	1.282	1.709.736	1.970.960	0,76	0,88
TIR	5.313	4.972	20.668.164	15.566.752	9,16	6,96
MITE	8.754	10.466	3.046.983	3.462.657	1,35	1,55
Total	34.358	35.801	67.495.012	58.236.314	29,91	26,04

RNA sequencing and analysis

The RNA-seq libraries were obtained from 2–4 µg of total RNA. For each parental and hybrid genotype, three biological replicates were collected. The RNA-seq libraries were produced using the Truseq stranded mRNA protocol and were sequenced on Illumina platform NextSeq 500 (2x150 bp, Paired-end) (Table 2).

The RNAseq reads were filtered to eliminate adapters and low-quality and short sequences, using BBDuk (Bushnell, 2014) with the next parameters: ktrim=r, k=23, hdist=1, tpe ftr=139 and trimq=10 (Table 2). Their quality was checked using FastQC (Andrews, 2010).

Table 2. Number of raw and filtered RNA-seq reads for each genotype and organ.

	REPLICATE	LEAF		FLOWER		FRUIT	
		Raw	**Filtered**	**Raw**	**Filtered**	**Raw**	**Filtered**
ALMOND	A1	23,323,658	22,010,056	106,091,346	105,483,610	83,459,172	83,007,998
	A2	56,824,886	53,831,278	105,497,490	105,071,222	87,014,356	86,236,020
	A3	25,343,256	24,740,970	93,630,526	93,081,020	93,591,264	93,053,680
HYBRID	H1	46,460,290	44,868,054	105,167,386	104,491,142	98,600,878	97,474,008
	H2	43,971,892	42,782,980	105,822,882	105,169,990	90,435,172	89,625,494
	H3	38,049,008	35,688,258	88,602,688	87,826,878	87,013,114	86,635,092
PEACH	P1	44,440,218	43,145,450	88,888,540	88,154,340	90,916,794	89,616,196
	P2	43,982,836	43,366,238	91,325,320	90,869,576	105,986,924	105,009,004
	P3	43,729,274	43,015,870	91,863,878	91,271,604	96,953,916	95,767,382
AVERAGE		40,680,590	39,272,128	97,432,228	96,824,487	92,663,510	91,824,986

We performed two different RNA-seq analyses. The first analysis included only leaf samples from the three genotypes (peach, almond and their hybrid). The second analysis included samples of the three organs from the three genotypes (peach, almond and their hybrid). To analyzed TE transcription, we used the methodology described in Vendrell-Mir *et al.* (2020) with minor modifications. After filtering reads, they were mapped to the TE annotation using Bowtie2 (Langmead & Salzberg, 2012). All the mapped reads were extracted using Samtools (Danecek *et al.*, 2021). These reads were assembly to contigs using Trinity (Grabherr *et al.*, 2011). We aligned them to our TE annotation library using BLASTn with an e-value cutoff of 10-20 and word size of 400. The most similar genomic copy of our annotation to each assembled contig was identified as a family representative (Supplementary Data S8 of de Tomás *et al.*, 2022), using a length coverage cut-off of 80% for retrotransposons and 40% for DNA transposons. Next, RNA-seq reads were aligned to peach transcript models concatenated with peach and almond representative TEs using Bowtie2 (Langmead & Salzberg, 2012) through the RSEM package (Li & Dewey, 2011) in default parameters. Only reads aligned in the sense strand were kept for quantification. Additionally, in the first RNA-seq analysis (with only leaf samples), we performed this latter alignment using almond transcript models concatenated with peach and almond representative TEs. The objective was to ensure that there was not significant impact due to the genome used.

Differential expression analysis between peach and almond was performed using DESeq2 (Love *et al.*, 2014) with a Log fold-change cut-off of one. False Discovery Rate (FDR) was employed for multiple-testing correction in the differential expression analysis involving RNAseq leaf samples exclusively. On the other hand, Bonferroni correction was utilized for the analysis involving RNAseq of the three organs due to its higher stringency and the higher data coverage available. DESeq2 regularized log (rlog) values were used as normalized expression data.

Validation of retrotransposon and gene differential transcription on leaves

Quantitative RT-PCR analyses were performed using three independent RNA extractions per genotype. The cDNAs were synthesized using SuperScript® III

Reverse Transcriptase (Invitrogen™, Carlsbad, CA, USA). The primers used were designed following the next parameters: size of 20-22 bps, product size of 100-200 bps, GC percentage of 40-60%, maximum conservation between peach and almond sequences and between all the copies of each cluster (Table 3). Each qRT-PCR reaction consisted of 5 µl of Roche's SYBR green Master Mix (Roche Applied Science), 0,3 µl of 10 µM forward primer, 0,3 µl of 10 µM reverse primer, 40 ng of cDNA in a total volume of 10 µL. The qRT-PCR were performed in a Roche LightCycler II with the initial denaturation step of 5 min at 95°C, followed by 40 cycles (10 s at 95°C, 10 s at 56°C, and 10 s at 72°C). The Translation Elongation Factor (TEF2) and the RNA Polymerase II (RPII) were used as internal controls to normalize the expression of the tested LTR retrotransposons (Tong *et al.*, 2009). Two biological replicates with two technical replicates were used with negative reverse transcriptase and non-template controls. The relative levels of retrotransposon and gene expression were calculated using the 2 − ΔΔCt method. The specificity of the primers and their product length were verified by agarose gel electrophoresis. The primers for qRT-PCR are listed in Table 3.

Table 3. DNA primers used for validation of retrotransposon and gene differential transcription on leaves using qRT-PCR.

PRIMER NAME	SEQUENCE	ANALYZED
P003_M_F	ACYTGGCAGTGTCCAACTCA	LTR_3
P003_M_R	TCTATAGCCCACTTCATGAG	
P048_F	GGCATTTGCTAGYCTYAGTG	LTR_48
P048_M_R	AACAYTTTGGYTTRCCCTTG	
P053_F	CTGATTCCTTGCTCATAGCA	LTR_53
P053_M_R	TCATCGATGATCACTCTCSG	
P081_M_F	GTGGTTCTACTTGCATATGC	LTR_81
P081_R	TATCTCCACATCCTTTGGCC	
P088_M_F	TGTGTCTCAATTCAGTTGGC	LTR_88
P088_R	TTACATGAGAAGGGAATGCC	
P108_F	GGTTAGATCTCATGAAGGGA	LTR_108
P108_M_R	GTTCCTTCCAATTCTTCCAC	

P124_F	GGATGAAGCTTGGTGTGATG	LTR_124
P124_M_R	CAAAGTCCACYTTCTCCCAT	
P138_F	GTTCCTCTTCAATTGGGTCC	LTR_138
P138_M_R	GGTTGAACATCCTTAKTTGG	
P141_M_F	GCATCACACATTTTGTRCTC	LTR_141
P141_R	GTCCTACTCATGCTGTGAAG	
821_1F	GGCTATGTCTTTGTAATGGG	*Prupe.1G388900*
821_2R	TTTGGAGGGAGGCCAATAGC	
1022_1F	CGACACTGTTGTGAAACCTC	*Prupe.1G547300*
1022_2R	TCTTCCTGCCCTAACCCAAG	
1616_1F	TCCAACTTCTGGCCAATTGA	*Prupe.2G111400*
1616_2R	ATATGGAACATATGCAACGC	
1865_1F	CACATGAGCAAGGAGACCTT	*Prupe.2G269400*
1865_2R	GCAAGTATGATATCTTCCAA	
948_1F	GACTGGCATCCGAGGAAGAA	*Prupe.5G169500*
948_2R	TAATTCCGGTCTTCACGATC	
4241_1F_M	CAACTCGGGCGTAATCAATC	*Prupe.6G220500*
4241_2R_M	GATGATCCAGAATACCAGCT	
TEF2_F	GGTGTGACGATGAAGAGTGATG	*Prupe.4G138700* (TEF2)
TEF2_R	TGAAGGAGAGGGAAGGTGAAAG	
RPII_F	TGAAGCATACACCTATGATGATGAAG	*Prupe.8G132000* (RPII)
RPII_R	CTTTGACAGCACCAGTAGATTCC	

Identification of the expressed TE copies on leaves

The identification of the expressed TE copies was performed using RT-PCR using different sets of primers (Table 4). 1 µg of total RNA was reverse transcribed using SuperScript® III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA). The primers used were designed with the next parameters: size of 20-22 bps, product size of 400-1400 bps, GC percentage of 40-60%, maximum conservation between peach and almond sequences and between all the copies of each cluster (Table 4). Each RT-PCR reaction consisted of 2 µl of DreamTaq DNA Polymerase Buffer (Thermo Scientific™, Carlsbad, CA, USA), 1 µl of 2mM dNTPs, 1 µl of 10 µM forward primer, 1 µl of 10 µM reverse primer, 0,15 µl of DreamTaq DNA Polymerase (Thermo Scientific™, Carlsbad, CA, USA), 20 ng of cDNA obtained by reverse transcription in a total volume of 20 µL. PCR cycling conditions were 2 min at 95°C, followed by 35 cycles of 95°C for 30 s, 30 s at the annealing Tm and 1 min/kb at 72°C, and a final step of 10 min at 72°C. Negative reverse transcriptase and non-template controls were used. PCR results were observed using a 1% (w/v) agarose gel. PCR fragments were extracted using Macherey-Nagel™ NucleoSpin™ Gel and PCR Clean-up Kit (Fisher Scientific, UK), cloned into the pGEM-Teasy plasmid using pGEM®-T Easy Vector Systems kit (Promega Corporation, Madison, WI, USA), introduced into *E.coli* TOP10 cells and amplified on LB plates containing carbenicillin, X-Gal and IPTG. 8 colonies were selected using the blue/white screening and grew each one in 3ml of LB overnight. Finally, the DNA was extracted using Macherey-Nagel™ NucleoSpin Plasmid QuickPure™ Kit (Fisher Scientific, UK) and their inserts were sequenced by Sanger Sequencing using an ABI 3730 DNA Analyzer for capillary electrophoresis (40 capillaries) and fluorescent dye terminator detection (BigDye ®Terminator) in the DNA Capillary Sequencing Facility of CRAG. Sequences were compared with the parental genomes and the expressed copies were identified only if the amplified sequences were more than 99,5% identical to the genomic sequence.

Table 4. DNA primers used for the identification of the expressed TE copies using RT-PCR.

PRIMER NAME	SEQUENCE
P003_F	AAACAACCATGGTTGGAAGC
P003_R	GAGGTGTATTYTTTGGTGAA
P028_F	ATGCATMARTGTGTACCTCA
P028_R	CTAGGAATGRAGTTCATGGA
P048_F	GGCATTTGCTAGYCTYAGTG
P048_R	ACACCATTTTGYTGTRGTGT
P053_F	CTGATTCCTTGCTCATAGCA
P053_R	CAATGAAGAGTYTTGGGTGT
P081_F	ACTCTGGCCCTGTTGAGCAA
P081_R	TATCTCCACATCCTTTGGCC
P088_F	TAATGGTGTCCAATCTGGCT
P088_R	TTACATGAGAAGGGAATGCC
P108_F	GGTTAGATCTCATGAAGGGA
P108_R	GATGCTAGGCTCTGCGTGTA
P124_F	GGATGAAGCTTGGTGTGATG
P124_R	AGATAAGTTGTCCATAGAAC
P138_F	GTTCCTCTTCAATTGGGTCC
P138_R	GCAGCCAAATCAAGAYCATG
P141_F	CCTTTAGCTACTAACCTGGC
P141_R	GTCCTACTCATGCTGTGAAG

Characterization of a powdery mildew resistance candidate gene

The presence of a polymorphic insertion within a candidate powdery mildew resistance gene was analyzed using three primer combinations, that enabled us to detect its presence or absence. The primer design adhered to the following parameters: a size of 20-22 bps, a GC percentage of 40-60%, and maximal conservation between peach and almond sequences (Table 5). Each PCR reaction comprised 2 µl of Dreamtaq DNA Polymerase Buffer (Thermo Scientific™, Carlsbad, CA, USA), 1 µl of 2mM dNTPs, 1 µl of 10 µM forward primer, 1 µl of 10 µM reverse primer, 0.15 µl of DreamTaq DNA Polymerase (Thermo Scientific™, Carlsbad, CA, USA), 20 ng of DNA, in a total volume of

20 µL. PCR cycling conditions involved an initial 2 min at 95°C, followed by 35 cycles of 95°C for 30 s, 30 s at the annealing temperature (Tm), and 1 min/kb at 72°C, concluding with a final step of 10 min at 72°C. PCR results were visualized through 1% (w/v) agarose gel electrophoresis.

Table 5. DNA primers used for to detect a polymorphic insertion using PCR.

PRIMER NAME	SEQUENCE
1616_1F	TCCAACTTCTGGCCAATTGA
1616_3R	CAGCTTACATAGTGTGTTTG
1616_4F	GGTGTACTGCAAAACAGAAG
1616_5R	AAGTAAGGACTTCTATCTCC

The transcripts of this gene were analyzed by RT-PCR and Sanger sequencing. 1 µg of total RNA was reverse transcribed using SuperScript® III Reverse Transcriptase (Invitrogen, Carlsbad, CA, USA). The primers used were designed with the next parameters: size of 20-22 bps, GC percentage of 40-60%, maximum conservation between peach and almond sequences (Table 6). Each RT-PCR reaction consisted of 2 µl of DreamTaq DNA Polymerase Buffer (Thermo Scientific™, Carlsbad, CA, USA), 1 µl of 2mM dNTPs, 1 µl of 10 µM forward primer, 1 µl of 10 µM reverse primer, 0,15 µl of DreamTaq DNA Polymerase (Thermo Scientific™, Carlsbad, CA, USA), 20 ng of cDNA obtained by reverse transcription in a total volume of 20 µL. PCR cycling conditions were 2 min at 95°C, followed by 35 cycles of 95°C for 30 s, 30 s at the annealing Tm and 1 min/kb at 72°C, and a final step of 10 min at 72°C. Negative reverse transcriptase and non-template controls were used. PCR results were observed using a 1% (w/v) agarose gel. PCR fragments were extracted using Macherey-Nagel™ NucleoSpin™ Gel and PCR Clean-up Kit (Fisher Scientific, UK) and were sequenced by Sanger Sequencing using an ABI 3730 DNA Analyzer for capillary electrophoresis (40 capillaries) and fluorescent dye terminator detection (BigDye ®Terminator) in the DNA Capillary Sequencing Facility of CRAG.

Table 6. DNA primers used for the detection and analysis of the transcripts of the analyzed gene.

PRIMER NAME	SEQUENCE
1616_1F	TCCAACTTCTGGCCAATTGA
1616_2R	ATATGGAACATATGCAACGC
1616_3R	CAGCTTACATAGTGTGTTTG
1616_4F	GGTGTACTGCAAAACAGAAG
1616_5R	AAGTAAGGACTTCTATCTCC
1616_6F	GGATTGATAACTCCTTCGGG
1617_7R	GGAATGGCTTTCCTTCAGAC
16167_8F	GGAGATAGAAGTCCTTACTT
1616_9R	GGTTGGGAAGACATGCTTGA
1616_10R	GCTCGAGCAGTTTACCAGGA

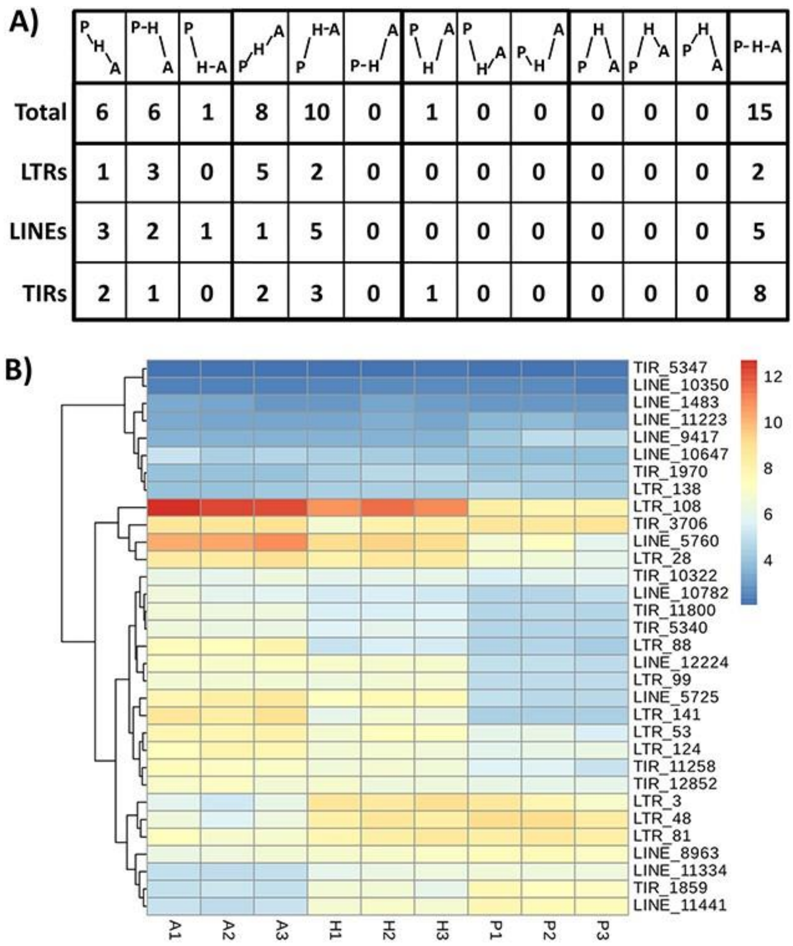
1.4. Results

RNA-seq on leaves

Analysis of the potential changes in the transcription of the transposable elements in leaves of the peach x almond hybrid

To study the potential activation of TEs by the interspecific cross, we performed an RNA-seq analysis of the expression in mature leaves of almond, peach and the hybrid. We analyzed the possible expression of LTR retrotransposons, LINEs and TIRs (Figure 1A). We found significant transcription levels for 47 TE families: 13 LTR retrotransposon, 17 LINE and 17 TIR. Among them, we found significant differential transcription between the almond, peach and/or the hybrid in 32 families (Figure 1A and B): 11 LTR retrotransposon, 12 LINE and 9 TIR. In most of the cases, the differential expression is due to differences between peach and almond and the expression in the hybrid is intermediate. In 18 families the expression was significantly higher in almond than in peach, and in 13 was the opposite. Only one of the TE families showed lower significant expression in the hybrid than in the two parental (TIR_3706) and none was expressed at a higher level than in the two parental species.

A more detailed analysis of the transcribed LTR retrotransposon families showed that in one case (LTR_99) only the almond genome contains full-length copies and this is correlated with a higher expression in almond (Table 7). In 8 of the other 10 transcribed families the species with higher levels of transcription is the one with higher copy number and, in the cases in which that does not happen (LTR_48 and LTR_124), the copy numbers are very similar. On the other hand, all the transcribed LTR retrotransposon families except one (LTR_138) contain relatively young copies with an estimated age of 1,4 Mya or less. In all the families, the parental species with the younger element is the one showing higher transcription (Table 7).



	P \H \A	P-H \ A	P \ H-A	A /H P/	H-A / P	A / P-H	P A \ / H	P \ A H/	A P\ / H	H / \ P A	H / \A P	H P/ \ A	P-H-A
Total	6	6	1	8	10	0	1	0	0	0	0	0	15
LTRs	1	3	0	5	2	0	0	0	0	0	0	0	2
LINEs	3	2	1	1	5	0	0	0	0	0	0	0	5
TIRs	2	1	0	2	3	0	1	0	0	0	0	0	8

Figure 1. Transcriptomic analysis of transposable elements in leaves of almond, peach and the hybrid. A) Patterns of transcription of the TE families. Vertical higher position indicates more transcription (P, peach; H, hybrid; A, almond). Total means the total number of families in peach and almond genomes containing at least one full-length element. B) Heat-map of the transcription levels of the TE families showing differential expression. Higher expression is indicated in red and lower expression in dark blue. We show the results of three replicates per sample.

Table 7. Differentially transcribed LTR retrotransposon families. A, almond; P, peach; H, hybrid.

FAMILY	TYPE	PATTERN OF TRANSCRIPTION	NUMBER OF FULL COPIES		MINIMAL ESTIMATED TIME OF INSERTION (MYA)	
			ALMOND	PEACH	ALMOND	PEACH
3	Gypsy	A < (P = H)	11	27	1.7	0.0
28	Gypsy	(A = H) > P	16	10	1.7	0.6
48	Copia	A < (P = H)	26	23	0.0	0.0
53	Copia	A > H > P	26	12	0.0	0.0
81	Copia	A < (P=H)	7	8	1.4	1.4
88	Copia	A > (H = P)	10	5	0.8	0.8
99	Copia	(A = H) > P	3	0	1.4	-
108	Gypsy	A > H > P	14	6	0.0	0.5
124	Unclassified	A > H > P	3	4	10.4	4.3
138	Unclassified	A < H < P	4	11	0.8	0.6
141	Copia	A > (H = P)	5	2	1.3	4.6

To validate the RNA-seq data we performed qRT-PCR analysis for nine of the LTR retrotransposon families differentially expressed and the results confirmed their expression profile showing a general agreement with the transcript abundance estimated by RNA-seq (Figure 2), except for LTR_124, which contains the oldest copies (Table 7). There may be remains of copies that could be inferred in our qRT-PCR result.

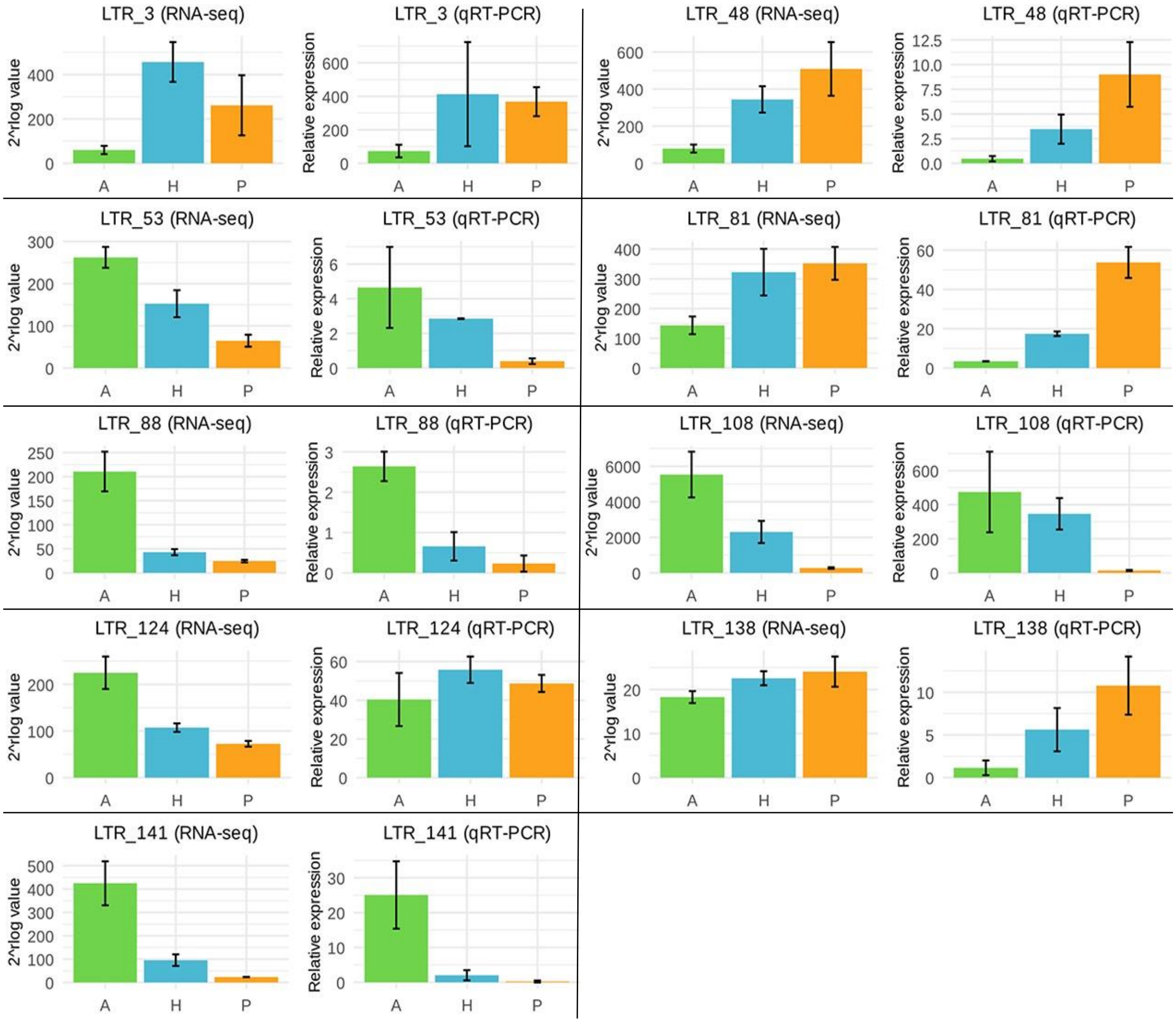


Figure 2. Transcription levels of LTR retrotransposons differentially expressed in leaves of almond (A, green), peach (P, orange) and the F1 hybrid (H, blue) analyzed by qRT-PCR and RNA-seq. The LTR retrotransposon family is indicated in the top.

Next, we tried to identify which of the copies of the differentially transcribed LTR retrotransposon families were the ones that produce transcripts in leaves. For most of the LTR retrotransposon families with differential expression several copies are expressed but from four of them we were able to determine a single copy responsible of all, or most, of the transcription, being in most cases recent insertions present close to genes or inside a gene. In the family LTR_3 the expressed copy is in chromosome 6 of peach (27,139,890–27,148,350), is estimated to be 1,8 Mya old and is located inside a gene (in an intron) in the same orientation. In the family LTR_81 the expressed copy is in the chromosome 3 of peach (19,702,391–19,706,564), is estimated to be 1,9 Mya old and is located 3,2 kB from the closest gene. In the family LTR_124 the expressed copy is in the chromosome 7 of almond (21,257,163–21,260,133), is estimated to be 10,4 Mya old and is located inside a gene (in an intron) in the same orientation. This family LTR_124 is an exception and is not young, unlike the rest. Finally, in the family LTR_141 the expressed copy is in the chromosome 8 of almond (10,143,038–10,147,556), is estimated to be 1,3 Mya old and is located 35 bases downstream a gene but in opposite orientation.

Another interesting case is the family LTR_155. This family is one of the two expressed LTR families without significant differential expression, so we classified it in the category of families with the same expression in both parents as in the hybrid (Figure 1A). However, we have observed a higher expression of this family in the hybrid and peach than in almond, although it was not significant (Table 8). This family contains 4 completed copies for almond and 2 for peach. We have identified expression for both complete copies in peach. The first copy is in chromosome 7 of peach (11,439,955-11,444,586) and is estimated to be 7.4 Mya old. This copy is within the first intron of a gene. Its orthologous copy in almond has truncated transcription due to an additional insertion within this copy, affecting its transcription. The second copy of peach is also expressed and is located near a gene (at a distance of 750 bps from the ATG) in the chromosome 8 (1,334,298-1,338,928). It is very young because it is estimated to be 0.96 Mya. However, we have been unable to find its orthologous region in almond.

Table 8. RNA-seq expression (Deseq2 regularized log values) and Relative expression by qRT-PCR of family LTR_155.

	ALMOND	**HYBRID**	**PEACH**
RNA-seq	9.179	9.874	9.308
qRT-PCR	34.511	42.629	97.935

We have found cases of copies that do not present differential expression in the sense orientation but are transcribed in antisense. For example, a copy of the family LTR_82, present in both peach and almond with an estimated age of 2.4 Mya old, is transcribed within a gene. This transcription occurs in the opposite direction to the gene because the strand of the gene and the strand of the copy are opposite. This copy is located in the chromosome 5 of peach (174,911-

176,479). The same occurs in LTR_138, where a copy is transcribed antisense in peach but is absent in almond. This copy is inserted within an intron of a gene, oriented in the opposite direction to the gene. It is located in chromosome 2 (4,542,106-4,545,556) and is estimated to be 0,6 Mya old.

All these data suggest that the merging of the genomes of peach and almond in a hybrid does not greatly deregulate the expression of TEs and that the differential expression of TEs between these two genomes is mainly due to the presence or absence of transcriptional active copies in each of them.

Analysis of the potential changes in gene transcription in the peach x almond hybrid on leaves

We analyzed the possible changes in gene expression in the hybrid. As already mentioned, the peach and almond genome show a high degree of sequence identity (mean of 97.99% in regions aligning 1:1) (Donoso *et al.*, 2016). Therefore, to facilitate the comparison of the level of expression, we decided to map the RNA-seq reads from peach, almond and the hybrid to a single gene model dataset, that of peach or, alternatively, that of almond. A global comparison of the expression levels of 13,620 genes orthologous between peach and almond showed an almost perfect correlation between the two deduced profiles (Pearson correlation coefficient = 0.99) (Figure 3).

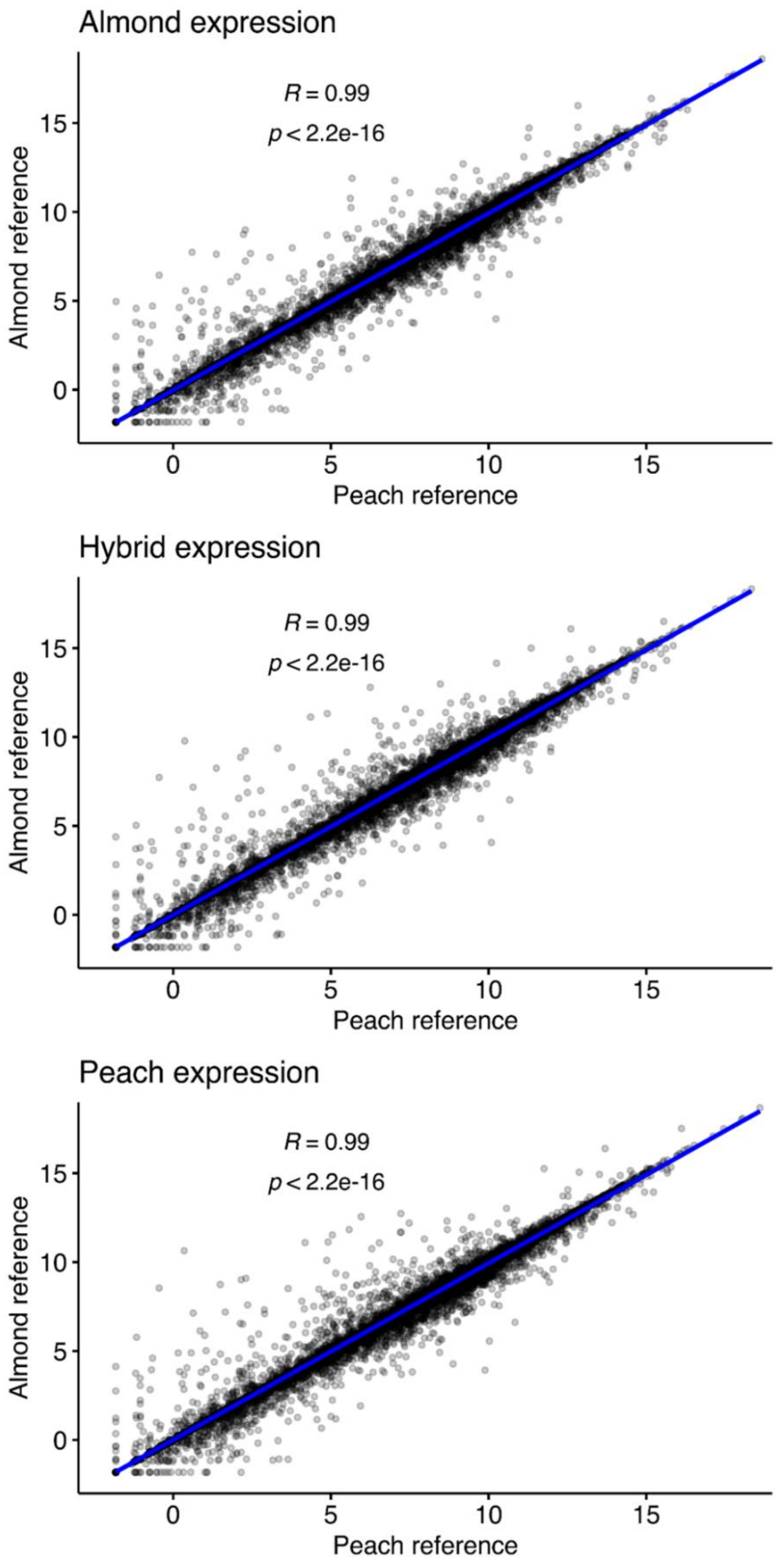


Figure 3. Global comparison of the expression between peach and almond orthologous genes using almond and peach genomes.

We present here the results of the analysis performed using the peach gene models (Figure 4) and the almond gene models (Figure 5). From the 26,873 genes present in the annotated 'Lovell' peach genome (Verde *et al.*, 2017) (Figure 4), we found that 22,274 of them are significantly expressed in at least one of the three genotypes (peach, almond and hybrid). For 17,439 genes (78.3%) the levels of transcription were similar in the three genotypes. For 2,389 genes (10.7%) the expression was higher in peach respect to almond and in the F1 hybrid the expression was intermediate or similar to one of the two parents. For 2,234 genes (10.0%) the expression was higher in almond respect to peach and in the F1 hybrid was intermediate or similar to one of the two parents. In 152 genes (0.7%) the expression was higher in the hybrid than in the two parental and in 60 genes (0.3%) the expression was lower in the hybrid.

From the 27,044 genes present in the annotated 'Texas' almond genome (Alioto *et al.*, 2020) (Figure 5), we found that 21,074 of them are significantly expressed in at least one of the three genotypes (peach, almond and hybrid). For 17,116 genes (81.2%) the levels of transcription were similar in the three genotypes. For 1,729 genes (8,2%) the expression was higher in peach respect to almond and in the F1 hybrid the expression was intermediate or similar to one of the two parents. For 2,041 genes (9.7%) the expression was higher in almond respect to peach and in the F1 hybrid was intermediate or similar to one of the two parents. In 148 genes (0.7%) the expression was higher in the hybrid than in the two parental and in 40 genes (0.2%) the expression was lower in the hybrid.

The summary of the results using the two annotations are presented in Table 9.

Table 9. Summary of gene transcription analysis using almond and peach annotation.

	ALMOND	**PEACH**
Annotated genes	27,044	26,873
Genes significantly expressed in at least one of the three genotypes	21,074	22,274
Differential expression genes	3,958	4,835
Same expression genes	17,116	17,439
Differential expression genes (%)	18.8	21.7
Same expression genes (%)	81.2	78.3

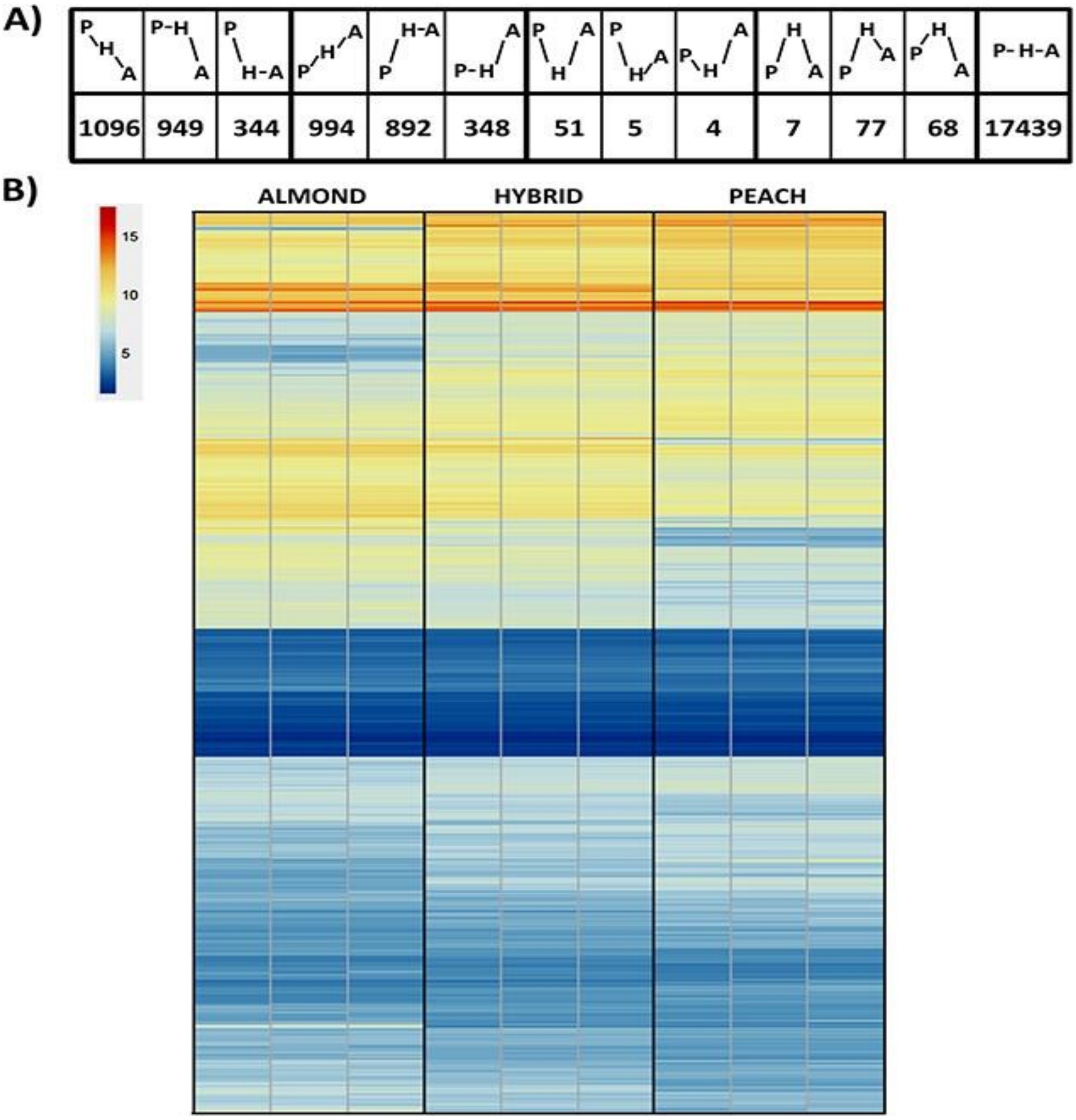


Figure 4. Transcription of genes in leaves of almond, peach and the hybrid using peach transcripts (Verde *et al.*, 2017). A) Patterns of transcription of the genes. Vertical higher position indicates more transcription (P, peach; H, hybrid; A, almond). B) Average transcription levels of the genes showing significant differential expression in peach, the F1 hybrid and almond. Higher expression is indicated in red and lower expression in dark blue.

