Impact of transposition on the generation of genetic variability in Prunus crop species

Fabio Barteri

PhD Dissertation

Impact of transposition on the generation of genetic variability in *Prunus* crop species.

Candidate: Fabio Barteri
Director: Dr. Josep Mª Casacuberta

UAB
Universitat Autònoma
de Barcelona

CRAG
CENTRE FOR RESEARCH
IN AGRICULTURAL GENOMICS

Doctorat en
Biologia i
Biotecnologia
Vegetal

That's me, on the beach side combing the sand, metal meter in my hand, sporting a pocket full of change.

NOFX

Preface.

Among the intricacies of the long and vibrant discussion on Transposable Elements (TEs) in modern biology, we'd probably need two very simple and catchy Barbara McClintock's quotes only to exhaustively summarize the most interesting features of these genetic elements. In two separate occasions, the Northern American scientist described the transposons she discovered as "jumping genes" and "controlling elements", to indicate their capability to change genomic location and to provoke a measurable impact on the phenotype. Even though current knowledge on genetics and TEs should get us to be at least rather cautious with defining these elements merely as "genes", the two expressions McClintock used are still functional in highlighting the two points of major interest in the genomic investigation on TEs: how they move, and what's the impact of their movement on the genes and their regulation.

From their discovery in 1951, research on transposons brought to light the several molecular mechanisms that allow their movement and proliferation and described many cases in which transposons are responsible for disruptive mutations or have provided new regulatory regions that changed gene regulation. Many of these results have highlighted a link between transposition and adaptation. Sometimes, the mere landing of a transposon close enough to a gene to impact its regulation can result in an adaptive change. Several studies have highlighted how some insertions are likely to play a role in adaptive evolution by changing gene regulation (Hirsch and Springer 2017). Transposons genes can be imported as novel host genes that are positively selected for conferring phenotypic benefits, or some portion of the transposon sequence can be maintained because it becomes part of a regulatory region, a phenomenon that has been defined "exaptation" (Hoen and Buerau 2012)

These findings enforce the idea that transposition has played a pivotal role in the structural and functional evolution of eukaryotic genomes, highlighting the need to evaluate their impact over biological diversity at a global scale. In this, the most important variables that have to be related are two. The first one is the change of the TEs' composition in the genome over the time, or else the balance between the creation of new insertions and their elimination, that is defined transpositional dynamics. The second variable is the genetic variability, defined as the presence and the creation of genetic differences, and usually evaluated at different levels, such as difference between cells or tissues in the same individual (somatic variability), difference between individuals or varieties within the same population or species (intraspecific variability) and even difference between two species (interspecific variability). Deciphering the relationship between transpositional dynamics and genetic variability has the potential to provide us with a fair description of the impact of transposition on the establishment of biological diversity, that is the base of the introduction of evolutionary innovation.

In this doctoral thesis we will investigate the relationship between transposition and genetic variability in peach (Prunus persica) and almond (Prunus dulcis), two sister crops that belong to the Rosaceae family (DeVore, et al., 2007). These two diploid species are so close that we can view them as "on the edge of speciation". Although we place their separation around 6 million years ago, about 50% of fecundation between individuals from these two species still results in viable and fertile offspring (Dirlewanger, 2004). Despite their high evolutionary proximity, peach and almond still display two fundamental differences at genomic level. The first one is the nucleotide diversity, that is seven folds higher in almond than peach (Velasco et al. 2014). The second relevant difference is the reproductive strategy. Whilst the most of cultivated almond varieties are made self-incompatible by a molecular mechanism that is specific for the Prunus genus (Tao, et al., 2010), peach is auto-fertile due to the secondary loss of this mechanism. These elements altogether enforce the outlook for these two species as good model to study how the structural genomic variation associated with transposition can affect genetic variability at different levels.

This dissertation unfolds around two main aspects. Under a biological point of view, we will assess the impact of transposition on the differentiation between peach and almond, a work that has been part of a collective project aimed at the assembly and annotation of the almond genome (Alioto et al. 2019). On a more technical fashion, a part of this thesis will instead focus on the technical evaluation of bioinformatics tools that are aimed at the identification of TEs from DNA-seq data and will proposed a possible workflow to quantitate TEs transcription via RNA-seq metanalysis.

Table of contents.

Introduction

Part A – The Mobile DNA. Transposable Elements and their impact on plant genomes.

As it happens in many major breakthroughs in science, the existence of Transposable Elements (TEs) emerged during an investigation that was focused on an apparently distinct phenomenon. When the first transposable element was described, Barbara McClintock was actually studying the chromosome breakage in Maize (McClintock, 1950). A particular breakage point on Maize chromosome 9 was identified as "dissociation" locus or Ds. This locus was found capable to change its genomic position under the dominant effect of another locus, the "activator" or Ac. In a famous article published on PNAS in 1950, the Ac/Ds system was proposed as the first transposable element ever described. This "jumping gene" was able to change its genomic position, was responsible for the induction of chromosomic breakages, and its movement was associated with a plethora or distinct phenotypic outcomes, including the mosaic colors in Maize kernels and leaves (McClintock 1950).

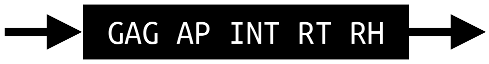
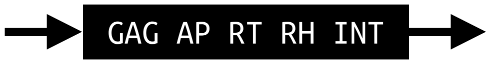
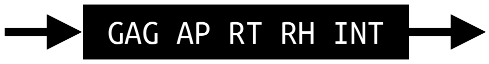
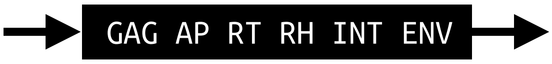
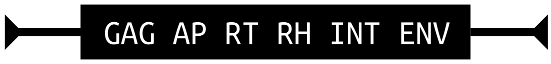
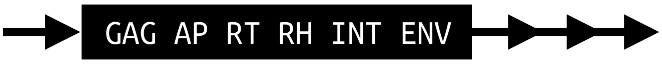
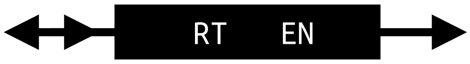
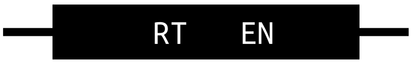
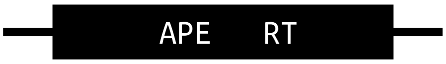
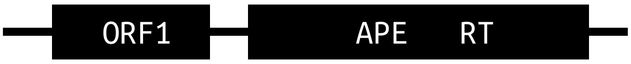
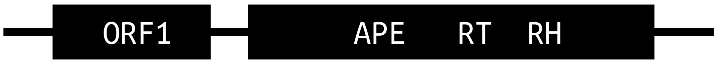
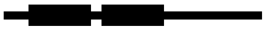
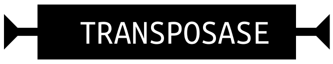
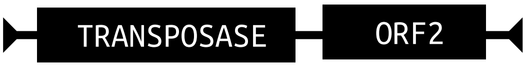
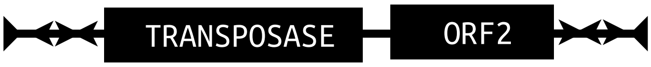
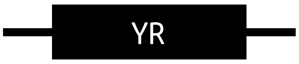
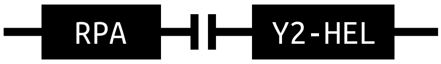
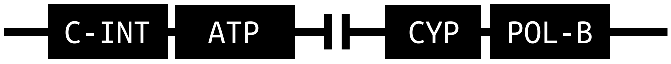
Despite some initial resistance, the scientific community slowly acknowledged increasing consideration to these mobile elements, that culminated in the awarding of the Nobel Prize to their discoverer Barbara McClintock in 1983, and that brought along important information about their diffusion, impact, evolutionary origins and transpositional mechanisms. Since the beginning of the history of research on TEs, what we now nowadays as a major component of eukaryotic genomes have been associated to three main features. The first one is mobility, or else the fact that they can change their chromosomic position as the Ds elements where find to do in Maize. The second one is the close connection with chromosomal structural features, that was made explicit by the Ac/Ds induction of chromosomal breakages, and that was later proven in the role that TEs seem to have in the definition of chromosomal structure (Klein and O'Neill 2018). The third one is the effect on the biological diversity. Ac/Dc movement was associated with the rise of some particular phenotypes, such as the mosaic pigmentation on kernels and leaves. This result has been the first of many findings highlighting the impact of transposition on the generation of genetic and phenotypic diversity (Fechotte, Jiang and Wessler 2002) .

In this first part of the introduction to this thesis, we will resume the current knowledge on TEs focusing on these three main features. First, we will concentrate on their mobility and on the mechanisms that permit it as a key feature for their classification and as a major effector of their change in number over the time, the transpositional dynamics. Then, we will go through their impact over the host genome on the double rail of highlighting the connection with chromosome side and the direct effect on genetic diversity on the other side.

I. Transposable Elements: structure and classification

The first attempt to classify Transposable Elements (TEs) dates back to 1989, when David J. Finnegan from the University of Edinburgh divided them in two classes (Class I and Class II) on the base of the molecular intermediate they produce at the moment of transposition (Finnegan 1992). Whilst Class II transposons are excised as they transpose, Class I elements do not. Their transpositional mechanism starts with the transcription of an RNA intermediate that is then retro-transcribed into DNA at the moment of the insertion. For this reason, it is common to refer to these elements as Retro-Transposons (RTs) (Goodler and Kazazian, 2008). This classification has been widely accepted for its simplicity and for resulting rather intuitive because of the parallelism with common personal computer actions "cut and paste" and "copy and paste" (Feschotte et al., 2002). However, in the mid 90s some small DNA transposons, the MITE (Miniature Inverted-repeats Tranpsosable Elements) were discovered as the first non-autonomous TEs (Wessler, Bureau and White 1995; Feschotte et al., 2002). This discovery, along with the increasing information on the different molecular mechanisms underlying transposition, highlighted the need to be able to classify the TEs. A different approach to classify autonomous elements has been done on the base of the enzymological categories of the proteins they encode (Curcio and Derbyshire, 2003). The two classifications, the molecular intermediate- based and the enzymological – based coexisted for some years till 2007, when a milestone paper first-authored by University of Zurich's Tomas Wicker proposed the TEs classification system that is still currently in wide use (Wicker et al., 2007).

Wicker et al. classification is shown in Figure 1. It is a hierarchical classification that takes into account the presence of a RNA intermediate, the genomic structure of the TEs and the molecular complex they encode for. Two main classes are defined on the base of whether an RNA intermediate is produced (Class I, RNA-Elements or Retro Transposons) or not (Class II or DNA-Elements). A further division into sub-classes is operated on the base of the transpositional mechanism, depending on whether the transposon copies itself into a new element that is then inserted in the genome (copy-and-paste mechanism) or it excides from the original position to insert elsewhere (cut-and-paste mechanism). Since all the Class I elements share a copy-and-paste transposition, this further division only makes sense within the Class II (DNA transposons), where we distinguish those TEs that have a cut-and-paste transposition (Subclass I) from the ones that have a copy-and-paste strategy (Subclass II). Subclasses are then further divided into Orders of elements sharing the same genomic structure. In Figure 1 we can easily recognize the two major Class I and Class II orders, the LTR-Retro Transposons (LTR-RTs), Class I elements that are characterized by the presence of long terminal directed repeats, and the Class II Terimnal Inverted Repeats (TIR) order, that recoil DNA transposons that are characterized by the presence of inverted repeats at the 5' and 3'

Classification			Structure
Order	Superfamily(ies)	Code	
Cass I (retrotransposons)			
LTR	Copia	RLC	 GAG AP INT RT RH
	Gypsy	RLG	 GAG AP RT RH INT
	Bel-Pao	RLB	 GAG AP RT RH INT
	Retrovirus, ERV	RLR, RLE	 GAG AP RT RH INT ENV
DIRS	DIRS	RYD	 GAG AP RT RH INT ENV
	Ngaro, VIPER	RYN, RYV	 GAG AP RT RH INT ENV
PLE	Penelope	RPP	 RT EN
LINE	R2	RIR	 RT EN
	RTE	RIT	 APE RT
	L1, Jockey	RIL, RIJ	 ORF1 APE RT
	I	RII	 ORF1 APE RT RH
SINE	tRNA, 7SL, 5S	RST, RSL, RSS	
Cass II (DNA transposons) - Subclass 1 (“cut-and-paste” transpositional strategy)			
TIR	Mariner, hAT, Mutator, Merlin, Translib, P, PiggyBac	DTT, DTA, DTM, DTE, DTR,DTP, DTB	 TRANSPOSASE
	PIF-Harbinger	DTH	 TRANSPOSASE ORF2
	CATCA	DTC	 TRANSPOSASE ORF2
Crypton	Crypton	DYC	 YR
Cass II (DNA transposons) - Subclass 2 (“copy-and-paste” transpositional strategy)			
Helitron	Helitron	DHH	 RPA Y2-HEL
Maverick	Maverick	DMM	 C-INT ATP CYP POL-B

Legend.

Structural features

→	Long terminal repeat
▶—◀	Terminal inverted repeats
YR	Coding region
—	Non-coding region
—▬—	Diagnostic feature in coding region
⊣⊢	Region that can contain one or more additional ore

Protein Coding domains

AP	Aspartic proteinase	**INT**	Integrase
APE	Apurinic endonuclease	**ORF**	Open reading frame of unknown function
ATP	Packaging ATPase	**POL-B**	DNA polymerase B
C-INT	C-integrase	**RH**	RNAse H
CYP	Cysteine protease	**RPA**	Replication protein A (found only in plants)
EN	Endonuclease	**RT**	Reverse transcriptase
ENV	Envelope protein	**TRANSPOSASE**	Transposase
GAG	Capsid protein	**YR**	Tyrosine recombinase
HEL	Helicase	**Y2**	Tyrosine recombinase with YY motif

Figure 1 - Wicker classification for TEs (Wicker et al., 2007)

Wicker's TEs classification sorts TEs into classes and subclasses according to their replication mechanism, then into orders in reason of their structure and finally into superfamilies according to the organisation of the open reading frames. The table is elaborated on the one published by Wicker et al. on Nature Genetics in 2007.

termination of the element. Finally, the organization of the coding domains defines the Superfamilies. Studies that focus on LTR-RTs use to revolve on two main superfamilies, the Ty-1-Copia-like or Copia and the Ty-3-Gypsy-like or Gypsy (Boeke et al., 1985). These two superfamilies differ in the organization of the POL open reading frame. Further classifications into families are built upon sequence similarity. Usually, the so-called "80-80-80 rule" applies. TEs sharing the 80% of sequence identity, over a length of more than 80 nucleotides that constitutes the 80% of whole sequence length, are grouped in the same family (Wicker et al., 2007). Anyways, the Wicker classification limits to the superfamily only. TEs are usually very diverse and share poor sequence similarity (Le et al., 2000). Besides any known genome was found bearing TEs (Bourque et al., 2018), "80-80-80" families are mostly species-specific or shared between few sister species (Le et al., 2000). If we add to this that any species might reach to a high number of different families and singletons, we can understand that a global classification of TE families would result in an unrealistic count of thousands different families.

Class I transposable elements or Retrotransposons.

Retrotransposons (RTs) are TEs that need an RNA intermediate to complete their transpositional process. Figure 2-A shows a simplified scheme of the main steps of RTs' transpositional cycle. A RT copy is first transcribed to produce a RNA that will form a ribonuclear complex along with the proteins translated from the RT's ORF. These proteins are responsible for the retro-transcription of the RNA intermediate into a DNA copy of the original transposon, and for its integration in the host genome. This integration can happen randomly in the genome or be directed towards some specific regions, such as heterochromatic regions (Bennentzen, 2014).