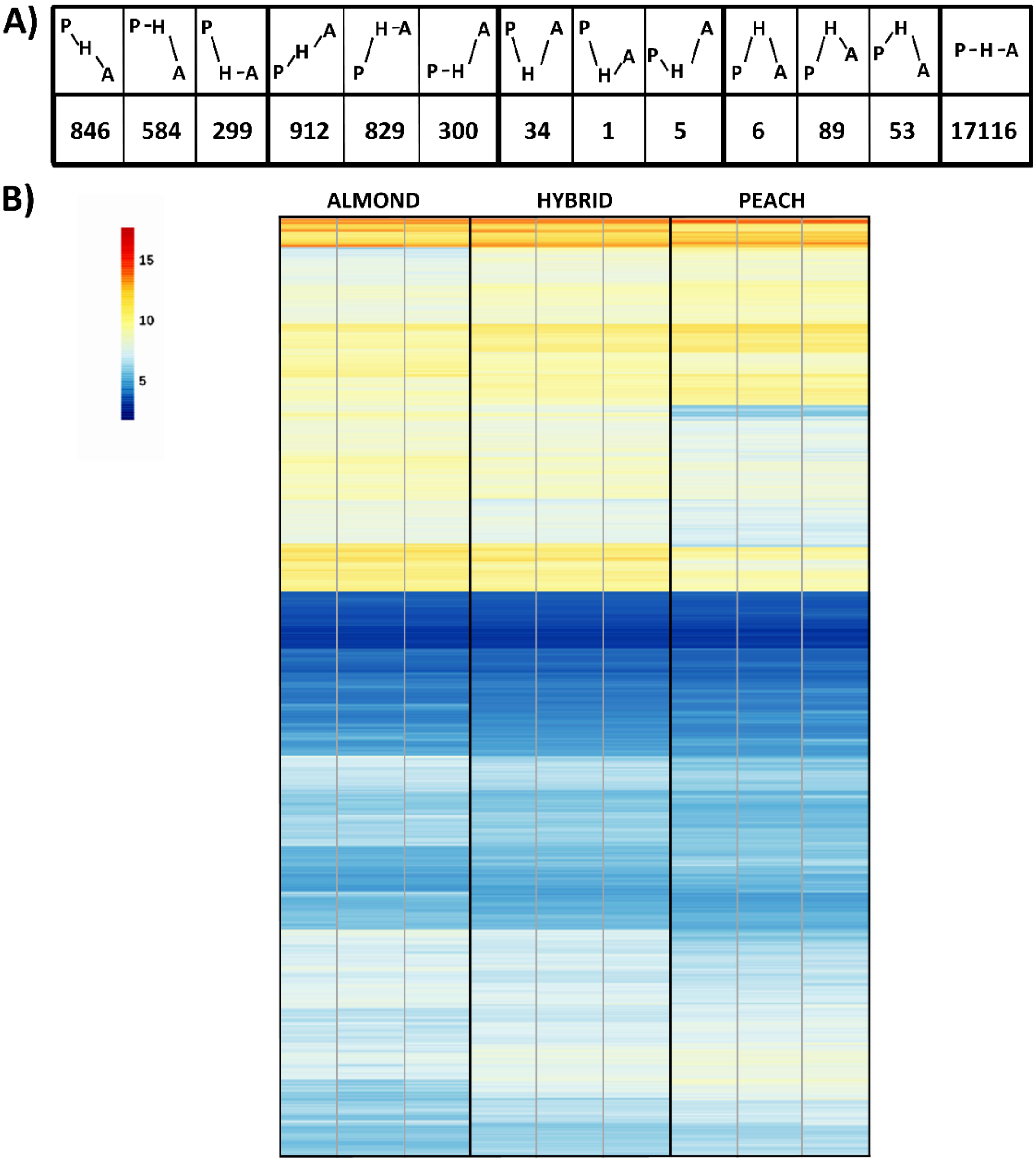


Figure 5. Transcription of genes of genes in leaves of almond, peach and the hybrid using almond transcripts (Alioto *et al.*, 2020). A) Patterns of transcription of the genes. Vertical higher position indicates more transcription (P, peach; H, hybrid; A, almond). B) Average transcription levels of the genes showing significant differential expression in peach, the F1 hybrid and almond. Higher expression is indicated in red and lower expression in dark blue.

To validate the RNA-seq data of genes we performed qRT-PCR analysis for six genes differentially expressed and the results confirmed their expression profile showing a general agreement with the transcript level estimated by RNA-seq (Figure 6).

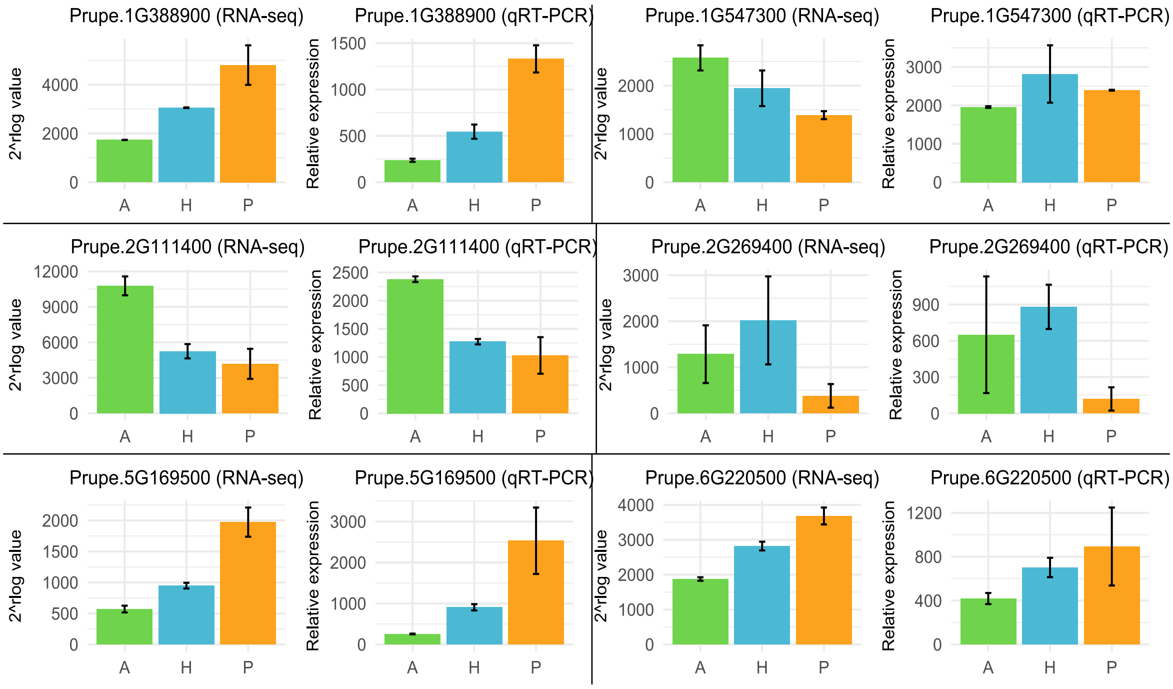


Figure 6. Transcription levels of genes in leaves of almond (A, green), peach (P, orange) and the F1 hybrid (H, blue) analyzed by qRT-PCR and RNA-seq. The name of the gene is indicated in the top.

Among the genes whose expression is significantly lower in the hybrid there are two genes, *Prupe.1G332600.1* and *Prupe.1G334500.1* (Table 10), annotated as potentially encoding an “RNA-dependent RNA polymerase, eukaryotic-type” showing sequence similarity to the Arabidopsis RDR1 (AT1G14790), one of the enzymes involved in the production of sRNAs, in the defense against viruses (Leibman *et al.*, 2018) and in gene regulation and DNA methylation (Wang *et al.*, 2014). We next analyzed other genes possibly involved in DNA methylation encoding DNA methyltransferases or DNA demethylase (Table 10) but none of them showed differential expression among peach, almond and the hybrid.

Table 10. Transcription of genes encoding proteins possibly involved in DNA methylation using peach transcripts annotation.

	ALMOND		HYBRID		PEACH	
	rlog value	**SD**	**rlog value**	**SD**	**rlog value**	**SD**
RNA-dependent RNA polymerase, eukaryotic-type						
Prupe.1G332600.1	5,89	0,34	5,09	0,16	5,57	0,42
Prupe.1G334500.1	6,11	0,43	5,45	0,19	6,04	0,28
DNA methyltransferase						
Prupe.7G183100.1	7,08	0,30	6,64	0,19	6,67	0,30
Prupe.6G011600.1	2,70	0,03	2,71	0,06	2,96	0,15
Prupe.8G038800.1	9,62	0,33	9,21	0,15	8,97	0,15
Prupe.6G322700.1	10,89	0,10	10,90	0,14	10,91	0,26
DNA demethylase						
Prupe.7G118000.1	11,37	0,30	11,36	0,13	11,26	0,09
Prupe.7G005000.1	12,34	0,13	12,41	0,15	12,11	0,27
Prupe.6G119100.1	6,48	0,37	6,63	0,10	7,02	0,09

Relationship between differential gene expression and polymorphic TEs

We examined the relationship between differential expression in almond and peach and the presence of nearby transposons in their upstream region (1 Kb). We focused on the 4,429 genes that had differential expression between the two parentals (Figure 4). Our analysis showed that among these 4,429 genes, 307 of them (6.9% of the genes with differential expression) had a non-polymorphic insertion in both genomes. Also, our analysis showed that 190 genes had a polymorphic insertion, 163 present in the peach and 27 present in almond genome (4.2% of the total). Among these 190 genes, 107 had an LTR retrotransposon, 61 had MITEs, 16 had TIRs and 6 had LINEs. However, the total number of genes harboring non-polymorphic transposons (LTR retrotransposons, LINEs, TIRs and MITEs) in the upstream region of the two both species was 1807, the total number of genes with insertions only in peach was 1038 and the total number of genes with insertions only in almond was 176. This indicates that the capability to detect polymorphic insertions was limited, particularly in almond, that is very heterozygous (Table 11).

Table 11. Count of analyzed genes using the annotation of peach.

	Count
Analyzed genes	22274
Genes with non-polymorphic TEs	1807
Genes with polymorphic TEs	1214 (176 present in almond, 1038 present in peach)
Genes with differential expression	4429
Genes with differential expression and with non-polymorphic TEs	307
Genes with differential with polymorphic TEs	190 (27 present in almond, 163 in peach)

We conducted an additional analysis to determine whether the presence of an insertion increased or decreased the expression of the nearby gene. We examined the differential expression among the 163 genes that had a polymorphic insertion present in peach and absent in almond (Figure 7). Among these 163 genes, we observed that 93 exhibited higher expression in peach compared to 70 that displayed higher expression in almond.

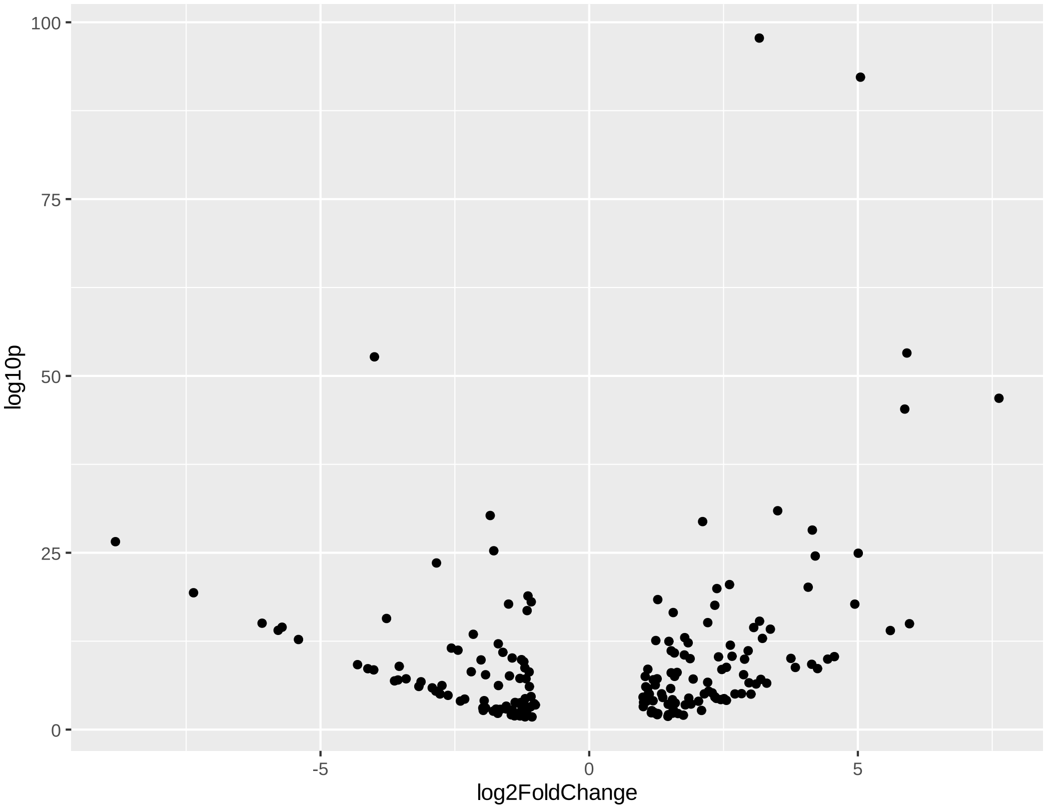


Figure 7. Differential expression on genes with a polymorphic insertion present in peach and absent in almond. Positive log2FoldChange indicate more expression in peach, negative log2FoldChange indicate more expression in almond.

RNA-seq on leaves, flowers and fruits

Analysis of the potential changes in the transcription of the transposable elements in leaves, flowers and fruits of the peach x almond hybrid

The RNA-seq results from the leaves allowed us to determine that significant transcriptional changes were not occurring in the almond and the peach genomes, and their F1 hybrid. However, we wanted to analyze other organs, that present bigger differences between peach and almond, in order to know if there is transposon activation in the hybrid. To explore this possibility, we performed an RNA-seq analysis including leaf, flower, and fruit samples. The Principal Component Analysis (PCA) show that replicates of each sample are very consistent and do not have a strong variability (Figure 8). The assembly of transposons using a larger number of samples improved our ability to identify transcribed transposons compared to the previous RNA seq analysis.

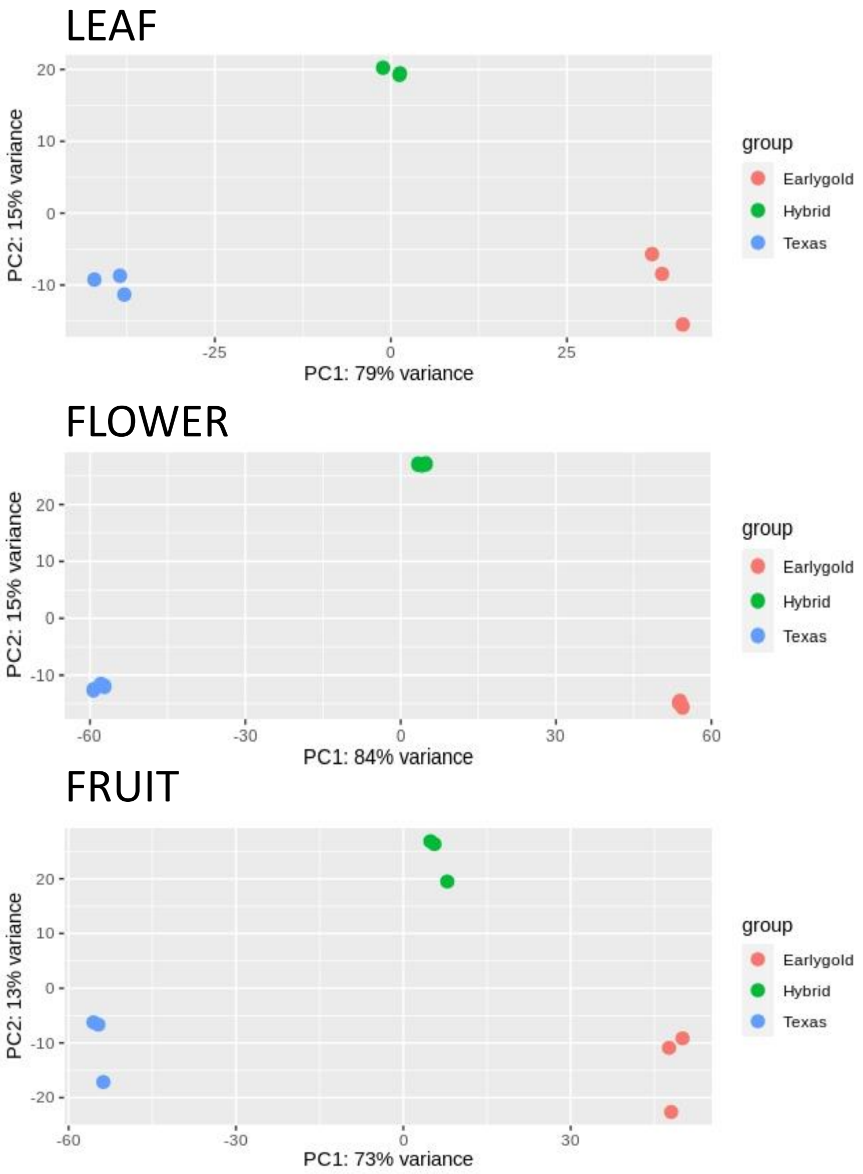
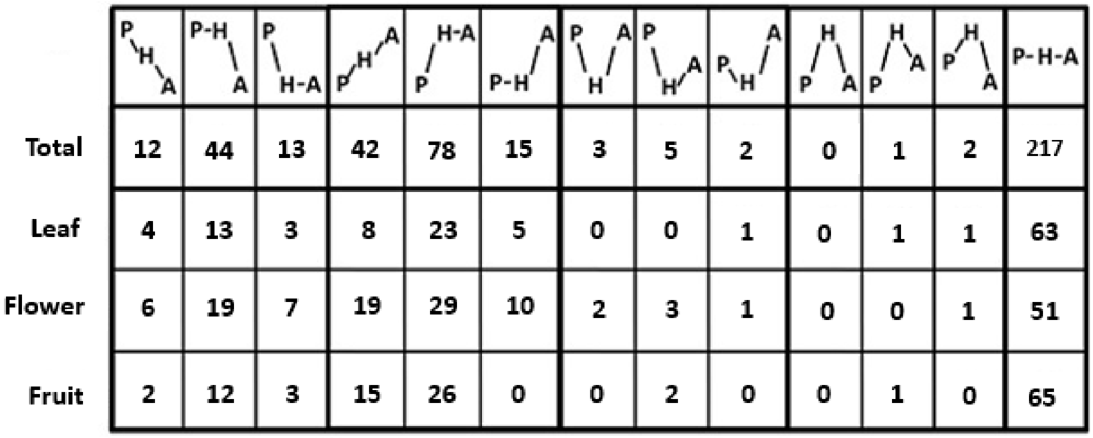


Figure 8. Principal Component Analysis (PCA) for the three replicates of leaf, flower and fruit samples. Each genotype is presented: Earlygold (peach), Hybrid (F1) and Texas (almond).

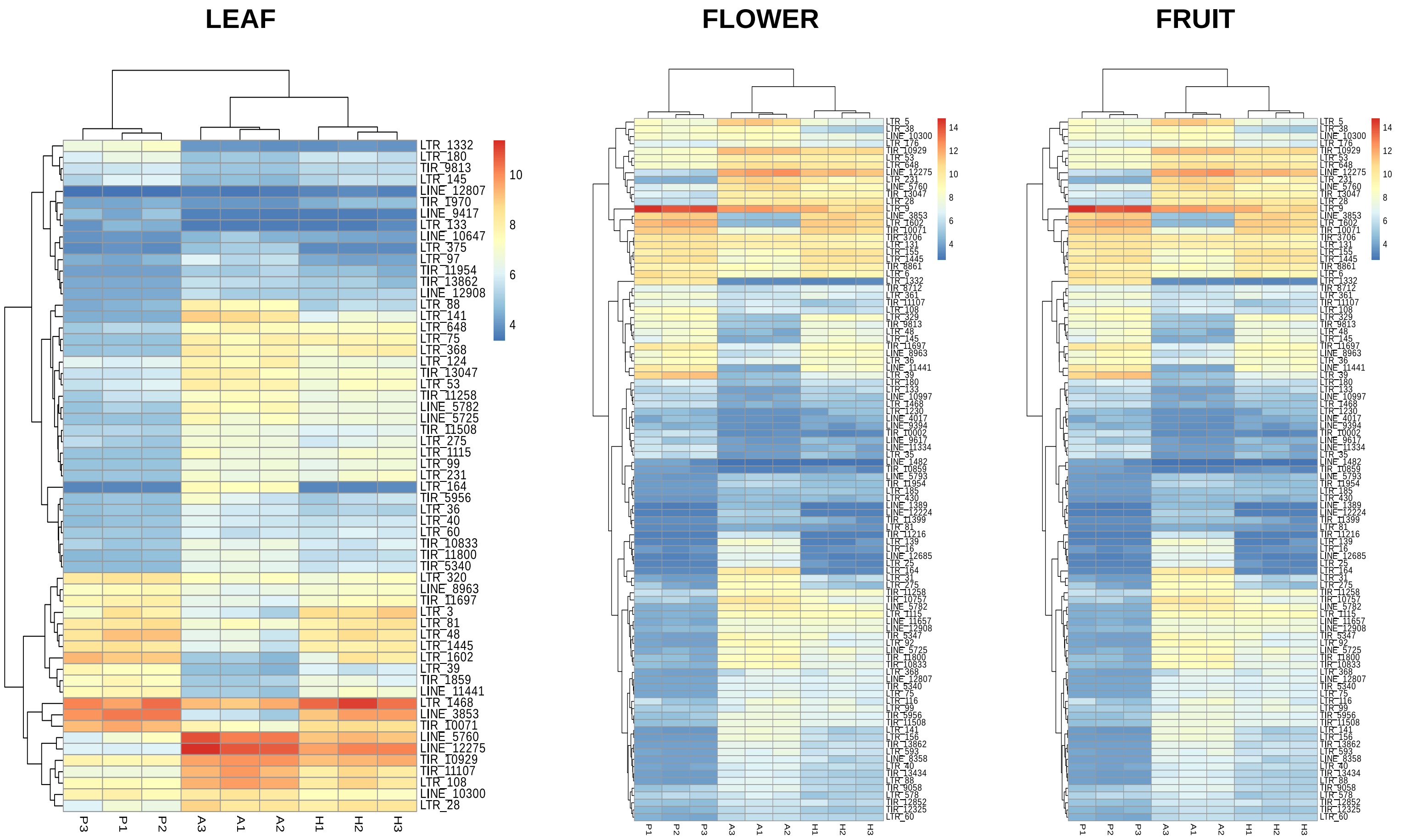
We found significant transcription (DESeq rlog value > 2) in 122 families in leaf, 126 families in fruit and 148 families in flower. The same pattern is observed across all organs. The hybrid is transcribed at a similar level to both parents, either like one of the parents or at an intermediate point between them (Table 12). Cases where this does not occur are sporadic: 3 families in leaf, representing 2.46% of the 122 leaf families (1 family with reduced expression in the hybrid and 2 with elevated expression), 7 families in flower, representing 4.73% of the 148 expressed flower families (6 with reduced expression in the hybrid and 1 with elevated expression) and 3 families in fruit, representing 2.38% of the 126 expressed fruit families (2 with reduced expression in the hybrid and 1 with elevated expression). The differences between the hybrid and parents expression of these families are not very high in the three organs (Figure 9).

Table 12. Patterns of transcription of the TE families for each organ. "Total" is the summation of TE families detected in the three organs, which, in some cases, overlap and are the same for the three organs. Vertical higher position indicates more transcription (P, peach; H, hybrid; A, almond).



	P \H \A	P-H \ A	P \ H-A	A H/ P/	H-A / P	A / P-H	P A \ / H	P \ A H/	A P\H/	H / \ P A	H / \A P	H P/ \ A	P-H-A
Total	12	44	13	42	78	15	3	5	2	0	1	2	217
Leaf	4	13	3	8	23	5	0	0	1	0	1	1	63
Flower	6	19	7	19	29	10	2	3	1	0	0	1	51
Fruit	2	12	3	15	26	0	0	2	0	0	1	0	65

Figure 9. Heat-map of the transcription levels of the TE families with differential expression. Higher expression is indicated in red and lower expression in dark blue. We show the results of three replicates per sample (P, peach; H, hybrid; A, almond).



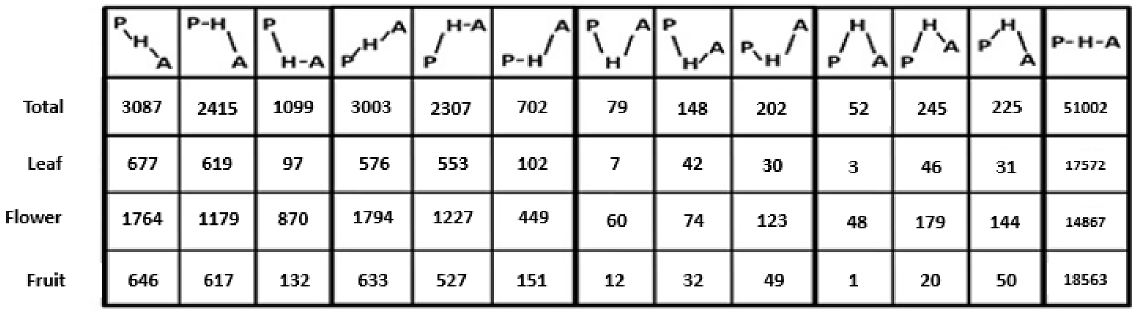
Analysis of the potential changes in the transcription of the genes in leaves, flowers and fruits of the peach x almond hybrid

In this analysis, we exclusively used the peach gene annotation as a reference due to our RNA-seq analysis on leaves demonstrated that the results were very similar using both parental genome annotations. From the 26,873 genes present in the annotated 'Lovell' peach genome (Verde *et al.*, 2017) (Figure 4), we found that 20,354 of them are significantly expressed (rlog value > 2) in leaf, 22,778 in flower and 21,433 in fruit.

We observe similar results for the three organs. Among the 20,354 expressed genes in leaf, 79 genes presented less expression in the hybrid compared to both parents (0.39%), while 80 genes presented higher expression in the hybrid compared to both parents (0.39%). Among the 22,778 expressed genes in flower, 257 presented less expression in the hybrid compared to both parents (1.13%) and 371 genes had more expression in the hybrid compared to both parents (1.63%). Within the 21,433 expressed genes in fruit, 93 had less expression in the hybrid compared to both parents (0.43%) and 71 had more expression in the hybrid compared to both parents (0.33%).

In general, just as is the case with transposons and with RNA-seq analysis conducted only on leaves, the hybrid is transcribed at a level similar to that one of two of the parents, or at an intermediate point between the two parents. These findings once again suggest that there is no transcriptional activation in the hybrid (Table 13). In the case of genes that are transcribed more or less in the hybrid compared to the two parents, we also do not observe big differences in the transcription level.

Table 13. Patterns of transcription of the genes for each organ. Total is the summation of genes detected in the three organs, which, in some cases, overlap and are the same. Vertical higher position indicates more transcription (P, peach; H, hybrid; A, almond).



	P \H \A	P-H \ A	P \ H-A	A /H/ P	H-A / P	A / P-H	P A \ / H	P \ A H/	A P\ / H	H / \ P A	H / \A P	H P/ \ A	P-H-A
Total	3087	2415	1099	3003	2307	702	79	148	202	52	245	225	51002
Leaf	677	619	97	576	553	102	7	42	30	3	46	31	17572
Flower	1764	1179	870	1794	1227	449	60	74	123	48	179	144	14867
Fruit	646	617	132	633	527	151	12	32	49	1	20	50	18563

Within the differentially expressed genes, we encounter an intriguing case of differential expression between almond and peach. This gene called *Prupe.1G446400* exhibits higher expression in peach than in almond across all three organs with higher expression levels in the fruit. This gene encodes a polyketide cyclase / dehydrase, involved in lipid transport (PF1060) and a lachrymatory-factor synthase. It might be associated with fruit aroma and taste, potentially explaining its higher expression in peach fruit compared to almond (Table 14).

Table 14. Transcription levels (DESeq rlog values) of the *Prupe.1G446400* gene in almond and peach for each organ.

	ALMOND		**PEACH**	
	rlog value	**SD**	**rlog value**	**SD**
LEAF	7.504	0.273	8.424	0.208
FLOWER	6.800	0.242	7.967	0.124
FRUIT	10.066	0.456	11.703	0.094

An interesting gene with a polymorphic insertion between peach and almond

One of the genes exhibiting differential expression in leaves in our RNAseq analysis, validated through qRT-PCR (Figure 6), is the gene *Prupe.2G111400*. This gene encodes an ABC transporter. ABC transporters have been associated with pathogen resistances in the literature. We identified *Prupe.2G111400* as one of the genes harboring a polymorphic insertion upstream (less than 1 Kb from the

ATG) in peach. This gene was among the 163 genes displaying differential expression between almond and peach and having a polymorphic insertion present in peach, that are included in the Table 11 of the section titled "Relationship between differential gene expression and polymorphic TEs". This insertion, annotated as Ppersica_LTRretrotransposon1616 (belonging to cluster LTR_1363), is an LTR retrotransposon Copia, measuring approximately 7 Kb and estimated to be inserted around 2.93 MYA.

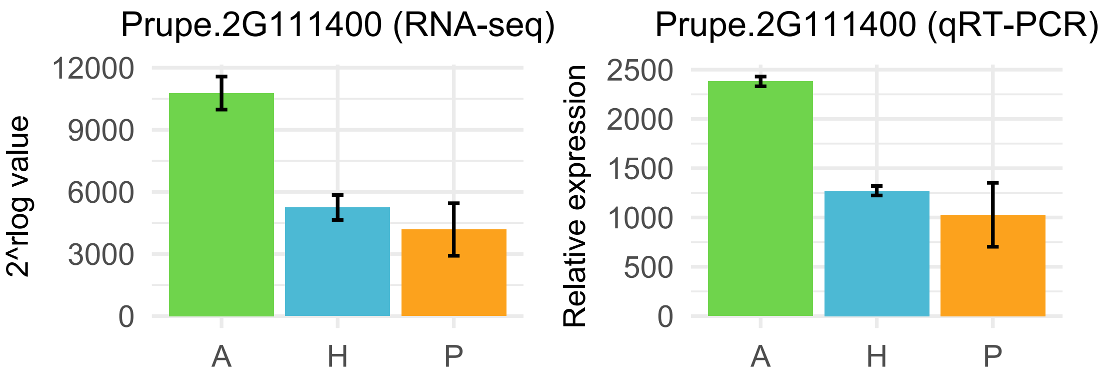


Figure 10. Transcription levels of *Prupe.2G111400* in leaves of almond (A, green), peach (P, orange) and the F1 hybrid (H, blue) analyzed by qRT-PCR and RNA-seq.

Consequently, we validated the presence of this polymorphic insertion using PCR in 'Early Gold', employing primer combinations that enabled us to detect its presence or absence (Figure 11A). We determined that this insertion is homozygous in 'Early Gold' peach, absent in 'Texas' almond as indicated by the gene annotation of the genome of peach and the genome of almond (Verde *et al.*, 2017; Alioto *et al.* 2020), and, as expected, it is also present in the F1 hybrid (Figure 11B).

We examined the annotation of this gene in the 'Texas' almond genome and observed a distinct annotation. The gene *Prupe.2G111400*, along with the gene *Prupe.2G111500*, were annotated as a single gene in the almond genome, named *Prudul26A016647*. The insertion point of the insertion exactly coincided with an intron of the annotated almond gene *Prudul26A016647* (Figure 12). This possibly led to the gene being annotated as two separate genes in the peach genome, despite being a single gene. Furthermore, we observed that in other *Prunus* genomes, such as the *Prunus armeniaca* genome, this gene was annotated in the same manner as in almond, as a single gene. As expected, considering its estimated age of 2.7 Mya, we observed that this insertion was not annotated in the *P. armeniaca* genome, nor in the genomes of other *Prunus* species closer to peach, such as *P. mira* and *P. davidiana*. It was also not annotated in almond 'Lauranne'. However, we did find that the insertion was present in 'Chinese Cling' peach, just like in ‘Lovell’ and ‘Early Gold’ peach.

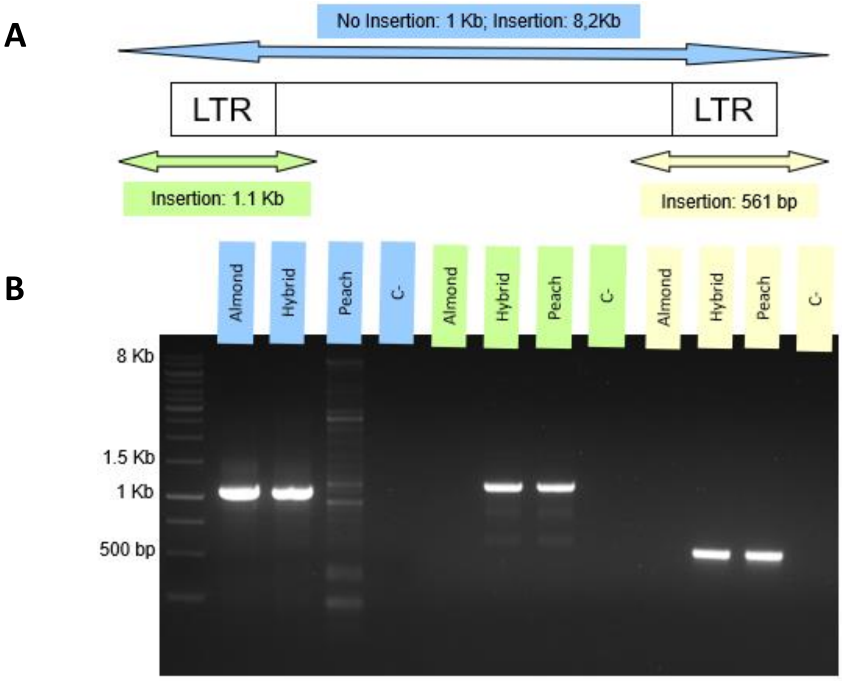


Figure 11. PCR validation of the presence or absence of insertion Ppersica_LTRretrotransposon1616 in almond, peach, and the hybrid between both species. **A)** Scheme of the three primer combinations utilized in the PCRs to amplify the insertion and the expected product sizes. The blue primer combination amplifies the entire LTR retrotransposon or the absence of the LTR retrotransposon. The green primer combination and the yellow primer combination amplify each end of the transposon. **B)** PCR results for each primer combination described in section A. The size of the ladder bands is indicated. C- indicates negative control with water.

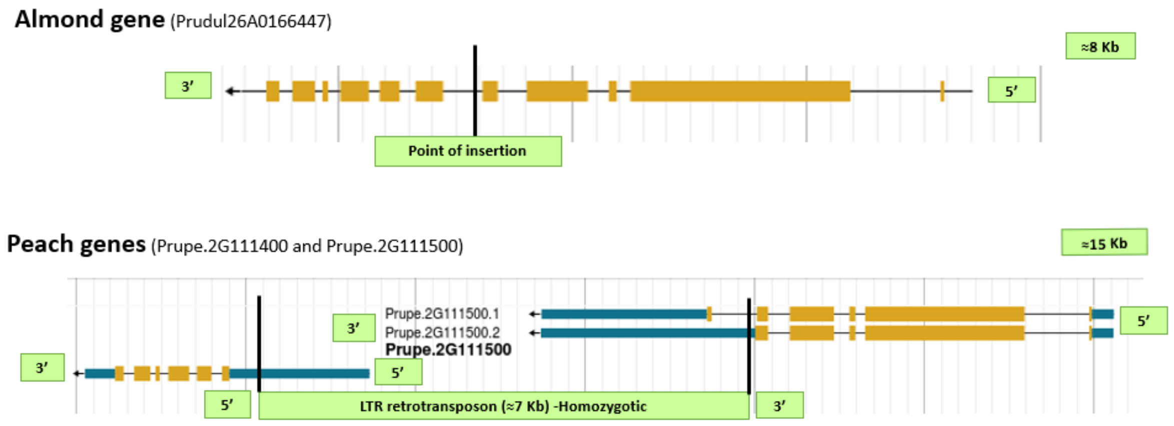

Almond gene (Prudul26A0166447)
≈8 Kb
3'
5'
Point of insertion
Peach genes (Prupe.2G111400 and Prupe.2G111500)
≈15 Kb
Prupe.2G111500.1
Prupe.2G111500.2
Prupe.2G111500
3'
5'
3'
5'
5'
LTR retrotransposon (≈7 Kb) -Homozygotic
3'

Figure 12. Schematic representation of the annotation of the polymorphic gene analyzed in the almond genome (as a single gene named *Prudul26A016647*) and in the peach genome (as two genes named *Prupe.2G111400* and *Prupe.2G111500*).

Considering this alteration, we chose to retrieve the expression data of the almond gene using the RNA-seq analysis on leaves that utilized the almond genome as a reference. In Table 15, I present the transcription data of this gene, employing both the almond and peach genomes as references. As observed, with the accurate annotation, the expression remains distinctly differential, with higher levels in almond than in peach.

Table 15. Transcription level for the analyzed polymorphic gene (DESeq rlog values) in the RNA-seq analyses for leaves using almond genome or peach genome as reference.

RNAseq	GENE NAME	ALMOND	PEACH
Almond genome	*Prudul26A016647*	14,59	13,16
Peach genome	*Prupe.2G111400*	13,39	11,98
	Prupe.2G111500	13,76	12,01

This polymorphic gene, based on the literature, was observed to be situated in the region where the *Vr3* gene is located. In this region, Marimon *et al.* (2020) conducted fine mapping and proposed a region encompassing 27 candidate genes for being the resistance gene against powdery mildew. Among these genes, two were identified as *Prupe.2G111400* and *Prupe.2G111500*. Given the differential expression between peach and almond in these genes and the presence of a LTR retrotransposon within an intron of the gene, we speculated that this insertion could indeed impact transcription, at least at the transcriptional level. Among the two candidate genes with the most substantial evidence to potentially be the *Vr3* gene as proposed by Marimon *et al.* (2020), we observed differential expression in one of them, in the gene *Prupe.2G112800*. This gene exhibited higher expression in peach compared to almond.

Furthermore, we aimed to analyze the transcripts of this gene in both peach and almond. However, we noticed that there were multiple transcripts, making it challenging to draw definitive conclusions. However, we did observe differences between peach and almond. For instance, in the two exons flanking the insertion, we noticed differences. Using a primer combination that amplified those two exons (primers 1616_5R and 1616_6F), I successfully amplified 5 times the sequence in almond but not in peach. Additionally, in our RNA-seq data, we also noted transcripts that lacked certain exons and that transcription levels are not uniform across all exons (Figure 13).

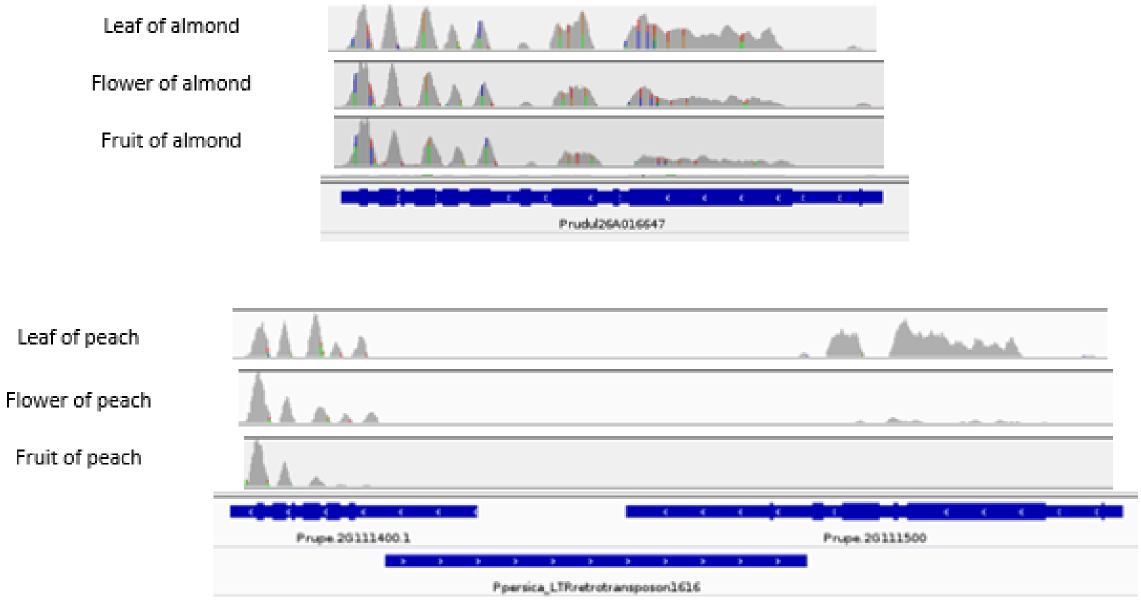
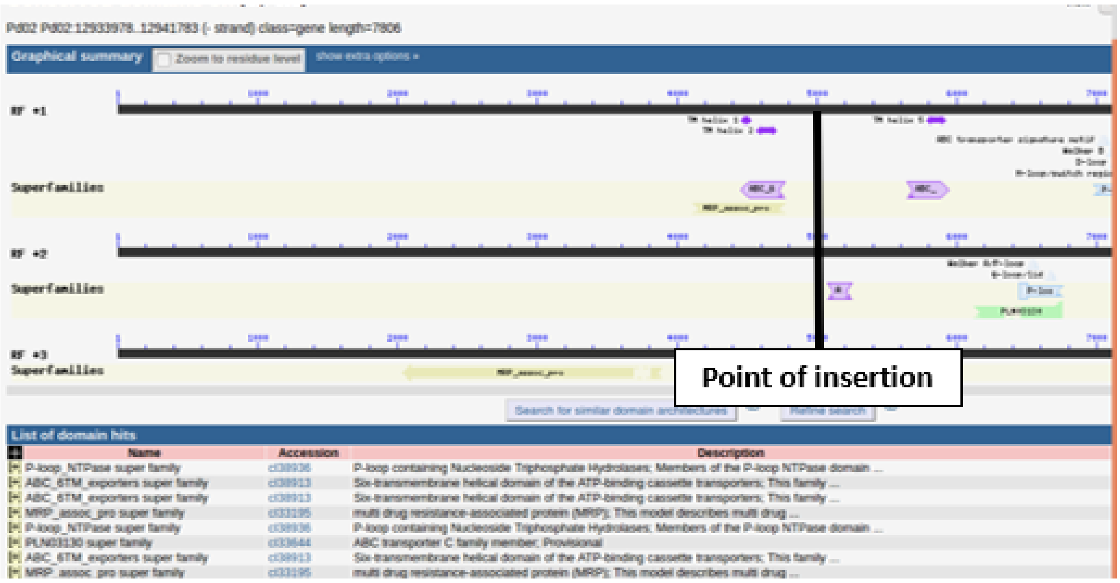


Figure 13. Transcription profiles from our RNA-seq data analysis in leaves, flowers, and fruits for peach and almond, based on their respective annotations of our candidate gene for powdery mildew resistance.