The enzyme that catalyzes reverse transcription, the Retrotranscriptase was discovered and insulated in 1972 (Temin and Baltimore 1972). It is an RNA- based DNA polymerase that catalyzes the synthesis of a DNA chain upon an RNA template and is conserved through all the RTs superfamily and shared with retroviruses and Multi copy Single stranded DNA (msDNA or retrons) (Temin and Mizutani, 1970). Phylogenetic studies focused on this protein have clarified the evolutionary relationships between RTs, RNA viruses and other genetic elements. The phylogenetic tree in Figure 2-B was published by Yue Xiong and Thomas H. Eickbush in 1990, and it is based on the multiple alignment of 82 Retrotranscriptase primary sequences, with 15 RNA-dependent RNA polymerases serving as outgroup. These latter enzymes are shared by several RNA viruses that do not need a DNA intermediate in their life cycle, such as the Coliphages (Remaut, Stanssens and Fiers, 1990), the Sinbad-like viruses (Copeland et al., 2005) and the polio-like viruses (Wimmer, Hellen and Cao, 1993). The acquisition of LTR termination gave rise to the LTR-RTs and to three distinct virus groups: retroviruses (Linial and Brail, 1982), hepadnaviruses (Howard, 1986) and caulimoviruses (Blanc et al., 2001). The Retrotranscriptases grouped in the other branch are shared by non-LTR retrotransposons (Eickbush, 1992) and some mobile elements such as msDNA, type II self-splicing introns, mitochondrial RTLs and Nc-P1 elements (Lampson, Inouye and Inouye, 2005). The evolutionary relationships are drawn in Figure 2-C highlighting the introduction of

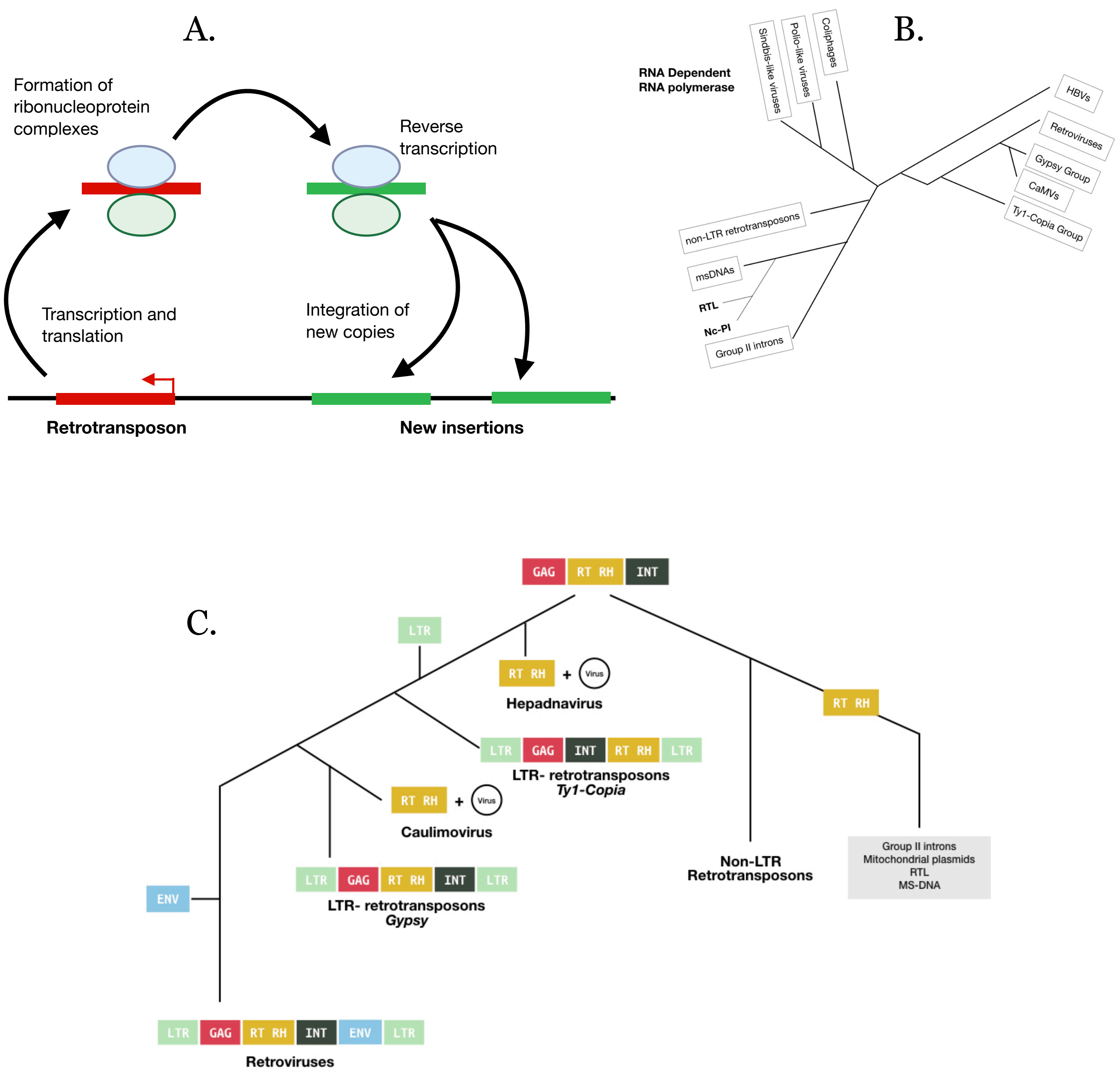


Figure 2 - Class I Retro Transposons transpositional mechanism and evolution

A. Transpositional mechanism of Class I Retro Transposons requires the formation of a ribonucleoproteic intermediate that is retro-transcribed in a new insertion.
B. Retro Transposons evolutionary tree based on retrotranscriptase alignment from 82 mobile elements and viruses, and compared with 15 RNA- dependent RNA polymerases as outgrup clarifies the evolutionary relationships between retro elements and viruses.
C. Introduction of important structural features such as the LTRs and the ENV (envelope) coding regions during retrotransposons and retroviruses evolution.

structural and functional innovation during retrovirus and retrotransposons differentiation. In the right branch we see the rise of non-LTR Retrotransposons and the subsequent differentiation of Group II introns and mitochondrial mobile elements (plasmids, RTLs and msDNAs). The left branch shows how the differentiation of LTR retrotransposons and Retroviruses was marked by the introduction of structural and functional features. The appearance of the LTR ends marked the difference between LTR retrotransposons and hepadnaviruses (Howard, 1986). Whilst these latter are proposed to be the result of the fusion with a DNA virus genome, the introduction of the long terminal repeat marked the separation of LTR-RTs . From LTR-RTs, the relationship with viruses is highlighted in two split events. First, the separation between Gypsy LTR-RTs and the Cauliflower Mosaic Virus (Caulimovirus or CMV) was probably due to the fusion of a Gyspy-like sequence with a dsDNA virus (Wagstaff et al. 2012), then the importation of the Envelope coding region, that superintend the formation of viral envelope brought to the appearance of Retroviruses (Eickbush and Malik, 2002).

If we focus our attention back to Figure 1, we can appreciate how the Wicker classification works. The discriminant to define an order is the genomic structure, which most evident feature is the disposition of the internal repetitions. LTR-RTs share with DIRS for the disposition of their repetitions, that are also present in the Penelope Retro- Transposons (PLE-RTs), but absent in the Long and Short Interspersed Nuclear Elements (LINE and SINE). Within the orders, superfamilies are defined by the organization of the open reading frames. Within LTR-RTs, the two main superfamilies Copia and Gypsy differ for the disposition of Integrase (INT), Retro Transcriptase (RT) and RNAse-H (RH). Gyspy elements, that are on the average twice longer than copia (about 8-12 Kb instead of 6 kb) (Kumar et al, 1999), have generated several derivates. Bel-Pao elements have been identified in animals only and are considered a shorter (5kb on the average) Gypsy derivate (De la Choux and Wagner, 2011). Retroviruses and Endogenous Retro Viruses (ERV) share the gene disposition with the Gypsy but are provided with a further coding region for viral envelope (ENV). If the DIRS order differs for the number of repetitions flanking the coding regions, its three families (DIRS, Ngaro and VIPER) share a coding region that is similar to the one from Gypsy LTRs plus a gene that encodes for the DNA endonuclease Tyrosine Recombinase (YR). This enzyme catalyzes site-specific DNA breakage that is used to direct the site of insertion, and it is shared with Crypton Class II TEs (Poulter and Goodwin 2005). Long and Short interspersed Elements (LINE and SINE) lack of repetitive regions that flank the coding region. LINE superfamilies code for several enzymes, mostly the RT and the Apurinic Endonuclease (APE) (Seki et al. 1993). Jockey, L1 and I superfamilies are provided with a 5’ ORF with unknown function (Flavell 1991). SINE elements are non-autonomous transposons that opportunistically use the LINE machinery to complete their retro transposition (Shedlock and Okada, 2000).

Class II DNA- transposons

Those TEs which transpositional cycle does not require an RNA intermediate are included in the Class II, and commonly called DNA transposons (Feschotte and Pritham, 2007). Differently from retrotransposons, these elements can be further divided into 2 subclasses (Subclass I and Subclass II) according to their transpositional mechanism (Wicker et al., 2007). Retrotransposons and DNA transposons of the Subclass II transpose with a non-conservative or replicative mechanism, where the element is duplicated to produce a new copy that will insert the genome on its turn. What distinguishes the retrotransposons is the presence of an RNA intermediate, that is not needed in Subclass II DNA-transposons. Ideally, each transpositional event of this kind involves the increase of the number of TEs in the genome in a "non-conservative" fashion. Conversely, Subclass I DNA transposons are excided from their original location to be displaced elsewhere in the genome, in a way that is "conservative" since it implies no increase of the number of TEs in the genome. In a simpler way, these two mechanisms are usually referred as "copy and paste" and "cut and paste" transposition (Derbyshire and Grindley 1986).

Subclass I. The first subclass in which we divide the DNA transposons, thus collects all the known "cut and paste" TEs. A swift look to the Wicker's table in Figure I allows us to appreciate how most of the superfamilies that are mentioned for this subclass share the same genomic structure, with two Tandem Inverted Repeats that flank the coding region and that name the same order (TIR). Besides the variations that allow to operate the distinction into several families, all the shown TIR superfamilies bear a coding region for the enzyme known as Transposase (Tpase). This enzyme is included in the same superfamily as the integrase (Rice and Baker 2001), and its function is to excide, transport and insert TIR TEs.

In Figure 3 we summarize the structure of the transposase (PDB id: 1mus) bond to DNA (Figure 3-B and C), along with the transpositional mechanism that it catalyzes. The model presented in Figure 3 comes from the crystal structure of the Tpase sythesized by the prokariotic DNA TE Tn5, that is widely used as a general model for DNA transposition (De la Cruz et al. 1993). Transpositional cycle starts as two molecules of this enzyme bind the DNA on the two inverted repeats at the end of the TE, to eventually dymerize and force the transposon into forming a DNA loop. The DNA segment containing the two TIR and the Tpase gene is then excised and brought to a target site, where the strand is inserted through DNA cut and repair (Hallet and Sherratt 1997).

The ac/ds mobile system discovered by Barbara Mc Clintock was later proven to be a TIR element classified within the hAT order (Weil and Kuntz 2000).

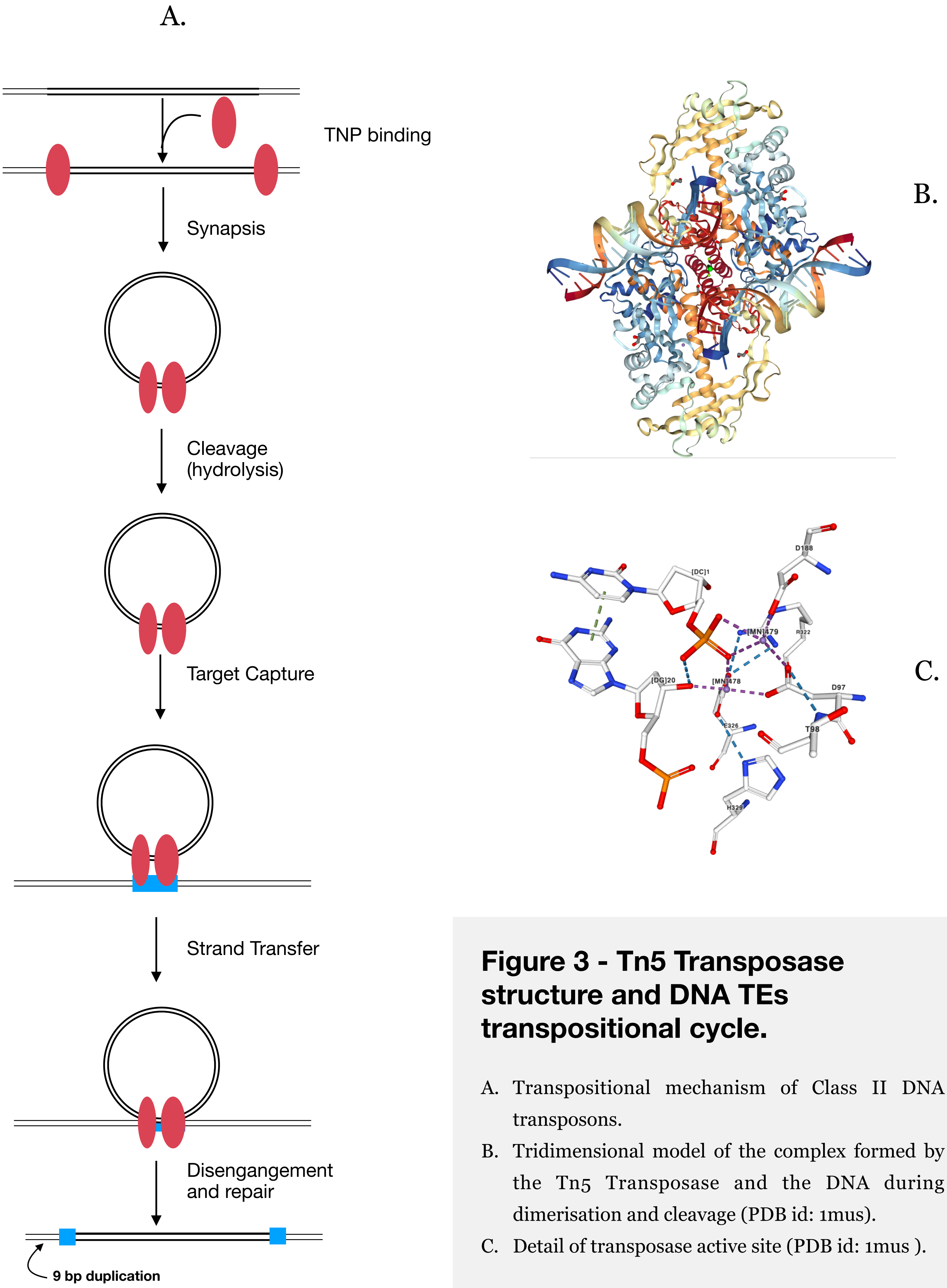


Figure 3 - Tn5 Transposase structure and DNA TEs transpositional cycle.

A. Transpositional mechanism of Class II DNA transposons.
B. Tridimensional model of the complex formed by the Tn5 Transposase and the DNA during dimerisation and cleavage (PDB id: 1mus).
C. Detail of transposase active site (PDB id: 1mus).

Although TIR elements constitute most of the DNA transposons having a cut-and-paste transpositional mechanism (Feschotte and Pritham, 2007), there is another order of mobile elements which transpositional cycle is conservative. Described in 2007 (Pritham et al. 2007), the Maverick elements are basically tyrosyne recombinase (YR) genes that are flanked by short direct repeats about 6-8 nt long. YR recognizes the repetitions and excides the element that circularise to be inserted elsewhere in the genome (Esposito and Scocca 1997).

Subclass II. Not all the DNA elements traspose conservatively. Helitrons and Mavericks are two orders of DNA transposons that duplicate their sequence into a DNA intermediate to complete their transpositional cycle.

Helitrons duplication has been described as very similar to the “rolling circle” replication of prokaryotic plasmids. This mechanism is particularly prone to extend transposition to the flanking regions that are duplicated along with the TE. When this happens involving a whole gene sequence, the phenomenon is called “gene capture” and it is deemed to be an important cause of gene duplication. Despite little in known about their evolutionary history and an actual connection with prokaryotic plasmids has never been demonstrated, the interest in these elements is relevant because of their potential involvement in gene duplication (Kapitonov and Jurka 2007).

Conversely, the evolutionary history of Maverick elements has been more widely studied and linked with several mobile elements and viruses including the adenoviruses (Fischer and Suttle 2011). With a length of about 15 kb and about 10 ORFs, they replicate their sequence with a DNA polymerase they encode and insert the genome by an integrase-like protein (Pritham, Putliwala and Feschotte 2007).

Partial and non-autonomous elements

Over the time, transposons can lose their structural integrity. The accumulation of single nucleotide mutations or major rearrangements, such as large deletions or insertions, can result in the loss of a TE's capability to produce the proteins that are needed to complete its transpositional cycle, and its capability to transpose (Naville et al. 2019). Anyways, the molecular machineries can still recognize these partial elements till these ones keep some specific structure in a good state of conservation and promote their transposition. As already introduced, the transpositional cycle of TIR elements starts with the binding of two Transposases on the two terminal inverted repeats of the TE. These two proteins dimerize eventually to excide the transposon from its insertion site (Since no other structure is needed in cis to successful complete the transposition, a TIR element that has lost its internal region and its capability to transpose but is still defined by the two terminal inverted repeats, can transpose if recognized by the protein machinery synthesized by another transposon (Wessler, Bureau and White 1995; Naville et al. 2019). This means that even if a transposon is not able to complete its transposition “autonomously”, its movement could be complemented by the proteins encoded by other elements. In other terms, TEs can be divided into autonomous and non-autonomous elements depending on their capability to transpose autonomously (Wicker et al. 2007).

Despite non-autonomous elements have been named and classified according to their structure, Wicker and collaborators decided to include not them in their classification, but to mention them along with their “entire” counterparts because poorly informative on TEs diversity. Anyway, even if these elements are not really functional to a global TEs classification, they still have a high impact on eukaryote’s genetic variability.

There are two kinds of non-autonomous elements that generate from LTR retroelements. Nested LTR-RTs insertions can give rise to the Large Retrotransposon derivatives (LARDS) that where originally found in Barley and in other Triticaceae. They have long LTR terminations (bout 4.5 kb) and an internal non coding domain. Another class of non-autonomous retroelements is represented by the terminal-repeat retrotransposons or TRIMs. Originally identified in Arabidopsis, these elements are small LTR-RTs 540 bp long on the average. In plants, their placement in promoters and introns indicates a potentially relevant impact on genetic variability (Witte et al. 2001).

Perhaps, the most common and important non-autonomous elements are defective derivates of the Class II - TIR elements. TIRs can lose their internal region by deletion, just like it happens to the LTR elements. The element that is generated is barely a small tandem of inverted repeats that is defined Miniature Inverted Tandem-repeats Element or MITE (Wessler, Bureau and White 1995). The first MITE identified was found in a maize gene named waxy. This gene encodes for an ADP-glucose transferase responsible for amylose biosynthesis in endosperm and pollen and derives its name from the "waxy" appearance of mutant kernels, that reflects the lack of amylose in the endosperm. When a 192 bp long element was found in its introns, this element was first called "Tourist", and appeared as a small non-coding DNA region bordered by two inverted repeats (Bureau and Wessler 1992).

Along with LTR retrotransposons, MITE elements are the most active and impactful TEs in plants. They are present in high copy number in plant genomes, even though the mechanism in which they amplify is still unclear considering their conservative transposition. They are often found in gene-rich regions and bearing TF binding motifs (Casacuberta and Santiago 2003, Lu et al. 2012). In other words, these small elements potentially affect plants' gene regulation by rewiring new genes into gene network and give rise to important genetic innovation (Morata et al. 2018).

II Transpositional dynamics and impact of TEs on eukaryotic genomes.

Wicker's classification effectively suffices in sorting the most of known TEs into superfamilies, order and classes, fairly describing the diversity of these elements. This diversity reflects the different strategies that have evolved to allow TEs to be maintained in the genome, and it is clear that the movement is a key feature to understand the impact of transposons on chromosome structure and genetic variability (Bourque et al. 2018). In this section, we will review the concept of transposon dynamics as the change of genomic transposon content and positioning over the space and the time (Doolittle 2012, González and Petrov 2017), to later focus on the effect of this changes over the structure of eukaryotic genomes. The effect of the transposition on the genetic variability will be left to the next section.

The dynamics of transposable elements

In genetics, the concept of "dynamics" is usually adopted to indicate genotype's changes during evolution and between different genetic assets (Orgogozo et al. 2015). A genotype can change over the time from cell to cell, as it happens in cancer development (Merlo et al. 2006), among individuals of a population or different populations of the same species (Violle et al. 2012) and, of course, as between different species before and after their separation (Ioeger et al. 1990, Rieseberg and Willis 2007). Genetic dynamics describes the change of the genotype (or a subset of it) in a space that is defined by the context where we evaluate the variability; a single organism, a population, an entire species or different species. That is why some authors refer to the genetic dynamics as the change of the genotype "over time and space" (Lewontin 1974, Lachance 2008, Orgogozo et al. 2015). As for the genes, we can sort all the possible changes into 4 distinct categories.

1. Changes in their presence in a genome, due to novel gene formation (Kaessmann 2010), gene loss (Abalat and Cañestro 2016) or horizontal transfer (Keeling and Palmer 2008).

2. Changes in function, regulation and connection of existing genes, due to fusion with other genes (Mertens et al. 2015, Gao et al. 2018), mutations of regulatory regions that can change gene's regulation (Abbas et al. 2006) and/or rewiring in different gene networks (Tu et al. 2018).

3. Changes in the positioning in the genome, that may be the consequence of chromosomal rearrangements (Zhang et al. 1999) or transposon gene capture (Barbaglia et al 2012).

4. Changes in the number of copies, that can be the result of gene duplication (Kaessmann 2010).

In a similar way, but taking some important differences into account, we can describe the dynamics of TEs. These genetic elements are characterized by their movement, that results in an increase of the copy number in non-conservative, "copy and paste" transposition (Bennetzen 2014) or in a simple displacement of the location of an element, as it happens in the "cut and paste" conservative transposition (Reznikov 2003). Just like the genes, they can be transferred horizontally in a new genome or be lost in several ways

(Copeland et al. 2005), they can change their functionality after mutation (Bourque et al. 2018) and, of course, they can move and also increase their copy number. Perhaps, the most relevant difference between the genetic dynamics of transposons and genes is that whilst these latter have no active control on their dynamics, mobile elements can actively change their genomic location or their copy number. However, we need not to forget that transposition has a relevant but still partial impact on the global genetic dynamics of TEs, that can in fact undergo to mutations and removal like any other DNA portion. The dynamics of TEs is hence the result of transposition and transposon elimination, two forces that we're about to argue as often counteracting during evolution.

Impact of TEs on chromosome structure

Eukaryotic genomes are covered by TEs in a proportion that is variable, but almost always rather considerable. Having been found in virtually any known eukaryotic genome, in plants their coverage ranges from the 25% of total genome length in Arabidopsis thaliana to the more than 75% of Zea mays genome (Flutre, Permal and Quesneville, 2012). In the chromosomes, transposons usually concentrate asymmetrically in highly heterochromatic regions, such as the centromere and telomers (Slotkin and Martienssen 2007). Both the proportion of TEs in eukaryotic chromosomes and their asymmetric concentration are the result of two counteracting forces. On one side, TEs tend to expand and invade the genomes and, on the other side, the genomes "fight back" to preserve their stability, fitness and functionality. We have in fact defined transposon dynamics as the result of this "struggle" between TEs invasion and the molecular mechanisms that the genomes deploy to limit transposon proliferation and preserve their functionality.

More in detail, Figure 4 provides the reader with a swift overview of how and how much the TEs have shaped eukaryotic chromosomes during evolution. Panel A shows the density of genes (blue line), class I transposons (pink line) and class II transposons (green line) in the 8 peach chromosomes. We will see in Chapter 2 that Peach genome is covered by transposable elements for about the 40% of total length. In each chromosome we can recognize that the concentration of genes looks lower in those regions that are more densely covered by TEs, and vice-versa. The black dashed lines per each chromosome indicate the putative positioning of the centromere (Verde et al. 2013). Peach is just an example, but its TEs coverage and distribution are comparable to the ones that have been described in other angiosperms and crops (Flutre, Permal and Quesneville, 2012).

Why do the transposons tend to accumulate in the centromere, and possibly far from the genes? We need to take into account what happens as a transposon “lands” in a specific site to the structure and genetics of the DNA region that is inserted.

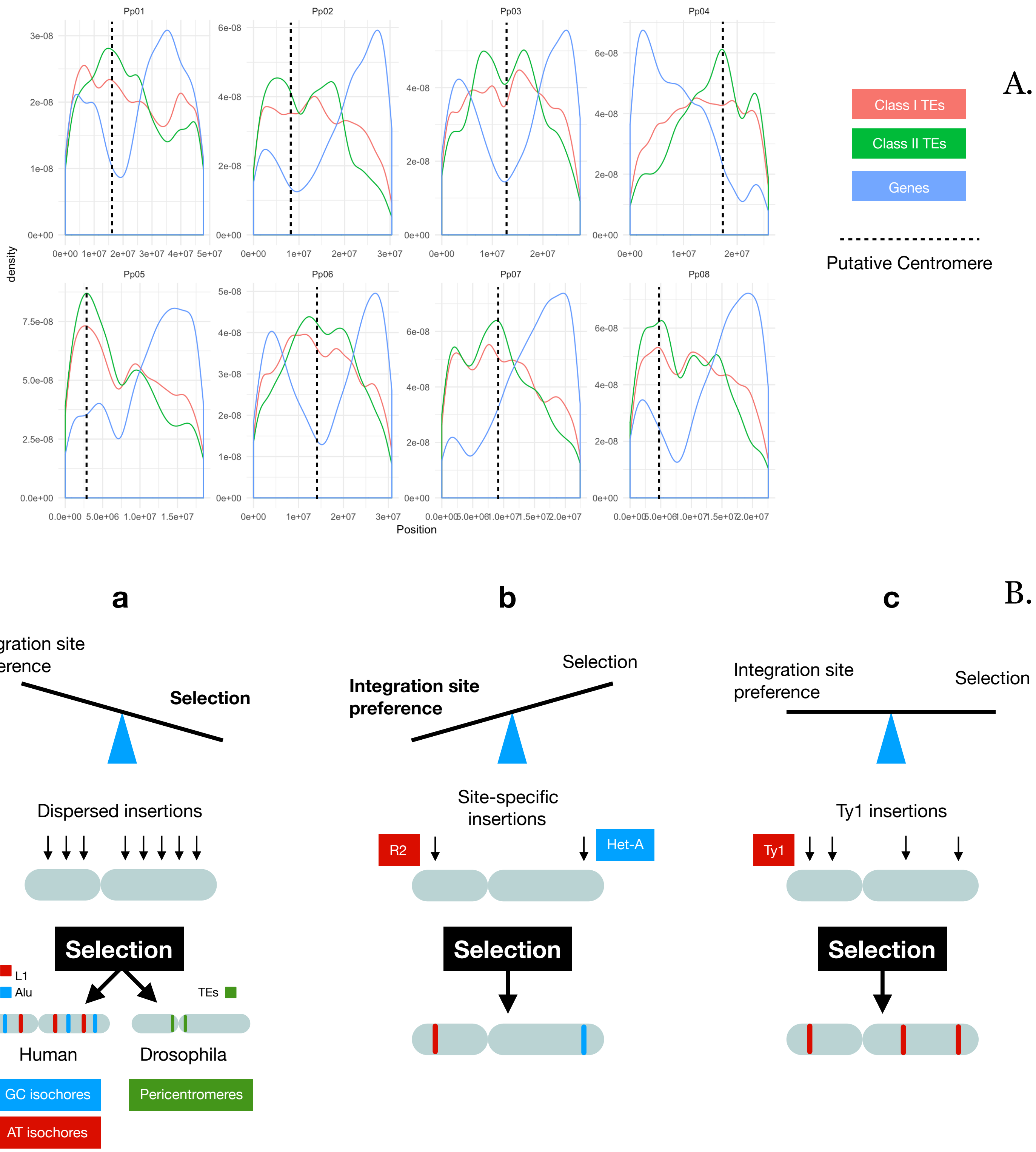


Figure 4 - Impact of TEs on genome structure.

A. (Based on Verde et al. 2013) Gene and TEs density along the 8 chromosomes in Peach (*Prunus persica*) genome. Genes are shown in **blue**, Class I TEs (retro-elements) are in **pink** and Class II TEs in **green**.

B. (Based on Sultana et al. 2017) What causes the asymmetric accumulation of TEs in eukaryotic genomes? Here, the different contribution of Integration Site Preference (ITS) and Selection are compared in three case-limits. In **a** selection is more important than integration site preference, whilst in **b** the integration site preference determines the asymmetry. In **c**, both selection and integration site preference contribute to the accumulation of insertions.

Under a structural and epigenetic point of view, the insertion of a TE causes both the epigenetics and the local status of the chromatin to change in the region that flanks the insertion site (Slotkin and Martienssen 2007). The epigenetic silencing of TEs is a well-known and widely described mechanism of permanent transcriptional silencing that is aimed at preventing transposon proliferation. This silencing occurs as a consequence of strong methylation in plants and transcriptionally repressive histone modifications (Lipmann et al. 2004, Hollister 2009). Changes that can be extended beyond the borders of the transposons or have consequences on the flanking regions, and that are hence impactful on the local gene regulation (Le et al. 2015).

Under a genetic point of view, a new insertion might hit the coding region of a gene, resulting in its disruption and inactivation (Bellen et al. 2004), or within an intron causing possible changes in post-transcriptional regulation (Adachi, Watanabe-Fukunaga and Nagata 1993). Also, TEs can provide new regulatory regions if they insert close enough to a gene (Morata et al. 2018).

The accumulation of insertions in a DNA region can hence result in a series of structural, epigenetic and genetic changes that can strongly modify the expression and regulation of the genes (Bourque et al. 2018). Even though numerous recent evidences confirm that these changes might result being adaptive (Schrader and Schmitz 2018, Lerat et al. 2018), in most of the cases they end up jeopardizing important genetic functions (Lerat et al. 2018). For these reasons, insertions in gene-rich euchromatic regions are usually negatively selected (Pasyukova et al. 2004), and transposons tend to accumulate in heterochromatic regions with a lower gene density (Flutre, Permal and Quesneville, 2012). Indeed, this accumulation itself favours the establishment and consolidation of heterochromatin. Transposons are heavily marked for repressive histone modification, that promote the transition to a closed chromatin state, and the accumulation of insertions in a given region causes this region to turn heterochromatic (Lippman et al. 2004). The segregation between TEs and genes that we observe in the peach genome (Figure 4-A), and that is common among eukaryotes, is indeed the result of the interplay between TEs proliferation and its repression operated by the genome, a long challenge that resulted into shaping the main structural features of eukaryotic chromosomes.

How is this segregation achieved? We have suggested that the insertions in gene-rich euchromatic regions are usually negatively selected because of their strong and potentially harmful impact on gene regulation (Pasyukova et al. 2004), but selection is not the only cause of asymmetric accumulation of insertions. TEs can be selective for insertion sites with a specific consensus sequence or chromatin state. Figure 4-B was published in a 2017 review on the integration site specificity by transposons and integrating viruses (Sultana et al. 2017) and explains fairly well the balance between the contribution of TE's insertional specificity and selection over the establishment of transposons asymmetric accumulation. The first case shown on the left in (a) is when insertion is random, and selection operates by filtering TEs on the base of their chromosomal positioning. In H. sapiens, the non-LTR elements L1 and Alu are enriched in GC and AT rich DNA isochores (Walker and Hurst 2001, Gu et al. 2016). Since this specificity is not observed in novel insertions, the authors suggest that a post-insertional selection mechanism acts to filter L1 and Alu elements (Wagstaff 2012). In Plants, TEs tend to accumulate in the pericentromeric region. Besides those few families that have been proposed to make use of selective insertion mechanisms (and that we're going to discuss), there is no evidence for such a capability in the most of the known TEs, which asymmetrical distribution is deemed to be due to filtering post-insertional selection (Le et al. 2000). Among those TEs that are instead capable to direct their insertion towards certain genomic regions, we can recognise two main mechanisms leading to insertional specificity. In some TE families, insertions are attracted by a particular chromatin or epigenetic status, whilst some other TEs can detect specific DNA motives (Sultana et al. 2017). These two strategies are represented in Figure 4-B.

In (b) the case of Drosophila's HeT-A transposon is reported as an example of epigenetically-driven insertion selectivity. HeT-A is a non-LTR retrotransposon found to be very active in Drosophila and which accumulation in telomeres has been associated with an increase of their stability and resistance to mechanic solicitations (Biessmann et al. 1992). After retro-transcription, the transposon is localized to the nucleus by the open reading frame 1 protein (ORF1p). Before insertion, ORF1p binds the Verrocchio protein, a structural component of Drosophila telomeres that prevents their fusion. This interaction directs HeT-A insertion specifically towards the telomeres (Cicconi et al. 2017). This is not the only case in which the interaction between a retrotransposon protein and a DNA binding protein causes insertional selectivity. In LTR retroelements, for instance, the integrase can interact with several chromatin components. In Ty5 S. cerevisiae element, the phosphorylation of the integrase enhances its binding to the heterochromatin protein silent information regulator 4 (Sir4), increasing the frequency of insertion of this element in heterochromatic regions (Baller, Gao and Voytas 2011). Chromoviruses and their derived elements are characterized by the presence of the chromodomain, an integrase domain that interacts with repressive histone marks such as H3K9me2 and H3K9me3, allowing targeting to heterochromatin (Bannister et al. 2001).

Yeast's Ty1 LTR retroelement binds specifically tRNA-encoding DNA (tDNA) and some other sub-telomeric repetitions. In this case, that is drawn in Figure 4-B (c), the specificity goes for a DNA motiv, that is repeated. The resulting asymmetric disposition is hence due to both the insertional specificity and the filtering post-insertional selection (Curcio, Lutz and Lesage 2015).

TEs cover an averagely relevant portion of eukaryotic chromosomes, attract the heterochromatin formation and dispose asymmetrically along the chromosome, either because of their insertional specificity or as consequence of purifying selection. In the light of these three considerations, is rather difficult to help but linking the TEs expansion during evolution to the shaping of eukaryotic chromosomes (Bourque et al. 2018). Anyways, despite many forces contribute to "push" the TEs towards telomeres and centromeres, far away from gene-rich euchromatic regions, the most of insertions are random, and the possibility for a gene to be inserted or flanked by a TE is non-zero. As we're going to discuss in the next section, the full understanding of the role that these elements played in eukaryotes' evolution behooves us to focus on what happens when transposons and genes come to close liaison.Impact of TEs on gene regulation

The relationship between transposition and gene regulation is all but straightforward. Besides the case in which a TE inserts a gene within its coding region and destroys its capability to encode, the effects of a novel insertion on the insertion site's genetic context are not easily predictable (Le et al. 2015). As we discussed what underlies the asymmetric accumulation of TEs in eukaryotic chromosomes, we pointed out that these elements are usually silenced epigenetically (Fultz et al. 2015, Hokamoto and Hirochika 2001, Hollister and Gaut 2009), and that this silencing has an effect on the local epigenetic status and chromatin conformation of the region they insert. Of course, even though the contribution of the novel element to the definition of the local epigenetics is potentially relevant, many factors can affect the epigenetic status of a DNA region, and the actual contribution of a TE to the global picture is not so immediate to evaluate (Lippman et al. 2004). In the last decades anyways, along with the increasing breakthroughs on TEs, many examples of how these elements changed gene regulation and phenotype have been highlighted.

In this section we will try to sort them into three main categories that are defined by the distance between the insertion and the closest gene. When an insertion happens within a gene sequence, this usually results in a loss of function (Klein and O'Neill 2018). If a TE lands near to a gene, on one hand it can provide new regulatory regions that can directly affect gene expression, as it happens in the case of the color of grapevine fruit skin (This et al. 2007) and, on the other hand, the epigenetic changes that are observed consequently to an insertional event can potentially impact the epigenetic regulation of the genes that are close to the insertion. An exemple of this can be the mail-phenotype promoter gene in flowers CmWIP1 in melon, which expression is mediated by the presence of a transposon upstream (Martin et al. 2009). In both cases, genes can undergo to a modification of usual expression levels, or even to a rewiring into different network if a new regulatory region is provided (Morata et al. 2018).