Regarding the function of our candidate gene for the powdery mildew resistance gene *Vr3*, it is annotated as an ABC transporter in both peach (for both genes) and almond genomes. Specifically, it is an ABCC transporter and possesses a structure with nucleotide binding domain (NBD) and transmembrane domain

(TMD) regions. We have observed that the insertion is located within a TMD, which truncates and separates this gene's domain (Figure 14). This could potentially have an impact in its function. ABCC transporters like this have been previously associated with powdery mildew resistance (Krattinger *et al.*, 2009), making it a strong candidate for the *Vr3* gene (Marimon *et al.*, 2020).



Pd02 Pd02:12933978..12941783 (- strand) class=gene length=7806

Name	Accession	Description
P-loop_NTPase super family	cl38936	P-loop containing Nucleoside Triphosphate Hydrolases; Members of the P-loop NTPase domain ...
ABC_6TM_exporters super family	cl38913	Six-transmembrane helical domain of the ATP-binding cassette transporters; This family ...
ABC_6TM_exporters super family	cl38913	Six-transmembrane helical domain of the ATP-binding cassette transporters; This family ...
MRP_assoc_pro super family	cl33195	multi drug resistance-associated protein (MRP); This model describes multi drug ...
P-loop_NTPase super family	cl38936	P-loop containing Nucleoside Triphosphate Hydrolases; Members of the P-loop NTPase domain ...
PLN03130 super family	cl33644	ABC transporter C family member; Provisional
ABC_6TM_exporters super family	cl38913	Six-transmembrane helical domain of the ATP-binding cassette transporters; This family ...
MRP_assoc_pro super family	cl33195	multi drug resistance-associated protein (MRP); This model describes multi drug ...

Figure 14. Structure and domains of our candidate gene for powdery mildew resistance (NCBI Conserved Domain Search).

1.5. Discussion

Hybridization is a very relevant and relatively frequent process in plant evolution (Nieto Feliner *et al.*, 2020), that has also been used in plant breeding. For example, crosses of different varieties to produce hybrids presenting superior phenotypes as compared with their parents (hybrid vigor) is widely used in crops (Rajendrakumar *et al.*, 2015) and interspecific hybridization with crop wild relatives, or highly variable related species, is frequently used to expand the species variability used for breeding (Warschefsky *et al.*, 2014). Peach is a self-fertile and naturally self-pollinating species with very low genetic variability (Micheletti *et al.*, 2015). The use in breeding programs of interspecific crosses with other species of the genus *Prunus* has been historically used to increase peach genetic variability (Foolad *et al.*, 1995). In the last years a growing interest has emerged in the use of these related species for peach breeding, mainly as a source of pathogen resistances (Martínez-Gomez *et al.*, 2004). Almond has become an interesting choice for introgressing new genes into peach, mainly due to the high genetic variability present in almond and also for fruit quality traits (Donoso *et al.*, 2016). In addition to the combination of diverged alleles and different genes, the merging of two different genomes can also be accompanied by epigenetic and structural changes that can be so widespread that have been defined as a *genomic shock* (Comai *et al.*, 2003).

We analyzed the possible genetic and epigenetic changes associated with the crossing of peach and almond to produce an interspecific hybrid. The transcription analysis in leaves was published in a paper authored by de Tomás *et al.* (2022) (see annexes), along with the data from a leaf methylation analysis conducted by Dr. Bardil. Transposons are a primary target of epigenetic mechanisms, and DNA methylation is the main epigenetic modification associated with TE silencing (Deniz *et al.*, 2019). Methylation results show that there are no major differences in the methylation of TEs between the two parental species or between both parents and the hybrid (de Tomás *et al.*, 2022). de Tomás *et al.* (2022) found some differentially methylated regions that overlap with LTR retrotransposons, the main order of TEs in peach and almond (Alioto *et al.*, 2020). In some cases, as for the CHG context, and in particularly when comparing the hybrid with peach, most of the DMRs of the same TE family are demethylated in the hybrid, suggesting a possible weakening of the epigenetic silencing and an increased potential for activation associated with the interspecific cross. However, this demethylated trend has no parallel for the CHH context, where different copies of the same LTR retrotransposon family can show hypermethylated and hypomethylated DMRs in the hybrid. Moreover, the analysis of the transcription of the LTR retrotransposons in leaves in the two parents and in the hybrid did not show the reactivation of any of the LTR retrotransposon families after hybridization (Figure 1). In addition, we neither found transcriptional activation of other types of TEs in leaves of the hybrid (Figure 1). These results suggest that the cross of peach and almond did not result in important changes in the regulation of TEs in general and in the LTR retrotransposons in particular. Among the genes that show a reduced expression in the hybrid there are two genes potentially encoding for an RNA-dependent RNA polymerase showing similarity with the RDR1 protein from Arabidopsis, which is involved in the production of sRNAs, viral defense and DNA methylation (Table 10) (Leibman *et al.*, 2018; Wang *et al.*, 2014). However, a closer look at more genes involved in DNA methylation dynamics (Rothkegel *et al.*, 2021) did not reveal any difference of expression. Only 1% of the genes showed transgressive expression in the hybrid which reinforces the idea that no major genomic changes are induced by the merging of the peach and almond genomes in the hybrid.

When we extended our analysis to other organs with the hypothesis that we might detect more changes, we observed that, in general, transcription patterns in flowers and fruits were similar to those in leaves, indicating the absence of transposon activation in the hybrid (Table 10 and Table 13). Nevertheless, we detected more expressed genes and transposon families in flowers compared to other organs, as has been observed previously. For instance, Vicient (2010) described different profiles of TE transcription in various maize organs and conditions, detecting more expressed transposon families in flowers compared to leaves. The transcriptional activity of various Transposable Elements is particularly elevated in the sperm cells. The fact that more TEs and genes are expressed in flowers could be due to a major cell type variability in this organ compared to others like leaves or fruits.

There are many examples were interspecific crosses result in genome demethylation and/or TE activation (Senerchia *et al.*, 2015, Ungerer *et al.*, 2006). In consequence, the lack of signs of a “genomic shock” in the peach x almond hybrid may seem surprising. However, not always interspecific crosses result in genome demethylation and/or TE activation as it has been shown, for example, in crosses between *Arabidopsis thaliana* and *Arabidopsis lyrata* (Göbel *et al.*, 2018). It has been proposed that when merging two different genomes in a hybrid the intensity of the genome rearrangement and TE mobilization could depend on the TE load imbalance and the phylogenetic distance between the parents (Mhiri *et al.*, 2019). Almond and peach are *Prunus* species of the same subgenera, *Amygdalus*, and have diverged only six million years ago (Alioto *et al.*, 2020). Considering that the mean generation time for these species is 10 years, this explains the low divergence of their genome sequences which is as low as 20 nucleotide substitutions per Kbp (Alioto *et al.*, 2020). In addition, the two genomes are also very similar in the proportion and types of TEs they content, sharing the majority of TE families and many individual TE insertions (Alioto *et al.*, 2020). Therefore, the small phylogenetic distance between peach and almond and their shared TE load could be the reason for the absence of a detectable *genomic shock* associated with their interspecific cross.

In conclusion, my work shows that the merging of peach and almond genomes in an interspecific hybrid has not a major impact on gene expression and is not associated with TE reactivation, which is perfectly compatible with the observed absence of general alterations in DNA methylation levels observed by Dr. Bardil (de Tomás *et al.*, 2022). The absence of alterations in the hybrid may facilitate the use of almond as a source of new genetic variability for breeding the low variable peach species.

When we examined the relationship between differential gene expression and polymorphic transposable elements (TEs), we identified 190 genes that displayed differential expression and had a polymorphic insertion between peach and almond, located at less of 1 kb upstream of the ATG. Despite this, the majority of these polymorphic insertions were present in peach (163), while in almond, there were only 27 (Table 11). This count aligns with the total number of genes that had a nearby polymorphic insertion, which was significantly higher in peach compared to almond. This difference could imply the challenge of detecting insertions in a species like almond, known for its high heterozygosity (see Chapter 2). Among these genes with polymorphic insertions and differential expression, we noted that 93 exhibited higher expression when the insertion was absent and 70 when it was present (Figure 7). The presence of a TE insertion could interrupt and inactivate the promoter of the gene but, in some cases, transposons can provide novel promoter elements and induce changes in the expression of the close genes (Hirsch & Springer, 2017).

In this chapter, we also identified a polymorphic gene between peach and almond, characterized by an LTR retrotransposon insertion present in peach but absent in almond. This gene exhibited differential expression between the two species, with almond displaying higher expression levels. Remarkably, this gene was located within the genomic region where the powdery mildew resistance gene *Vr3* had been mapped (Marimon *et al.*, 2020). Consequently, based in my RNA-seq and RT-PCR analyses, *Prudul26A016647* could be a promising candidate for improving peach resistance to fungal diseases (Figure 10).

Functionally, this gene encodes ABCC transporter and, in peach, the LTR retrotransposon is inserted within their DNA sequences which encode its transmembrane domains (TMD), suggesting a potential impact on its functionality. ABCC transporters have been described in the literature to play a role in pathogen resistance. For instance, Underwood & Somerville (2017) describe how the *Arabidopsis* PEN3 AVC transporter accumulates at sites of pathogen detection and participates in defense against various pathogens, including powdery mildew. Similarly, Krattinger *et al.* (2009) describes how a putative ABC transporter contributes to durable resistance against multiple fungal pathogens in wheat, including powdery mildew. Antimicrobial plant secondary metabolites constitute a crucial defense mechanism against both host and non-host pathogens, with increasing evidence indicating that ABC transporters play a role in their secretion (Kant *et al.*, 2011).

Considering these factors, we regard this gene as a strong candidate, alongside the two genes proposed by Marimon *et al.* (2020) with big evidence. These genes can be studied in the future, possibly through genetic transformation methods, despite the challenge posed by the recalcitrant nature of peach (Zong *et al.*, 2019).

CHAPTER 2:

STUDY OF TRANSPOSABLE ELEMENTS OF A NEW PHASED VERSION OF ALMOND GENOME ‘TEXAS’

2.1. Introduction

Almond [*Prunus dulcis* (Mill.) D.A. Webb (syn. *Prunus amygdalus* Batsch., *Amygdalus communis* L.)] is a fruit-bearing tree belonging to the genus *Prunus* of the family *Rosaceae*, along with other species as peach [*Prunus persica* (L.) Batsch], apricots (*Prunus armeniaca* L.), sweet cherries [*Prunus avium* (L.) L.], japanese plums (*Prunus japonica* Thunb.) and european plums (*Prunus domestica* L.).

Almonds have a wide range of applications, primarily as human food (as nut) due to their high nutritional and culinary quality. They are an important source of macronutrients and phytonutrients, such as vitamin E, folate and oleic acid (Gradziel, 2020). Additionally, almonds have been used as food additives for flavoring (Facciola, 1990), serve as bee plant for honey production (Ortega-Sada, 1999), and provide materials in the form of lipids, such as almond oil for pharmaceutical purposes (Markle, 1998). Furthermore, almond have found diverse application in medicine (McGuffin *et al.*, 2000) and have even been used as vertebrate poisons due to their bitterness (Cooper & Johnson, 1998).

This species is native to central Asia but it is one of the oldest domesticated tree species, dating back around 5,000 years ago (Zohary *et al.*, 2012). It is now extensively cultivated in Mediterranean climate regions around the world (FAO, 2022), including California, which accounts more than half of the global almond production, followed by Spain and Australia.

As mentioned in the General Introduction section, almond is a diploid species with 8 chromosomes and a compact genome of 300 Mbps. It presents a high level of heterozygosity, approximately seven times more variable than other *Prunus* species such as peach. This high heterozygosity is mainly due to the fact that most almond varieties are self-incompatible, which contributes to a higher level of genetic variability (Velasco *et al.*, 2016).

Currently, almond has genomes available for three varieties: ‘Lauranne’ (Sánchez-Pérez *et al.*, 2019), ‘Texas’ (Alioto *et al.*, 2020) and ‘Nonpareil’ (D’Amico-Willman *et al.*, 2022) (Table 1 of General Introduction). None of these genomes are sequenced in phases, meaning they lack the separate reconstruction of sequences corresponding to the two copies of each chromosome. This is a disadvantage when working with the highly heterozygous genome. Phased genomes are very useful for studying structural variations (SV) between haplotype alleles, including chromosomal rearrangements and single-nucleotide polymorphisms (SNPs), as well as gene variations like presence-absence variations (PAVs), allele-specific expression (ASE) and dominant-recessive alleles that may be associated with traits with of agronomic interest (Guk *et al.*, 2022).

The absence of phased genomes is not unique to almonds genomes. It is the same for all the *Prunus* species, and for most of the available genomes. As explained by Duitama (2023), recent genome assemblies are a combination of the two underlying chromosomes copies present in the sequenced individual due to the challenges in creating phased genome assemblies.

Nevertheless, thanks to sequencing technologies that provide long and accurate reads, including Single-molecule real-time sequencing (SMRT), and Oxford Nanopore Technologies (ONT), with chromosome-scale mate-pair reads such as high-throughput chromatin conformation capture (Hi-C), phased genomes have become feasible (Guk *et al.*, 2022).

Haplotype-resolved genomes have been more frequently accomplished in humans (Cao *et al.*, 2015) and animals such as domestic cat and Asian leopard cat (Bredemeyer *et al.*, 2021). In contrast, their use is not common in plants due to their highly non-inbred nature and complex genomic structures (Guk *et al.*, 2022). However, the recent advances of the sequencing technologies mentioned before have led to some haplotype-resolved genomes assemblies for plants such the rosaceous specie apple ‘Gala’ (Sun *et al.*, 2020), lemon (Di Guardo *et al.*, 2021), patchouli (Shen *et al.*, 2022), vanilla (Piet *et al.*, 2022) and kiwifruit (Han *et al.*, 2023).

As mentioned earlier, alelle-specific expression (ASE) can be measured in phased genomes. ASE quantifies the relative expression of the two alleles in a diploid individual. This expression imbalance could potentially play a role in generating differences in traits and the development of diseases among individuals (Fan *et al.*, 2020).

Additionally, there are now available DNAseq short reads data of almond in public databases that can facilitate the study of the diversity and variability of this species. For example, the detection of TE insertion Polymorphisms (TIPs) can now be performed using more than 80 tools, as described in TE Hub Consortium *et al.* (2021). Some of these tools have been compared through benchmarking in TE insertion detection (Vendrell-Mir *et al.*, 2019).

In this second chapter, a new and phased assembly of genome of the almond ‘Texas’ (syn. ‘Texas Prolific’ and ‘Mission’) called “Texas v.3.0” is presented and compared with the old version “Texas v.2.0” (also known as pdulcis26) (Alioto *et al.*, 2020). ‘Texas’ is one of the oldest US cultivars, obtained in Huston, Texas, USA, as a seedling of French cultivar ‘Languedoc’ (Kester *et al.*, 1991), and along with ‘Nonpareil’, ‘Tuono’ and ‘Cristomorto’ one of the four major contributors to modern almond breeding worldwide (Pérez de los Cobos *et al.*, 2021).

Specifically, we will present the assembly of the Phase 0 (P0) and Phase 1 (P1), and the gene and Transposable Element annotation. We will deepen in the powdery mildew resistance region described in the Chapter 1, that is improved in the new assembly. Furthermore, benefiting of this new phased assembly, we will present the impact of homozygous and heterozygous Transposable Elements on gene expression of different organs (leaves, flowers and fruits) and the diverse patterns of allele-specific expression (ASE) during almond development. Concurrently, we will present the genetic variability of 40 almond public accessions of DNAseq short reads to obtain an evolutionary perspective.

This work is a collaboration between some members of my research group (Dr. Raúl Castanera, Dr. Josep M. Casacuberta and my PhD supervisor Dr. Carlos M. Vicient) and the IRTA *Prunus* Group (Dr. Iban Eduardo, Dr. M. José Aranzana and Dr. Pere Arús).

The genome assembly and gene annotation were undertaken by Dr. Valentino Ruggieri, as external. The Transposable Element analysis is the product of a collaborative work between Dr. Raúl Castanera and myself. I will especially present my results in this chapter, but in some sections, I will present the results obtained by Dr. Castanera for the good understanding. In particular, annotation was conducted by Dr. Castanera and myself. The detection of completed copies and the obtention of structural and genetic variation data between the two phases was performed by Dr. Castanera. This data was used in my subsequent analysis. The sample recollection and the RNA extractions for the RNA seq analysis were carried out by myself. For the RNA seq analysis, I executed the read filtering and read mapping against the genomes. The quantification of the mapped reads for the general gene expression was performed by myself and the quantification of the mapped reads for each allele of the genes was performed by Dr. Castanera. Finally, I performed the comparative analysis between the "Texas v.2.0" and "Texas v3.0" assemblies of the powdery mildew resistance region described in Marimon *et al.* (2020), the examination of the impact of homozygous and heterozygous TEs on gene expression, the ASE analysis and the analysis of genetic variability in almond cultivars.

The results of this Chapter 2 will be included in a paper currently in preparation.

2.2. Objectives

The objectives of this study are as follows:

- Annotate Transposable Elements in the new version of the phased genome of almond.

- Compare “Texas v.2.0” and “Texas v3.0” assemblies within an interesting agronomic region, associated with a powdery mildew resistance locus.

- Study the relationship of the Transposable Elements on the gene expression of almond through the study of the homozygous and heterozygous insertions near to genes.

- Analyze the allelic-specific expression (ASE) during the almond development.

- Analyze the cultivar variability of 40 almond accessions.

2.3. Material and methods

Genome assembly

The *P. dulcis* 'Texas' genome assembly derives from PacBio long reads and Hi-C Illumina short reads release, including both genome phases (P0 and P1).

Hapo-G was used to polish contigs removing sequencing errors (Aury & Istace, 2021), and Falcon-Phase to reconstruct the complete phases corresponding to each of the two haplotypes (Kronenberg *et al.*, 2021). Hi-C data was further used to produce scaffolds, which were mapped to linkage groups the genetic map of 'Texas' almond x 'Early gold' peach F2 progeny (Donoso *et al.*, 2015), resulting in 92% of the genome assembly (232.5 Mbp) successfully assigned to the 8 linkage groups. A fraction of the unplaced contigs were subsequently integrated in the pseudomolecules based on the synteny with the previous 'Texas' assembly (Alioto *et al.*, 2020), resulting in 98% of the genome anchored to chromosomes. Finally, we obtained a phased genome assembly (Texas v.3.0) that spanned 254.02 Mb for Phase 0 (P0) and 252.65 Mb for Phase 1 (P1).

Gene annotation

Gene annotation was performed in phase 1 using a custom pipeline based on MAKER2, combining transcriptome-based, protein-based, and *ab-initio* based gene prediction (Holt & Yandell, 2011).

RNA-Seq datasets were retrieved from public collections corresponding to almond buds, fruits and roots (SRR11251343, SRR11251344, SRR11251345, SRR10189207, SRR10189208, SRR10189209, SRR6815287, SRR6815288, SRR6815289) and our collections spanning different organs (that I presented in the Chapter 1 and I will present again in the RNAseq section of this chapter). Reads were filtered and trimmed by Trimmomatic v0.39 (Bolger *et al.*, 2014) to remove low quality reads portions.

The haplotype Texas P0 was annotated by transposing the previously predicated gene models of haplotype Texas P1 on the genome assembly of Texas P0. This strategy allowed to preserve the synteny and correspondence between the genes of the two phases. Indeed, more than 90% of the gene models passed through.

However, a minority of gene models, representing less than 10%, failed to be correctly transferred (due to differences in haplotype sequences, repetitive regions, complex genomic regions that caused unmapped genes or errors in ORF or CDS structure validity). These genes were re-mapped using the est2genome tool build within MAKER2 (the approach uses blastn and exonerate to correctly spice transcripts on the assembly).

To assign functional description, GO terms and KEGG pathway information to the new gene models, sequences (transcripts/proteins) were functionally annotated using TRINOTATE v.2 (Bryant *et al.*, 2017).

Transposable Element annotation

Extensive de-novo TE Annotator (EDTA) pipeline was run independently on each Texas v.3.0 phase to obtain individual TE libraries and genome coordinates of LTR-retrotransposons (based on LTR-retriever) (Ou *et al.*, 2019). Redundancy among the two libraries was eliminated by running CD-HIT at 80% identity cut-off. Unclassified consensuses and/or sequences with length < 200 bp were filtered out, and the resulting library was complemented with LINE coding REPET consensuses from the previous almond TE annotation (Alioto *et al.*, 2020) to compensate EDTA low sensitivity on the detection of this TE order.

A first round of RepeatMasker was run using this preliminary TE library (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-4.0. 2013-2015 http://www.repeatmasker.org). LTR, LINE and TIR consensuses that did not have a full-length RepeatMasker match (> 80% of TE length) were removed. For MITEs and other non-coding consensuses elements, we only retained consensuses with three or more complete matches in the genome. Helitron consensuses without coding domains were filtered out too. Finally, the library was complemented with peach-specific REPET coding consensuses (< 80% identity based on CD-HIT clustering) (de Tomás *et al.*, 2022). A second round of RepeatMasker was performed with the curated TE library and integrated with the EDTA structural annotation of intact LTR-retrotransposons. Specifically, LTR-retrotransposon RepeatMasker matches overlapping EDTA intact elements were removed from the final annotation.

Identification of complete TE copies

We relied on elements retrieved from EDTA phase-1 as starting material to detect potential complete elements (*EDTA.intact.gff3 file). For LTR retrotransposons we kept all EDTA elements due to the very low false negative ratio of LTR-retriever (elements carrying LTR, TSD and coding domains). For TIR and Helitrons, we extracted the sequence of all putative intact copies and re-classified them with TEsorter (Zhang *et al.*, 2022). We kept elements where EDTA and TEsorter classification matched at the order level. MITE elements were extracted and clustered with CD-HIT at 80% identity. We kept only elements present in clusters of three or more copies.

Identification of structural variants

We used Minimap2 to align Texas P0 and P1 assemblies (parameters: -ax asm5) (Li, 2018), and SVIM-asm (Heller & Vingron, 2021) was used to detect structural variants (default parameters). Heterozygous TEs were detected by performing a reciprocal intersection of TE annotations with the deletions found in P0 or P1 using BEDTools (Quinlan & Hall, 2010). A TE was considered heterozygous if it spanned more than 50 % of an overlapping deletion and the deletion covered at least 80% of the TE (bedtools parameters -F 0.5, -f 0.8). A TE was considered homozygous if it was completely covered by the genome alignment.

Comparative analysis of the powdery mildew resistance locus

The powdery mildew resistance locus, known as the *Vr3* gene, was successfully mapped to a genomic region situated between the markers Indel16912 and SNP_17184692. This mapping was accomplished through the analysis of interspecific populations derived from almond ('Texas') and peach ('Early Gold') crosses (Marimon *et al.*, 2020).

The physical locations of these markers were identified within the genomic sequence of the 'Lovell' peach (Verde *et al.*, 2017). The markers were also localized within both the old version (Texas v.2.0) (Alioto *et al.*, 2020) and the new phased version (Texas v.3.0) of the 'Texas' almond genome using Blastn.

Subsequently, the specific genomic regions encompassing these markers were extracted utilizing BEDtools (Quinlan and Hall, 2010).

In order to facilitate comparative analysis of the assembly of this region among the distinct genomes (‘Lovell’ and the both phases of ‘P0’), these extracted regions were aligned using D-GENIES (Cabanettes & Klopp, 2018). Additionally, Liftoff software was used to align gene annotations from the reference genome (Texas v.3.0) to a target genome (Texas v.2.0) and vice versa, facilitating our comparison of gene annotations in this region (Shumate & Salzberg, 2021).

Plant material for RNAseq samples and growth conditions

Leaves, flowers, and fruits from the *Prunus dulcis* 'Texas' tree were collected from the Experimental Station of Lleida located in Gimenells (Catalonia, Spain), kindly provided by the Institute of Agrifood Research (IRTA). The ‘Texas’ almond tree was cultivated in the field and regularly watered at the same time of the day. These samples are the same that we presented in the Chapter 1.

Fully expanded leaves were collected at the end of September, flowers in the pink stage were collected on February and immature fruits with approximately 2cm in diameter, were collected during the initial week of May. For each type of sample, a composite pool of 7 leaves, 10 flowers and 4 fruits was generated. The samples of each organ were harvested from three replicates of separate branches of the same tree.

In order to maintain sample integrity during transportation, dry ice and nitrogen liquid were employed. Finally, the sampled were stored in a -80º freezer.

RNA isolation

0.15 grams of each sample were ground using liquid nitrogen and a mortar, until it was transformed into a fine powder. Total RNA was extracted using the Maxwell RSC Plant RNA Kit and the Maxwell RSC instrument (Promega Corporation, Madison, WI, USA). Complete DNA removal was obtained using the DNA-free DNA Removal Kit (Invitrogen™, Carlsbad, CA, USA).

To assess the quality and purity of the RNA samples, the density ratios (260/280 and 260/230) were evaluated using NANODROP ND-1000 spectrophotometer (Thermo Fischer Scientific). The ranges of these density ratios are approximately 1.8 and 2, indicating desirable levels of RNA purity. The RNA Integrity Number (RIN) was calculated using Agilent 2100 Bioanalyzer. The determined RIN value for each RNA sample should be higher than 7.

RNAseq analysis

The RNAseq reads were filtered using BBDuk (Bushnell, 2014) with the next parameters: ktrim=r, k=23, hdist=1, tpe ftr=139 and trimq=10. Their quality was checked using FastQC (Andrews, 2010). To analyze gene transcription, RNA-seq reads were aligned to the P0 and P1 sequences using HISAT2 (Langmead & Salzberg, 2015) and manipulated with SAMtools to produce bamfiles (Danecek *et al.*, 2021).

To obtain the count of mapped reads on genes, FeatureCounts was used (Liao *et al.*, 2014). The normalization of the general gene expression data was performed using DESeq2, specifically we used DESeq2 regularized log (rlog) values (Love *et al.*, 2014). A comparative transcriptome analysis was conducted among genes lacking nearby insertions, genes with heterozygous insertions nearby, and genes with homozygous insertions nearby in P0. The same comparative transcriptome analysis was repeated in P1. Nearby indicates that the insertion is located within 1 Kb upstream.

For the allele-specific expression (ASE), SNPs presented in gene coding regions without surrounding INDELS (at <50bpS) that could allow us to differentiate the expression of the two alleles (reference and alternative alleles) were searched. Next, to obtain the allele expression count of these genes with two differentiated alleles, ASEReadCounter was utilized (Castel *et al.*, 2015). The normalization of the expression data was performed using DESeq2, specifically we used DESeq2 regularized log (rlog) values. Allelic expression analysis (differential expression analysis between the reference and the alternative alleles) was conducted using DESeq2 with a Log fold-change cut-off of one (Love *et al.*, 2014). False Discovery

Rate (FDR) was employed for multiple-testing correction. In the heatmap of the allelic expression, we present the percentage of mapped reads from each allele.

Analysis of the cultivar variability

The analysis of almond cultivar variability was conducted on a total of 40 almond cultivars (Table 1). These selected accessions were re-sequenced using Paired-End Illumina sequencing. The assortment of chosen varieties collectively represents a diverse range of major cultivation regions across Europe (Spain, Italy and France), the USA, and also includes representative varieties from China.

To initiate the data processing pipeline, the SRA for these samples were initially acquired through the utilization of SRA Toolkit (Leinonen *et al.*, 2011). Subsequently, the data were transformed into fastq format using fasterq-dump tool. Following this, a data quality control step was executed using BBDuk (Bushnell, 2014) with specific parameters, which included a minimum length of 25 bps. In order to achieve uniform coverage across the distinct varieties, a subsampling approach was implemented. This process was executed to attain a maximum coverage of 40X for all the selected cultivars and was accomplished using Seqtk.

Following the preprocessing steps, the resulting reads for each individual cultivar were aligned to the almond reference genome (P0) using the BWA Aligner (Li, 2013). Finally, to investigate the presence of the Transposable Elements across the various almond cultivars, the TE annotation previously established for this 'Texas v.3.0.' were searched in the varieties. This was executed using PoPoolation TE2 (Kofler *et al.*, 2016) with the mode 'joint' and the TE annotation of the Phase 0 of 'Texas v.3.0.' as reference. TIPs with a zygosity lower than 0.25 in all the samples and TIPs with Non data in more than 20% of the samples were excluded to avoid false positives. The specific parameters are the same that are used in Castanera *et al.*, 2023: --min-count 5, -max-otherte-count 2, -max-structvar-count 2. The TIP matrix was transformed to binary form (0,1) using zygosity cutoff of 0.05 to consider the presence of the insertion. It enables the identification and characterization of Transposon Insertion Polymorphisms (TIPs) within each cultivar. Specifically, homozygous and heterozygous insertions were also analyzed, allowing us to determine whether they were fixed in the population.

Table 1. The different cultivars used in the population study (accession run, variety and precedence).

RUN	VARIETY	COUNTRY	CONTINENT
ERR3366583	Ai-2	France	Europe
ERR3366585	Belle dAurons-2	France	Europe
ERR3366586	Cristomorto	Italy	Europe
ERR3366588	Desmayo largueta-2	Spain	Europe
ERR3366589	Falsa Barese	Italy	Europe
ERR3366590	Genco	Italy	Europe
ERR3366591	Marcona	Spain	Europe
ERR3366592	Non-pareil	USA	North America
ERR3366593	Ripon	USA	North America
ERR3366594	Vivot	Spain	Europe
ERR4093803	Texas	USA	North America
ERR4762264	Del Cid	Spain	Europe
SRR3141032	Shuang Guo	China	Asia
SRR3141040	Zhi pi	China	Asia
SRR3141049	Gong Ba Dan	China	Asia
SRR3141057	Wan Feng	China	Asia
SRR3141065	Ai Feng	China	Asia
SRR3141073	Ba Dan Wang	China	Asia
SRR3141083	Huang Shuang	China	Asia
SRR3141098	A Yue Hun Zi	China	Asia
SRR3141113	Tao Ba Dan	China	Asia
SRR3141181	Da Ba Dan	China	Asia
SRR3141192	Ye Er Qiang	China	Asia
SRR3141204	Bian Zui He	China	Asia
SRR3141229	Ao 2 #	USA	North America
SRR4036105	#53	USA	North America
SRR4036108	Tardy Nonpareil	USA	North America
SRR4045222	DPRU 1207.2	USA	North America
SRR4045223	Languedoc	USA	North America
SRR4045224	DPRU 2331.9	USA	North America

SRR4045225	BE-1609	USA	North America
SRR4045226	Tuono	USA	North America
SRR4045227	DPRU 2374.12	USA	North America
SRR4045228	Badam	USA	North America
SRR4045229	DPRU 1462.2	USA	North America
SRR7010336	Lauranne	France	Europe
SRR7010337	Alnem1	France	Europe
SRR765679	Ramillete	Spain	Europe
SRR765850	D05-187	Spain	Europe
SRR765861	S3067	Spain	Europe

2.4. Results

Assembly of phased genome Texas v.3.0 and gene annotation

The new version of the phased genome of ‘Texas’ (Texas v.3.0) was assembled using PacBio long reads and Hi-C Illumina short reads. A total of 250 Mb in 362 primary contigs and 128 Mb in 1,345 haplotigs were assembled. The phased genome assembly spanned 254.02 Mb for Phase 0 (P0) and 252.65 Mb for Phase 1 (P1). Each phase contained 80 scaffolds and 99 contigs, and 97.3 % of the total sequence was anchored to eight pseudomolecules (Table 2). The new pseudomolecules showed overall high synteny with the already published Texas v.2.0 assembly (Alioto *et al.*, 2020), although with some interruptions with inversions, translocations, insertions and deletions.

Table 2. Genome assembly for each phase: length, number of scaffolds/contigs and gaps for each chr.

CHR.	P0 - LENGTH	P1 - LENGTH	CONTIGS/SCAFFOLDS	TOTAL GAPS
Chr0	6,870,097	6,646,489	20/1	22
Chr01	50,385,530	50,520,004	13/13	60
Chr02	30,936,755	31,007,932	4/10	39
Chr03	30,529,542	30,467,225	14/9	46
Chr04	28,248,075	27,889,346	12/10	38
Chr05	22,392,796	22,089,746	11/4	27
Chr06	32,392,687	32,350,728	7/11	46
Chr07	24,805,092	24,914,025	11/10	38
Chr08	27,457,878	26,760,049	7/12	37
Total	254,018,452	252,645,544	80/99	353

In comparison to the Texas v.2.0 reference genome, which is a collapsed representation of the two haplotypes, the new assembly, Texas v.3.0, contains up to 13.2% more contig sequence (30Mb more for P0 and 29.93 Mb for P1). This increase in assembled sequence is homogeneously distributed among the 8 chromosomes and is accompanied by a concomitant reduction of unplaced contig sequence. The sequence contiguity is strongly improved, with an average of 11.5x higher contig N50, with respect to Texas v.2.0, and this improvement correlates with a strong increase of the LTR Assembly Index (LAI) score (Ou *et al.*, 2018), a common indicator of assembly continuity, that is over 20 for both phases (Table 3), a figure that corresponds to the category of “gold quality genome” as proposed by the developers (Ou *et al.*, 2018).

A search for the 166 bp centromeric repeat previously described for *Prunus* species (including almond) (Melters *et al.*, 2013), shows that the number of copies of this repeat is increased by 10.9-fold in Texas v.3.0 with respect to Texas v.2.0, indicating a much better assembly of centromeric regions. This repeat sequence localizes in sharp single peaks in five out of the eight chromosomes, which potentially correspond to the centromeres. Additionally, the reference motif of telomeric sequences in *Arabidopsis thaliana* (TTTAGGG) was searched (Richards & Ausubel, 1988). However, they were not identified, indicating that the genome Texas v.3.0 is not telomere-to-telomere, as the previous version Texas v.2.0.

Table 3. Comparison between the genome assembly and annotation statistics of both phases of Texas v.3.0 genome and Texas v.2.0 genome.

FEATURE	**PHASE 0**	**PHASE 1**	**V.2.0**
Assembly length (Mb)	254.02	252.65	227.59
Pseudomolecule N50 (Mb)	30.53	30.47	24.8
Contig	362	362	4,395
Contig L50	62	61	511
Contig N50 (Mb)	1.21	1.19	0.104
Max. Contig length (Mb)	7.01	7.01	1.31
Percent anchored to pseudomolecules	98	98	91.47
Gap (%)	0.01	0.01	1.72
LAI index	20.58	20.92	8.15
BUSCO complete genes (%)	96.9	97.7	95.4
BUSCO fragmented genes (%)	1.4	0.9	1.0
BUSCO missing genes (%)	1.7	1.4	3.6
Number of protein-coding genes	28,625	29,616	27,969
Genes with Pfam domain *	22,892 (79%)	23,413 (79%)	21,582 (77%)
Gene density (genes/Mb)	113	117	123
Mean CDS length	1,153	1,122	1,244
Mean exons per transcript	5.3	5.3	5.4

* e-value < 0.05 | FDR < 5%.

The results of BUSCO (Benchmarking Universal Single-Copy Orthologs, Simão *et al.*, 2015) evidenced an increased completeness at the gene level in comparison to Texas v.2.0, with 96.9 to 97.7 % of BUSCO complete genes in P0 and P1, respectively (95.4 % in Texas v2.0), and less than 2% of BUSCO missing genes. This improvement was reflected in an increased number of annotated protein-coding genes in comparison to Texas v2.0 (Table 3).

Transcriptomic data from multiple almond developmental stages (root, leaf, flower bud, flower and fruit) was used to annotate 29,616 protein-coding genes and 534 tRNAs on Phase-1 (96.7 % successfully lifted to Phase-0). To identify genes specific to this new gene annotation, we used Liftoff (Shumate & Salzberg, 2021) to map the annotation of Texas v.3.0 to Texas v.2.0. Texas v.3.0 contains 2,518 additional genes as compared with the gene annotation of Texas v.2.0, 79.7 % of them harboring a PFAM conserved domain. The most abundant functions of the proteins putatively encoded by these genes were ubiquitin-like proteases (257), FAR1-related proteins (61), disease resistance proteins (59) and putative transcription factors (13). When mapping the Texas v.2.0 annotation to the new assembly, we identified 926 genes that failed to be lifted (77,3% carrying PFAM domain). The most abundant functions among those genes are protein kinases (76) and Leucine-rich repeat domain containing proteins (67).

Transposable Elements annotation

Using the EDTA pipeline followed by additional filtering steps to eliminate false positives, and complemented with the transposable element (TE) library of Texas v.2.0 (Alioto *et al.*, 2020), we identified 1,022 non-redundant TE consensus sequences, which represent putative TE families. This set of consensuses was used to annotate TEs in the genome by homology search using RepeatMasker, and the results were integrated with a filtered EDTA-based structural annotation of complete elements to produce the final TE annotation. As expected, and as previously described for Texas v.2.0 (Alioto *et al.,* 2020), TEs and genes show a complementary distribution (Figure 1), with TEs concentrating in low gene-density regions such as the regions surrounding the putative centromeres.

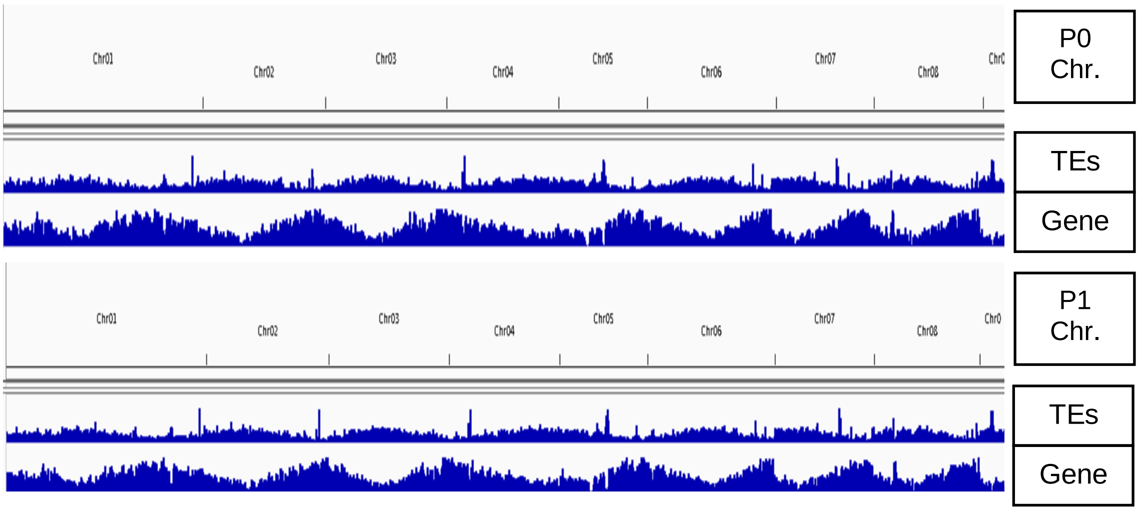


Figure 1. Distribution of Transposable Elements (TEs) and Genes across the different chromosomes (chr.) of the two phases of the version ‘Texas v.3.0’ of almond genome.