When a TE inserts a gene: nectarines and seedless apples.

What happens when an inserting transposon lands into a gene? Plant science literature provides us with a couple of clear examples that, as fruit consumers, we might be quite familiar with. In the next paragraphs we will discuss how two very well-known fruit phenotypes, the nectarines and the seedless apples, are caused by the genetic effects of transposition, and more precisely by the disruptive effect of the insertion of a TE in a gene.

Despite in several languages the terms to describe "nectarines" and "peaches" are different, suggesting that these fruits belong to separate species, nectarines are actually peach (Prunus persica) fruits. Their particular phenotype, or else a glabrous skin that lacks the usual peach's trichomes, can arise as a somatic branch-specific mutation (Delgado et al. 2013). The commercial interest in this phenotype is due to the consumers' acceptance of these fruits. This interest caused the growers to breed nectarine varieties, and plant scientists to focus on the genetic mechanism underlying this phenotype. Some recent works have highlighted how transpositional activity can be linked to the loss of fruit skin trichomes.

The gene MYB25 codes for a helix-turn-helix transcriptional factor that probably initiates the developmental process that leads to fruit skin trichome formation in P. persica, and it is homologous to the GLABRA I gene in Arabidopsis and other Brassicaceae. Despite the molecular process underlying fruit trichome development has not been fully clarified yet, it has been demonstrated that the knockout of MYB25 suffices to lead to the nectarine phenotype, suggesting that this gene works as an upstream activator of the whole trichome development. This gene belongs in fact to a monogenic trait (Glabrous, G) that is mapped on peach's chromosome 5. In 2014, Vendramin et al. demonstrated that the nectarine phenotype co-segregates with a large insertion in the 3rd exon of the gene MYB25. This insertion was proven to be a Class I copia-like LTR Retro Transposon (LTR-RT), it is shared by all of the 95 nectarine varieties and determine a glabrous recessive phenotype. Peach is a diploid species, and the nectarine phenotype is caused by the double recessive (g/g) for the G trait. The copia insertion destroys the gene coding region causing a loss of function. However, being heterozygous, it determines a G/g genotype that still results in a non-glabrous phenotype. (Vendramin et al. 2014).

In higher plants, the evolution of edible fruit is linked with the need for a major seed dispersal and enables frugivore animals to be the vector for the spread of new potential individuals (Beck and Wall 2010). Fruit development is hence initiated by pollination and fertilization, that stimulate the cell division of some specific floral tissues. In some cases anyway, the fruit can develop as parthenocarpic. In botany, this term is used to describe "virgin fruits", or else those fruit lacking seeds and which development from the floral buds doesn't require any prior pollination and fertilization (Gillaspy, Ben-David, and Gruissem 1993). In fruit production, seedless fruit variants are highly popular among consumers, and several apple (Malus domestica) varieties, such as Rae Ime, Spencer Seedless and Wellington Bloomless became quite popular for their lack of seeds (Brown 2003).

At the beginning of the 2000s, some genetic studies have identified one single gene (MdPI) to prevent the formation of parthenocarpic fruits in M. domesica. This gene codes for a MADS-box transcriptional factor that is homologous to the pistilata factor (PI) in Arabidopsis thaliana. In Arabidopsis, the loss of its function causes the plant to produce flowers which petals are transformed into sepals and stamens into carpels, and a strikingly similar phenotype is observed in these apple varieties. Further molecular analyses returned that these varieties shared the insertion of an LTR-RT in MdPI gene region. Differently from what happens in nectarine anyways, the insertion was found in two different intronic regions (the 4th and 6th intron) and not in exons. The insertion was found to suppress protein production (Yao, Dong and Morris 2001).

When a TE inserts next to a gene: sex determination in melon, grapes' color and cold stress response in rice.

In previous chapters, we have discussed how a novel insertion can change the regulation of neighboring genes by providing new regulatory regions or changing the local epigenetic profile (Lippmann et al. 2004). In recent years, the research on plant genomes returned some striking examples of how a TE can change the regulation of a gene that stands in the proximity of its insertion site. We will choose three specific examples that will give the sense on how a transposon can change the epigenetic regulation of a gene or providing new regulatory regions that change their expression levels or their wiring into regulatory networks.

Angiosperms' flowers can be either male, female or hermaphrodite. A male flower will bear pollen-producing staminate, a female flower will have ovule-producing carpellate and a hermaphrodite or "perfect" flower will be provided with both male and female organs. A single plant can bear male, female and hermaphrodite flowers in different combinations. In monoecious species a single individual expresses mixed flowers, whilst in dioecious species, single organisms can express unisexual flowers of one gender only (Dahlgren 1989). Commercial melon (Cucumis melo L.) is a monoecious species in which each plant expresses perfect flowers combined with male or female flowers (Garcia-Mas et al. 2012). Each melon variety is actually characterized by the sexual composition of its flowers, and we can divide them into andromonoecious and gynomonoecious depending on whether male or female flowers are expressed (Pitrat 2008).

In this species, the gene WIP1 (CmWIP1) is a transcriptional factor involved in sex determination. Its expression leads to carpel abortion, resulting in the development of unisexual male flowers (Martin et al. 2009). This gene results inactivated in gynomonoecious varieties by the presence of what Martin et al. identified in 2009 as a transposable element. We have pointed out several times how the insertion of a transposon can impress an epigenetic change in the region surrounding the insertion site and that this might occur as a consequence of the strong methylation that might exceed the same borders of the TE and involve the flanking regions. In this case, the gynomonoecious trait co-segregates with an insertion flanking the WIP1 promoter, that was found methylated. Martin and co-workers propose that this methylation leads to the inactivation of the WIP1, thus abolishing the suppression of the female determination pathway (Martin et al. 2009).

With a more direct effect on their regulation, TEs can provide the genes they flank with new regulatory sequences. A recent paper published in the lab that guested this Ph.D. project showed how the spread of MITE elements in peach (Prunus persica) and Chinese Plum (Prunus mume) might have involved a global re-wiring of new genes into several regulatory network because of the enrichment of these transposons with different Transcription Factor Binding Sites (TFBS). Adding a new regulatory element to a gene has two potential consequences: the change in the expression level and the change in regulation. A new promoter can either or both re-modulate the level of expression of the controlled gene and change its activation pattern, and plant science provides us again with some relevant examples (Morata et al. 2018).

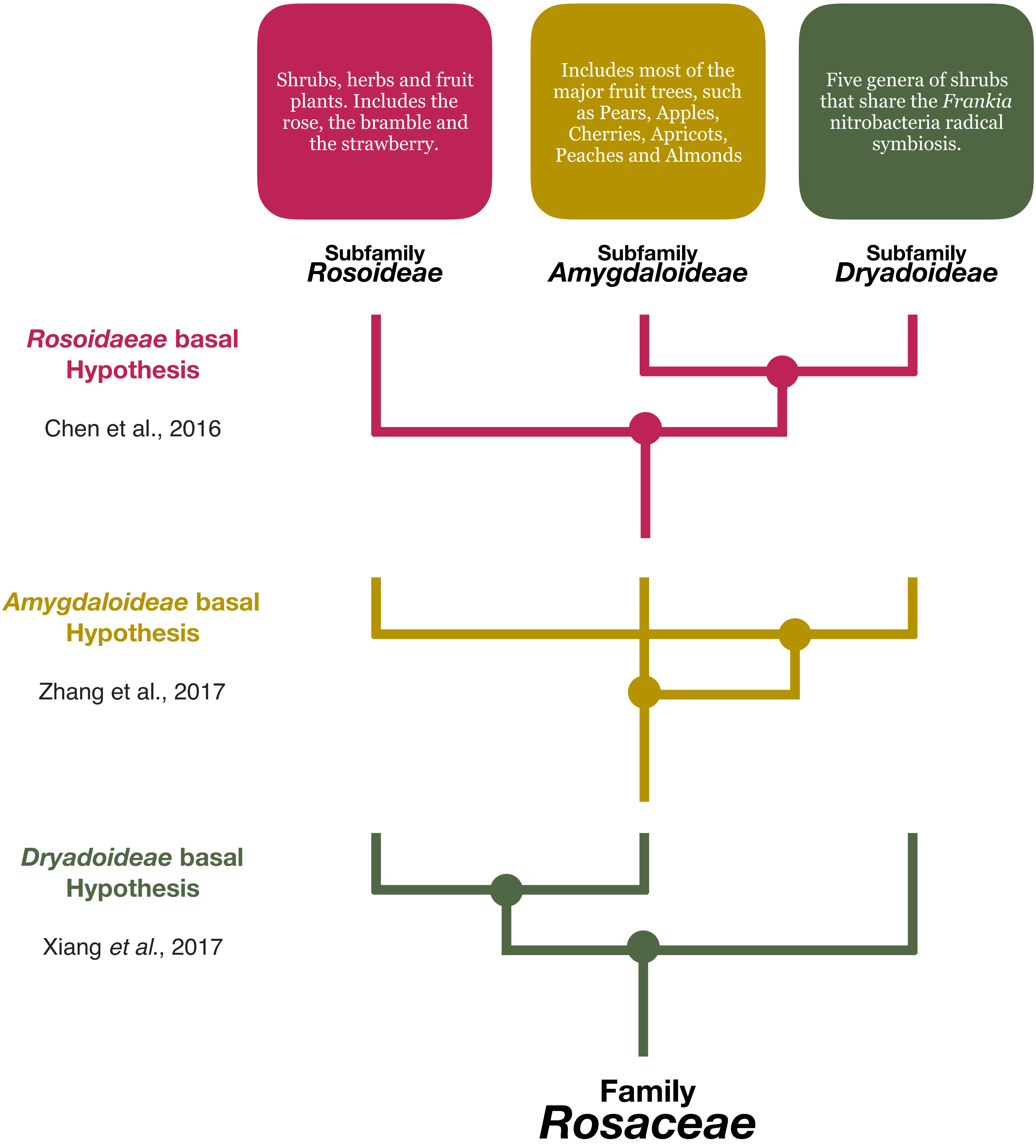
“Grapevine” is a term of common use to indicate the 79 accepted species in the Vitis genus. Some of these species are bred as crops for wine production or direct consumption and selected for phenotypic traits of interest, such as the color of the skin of the fruit (Güner et al. 2008, Lachman et al. 2015). In 2007, a study published by This et al. clarified that the different color shades present in commercial grapevine are due to a TE. The color of grapevine drupes is due to the accumulation of anthocyanins in the fruit skin (Castellarin et al. 2007). The gene that activates the anthocyanin biosynthesis pathway is a MYB- Helix Turn Helix transcriptional factor identified as VvmybA1 (Jeong et al. 2006). In the regulatory region of this gene, a Ty3-gypsy-like LTR retroelement named Gret1 was identified in the promoter region of this gene in those grapevine varieties having poor or absent coloration of the drupe skin (white grapevine or blanc). This led to the conclusion that grapevines with red/dark skin (noir and gris) express the native phenotype, whilst the decrease of the pigmentation in blanc varieties was due to the change in promoter sequence caused by Gret1 insertion. Intriguingly, varieties harbouring an intermediate light red/pink color (rouge and rose), were found having a Gret1 solo-LTR in the promoter region. Solo-LTRs are the result of internal recombination, that occurs quite often in LTR-Retroelements and that results in the deletion of most of the transposon except for one of the two LTRs . In this case, the anthocyanin production is attenuated but not completely shut, leading to a light red /pink color. The authors of this work propose that the change of the promoter sequence caused by the Gret1 insertion and the further re-modulation caused by its partial depletion influence the expression level of VvmybA1 (This et al. 2007).

Besides the promoter, genes’ upstream regions usually host regulator sequences that serve to activate or repress the expression in response to specific stimuli that are mediated in trans by DNA-binding proteins. TEs can provide new binding sites that are recognised by these proteins and that can submit the gene to a new regulatory network. This is the case of the mPing transposon identified in rice. As reported in the aforementioned work by Morata et al., MITE elements can diffuse TFBSs through the genome and rewire the genes to form novel regulatory networks. The mPing element is a MITE as well, and it contains putative cold-stress responsive cis-elements within its sequence (Naito et al. 2009). As extensively reported by Yasuda et al. in 2013, the insertion of a mPing next to a gene suffices to render this gene cold-inducible.

Part B – Overview on the *Rosaceae* fruit crops.

The family Rosaceae collects about 5000 species and 91 genera of flowering eu-dicotyledon plants. The family is divided in three subfamilies, which have been proven to be monophyletic, but which mutual evolutionary relationships still remain controversial (Soundararajan Won and Kim 2019). Subfamily Rosoideae consists of about 850 species, that include shrubs as the rose, herbs, and fruit plants such as strawberries and brambles (Hufford 1992). The small subfamily Dryadoideae only collects four genera of shrubs, and small trees. These species share root nodules that host the nitrogen fixing bacteria Frankia and a fixed number of 9 chromosomes (Gajewski 1999). The third genus we mention is the Amygdaloideae, a wide and diverse genus that collects several relevant fruit crops such as Pears, Apples, Cherries, Apricots, peaches and almonds and ornamental plants such as Spiraea and Aruncus. A distinctive trait of this genus is the shape of the fruit, that is usually fleshy and soft outside with a hard-shelled seed inside (Das Ahmed and Singh 2017). This seed is usually called stone or pit, reason why these fruits are called stone fruits or drupes, even though a differentiation is made for the fruits of Pear and Apples that carry smaller seeds and are often called pome fruits (Velasco et al 2010).

The phylogenetics of this family is rather controversial. Although there is a wide agreement on the boundaries of this family, the internal subdivisions have been subject of a long debate. The definition of the three families was achieved after a controversy that crossed all the 1960s and 1970s and even if there is some consent on dividing the family into the three subfamilies we have described, there is still debate on the number of the genera and on their mutual evolutionary relationship (Potter et al. 2007). Three hypotheses have been proposed, one for each family that is considered as being the basal group. In Figure 5, we compare the three hypotheses. The Rosoideae basal hypothesis was the first to be proposed in 1994 by Morgan et al., and has been recently confirmed by the phylogenetic analysis of mitochondrial proteins (Morgan et al. 1994). The Amygdaloideae basal hypothesis is more recent and found some confirmation in 2017 through plastid whole genome alignment analysis (Zhang et al. 2017) but, in the same year, an alternative hypothesis was proposed by Xiang et al. Based on the comparison between nuclear transcriptomes, the Dryadoideae have been proposed as the basal group (Xiang et al. 2017).



Shrubs, herbs and fruit plants. Includes the rose, the bramble and the strawberry.
Includes most of the major fruit trees, such as Pears, Apples, Cherries, Apricots, Peaches and Almonds
Five genera of shrubs that share the *Frankia* nitrobacteria radical symbiosis.
Subfamily *Rosoideae*
Subfamily *Amygdaloideae*
Subfamily *Dryadoideae*
Rosoidaeae basal Hypothesis
Chen et al., 2016
Amygdaloideae basal Hypothesis
Zhang et al., 2017
Dryadoideae basal Hypothesis
Xiang *et al.*, 2017
Family *Rosaceae*

Figure 5 - Classification and evolution of *Rosaceae* subfamilies

Rosaceae subfamilty is divided into three major subfamilies: *Rosoideae, Amigdaloideae* and *Dryadoideae*. These three families have been proposed monophyletic (), but their evolutionary relationship remains unclear. Here, we compare the three hypotheses proposing different connections.

The Prunus genus

Most of this thesis work was developed within a collective effort to sequence and annotate the almond (Prunus dulcis) genome (Alioto et al. 2019) and is based on the comparison between the transpositional dynamics of almond with the one in its sister species peach (Prunus persica). Therefore, to give an overview of our workframe, it ought us to focus our discussion on the genus that collects the species of interest for this project, or else the Prunus genus.

Included in the Amygdaloideae subfamily along with other crops genera such as Apple (genus Malus) and Pear (genus Pyrus), Prunus is the widest and most diverse genus in the subfamily. The group was proven monophyletic by a study that compared DNA sequence polymorphisms in cloroplast tnrL-tnlF spacer in 48 species (Bortiri et al. 2001). The study proposed a common Eurasian origin for the genus. Historically, the genus was treated by many authors as broken into several genera, and this work enforced the view of a unique genus divided into more sub-genera, already proposed by the american dendrologist Alfred Rehder in 1940 (Rollins 1951). Rehder proposed to divide Prunus into five subgenera: Amygdalus, Prunus, Cerasus, Padus and Laurocerasus. Later on, the subgenus Lithocerasus was proposed and included as accepted subgenus (Gradziel 2003).

Subgenus Prunus. It includes Apricots (Prunus armeniaca) and Plums (Prunus domestica) and it is characterised by the presence of a groove on one side of the fruit. Adult plants grow as trees (Das Ahmed and Singh 2011).

Subgenus Cerasus. The subgenus collecting the “cherries”, or more specifically the Sour Cherry (Prunus cerasus) and the Sweet Cherry (Prunus avium). Fruits are not grooved, and the stone is smooth. Adult plants grow as trees (Ohta et al. 2005).

Subgenus Lithocerasus. The best known species of this subgenus is the North-American shrub known as Sand Cherry (Prunus pumila). Fruits are not grooved and the stone is smooth. Adult plants grow as shrubs (Shimada et al. 2001).

Subgenus Padus. Bird Cherries (Prunus padus) are trees that are known to produce flowers in racemes. Divided in the two main varieties European and Asian Bird Cherries, they are diffused in the humid continental climates of Eurasia. Their fruit have a deep astringent flavour that is not perceived by the birds, the only animals that eat the small not grooved fruits (Kader and Proebsting 1992).

Subgenus Laurocerasus. Whilst all the sub-genera collect deciduous species, the subgenus Laurocerasus is mostly constituted by evergreen species. The two cultivated species (Prunus carolinana in North America and Prunus laurocerasus in Europe and the UK) are appreciated for their ornamental qualities (Kalkman 1965).

Subgenus Amygdalus. This latter is the genus that collects the species that we're going to focus on during this dissertation: peach (Prunus persica) and almond (Prunus dulcis). Along with other closely related species, peaches and almonds flower in early spring to produce a fruit that is grooved on one side and a characteristic deeply grooved stone (Verde et al. 2013, Velasco et al. 2014).

The evolution of peach and almond

Charles Darwin mentioned peach and almond as an example of how species can separate and change. Its famous quote from 1868 says that "peaches are almonds modified in a beautiful manner", and it intended to remark how the two crops have undergone different modifications from the same ancestor (Yazbek and Al-bein 2014). Over the time, several botanical, paleontological and genetic (Verde et al. 2013, Velasco et al. 2014, Yazbek and Oh 2013) insights have shed some light on the evolutionary history of these two crops and of the Amygdalus subgenus they belong.

Available results are consistent in indicating that the geological movements caused by the collision between the Indian and Eurasian tectonic plates, and that caused the birth of the Himalaya (Wagner 1993), have provided the physical conditions for the separation between peach and almond to happen (Yu et al., 2018). Around 10 Million Years Ago (MYA), during the second half of the Miocene, the Indian plate collided with the Eurasian plate, triggering the rise of the mountain system that includes the Himalaya and that is known as the Tibetan plateau (Yang et al. 2016). This event favored the geographical isolation of part of the population of the ancestor species, that occurred in concomitance with the radical regional climate change that was caused by the geological rearrangements (Yu et al., 2018). These rearrangements defined a region that corresponded to the current South-Western China, where the ecological changes caused a reduction in the population of the ancestor and impressed those morphological and biological changes that led to the speciation of peach.

Similarities and differences between peach and almond

The comparison between peach (Prunus persica) and almond (Prunus dulcis) is particularly interesting for an evolutionary study because of the presence of sharp and recognizable differences at various levels even though the two species are barely separated. We have already discussed how there is some wide agreement on the idea that the separation between these two species was consequent to the rise of the Tibetan plateau (Yang et al. 2016), and the actual speciation has been proposed to be eventual to that event, with all the authors placing it between 4-10 MYA (Verde et al. 2013, Velasco et al. 2014, Yu et al. 2018, Alioto et al. 2019). However, breeding studies have confirmed that is possible to cross the two species and obtain vital and fertile crosses, with a rough success probability of the 50% (Jáuregui et al. 2001), confirming peach and almond high proximity. We will now swiftly go through what is common between the two crops and what differ under different point of view.

Botany. peach and almond adult plants grow into trees that can reach a 7 m height, with an average of 3-4 m tall and wide for cultivated varieties. Leaves are lanceolate, 7-16 cm long and 2-3 cm broad with a pinnate veining. Flowers are solitary or paired, 2.5 - 3 cm diameter, from white to pink and with five petals. peach and almond trees are so similar that can be distinguished only by some details. almond trees are bigger in size and stronger in vigor, being way less affected by diseases and pests than peach trees. Even though the flowers are very similar, almond flowers are slightly smaller and fragrant, differently from the unscented peach flowers (Gradziel 2009). Of course, the best-known difference between peach and almond regards the fruit. peach fruits ripe into a fleshy and juicy fruit with a developed mesocarp, whilst almond fruits have a smaller and dry mesocarp, but grow a fleshy endosperm. About this latter part of the fruit, one relevant difference is in the concentration of cyanide, that is high in peach and bitter almond seeds, but absent in the sweet commercial almond.

Flowering and reproduction. In the Mediterranean, peach and almond trees are the first to flower in late winter/ early spring. They are both monoecious species that produce hermaphrodite flowers (George and Niessen 1992), but they differ in the reproductive strategy. The species of the Prunus genus are mostly self-incompatible and share a common molecular mechanism that prevents the auto-fecundation (Tao et al. 2007). Whilst the very most of almond cultivated varieties are self-incompatible, peach varieties are mostly auto-fertile, and this condition seem to have arisen eventually.

Genomics. The Prunus genus is characterized by a 8 chromosome genome that is sized between 200 and 300 MBases (Verde et al. 2013, Alioto et al. 2019). peach and almond genomes are about 250 MBases long with 8 chromosomes as well. The two species are strictly diploid, and the neatest difference at genomic level described so far was proposed by Velasco et al. in 2014, that have calculated a 7-fold higher intraspecific nucleotide diversity in almond than peach. The reasons of such a lower diversity have been proposed to be found in either or both of the lack of self-incompatibility mechanism and the bottleneck that has preceded peach speciation (Velasco et al. in 2014).

Chapter 1 - Software for Transposable Elements detection from DNA-seq data. A benchmarking.

1.1 Introduction

Transposable Elements (TEs) usually cover a large proportion of the genome of many species, and the research on their impact over evolution and genomic regulation is strongly dependent on the possibility to identify and characterize them de novo by computational means from genomic data. To fully understand how TEs can be identified from the sequenced genome, we need to discuss how genome data is processed and made available (Ewing 2015).

The revolutionary introduction of high throughput DNA sequencers during the second half of the 1990s brought to a new era the genomic research. The new technology was based on the possibility to fraction the DNA and sequence a high amount of short fragments (50-200 bp). Computational alignment algorithms would have been deployed eventually to reconstruct the whole sequence. About ten years later, a new generation of sequencers was able to read longer DNA fragments (15000 bp), increasing sensitively the power of sequencing resolution (Goodwin, McPherson and McCombie 2016). From Sanger sequencing (Sanger and Coulson 1975), technological development has brought radical changes during the last 20 years. Usually, we refer to the technology based on the sequencing of short DNA fragments or "short reads" as Next Generation Sequencing or NGS (Schuster 2008). The long-reads based technologies are usually associated with companies like PacBio™ and Oxford Nanopore™ and globally indicated as Third Generation Sequencing or TGS (Schadt, Turner and Kasarskis 2010). Their diffusion increases proportionally to the decrease in cost per sequence. Even though the TGS technologies have a better resolution, their relatively higher costs have slowed down their spread (van Dijk at al. 2018). Most of the genomic data is now available as NGS short-reads distributions. The sequencing of genomic DNA results in data distributions that are usually defined as DNA-seq or Whole Genome Sequencing - WGS. The possibility to rapidly sequence whole genomes increased the capability to resolve the sequence of entire genomes. The short reads are aligned and piled together to be assembled into one sequence per chromosome (pseudomolecule) or large fractions of it (scaffold), and the sequence is provided as WGS distribution.

In the last 20 years reference genomes have been produced for many different species. However, in most cases for each species only one or few assembled genomes are available, whilst most of genomic variation is studied through the combination of high-quality and high-throughput sequencing approaches (Alkan, Sajjadian, Eichler 2011; Cirulli and Goldstein 2010). For instance, a typical population genomics study will involve the sequencing of several tens of samples coming from different individuals, just as a cancer genomics study will require the sequencing from different cells or tissues extracted from different patients (Mardis and Wilson 2009). These samples will be treated as non-assembled WGS to be aligned to the reference genomes to detect their variability.

This has the consequence that most of the genomic data is available as non-assembled WGS files, that contain the information about the nucleotide diversity and structural variation within a species, whereas the variation that is often the consequence of the movement of mobile elements. Hence, the identification of TEs from genomic data can't be a mere matter of annotating TEs loci on an assembled genome fasta files but requires tools to detect mobile elements from non-assembled WGS data (Ewing 2015). To understand how transposition contributed to the generation of genetic variability, we need to know which transposable elements are present in all the samples from a given population and which of them are present in some samples only and are polymorphic in the population (Handsaker et al 2011).

The software packages for TE insetion polymorphisms identification from WGS files are designed to link the structural variation (SV), that is detected by comparing WGS samples to a genome of reference, to transposition. Usually, after the identification of proofs of SV from aligned WGS datasets in SAM or BAM format, these programs try to select those putative structural changes that are connected to transposition by making use of either or both of a TEs annotation on the reference genome and the consensus sequence in fasta or one or more TE families. Some programs use this information to highlight the absence in samples of TEs annotated in the reference, some programs try to identify TEs that are present in the sample but not annotated in the reference, even though the most of the currently available solutions try to do both (Goerner-Potvin and Bourque 2018).. Also, whilst some programs prefer to focus on one TE family at the time, requiring the sequence of the consensus (or an instance) of a specific family, other programs return the SV associated to more family at the same run. Basing on whether a package returns the presence of annotated transposons or detects new transposons from WGS samples, and on the number of detected families per run, we operated a bold classification of the most used available packages in Figure 1.1.

The programs that detect multiple families link the SV found in aligned WGS files with transposition on the basis of their overlap with a TEs annotation that is provided as input. Those packages that limit their output to the detection of the absence of reference insertions, such as T-lex2, base their prediction on the hallmarks of gaps in samples, such as the split reads (Stewart et al, 2011, Fiston-lavier et al. 2015), that

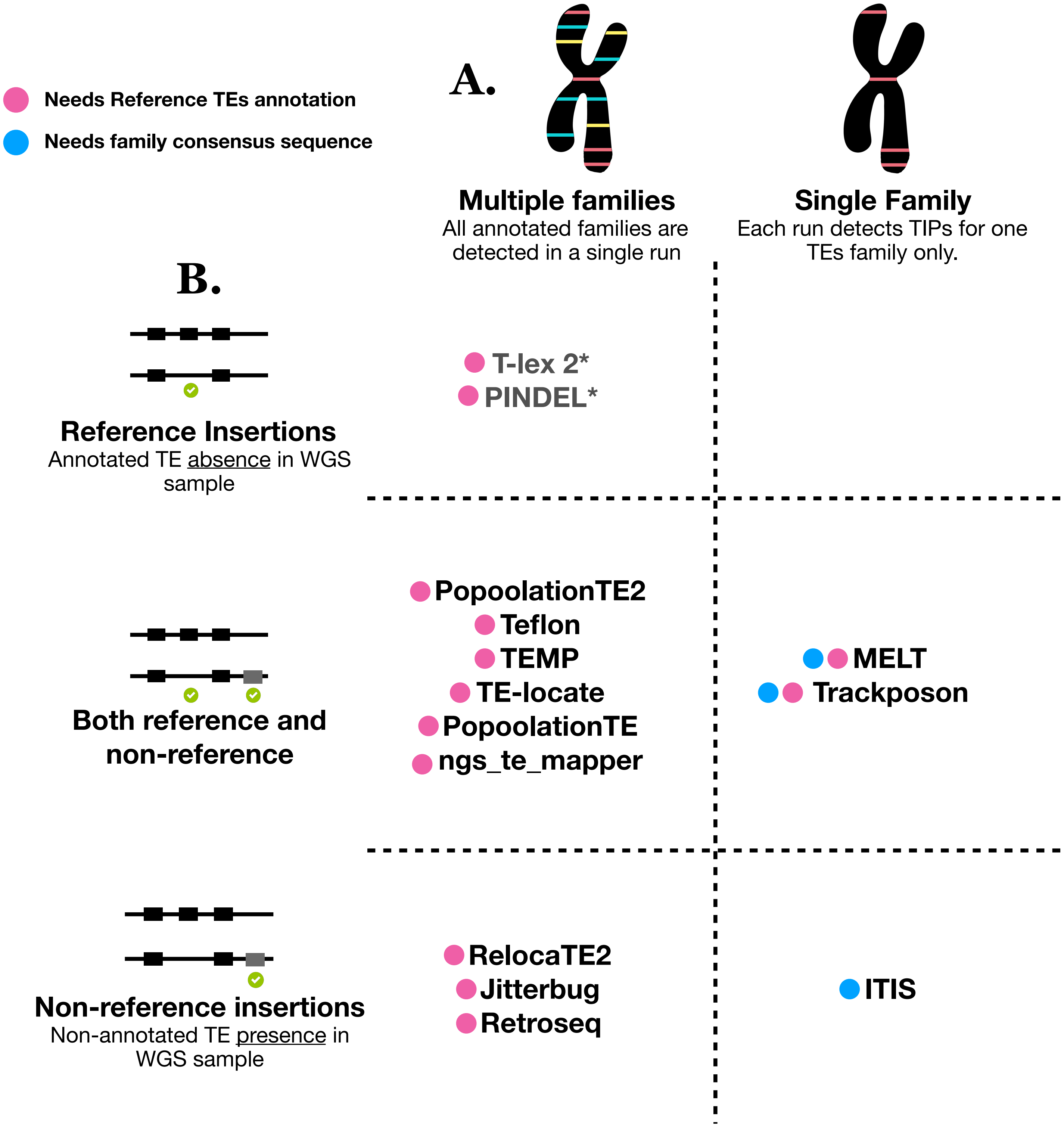

● Needs Reference TEs annotation
● Needs family consensus sequence
A.
Multiple families
All annotated families are detected in a single run
Single Family
Each run detects TIPs for one TEs family only.
B.
Reference Insertions
Annotated TE absence in WGS sample
● T-lex 2*
● PINDEL*
Both reference and non-reference
● PopoolationTE2
● Teflon
● TEMP
● TE-locate
● PopoolationTE
● ngs_te_mapper
●● MELT
●● Trackposon
Non-reference insertions
Non-annotated TE presence in WGS sample
● RelocaTE2
● Jitterbug
● Retroseq
● ITIS

Figure 1.1 - WGS TE identification tools sorted by output.

A. **Number of detected families.** WGS TE detection tools can either return TIPs that belong to different families or be focused on one family per run.

B. ***Reference* or *non-reference* insertions.** TE annotation- based programs can be further sorted into programs detecting the presence or absence of reference TEs in the sample or the presence of TEs in the samples that are not annotated in the reference.

* These packages have not been included in this benchmarking.

overlap to transposon annotation. Those packages that try to detect the presence of TEs in WGS samples that are not annotated on the reference genomes rely on the same principle, except for the difference that they look for structural hallmarks that indicate an insertion. The package Jitterbug, for instance, bases its predictions of structural variation on reads that indicate the possible presence of insertion, such as soft-clipped and discordantly mapping reads. The putative SVs that are finally linked to the presence of a TE are those ones having a minimum number of discordant read couples where one read maps to an annotated TE (Henaff et al. 2015). As we can see in Figure 1.1, most of the available packages are designed to tackle the detection of annotated and non-annotated TEs. The packages that are designed to run with a single family request the user to provide a fasta sequence that is representative of the family (an instance, a consensus or a centroid). MELT requires also an annotation in bed format of all the instances of the family.

A further classification can be made by sorting the software in the light of the length of the reads they accept as input. Actually, the length of the reads represents an important limiting factor for WGS- based TEs identification, as long-reads allow, in principle, a better detail in the annotation of repetitive regions. The package LorTE can accept Pac-Bio@ TGS long reads (Disdero and Filée 2017) and the creation of SV detection tools from long-reads WGS samples is increasing along with the spread of this technology. However, as most resequencing data of populations and varieties is at present based on short-read technologies, this chapter unfolds around the methods that are classified in Figure 1.1 and that are designed to work on short-reads WGS data. Although long-reads technology is, by all means, promising in increasing our capability to detect structural variation from WGS data, the very most of available DNA-seq datasets are based on short-reads (Leinonen et al 2010). To date, the efforts to validate the predicted polymorphic TEs outputted by these packages (Rishiwar et al. 2016) have been partly jeopardized by the difficulty to obtain a comparison based on real data instead of simulated ones (Hoen et al. 2015).

In this, the assemblies produced for the Rice genome (Oryza sativa) can turn as rather useful. Having been the first crop to be assembled in the year 2000, different rice genome have been sequenced and assembled to better investigate its high intraspecific diversity (Jackson 2016). Among the available reference assemblies, the one produced for Nipponbare (Oryza sativa var. Japonica) (Kawara et al. 2016) and Minghui 63 (Oryza sativa var Indica) (Zhang et al 2016) varieties are assembled with outstanding quality. Along with the genome assembly in fasta format, the WGS datasets that served as the base for the assembly are available as well. The presence of two assemblies of genomes of the same species along with the raw WGS data represents a very favourable setting to run an effective validation of WGS TEs identification tools, that will take the comparison between the assembled FASTA genome files as a golden standard to compare their results.

1.2 Scope of the Chapter

This chapter will discuss the results of a comparative benchmarking and validation analysis of 12 WGS-based TEs detection tools. After having annotated Class I LTR-Retro Transposons (LTR-RTs) and Miniature Inverted-repeat Transposable Elements (MITE) on the reference genome of Nipponbare (Nip) and Minghui 63 (MH63) Rice varieties, fixed and polymorphic elements have been identified on the chromosome 5 of both assemblies through pairwise alignment and manual curation. This information is then used as a golden standard to validate the prediction of inspected tools. The raw WGS files that were used to perform the two reference genomes have been aligned with the homologous assembly (MH63 WGS to Nip reference) at different coverage levels. Resulting alignments have been used as inputs to the 9 packages along with either MITE and LTR-RTs annotations or the consensus sequences depending on package requirements (Figure 1.1).

The performance of each program is compared as the ratio of true positives over the total identified TEs (precision), as the ratio of true positive over the total TEs in the golden standard (recall) and as F1 score (F1 = 2 * precision*recall/precision+recall) (Powers and Aliab 2011).