TE sequences span 32.9% of the assembly in both P0 and P1 and Texas v3.0 contains an additional 17Mb of TE annotated sequence as compared with Texas v.2.0 (Figure 2A).

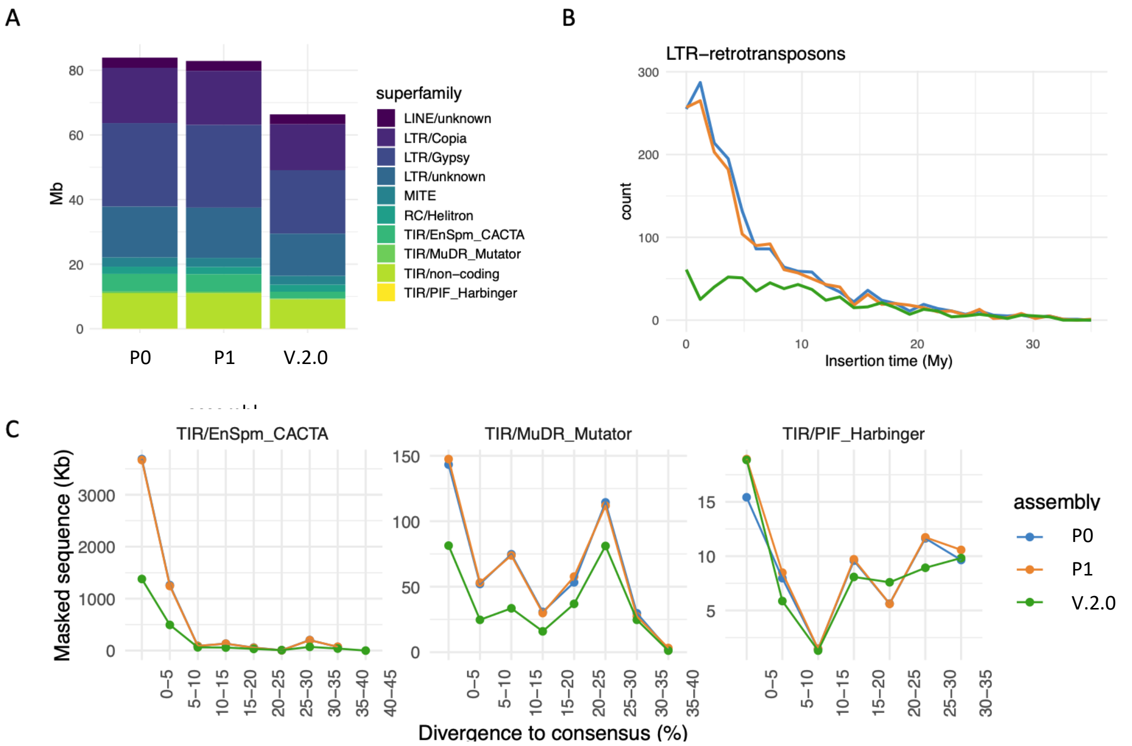


Figure 2. A) TE content in P0 and P1 of Texas v.3.0 and Texas v.2.0 genome assemblies (Mb of sequence). B) Divergence between LTR of intact LTR-retrotransposons. C) Divergence of all TIR TE copies vs their respective consensus sequence. (Figure made by Dr. Raúl Castanera).

The Texas v.3.0 TE annotation contains about twice the number of complete elements as compared with Texas v.2.0 (Table 4). This increase is particularly important for Gypsy LTR-RTs as their number in Texas v3.0 is three-fold that of Texas v2.0. This could be due to the already mentioned improved assembly of the regions putatively containing the centromeres, as Gypsy LTR-RTs tend to concentrate in these regions of plant genomes (Alioto *et al.*, 2020). In order to investigate this, we aligned the sequence of the two phases (P0 and P1) of Texas v.3.0 to Texas v.2.0 and asked whether the regions flanking the Gypsy LTR-RTs of Texas v.3.0 were present in Texas v.2.0. In 51.6% of the cases the flanks are absent from the Texas v.2.0 assembly, indicating that these regions, and the Gypsy LTR-RTs sitting in these regions, were not included in the old assembly. In addition, the new assembly also contains many more complete Copia LTR-RT elements and LTR-RTs in general, as well as more complete TIR transposons (Table 4). An analysis of the age of LTR-RTs inferred from the intra-element LTR comparison, showed that an important fraction of the LTR-RTs newly annotated in this assembly are young LTR-retrotransposons insertions (< 5 My) (Figure 2B).

Table 4. Transposable element content in *P. dulcis* 'Texas'.

	PERCENTAGE OF GENOME SIZE (%)			**NUMBER OF COMPLETE ELEMENTS ***		
Order / Superfamily	**P0**	**P1**	**Texas v.2.0**	**P0**	**P1**	**Texas v.2.0**
TIR	6.7	6.7	5.0	355	348	240
MITE	1.1	1.1	1.2	620	624	511
HELITRON	0.9	0.9	0.9	9	9	10
LTR/Gypsy	10.2	10.1	8.6	519	479	126
LTR/Copia	6.7	6.6	6.2	790	743	340
LTR/Unknown	6.2	6.2	5.7	468	464	236
LINE	1.2	1.2	1.3	NA	NA	NA
Total	33.0	32.8	29.2	2761	2667	1463

*Containing structurally intact features.

Within the TIR order, the most important differences between the two assemblies were found in the EnSmp/CACTA and MuDR superfamilies. In particular, the new assembly contained a 2.5-fold increase in EnSmp/CACTA sequence over Texas v.2.0 (5.4 Mb vs 2.2 Mb) (Figure 2A). An analysis of the divergence of every TIR copy versus its respective TE consensus sequence, which can be used as an indication of the element's age, revealed that an important fraction of the EnSmp/CACTA elements newly annotated in this new assembly are young elements, similarly to what we found for LTR-RTs (Figure 2C).

The genome Texas v.3.0 improves the assembly and annotation of the powdery mildew resistance gene region

The powdery mildew resistance locus, known as the *Vr3* gene, was characterized in Chapter 1 using the 'Lovell' peach genome (Verde *et al.*, 2017) and the compacted genome of ‘Texas’ v.2.0 (Alioto *et al.*, 2020). Marimon *et al.* (2020) localized this gene between the two markers (Indel16912 and SNP_17184692) in positions Chr 2:16,912,811 to 17,184,692 of the peach genome. This region spans 272 kb and encompasses 27 annotated genes.

We determined that the localization of the *Vr3* gene is at positions Chr02:16,026,755-16,263,319 in Texas v.3.0 (P0) (a region of 236 Kb), and at positions Chr02:12,907,187-13,129,481 in Texas v.2.0 (a region of 222 Kb where there are 23 annotated genes). We then conducted a comparison of this region through alignments. It was observed that the alignment between Texas v.3.0 and Texas v.2.0 is not perfect, revealing a region present in Texas v.3.0 that is absent in Texas v.2.0, and vice versa (Figure 3A). Consequently, we chose to align both genomes with the peach genome and observed a more consistent alignment between Texas v.3.0 and peach (Figure 3B), compared to the alignment between Texas v.2.0 and peach (Figure 3C). This suggests that this region of significant agronomic interest is better resolved in the new genome, as opposed to Texas v.2.0.

Moreover, an interesting discovery was made in Texas v.3.0, where a duplication of a region from the peach genome was identified. Furthermore, a comparison was undertaken between P0 and P1 of ‘Texas v.3.0, revealing a flawless alignment and highly identical sequences. Consequently, the sequence in this region is non-polymorphic.

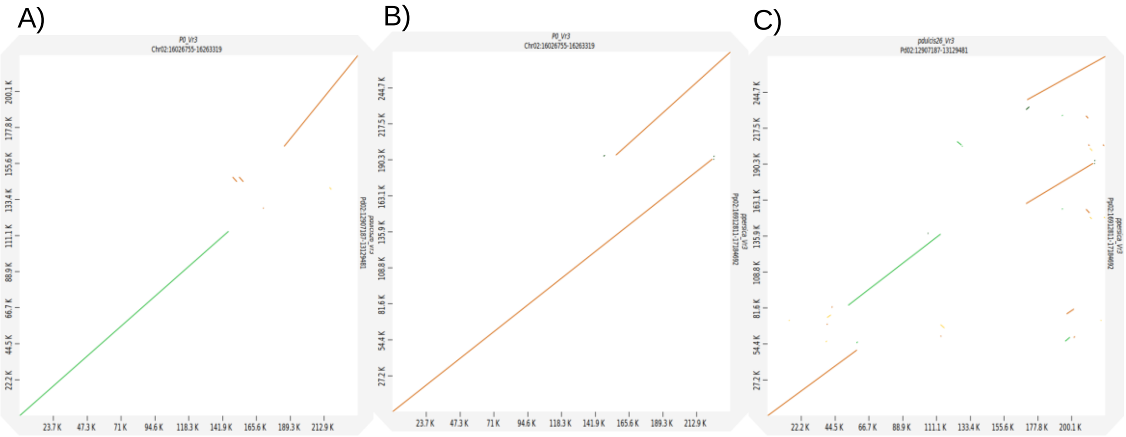


Figure 3. A) Alignment between Texas v.3.0 (P0) and Texas v.2.0 (pdulcis26) genomes. B) Alignment between Texas v.3.0 (P0) and Peach 'Lovell' genomes. C) Alignment between Texas v.2.0 (pdulcis26) and Peach 'Lovell' genomes. Light green indicate identity between 50 and 70% and brown indicate identity between 25 and 50%.

Within the Vr3 region of the Texas v.3.0 genome, there are 29 annotated genes (Table 5), which is an increase of six genes compared to the previous version, Texas v.2.0, and two genes more than what is found in the peach genome 'Lovell'. Among these 29 genes, our Chapter 1 candidate, the ABC transporter (*TexasF0_G8087*), has been annotated uniquely, distinct from its annotation in the peach genome as two different genes.

Table 5. Annotated genes in the *Vr3* locus in P0 of Texas v.3.0.

GENE	FUNCTION
TexasF0_G8081	agamous-like MADS-box protein AGL82
TexasF0_G8082	Uncharacterized protein
TexasF0_G8083	Uncharacterized protein
TexasF0_G8084	Germin-like protein
TexasF0_G8085	Germin-like protein
TexasF0_G8086	Uncharacterized protein
TexasF0_G8087	AAA-type ATPase family protein
TexasF0_G8088	ABC-type xenobiotic transporter
TexasF0_G8089	PMD domain-containing protein
TexasF0_G8090	EF-TU receptor (Fragment)
TexasF0_G8091	Disease resistance protein TIR-NBS-LRR class
TexasF0_G8092	LRR and NB-ARC domains-containing disease resistance protein
TexasF0_G8093	Disease resistance protein RGA4

TexasF0_G8094	PMD domain-containing protein
TexasF0_G8095	ULP_PROTEASE domain-containing protein
TexasF0_G8096	Putative P-loop containing nucleoside triphosphate hydrolase, leucine-rich repeat domain, L
TexasF0_G8097	probable disease resistance protein RPP1
TexasF0_G8098	Disease resistance protein RGA4
TexasF0_G8099	haloacid dehalogenase
TexasF0_G8100	Endoglucanase
TexasF0_G8102	Diaminohydroxyphosphoribosylaminopyrimidine deaminase (Fragment)
TexasF0_G8103	DNA replication ATP-dependent helicase/nuclease
TexasF0_G8104	Sister chromatid cohesion protein DCC1
TexasF0_G8105	Zinc ion-binding protein
TexasF0_G8106	Protein ECERIFERUM 1-like
TexasF0_G8107	C-JID domain-containing protein (Fragment)
TexasF0_G8108	TIR domain-containing protein
TexasF0_G8109	Peroxidase
TexasF0_G8110	Putative aldehyde oxygenase (Deformylating)

TEs are at the origin of a major fraction of the heterozygous structural variation

We compared the two phases of Texas v.3.0, P0 and P1, and found 408,670 SNPs, 152,334 INDELs (< 40bp) and 8,183 structural variants (SVs, length > 40 bp). This corresponds to one SNP every 313bp, one INDEL every 840 bp, and one SV every 10,042 bp of the 128 Mb spanned by the haplotigs.

Among the 8,183 SV detected, the vast majority (93,6 %) were insertions and deletions. An important fraction of the insertions/deletions (32%) overlap almost perfectly (> 80%) with a TE annotation, in particular for the large insertion/deletions, suggesting that they correspond to heterozygous TE insertions (1,314 specific of P0 and 1,258 specific to P1). In addition, we identified 29% of SVs that partially overlap with TEs (with deletion/insertion covering < 80% of TE length). These cases are not likely the result of transposition but may be the result of TE internal deletions or rearrangements. In any case this suggests that a major fraction of the heterozygous structural variation is TE-related.

We detected heterozygous TE insertions from all the different TE orders, with LTR-RTs being the most abundant (66 % of the total). The Texas v.2.0, as well as all the other publicly available almond genome assemblies, is are collapsed representations of the two haplotypes. We hypothesized that heterozygote TE insertions may thus be underrepresented in unphased assemblies, which could be one of the reasons explaining the difference in the number of TEs between Texas v.3.0 and Texas v.2.0. Indeed, an analysis of the 3,148 TEs missing in Texas v.2.0 showed that 94.1% of them are heterozygous.

To test the potential impact of TE insertions on gene expression we obtained RNASeq data from almond ‘Texas’ immature fruits and flowers which we complemented with the already available RNASeq data from leaves (presented in Chapter 1 and in de Tomás *et al.*, 2022). We compared the expression of genes that do not contain a TE insertion in the proximal upstream region (1Kb) with that of genes carrying a homozygous or heterozygous TE in this region. The analysis was conducted using a total of 24,394 genes, which are annotated genes with at least one mapped RNASeq read, excluding pseudogenes.

We observed the same pattern in the three RNASeq datasets obtained from different organs. In all cases, genes with homozygous TE insertions had lower expression than genes without TEs ($p < 0.05$), which suggests that TE insertions in the upstream regions of genes have, in general, a negative effect on gene expression. Interestingly, this trend is reversed for the genes harboring a heterozygous insertion in the proximal upstream region, which have a higher expression level ($p < 0.05$) than those without a TE insertion (Figure 4). This analysis was performed in P0 (Figure 4B), P1 (Figure 4C), and the combination of the two phases (Figure 4A). We observed the same pattern in both phases.

Furthermore, this analysis was repeated with all genes regardless of their number of mapped reads (a total of 29,183 genes), and the results follow the same pattern that the analysis of the 24,394 genes.

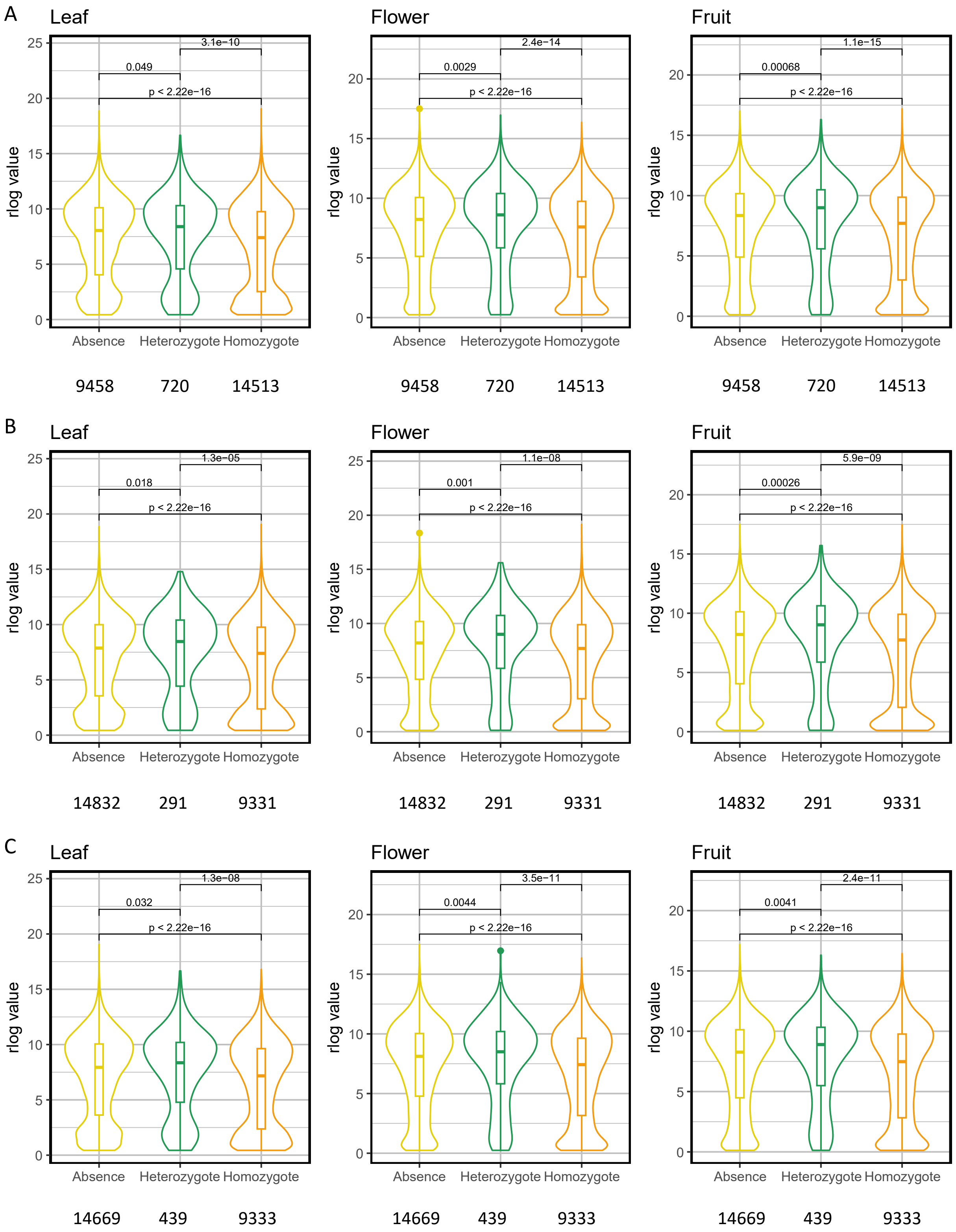


Figure 4. Relationship between gene expression levels (log scale) and the presence of homozygous and heterozygous TE insertions at 1Kb upstream gene TSS. Expression in the Y-axes is presented in logarithmic scale (Deseq2 regularized log values). The number of genes for each condition is included below each violin plot. In cases where genes have both homozygous and heterozygous insertions in the proximal region, they have been included in both groups. A, corresponds to the combination of both phases. B, correspond to Phase 0. C, correspond to Phase 1. p values are included between the different comparisons.

This could suggest an opposite impact of homozygous and heterozygous TE insertions, with heterozygous insertions activating gene expression. In order to test this hypothesis, we have analyzed the allele-specific expression of the genes containing a heterozygous TE insertion in the upstream proximal region (we detected a total of 31 genes). Despite not achieving statistical significance due to the limited number of genes (a total of 31 genes, comprising 20 genes with an insertion in P0 and 11 genes with an insertion in P1), our data clearly show that, in general, the allele without the TE insertion tends to be expressed at a higher level than the one containing the insertion (Figure 5), which confirms that TE insertions in the proximal upstream regions of genes have, in general, a negative effect on gene expression, irrespective of the zygosity level of the insertion.

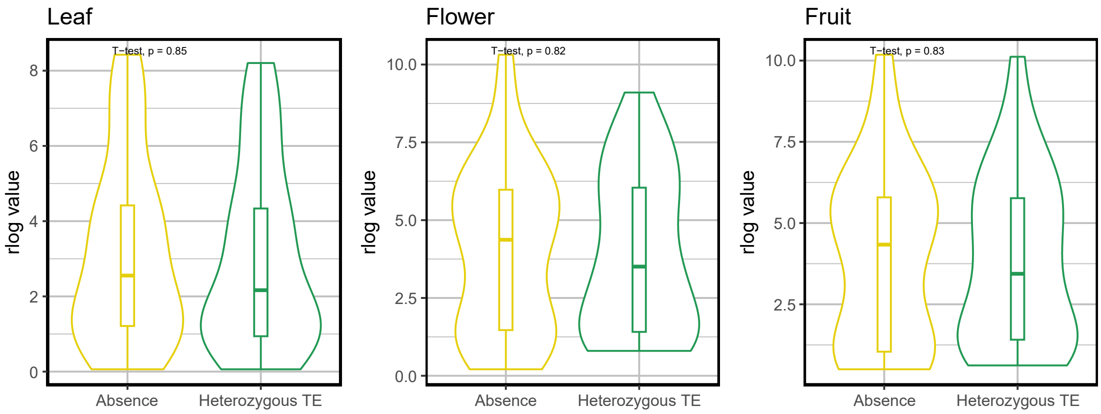


Figure 5. Relationship between allele-specific expression levels (log scale) of the allele without the TE insertion (Absence; left violins) and the allele with a heterozygous TE at 1Kb upstream gene TSS (Heterozygous TE; right violins) for each organ. Expression in the Y-axes is presented in logarithmic scale (Deseq2 regularized log values).

The difference of expression of genes with heterozygous TE insertions, could also be due to a preference of insertion of TEs into genes that are highly expressed. To test this hypothesis, we produced transcriptomic data on the same tissues in peach (the peach samples presented in the Chapter 1 and de Tomás *et al.*, 2022) and obtained the gene expression levels of almond-peach orthologous gene pairs. Then we looked for gene pairs where almond has a heterozygous TE insertion in the promoter region that is missing in the peach ortholog (likely a recent insertion occurred after the split of the two species). Our results clearly indicated that those peach genes had higher expression levels than the remaining genes (Figure 6).

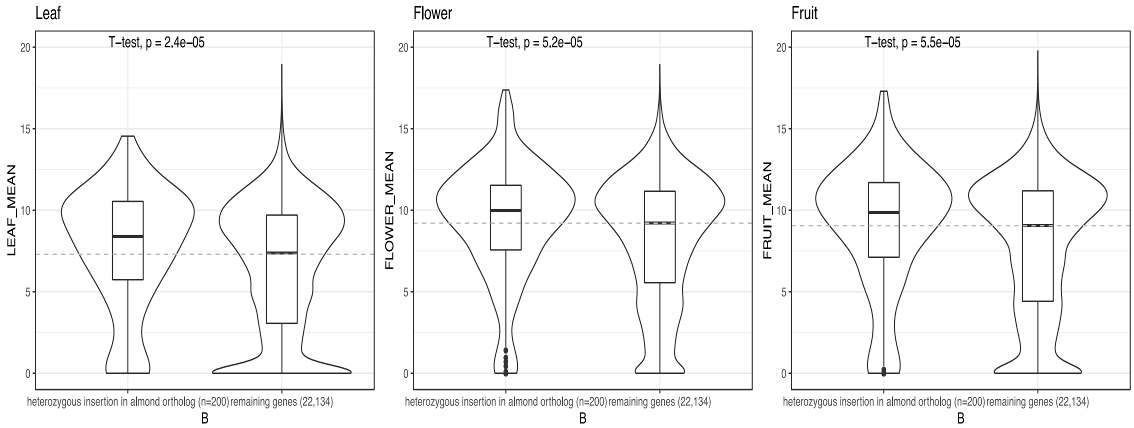


Figure 6. Expression levels of peach genes with (left violins) or without (right violins) heterozygous TE insertion in the promoter of the almond ortholog for each organ.

Patterns of allele-specific expression during almond development

In order to analyze the possible allelic specific expression (ASE) we searched for SNPs present in gene coding regions without surrounding INDELS (at < 50bp) that could allow us to differentiate the expression of the two alleles. We found 24,051 SNPs fulfilling this requirement in up to 6,939 genes, for which 6,182 showed detectable expression in at least one of the organs tested (leaves, flowers and immature fruit). We found that 579 genes (9.3 % of the expressed genes with informative SNPs) showed ASE in at least one organ (82 in leaf, 493 in flower and 271 in fruit). Only a small number of genes (68) showed the same pattern of allelic expression in the three organs, whereas 383 genes displayed ASE only in a single organ. However, the capacity to detect ASE greatly depends on the RNASeq coverage, which is higher in our flower and fruit data.

Therefore, and in order to minimize the possible bias introduced by differences of coverage on the RNASeq data of the three organs, we extracted the genes displaying ASE and more than 10 reads mapping to target SNPs in the three replicates of each organ (250 in total). We performed a hierarchical clustering and found four clusters of co-expressed alleles (Figure 7). Clusters 1 and 2 represent genes where one of the alleles is predominantly expressed in all organs. On the contrary, Clusters 3 and 4 contain genes that express different alleles in different organs. For example, some genes of cluster 3 specifically express the P0 allele in flowers, whereas some genes of cluster 4 express the P1 allele in flowers and the P0 in the other organs. Furthermore, a functional enrichment of the four clusters has been performed, but it has not been feasible due to the low number of genes for each cluster.

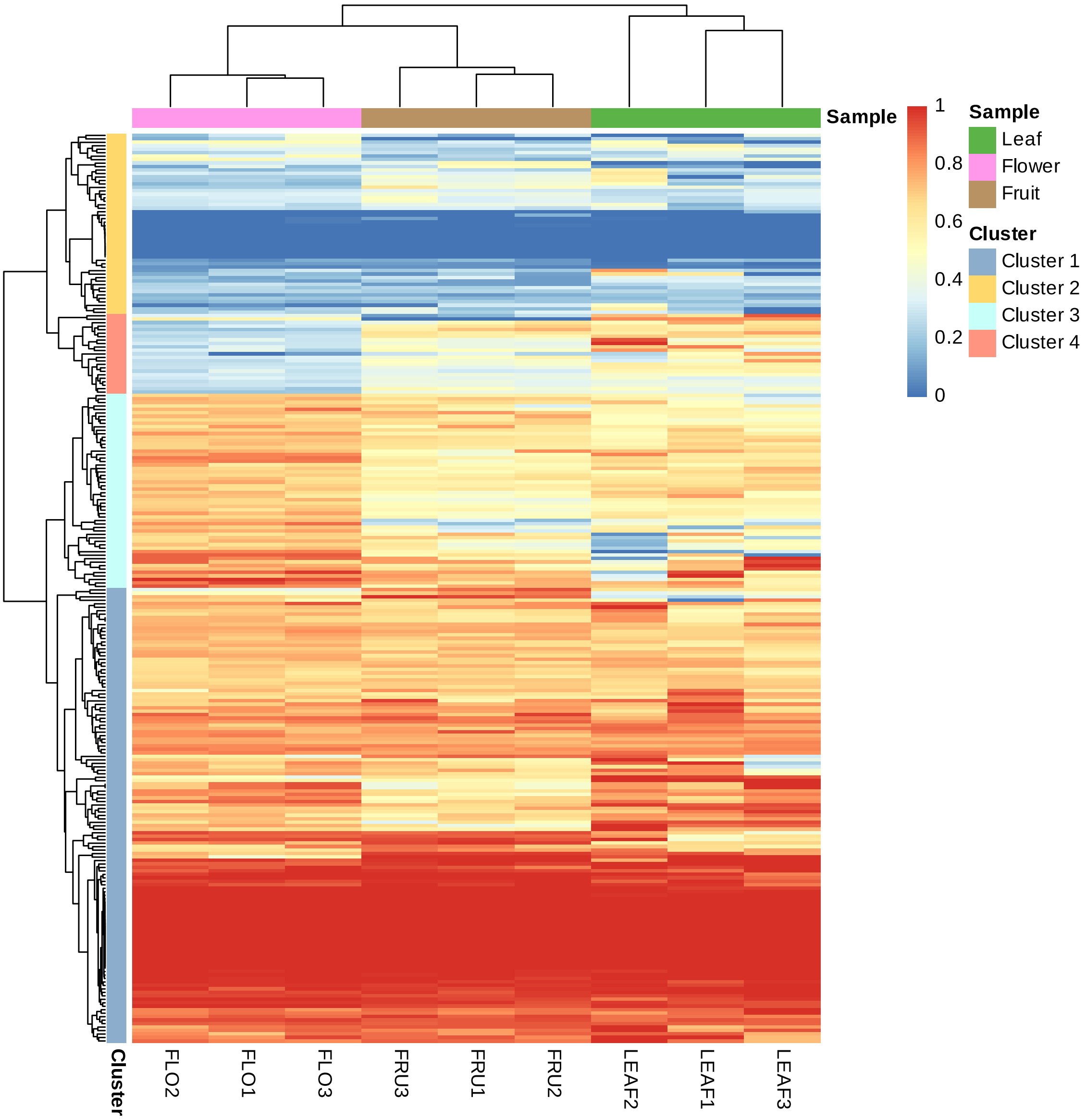


Figure 7. Heatmap representing the allele-specific expression profiles of the 250 genes with at least 10 mapped reads in every replicate. Each row represents a gene. Colors indicate the percentage of mapped reads from each allele (red = 100% P0, blue = 100% P1) over the total.

As mentioned in the previous section, we have detected a heterozygous insertion in 31 of 579 cases with ASE bias. Generally, an increase in expression is observed in the allele that lacks the insertion (Figure 5). Nevertheless, there are instances where the allele containing the insertion shows an increase in expression compared to the other allele. A clear example is the *G17311* gene (Cinnamoyl-CoA reductase 2-like), a pivotal enzyme in plant lignin synthesis, with roles in plant secondary cell wall development and environmental stress defense. This gene displays an LTR retrotransposons insertion (TE_84210) within 1kb upstream region in P0, while this insertion is absent in P1 (Figure 8). This heterozygous insertion could potentially produce its allelic expression bias, resulting in higher expression of the allele with the insertion across all organs (Table 6).

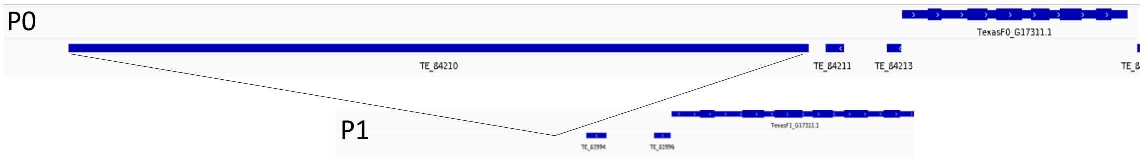


Figure 8. Example of gene (*G17311*) with allelic expression bias (P0 allele is more expressed) potentially caused by the insertion of an LTR-retrotransposon (TE_84210) in P0, that is not present in P1.

Table 6. Allelic specific expression (Deseq2 regularized log values) of *G17311* gene for each organ.

G17311	**LEAF**	**FLOWER**	**FRUIT**
P0	7.923	6.682	7.180
P1	7.263	5.352	4.291

Cultivar variability analysis

A genetic variability analysis was conducted on 40 public accessions of almond varieties spanning Europe (Spain, Italy and France), USA and China (Table 1). The annotation of the new almond genome version 3.0 (Texas) and the PopoolationTE software were utilized for this study. A total of 26,487 insertions were detected across these varieties, with a mean of frequency in the population per insertion of 0.803.

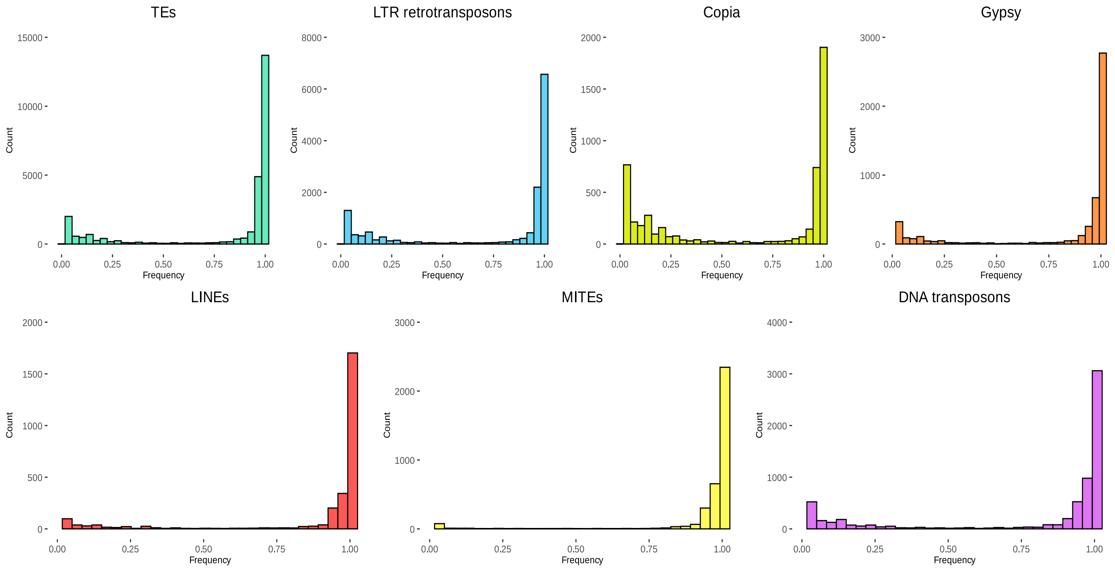


Figure 9. Count of detected insertions (for all the TEs, LTR Retrotransposons, Copia, Gypsy, LINEs, MITEs and DNA transposons) and their frequency in the almond population.

Among these insertions, various types Transposable Elements were identified, including Copia and Gypsy LTR retrotransposons, as well as LINEs, MITEs, and DNA transposons, especially TIRs. The distribution of these insertions for all types of TEs in the population is quite similar (Figure 10). The majority of insertions are highly fixed (with a frequency in the population equal to 1). However, a peak of insertions specific to a few varieties is also observed, especially in the case of the LTR retrotransposons. It is most evident in Copia LTR retrotransposons. Generally, Gypsy retrotransposons exhibit higher fixation (mean = 0.830) compared to Copia (mean = 0.634). This could be attributed to Gypsy elements being located in pericentromeric regions (Alioto *et al.*, 2020) and not being eliminated as easily as Copia elements.

Among the 26,487 detected insertions, we investigated the number of insertions that corresponded to homozygous and heterozygous insertions in ‘Texas’. We found that 21,595 insertions were homozygous (7,459 TIPs), while 224 insertions were heterozygous (217 TIPs). We observed that the heterozygous insertions are present at a much lower population frequency than the homozygous ones (mean heterozygous = 0.340; mean homozygous = 0.966) (Figure 10). However, some of the heterozygous insertions are present at high frequencies in the population and may be relatively old.

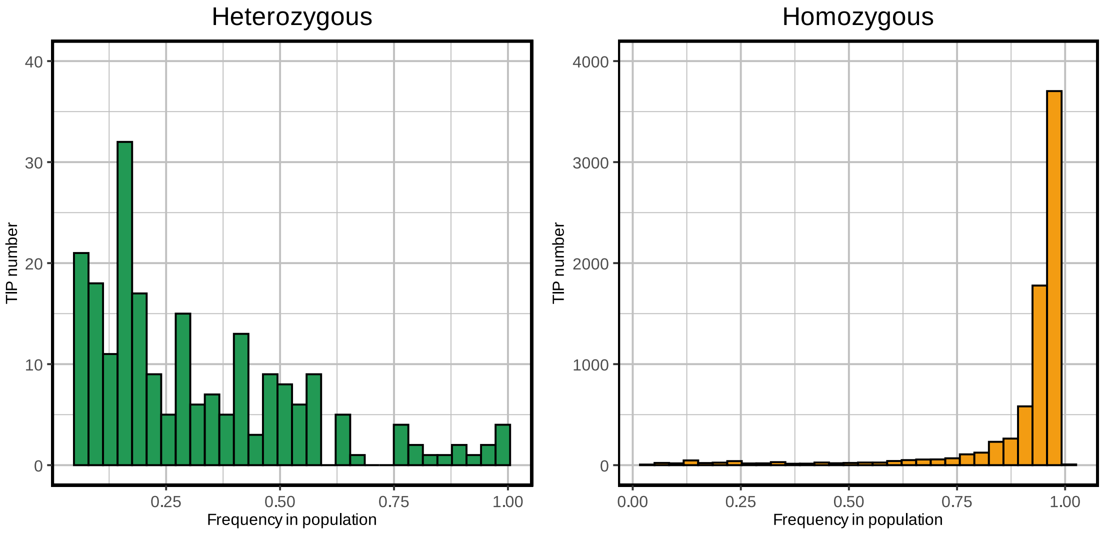


Figure 10. Number of Transposon insertion polymorphism (TIP) and their frequencies in a population of almond cultivars.

An analysis of the insertion time of LTR-RTs showed that the heterozygous insertions are in general more recent than the homozygous ones (mean heterozygous = 2.6 Mya, homozygous = 6.5 Mya, Wilcoxon $p < 0.05$) (Figure 11), suggesting that they have not had the time to become fixed.

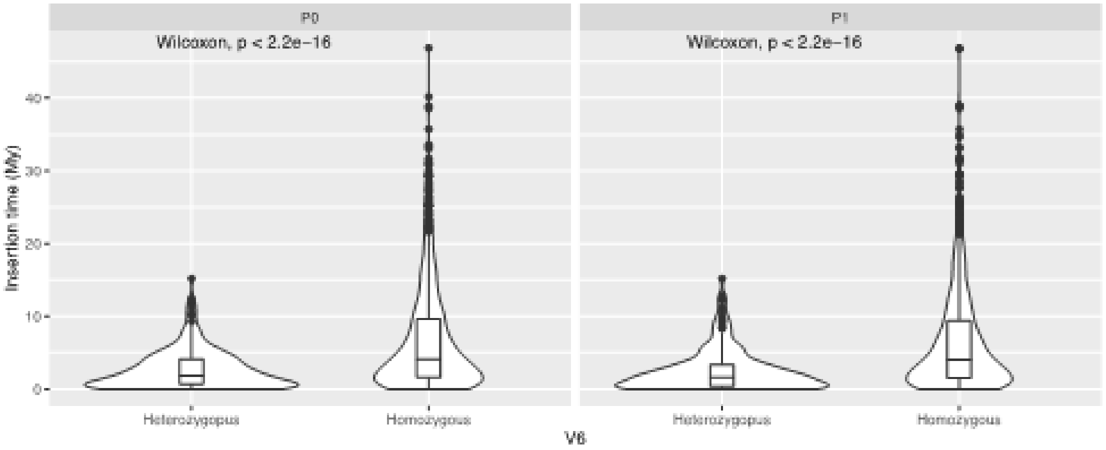


Figure 11. Distribution of heterozygous and homozygous LTR-retrotransposon insertion age (made by Dr. Raúl Castanera).

The analysis of the distribution of heterozygous and homozygous insertions shows that heterozygous insertions are in general closer to genes as compared with the homozygous insertions (Figure 12), which, as in general the heterozygous insertions are younger, may suggest an impact on gene coding or expression capacity of these TE insertions that are purged with time.

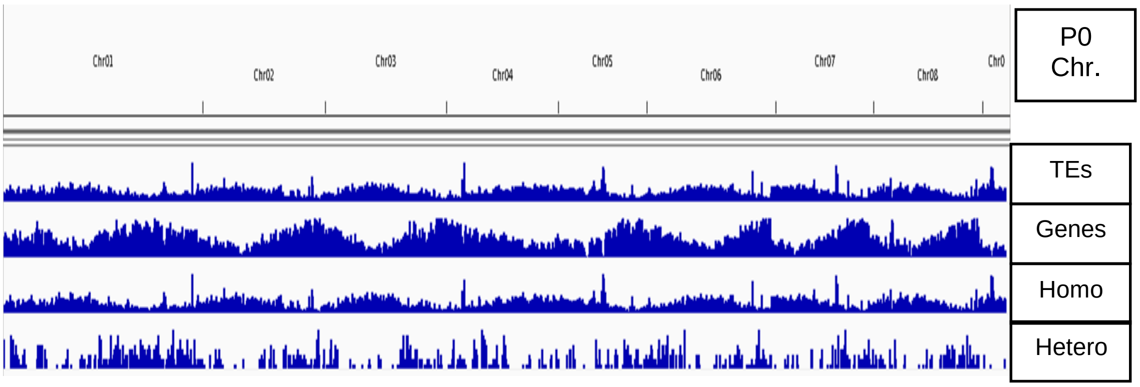


Figure 12. Distribution of Transposable Elements (TEs), Genes, homozygous (homo) and heterozygous (hetero) TEs across the different chromosomes (chr.) of the P0 of the version 'Texas v.3.0' of almond genome.

2.5. Discursion

The new version of the almond genome of the variety ‘Texas’, known as Texas v.3.0 is distinguished by having both P0 and P1 phases sequenced, unlike the older version of the genome Texas v.2.0 (also referred to as pdulcis26) (Alioto *et al.*, 2020). Recent advancements in sequencing techniques, particularly the utilization of long and accurate reads, have made this achievement possible (Guk *et al.*, 2022).

Within the family *Rosaceae*, we already have the ‘Gala’ apple genome (Sun *et al.*, 2020), which has been resolved into two distinct phases, but within the genus *Prunus*, Texas v.3.0 is the first haplotype-resolved genome. It represents a remarkable achievement considering the highly non-inbred nature and complex genomic structures of the plants (Guk *et al.*, 2022). This haplotype-resolved genome allows for a more thorough investigation of the almond species, which is highly heterozygous and exhibits greater variability compared to other species within the same genus (Velasco *et al.*, 2016). Moreover, it facilitates the study of transposable elements and their impact.

The assembly corresponds to a total of 254.02 Mb for Phase 0 and 252.65 Mb for Phase 1. Each phase comprises 80 scaffolds and 99 contigs, and 97.3% of the total sequence was anchored to eight pseudomolecules, demonstrating significant synteny with the previous genome Texas v.2.0 (Alioto *et al.*, 2020). Texas v.3.0 contains up to 13.2% more contig sequence homogeneously distributed among the 8 chromosomes, leading to an improved sequence contiguity (Table 2). Due to these improvements, this new genome could be considered a gold quality genome, following Ou *et al.* (2018) criteria, along with other good resolved genomes as ‘Nipponbare’ rice (MSUv7) (Kawahara *et al.*, 2013) and maize (B73 v4) (Jiao *et al.*, 2017). The new genome includes 28,625 annotated genes in P0 and 29,616 in Phase 1, indicating an increased number of annotated genes compared to Texas v2.0 (Alioto *et al.*, 2020), which contained 27,969 genes (Table 3).

This improved assembly can prove highly valuable for better characterizing regions of agronomic interest, such as the region described by Marimon *et al.* (2020), which contains the powdery mildew resistance gene *Vr3*. In our analysis, we observed an enhanced assembly of this region in the Texas v.3.0 genome and annotated six additional genes compared to Texas v.2.0 (Figure 3, Table 5). As mentioned in Chapter 1, almond is resistant to the powdery mildew while peach is susceptible. When aligning Texas v.3.0 with the 'Lovell' peach genome (Verde *et al.*, 2017), we identified a duplicated region in almond (Figure 3). It remains uncertain whether this region results from assembly issues or represents a true duplication. Nevertheless, duplications in genomes are a prominent factor in the diversity and evolution (Crow & Wagner, 2006) and could potentially play a role in powdery mildew resistance. For instance, Rajaraman *et al.* (2018) suggests that the *ARM1* gene is an example of gene neo-functionalization derived from a gene duplication event. This duplication has played a role in quantitative resistance against the powdery mildew in the *Triticeae* tribe. In any case, our discussion from the chapter 1 remains compatible with these new results because our candidate gene for powdery mildew resistance, the ABC transporter, is still annotated in this new genome (as *G8088*), and located within this region (Table 5). Therefore, it could potentially be a candidate for the *Vr3* gene.

The Texas v.3.0 TE annotation encompasses approximately twice the number of complete elements compared with Texas v.2.0 (Table 4). This increase is remarkable in LTR retrotransposons, MITEs, and TIRs. Regarding LTR retrotransposons, the increase is mainly attributed to Gypsy elements, which are commonly known to insert into pericentromeric regions according to the literature (Neumann *et al.*, 2011; Alioto *et al.*, 2020). For example, the Gypsy Retrotransposons called Centromeric Retrotransposon lineage of Chromovirus, also called Centromeric Retrotransposons of Maize (CRM) tend to be found in centromeric regions (Neumann *et al.*, 2011). They carry heterogenous domains at their integrase C-terminus, possibly related to their chromosomal distributions. These domains, including the chromodomain and CR motif, interact with the Centromere-specific histone H3 (CENH3) protein, implying the CRMs in centromere function (de Castro *et al.*, 2018). The enhancement of the assembly in these pericentromeric regions could have played an important role in this annotation improvement. It is confirmed by the better identification of the 166 bp centromeric repeat sequence (Melters *et al.*, 2013) in this genome than in Texas v.2.0. On the other hand, we have observed that a significant fraction of the annotated insertions (LTR retrotransposons and TIRs) are young (Figure 2), so they haven’t been eliminated by purifying selection, which aligns with their heterozygosity.

The comparison between the two phases (P0 and P1) allowed the identification of 408,670 SNPs, 152,334 INDELs, and 8,183 structural variants. The vast majority of the structural variants were insertions/deletions, with 32% of them perfectly overlapping with our TE annotation and 29% showing partial overlap. This suggests that a significant portion of heterozygous structural variation is associated with transposons. One of our hypotheses is that genomes with compacted assemblies have underrepresented heterozygous insertions, which would also influence the difference in the number of annotated insertions between Texas v.2.0 and Texas v.3.0 genomes. Our analysis reveals that 94% of the missing insertions in Texas v.2.0 are heterozygous, suggesting that for regions harboring heterozygous TE insertions, the empty haplotype was more frequently included in the Texas v.2.0 assembly.

Our observations highlight the impact of transposable elements (TEs) on gene expression. Our analysis demonstrates that genes harboring a homozygous TE insertion in their upstream region generally exhibit lower gene expression levels in almond compared to those lacking such insertions, implying a negative influence on gene expression. And the genes with near heterozygotes show higher expression than those without insertions (Figure 4). These comparisons were statistically significant and held true whether the heterozygous insertion was in Phase 0 or Phase 1. Furthermore, the analysis consistently followed the same pattern across the three studied organs: leaves, flowers, and fruits. These effects on gene expression could be caused because TEs can serve as novel alternative promoters, leading to the generation of alternative transcripts. Also, they can introduce novel cis-acting regulatory sites that function as enhancers or become integrated within existing enhancers, thus shaping the production of transcripts. Also, TEs can cause chromatin modifications within regions near genes, resulting in effects on gene expression levels (Hirsch & Springer, 2017).

Our results suggest an opposite impact of homozygous and heterozygous TE insertions. For this reason, we analyzed the unexpected impact of the heterozygotes in detail. Specifically, we observed that among genes exhibiting Allele-specific expression (ASE) imbalance and a nearby heterozygous insertion, the allele carrying the insertion tends to be less expressed than the allele without the insertion (Figure 5). These results reaffirm that insertions, in general, tend to have a negative effect on gene expression. Furthermore, we hypothesized that the heterozygous TE are inserted on highly expressed genes. It was confirmed by the analysis of the orthologs in peach (Figure 6). These heterozygous TEs could be lost over time due to purifying selection.

The ASE analysis showed among 6,939 genes with enough SNPs to differentiate the expression of both alleles. Among them, 6,182 had detectable expression in at least one organ. Within this group, 579 genes (9.3% of the 6,182 genes) displayed ASE in at least one organ: 82 genes in leaves, 493 in flowers, and 271 in fruits. The differences in the number of each organ due to differences in coverage among organs, with leaf samples having the lowest coverage. Only 68 genes showed the same ASE patterns across all organs, while 383 genes showed an organ-specific pattern. From these 579 genes, we extracted 250 with a minimum of mapped reads and performed a hierarchical clustering, revealing four clusters of co-expressed alleles (Figure 7). Clusters 1 and 2 represented alleles predominantly expressed in all organs, while clusters 3 and 4 indicated genes expressing different alleles in distinct organs. Functional enrichment analysis was not possible due their low number, but we conclude that looking into these 579 genes in detail and individually, which have different role in each organ, is interesting for the future.

Among these 579 genes, 31 (5.35%) had a near heterozygous insertion in the upstream region. As previously explained, the allele carrying the insertion generally exhibited lower expression than the one without it. Nonetheless, certain genes with ASE and a nearby heterozygous insertion increased expression in the allele with the insertion. For example, the gene *G17311*. This gene, associated with lignin synthesis, secondary cell development, and stress defense, displayed increased expression in the allele carrying the LTR retrotransposon in its upstream region (Figure 8, Table 6). It can be an interesting gene for agronomic traits, biofuel production and the pulping industry (Yoon *et al.*, 2015).

The variability analysis across the 40 almond varieties showed a total of 26,487 insertions, corresponding to various transposon groups. The majority of these insertions were fixed within the population at a mean of frequency of 0.803 (Figure 9). A clear peak in population frequency was observed at 1, with a smaller peak at low frequencies. This latter peak was prominent in Copia LTR retrotransposons, which had a population frequency of 0.634, compared to 0.830 for Gypsy LTR-RT. This distribution could be explained because the Gypsy are located in pericentromeric regions, as mentioned before (Neumann *et al.*, 2011; Alioto *et al.*, 2020), where they are less prone to elimination. Additionally, we investigated the number of insertions that were heterozygous and homozygous in ‘Texas’. We found that 21,595 insertions were homozygous (7,459 TIPs), while 224 were heterozygous (217 TIPs). The difference in the number can be attributed to the lower detectability of heterozygotes due to their lower zygosity. Homozygous insertions (mean of population frequency = 0.966) were more fixed within the population than heterozygous insertions (mean = 0.340) (Figure 10). These findings align well with our data on the age and distribution of homozygous and heterozygous LTR retrotransposons, indicating that heterozygous insertions, being younger and closer to genes (Figure 11; Figure 12), have not yet been purged by purifying selection.

Phased genomes offer valuable insights into various aspects of genome dynamics, including structural variations, SNPs, and ASE (Guk *et al.*, 2022). Texas v.3.0 has proven to be a useful tool for our analysis and for further exploration in a highly heterozygous species like almond (Velasco, 2016).

CHAPTER 3:

IDENTIFICATION, CLASSIFICATION, AND CHARACTERIZATION OF RECENTLY INSERTED ENDOGENOUS PARARETROVIRUSES, INCLUDING THE DISCOVERY OF THE NEW PUTATIVE GENUS *WENDOVIRUS*

3.1. Introduction

In the last decade, some studies based on a limited number of plant genomes have investigated the presence and diversity of the Endogenous Viral Elements (EVEs) called Endogenous Pararetroviruses (EPRVs). These studies evidenced EPRVs integrated forms among genomes of vascular plants, including clubmosses, ferns, and gymnosperms. The genus *Florendovirus*, while lacking known episomal forms, is a major component of flowering plant genomes.

EPRVs are usually located in hotspots, which can be unevenly distributed across chromosomes (Vassilieff *et al.*, 2023), particularly in heterochromatin and pericentromeric regions of chromosomes, and close to retrotransposons (Staginnus & Richert-Pöggeler, 2006; Yu *et al.*, 2019). Hence, similar to transposable elements, heterochromatin/pericentromeric regions might serve as safe regions for EPRVs. Within these regions, EPRVs could evade elimination mechanisms, and their presence would have neutral on the host, potentially persisting throughout long evolutionary periods (Vassilieff *et al.*, 2023).

The integration of any EVE into or near a gene can potentially modify gene transcription or modify mRNA processing, resulting in mutant phenotypes. Most of the described EPRVs are inserted in intergenic regions and have no apparent deleterious effect on the host. However, there are examples of EPRVs inserted inside genes with potential effects on gene expression, as the case of the grape vine (*Vitis vinifera*), which has several EPRVs inserted in introns (90% of their *Florendoviruses*) (Geering *et al.*, 2014). Another case with potential impact is the integration of an endogenous *Petuvirus* into a Citrus tristeza virus (CTV) resistance locus in the trifoliate orange (*Poncirus trifoliata*) genome. Additionally, some segments of EPRVs were discovered in the *Cg1g024630* gene of pummelo (*Citrus maxima*), which may be related to this CTV resistance gene locus (Yu *et al.*, 2019).

Most of the EPRVs are transcriptionally or translationally inactive because they are partial and/or comprise rearranged sequences and/or inactivating mutations. Often EPRVs form clusters resulting from the integration of several complete or partial copies in tandem or nested (Richert-Pöggeler *et al.*, 2003), sometimes leading to integration hotspots. These repetitive structures in tandem could originate from the integration of concatemers of viral DNA or from homologous combination between existing EPRVs, or between EPRVs and the episomal forms of a related virus (Vassilieff *et al.*, 2023).

Infrequently, these integrated sequences are transcriptionally active and the resulting RNAs can serve as precursors of extrachromosomal viral DNA and lead to systemic and vertically transmitted infections (Hohn *et al.*, 2008; Gayral *et al.*, 2008). Transcriptional activation can be driven by viral promoters present within the integrated element or plant promoters in the vicinity of the EPRV sequence (Lockhart *et al.*, 2000; Kuriyama *et al.*, 2020). Replication-competent EPRVs (infective EPRVs) have been reported only for interspecific hybrids of the plant genera *Musa*, *Petunia* and *Nicotiana*: endogenous banana streak viruses (eBSV) in bananas, endogenous petunia vein-clearing virus (ePVCV) in petunias, and endogenous tobacco vein-clearing virus (eTVCV) in Nicotiana (Vassilieff *et al.*, 2023).

On the other hand, EPRV derived RNAs can also be inducers for RNA interference (RNAi) and gene silencing mechanisms through the generation of small interfering RNAs (siRNAs) (Bertsch *et al.*, 2009; Ricciuti *et al.*, 2021), which act in different silencing pathways, including transcriptional gene silencing and post-transcriptional gene silencing (Richert-Pöggeler *et al.*, 2021).

As mentioned in the General Introduction, EVEs are often referred to as genomic “fossils” and are subject of study in the Paleovirology, representing remnants of past viral infections (Etienne, 2017). So, another important property of the EPRVs is that the EVEs can be used to calibrate the timing of virus evolution. If an EVE is orthologous across several species, this gives a minimum estimate for the age of the virus that integrated into the genome. (Aiewsakun & Katzourakis, 2015).

RNA-directed DNA polymerase (Reverse Transcriptase, RT) coding sequences are present in a wide variety of genetic elements and contains a relatively well conserved central domain, allowing its use for phylogenetic analyses (Hansen & Heslop-Harrison, 2004) and for searches for homologues of, for example, EPRVs in genome sequences (Diop *et al.*, 2018).

As mentioned earlier, previous studies have examined the EPRVs diversity in plant genomes based on the limited number of genome sequences available in each case (Geering *et al.*, 2014; Diop *et al.*, 2018). In particular, Geering *et al.* (2014) analyzed genomes from 35 plant species, while Diop *et al. (*2018) analyzed genomes from 72 species.

Nowadays, there has been a significant increase in both the number of sequenced plant genomes and the quality of these genome sequences. This progress can be attributed in part to advancements in genome assembly techniques, including the utilization of long read sequencing (Michael & VanBuren, 2020).

For this reason, we decided to screen 278 genomes corresponding to 267 species for the presence of EPRVs, obtaining a broader picture of the distribution of these endogenous elements. We identified the major EPRV lineages and analyzed their distribution in the different plant orders and genera. We also describe a new possible genus of *Caulimoviridae* present only as EPRVs we called *Wendovirus*. Finally, we studied in more detail the EPRVs of peach and almond, studying their distribution and transcription.

The majority of these results are presented in the scientific paper titled: Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (*Caulimoviridae*). This paper, authored by de Tomás & Vicient (2022) is published in the Plant Bioinformatics section of the journal Frontiers in Plant Research (see Annexes).

3.2. Objectives

The objectives of this study are as follows:

- Identify recently integrated Reverse Transcriptase domains of Endogenous Pararetroviruses in several plant genomes.

- Classify the Reverse Transcriptase domains discovered in different genera of *Caulimoviridae*.

- Calculate the minimum ages of the integration events reported in this study.

- Characterize the Endogenous Pararetroviruses belonging to a new putative genus of the *Caulimoviridae* family named *Wendovirus*.

- Characterize the distribution and transcription patterns of EPRVs in peach and almond plants, as well as in their F1 hybrid.

3.3. Material and methods

Identification of recently integrated reverse transcriptase domains

We built a library containing an assortment of 182 RT central domain amino acid sequences (Supplementary Data 1 of de Tomás & Vicient, 2022). This collection includes one sequence from *Retroviridae*, 14 from Ty3/Gypsy LTR retrotransposons of the six most abundant genera in plants (Athila, CRM, Galadriel, Ogre, Reina, Retand and Tekay), 104 from the eleven genera of *Caulimoviridae* with episomal forms (*Badnavirus*, *Caulimovirus*, *Vaccinivirus*, *Soymovirus*, *Cavemovirus*, *Solendovirus*, *Dioscovirus*, *Rosadnavirus*, *Tungrovirus*, *Petuvirus* and *Ruflodivirus*), and 63 from six groups of exclusively endogenous *Caulimoviridae* (*Florendovirus*, *Xendovirus*, *Yendovirus*, *Zendovirus*, *Gymnendovirus* and *Fernendovirus*) (hereafter referred to as operational taxonomic units (OTUs) following the nomenclature proposed by Diop *et al.* (2018). For further analyses, we selected ten sequences representatives of the *Caulimoviridae* groups listed in Table 1. The sequences are available in Supplementary Data 2 of de Tomás & Vicient (2022).

Table 1. List of *Caulimoviridae* RT sequences used in the tBLASTn analyses.

NAME	ABBREVIATION	ACCESION	GENUS	SIZE (aa)
Strawberry vein banding virus	SVBV	X97304.1	*Caulimovirus*	306
Blueberry fruit drop associated virus	BFDaV	NC_028462.1	*Vaccinivirus*	304
Blueberry red ringspot virus	BRRV	AF404509.2	*Soymovirus*	309
Tobacco vein-clearing virus	TVCV	AF190123.1	*Solendovirus*	303
Blackberry virus F	BVF	NC_029303.1	*Badnavirus*	314

Dioscorea nummularia associated virus	DNUaV	NC_040712.1	*Dioscovirus*	306
Rose yellow vein virus	RYVV	NC_020999.1	*Rosadnavirus*	313
Rice tungro bacilliform virus	RTBV	X57924.1	*Tungrovirus*	305
Petunia vein clearing virus	PVCV	NC_001839.2	*Petuvirus*	307
Prunus persica virus	PpersV_sc1	Geering *et al.*, 2014 (Supplementary data 1)	*Florendovirus*	309

We selected 278 genome assemblies corresponding to 267 species (Supplementary Data 3 of de Tomás & Vicient, 2022). As can observe in Table 2, two from *Bacteria*, one from *Chromista*, two from *Protozoa*, 13 from *Animal*, six from *Fungi* and 254 from *Plantae* kingdom. *Plantae* kingdom's genomes include three *Rodophyta*, seven *Chlorophyta*, three *Bryophyta*, one *Marchantiophyta* and 240 *Tracheophyta* genomes. *Tracheophyta* includes one *Lycopodiopsida*, four *Pinopsida*, 35 *Liliopsida* (11 families) and 200 *Magnoliopsida* (46 families) genomes. In the literature, the family *Caulimoviridae* is exclusively described in plants. Therefore, 24 genomes outside the *Plantae* kingdom were used as negative controls.

Table 2. The count of analyzed genome assemblies for each phylum.

KINGDOM	PHYLUM	ANALYZED GENOMES
Bacteria	*Firmicutes*	1
	Proteobacteria	1
Chromista	*Miozoa*	1
Protozoa	*Euglenozoa*	1
	Mycetozoa	1
Animalia	*Arthropoda*	6
	Chordata	5
	Mollusca	1
	Nematoda	1
Fungi	*Ascomycota*	5
	Basidiomycota	1
Plantae	*Rhodophyta*	3
	Chlorophyta	7
	Bryophyta	3
	Marchantiophyta	1
	Tracheophyta	240
Total		**278**

We compared the ten RT sequences with the 278 genome assemblies using tBLASTn with default parameters (except -e option set to 1e-10). Only the hits with at least 300 amino acid residues and no stop codons nor frameshifts were selected for further analysis. This criterion was applied to focus on recently integrated RT-EPRVs. To avoid the inclusion in the selection of tandem duplications, we removed a hit if it was located less than 1500 bp to another (Supplementary Data 3 of de Tomás & Vicient, 2022).

For each genome assembly, the selected set of RT sequences were clustered with the 182 RT selected reference domains and those having higher similarity with retrotransposons were removed from the analyses. RT sequences having higher similarity with *Caulimoviridae* were used for further analyses (Supplementary Data 4 of de Tomás & Vicient, 2022).

Classification of the recently integrated RT-EPRVS identified

First of all, we performed a cluster determination. The selected sequences from the genome assemblies were grouped using CD-HIT with a sequence identity cut-off of 60% (Cluster60) or of 100% (Cluster100), a bandwidth of alignment of 20 and a length of sequence to skip of 10. One sequence was then selected to be representative of each cluster60 (Supplementary Data 5 of de Tomás & Vicient, 2022), according to a control manual of the alignment of the different sequences performed with ClustalW (Larkin *et al.*, 2007). Only in the case of cluster60-8, we selected two sequences because the sequences in this cluster were clearly divided in two groups.

The cluster representative sequences were aligned with the representative sequences of episomal or endogenous *Caulimoviridae* (Supplementary Data 1 of de Tomás & Vicient, 2022) using MEGA-X (Kumar *et al.*, 2018). The resulting alignment was then used to build a phylogenetic reconstruction using the maximum likelihood (ML) method and 500 bootstrap replicates using MEGA-X. The resulting tree was then used as a reference to classify the EPRV-RTs found in the genome assemblies.

Calculation of the minimum ages of the integration events

The minimum ages of the integration events reported in this study were inferred by identifying the most distantly related pair of host species sharing a particular cluster60 of EPRVs and applying the estimated species divergence dates in TimeTree (Kumar *et al.*, 2022).

Characterization of the EPRVs structure

For the structure characterization, potential ORFs were predicted using ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/) and the presence of Pfam domains in their encoded polypeptides was confirmed using MOTIF Search (https://www.genome.jp/tools/motif/).

Detection of peach, almond and their F1 hybrid transcription

For the transcription detection, the RNA-seq samples described in Chapter 1 were used. It included 3 replicates each for leaves, flowers, and fruits of an 'Early Gold' peach tree, a 'Texas' almond tree, and their F1 hybrid called 'MB 1.37.' RNA-seq reads from the peach samples were aligned against the peach genome of ‘Lovell’ (Prunus persica Genome v2.0.a1; Verde *et al.*, 2017), and the RNA-seq reads from the almond samples were aligned against the almond genome of ‘Texas’ (Prunus dulcis Texas Genome v2.0; Alioto *et al.*, 2020). In the case of the F1 hybrid, the reads were aligned against both genomes. All the alignments were performed using Bowtie 2 (Langmead & Salzberg, 2012) and visualized using Integrative Genomics Viewer (IGV) (Robinson *et al.*, 2023) to determine the presence of transcription.

3.4. Results

Distribution of genomic sequences encoding Reverse Transcriptase domains of recently inserted Endogenous Pararetroviruses (*Caulimoviridae*)

The objective of the work was to determine the presence of sequences encoding complete conserved RT domains corresponding to Endogenous Pararetrovirus (*Caulimoviridae*) within a collection of 278 publicly available genome sequence assemblies from plant species and using 24 non-plant genome assemblies as negative controls (Table 2; Supplementary Data 3 of de Tomás & Vicient, 2022).

To identify them, we used a custom designed tBLASTn-based discovery pipeline, using as a probe a collection of 10 representative RT sequences of the different *Caulimoviridae* genera and OTUs (Table 1; Supplementary Data 2 of de Tomás & Vicient, 2022). To give priority to the recently inserted copies, we only select sequences encoding RT domains of at least 300 amino acids that contain uninterrupted reading frames. Frequently EPRVs are inserted in tandemly arranged structures. To remove these duplications, when a RT coding region was located less than 1500 bp of another we only kept one of them. Due to their high sequence similarity, this first selection also contained RT sequences from Ty3/gypsy LTR retrotansposons (*Metaviridae*). To remove them, EPRVs were confirmed by phylogenetic analyses. They were aligned with RT sequences of representative *Caulimoviridae* and LTR retrotransposons (Table 1; Supplementary Data 1 of de Tomás & Vicient; 2022). Those sequences showing higher similarity with the *Metaviridae* than with *Caulimoviridae* were removed. Finally, we obtained 11,527 RT-EPRV sequences. These sequences were designated using the first three letters of their genus, followed by the first three letters of their species. In cases where various varieties were analyzed, the initial letter of the variety was added, followed by a hyphen and a number. For instance, PruPerL-01 corresponds to the first sequence of *Prunus persica* 'Lovell'. (Supplementary Data 4 of de Tomás & Vicient, 2022).

Table 3. The count of genome assemblies analyzed for each phylum, along with the number of analyzed genome assemblies containing identified RT-EPRVs.

KINGDOM	PHYLUM	ANALYZED GENOMES	GENOMES WITH RT-EPRVs
Bacteria	*Firmicutes*	1	0
	Proteobacteria	1	0
Chromista	*Miozoa*	1	0
Protozoa	*Euglenozoa*	1	0
	Mycetozoa	1	0
Animalia	*Arthropoda*	6	0
	Chordata	5	0
	Mollusca	1	0
	Nematoda	1	0
Fungi	*Ascomycota*	5	0
	Basidiomycota	1	0
Plantae	*Rhodophyta*	3	0
	Chlorophyta	7	0
	Bryophyta	3	0
	Marchantiophyta	1	0
	Tracheophyta	240	206
Total		**278**	**206**

None of the analyzed genomes outside *Plantae* Kingdom contain RT-EPRV sequences, showing no activity in our negative controls. Among the genomes of the *Plantae* kingdom, we did not find RT-EPRVs in *Chlorophyta*, *Rodophyta*, *Bryophyta* or *Marchantiophyta* (Table 3).

As seen in Table 4, among the *Tracheophyta* species, we did not find RT-EPRVs in the class *Lycopodiopsida* (*Selaginella moellendorffii*), but we found RT-EPRVs in 206 genomes (202 species) of all *Tracheophyta* classes (*Pinopsida*, *Liliopsida* and *Magnoliopsida*), confirming previous results (Gong & Han, 2018). All the four *Pinopsida* genomes analyzed contain RT-EPRV sequences (between 4 in *Pinus glauca* ‘PG29’ and 46 in *Pinus picea*). We included 35 genomes of species of the class *Liliopsida* and we found RT-EPRV sequences in 22 of them (63%) (between 1 in *Paspalum vaginatum* and 63 in *Chasmanthium laxum*). Finally, we found RT-EPRV sequences in 180 of the 200 *Magnaliopsida* genomes (90%) (between 1 in different species as *Arabidopsis helleri* or *Populus tremula* and 1,186 in *Capsidum annum*).

Table 4. The count of genome assemblies analyzed for the different classes of the *Tracheophyta* phylum, along with the number of analyzed genome assemblies containing identified RT-EPRVs and the number of RT-EPRVs for each genome.

PHYLUM	CLASS	ANALYZED GENOMES	GENOMES WITH RT-EPRVs	RT-EPRVs BY GENOME
Tracheophyta	*Lycopodiopsida*	1	0	0
	Pinopsida	4	4	4-46
	Liliopsida	35	22	0-63
	Magnoliopsida	200	180	0-1,186
Total		**240**	**206**	0-1,186

When comparing the results with the genomes of species belonging to the same genus, the results obtained are, in general, similar. For example, the genomes of the two species of *Kalanchoe* contain 20 and 34, the two of *Vitis* contain 24 and 29 and the three of *Solanum* between 29 and 35.

The same happens between varieties of the same species. For example, the genomes of the two varieties of *Pinus glauca* contain 4 and 6, the two of *Cerasus x kanzakura* contain 27 and 38, and the two of *Citrullus lanatus,* 3 and 4.

However, this is not always the case, and we can observe important differences in the number of RT-EPRVs in species of the same genus and species. For example, in the genera *Arachis* (between 56 in *A. ipaensis* and 473 in *A. hypogaea*), *Prunus* (between 3 in *P. dulcis* ‘Lauranne’ and 144 in *P. domestica*), *Rosa* (between 76 in *R. multiflora* and 340 in *R. chinensis*), *Citrus* (between 63 in *C. sinensis* ‘Ridge Pineapple’ and 412 in *C. maxima*) and *Nicotiana* (between 12 in *N. attenuata* and 130 in *N. tabacum*).

We can observe important differences between the number of RT-EPRVs in varieties of the same species too. For example, in *Prunus dulcis*, between 3 in ‘Lauranne’ and 35 in ‘Texas’; in *Citrus sinensis*, between 63 in ‘Ridge Pineapple’ and 108 in ‘Valencia’; in *Malus domestica*, between 16 in ‘Golden Delicious’ and 43 in ‘HFTH1’.

Some of these differences between species of the same genus and between varieties of the same species can be due to differences in the quality of the genome assemblies. For example, the presence of undetermined nucleotides can give rise to a reduction in the number of RT-EPRVs we detected. However, there are cases in which the best quality genome is the one with the least number of sequences. For example, we included three species of the genera *Arabidopsis* and the genome with the least number of sequences (does not have any RT-EPRVs identified), is the one with the best quality (*A. thaliana*), while *A. halleri* has 1, and *A. lyrata* has 22. All these results suggest that in some of the species have been very recent integrations of EPRVs, possibly occurring after speciation.

Classification of the RT-EPRVs present in plant genomes

To provide a classification, RT-EPRV sequences with at least 60% amino acid identity to each other were grouped, yielding a total of 57 clusters. These clusters were listed from Cluster-0 to Cluster 56, ordered by the number of sequences. Cluster 0 was the most populous with 3,207 RT-EPRVs sequences, while Clusters 55 and 56 were the least populous, each containing only 1 RT-EPRV sequence. There are 17 clusters with more than 100 RT-EPRVs (from Cluster 0 to Cluster 16) and the clusters include between 1 and 114 host species. The total number of sequences and genomes represented in each cluster varies greatly. (Table 5).

Table 5. Information of the Cluster60, which includes the Cluster number (Cluster N.), number of RT-EPRVs sequences (N. RT-EPRVs), number of plant classes (N. Classes), number of plant orders (N. Orders), number of plant families (N. Families), number of plant genus (N. Genus), Number of plant species (N. Species), the two most divergent related host species (A and B) and the maximum age of divergence (Max. Age) in Million Years (MY).

Cluster	**Cluster N.**	**N. RT-EPRVs**	**N. Classes**	**N. Orders**	**N. Families**	**N. Genus**	**N. Species**	**A**	**B**	**Max. Age (MY)**
BADNAVIRUS		**80**	**3**	**10**	**11**	**12**	**13**			
Badnavirus-01	20	75	2	9	10	12	12	*Dioscorea*	*Amborella*	191
Badnavirus-02	43	5	1	2	2	2	2	*Phalaenopsis*	*Musa*	117
CAULIMOVIRUS		**38**	**1**	**4**	**4**	**6**	**9**			
Caulimovirus-01	28	36	1	3	3	5	8	*Helianthus*	*Arabidopsis*	118
Caulimovirus-02	52	2	1	1	1	1	1	*Gossypium*	*Gossypium*	0
DIOSCOVIRUS		**144**	**2**	**5**	**5**	**7**	**9**			
Dioscovirus-01	23	49	1	3	3	4	5	*Cynara*	*Cajanus*	118
Dioscovirus-02	25	43	1	1	1	1	2	*Dioscorea*	*Dioscorea*	0
Dioscovirus-03	31	24	1	1	1	2	2	*Glycine*	*Vigna*	23
Dioscovirus-04	34	16	1	1	1	1	1	*Macadamia*	*Macadamia*	0
Dioscovirus-05	35	12	1	1	1	1	2	*Dioscorea*	*Dioscorea*	0
PETUVIRUS		**1693**	**2**	**14**	**16**	**47**	**66**			
Petuvirus-01	1	1202	1	5	5	10	19	*Arachis*	*Citrus*	108

Petuvirus-02	14	131	1	9	9	16	18	*Amborella*	*Helianthus*	191
Petuvirus-03	15	129	1	3	4	6	9	*Coffea*	*Gossypium*	118
Petuvirus-04	19	78	1	1	1	11	13	*Brassica*	*Rorippa*	27
Petuvirus-05	22	52	1	1	1	1	1	*Ipomoea*	*Ipomoea*	0
Petuvirus-06	27	39	1	1	1	7	9	*Arachis*	*Cicer*	59
Petuvirus-07	30	24	1	3	3	4	6	*Populus*	*Gossypium*	108
Petuvirus-08	33	18	1	1	1	3	8	*Citrus*	*Atalantia*	18
Petuvirus-09	36	12	1	2	2	2	2	*Durio*	*Macadamia*	123
Petuvirus-10	39	8	1	1	1	1	1	*Eucalyptus*	*Eucalyptus*	0
SOLENDOVIRUS		**1124**	**1**	**2**	**2**	**5**	**8**			
Solendovirus-01	3	1124	1	2	2	5	8	*Nymphaea*	*Nicotiana*	179
SOYMOVIRUS		**454**	**1**	**5**	**6**	**12**	**14**			
Soymovirus-01	6	391	1	1	1	1	3	*Arachis*	*Arachis*	0
Soymovirus-02	24	49	1	4	5	6	6	*Lactuca*	*Cleome*	118
Soymovirus-03	42	6	1	1	1	1	1	*Chenopodium*	*Chenopodium*	0
Soymovirus-04	44	5	1	1	1	3	3	*Brassica*	*Cakile*	13
Soymovirus-05	48	3	1	1	1	1	1	*Medicago*	*Medicago*	0
TUNGROVIRUS		**308**	**2**	**5**	**5**	**10**	**32**			

Tungrovirus-01	8	251	1	3	3	10	29	*Prunus*	*Vitis*	117
Tungrovirus-02	29	32	1	1	1	1	1	*Lindenbergia*	*Lindenbergia*	0
Tungrovirus-03	38	9	1	1	1	1	1	*Cinnamomum*	*Cinnamomum*	0
Tungrovirus-04	46	4	1	1	1	1	2	*Malus*	*Malus*	0
Tungrovirus-05	54	2	1	1	1	1	1	*Citrus*	*Citrus*	0
FLORENDOVIRUS		**6162**	**3**	**29**	**40**	**91**	**151**			
Florendovirus-01	0	3207	2	27	34	70	114	*Asparagus*	*Amborella*	191
Florendovirus-02	2	1188	1	6	8	21	35	*Brassica*	*Nicotiana*	118
Florendovirus-03	4	949	2	21	27	38	47	*Asparagus*	*Amborella*	191
Florendovirus-04	7	317	1	2	2	2	3	*Coffea*	*Lindenbergia*	77
Florendovirus-05	12	133	1	1	1	3	5	*Arachis*	*Lotus*	59
Florendovirus-06	13	132	1	2	2	5	8	*Lindenbergia*	*Nicotiana*	79
Florendovirus-07	16	120	1	8	9	13	18	*Amborella*	*Brassica*	191
Florendovirus-08	18	79	1	2	2	7	8	*Glycine*	*Manihot*	101
Florendovirus-09	41	7	1	1	2	2	2	*Capsicum*	*Nicotiana*	24
Florendovirus-10	47	4	1	1	1	2	2	*Cucumis*	*Momordica*	48
Florendovirus-11	51	2	2	2	2	2	2	*Asparagus*	*Prunus*	160
GYMNENDOVIRUS		**95**	**1**	**1**	**1**	**2**	**3**			

Gymnendovirus-01	17	95	1	1	1	2	2	*Pinus*	*Picea*	130
WENDOVIRUS		**282**	**1**	**7**	**7**	**10**	**17**			
Wendovirus-01	9	200	1	1	1	4	11	*Citrus*	*Atalantia*	18
Wendovirus-02	21	70	1	2	2	3	3	*Helianthus*	*Coffea*	101
Wendovirus-03	40	7	1	2	2	3	4	*Citrus*	*Solanum*	118
Wendovirus-04	49	3	1	1	1	1	1	*Lindenbergia*	*Lindenbergia*	0
Wendovirus-05	55	1	1	1	1	1	1	*Olea*	*Olea*	0
Wendovirus-06	56	1	1	1	1	1	1	*Portulaca*	*Portulaca*	0
XENDOVIRUS		**65**	**1**	**6**	**6**	**8**	**10**			
Xendovirus-01	26	41	1	4	4	6	8	*Vaccinium*	*Rosa*	118
Xendovirus-02	32	19	1	1	1	1	1	*Olea*	*Olea*	0
Xendovirus-03	45	5	1	1	1	1	1	*Ipomoea*	*Ipomoea*	0
YENDOVIRUS		**334**	**2**	**6**	**7**	**17**	**23**			
Yendovirus-01	10	190	1	1	1	9	11	*Oryza*	*Eleusine*	47
Yendovirus-02	11	142	2	5	5	8	12	*Dioscorea*	*Solanum*	160
Yendovirus-03	50	3	2	2	2	2	2	*Ananas*	*Nymphaea*	179
ZENDOVIRUS		**781**	**1**	**2**	**2**	**5**	**19**			
Zendovirus-01	5	768	1	1	1	4	18	*Fragaria*	*Rubus*	41

Zendovirus-02	37	11	1	1	1	2	4	*Fragaria*	*Rosa*	31
Zendovirus-03	53	2	1	1	1	1	1	*Pistacia*	*Pistacia*	0

We performed a phylogenetic analysis using representative sequences of each cluster groups (Table 6; Supplementary Data 5 of de Tomás & Vicient, 2022) and representatives of all *Caulimoviridae* genera and OTUs (Table 1; Supplementary Data 1 of de Tomás & Vicient, 2022). The Cluster 08 is represented by two different representatives (A and B) due to clear differentiation of the cluster into two separate groups.

Table 6. Representative sequences of each cluster and their corresponding class.