Our goal is to draw a fair view of the actual effectiveness of these methods in detecting TEs from WGS data and to compare them under the variation of the input coverage and with different superfamilies of TEs (LTR-RTs and MITEs).

1.2.1 Note on the contribution of the candidate

What is presented here is the result of a collective work for which the PhD candidate has contributed. Besides the results from the whole project are presented, it is necessary to inform the commission and the readers that my contribution to this project is specific to the following points:

- Software selection and review of algorithmic structure
- De-novo annotation of LTR-RTs from Oryza sativa v. japonica (Nip) reference genome.
- Run and Benchmark of the Jitterbug and MELT pipelines

After having classified and discussed the tools included in the benchmarking, the dissertation will hence focus on the results that have been obtained by the candidate to finally provide a swift overview on the global results of the project.

1.3 Materials and Methods

1.3.1 LTR and MITE annotation on reference genomes

LTR-retrotransposons were identified by running LTRharvest with default parameters. The internal conserved domains of these elements were obtained running HMMscan from the HMMer suite (Potter et al. 2018), and only coding elements were retained for further analyses. The identified elements were clustered with Silix (Miele, Penel and Duret 2011) in sets of sequencing that mutually shared the 80% of sequence identity over the 80% of sequence length. All the elements in each family were aligned with Mafft (Katoh and Standley 2013) and trimmed with Trimal (Capella-Gutierrez, Silla-Martinez and Gabaldón 2009). Consensus sequences were built from the alignments using the EMBOSS package (Rice, Longden and Bleasby 2000). MITE-hunter (Han and Wessler 2010) was run on Nip and MH63 assemblies to detect potential MITEs families, which were then combined with the high-quality predictions available in PMITE database (families carrying TSD) (Chen et al. 2014). Clustering at 90% was performed to remove redundancy using cd-hit2 (Li and Godzik 2006) and produce a final library. RepeatMasker (http://www.repeatmasker.org/) was run to annotate all regions having significant homology with any of the MITE families. The annotations were further screened to discriminate full-length elements (consensus length ± 20%) from truncated hits.

1.3.2 Determination of Benchmarking Standards

Several assemblies have been produced for the Rice (Oryza sativa) genome. We have selected two of them for their better quality; Nipponbare (Nip, Oryza sativa subsp. japonica) (Kawahara et al. 2013) and Minghui 63 (MH63, Oryza sativa subsp. indica) (Zhang et al. 2016). We used Nip as a reference to obtain a curated dataset of “reference” orthologous and “non-reference” (specific to MH63) insertions. From the insertions found in MH63, we selected 500 nt upstream and downstream flanking regions and mapped to the Nipponbare reference through NCBI Blast alignment (Altschul et al. 1990). The distance between the upstream and downstream flanking hits gave information on the possible conservation of the MH63 insertions in Nip (distance similar to the size of the insetion), or the absence of MH63 insertion in Nip (distance close to zero). The mapping of genome windows to MH63 genome was performed using BBmap (https://sourceforge.net/projects/bbmap/). Intersections between annotations have been done with BEDtools (Quinlan and Hall 2010).

1.3.3 Polymorphism predictions

Predictions of polymorphisms were done using the 12 tools that are described in Table 3.1. Default parameters or recommendations of the authors have been used to run the programs.

1.3.4 Evaluation parameters

The ability of each tool to detect MITEs and LTRs was evaluated in terms of number of True Positives (TP), false positives (FP) and False negatives (FN). The comparison between the annotation on the reference genome and the actual gaps identified through whole genome alignment allowed us to build a curated dataset of True and False Positives. The result of each tool was compared with these two datasets and assigned to a value fo TP, FP and FN according to the following criteria:

- **True Positives, TP** are insertions detected by any tool matching with our curated dataset of TPs.
- **False Positives, FP** are insertions detected by any tool matching with our curated dataset of FPs.
- **False Negatives, FN** are insertions present in our curated dataset of TPs, not detected by the evaluated tool

These parameters was then processed to calculate the following performance descriptors:

- **Sensitivity** = TP/(TP+FN)
- **Precision** = TP/(TP+FP)
- **F1 Score** = 2 x [(Precision x Sensitivity) / (Precision + Sensitivity)] 1.4 Results

1.4 Results

1.4.1 Tools selected within the benchmarking project.

The tools that have been selected for this benchmarking are reported and briefly discussed in Table 1.2. For each tool, we describe the technical features in A and provide the coordinates to retrieve and test the software in B. The features we describe are basically three. First, the characteristics of the output are reported in the pink- labelled columns. The number of detected families per run (single for one family per run and multiple for 2 or more families) and whether the program reports the absence in samples of insertions that are annotated in the reference or the presence in samples of insertions that are not annotated in the reference. File formats columns' headers are highlighted in blue. These columns report the information on which file is accepted as input (either fastq or bam) and what kind of output is returned (bed, vcf, gff or more than one format).

The last columns describe the perceived difficulty in the installation and input setup and are highlighted in yellow. We have labeled as easy those packages which installation is automatic or semi-automatic, medium those packages that requires several dependencies or a specific versions of packages, and difficult those packages with advanced users operations required (e.g. compiling from source code).The difficulty associated to the input preparation was labeled in reason of whether the package accepts common file formats - such as bed annotations or sequences in fasta files (easy) - or requires some specific (medium) or very specific (difficult) input file formatting.

The tools that are included in this benchmarking are 12. ITIS (Jiang et al. 2015), MELT (Gardner et al. 2017) and Trackposon (Carpentier et al. 2019) are designed to detect TIFs associated with a single family per run. ITIS is designed to detect non-reference insertions, whilst MELT and Trackposon return both reference and non-reference insertions. Among the tools that are designed to detect multiple insertions RelocaTE2 (Robb et al. 2013), Jitterbug (Henaff et al. 2015) and Retroseq (Keane, Wong and Adams 2013) are specific for returning non-reference insertions. Instead, Popoolation TE2 (Kofler, Gómez-Sánchez and Schlötterer, 2016), Teflon (Adrion et al., 2017), TEMP (Zhuang et al., 2014), TE-locate (Platzer, Nizhynska and Long, 2012), Popoolation TE (Kofler, Gómez-Sánchez and Schlötterer, 2016) and ngs_TE_mapper (Linheiro and Bergman 2012) return both reference and non-reference insertions. The package McClintock (Nelson, Linhero and Bergman 2017) deserves a special mention since it is not an actual TIF discovery tool but a computational framework that merges several packages and compares the results. Retroseq, TEMP, TE-locate, PopoolationTE and ngs_te_mapper have been run within McClintock.

A. Software Specifications

Package Name	Output		File formats		Perceived difficulty	
	Number of detected families per run	Reference or non-reference insertions	Input (WGS)	Output	Installation	Input setup
ITIS	Single	Non-reference	Fastq	Bed	Easy	Medium
MELT	Single	Both	Bam	Vcf	Easy	Medium
Trackposon	Single	Both	Fastq	Bed	Easy	Easy
RelocaTE2	Multiple	Non-reference	Fastq	Gff	Easy	Easy
Jitterbug	Multiple	Non-reference	Bam	Gff	Medium	Medium
PopoolationTE2	Multiple	Both	Fastq	Multiformat	Easy	Easy
Teflon	Multiple	Both	Fastq	Multiformat	Medium	Medium
McClintock*	Multiple	Both	Fastq	Bed	Easy	Difficult
Retroseq**	Multiple	Non-reference	Bam	Vcf	Easy	Difficult
TEMP**	Multiple	Both	Bam	Multiformat	Easy	Difficult
TE-locate**	Multiple	Both	Sam	Multiformat	Easy	Difficult
PopoolationTE**	Multiple	Both	Fastq	Multiformat	Easy	Difficult
ngs_te_mapper**	Multiple	Both	Fastq	Bed	Easy	Difficult

B. Software availability

Package Name	Documentation and/or download	Bibliographic Reference
ITIS	https://github.com/Chuan-Jiang/ITIS	Jiang et al. 2015
MELT	http://melt.igs.umaryland.edu/manual.php	Gardner et al. 2017
Trackposon	http://gamay.univ-perp.fr/~Panaudlab/TRACKPOSON.tar.gz	Carpentier et al. 2019
RelocaTE2	https://github.com/JinfengChen/RelocaTE2	Robb et al. 2013
Jitterbug	https://github.com/elzbth/jitterbug	Henaff et al. 2015
Retroseq**	https://github.com/tk2/RetroSeq	Keane, Wong and Adams 2013
McClintock*	https://github.com/bergmanlab/mcclintock	Nelson, Linhero and Bergman,2017
PopoolationTE2	https://sourceforge.net/p/popoolation-te2/wiki/Manual/	Kofler, Gómez-Sánchez and Schlötterer, 2016
Teflon	https://github.com/jradrion/TEFLoN	Adrion et al., 2017
TEMP**	https://github.com/JialiUMassWengLab/TEMP	Zhuang et al., 2014
TE-locate**	https://sourceforge.net/projects/te-locate/	Platzer, Nizhynska and Long, 2012
PopoolationTE**	https://omictools.com/popoolation-te-tool	Kofler, Gómez-Sánchez and Schlötterer, 2016
ngs_te_mapper**	https://omictools.com/ngs-te-mapper-tool	N.A.

Table 1.1- List of tested packages

A. The table describes the main features of the software that was tested in this benchmarking. For each package, it is described wether the algorithm detects one or more families per run (column 2), whether identifies *reference*, *non-reference* insertions or *both* (Column 3), along with the file formats for input and output (Column 4). The last column (Column 5) describes the perceived difficulty during installation and the preparation of the input. In **B.** the links to download the packages and their documentation are provided.

* *McClintock* is an integrated pipeline that runs different packages to identify TEs

** These packages have ben ran within *McClintock*

1.4.2 Detection strategy in MELT and Jitterbug

Although all the software packages that are mentioned in Table 1.1 have been tested in this benchmarking project, the contribution of the candidate authoring this manuscript revolves specifically on the testing of two packages: Jitterbug and MELT. A thorough look to their functioning can be rather informative on how the detection of structural variation that is potentially associated to transposition is implemented in this kind of software.

Jitterbug was developed by Henaff et al. in 2014 within a collaboration between the laboratory that hosted this PhD project and the Centre for Genomic Regulation (Centre de Regulaciò Genomica - CRG). The scope of this software is to detect insertions from a WGS sample that are not annotated in the reference. Jitterbug needs an annotation of the TEs in gff format and allows the detection of insertions from more than one family.

MELT (Gardner et al. 2017) was developed at the University of Maryland within the 1000 Human Genomes Project. It is designed to work on large populations and thought to be run on human WGS samples, even though it is possible to set it up to run it on any species. The program requires several input files that are compressed into an in-house compressed file which filename ends with a ".mei" extension. The two main inputs to be provided are the consensus sequence of a specific TE family and the annotation file of the instances of this family in bed format. If the consensus sequence is used to detect non-reference insertions, the bed file is used to detect polymorphisms associated to the TEs annotated in the reference. For human TEs, specific .mei files ready for use are provided along with the program. To run MELT on other species, the user needs to create the .mei files by a script that is released by the same MELT developers.

The detection mechanism in Jitterbug and MELT is summarized in Figure 1.2. Both programs make use of pair-end short reads. This kind of reads is the result of the sequencing of the two ends of a DNA fragment. During alignment, they generally lay at a distance of a few tens of nucleotides (depending on the length of the fragment), or else aligned in two different positions in the genome. These cases can be relevant to the detection of TE polymorphisms because the distance between mate-reads can be changed after an insertion. The discordant mate reads can be due to the presence of an insertion in the sample that is not present in the reference, and to prove it, we need to verify whether the discordant read aligns to a Transposable Element.

In Jitterbug (Figure 1.2-A), the discordant reads having one mate aligned to an annotated transposon are defined as supporting reads and are used as a marker of a putative insertion. When the coverage is high enough, there can be reads spanning the insertion site. Those reads are partially aligned to the refrence genome. These reads, that are soft-clipped by the detection program, are used as proof of insertion. If the presence of insertion is flagged by a discordant read which mate aligns to an annotated TE, the soft-

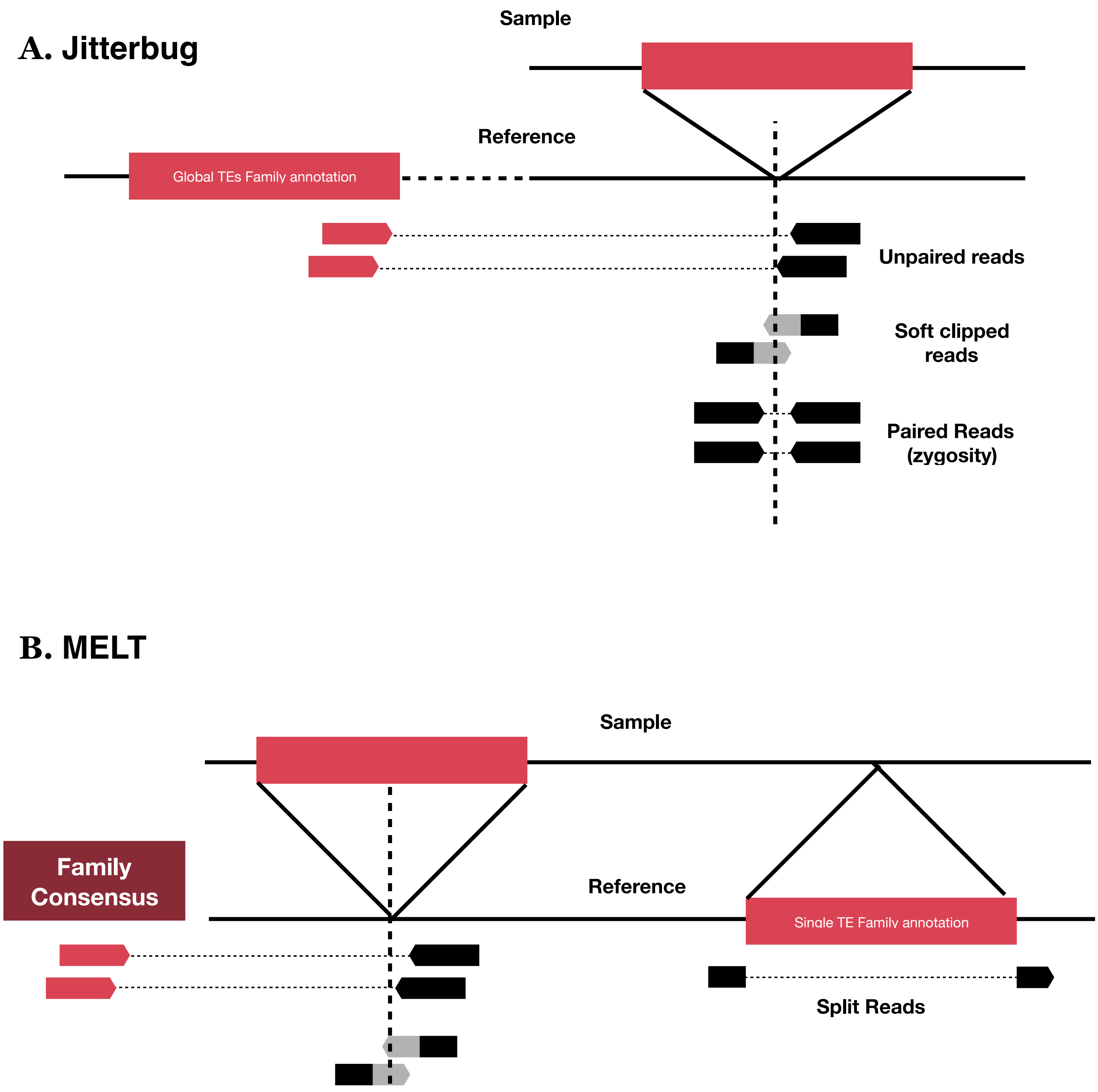


Figure 1.2 - MELT and Jitterbug algorithms

A - Jitterbug algorithm. Jitterbug detects non-reference insertions from pair-end WGS bam files by identifying those unpaired reads that align a TE (supporting reads). Soft-clipped reads that pile up with the supporting reads define the site of insertion.
B - MELT algorithm. MELT is designed to detect both non-reference and reference insertions. Whilst these latter are detected by the presence of split reads overlapping annotated TEs, the detection mechanism for non-reference insertions is similar to the one implemented in Jitterbug, even though MELT selects the split-reads that align to the consensus sequence of a single family as supporting reads.

clippedreads define the insertion site. Finally, the properly paired reads that align the region are used to infer the zygosity of the insertion (Henaff et al. 2015).

In MELT (Figure 1.2 - B), the detection of non-reference insertions is based on unpaired and soft-clipped reads as well. Like in Jitterbug, unpaired reads are used to detect insertion and soft-clipped reads to define the insertion point. The relevant difference here is that MELT aligns the unpaired reads to the family consensus sequence that is provided by the user. MELT can also detect the reference insertions, or else those TEs that are present in the reference but absent in the sample. This detection is based on the "split reads", or else those reads that result split in two relatively distant loci, that define the edges of the insertion.

In both packages, it is possible to proceed with further filtering based on the number of supporting reads, the definition of the edges of the insertion, and the overall quality of the prediction.

1.4.3 LTR-retrotransposon landscapes in Nipponbare and Minghui 63 rice genome assemblies

LTR retrotransposons are a subclass of Class I transposable elements or retrotransposons that have been found very active in plants (Casacuberta and Santiago 2003), and that are relatively easy to annotate due to their conserved structure. This makes them very suitable to test tools' capability to detect insertional variability from plant genome WGS datasets, and we decided to perform a de-novo annotation in the two rice assemblies we were comparing, Nipponbare (Nip) (Kawahara et al 2013) and Minghui 63 (MH63) (Zhang et al. 2016).

We submitted the two assemblies in fasta format to the package LTR Harvest (Ellinghaus, Kurtz and Willhoeft 2008) to retrieve a list of putative entire LTR retrotransposons. Elements from this first annotation have been first translated in protein sequence by EMBOSS package TranSeq (https://www.ebi.ac.uk/Tools/st/emboss_transeq/), in order to be scanned by HMMscan program from HMMer suite (http://hmmer.org) to identify putative coding regions by aligning translated nucleotide sequences to Hidden Markov Model alignments of TE proteins downloaded from the Gypsy database (http://gydb.org/), which, inspite of its name, contains information on proteins of different retrotransposon classes (gypsy and copia), as well as on viruses. Only the elements that have been found bearing a TE protein coding region were kept for further analysis and classified into Gypsy, Copia or Unclassified in the light of the 5'-3' order of the integrase (INT) and retro-transcriptase (RT) coding regions. Later, we annotated partial elements looking for sequences having similarity to the annotated retrotransposons by RepeatMasker. The results of this annotation on Nip and MH63 are reported in Table 1.2. The sum of all the entire elements and the partial fragments found by RepeatMasker is 131,905 elements in Nipponbare and 117,362 elemenents in MH63. Full-length LTR-elements are 3733 in Nipponbare and 3787 in MH63, divided into 1354 (Nip) and 1303 (MH63) Gypsy-like, 944 (Nip) and 759 (MH63) Copia-like and   1435

TE classification	**Nipponbare (Nip)** *Oryza sativa v. Japonica*	**Minghui 63 (MH63)** *Oryza sativa v.*
LTR-all (1)	131,905	117,362
LTR full-length (2)	3733	3787
LTR-gypsy	1354	1303
LTR-Copia	944	759
LTR-Unclassified (3)	1435	1725

Table 1.2 - Annotation of LTR-Retrotransposons and MITEs in rice assemblies

Result of the genome-wide LTR-RTs annotation in Nipponbare and Minghui 63 rice genome assemblies.

[1] Repeatmasker fragments. Includes both entire and truncated elements

[2] High confidence elements containing intact LTR. TSD and coding domains

[3] Intact elements whose poor coding domain conservation doesn't allow proper classification

(Nip) and 1725 (MH63) elements. The two genome assemblies thus have a similar LTR-RTs load, consistently with the fact that they have been built upon DNA samples from different varieties of the same species.

1.4.4 Annotation of standard datasets for tool benchmarking

A strong limitation for an unbiased benchmarking of tools that detect structural variability associated to transposition is the lack of validated panels of insertions, such as high-quality TE polymorphism datasets which insertions are, for instance, experimentally validated. For this reason, technical evaluations of software performance are frequently based on simulated insertions that can hardly replicate the complexity of a natural TEs landscape. A realistic picture of insertional polymorphism that can be obtained in silico might result from the pairwise alignment of the orthologous regions from the two genome assemblies. Pairwise alignments allow to detect large indels that are potentially due to transposition and, ideally, the overlap of the indels from a whole genome alignment and the TEs annotation should suffice in providing a list of TE polymorphisms between two genomes. At first, we aligned orthologous regions and then we complemented this alingment with the manual analysis of the prediction of the different tools.

As shown in Panel A from Figure 1.3, all the tools have been run using the Nipponbare (Nip) rice genome as a reference and the reads used to assemble Minghui 63 (MH63) as query, at increasing coverage levels of 5X, 10X, 20X and 40X. The first step was to divide the two genomes in 500 bp windows (Panel B) and align the orthologous windows. Later, we applied the window-based approach that is described in Figure 2-E and 2-D to all prediction to detect orthologous loci carrying reference and non-reference insertions. This resulted in a high-quality dataset of True Positives (TP) and False Positives (FP).

As for the LTR retrotransposons, a further curation of the standard dataset was needed due to the relatively high frequency of nested insertions and the high amount of truncated and degenerated elements. So, the LTR standard was reduced to the insertions found on Chromosome 5 which were manually curated. The resulting LTR standard consisted of 478 reference and 104 non-reference insertions.

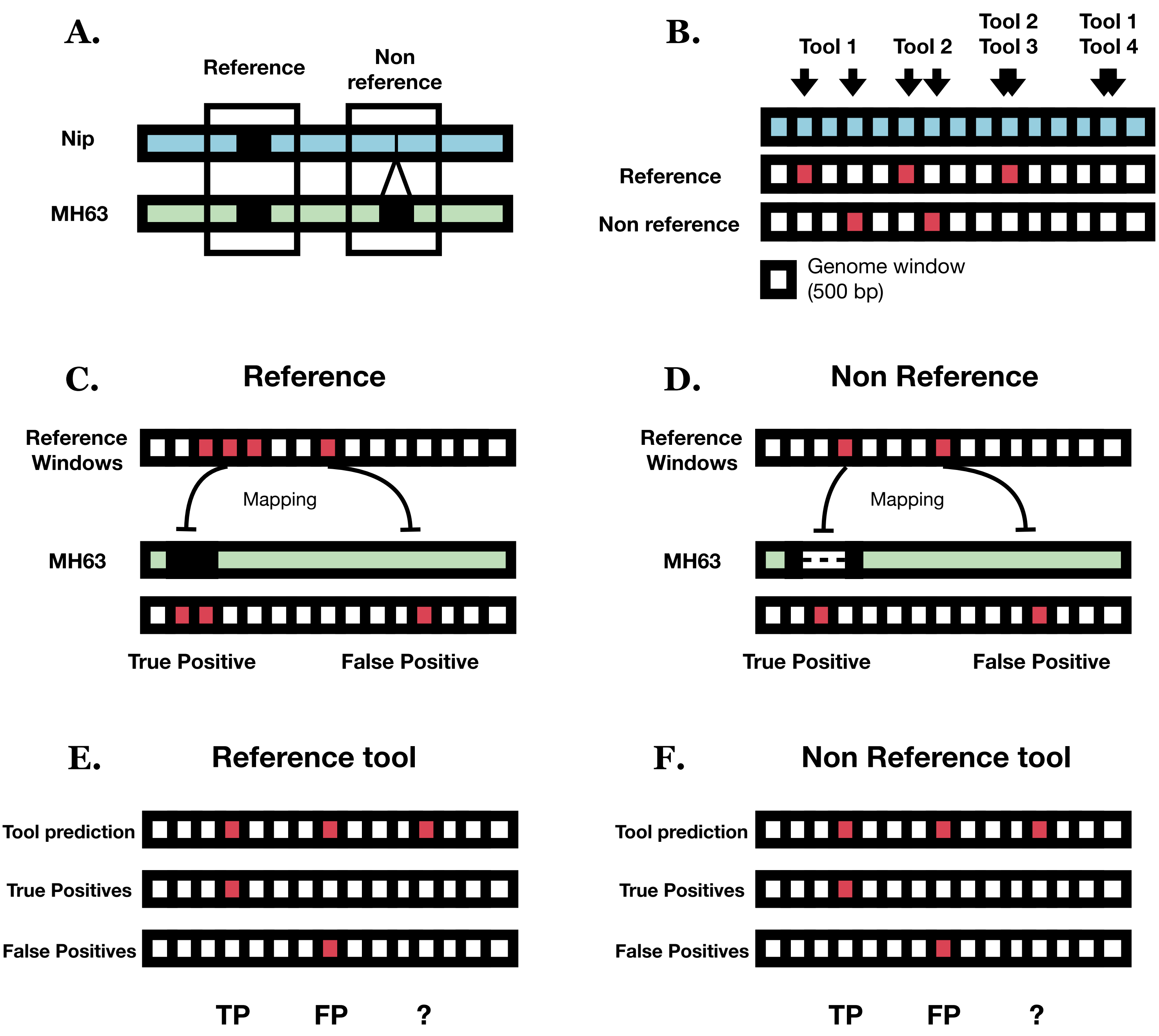


Figure 1.3 - Benchmarking standard annotation

A. Examples of “reference” and “non-reference” insertions.
B. The insertions predicted by the different tools (in IRGSP coordinates) are intersected with windows of 500bp spanning the entire IRGSP genome, to obtain reference and non-reference windows.
C. D. Reference and non-reference windows are mapped to the MH63 genome. Are considered True Positives (TP) all those cases in which a reference window aligns the MH63 genome with no indels and a non-reference windows aligns the MH63 genome with indels in a range of 500-25,000 bp.
E. F. The result of each single tool is crossed with the annotation of true and false positives for reference and non-reference insertions.

1.4.5 Comparing MELT and Jitterbug performance

Nipponbare (Nip) and Minghui 63 (MH63) O. sativa genomes are available in the form of assembled reference and unassembled WGS distributions (Kawahara et al 2013; Zhang et al. 2016). As described in Figure 1.3, we aligned the MH63 reads to the Nip reference genome to run the packages we were testing. Since we were interested in evaluating the performance of these programs on samples of different coverage, we made use of the packages included in the Samtools (http://samtools.sourceforge.net/) suite to generate distributions at different coverages (5X, 10X, 20X and 40X), and we repeated each program's run on each resulting distribution.

As already discussed, Jitterbug is a program that has been designed to detect non-reference insertions of multiple families. MELT, instead, detects reference and non-reference insertions from a single TE family. If on one side a comparison between the performance of these two packages, as the one that is proposed in Figure 1.4, must take into account the relevant differences in terms of input and scopes between the two programs, on the other side this comparison can be rather explanatory on the difference between the two annotation strategies (single vs. multi-family), and the overall reliability of their results.

The test on Jitterbug has been based on the full-length LTR-RTs annotation on Nip and ran on MH63 WGS samples of different coverages (5X, 10X, 20X and 40X). To run MELT, we needed to identify a specific LTR-RTs family and build its consensus sequence. We performed a blast between all the 3733 full-length LTR-RTs identified in Nip (Table 1.2) by NCBI-BlastN (Ye, McGinnis and Madden 2013) and used the package SiLiX (http://lbbe.univ-lyon1.fr/-SiLiX-?lang=en) to compute sets of sequences that mutually shared 80% of sequence similarity over 80% of the sequence length. We selected the most numerous family, a set of 203 Gypsy-like elements (family Gypsy-1) and computed the consensus sequence via EMBOSS- Cons package (http://www.bioinformatics.nl/cgi-bin/emboss/cons). The insertions detected by both programs have been compared with the standard to extract the number of false positives (FP), false negatives (FN) and true positives (TP).

In Figure 1.4-A we compare the results of Jitterbug and MELT. The first important result to notice is the "gold standard", reported in the yellow square at the top right of each table. This number represents the number of LTR-RTs non-reference insertions that have been manually annotated on MH63 chromosome 5. Jitterbug gold standard refers to non-reference insertions from all the TEs families (239 insertions), whilst MELT's standard is limited to the non-reference insertions that are assigned to Gypsy-1 family (33). Both table report values for True Positives (TP), False positives (FP) and False Negatives (FN) along with the evaluation of Sensitivity (S), Precision (P1) and the F1 score (F1), In Figure 1.4-B, Sensitivity, Precision and F1 are compared through a segment-scatter plot.

Sensitivity is defined as the ratio between the TP and the sum of TP and FN, hence representing the proportion of detected insertions over the actual ones. This value is plotted as a percentage in Figure 1.4-A. In both programs, sensitivity grows along with the sample coverage but it remains lower than the 50%.

A.

JITTERBUG			Gold standard* 239	
	5x	10x	20x	40x
Total Detected	11	29	50	69
True Positives (TP)	8	25	38	44
False Positives (FP)	4	5	16	32
False Negatives (FN)	231	214	201	195
Sensitivity TP/(TP+FN)	0.03	0.10	0.16	0.18
Precision TP/(TP+FP)	0.89	0.96	0.90	0.86
F1 Score 2*[(P*S)/(P+S)]	0.6	0.19	0.27	0.3
Runtime (CPU Time)		12.8		

MELT			Gold standard** 33	
	5x	10x	20x	40x
Total Detected	5	9	18	29
True Positives (TP)	3	5	11	17
False Positives (FP)	2	4	7	13
False Negatives (FN)	30	28	22	16
Sensitivity TP/(TP+FN)	0.09	0.15	0.33	0.52
Precision TP/(TP+FP)	0.6	0.55	0.61	0.57
F1 Score 2*[(P*S)/(P+S)]	0.20	0.26	0.50	0.67
Runtime (CPU Time)		13.5		

* Non-reference LTR-RTs insertions annotated on MH63 chromosome 5

** Non-reference LTR-RTs insertions annotated on MH63 chromosome 5 that belong to Family Gypsy-1

B.

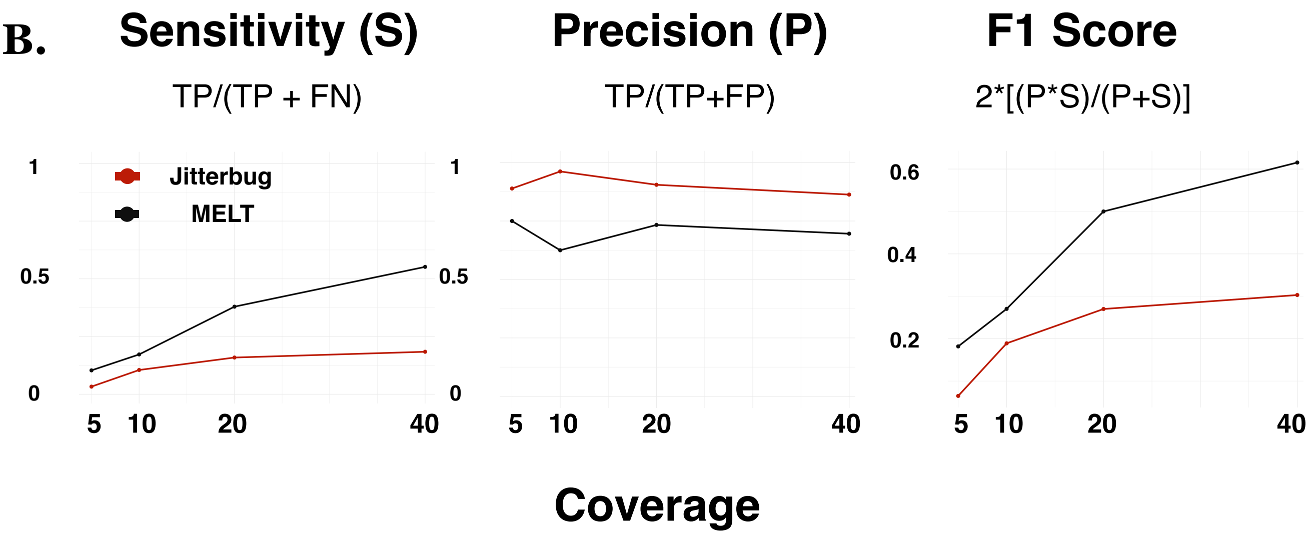


Figure 1.4 - MELT and Jitterbug performance for non-reference insertions on full-length LTR-RTs

Comparison between MELT and Jitterbug performance on full-length LTR-RTs non-reference insertions. The number of True Positive (TP), False Positives (FP) and False Negatives (FN) and the values of Sensitivity (S), Precision (P) and F1 Score have been evaluated on distribution of increasing coverage (5X, 10X, 20X, 40X). In (A) the values are reported in a table, in (B) we compare S, P and F1 through a segment-scatter plot.

Jitterbug scores a minimum sensitivity at 5x coverage identifying barely 3% of the real insertions. This number raises up to 18% in the highly covered samples (40x). In MELT, the minimum sensitivity is registered on 5x coverage (10,34%), and the level increases up to the 55,17% in the 40x sample. These values are considerably lower than the ones that were precedently available from tests on simulated reads published along with the package. As for Jitterbug, Henaff et al. published the result on a benchmarking made on simulated data. The sensitivity at 10x, 20x and 40x coverage was proposed to be constant and ranging around 85%, a result that is 5 times higher than the one we report in this work based on real data. Likewise, MELT was reported to reach an average sensitivity of 95.31% out of 50 runs on Alu families on simulated data (Gardner et al. 2017).

Conversely, precision is defined as the ratio of TP over the sum between TP and FP, and it represents the proportion of the true positives over the total predicted insertions. On this front, the result we report in Figure 1.4 -A differs less radically from the tests made on simulated reads. Jitterbug precision doesn't variate much with the increase of the coverage, ranging from a minimum of 0.86 at 40x and a maximum of 0.96 at 10x, a value that is comparable to the one published in 2015 as "positive predictive value" which was 92.2% of true positives on the overall result (Henaff et al. 2015). Instead, MELT's precision is lower. The maximum value is reached on the 5x sample, where the proportion of true positives on the global result is 0.6 whilst the lowest precision is scored on the 10x sample (0.55). In terms of absolute numbers, this is due to the fact that MELT identifies only 3 TP and 2 TP in 5x and only 5 TP and 4 FP at 10x.

Finally, Figure 1.4 – B shows the comparison of the F1 score or F-measure (Powers and Ailab 2011) which is the harmonic mean between precision and sensitivity and it is used to provide a global assessment of program functioning.