Cluster N.	Model seq	Class	Cluster N.	Model seq	Class
0	CarIll-061	*Florendovirus-01*	**28**	AraHal-1	*Caulimovirus-01*
1	CitMax-156	*Petuvirus-01*	**29**	LinPhi-001	*Tungrovirus-02*
2	CitLim-160	*Florendovirus-02*	**30**	CorCit-01	*Petuvirus-07*
3	SolLyc-23	*Solendovirus-01*	**31**	VigUng-01	*Dioscovirus-03*
4	AmbTri-17	*Florendovirus-03*	**32**	OleEur-03	*Xendovirus-02*
5	FraAna-02	*Zendovirus-01*	**33**	AtaBux-026	*Petuvirus-08*
6	AraDur-07	*Soymovirus-01*	**34**	MacInt-008	*Dioscovirus-04*
7	CofAra-115	*Florendovirus-04*	**35**	DioAla-02	*Dioscovirus-05*
8A	PruArmS-05	*Tungrovirus-01*	**36**	MacInt-079	*Petuvirus-09*
8B	VitRip-01	*Tungrovirus-01*	**37**	RosChi-116	*Zendovirus-02*
9	AtaBux-075	*Wendovirus-01*	**38**	CinKan-04	*Tungrovirus-03*
10	HorVul-01	*Yendovirus-01*	**39**	EucGra-12	*Petuvirus-10*
11	OleEur-01	*Yendovirus-02*	**40**	SolLyc-10	*Wendovirus-03*
12	PhaLun-25	*Florendovirus-05*	**41**	NicTab-032	*Florendovirus-09*
13	CapAnn-0562	*Florendovirus-06*	**42**	CheQui-1	*Soymovirus-03*
14	PisVer-8	*Petuvirus-02*	**43**	PhaEqu-1	*Badnavirus-02*
15	CofAra-051	*Petuvirus-03*	**44**	CakMar-03	*Soymovirus-04*
16	AraHyp-223	*Florendovirus-07*	**45**	IpoNil-01	*Xendovirus-03*
17	PicAbi-20	*Gymnendovirus-01*	**46**	MalSyl-71	*Tungrovirus-04*
18	LotJap-024	*Florendovirus-08*	**47**	MomCha-05	*Florendovirus-10*
19	IsaTin-22	*Petuvirus-04*	**48**	MedTru-1	*Soymovirus-05*
20	AtaBux-025	*Badnavirus-01*	**49**	LinPhi-021	*Wendovirus-04*
21	CofAra-180	*Wendovirus-02*	**50**	AnaCom-1	*Yendovirus-03*
22	IpoTri-55	*Petuvirus-05*	**51**	PruMum-01	*Florendovirus-11*
23	IpoTri-04	*Dioscovirus-01*	**52**	GosHir-01	*Caulimovirus-02*
24	CajCaj-37	*Soymovirus-02*	**53**	PisVer-1	*Zendovirus-03*
25	DioAla-03	*Dioscovirus-02*	**54**	CitMed-040	*Tungrovirus-05*
26	DurZiv-7	*Xendovirus-01*	**55**	OleEur-30	*Wendovirus-05*
27	AraHyp-143	*Petuvirus-06*	**56**	PorAmi-2	*Wendovirus-06*

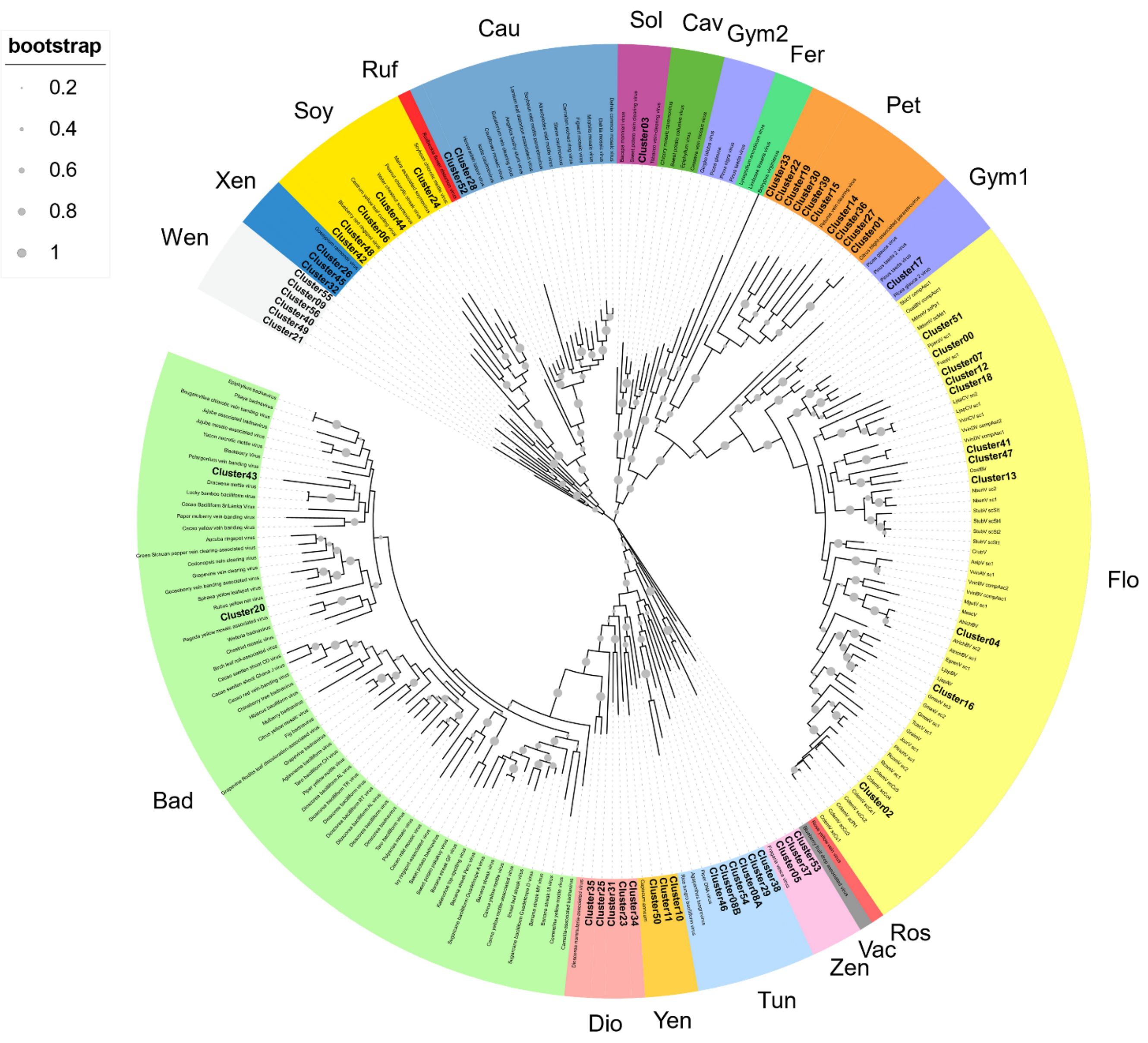
Our phylogenetic analysis clustered together all the previous known sequences corresponding to the same genera and OTU of the *Caulimoviridae*, confirming the robustness of the analysis (Figure 1).

This phylogenetic reconstruction allowed us to determine the diversity and nature of our collection of 11,527 RT-EPRV sequences. They were separated into 13 groups. 30 clusters of these RT-EPRVs were associated with sequences of

Caulimoviridae with episomal forms: 10 *Petuvirus*, 5 *Dioscovirus*, 5 *Soymovirus*, 5 *Tungrovirus*, 2 Badnavirus, 2 *Caulimovirus* and 1 *Solendovirus*. We did not find any representative of the genera *Cavemovirus*, *Rosadnavirus* or *Vaccinivirus*, and neither from the recently proposed genera *Ruflodivirus*. This result suggests that the virus species of these genera do not carry out endogenization, at least not recently or as frequently, or they only do it in a small range of species whose complete genomic sequence is not yet available (Figure 1).

Of the rest, 21 clusters corresponded to OTUs from which only endogenous forms have been found: 11 *Florendovirus*, 3 *Xendovirus*, 3 *Yendovirus*, 3 *Zendovirus* and 1 *Gymnendovirus*. And the remaining 6 clustes were associated with each other, forming a new OTU we called *Wendovirus* (Figure 1).

Figure 1. Phylogenetic relationships within the episomal and endogenous *Caulimoviridae*. Phylogram obtained from a maximum likelihood analysis with protein sequence data from RT conserved domains using 500 bootstrap replications. The size of the point indicated the bootstrap support of the tree branch. Known episomal and Endogenous Pararetrovirus are shown in grey and small letters. New endogenous Clusters60 are shown in bold letters. The color of the branch indicates the genus of *Caulimoviridae*; Bad, *Badnavirus*; Dio, *Dioscovirus*; Yen, *yendovirus*; Tun, *tungrovirus*; Zen, *zendovirus*; Vac, *vaccinivirus*; Ros, *rosadnavirus*; Flo, *florendovirus*; Gym1 and Gym2, *gymnendovirus*1 and 2; Pet, *petuvirus*; Fer, *fernendovirus*; Cav, *cavemovirus*; Sol, *solendovirus*; Cau, *caulimovirus*; Ruf, *ruflodivirus*; Soy, *soymovirus*; Xen, *xendovirus*; and Wen, *wendovirus*.



We observed important differences between genera for both the number of RT-EPRV sequences and the diversity of species in which they were found (Table 5). *Florendovirus* are clearly the most abundant (with 6,162 RT-EPRVs) followed by *Petuvirus* (1,693), *Solendovirus* (1,124) and *Zendovirus* (with 781). However, whereas *Florendovirus* is present in genomes of 40 families of species, *Petuvirus* is present in 16 and *Solendovirus* and *Zendovirus* in only two. Interestingly, although we only detected 80 RT-EPRV sequences corresponding *Badnavirus*, they present a wide distribution (3 Classes, 10 Orders and 11 Families). On the opposite, *Gymnendovirus* are only present in *Pinopsida*.

If we look at the different classes of plants, we observed important differences in Table 7 and Table 8. *Pinopsida* only contains *Gymnendovirus*. *Magnolids* contains *Badnavirus*, *Petuvirus*, *Solendovirus*, *Tungrovirus*, *Florendovirus* and *Yendovirus*. *Liliopsida* contains *Badnavirus*, *Dioscovirus*, *Florendovirus* and *Yendovirus*. Finally, *Magnaliopsida* contains all the genera except *Gymnendovirus*. If we look at the distribution of the clusters in the different plant species, we observed a wide diversity. Some of them are exclusively present in one class. For example, *Gymnendovirus-1* is only present in *Pinopsida*, *Tungrovirus*-3 is only present in Magnolids, *Badnavirus*-2, *Dioscovirus*-2 and -5 and *Yendovirus*-1 are only present in *Liliopsida*, and many clusters are only present in *Magnoliopsida*.

On the opposite, *Badnavirus-1*, *Florendovirus-1* and *Florendovirus-3* are present in *Magnolids*, *Liliopsida* and *Magnoliopsida*. Looking at more detail, 31 of the 57 clusters are present in genomes of only one family of plants, whereas two are present in genomes of more than 20 plant families (both *florendovirus*). These differences of distribution are reflected in the Maximum Age Value (Table 5), which depends on the maximum phylogenetic distance between the species present in the cluster.

Table 7. Distribution of Cluster60 with episomal forms in plant families: *Badnavirus* (Badn), *Caulimovirus* (Caul), *Dioscovirus*, *Petuvirus*, *Solendovirus* (SL), *Soymovirus* and *Tungrovirus*.

■ 1 ■ 2-5 ■ 6-10 ■ 11-40 ■ >40

			BADN		CAUL		DIOSCOVIRUS					PETUVIRUS										SL	SOYMOVIRUS					TUNGROVIRUS				
Class	Order	Family	1	2	1	2	1	2	3	4	5	1	2	3	4	5	6	7	8	9	10	1	1	2	3	4	5	1	2	3	4	5
Pinopsida	Pinales	Pinaceae																														
Liliopsida	Acorales	Acoraceae																														
	Alismatales	Lemnaceae																														
		Zosteraceae																														
	Dioscoreales	Dioscoroceae	1					22			6																					
	Asparagales	Asparagaceae																														
		Orchidaceae		2																												
	Arecales	Arecaceae																														
	Poales	Bromeliaceae																														
		Joinvilleaceae																														
		Poaceae																														
	Zingiberales	Musaceae	5	1																												
Magnoliopsida	Amborellales	Amborellaceae	1										3																			
	Nymphaeales	Nymphaeaceae	10																			6										
	Laurales	Lauraceae																														
	Ranunculales	Papaveraceae																												9		
		Ranunculaceae																														
	Proteales	Nelumbonaceae																														
		Proteaceae								16			27							11												
	Saxifragales	Crassulaceae																														
	Vitales	Vitaceae																										8				
	Celastrales	Celastraceae																														
	Malpighiales	Euphorbiaceae										4																				
		Linaceae																														
		Salicaceae			1													2														
	Fabales	Fabaceae					1		1			2					2						23	1			1					
	Rosales	Rhamnaceae	2																													
		Moraceae										1																				
		Rosaceae	1																									5			1	
	Urticales	Cannabaceae											2																			
	Cucurbitales	Cucurbitaceae																														
	Fagales	Juglandaceae	1																													
		Fagaceae											10																			
		Betulaceae																														
	Myrtales	Lythraceae																						20								
		Myrtaceae										1	2	1				3			4											
	Sapindales	Anacardiaceae											2																			
		Rutaceae	1									1194							2									1				1
	Malvales	Malvaceae				1							1	2				1		1												
	Brassicales	Brassicaceae			1								1		3									1		1						
		Cleomaceae																						3								
		Caricaceae																														
	Caryophyllales	Amaranthaceae																							2							
		Portulacaceae																														
	Cornales	Hydrangeaceae																														
	Ericales	Ericaceae																														
		Theaceae	35																													
	Gentianales	Rubiaceae												47																		
	Solanales	Convolvulaceae					21									26																
		Solanaceae												3								140										
	Lamiales	Lamiaceae																														
		Oleaceae																														
		Pedaliaceae																														
		Phrymaceae																														
	Scrophulariales	Scrophulariaceae																											32			
	Asterales	Asteraceae	3		6		1						3											1								
	Apiales	Apiaceae																														

Table 8. Distribution of Cluster60 with endogenous forms in plant families: *Florendovirus*, *Gymnendovirus* (G), *Wendovirus*, *Xendovirus* (Zendov), *Yendovirus* (Yendov) and *Zendovirus* (Zendov).

■ 1 ■ 2-5 ■ 6-10 ■ 11-40 ■ >40

			FLORENDOVIRUS											G	WENDOVIRUS						XENDOV			YENDOV			ZENDOV		
Class	Order	Family	1	2	3	4	5	6	7	8	9	10	11	1	1	2	3	4	5	6	1	2	3	1	2	3	1	2	3
Pinopsida	Pinales	Pinaceae												24															
Liliopsida	Acorales	Acoraceae	15																										
	Alismatales	Lemnaceae																											
		Zosteraceae																											
	Dioscoreales	Dioscoroceae																							3				
	Asparagales	Asparagaceae	19		40								1																
		Orchidaceae			3																								
	Arecales	Arecaceae	1		17																								
	Poales	Bromeliaceae																							1	2			
		Joinvilleaceae																											
		Poaceae	1																					9					
	Zingiberales	Musaceae																											
Magnoliopsida	Amborellales	Amborellaceae	59		8				10																				
	Nymphaeales	Nymphaeaceae																								1			
	Laurales	Lauraceae	7		1																								
	Ranunculales	Papaveraceae	2		34																								
		Ranunculaceae			38																								
	Proteales	Nelumbonaceae																											
		Proteaceae	70																										
	Saxifragales	Crassulaceae			27																								
	Vitales	Vitaceae	18		1																								
	Celastrales	Celastraceae	1		6																				4				
	Malpighiales	Euphorbiaceae	33	33	17				1	1																			
		Linaceae																											
		Salicaceae		3																									
	Fabales	Fabaceae	24	1	2		8		3	4																			
	Rosales	Rhamnaceae	18																										
		Moraceae																											
		Rosaceae	26		1								1								1						16	1	
	Urticales	Cannabaceae																											
	Cucurbitales	Cucurbitaceae	4		1							1																	
	Fagales	Juglandaceae	49																										
		Fagaceae	408																										
		Betulaceae	63																										
	Myrtales	Lythraceae	30							4																			
		Myrtaceae								20																			
	Sapindales	Anacardiaceae	1		1																								1
		Rutaceae	1	91						1					18		1												
	Malvales	Malvaceae	4							1																			
	Brassicales	Brassicaceae	2	1	1					1																			
		Cleomaceae	3	7																									
		Caricaceae																											
	Caryophyllales	Amaranthaceae	1																										
		Portulacaceae			4															1									
	Cornales	Hydrangeaceae	155		185																14								
	Ericales	Ericaceae	40		36																13								
		Theaceae	73		4																								
	Gentianales	Rubiaceae	2	2		300										2									29				
	Solanales	Convolvulaceae																					3						
		Solanaceae	1	1	2			15			1						1								5				
	Lamiales	Lamiaceae	96		58																								
		Oleaceae																	1			19			31				
		Pedaliaceae			7																								
		Phrymaceae			31																								
	Scrophulariales	Scrophulariaceae	42		3	17		12										3											
	Asterales	Asteraceae	1		15											13													
	Apiales	Apiaceae	9		11																								

Very recent EPRV amplification in plant genomes

The above results suggest that, at least in some species, there has been a recent amplification in the number of EPRV sequences inserted in their genomes. To try to delve further into this aspect, we decided to select those cases in which 100% identical RT-EPRV sequences were present in 10 or more copies in the same genome. Using this highly restrictive criterion, we detected 31 clusters grouping a total of 1,534 sequences (Table 9).

Table 9. Information of the Cluster100, which includes the Cluster number (Cluster 100 N.), number of RT-EPRVs sequences (N. RT-EPRVs), the host plant genome and the RT-EPRV group.

Cluster 100 N.	N. RT-EPRVs	Genome	RT-EPRV Group
1	951	*Capsicum annuum*	*Solendovirus-01*
2	77	*Lotus japonicus*	*Florendovirus-01*
3	53	*Citrus maxima*	*Petuvirus-01*
4	43	*Hydrangea quercifolia*	*Florendovirus-01*
5	27	*Citrus medica*	*Petuvirus-01*
6	26	*Citrus medica*	*Petuvirus-01*
7	24	*Salvia splendens*	*Florendovirus-03*
8	22	*Ipomoea triloba*	*Petuvirus-05*
9	21	*Capsicum annuum*	*Yendovirus-02*
10	20	*Capsicum annuum*	*Florendovirus-03*
11	20	*Atalantia buxifolia*	*Petuvirus-01*
12	19	*Fortunella hindsii*	*Florendovirus-02*
13	19	*Atalantia buxifolia*	*Petuvirus-01*
14	16	*Helianthus annuus*	*Wendovirus-02*
15	16	*Ipomoea triloba*	*Dioscovirus-01*
16	14	*Fortunella hindsii*	*Florendovirus-02*
17	13	*Lactuca sativa*	*Florendovirus-03*
18	12	*Castanea dentata*	*Florendovirus-01*
19	12	*Atalantia buxifolia*	*Florendovirus-02*
20	12	*Nicotiana tabacum*	*Solendovirus-01*
21	12	*Lindenbergia philippensis*	*Tungrovirus-02*
22	11	*Lactuca sativa*	*Caulimovirus-01*
23	11	*Lotus japonicus*	*Florendovirus-01*
24	11	*Atalantia buxifolia*	*Florendovirus-02*
25	11	*Capsicum annuum*	*Florendovirus-03*

26	11	*Capsicum annuum*	*Solendovirus-01*
27	10	*Fragaria nilgerrensis*	*Florendovirus-01*
28	10	*Arachis hypogaea*	*Florendovirus-01*
29	10	*Nicotiana sylvestris*	*Solendovirus-01*
30	10	*Hordeum vulgare*	*Yendovirus-01*
31	10	*Rosa chinensis*	*Zendovirus-01*

These clusters (clusters100) involve 19 genomes. Only one corresponds to a *Liliopsida* (*Hordeum vulgare*) and the remaining 18 are genomic sequences of *Magnaliophyta*. Nine EPRV OTUs are represented in the Clusters100 including *Caulimovirus*, *Dioscovirus*, *Florendovirus*, *Petuvirus*, *Solendovirus*, *Tungrovirus*, *Yendovirus*, *Zendovirus* and the newly proposed *Wendovirus*. Cluster 100-10 is particularly noteworthy as it includes 951 RT-EPRVs sequences present in the genome of pepper (*Capsicum annuum*). Another four groups also correspond to the same genome, with a total of 1,014 sequences (962 are *Solendovirus*, 31 are *Florendovirus* and 21 *Yendovirus*). In total, we found 1,183 RTEPRV sequences in this genome and more than 81% are present in the Cluster 100 selection. This is a very clear indication of a relatively recent proliferation of EPRVs in the pepper genome.

Next, we perform a phylogenetic analysis of representatives of each Cluster-100 and from the described OTUs from *Caulimoviridae* (Figure 2). The sequences of some of the Clusters100 are very similar and, probably, they correspond to the same virus. This is the case of Clusters100-1 and -26 (*Solendovirus* of *Capsicum annuum*), Clusters100-11 and -13 (*Petuvirus* of *Atalantia buxifolia*) and Clusters100-5 and -6 (*Petuvirus* of *Citrus medica*). The sequences of Clusters100-12 and -16 (*Florendovirus* of *Fortunella hindsii*) and of Clusters100-19 and -24 (*Florendovirus* of *Atalantia buxifolia*) are also near identical. The sequences of the Clusters100-20 and 29, that correspond to two different but closely related species (*Nicotiana tabacum* and *Nicotiana sylvestris*), are also almost identical, which suggests that they could come from the same virus capable of infecting both species. Figure 2 also shows that some of the endogenous sequences grouped in Clusters100 are very similar to the sequences of episomal pararetroviruses.

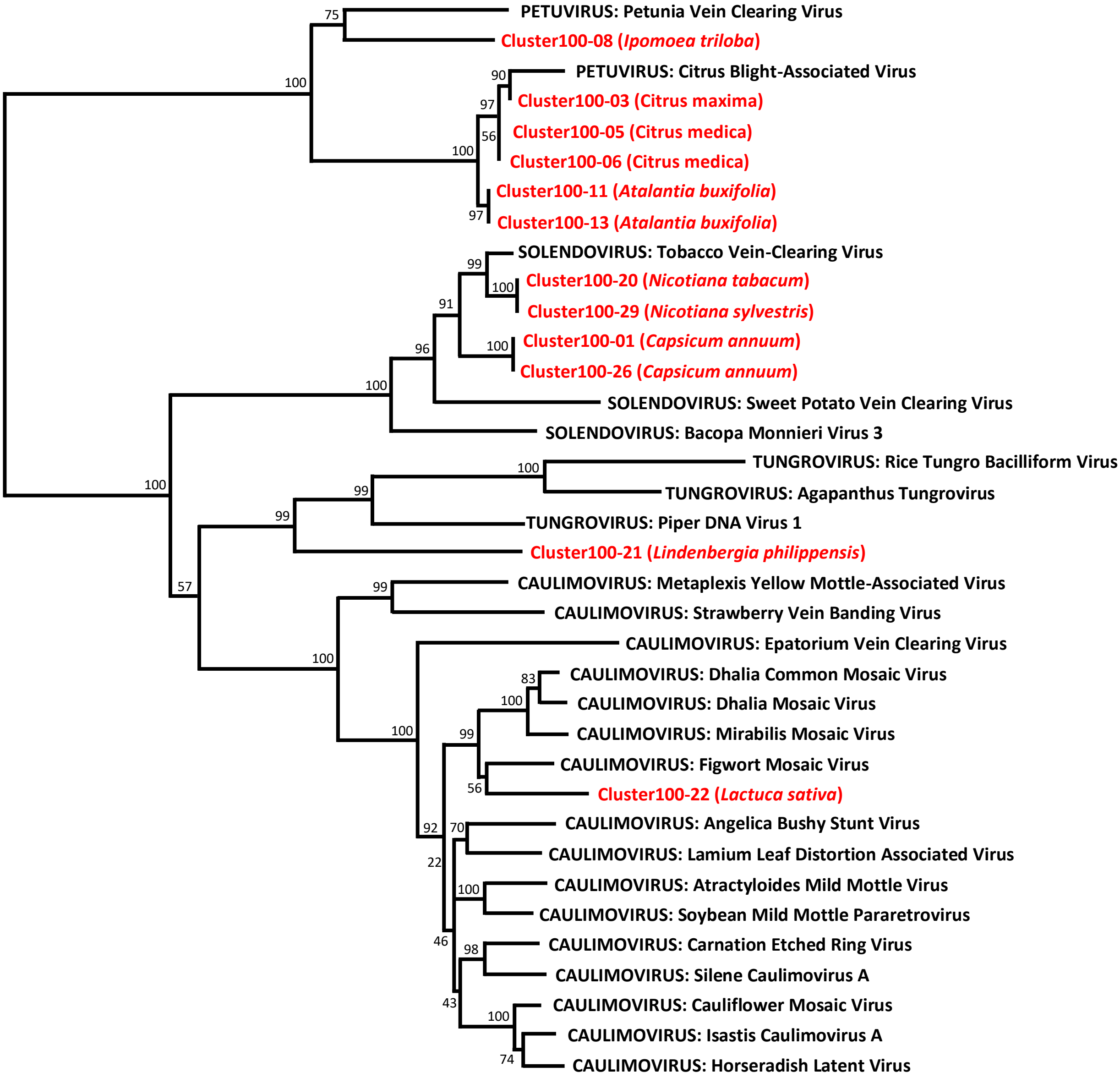

75
PETUVIRUS: Petunia Vein Clearing Virus
Cluster100-08 (Ipomoea triloba)
100
90
PETUVIRUS: Citrus Blight-Associated Virus
97
Cluster100-03 (Citrus maxima)
56
Cluster100-05 (Citrus medica)
100
Cluster100-06 (Citrus medica)
Cluster100-11 (Atalantia buxifolia)
97
Cluster100-13 (Atalantia buxifolia)
SOLENDOVIRUS: Tobacco Vein-Clearing Virus
99
100
Cluster100-20 (Nicotiana tabacum)
91
Cluster100-29 (Nicotiana sylvestris)
96
100
Cluster100-01 (Capsicum annuum)
100
Cluster100-26 (Capsicum annuum)
SOLENDOVIRUS: Sweet Potato Vein Clearing Virus
SOLENDOVIRUS: Bacopa Monnieri Virus 3
100
TUNGROVIRUS: Rice Tungro Bacilliform Virus
100
99
TUNGROVIRUS: Agapanthus Tungrovirus
99
TUNGROVIRUS: Piper DNA Virus 1
Cluster100-21 (Lindenbergia philippensis)
57
99
CAULIMOVIRUS: Metaplexis Yellow Mottle-Associated Virus
CAULIMOVIRUS: Strawberry Vein Banding Virus
100
CAULIMOVIRUS: Epatorium Vein Clearing Virus
83
CAULIMOVIRUS: Dhalia Common Mosaic Virus
100
CAULIMOVIRUS: Dhalia Mosaic Virus
100
99
CAULIMOVIRUS: Mirabilis Mosaic Virus
CAULIMOVIRUS: Figwort Mosaic Virus
56
Cluster100-22 (Lactuca sativa)
92
70
CAULIMOVIRUS: Angelica Bushy Stunt Virus
22
CAULIMOVIRUS: Lamium Leaf Distortion Associated Virus
100
CAULIMOVIRUS: Atractyloides Mild Mottle Virus
CAULIMOVIRUS: Soybean Mild Mottle Pararetrovirus
46
98
CAULIMOVIRUS: Carnation Etched Ring Virus
CAULIMOVIRUS: Silene Caulimovirus A
43
CAULIMOVIRUS: Cauliflower Mosaic Virus
100
CAULIMOVIRUS: Isastis Caulimovirus A
74
CAULIMOVIRUS: Horseradish Latent Virus

Figure 2. Phylogenetic relationships of representative sequences of the Cluster100. Representative sequences of the RT-EPRV Cluster100 (in red) were aligned with RT sequences of pararetroviral elements (in black), and a phylogenetic tree was constructed using the NJ method and 1000 bootstrap replications.

For example, the RT sequence of the citrus blight associated virus is highly similar to the sequences of Cluster100-3, -5 and -6, all of them belonging to genomes of the genus *Citrus*, and the sequence of the tobacco vein clearing virus is similar to Clusters100-20 and -29, belonging to genomes of the genus *Nicotiana*.

Wendovirus, a new group of *Caulimoviridae*

Six of the 60 clusters and one of the 100 clusters correspond to a new group of endogenous *Caulimoviridae* with distinctive characteristics that, following the nomenclature proposed by Diop *et al.* (2018) using the last letters of the alphabet (*Zendovirus*, *Xendovirus* and *Yendovirus*), we have called them *Wendovirus* (Table 5 and Table 9). We were able to reconstruct the structure of the *Wendovirus* for seven genomes corresponding to Cluster 60 (Figure 3; Supplementary Data 6 of de Tomás & Vicient, 2022).

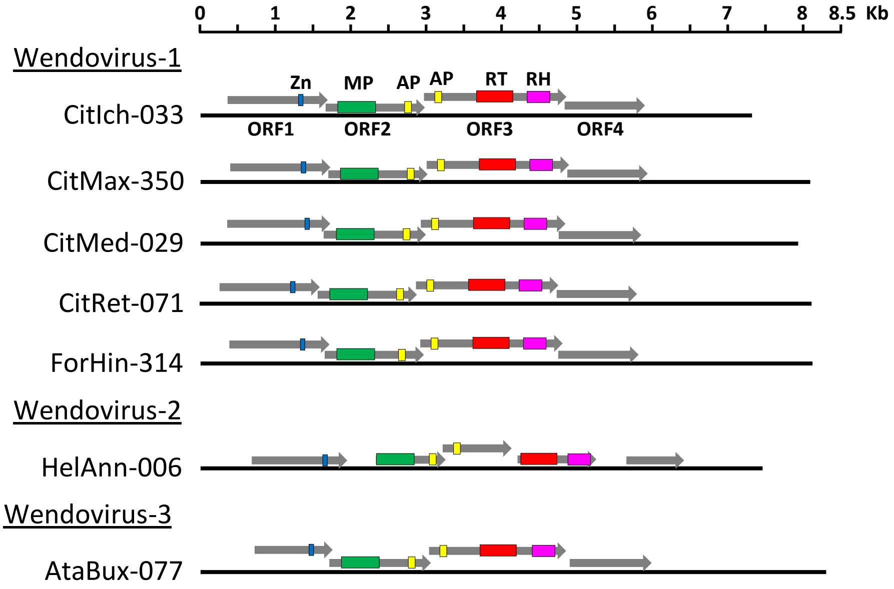


Figure 3. Schematic representation of *wendovirus* Endogenous Pararetrovirus. A scaled linear view of the genome organization of *Wendovirus*. Grey arrows mark open reading frames and colored regions within ORFs are conserved protein domains: blue, zinc finger typically present in the coat proteins; green, Movement Protein; yellow, Aspartic Proteinase; red, Reverse Transcriptase; pink, RNaseH.

The structure was very similar in all of them, with four partially overlapping ORFs. Comparisons with protein motif databases allowed us to find different conserved domains (Supplementary Data 6 of de Tomás & Vicient, 2022). The ORF1 encodes for a zinc finger motif, which is typical of the *Caulimoviridae* coat proteins. The ORF2 encodes for a movement protein and an AP. The ORF3 encodes a second AP, the RT and the RNAseH. Finally, the ORF4 encodes a protein without significant homologies to other reference proteins and without known protein domains but that is well-conserved in all the *wendovirus* elements.

The most noticeable aspect of these structures is the presence of two aspartic proteinase domains instead of one, as usual. They are located close to each other, but in two different ORFs (2 and 3). In the case of the HelAnn-006 element (*Wendovirus-02*), although the domains and their order are conserved, the ORF2 is shorter and the ORF3 is divided in two. The two aspartic proteases are different between them. When compared to databases, the highest similarities of these two aspartic proteinase domains are with members of *Caulimoviridae*. However, the two aspartic proteinases from different *wendoviruses* form a distinct grouping among themselves, separate from APs of other pararetroviruses (Figure 4).

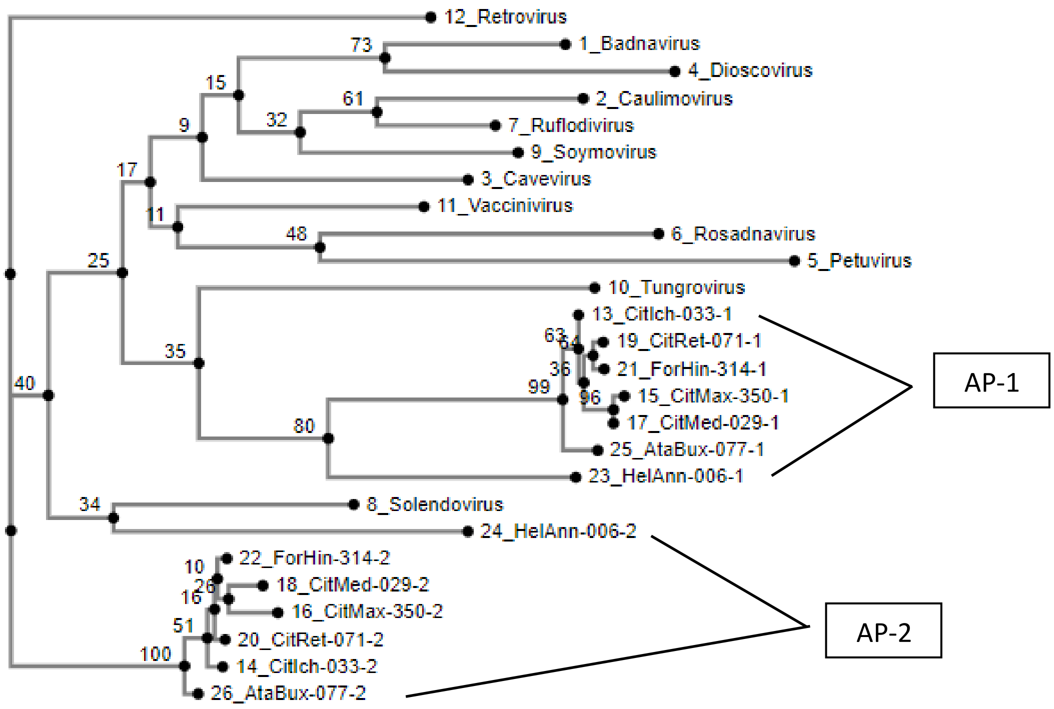


Figure 4. Phylogenetic tree using the NJ method and 1000 replicates with the two Aspartic proteases (AP1 and AP2) of *wendovirus* of seven genomes and Aspartic proteases of different known viruses. Retrovirus is used as outgroup.

On the other hand, given that the classification was conducted based on the RT sequence, the additional MP domain was isolated for the *Wendovirus*es, and we concluded that the MP domains of the *Wendoviruses* exhibited more homology among themselves than with the other members of *Caulimoviridae*, reinforcing the robustness of our study.

Distribution of RT-EPRVs in peach and almond

To analyze the distribution of RT-EPRVs within peach and almond genomes, we investigated the distribution of them across the various chromosomes constituting each genome. We compared this distribution of RT-EPRVs with gene and Transposable Elements annotations. Due to the limited number of RT-EPRVs, we decided to analyze the RT-EPRVs identified using more relaxed filters (with a size of 260 amino acids and the potential inclusion of stop codons). We obtained a total number of 72 RT-EPRVs in the 'Lovell' peach genome and 69 RT-EPRVs in the 'Texas' almond genome.

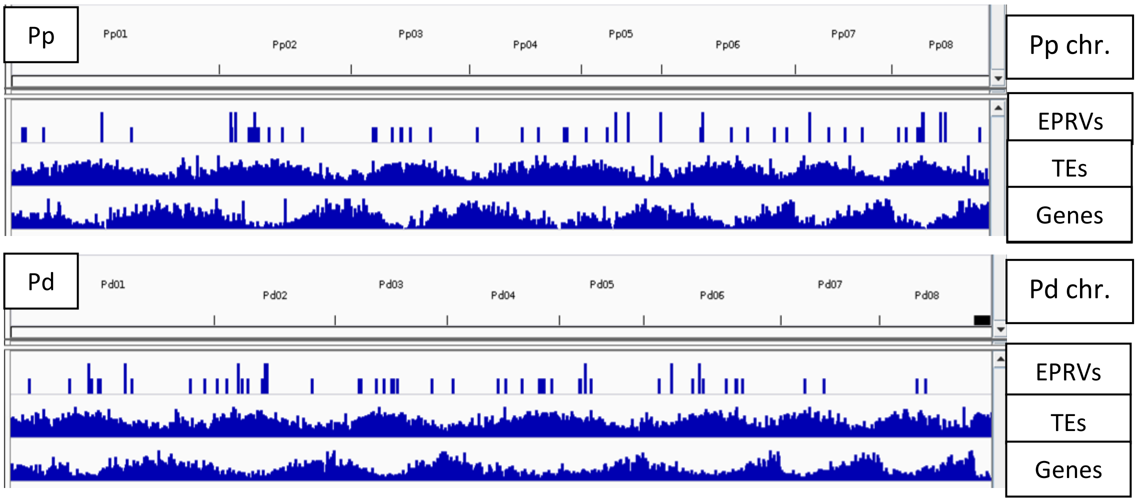


Figure 5. Distribution of Endogenous Pararetroviruses (EPRVs), Transposable Elements (TEs) and Genes across the different chromosomes (chr.) of *Prunus persica* (Pp) and *Prunus dulcis* (Pd).

Although the data's limitations, we did observe a tendency suggesting that EPRVs tend to insert within regions rich in transposable elements and intergenic regions (Figure 5). These results are in line with the literature, that described that EPRVs are usually located in hotspots, which can be unevenly distributed across chromosomes (Vassilieff *et al.*, 2023), particularly in heterochromatin and pericentromeric regions of chromosomes, and close to retrotransposons (Staginnus & Richert-Pöggeler, 2006; Yu *et al.*, 2019).

Transcription of peach, almond and their F1 hybrid

Infrequently, Endogenous Plant Retroviruses exhibit transcriptional activity (Hohn *et al.*, 2008; Gayral *et al.*, 2008). However, as discussed by Vassilieff *et al.* (2023), replication-competent EPRVs (infective EPRVs) have only been reported in interspecific hybrids of the plant genera *Musa*, *Petunia*, and *Nicotiana*. Therefore, we conducted a study involving the transcriptional analysis of 'Early Gold' peach, 'Texas' almond, and their F1 hybrid 'MB 1.37'. The RNA-seq data from replicates of each individual were aligned to the genomes of their respective species ('Lovell' peach and 'Texas' almond), and in the case of the hybrid, against both genomes.

We examined the positions of all identified RT-EPRVs within the 'Lovell' and 'Texas' genomes. Additionally, we analyzed the flanking positions, as only the RT domains had been identified initially. The results were conclusive: among the 38 EPRVs identified in 'Lovell' peach (including 37 classified as *Florendovirus*-1 and 1 as *Tungrovirus*-1), none showed transcriptional activity in 'Early Gold' or the F1 hybrid 'MB 1.37'. Similarly, among the 35 EPRVs identified in the first version of the genome of 'Texas' almond (comprising 34 *Florendovirus*-1 and 1 *Florendovirus*-3), no aligned reads were found, indicating a lack of transcription in both the almond and the F1 hybrid 'MB 1.37'.

3.5. Discussion

Endogenous viral elements (EVEs) are viral sequences integrated in host genomes that are inherited as host DNA sequences (Holmes, 2011). Some of the EVEs, are derived from viruses in which integration into the genome is part of their replication cycle, for example, mammalian retroviruses. However, many viruses in which integration into the genomic DNA is not a part of their normal replication cycle can also be found as EVEs, as is the case of the endogenous *Caulimoviridae* (Endogenous Pararetrovirus, EPRVs). The presence of EPRVs has been described in the genomes of different plant species (Hohn *et al.*, 2008).

In this work we have focused on determining the presence of EPRV sequences relatively recently integrated, based on the selection of elements with complete and conserved RT domains. Based on the RT domain sequence similarity we detected 11,527 sequences distributed in 57 clusters corresponding to 13 OTUs. Twelve of these groups had already been described (Diop *et al.*, 2018) and one is shown here for first time, we called *Wendovirus*. Contrary to what has been observed in other plant viruses as *Geminivirus* or *Nanovirus* (Barreat & Katzourakis, 2021), EVEs from *Caulimoviridae* are exclusively present in plants. Recently integrated RT-EPRVs are present in genomes of *Lycopodiopsida*, *Pinopsida*, *Liliopsida* and *Magnoliopsida*, but not necessary in all the genomes of these groups. For example, they are not present in the genomes of *Arabidopsis thaliana*, *Zea mays*, *Triticum aestivum*, *Phaseolus vulgaris*, *Theobroma cacao* or *Spinacia oleracea*. They are also absent in the *Selaginella moellendorffii* (*Marchantiophyta*) and in *Rhodophyta*, *Chlorophyta* or *Bryophyta*.

We have found that, in some cases, the integration events can be considered very recent. Once in the genome, the EPRV sequences begin to accumulate point random mutations, so, if the sequences are identical that means that they probably integrated recently in the genome. We have found multiple sequences encoding identical RT domains in different species being the most extreme case *Capsicum annuum* in whose genome we found up to 951 sequences encoding identical RT domains. Recent genome integrations of *Caulimoviridae* sequences have been described in some species, such as banana (Gayral *et al.*, 2010). It is interesting to note that, in some cases, these identical RT sequences correspond to groups that have only been detected as endogenous forms (*Florendovirus*, *Yendovirus*, *Zendovirus*, *Wendovirus*) suggesting that probably at least some of them may have their corresponding episomal virus species that have not been yet identified.

The distribution of the different clusters of EPRVs between species shows a great diversity. Some clusters are present exclusively in certain plants as, for example, *Gymnendovirus* in *Pinopsida*, *Zendovirus*-1 in the tribus *Potentilleae* and *Roseae*, *Soymovirus*-1 in the genus *Arachis* or *Wendovirus*-1, only present in *Rutaceae*. In other cases, such as *Florendovirus*-1 and 3, the distribution is very wide, including *Lilipsida* and *Magnoliopsida*. In general, the distribution of the different groups of EPRVs is consistent with the phylogeny, but not always. For example, *Petuvirus*-2 are present in *Amborella trichopoda* and in eight *Magnoliopsida* orders, *Florendovirus*-7 are present in *Amborella trichopoda* and in seven *Magnoliopsida* orders and *Solendovirus*1 are present in *Nymphaea colorata* and in *Solanaceae*.

A possible explanation for these species distributions is the horizontal transmission of the virus between species. There are data suggesting multiple viral jumps between different animal species in *Hepadnavirus* (Dill *et al.*, 2016), and previous data also suggests such horizontal transfers can occur for EPRVs in plants (Diop *et al.*, 2018; Gong and Han, 2018).

We have detected differences in the number of EPRVs in the different genomes. Sometimes the differences are also observed comparing the genomes of species of the same genus or varieties of the same species. The number of EPRVs observed results from the combination of the virus integration and the mechanisms of amplification or reduction of the integrated sequences. First, *Caulimoviridae* integration requires the presence of viruses that are infectious for the species and that the defense mechanisms of the plant are not able to eliminate, or not completely. Second, the main integration mechanism is thought to involve illegitimate recombination, which requires the existence of DNA double-strand breaks and subsequent repair mechanisms (Richert-Pöggeler *et al.*, 2021). Furthermore, to be transmitted, integration must occur in reproductive cells. Third, once integrated, EPRVs, copies are inactivated by sequence degeneration or fragmentation, or by the insertion of transposable elements, and subjected to epigenetic silencing (reviewed by Richert-Pöggeler *et al.*, 2021). All these processes lead to the degeneration of the coding sequences. Finally, it has also been proposed that once integrated, the sequences can be amplified, and different mechanisms have been suggested such as transposition like retroelements, rolling circle amplification, unequal meiotic crossing-over of tandem arrays, or ectopic recombination between EPRVs on non-homologous chromosomes (reviewed by Richert-Pöggeler *et al.*, 2021). Variations in any of these processes together with the time elapsed since the last event of integration could explain the observed differences in the number of EPRVs in the analyzed genomes. Nor can we rule out that the different quality of the genome assemblies may also affect.

We have identified a new putative genus of the *Caulimoviridae*, tentatively named '*Wendovirus*'. *Wendovirus* genomes are about 7,7 Kb long and are present in the genomes of different *Magnaliopsida* species, especially in *Rutaceae* and in sunflower. Our phylogenetic analysis shows that *wendovirus* are related to *Xendovirus* and *Soymovirus*. They contain four ORFs that encode the typical protein domains in *Caulimoviridae*: Zinc-Finger, Movement Protein, Aspartic Proteinase, Reverse Transcriptase and RNAseH. A remarkable feature of *wendovirus* is the presence of two protease coding domains located in two different ORFs (Figure 3). Although both encode aspartyl proteases, the domains are different (PF13975 in ORF2 and PF00077 in ORF3) (Figure 4), so the hypothesis that their origin was a genomic duplication can be discarded. When compared to protein bases, all these described domains, including the two aspartic proteinase domains, show the greatest similarities against other members of *Caulimoviridae*. Therefore, it seems to be ruled out that the second proteinase domain could come from some other families of viruses. Recombination between EPRV fragments has been observed (Chabannes & Iskra-Caruana, 2013) and many viruses have modularly acquired domains and ORFs (Smyshlyaev *et al.*, 2013; Koonin *et al.*, 2015). Encapsidation of genomes (or genome fragments) of different species of *Caulimoviriridae* in the same capsid can lead to recombination and formation of chimeric genomes. Virus-like particles (VLPs) containing host RNAs were found to be produced during agroinfiltration of cucumber necrosis virus, some of them corresponding to retrotransposon or retrotransposon-like RNA sequences (Ghoshal *et al.*, 2015). On the other hand, template switching between two RNA molecules during reverse transcription has been shown for retroviruses, LTR retrotransposons and is proposed for *Caulimoviridae* (Froissart *et al.*, 2005; Tromas *et al.*, 2014; Sanchez *et al.*, 2017; Richert-Pöggeler *et al.*, 2021). Such an acquisition of ORFs likely contributed to the evolution of the *Wendovirus*, although the possible functions of this second proteinase domain remain unknown.

The distribution of EPRVs within peach and almond genomes is notably concentrated in intergenic regions (Figure 5), which are abundant in transposable elements. This observation reinforces the findings of previous analyses. These intergenic regions, as highlighted by Vassilieff *et al.* (2023), appear to serve as safe zones for EPRVs, potentially allowing them to evade elimination mechanisms. This presence might exert a neutral influence on the host, possibly persisting over extended evolutionary periods.

We have not detected any transcriptional activity in either the peach or almond genomes, nor in their F1 hybrid. Replication-competent EPRVs (infective EPRVs) have been exclusively documented in interspecific hybrids of the plant genera *Musa*, *Petunia*, and *Nicotiana*. However, in the context of the current study, as elucidated in Chapter 1 and in de Tomás *et al.* (2022), the crossing of two species so closely related as peach and almond may not induce the *genomic shock* necessary for such activation.

Nonetheless, an interesting challenge for the investigation involves analyzing these species or their hybrids to stress conditions. Stress has been identified as one of the factors that can trigger EPRV activation and transcription (Vassilieff *et al.*, 2023). Therefore, it remains an open question whether subjecting these organisms to abiotic and biotic stressors could lead to EPRV activation and subsequent transcription.

GENERAL DISCUSSION

GENERAL DISCUSSION

Interspersed repeats, such as TEs and EPRVs, are an important source of genetic variability in plants, which can have an impact on crop genomes. In this thesis, we have specifically studied TEs in the genomes of the genus *Prunus*. Specifically, in peach (*P. persica* 'Early Gold'), almond (*P. dulcis* 'Texas'), as well as in an F1 hybrid (*P. persica* x *P.dulcis* 'MB1.37') between these two species. Regarding EPRVs, we have examined them across genomes of different plant groups.

In Chapter 1 I identified different transcriptionally active TE families and genes in peach, almond and their hybrid. My initial hypothesis of a potential *genomic shock* when crossing two different species did not occur in this hybrid. This result differed from studies in other species (Senerchia *et al.*, 2015; Ungerer *et al.*, 2006), but it can be explained considering the relatively short time that both species diverged of these both species, approximately 6 MYA, the mean generation time for these species around 10 years, their high sequence conservation (only 20 nucleotide substitutions per Kbp), and the fact that they share most TE families and many specific insertions (Alioto *et al.*, 2020). These results can be beneficial because the absence of *genomic shock* can facilitate the introgression of alleles from almond into peach through Marker-Assisted Selection, which is already being utilized in breeding programs (Donoso *et al.*, 2016). This is very useful because almond is up to seven times more variable than peach (Velasco *et al.*, 2016) and can serve as a significant source of genetic variability. The major part of the leaf data of this chapter along with the comparative methylation analysis between peach, almond, and the hybrid conducted by Dr. Bardil, which also did not show significant differences between the hybrid and its parents, were included in de Tomás *et al.* (2022) (see Annexes). Furthermore, my transcriptomic results have been robust, considering that they remained consistent across different organs (leaf, flower and fruit).

Chapter 2 is based on the study of transposons using a new version of the Texas almond genome (Texas v.3.0). This genome includes the sequencing of both almond phases due to the advancement in sequencing technologies, such as the use of accurate and long reads (Guk *et al.*, 2022). The fact that there are sequences for both phases is highly beneficial for a considerably heterozygous species like almond. In this thesis, I have improved the detection and annotation of transposons in almond thanks to this new genome with both phases and a better assembly, identifying many transposons that were not detectable in the previous version.

In both Chapter 1 and 2, I have conducted a deep study of transposons in *Prunus*, particularly focusing on transcriptome characterization in peach and almond. While Chapter 1 focuses on the comparison of differential expression between peach and almond and relates it to polymorphic insertions present in one species and absent in the other, Chapter 2 compares the expression of each almond allele (allele-specific expression) and relates it to heterozygous insertions, which are present in one allele but not in the other. The analysis in the Chapter 1 was more limited and preliminary because we had a limited number of polymorphic insertions specially in almond because we used the previous genome sequence version (Texas v.2.0). I was able to identify genes exhibiting differential expression between the two *Prunus* species with polymorphic insertions nearby. Despite the low numbers, my results suggested a trend that genes with the insertion had lower expression compared to those without the insertion. I observed a similar trend in Chapter 2 results, where the allele with the insertion exhibited lower expression than the allele without the insertion. On a global scale in almond, I also observed the impact of transposons when comparing the transcription of genes without transposons nearby, genes with homozygous transposons, and genes with heterozygous transposons. In this analysis, I found that genes with a homozygous insertion in the upstream region had lower expression than those without nearby insertions, which is significant and consistent across all three organs. What's surprising about these results is that when there is a heterozygous insertion nearby, the gene has more expression than when there is an absence of insertion. We justified this by analyzing the orthologous genes in peach (those without an insertion) and observed that these genes had higher transcription levels. So, heterozygous transposons could be inserted into genes with higher transcription. It may be because they are young insertions and have low frequency in the population but may be eliminated in the future through purifying selection.

In Chapter 1, I identified a gene (*Prudul26A016647*) annotated as an ABC transporter that has an insertion in peach not present in almond. This gene also has differential expression between the two species, with almond showing higher expression. This gene is located in the region where the powdery mildew resistance gene *Vr3* has been mapped. It could be a promising candidate, along with the two genes proposed by Marimon *et al.* (2020), for future investigation. These genes can be studied through genetic transformation methods despite the challenge posed by the recalcitrant nature of peach (Zong *et al.*, 2019). Alternatively, close related species of the genus *Prunus* with better transformation rates than peach can be used to test the *Prudul26A016647* like *Prunus domestica* (Siderova *et al.*, 2019) or *Prunus salicina* (Urtubia *et al.*, 2008) In Chapter 2, in Texas v.3.0, gene assembly and annotation have also been improved. Specifically, in this region of potential agricultural and commercial interest, I observed enhanced gene assembly. Additionally, I detected a potential duplication in almond in this region, which could have an impact. Regarding our candidate gene for the *Vr3* gene, it remains in the same genomic position in the new version of the genome, reaffirming my results in Chapter 1. Furthermore, with this new genome that includes both phases, I have confirmed that the polymorphic insertion is not present in almond, not even as heterozygote, which agrees with my PCR results.

In Chapter 3, I studied another type of repetitive element, the EPRVs, by identifying their RT sequences, which are their most conserved domains. I conducted a large-scale analysis on 278 genomes, taking advantage of the increased number of genomes available in the databases. Most of the results are already presented in de Tomás & Vicient, 2022. We identified RT-EPRVs in 202 species within the *Tracheophyta* clade and grouped them into 13 genera. The distribution of different EPRV among species is diverse. *Florendoviruses* generally dominate as the most abundant and widely distributed EPRV, which aligns with Geering *et al.* (2014). Many of these RT-EPRVs were identical, suggesting recent insertions. An interesting case is found in the genome of *Capsicum annuum*, where there are 951 sequences encoding identical RT domains, indicating significant recent activity in this species. Among the 13 genera of EPRVs in which I grouped the identified sequences, one of them was newly proposed, called *Wendovirus*. The elements in this genus contain four ORFs, differing from other *Caulimoviruses* in that two of them encode distinct aspartyl-proteases instead of one, as usual. The two aspartic proteases domains show the greatest similarities against other members of *Caulimoviridae*. Therefore, a recombination between EPRVs could be the origin of this structural feature. Recombination between EPRV fragments has been observed (Chabannes & Iskra-Caruana, 2013) and many viruses have modularly acquired domains and ORFs (Smyshlyaev *et al.*, 2013; Koonin *et al.*, 2015). Encapsidation of genomes (or genomes fragments) of different species of Caulimoviridae in the same capsid can lead to recombination and formation of chimeric genomes (Ghoshal *et al.*, 2015). Template switching between two RNA molecules during reverse transcription has been shown for retroviruses, LTR retrotransposons and is proposed for Caulimoviridae (Froissart *et al.*, 2005; Tromas *et al.*, 2014, Sanchez *et al.*, 2017; Richert-Pöggeler *et al.*, 2021). Such an acquisition of ORFs likely contributed to the evolution of the *Wendovirus*, although the possible functions of this second proteinase domain remain unknown. The proposed genus *Wendovirus* has already been included in a recent review on Endogenous Pararetroviruses (Vassilieff *et al.*, 2023).

In addition to TEs and EPRVs sharing a repetitive nature and sometimes being found in tandem, I have found similarities in their distribution throughout the genomes. In species such as peach and almond, I have observed that both TEs and EPRVs tend to be located in intergenic regions, likely because they do not affect genes and do not result in deleterious alleles, allowing them to evade purifying selection. The analysis of the 278 genomes in Chapter 3 does not include the new almond genome Texas v.3.0. However, we later observed a significantly higher number of identified EPRVs in Texas v.3.0 compared to Texas v.2.0, similar to what happened with the TEs. This is likely due to the improved assembly of pericentromeric regions and to the inclusion of heterozygous elements.

In terms of transcription, while I have identified TE families that are transcriptionally active in peach and almond in Chapter 1, we describe in Chapter 3 the absence of transcription in the EPRVs in the parents and the hybrid. This absence of EPRV activation in the hybrid is consistent with the transcription results in genes and TEs presented in the Chapter 1. This confirms the absence of *genomic shock* between peach and almond, possibly for the same reasons mentioned before. However, as described in Vassilieff *et al.* (2023), EPRVs, like transposons, can become activated under stress conditions, both biotic and abiotic. Therefore, we believe it could be interesting to analyze their transcription in situations of biotic or abiotic stress to better understand their regulatory mechanisms and potential functional roles.

I conclude that both TEs and EPRVs are repetitive elements that can have an impact on genomes. The improvement in technology and genome assemblies, as exemplified by Texas v.3.0 genome, facilitates the study of these elements. Both are a significant source of genetic variability and may be related to traits of agricultural interest. There are many examples of transposons related to commercially relevant traits described in the literature, such as in the case of nectarines (Vendramin *et al.*, 2013). However, EPRVs have not been as extensively studied, although there are examples of their insertion into interesting locus. For instance, the integration of an endogenous *Petuvirus* into a Citrus tristeza virus (CTV) resistance locus in the trifoliate orange genome (Yu *et al.*, 2019) could, in some way, lead to a new phenotype. For this reason, I consider the study of both elements important and in common, because they are related since they share their locations in the genome.

CONCLUSIONS

CONCLUSIONS

1. Our results show that the merging of the ‘Early Gold’ peach and ‘Texas’ almond in MB1.37 F1 hybrid does not result in a genomic shock. There is not a significant change in the transcription of transposable elements and genes in leaves, flowers and fruits. This absence of major changes may facilitate using interspecific peach x almond crosses for peach improvement.
2. Polymorphic transposable elements contribute to the differential gene expression observed between peach and almond.
3. We have identified a gene displaying differential expression and a polymorphic LTR retrotransposon between peach and almond. This gene *Prudul26A016647* is located in the genomic region previously associated with the *Vr3* gene, known for resistance to powdery mildew. Our analysis strongly suggests that this gene is a promising candidate for being the *Vr3* gene.
4. The new version of the almond genome of the variety ‘Texas’, known as Texas v.3.0 is distinguished by having both phases sequenced, unlike the older version of the genome Texas v.2.0. This new genome improves the assembly and the gene annotation of almond.
5. Texas v.3.0 can prove highly valuable for better characterizing regions of agronomic interest, such as the powdery mildew resistance gene *Vr3*. This region is better assembled and gene annotated. We identified a duplication in this region which could potentially play a role in powdery mildew resistance.
6. Transposable Element annotation of Texas v.3.0 encompasses approximately twice the number of complete elements compared with Texas v.2.0. This increase is remarkable in LTR retrotransposons, MITEs, and TIRs.
7. Transposable elements have impact on gene expression in almond. Our analysis demonstrates that genes harboring a homozygous TE insertion in their upstream region generally exhibit lower gene expression levels in almond compared to those lacking such insertions, implying a negative influence on gene expression. And the genes with near heterozygotes

show higher expression than those without insertions because the heterozygous insertions prefer to insert on genes with high expression.

8. 579 genes present allele-specific expression in at least one organ: 82 genes in leaves, 493 in flowers, and 271 in fruits. There are four clusters of co-expressed alleles. Clusters 1 and 2 represented alleles predominantly expressed in all organs, while clusters 3 and 4 indicated genes expressing different alleles in distinct organs.
9. Our analysis of 40 almond cultivars showed a total of 26,487 insertions, corresponding to various transposable elements types. Homozygous insertions were more fixed within the almond population than heterozygous insertions, which are younger and nearer to genes.
10. Endogenous Pararetroviruses (*Caulimoviridae* family) are exclusive elements found within host plant genomes. Our investigation revealed 11,527 RT-EPRVs across 202 species spanning the entire *Tracheophyta* clade. These sequences grouped into 57 clusters and were categorized into 13 OTUs, comprising both episomal and endogenous representatives.
11. The presence of multiple identical RT-EPRVs sequences in certain genomes indicates recent integration events.
12. The distribution of different EPRV clusters among species exhibits considerable diversity. While *Florendoviruses* generally dominate as the most abundant and widely distributed EPRVs, some clusters are detected in specific plants.
13. A new genus, proposed as *Wendovirus*, has been discovered within the *Caulimoviridae* family. *Wendoviruses* are characterized by the presence of four open reading frames encoding typical protein domains seen in *Caulimoviridae*, including two distinct aspartic proteinases.
14. The distribution of EPRVs within peach and almond genomes is non-uniform. Particularly, they tend to integrate within intergenic regions enriched with Transposable Elements.
15. Transcriptional activity of EPRVs is clearly absent within the genomes of both peach and almond, as well as in their F1 hybrid. These findings suggest that the closely related taxonomy of peach and almond may not induce the genomic shock necessary for EPRV activation.

BIBLIOGRAPHY

BIBLIOGRAPHY

Aiewsakun, P., & Katzourakis, A. (2015). Endogenous viruses: Connecting recent and ancient viral evolution. *Virology*, *479–480*, 26–37. doi:10.1016/j.virol.2015.02.011

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., … Arús, P. (2020). Transposons played a major role in the diversification between the closely related almond and peach genomes: results from the almond genome sequence. *The Plant Journal: For Cell and Molecular Biology*, *101*(2), 455–472. doi:10.1111/tpj.14538

Alvarado, Q.H. & González, R.I. (1999). Manual del cultivo del Melocotón. (1st edition, p:38). Guatemala: PROFRUTA-MAGA.7

Andrews, S. (2010). FastQC: A Quality Control Tool for High Throughput Sequence Data [Online]. Available online at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Aranzana, M. J., Illa, E., Howad, W., & Arús, P. (2012). A first insight into peach [Prunus persica (L.) Batsch] SNP variability. *Tree Genetics & Genomes*, *8*(6), 1359–1369. doi:10.1007/s11295-012-0523-6

Arús, P., Aranzana, M. J., Howad, W., Eduardo, I., Donoso, J. M., López-Girona, E., & Serra, O. (2015). The peach genome and its applications. *Acta Horticulturae*, (1100), 29–33. doi:10.17660/actahortic.2015.1100.3

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

Aury, J.-M., & Istace, B. (2021). Hapo-G, haplotype-aware polishing of genome assemblies with accurate reads. *NAR Genomics and Bioinformatics*, *3*(2), lqab034. doi:10.1093/nargab/lqab034

Baird, W.V., Estager, A.S., & Wells, J.K. (1994). Estimating nuclear DNA content in peach and related diploid species using laser flow cytometry and DNA hybridization. *J. Am. Soc. Hortic. Sci., 119*, 1312–1316.

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA*, *6*(1). doi:10.1186/s13100-015-0041-9

Barreat, J. G. N., & Katzourakis, A. (2022). Paleovirology of the DNA viruses of eukaryotes. *Trends in Microbiology*, *30*(3), 281–292. doi:10.1016/j.tim.2021.07.004

Bassi, D., & Monet, R. (2008). *The Peach: Botany, Production and Uses* (D. R. Layne & D. Bassi, Eds.). Oxfordshire: CABI Publishing.

Bejarano, E. R., Khashoggi, A., Witty, M., & Lichtenstein, C. (1996). Integration of multiple repeats of geminiviral DNA into the nuclear genome of tobacco during evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(2), 759–764. doi:10.1073/pnas.93.2.759

Belshaw, R., Pereira, V., Katzourakis, A., Talbot, G., Paces, J., Burt, A., & Tristem, M. (2004). Long-term reinfection of the human genome by endogenous retroviruses. *Proceedings of the National Academy of*

Sciences of the United States of America, *101*(14), 4894–4899. doi:10.1073/pnas.0307800101

Bertsch, C., Beuve, M., Dolja, V. V., Wirth, M., Pelsy, F., Herrbach, E., & Lemaire, O. (2009). Retention of the virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biology Direct*, *4*(1), 21. doi:10.1186/1745-6150-4-21

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*, *30*(15), 2114–2120. doi:10.1093/bioinformatics/btu170

Bouhadida, M., Martín, J. P., Eremin, G., Pinochet, J., Moreno, M. Á., & Gogorcena, Y. (2007). Chloroplast DNA diversity in Prunus and its implication on genetic relationships. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *132*(5), 670–679. doi:10.21273/jashs.132.5.670

Bredemeyer, K. R., Harris, A. J., Li, G., Zhao, L., Foley, N. M., Roelke-Parker, M., … Murphy, W. J. (2021). Ultracontinuous Single Haplotype Genome Assemblies for the Domestic Cat (Felis catus) and Asian Leopard Cat (Prionailurus bengalensis). *The Journal of Heredity*, *112*(2), 165–173. doi:10.1093/jhered/esaa057

Bryant, D. M., Johnson, K., DiTommaso, T., Tickle, T., Couger, M. B., Payzin-Dogru, D., … Whited, J. L. (2017). A tissue-mapped axolotl DE Novo transcriptome enables identification of limb regeneration factors. *Cell Reports*, *18*(3), 762–776. doi:10.1016/j.celrep.2016.12.063

Buggs, R. J. A., Doust, A. N., Tate, J. A., Koh, J., Soltis, K., Feltus, F. A., … Soltis, D. E. (2009). Gene loss and silencing in Tragopogon miscellus (Asteraceae): comparison of natural and synthetic allotetraploids. *Heredity*, *103*(1), 73–81. doi:10.1038/hdy.2009.24

Bushnell, B. (2014). BBMap: A fast, accurate, splice-aware aligner. Department of Computer Science, University of California, Berkeley. URL: https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/

Byrne, D. H. (1990). Isozyme variability in four diploid stone fruits compared with other woody perennial plants. *The Journal of Heredity*, *81*(1), 68–71. doi:10.1093/oxfordjournals.jhered.a110927

Cabanettes, F., & Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ*, *6*(e4958), e4958. doi:10.7717/peerj.4958

Cao, H., Wu, H., Luo, R., Huang, S., Sun, Y., Tong, X., … Wang, J. (2015). De novo assembly of a haplotype-resolved human genome. *Nature Biotechnology*, *33*(6), 617–622. doi:10.1038/nbt.3200

Cao, K., Yang, X., Li, Y., Zhu, G., Fang, W., Chen, C., … Wang, L. (2021). New high-quality peach (*Prunus persica* L. Batsch) genome assembly to analyze the molecular evolutionary mechanism of volatile compounds in peach fruits. *The Plant Journal: For Cell and Molecular Biology*, *108*(1), 281–295. doi:10.1111/tpj.15439

Casacuberta, J. M., & Santiago, N. (2003). Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. *Gene*, *311*, 1–11. doi:10.1016/s0378-1119(03)00557-2

Castanera, R., Morales-Díaz, N., Gupta, S., Purugganan, M., & Casacuberta, J. M. (2023). Transposons are important contributors to gene expression variability under selection in rice populations. *ELife*, *12*. doi:10.7554/elife.86324

Castanera, R., Vendrell-Mir, P., Bardil, A., Carpentier, M.-C., Panaud, O., & Casacuberta, J. M. (2021). Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. *The Plant Journal: For Cell and Molecular Biology*, *107*(1), 118–135. doi:10.1111/tpj.15277

Castel, S. E., Levy-Moonshine, A., Mohammadi, P., Banks, E., & Lappalainen, T. (2015). Tools and best practices for data processing in allelic expression analysis. *Genome Biology*, *16*(1). doi:10.1186/s13059-015-0762-6

Catlin, N. S., & Josephs, E. B. (2022). The important contribution of transposable elements to phenotypic variation and evolution. *Current Opinion in Plant Biology*, *65*(102140), 102140. doi:10.1016/j.pbi.2021.102140

Chabannes, M., & Iskra-Caruana, M.-L. (2013). Endogenous pararetroviruses—a reservoir of virus infection in plants. *Current Opinion in Virology*, *3*(6), 615–620. doi:10.1016/j.coviro.2013.08.012

Chen, C., & Okie, W. R. (2015). Novel peach flower types in a segregating population from 'Helen Borchers.' *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *140*(2), 172–177. doi:10.21273/jashs.140.2.172

Chiba, S., Kondo, H., Tani, A., Saisho, D., Sakamoto, W., Kanematsu, S., & Suzuki, N. (2011). Widespread endogenization of genome sequences of non-retroviral RNA viruses into plant genomes. *PLoS Pathogens*, *7*(7), e1002146. doi:10.1371/journal.ppat.1002146

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*, 34–48. doi:10.1016/j.ympev.2014.02.024

Chu, H., Jo, Y., & Cho, W. K. (2014). Evolution of endogenous non-retroviral genes integrated into plant genomes. *Current Plant Biology*, *1*, 55–59. doi:10.1016/j.cpb.2014.07.002

Comai, L., Madlung, A., Josefsson, C., & Tyagi, A. (2003). Do the different parental 'heteromes' cause genomic shock in newly formed allopolyploids? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *358*(1434), 1149–1155. doi:10.1098/rstb.2003.1305

Cooper, M., Johnson, A. W. (1998). *Poisonous plants and fungi in Britain: Animal and human poisoning*. Norwich, England: Stationery Office Books.

Crow, K. D. & Wagner, G.P. (2006). What is the role of genome duplication in the evolution of complexity and diversity? *Molecular Biology and Evolution*, *23*(5), 887–892. doi:10.1093/molbev/msj083

Dahal, G., Hibino, H., & Aguiero, V. M. (1997). Population characteristics and tungro transmission by Nephotettix virescens (Hemiptera: Cicadellidea) on selected resistant rice cultivars. *Bull Entomol Res*, *87*, 387–395.

D'Amico-Willman, K. M., Ouma, W. Z., Meulia, T., Sideli, G. M., Gradziel, T. M., & Fresnedo-Ramírez, J. (2022). Whole-genome sequence and methylome profiling of the almond [*Prunus dulcis* (Mill.) D.A. Webb] cultivar 'Nonpareil.' *G3 (Bethesda, Md.)*, *12*(5). doi:10.1093/g3journal/jkac065

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., … Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2). doi:10.1093/gigascience/giab008

de Castro Nunes, R., Orozco-Arias, S., Crouzillat, D., Mueller, L. A., Strickler, S. R., Descombes, P., … Guyot, R. (2018). Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions. *Frontiers in Plant Science*, *9*. doi:10.3389/fpls.2018.00175

de Tomás, C. (2016). Introgressió d'un gen de resistència al oïdi procedent del ametller en el presseguer mitjançant la Selecció Assistida per Marcadors (SAM). [Bachelor's Degree Final Project, Universitat de Girona].

de Tomás, C., Bardil, A., Castanera, R., Casacuberta, J.M., & Vicient, C.M. (2022). Absence of major epigenetic and transcriptomic changes accompanying an interspecific cross between peach and almond. *Hortic Res, 9*, uhac127. doi: 10.1093/hr/uhac127.

de Tomás, C., & Vicient, C.M. (2022). Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (Caulimoviridae). *Front Plant Sci, 13*, 1011565. doi: 10.3389/fpls.2022.1011565.

Deniz, Ö., Frost, J. M., & Branco, M. R. (2019). Regulation of transposable elements by DNA modifications. *Nature Reviews. Genetics*, *20*(7), 417–431. doi:10.1038/s41576-019-0106-6

Dill, J. A., Camus, A. C., Leary, J. H., Di Giallonardo, F., Holmes, E. C., & Ng, T. F. F. (2016). Distinct viral lineages from fish and amphibians reveal the complex evolutionary history of hepadnaviruses. *Journal of Virology*, *90*(17), 7920–7933. doi:10.1128/jvi.00832-16

Ding, M., & Chen, Z. J. (2018). Epigenetic perspectives on the evolution and domestication of polyploid plant and crops. *Current Opinion in Plant Biology*, *42*, 37–48. doi:10.1016/j.pbi.2018.02.003

Diop, S. I., Geering, A. D. W., Alfama-Depauw, F., Loaec, M., Teycheney, P.-Y., & Maumus, F. (2018). Tracheophyte genomes keep track of the deep evolution of the Caulimoviridae. *Scientific Reports*, *8*(1). doi:10.1038/s41598-017-16399-x

Donoso, J. M., Picañol, R., Serra, O., Howad, W., Alegre, S., Arús, P., & Eduardo, I. (2016). Exploring almond genetic variability useful for peach improvement: mapping major genes and QTLs in two interspecific almond × peach populations. *Molecular Breeding: New Strategies in Plant Improvement*, *36*(2). doi:10.1007/s11032-016-0441-7

Donoso, José Manuel, Eduardo, I., Picañol, R., Batlle, I., Howad, W., Aranzana, M. J., & Arús, P. (2015). High-density mapping suggests cytoplasmic male sterility with two restorer genes in almond × peach progenies. *Horticulture Research*, *2*(1), 15016. doi:10.1038/hortres.2015.16

Doyle, J.J., & Doyle, J.L. (1990). Isolation of plant DNA from fresh tissue. *Focus*, 12, 13–5

Duitama, J. (2023). Phased Genome Assemblies. In *Methods in Molecular Biology* (pp. 273–286). New York, NY: Springer US.

Eduardo, I., Cantín, C.M., Batlle, I., & Arús, P. (2015). Integración de los marcadores moleculares en un programa de mejora de variedades de melocotonero. *Fruticultura*, 44, 7-17.

Etienne, L. (2017). Paleovirology: looking back in time to better understand and control modern viral infections. *Virologie*, *21*(6), 245–246. doi:10.1684/vir.2017.0715

Facciola, S. (1990). *Cornucopia: A source book of edible plants*. Kampong Publications.

Falchi, R., Vendramin, E., Zanon, L., Scalabrin, S., Cipriani, G., Verde, I., … Morgante, M. (2013). Three distinct mutational mechanisms acting on a single gene underpin the origin of yellow flesh in peach. *The Plant Journal: For Cell and Molecular Biology*, *76*(2), 175–187. doi:10.1111/tpj.12283

Fan, J., Hu, J., Xue, C., Zhang, H., Susztak, K., Reilly, M. P., … Li, M. (2020). ASEP: Gene-based detection of allele-specific expression across individuals in a population by RNA sequencing. *PLoS Genetics*, *16*(5), e1008786. doi:10.1371/journal.pgen.1008786

Fedoroff, N. V. (1989). Maize transposable elements. In M. M. How (Ed.), *Mobile DNA* (Vol. 1, pp. 375–416). American Societv for Microbiolow.

Feschotte, C., & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. *Nature Reviews. Genetics*, *13*(4), 283–296. doi:10.1038/nrg3199

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: where genetics meets genomics. *Nature Reviews. Genetics*, *3*(5), 329–341. doi:10.1038/nrg793

Franck, A., Guilley, H., Jonard, G., Richards, K., & Hirth, L. (1980). Nucleotide sequence of cauliflower mosaic virus DNA. *Cell*, *21*, 285-294.

Finnegan, D. J. (1989). Eukaryotic transposable elements and genome evolution. *Trends in Genetics: TIG*, *5*, 103–107. doi:10.1016/0168-9525(89)90039-5

Foolad, M. R., Arulsekar, S., Becerra, V., & Bliss, F. A. (1995). A genetic map of Prunus based on an interspecific cross between peach and almond. *Theoretical and Applied Genetics*, *91*(2), 262–269. doi:10.1007/bf00220887

Froissart, R., Roze, D., Uzest, M., Galibert, L., Blanc, S., & Michalakis, Y. (2005). Recombination every day: Abundant recombination in a virus during a single multi-cellular host infection. *PLoS Biology*, *3*(3), e89. doi:10.1371/journal.pbio.0030089

Fultz, D., & Slotkin, R. K. (2017). Exogenous transposable elements circumvent identity-based silencing, permitting the dissection of expression-dependent silencing. *The Plant Cell*, *29*(2), 360–376. doi:10.1105/tpc.16.00718

Gafni, A., Calderon, C. E., Harris, R., Buxdorf, K., Dafa-Berger, A., Zeilinger-Reichert, E., & Levy, M. (2015). Biological control of the cucurbit powdery mildew pathogen Podosphaera xanthii by means of the epiphytic fungus Pseudozyma aphidis and parasitism as a mode of action. *Frontiers in Plant Science*, *6*, 132. doi:10.3389/fpls.2015.00132

García-Tejero, I. F., Rubio, A. E., Viñuela, I., Hernández, A., Gutiérrez-Gordillo, S., Rodríguez-Pleguezuelo, C. R., & Durán-Zuazo, V. H. (2018). Thermal imaging at plant level to assess the crop-water status in almond trees (cv. Guara) under deficit irrigation strategies. *Agricultural Water Management*, *208*, 176–186. doi:10.1016/j.agwat.2018.06.002

Gayral, P., Blondin, L., Guidolin, O., Carreel, F., Hippolyte, I., Perrier, X., & Iskra-Caruana, M.-L. (2010). Evolution of endogenous sequences of *banana streak virus*: What can we learn from banana (*Musa* sp.) evolution? *Journal of Virology*, *84*(14), 7346–7359. doi:10.1128/jvi.00401-10

Gayral, P., Noa-Carrazana, J.-C., Lescot, M., Lheureux, F., Lockhart, B. E. L., Matsumoto, T., … Iskra-Caruana, M.-L. (2008). A single *Banana streak virus* integration event in the Banana genome as the origin of infectious endogenous pararetrovirus. *Journal of Virology*, *82*(13), 6697–6710. doi:10.1128/jvi.00212-08

Geering, A. D. W., Pooggin, M. M., Olszewski, N. E., Lockhart, B. E. L., & Thomas, J. E. (2005). Characterisation of Banana streak Mysore virus and evidence that its DNA is integrated in the B genome of cultivated Musa. *Archives of Virology*, *150*(4), 787–796. doi:10.1007/s00705-004-0471-z

Geering, Andrew D. W., Maumus, F., Copetti, D., Choisne, N., Zwickl, D. J., Zytnicki, M., … Teycheney, P.-Y. (2014). Endogenous florendoviruses are major components of plant genomes and hallmarks of virus evolution. *Nature Communications*, *5*(1), 5269. doi:10.1038/ncomms6269

Ghoshal, K., Theilmann, J., Reade, R., Maghodia, A., & Rochon, D. (2015). Encapsidation of Host RNAs by Cucumber Necrosis Virus Coat Protein during both Agroinfiltration and Infection. *Journal of Virology*, *89*(21), 10748–10761. doi:10.1128/jvi.01466-15

Glawe, D. A. (2008). The powdery mildews: a review of the world's most familiar (yet poorly known) plant pathogens. *Annual Review of Phytopathology*, *46*(1), 27–51. doi:10.1146/annurev.phyto.46.081407.104740

Göbel, U., Arce, A. L., He, F., Rico, A., Schmitz, G., & de Meaux, J. (2018). Robustness of transposable element regulation but no genomic shock observed in interspecific Arabidopsis hybrids. *Genome Biology and Evolution*, *10*(6), 1403–1415. doi:10.1093/gbe/evy095

Gong, Z., & Han, G.-Z. (2018). Euphyllophyte paleoviruses illuminate hidden diversity and macroevolutionary mode of Caulimoviridae. *Journal of Virology*, *92*(10). doi:10.1128/jvi.02043-17

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. doi:10.1038/nbt.1883

Gradziel, T. M. (2003). Interspecific hybridations and subsequent gene introgression within prunus subgenus amygalus. *Acta Horticulturae*, *622*, 249–255.

Gradziel, Thomas M. (2020). Redomesticating almond to meet emerging food safety needs. *Frontiers in Plant Science*, *11*. doi:10.3389/fpls.2020.00778

Gradziel, Thomas M. (2022). Transfer of self-fruitfulness to cultivated almond from peach and wild almond. *Horticulturae*, *8*(10), 965. doi:10.3390/horticulturae8100965

Groszmann, M., Greaves, I. K., Albertyn, Z. I., Scofield, G. N., Peacock, W. J., & Dennis, E. S. (2011). Changes in 24-nt siRNA levels in Arabidopsis hybrids suggest an epigenetic contribution to hybrid vigor. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(6), 2617–2622. doi:10.1073/pnas.1019217108

Guardo, M. D., Moretto, M., Moser, M., Catalano, C., Troggio, M., Deng, Z., … Gentile, A. (2021). The haplotype-resolved reference genome of lemon (Citrus limon L. Burm f.). *Tree Genetics & Genomes*, *17*(6). doi:10.1007/s11295-021-01528-5

Guk, J.-Y., Jang, M.-J., Choi, J.-W., Lee, Y. M., & Kim, S. (2022). *De novo* phasing resolves haplotype sequences in complex plant genomes. *Plant Biotechnology Journal*, *20*(6), 1031–1041. doi:10.1111/pbi.13815

Gutzat, R., & Mittelsten Scheid, O. (2012). Epigenetic responses to stress: triple defense? *Current Opinion in Plant Biology*, *15*(5), 568–573. doi:10.1016/j.pbi.2012.08.007

Ha, M., Lu, J., Tian, L., Ramachandran, V., Kasschau, K. D., Chapman, E. J., … Chen, Z. J. (2009). Small RNAs serve as a genetic buffer against genomic shock in Arabidopsis interspecific hybrids and allopolyploids. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17835–17840. doi:10.1073/pnas.0907003106

Han, X., Zhang, Y., Zhang, Q., Ma, N., Liu, X., Tao, W., … He, H. (2023). Two haplotype-resolved, gap-free genome assemblies for Actinidia latifolia and Actinidia chinensis shed light on the regulatory mechanisms of vitamin C and sucrose metabolism in kiwifruit. *Molecular Plant*, *16*(2), 452–470. doi:10.1016/j.molp.2022.12.022

Hansen, C., & Heslop-Harrison, J. S. (2004). Sequences and phylogenies of plant pararetroviruses, viruses, and transposable elements. In *Advances in Botanical Research* (pp. 165–193). Elsevier.

Harper, G., Hull, R., Lockhart, B., & Olszewski, N. (2002). Viral sequences integrated into plant genomes. *Annual Review of Phytopathology*, *40*(1), 119–136. doi:10.1146/annurev.phyto.40.120301.105642

Hegarty, M. J., Barker, G. L., Wilson, I. D., Abbott, R. J., Edwards, K. J., & Hiscock, S. J. (2006). Transcriptome shock after interspecific hybridization in Senecio is ameliorated by genome duplication. *Current Biology: CB*, *16*(16), 1652–1659. doi:10.1016/j.cub.2006.06.071

Heller, D., & Vingron, M. (2021). SVIM-asm: structural variant detection from haploid and diploid genome assemblies. *Bioinformatics (Oxford, England)*, *36*(22–23), 5519–5521. doi:10.1093/bioinformatics/btaa1034

Hirsch, C. D., & Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochimica et Biophysica Acta. Gene Regulatory Mechanisms*, *1860*(1), 157–165. doi:10.1016/j.bbagrm.2016.05.010

Hodel, R., Zimmer, E., & Wen, J. (2021). A phylogenomic approach resolves the backbone of Prunus (Rosaceae) and identifies signals of hybridization and allopolyploidy. *Molecular Phylogenetics and Evolution*, *160*(107118), 107118. doi:10.1016/j.ympev.2021.107118

Hohn, T. (2013). Plant pararetroviruses: interactions of cauliflower mosaic virus with plants and insects. *Current Opinion in Virology*, *3*(6), 629–638. doi:10.1016/j.coviro.2013.08.014

Hohn, T., Richert-Pöggeler, K. R., Staginnus, C., Harper, G., Schwarzacher, T., Teo, C. H., … Hull, R. (2008). Evolution of integrated plant viruses. In *Plant Virus Evolution* (pp. 53–81). Berlin, Heidelberg: Springer Berlin Heidelberg.

Hohn, T., & Rothnie, H. (2013). Plant pararetroviruses: replication and expression. *Current Opinion in Virology*, *3*(6), 621–628. doi:10.1016/j.coviro.2013.08.013

Holmes, E. C. (2011). The evolution of endogenous viral elements. *Cell Host Microbe*, *10*, 368–377. doi:10.1016/j.chom.2011.09.002Holmes

Holt, C., & Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*, *12*(1). doi:10.1186/1471-2105-12-491

Hoyt, S. J., Storer, J. M., Hartley, G. A., Grady, P. G. S., Gershman, A., de Lima, L. G., … O'Neill, R. J. (2022). From telomere to telomere: The transcriptional and epigenetic state of human repeat elements. *Science (New York, N.Y.)*, *376*(6588), eabk3112. doi:10.1126/science.abk3112

Hull, R. (2007, December 21). Caulimoviridae (Plant Pararetroviruses). *ELS*. doi:10.1002/9780470015902.a0000746.pub2

Inglis, P. W., Pappas, M., & Resende, L. V. (2018). Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for highthroughput SNP genotyping and sequencing applications. *PLoS One*, *13*.

Ingram, C. (1948). Ornamental Cherries. *Country Life*. doi:10.2307/4119752

Jáuregui, B., de Vicente, M. C., Messeguer, R., Felipe, A., Bonnet, A., Salesses, G., & Arús, P. (2001). A reciprocal translocation between 'Garfi' almond and 'Nemared' peach. *Theoretical and Applied Genetics*, *102*(8), 1169–1176. doi:10.1007/s001220000511

Jelenkovic, G., & Harrington, E. (1972). Morphology of the pachytene chromosomes in Prunus persica. *Canadian Journal of Genetics and Cytology*, *14*, 317–324.

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*(7659), 524–527. doi:10.1038/nature22971

Jiménez-Ruiz, J., Ramírez-Tejero, J. A., Fernández-Pozo, N., Leyva-Pérez, M. de la O., Yan, H., Rosa, R. de la, … Luque, F. (2020). Transposon activation is a major driver in the genome evolution of cultivated olive trees (*Olea europaea* L.). *The Plant Genome*, *13*(1). doi:10.1002/tpg2.20010

Jung, S., Lee, T., Cheng, C.-H., Buble, K., Zheng, P., Yu, J., … Main, D. (2019). 15 years of GDR: New data and functionality in the Genome Database for Rosaceae. *Nucleic Acids Research*, *47*(D1), D1137–D1145. doi:10.1093/nar/gky1000

Jung, S., Main, D., Staton, M., Cho, I., Zhebentyayeva, T., Arús, P., & Abbott, A. (2006). Synteny conservation between the Prunus genome and both the present and ancestral Arabidopsis genomes. *BMC Genomics*, *7*(1). doi:10.1186/1471-2164-7-81

Kang, J., Park, J., Choi, H., Burla, B., Kretzschmar, T., Lee, Y., & Martinoia, E. (2011). Plant ABC transporters. *The Arabidopsis Book*, *9*, e0153. doi:10.1199/tab.0153

Kapitonov, V. V., & Jurka, J. (2007). Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends in Genetics: TIG*, *23*(10), 521–529. doi:10.1016/j.tig.2007.08.004

Kawahara, Y., de la Bastide, M., Hamilton, J. P., Kanamori, H., McCombie, W. R., Ouyang, S., … Matsumoto, T. (2013). Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. *Rice (New York, N.Y.)*, *6*(1). doi:10.1186/1939-8433-6-4

Kawakami, T., Dhakal, P., & Katterhenry, A. N. (2011). Transposable element proliferation and genome expansion are rare in contemporary sunf lower hybrid populations despite widespread transcriptional activity of LTR retrotransposons. *Genome Biol Evol*, *3*, 156–167.

Kester, D. E., Gradziel, T. M., & Grasselly, C. (1991). ALMONDS (PRUNUS). *Acta Horticulturae*, (290), 701–760. doi:10.17660/actahortic.1991.290.16

Kim, D., Langmead, B., & Salzberg, S. L. (2015). HISAT: a fast spliced aligner with low memory requirements. *Nature Methods*, *12*(4), 357–360. doi:10.1038/nmeth.3317

Kiselev, K.V., Aleynova, O.A., Ogneva, Z.V., V., Suprun, A.R., & Dubrovina, A. S. (2021). 35S promoter-driven transgenes are variably expressed in different organs of Arabidopsis thaliana and in response to abiotic stress. *Molecular Biology Reports*. doi:10.1007/s11033-021-06235-x

Kobayashi, S., Goto-Yamamoto, N., & Hirochika, H. (2004). Retrotransposon-induced mutations in grape skin color. *Science (New York, N.Y.)*, *304*(5673), 982–982. doi:10.1126/science.1095011

Kofler, R., Gómez-Sánchez, D., & Schlötterer, C. (2016). PoPoolationTE2: Comparative population genomics of transposable elements using pool-seq. *Molecular Biology and Evolution*, *33*(10), 2759–2764. doi:10.1093/molbev/msw137

Koonin, E. V., Dolja, V. V., & Krupovic, M. (2015). Origins and evolution of viruses of eukaryotes: The ultimate modularity. *Virology*, *479–480*, 2–25. doi:10.1016/j.virol.2015.02.039

Koonin, E. V., Krupovic, M., & Agol, V. I. (2021). The Baltimore classification of viruses 50 years later: How does it stand in the light of virus evolution? *Microbiology and Molecular Biology Reviews: MMBR*, *85*(3). doi:10.1128/mmbr.00053-21

Krattinger, S. G., Lagudah, E. S., Spielmeyer, W., Singh, R. P., Huerta-Espino, J., McFadden, H., … Keller, B. (2009). A putative ABC transporter confers durable resistance to multiple fungal pathogens in wheat. *Science (New York, N.Y.)*, *323*(5919), 1360–1363. doi:10.1126/science.1166453

Kronenberg, Z. N., Rhie, A., Koren, S., Concepcion, G. T., Peluso, P., Munson, K. M., … Kingan, S. B. (2021). Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nature Communications*, *12*(1). doi:10.1038/s41467-020-20536-y

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Molecular Biology and Evolution*, *35*(6), 1547–1549. doi:10.1093/molbev/msy096

Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., … Hedges, S. B. (2022). TimeTree 5: An expanded resource for species divergence times. *Molecular Biology and Evolution*, *39*(8). doi:10.1093/molbev/msac174

Kuriyama, K., Tabara, M., Moriyama, H., Kanazawa, A., Koiwa, H., Takahashi, H., & Fukuhara, T. (2020). Disturbance of floral colour pattern by activation of an endogenous pararetrovirus, petunia vein clearing virus, in aged petunia plants. *The Plant Journal: For Cell and Molecular Biology*, *103*(2), 497–511. doi:10.1111/tpj.14728

Lambert, P. (2018). *Pest and pathogen resistance in peach. Organic Farming Reasearch and Perspectives (ORGANIST)*. Milan, Italy.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. doi:10.1038/nmeth.1923

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., … Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, *23*(21), 2947–2948. doi:10.1093/bioinformatics/btm404

Leibman, D., Kravchik, M., Wolf, D., Haviv, S., Weissberg, M., Ophir, R., … Gal-On, A. (2018). Differential expression of cucumber RNA-dependent RNA polymerase 1 genes during antiviral defence and resistance. *Molecular Plant Pathology*, *19*(2), 300–312. doi:10.1111/mpp.12518

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(Database), D19–D21. doi:10.1093/nar/gkq1019

Li, B., & Dewey, C. N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, *12*(1). doi:10.1186/1471-2105-12-323

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Retrieved from http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics (Oxford, England)*, *34*(18), 3094–3100. doi:10.1093/bioinformatics/bty191

Li, X.-W., Meng, X.-Q., Jia, H.-J., Yu, M.-L., Ma, R.-J., Wang, L.-R., … Aranzana, M. J. (2013). Peach genetic resources: diversity, population structure and linkage disequilibrium. *BMC Genetics*, *14*(1), 84. doi:10.1186/1471-2156-14-84

Lian, X., Zhang, H., Jiang, C., Gao, F., Yan, L., Zheng, X., … Feng, J. (2022). De novo chromosome-level genome of a semi-dwarf cultivar of Prunus persica identifies the aquaporin PpTIP2 as responsible for temperature-sensitive semi-dwarf trait and PpB3-1 for flower type and size. *Plant Biotechnology Journal*, *20*(5), 886–902. doi:10.1111/pbi.13767

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, *30*(7), 923–930. doi:10.1093/bioinformatics/btt656

Lockhart, B. E., Dahal, G., Menke, J., & Olszewski, N. E. (2000). Characterization and genomic analysis of tobacco vein clearing virus, a plant pararetrovirus that is transmitted vertically and related to sequences integrated in the host genome. *The Journal of General Virology*, *81*(6), 1579–1585. doi:10.1099/0022-1317-81-6-1579

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. doi:10.1186/s13059-014-0550-8

Lower, S. E., Dion-Côté, A.-M., Clark, A. G., & Barbash, D. A. (2019). Special Issue: Repetitive DNA sequences. *Genes*, *10*(11), 896. doi:10.3390/genes10110896

Marimon, N., Luque, J., Arús, P., & Eduardo, I. (2020). Fine mapping and identification of candidate genes for the peach powdery mildew resistance gene Vr3. *Horticulture Research*, *7*(1). doi:10.1038/s41438-020-00396-9

Markle, G. M. (1998). *Food and feed crops of the United States*.

Martínez-Gómez, P., Rubio, M., Dicenta, F., & Gradziel, T. M. (2004). Resistance to plum pox virus (Dideron isolate RB3.30) in a group of California almonds and transfer of resistance to peach. *Journal of the American Society for Horticultural Science. American Society for Horticultural Science*, *129*(4), 544–548. doi:10.21273/jashs.129.4.0544

Martinière, A., Bak, A., Macia, J.-L., Lautredou, N., Gargani, D., Doumayrou, J., … Drucker, M. (2013). A virus responds instantly to the presence of the vector on the host and forms transmission morphs. *ELife*, *2*, e00183. doi:10.7554/eLife.00183

Mason, A. S., & Batley, J. (2015). Creating new interspecific hybrid and polyploid crops. *Trends in Biotechnology*, *33*(8), 436–441. doi:10.1016/j.tibtech.2015.06.004

McCLINTOCK, B. (1950). The origin and behavior 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.36.6.344

McClintock, B. (1984). The significance of responses of the genome to challenge. *Science (New York, N.Y.)*, *226*(4676), 792–801. doi:10.1126/science.15739260

Mcguffin, M., Kartesz, J. T., Leung, A. Y., & Tucker, A. O. (2000). *Herbs of commerce*. Silver Spring, Maryland: American Herbal Products Association.

Mehrotra, S. (2014). Goyal Repetitive sequences in plant nuclear DNA: types, distribution, evolution and function Genom. *Proteom. Bioinf*, *12*, 164–171.

Melters, D. P., Bradnam, K. R., Young, H. A., Telis, N., May, M. R., Ruby, J., … Chan, S. W. L. (2013). Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biology*, *14*(1), R10. doi:10.1186/gb-2013-14-1-r10

Mhiri, C., Parisod, C., Daniel, J., Petit, M., Lim, K. Y., Dorlhac de Borne, F., … Grandbastien, M.-A. (2019). Parental transposable element loads influence their dynamics in young *Nicotiana* hybrids and allotetraploids. *The New Phytologist*, *221*(3), 1619–1633. doi:10.1111/nph.15484

Michael, T. P., & VanBuren, R. (2020). Building near-complete plant genomes. *Current Opinion in Plant Biology*, *54*, 26–33. doi:10.1016/j.pbi.2019.12.009

Micheletti, D., Dettori, M. T., Micali, S., Aramini, V., Pacheco, I., Da Silva Linge, C., … Aranzana, M. J. (2015). Whole-genome analysis of diversity and SNP-major gene association in peach germplasm. *PloS One*, *10*(9), e0136803. doi:10.1371/journal.pone.0136803

Mnejja, M., Garcia-Mas, J., Audergon, J.-M., & Arús, P. (2010). Prunus microsatellite marker transferability across rosaceous crops. *Tree Genetics & Genomes*, *6*(5), 689–700. doi:10.1007/s11295-010-0284-z

Moreno, A., Hébrard, E., Uzest, M., Blanc, S., & Fereres, A. (2005). A single amino acid position in the helper component of Cauliflower mosaic virus can change the spectrum of transmitting vector species. *Journal of Virology*, *79*(21), 13587–13593. doi:10.1128/jvi.79.21.13587-13593.2005

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., … Macas, J. (2011). Plant centromeric retrotransposons: a structural and cytogenetic perspective. *Mobile DNA*, *2*(1), 4. doi:10.1186/1759-8753-2-4

Ni, Z., Kim, E.-D., Ha, M., Lackey, E., Liu, J., Zhang, Y., … Chen, Z. J. (2009). Altered circadian rhythms regulate growth vigour in hybrids and allopolyploids. *Nature*, *457*(7227), 327–331. doi:10.1038/nature07523

Nieto Feliner, G., Casacuberta, J., & Wendel, J. F. (2020). Genomics of evolutionary novelty in hybrids and polyploids. *Frontiers in Genetics*, *11*. doi:10.3389/fgene.2020.00792

Okie, W. (2003). *Encyclopedia of Fruits and Nuts. C A B Intl* (J. Paulii, Ed.).

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), 3837. doi:10.3390/ijms20153837

Ortega Sada, J. L. (1999). *Flora de Interes Apicola y Polinizacion de Cultivo*. Mundiprensa.

Ou, S., Su, W., Liao, Y., Chougule, K., Agda, J. R. A., Hellinga, A. J., … Hufford, M. B. (2019). Benchmarking Transposable Element Annotation Methods for Creation of a Streamlined, Comprehensive Pipeline. *Genome Biol*, *20*(1).

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

Pacheco, I., Eduardo, I., Rossini, L., Vecchieti, A., Bassi, D.(2009). Qtl mapping for peach (Prunus Persica l. Batsch) resistance to powdery mildew and brown rot. Retrieved July 16, 2023, from Geneticagraria.it website: http://www.geneticagraria.it/attachment/SIGA_2009/5_09.pdf

Parisod, C., Salmon, A., Zerjal, T., Tenaillon, M., Grandbastien, M.-A., & Ainouche, M. (2009). Rapid structural and epigenetic reorganization near transposable elements in hybrid and allopolyploid genomes in *Spartina*. *The New Phytologist*, *184*(4), 1003–1015. doi:10.1111/j.1469-8137.2009.03029.x

Pascal, T., Aberlenc, R., Confolent, C., Hoerter, M., Lecerf, E., Tuéro, C., & Lambert, P. (2017). Mapping of new resistance (Vr2, Rm1) and ornamental (Di2, pl) Mendelian trait loci in peach. *Euphytica; Netherlands Journal of Plant Breeding*, *213*(6). doi:10.1007/s10681-017-1921-5

Pascal, T., Pfeiffer, F., & Kervella, J. (2010). Powdery mildew resistance in the peach cultivar Pamirskij 5 is genetically linked with the gr gene for leaf color. *HortScience: A Publication of the American Society for Horticultural Science*, *45*(1), 150–152. doi:10.21273/hortsci.45.1.150

Passos, D. O., Li, M., Craigie, R., & Lyumkis, D. (2021). Retroviral integrase: Structure, mechanism, and inhibition. In *Viral Replication Enzymes and their Inhibitors Part B* (pp. 249–300). Elsevier.

Paz, R. C., Rendina González, A. P., Ferrer, M. S., & Masuelli, R. W. (2015). Short-term hybridisation activates Tnt1 and Tto1 Copia retrotransposons in wild tuber-bearing*Solanum*species. *Plant Biology (Stuttgart, Germany)*, *17*(4), 860–869. doi:10.1111/plb.12301

Pellicer, J., & Leitch, I. J. (2020). The Plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. *The New Phytologist*, *226*(2), 301–305. doi:10.1111/nph.16261

Pérez de los Cobos, F., Martínez-García, P. J., Romero, A., Miarnau, X., Eduardo, I., Howad, W., … Batlle, I. (2021). Pedigree analysis of 220 almond genotypes reveals two world mainstream breeding lines based on only three different cultivars. *Horticulture Research*, *8*(1). doi:10.1038/s41438-020-00444-4

Piet, Q., Droc, G., Marande, W., Sarah, G., Bocs, S., Klopp, C., … Charron, C. (2022). A chromosome-level, haplotype-phased Vanilla planifolia genome highlights the challenge of partial endoreplication for accurate whole-genome assembly. *Plant Commun*, *3*.

Potter T Eriksson R, D. C., Evans, S., Oh, J. E. E., Smedmark D, R., Morgan, M., Kerr K, R., … Dickinson C, S. (2007). *Phylogeny and classification of Rosaceae*. *266*, 5–43. doi:10.1007/s00606-007-0539-9

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, *26*(6), 841–842. doi:10.1093/bioinformatics/btq033

Rafel Socias i Company, & Gradziel, T. M. (2017). *Almonds: Botany, production and uses*. Cabi.

Rajaraman, J., Douchkov, D., Lück, S., Hensel, G., Nowara, D., Pogoda, M., … Patrick Schweizer. (2018). Evolutionarily conserved partial gene duplication in the Triticeae tribe of grasses confers pathogen resistance. *Genome Biology*, *19*(1). doi:10.1186/s13059-018-1472-7

Rajendrakumar, P., Hariprasanna, K., & Seetharama, N. (2015). Prediction of heterosis in crop plants – status and prospects. *American Journal of Experimental Agriculture*, *9*(3), 1–16. doi:10.9734/ajea/2015/19263

Rao, S. R., Trivedi, S., Emmanuel, D., Merita, K., & Hynniewta, M. (2010). DNA repetitive sequences-types, distribution and function: a review. *J Cell Mol Biol*, *7*(2), 1–11.

Rehder, A. (1940). *Manual of cultivated trees and shrubs hardy in north America*. Portland, OR: Timber Press.

Ricciuti, E., Laboureau, N., Noumbissié, G., Chabannes, M., Sukhikh, N., Pooggin, M. M., & Iskra-Caruana, M.-L. (2021). Extrachromosomal viral DNA produced by transcriptionally active endogenous viral elements in non-infected banana hybrids impedes quantitative PCR diagnostics of banana streak virus infections in banana hybrids. *The Journal of General Virology*, *102*(11). doi:10.1099/jgv.0.001670

Richards, E. J., & Ausubel, F. M. (1988). Isolation of a higher eukaryotic telomere from arabidopsis thaliana. *Cell*, *53*, 127–136.

Richert-Pöggeler, K. R., Vijverberg, K., Alisawi, O., Chofong, G. N., Heslop-Harrison, J. S. (pat), & Schwarzacher, T. (2021). Participation of multifunctional RNA in replication, recombination and regulation of endogenous plant pararetroviruses (EPRVs). *Frontiers in Plant Science*, *12*. doi:10.3389/fpls.2021.689307

Robinson, J. T., Thorvaldsdottir, H., Turner, D., & Mesirov, J. P. (2023). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *Bioinformatics (Oxford, England)*, *39*(1). doi:10.1093/bioinformatics/btac830

Rothkegel, K., Espinoza, A., Sanhueza, D., Lillo-Carmona, V., Riveros, A., Campos-Vargas, R., & Meneses, C. (2021). Identification of DNA methylation and transcriptomic profiles associated with fruit mealiness in Prunus persica (L.) batsch. *Frontiers in Plant Science*, *12*. doi:10.3389/fpls.2021.684130

Ru, S., Main, D., Evans, K., & Peace, C. (2015). Current applications, challenges, and perspectives of marker-assisted seedling selection in Rosaceae tree fruit breeding. *Tree Genetics & Genomes*, *11*(1). doi:10.1007/s11295-015-0834-5

Sanchez, D. H., Gaubert, H., Drost, H.-G., Zabet, N. R., & Paszkowski, J. (2017). High-frequency recombination between members of an LTR retrotransposon family during transposition bursts. *Nature Communications*, *8*(1). doi:10.1038/s41467-017-01374-x

Sánchez-Pérez, R., Pavan, S., Mazzeo, R., Moldovan, C., Aiese Cigliano, R., Del Cueto, J., … Møller, B. L. (2019). Mutation of a bHLH transcription factor allowed almond domestication. *Science (New York, N.Y.)*, *364*(6445), 1095–1098. doi:10.1126/science.aav8197

Schnable, P. , & Wise, R.P.(1998). The molecular basis of cytoplasmic male sterility and fertility restoration. *Trends in Plant Science*, *3*(5), 175–180. doi:10.1016/s1360-1385(98)01235-7

Schnable, P. S., Ware, D., Fulton, R. S., Stein, J. C., Wei, F., Pasternak, S., … Wilson, R. K. (2009). The B73 maize genome: Complexity, diversity, and dynamics. *Science (New York, N.Y.)*, *326*(5956), 1112–1115. doi:10.1126/science.1178534

Scorza, R., Mehlenbacher, S. A., & Lightner, G. W. (1985). Inbreeding and coancestry of freestone peach cultivars of the Eastern United States and implications for peach germplasm improvement. *Journal of the American Society for Horticultural Science*, *110*, 547–552.

Scorza, R., & Okie, W. R. (1991). PEACHES (PRUNUS). *Acta Horticulturae*, (290), 177–234. doi:10.17660/actahortic.1991.290.5

Senerchia, N., Felber, F., & Parisod, C. (2015). Genome reorganization in F1 hybrids uncovers the role of retrotransposons in reproductive isolation. *Proceedings. Biological Sciences*, *282*(1804), 20142874. doi:10.1098/rspb.2014.2874

Shen, Y., Li, W., Zeng, Y., Li, Z., Chen, Y., Zhang, J., … Jiao, Y. (2022). Hong-Bin Wang . Chromosome-level and haplotype-resolved genome provides insight into the tetraploid hybrid origin of patchouli. *Nat Commun*, *13*(1).

Shumate, A., & Salzberg, S. L. (2021). Liftoff: accurate mapping of gene annotations. *Bioinformatics (Oxford, England)*, *37*(12), 1639–1643. doi:10.1093/bioinformatics/btaa1016

Sidorova, T., Mikhailov, R., Pushin, A., Miroshnichenko, D., & Dolgov, S. (2019). Agrobacterium-mediated transformation of Russian commercial plum cv. “startovaya” (Prunus domestica L.) with virus-derived hairpin RNA construct confers durable resistance to PPV infection in mature plants. *Frontiers in Plant Science*, *10*. doi:10.3389/fpls.2019.00286

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, *31*(19), 3210–3212. doi:10.1093/bioinformatics/btv351

Smyshlyaev, G., Voigt, F., Blinov, A., Barabas, O., & Novikova, O. (2013). Acquisition of an Archaea-like ribonuclease H domain by plant L1 retrotransposons supports modular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(50), 20140–20145. doi:10.1073/pnas.1310958110

Song, Q., Zhang, T., Stelly, D. M., & Chen, Z. J. (2017). Epigenomic and functional analyses reveal roles of epialleles in the loss of photoperiod sensitivity during domestication of allotetraploid cottons. *Genome Biology*, *18*(1). doi:10.1186/s13059-017-1229-8

Staginnus, C., & Richert-Pöggeler, K. R. (2006). Endogenous pararetroviruses: two-faced travelers in the plant genome. *Trends in Plant Science*, *11*(10), 485–491. doi:10.1016/j.tplants.2006.08.008

Sun, X., Jiao, C., Schwaninger, H., Chao, C. T., Ma, Y., Duan, N., … Fei, Z. (2020). Phased diploid genome assemblies and pan-genomes provide insights into the genetic history of apple domestication. *Nature Genetics*, *52*(12), 1423–1432. doi:10.1038/s41588-020-00723-9

Takahashi, H., Fukuhara, T., Kitazawa, H., & Kormelink, R. (2019). Virus Latency and the Impact on Plants. *Frontiers in Microbiology*, *10*. doi:10.3389/fmicb.2019.02764

Teycheney, P.-Y., Geering, A. D. W., Dasgupta, I., Hull, R., Kreuze, J. F., Lockhart, B., … Report Consortium, I. (2020). ICTV virus Taxonomy profile: Caulimoviridae. *The Journal of General Virology*, *101*(10), 1025–1026. doi:10.1099/jgv.0.001497

The TE Hub Consortium, Elliott, T. A., Heitkam, T., Hubley, R., Quesneville, H., Suh, A., & Wheeler, T. J. (2021). TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. *Mobile DNA*, *12*(1). doi:10.1186/s13100-021-00244-0

Tong, Z., Gao, Z., Wang, F., Zhou, J., & Zhang, Z. (2009). Selection of reliable reference genes for gene expression studies in peach using real-time PCR. *BMC Molecular Biology*, *10*(1). doi:10.1186/1471-2199-10-71

Tromas, N., Zwart, M. P., Poulain, M., & Elena, S. F. (2014). Estimation of the in vivo recombination rate for a plant RNA virus. *The Journal of General Virology*, *95*(3), 724–732. doi:10.1099/vir.0.060822-0

Underwood, W., & Somerville, S. C. (2017). Phosphorylation is required for the pathogen defense function of the Arabidopsis PEN3 ABC transporter. *Plant Signaling & Behavior*, *12*(10), e1379644. doi:10.1080/15592324.2017.1379644

Ungerer, M. C., Strakosh, S. C., & Zhen, Y. (2006). Genome expansion in three hybrid sunf lower species is associated with retrotransposon proliferation. *Curr Biol*, *16*.

Urtubia, C., Devia, J., Castro, Á., Zamora, P., Aguirre, C., Tapia, E., … Prieto, H. (2008). Agrobacterium-mediated genetic transformation of Prunus salicina. *Plant Cell Reports*, *27*(8), 1333–1340. doi:10.1007/s00299-008-0559-0

Vassilieff, H., Geering, A. D. W., Choisne, N., Teycheney, P.-Y., & Maumus, F. (2023). Endogenous caulimovirids: Fossils, zombies, and living in plant genomes. *Biomolecules*, *13*(7), 1069. doi:10.3390/biom13071069

Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary genomics of peach and almond domestication. *G3 (Bethesda, Md.)*, *6*(12), 3985–3993. doi:10.1534/g3.116.032672

Vendramin, E., Pea, G., Dondini, L., Pacheco, I., Dettori, M. T., Gazza, L., … Rossini, L. (2014). A unique mutation in a MYB gene cosegregates with the nectarine phenotype in peach. *PloS One*, *9*(3), e90574. doi:10.1371/journal.pone.0090574

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J. M., & Castanera, R. (2019). A benchmark of transposon insertion detection tools using real data. *Mobile DNA*, *10*(1). doi:10.1186/s13100-019-0197-9

Vendrell-Mir, P., López-Obando, M., Nogué, F., & Casacuberta, J. M. (2020). Different families of retrotransposons and DNA transposons are actively transcribed and may have transposed recently in Physcomitrium

(Physcomitrella) patens. *Frontiers in Plant Science*, *11*. doi:10.3389/fpls.2020.01274

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., … Rokhsar, D. S. (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., … Schmutz, J. (2017). The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. *BMC Genomics*, *18*(1). doi:10.1186/s12864-017-3606-9

Vicient, C. M. (2010). Transcriptional activity of transposable elements in maize. *BMC Genomics*, 11(1). doi:10.1186/1471-2164-11-601

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

Vilanova, S., Sargent, D. J., Arús, P., & Monfort, A. (2008). Synteny conservation between two distantly-related Rosaceae genomes: Prunus (the stone fruits) and Fragaria (the strawberry). *BMC Plant Biology*, *8*(1), 67. doi:10.1186/1471-2229-8-67

Wang, H.-Y., Tian, Q., Ma, Y.-Q., Wu, Y., Miao, G.-J., Ma, Y., … Liu, B. (2010). Transpositional reactivation of two LTR retrotransposons in rice-Zizania recombinant inbred lines (RILs): Reactivation of two LTR retrotransposons in rice. *Hereditas*, *147*(6), 264–277. doi:10.1111/j.1601-5223.2010.02181.x

Wang, N., Zhang, D., Wang, Z., Xun, H., Ma, J., Wang, H., … Liu, B. (2014). Mutation of the RDR1 gene caused genome-wide changes in gene expression, regional variation in small RNA clusters and localized alteration in DNA methylation in rice. *BMC Plant Biology*, *14*(1). doi:10.1186/1471-2229-14-177

Warschefsky, E., Penmetsa, R. V., Cook, D. R., & von Wettberg, E. J. B. (2014). Back to the wilds: Tapping evolutionary adaptations for resilient crops through systematic hybridization with crop wild relatives. *American Journal of Botany*, *101*(10), 1791–1800. doi:10.3732/ajb.1400116

Watkins, R. (1976). *Cherry, plum, peach, apricot and almond* (N. W. Simmonds, Ed.). London: Longman.

Wells, J. N., & Feschotte, C. (2020). A field guide to eukaryotic transposable elements. *Annual Review of Genetics*, *54*(1), 539–561. doi:10.1146/annurev-genet-040620-022145

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

Xiang, Y., Huang, C.-H., Hu, Y., Wen, J., Li, S., Yi, T., … Ma, H. (2016). Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. *Molecular Biology and Evolution*, msw242. doi:10.1093/molbev/msw242

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., & van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science (New York, N.Y.)*, *319*(5869), 1527–1530. doi:10.1126/science.1153040

Xu, Z.-S., Yang, Q.-Q., Feng, K., & Xiong, A.-S. (2019). Changing carrot color: Insertions in *DcMYB7* alter the regulation of anthocyanin biosynthesis and modification. *Plant Physiology*, *181*(1), 195–207. doi:10.1104/pp.19.00523

Yaakov, B., & Kashkush, K. (2011). Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. *Genome*, *54*(1), 42–49. doi:10.1139/G10-091

Yoo, M.-J., Szadkowski, E., & Wendel, J. F. (2013). Homoeolog expression bias and expression level dominance in allopolyploid cotton. *Heredity*, *110*(2), 171–180. doi:10.1038/hdy.2012.94

Yoon, J., Choi, H., & An, G. (2015). Roles of lignin biosynthesis and regulatory genes in plant development. *Journal of Integrative Plant Biology*, *57*(11), 902–912. doi:10.1111/jipb.12422

Yu, H., Wang, X., Lu, Z., Xu, Y., Deng, X., & Xu, Q. (2019). Endogenous pararetrovirus sequences are widely present in Citrinae genomes. *Virus Research*, *262*, 48–53. doi:10.1016/j.virusres.2018.05.018

Zhang, A., Zhou, H., Jiang, X., Han, Y., & Zhang, X. (2021). 124 Pan') Provides Insights into Its Good Fruit Flavor Traits. *Plants*, *10*(3).

Zhang, R.-G., Li, G.-Y., Wang, X.-L., Dainat, J., Wang, Z.-X., Ou, S., & Ma, Y. (2022). TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. *Horticulture Research*, *9*. doi:10.1093/hr/uhac017

Zohary, D., Hopf, M., & Weiss, E. (2012). *Domestication of Plants in the Old World*. Oxford University Press, Oxford.

Zong, X., Denler, B. J., Danial, G. H., Chang, Y., & Song, G.-Q. (2019). Adventitious shoot regeneration and Agrobacterium tumefaciens-mediated transient transformation of almond × peach hybrid rootstock 'Hansen 536.' *HortScience: A Publication of the American Society for Horticultural Science*, *54*(5), 936–940. doi:10.21273/hortsci13930-19

ACKNOWLEDGMENTS

En primer lugar, quiero agradecer a mi director de tesis, Carlos, por haberme dado esta oportunidad. Tu supervisión, criterio científico y disponibilidad total han sido constantes. Gracias por haberme ayudado siempre que lo he necesitado y aportarme tantas ideas. Has sido un apoyo científico y también personal a lo largo de estos años.

Vull donar les gràcies al Pep, per donar-me l'oportunitat, tutoritzar aquesta tesi i l'ajuda científica que he rebut als lab meetings i fora d'ells.

Agradezco también al resto de miembros de mi grupo. Empecé la tesis haciendo solamente biología molecular. La he acabado después de una pandemia mundial y de haberme iniciado en la bioinformática. Lo segundo, ha sido mucho más fácil gracias a Raúl, mi mentor bioinformático, que me ha ayudado en muchos análisis. Gracias a Pol, Noemia, Marc y Andrea, que han sido un apoyo y hemos pasado muy buenos momentos en el laboratorio y fuera. Gracias por brindarme vuestra amistad. A Lucía, Lana y Amalia, aunque hemos coincidido menos en el tiempo, lo compartido dentro y fuera del laboratorio ha sido de gran calidad. Y gracias también al resto de compañeros: Julia, Amélie, Fabio, Daniel, Miguel… En el laboratorio 102, además de con mi grupo de investigación, he compartido espacio con Igni (Nacho). Gracias por tu amistad, por darme el mejor protocolo de transformación y tanta vidilla. Gracias a Poonam, Unai, Carloscu, Juanjo, Rodrigo, Michela, Niki, Soraya, Sandra, Jonice… ¡A todos!

A Ibo, por haberme ayudado en la recolecta de material vegetal durante esta tesis. Al grupo de Prunus del IRTA por las discusiones sobre el nuevo genoma de almendro. Al tribunal de seguimiento de mi doctorado, que cada año me han dado buenos consejos e ideas. A los servicios del CRAG, en especial, secretaría, secuenciación, genómica y compras. Gracias a los amigos que he hecho durante estos años en el CRAG. A l'Arnau, el primer amic que vaig fer al CRAG. Una gran amistat que es manté a la distancia... A la Camila i la Violeta, molt bones amigues des del màster i que continuaran. Gracias Mariano por tu amistad desde 2020. Y a Laieta, Cristina, Nuria, Pedro, Ari, Yu, Andrea, Sandi, Elena, Christian, Lucia, Alba, José…

Gràcies als meus amics. A Marta, que nuestras conversaciones han sido muy enriquecedoras. Als amics de Girona, que sempre esteu tot i que passin els anys!

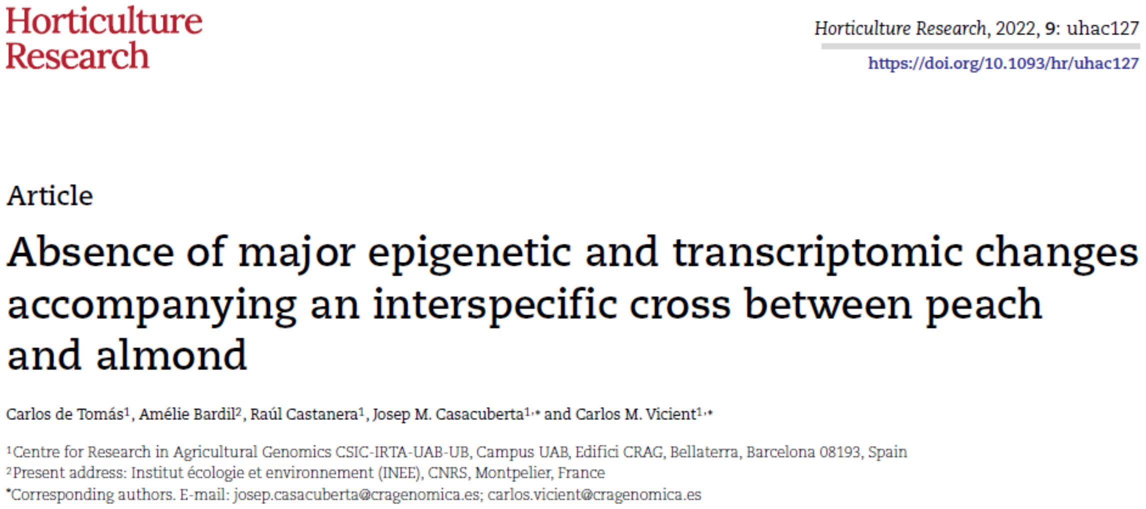
Por último, quiero agradecer a mi familia, incluyendo a mi gato Genís (si eres amante de los animales, lo entenderás), que no se ha separado de mí mientras escribía esta tesis. Gracias avis, aunque no estéis, soy quién soy, gracias a vosotros. Gracias tieta, sabes que eres mi segunda madre, también al tiet. Gracias Laura, mi querida hermana, y a Ignasi, que me habéis dado el mejor regalo con Inés y Abril. Ambas trajeron buena suerte. Inés nació cuando me ofrecieron empezar esta etapa y Abril en el ecuador. Gracias a mis padres. Sin ellos, no hubiese empezado, ni acabado esta tesis. Habéis estado incondicionalmente conmigo, cuanto todo iba bien, pero también cuando todo me desbordaba. Os quiero.

ANNEXES

ANNEXES

Articles published during the thesis:

de Tomás, C., Bardil, A., Castanera, R., Casacuberta, J.M., & Vicient, C.M. (2022). Absence of major epigenetic and transcriptomic changes accompanying an interspecific cross between peach and almond. *Hortic Res, 9*, uhac127. doi: 10.1093/hr/uhac127.



Horticulture Research

Horticulture Research, 2022, **9**: uhac127

https://doi.org/10.1093/hr/uhac127

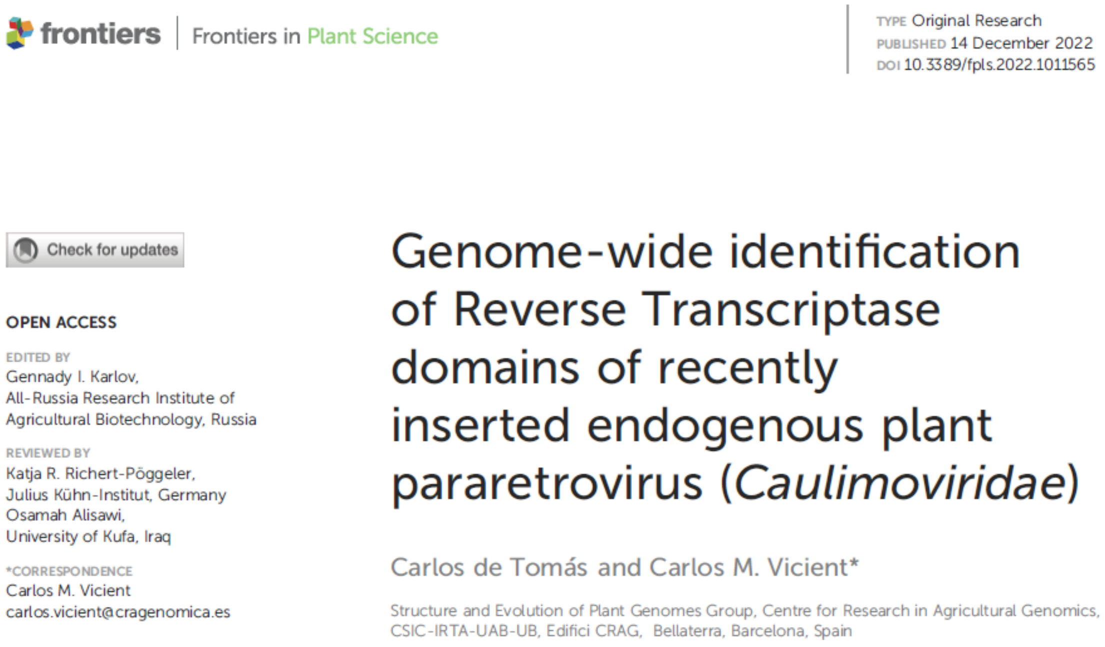
Article

Absence of major epigenetic and transcriptomic changes accompanying an interspecific cross between peach and almond

Carlos de Tomás[1], Amélie Bardil[2], Raúl Castanera[1], Josep M. Casacuberta[1,*] and Carlos M. Vicient[1,*]

[1]Centre for Research in Agricultural Genomics CSIC-IRTA-UAB-UB, Campus UAB, Edifici CRAG, Bellaterra, Barcelona 08193, Spain
[2]Present address: Institut écologie et environnement (INEE), CNRS, Montpelier, France
*Corresponding authors. E-mail: josep.casacuberta@cragenomica.es; carlos.vicient@cragenomica.es

de Tomás, C., & Vicient, C.M. (2022). Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (*Caulimoviridae*). *Front Plant Sci, 13*, 1011565. doi: 10.3389/fpls.2022.1011565.



frontiers | Frontiers in Plant Science

TYPE Original Research
PUBLISHED 14 December 2022
DOI 10.3389/fpls.2022.1011565

Check for updates

OPEN ACCESS

EDITED BY
Gennady I. Karlov,
All-Russia Research Institute of Agricultural Biotechnology, Russia

REVIEWED BY
Katja R. Richert-Pöggeler,
Julius Kühn-Institut, Germany
Osamah Alisawi,
University of Kufa, Iraq

*CORRESPONDENCE
Carlos M. Vicient
carlos.vicient@cragenomica.es

Genome-wide identification of Reverse Transcriptase domains of recently inserted endogenous plant pararetrovirus (*Caulimoviridae*)

Carlos de Tomás and Carlos M. Vicient*

Structure and Evolution of Plant Genomes Group, Centre for Research in Agricultural Genomics, CSIC-IRTA-UAB-UB, Edifici CRAG, Bellaterra, Barcelona, Spain