1.4.6 Overview of benchmarking results

The results presented in this chapter are relative to the candidate's contribution to a project whose purpose is to evaluate the effectiveness and reliability of the most common software packages for transposon detection in WGS files. This chapter reports the results obtained by the candidate presenting this manuscript, and that consist in the software review and classification, in the LTR-RTs annotation in Nip and MH63 and in the test run of the packages Jitterbug and MELT. This is a contribution to a project that unfolded within a lab-wide collaboration. In order to render the potential impact of this work, a swift overview of the outcome of the project is hence proposed in this paragraph.

The 12 packages classified in Figure 1.1 and summarized in Table 1.1 have been tested for the detection of reference and non-reference insertions of LTR-RTs and the Class II non-autonomous elements MITE (Miniature Inverted-repeat Transposable Elements). This choice is motivated by the fact that LTR-RTs and MITEs are the most active TEs in plant genomes and are responsible for several phenotypes (Casacuberta and Santiago 2003). If this explains the biological interest in focusing on these two TEs superfamilies, their different accumulation in the genome is relevant for technical reasons. MITEs have the tendency to insert next to the genes, while LTR-RTs are usually in heterochromatic regions and tend to produce nested insertions (Le et al. 2000). Testing the software in these two superfamilies involves different technical conditions which are interesting to compare.

The results are summarized in Figure 1.5, where the packages are classified into reference, non-reference and single-family dedicated software. In Panel A, we compare the packages detecting reference insertions, i.e. ngsTEmapper, Popoolation TE, Popoolation TE2, Teflon and TElocate. Among them, Teflon had the best sensitivity and overall performance (F1) in detecting MITE insertions, reaching a remarkable 74% sensitivity even at 10x coverage. At the highest coverage, most of the tools almost reached saturation. Regarding the detection of LTR-retrotransposons, the overall result is lower than in MITE detection. TElocate reached the maximum sensitivity, that was slightly higher than Teflon, and around 50%. Precision is high in all the software and with both MITEs and LTRs.

In Panel B, we compare the tools that detect non-reference insertions; Jitterbug, Popoolation TE, Popoolation TE2, Relocate 2, Teflon, TEMP, ngsTEmapper, retroseq and TElocate. This task looks generally more challenging since average performance on both MITE and LTR decreases in non-reference insertions detection. Teflon is the best performer on MITEs, reaching a 75% sensitivity at 40x, whilst the least sensitive is ngsTEmapper that stands below 10% with all coverages. With LTR-RTs, the sensitivities at 40x range from a minimum of 3% scored by ngsTEmapper up to 87.5% in Popoolation TE2. Note that Relocate2 was killed after 5 days running with 8 CPUs and 64GB of RAM while detecting LTR 40x. If sensitivity is comparable between MITE and LTR non-reference detection, what really variates is precision. It approximates saturation with all coverages in MITE detection for all packages, with the exception of TElocate which precision is affected by the increase of the coverage, but it variates a lot in LTR detection. In this, Jitterbug seems the only one keeping its precision above 75% with MITEs and LTR-RTs.

Panel C shows the results for the 3 single family detection tools, MELT, Trackposon and ITIS. Their performance is evaluated on non-reference MITE detection, and it’s considerably lower than the one marked by multi-family packages. In terms of sensitivity, the best performer is Trackposon which scores 61% at 40x, whilst both MELT and ITIS remain below 20%. MELT and Trackposon achieve a good precision, even though this drops in MELT when ran high covered sample.

Finally, Panel D compares the computational time needed for one run. Comparison is made by running the programs on MITEs in the sample at 10x coverage.

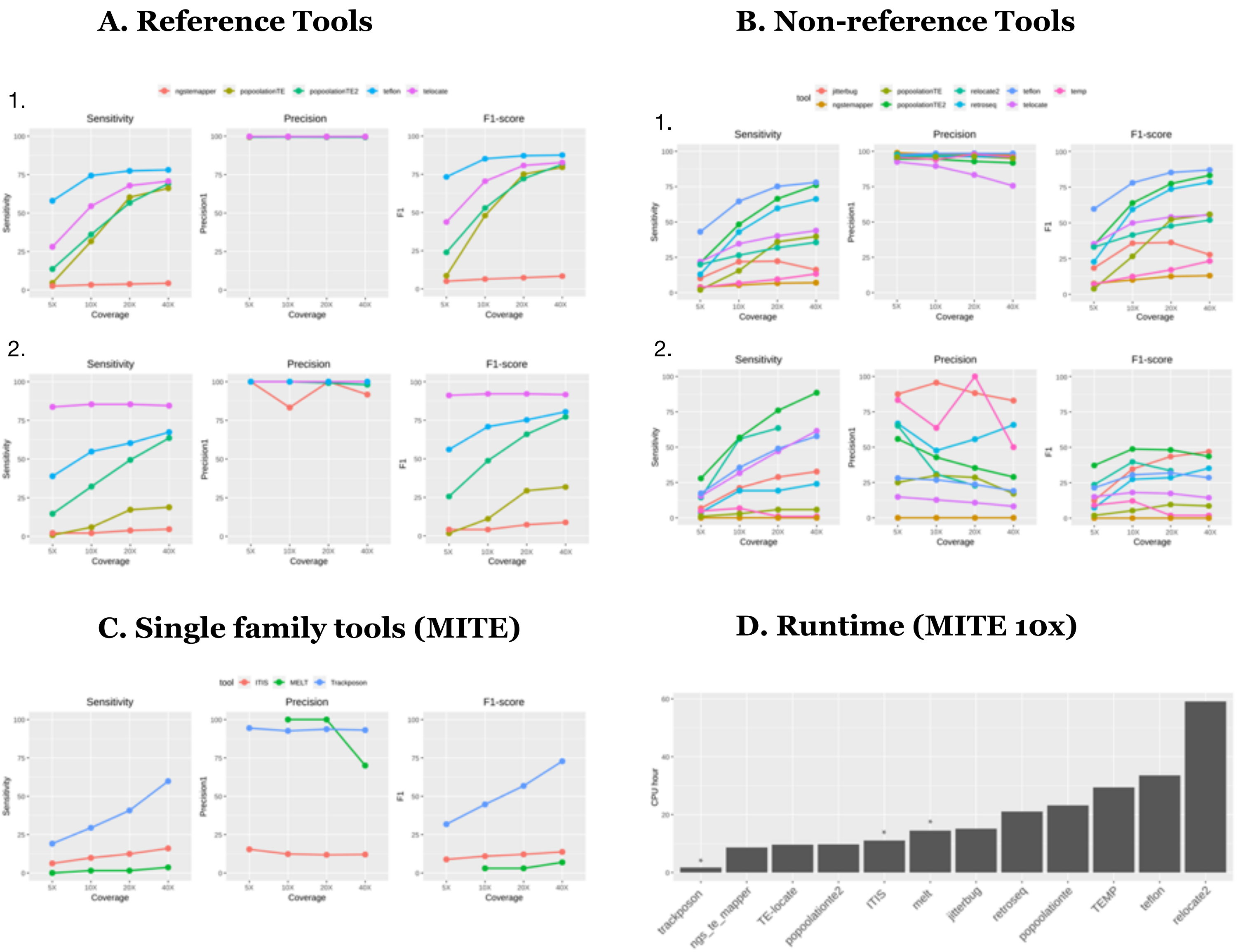


Figure 1.5 - Overview on benchmarking project results.

A. ngsTEmapper, Popoolation TE, Popoolation TE2, teflon and TElocate are compared as tools to detect **reference insertions**. Their performance on MITE is shown in 1., whilst their performance on full-length LTR retrotransposons is compared in 2..

B. Jitterbug, Popoolation TE, Popoolation TE2, relocate 2, teflon, TEMP, ngsTEmapper, retroseq and TElocate are compared as tools to detect **non-reference insertions**. Their performance on MITE is shown in 1., whilst their performance on full-length LTR retrotransposons is compared in 2.. Relocate2 LTR-40X was killed after 5 days running with 8 CPUs and 64GB of RAM.

C. MELT, ITIS and Trackposon are compared as single family tools on MITE.

D. Running time for each tool to perform the detection of MITEs in a 10X dataset. Family-specific tools are marked with an asterisk.

1.5 Discussion

The results from this benchmarking project shed a new light over several aspects of TEs detection from WGS samples. Scientists strive to build, improve and adapt bioinformatics tools to increase the reliability of their analyses. The choice of the right methods is important in all ambits of science, but it becomes crucial for those technologies that are relatively new and rapidly evolving, as in the case of genomic investigation of mobile DNA based on Next-Generation Sequencing. Genome scientists need to know to what extent their prediction can be realistic, and any effort in critically reviewing the effectiveness of available software to improve genomic pipeline is usually widely invoked (Henaff et al. 2015; Hoen et al. 2015). The relevance of this work is hence to be found on the possibility to share practical insights to improve a WGS-based detection workflow for TEs. This discussion will be structured around the 6 main considerations that can be drawn from this benchmark and that can have relevance to workflow design and implementation.

1. Our capability to detect transposable elements from DNA-seq is limited.

Differently from the previous test drives that were built upon simulated data (Richiwar, Marino-Ramirez and Jordan, 2016), this project relies on a standard dataset that is elaborated from real insertions detected from pairwise alignments. Simulated insertions can barely approximate the real complexity of an actual transpositional landscape, and even several developers (e.g. Jitterbug's Henaff et al. 2015) invoked a benchmark based on real data. As expected, the performances on real data are way lower than the ones proposed after benchmarks on simulated distribution, and this sheds light on the first consideration; our capability to detect insertions from WGS samples is limited, and lower than what we used to believe.

2. Sensitivity and Precision depend on coverage, but precision is affected by the algorithm implemented in the package as well.

We have chosen the two variables that are usually deployed to evaluate a bioinformatics package performance. Sensitivity, which is the proportion of actual insertions found, and precision, which is determined by the number of false positives included in the result. If sensitivity clearly increases proportionally to the coverage (Figures 1.4 and 1.5), indicating that more coverage implies more detection power, the precision results is quite constant with a tendency to the decrease. If precision is substantially high in those packages detecting reference insertions (Figure 1.5 A), the level variates a lot in non-reference insertions programs, but doesn't look variating proportionally to the coverage. This difference between sensitivity and precision variation with the coverage is consistent to the idea that the detection of insertions depends on structural hallmarks, such as the presence of unpaired reads (Figure 1.2), which presence increases as the coverage increases. Conversely, the attitude to include false positives, that defines precision, depends on the implementation of result filtering that is characteristic of each program.

Even though several packages tend to include more false positives at higher coverages, results indicate that precision keeps constant through the variation of coverage.

3. Detecting reference insertions is easier than inferring non-reference insertions.

In Figure 1.1 we operated a distinction between those packages that are designed to detect the presence or absence of TEs annotated in the reference (reference insertions) and those ones that are aimed at detecting TEs in samples that are not annotated in the reference (non-reference insertions), highlighting that most of the programs are actually designed to complete both tasks. The results of the project that are summarized in Figure 1.5 clearly indicate that the overall performance of inspected software on reference insertions (Figure 1.5 - A) is higher than the one shown on non-reference insertions (Figure 1.5 - B). Even with some differences between LTR and MITE in fact, we can appreciate how most of the reference tools converge to high F1 scores at 40x coverage, whilst the overall performance on non-reference insertions is lower. This is due to the different algrithms that are implemented for reference and non-reference insertions.

4. Focusing on one single family doesn’t necessarily imply a better result or a faster run.

The comparison between MELT and Jitterbug that is plotted in Figure 1.4 shows that MELT, a package that is designed to detect polymorphisms that are traceable to one single family, displays better overall performance than Jitterbug. This result is boosted by a relatively good outcome in terms of sensitivity (Figure 1.4 - A). On the other hand, though, the multi-family oriented Jitterbug demonstrates a better precision and a runtime for the complete LTR-RT dataset that is comparable to the one in MELT for a single LTR-RT family. If we extend the comparison between single-family and multi-family packages reported in Figure 1.5 A, B and C, we can’t see a striking difference in terms of performance. The run-times of the 3 single-family packages with their multi-family counterparts (Figure 1.5 - D) lay mostly on the left side of the graph but, still, this difference is nor neat nor high enough to conclude that single-family tools are faster than multi-family tools. In TEs research is a common practice to reduce the investigation to one single TE family in order to reduce the possible noise and the computation time, or to focus on those elements that are more active and impactful in one species (Gardner et al. 2015). In general, single-family tools are faster (Figure 1.5 - D) and represent a reasonable choice for those analysis that are focused on a single family. Anyways, a study that aims to identify polymorphisms that are relative to more families would benefit from the application of a generalistic tool.

5. Different TEs superfamilies are annotated with different performances.

A swift comparison between the performance on MITEs and LTR-RTs (Figure 1.5) suffices to understand that most of the inspected packages deal better with MITE detection. This difference is consequent to the different accumulation of LTRs and MITEs and to the different complexity of their landscape. As discussed in the introduction, LTRs are relatively large elements (6-10 Kb) that tend to concentrate in heterochromatic regions and to give rise to nested insertions (Cicconi et al. 2017). Conversely, MITEs are small elements (~500 bp) that mostly lay in gene-rich regions that are usually better assembled than the heterochromatic regions (Han and Wessler 2010). Hence, it's reasonable to expect that structural variation detection is easier in MITEs.

6. Third Generation Sequencing might drastically improve TEs detection.

By the end of the 2000s, the term Third Generation Sequencing (TGS) was introduced to describe those high-throughput sequencing devices that are able to return longer sequences than the usual NGS reads. Whilst a typical NGS "short" read is about 150 bp long, a TGS "long" read can reach the length of 15000 bp (van Dijk et al. 2017), increasing our DNA-seq resolution in coding and non-coding DNA. In a few years, this technology is expected to replace the Illumina® short-read datasets with long-reads samples that are way more effective in detecting the structural variation even in those repetitive regions that are more difficult to reconstruct (Treanberg and Salzberg 2012). In the last years, most of the efforts have been directed to the implementation of software tools that could detect transposon polymorphisms from the huge amount of NGS resequencing data, but we need to expect this to change

Chapter 2 - Peach and Almond transpositional and evolutionary dynamics

2.1 Introduction

In the introduction to this thesis we discussed how Transposable Elements (TEs) cover a large part of angiosperms genomes and have a relevant impact on the generation of the genetic variability in eukaryotes. The assessment of this impact can be achieved through a workflow that is made up of three fundamental steps: the annotation, the transpositional dynamics evaluation and the linkage with gene activity.

The annotation of TEs is increasingly recognised as a fundamental part in the characterisation of newly assembled genomes. Besides the biological interest in the TEs per se, these elements can difficult the correct annotation of the genes, that is probably the most crucial result in a genome characterisation. A better annotation of TEs will ease their distinction from the gene and allow a more correct gene annotation (Platt, Blanco-Berdugo, and Ray 2016). Here's why several efforts are being concentrated on the implementation and improvement of software solutions to perform de novo annotation and classification of TEs (Goerner-Potvin and Bourque 2018). These tools may be sorted on the base of their scope into two main classes. Tools like RepeatMasker (Smit, AFA, Hubley, R & Green, P. RepeatMasker Open-3.0. 1996-2010 http://www.repeatmasker.org/) and REPET (Flutre et al. 2011) will begin their workflow with the characterisation of repetitive and low complexity DNA. The identification of TEs is then obtained by comparing repetitive DNA with TEs consensus databases. In a more detailed fashion, another class of TEs annotation programs relies on the structural features that are conserved in some TEs superfamilies. MITE hunter (Han and Wessler 2010) and LTR harvest (Gremme, Steinbiss, and Kurtz 2013) scan genomic sequences to identify those structures that are relative to the Miniature Inverted-repeat Trasnposable Elements (MITEs) and the Long Terminal Repeat Retro-Transposons (LTR-RTs) respectively. Whilst the first class of programs provides a realistic evaluation of the global TEs genomic coverage, the structure-based pipelines perform better in the detailed definition of single insertions, providing annotations that are limited to one specific superfamily but yet more suitable to dedicated analyses.

Once the TEs annotation is obtained, the second step is to assess Transpositional Dynamics (TD), that is defined as TEs genetic variation through space and time (Estep, DeBarry, and Bennetzen 2013) and is determined by the balance between the transpositional activity and the counteracting mechanisms that the genome deploys to reduce the proliferation of mobile elements (Bardil, Tayalé, and Parisod 2015). Starting from a TEs annotation on the reference genome, we can obtain information on their TD by combining two different variables. The first one is the insertion time of single TE copies, that is inferred from the accumulation of mutations on the transposon sequence since its insertion (SanMiguel et al. 1998), and the second one is the insertional variability, or else the variation of TEs insertions within the same organism, between organisms of the same species or between species. When combined, insertion time and insertional variability can be informative on how TEs variated along the time of evolution and the space of an organism, a population or a set of related species, hence giving a fair overview on their TD.

As we have built an annotation of TEs that is provided with the information on their dynamics, we will be able to tell with some good approximation when a given TE copy has inserted and whether is fixed or polymorphic at somatic, intraspecific or interspecific level. This will allow us to thread into the third step, that is the linkage between transposition and genetic variability. TEs can interfere with the gene functioning in several ways. If an insertion that lands into a gene coding region will most likely destroy it with a knock-out effect on the gene, an insertion that hits the surroundings of a gene can modify its regulation (Casacuberta and Santiago 2003). TEs have been demonstrated providing new regulatory elements (Morata et al. 2018) and changing the local epigenetic status of the region they insert (Song and Cao 2017). An evaluation of the impact of transposition on genetic variability can start from the selection of those insertions that overlap or flank annotated genes. The TD of this transposon- induced mutations can tell us about the timing of their appearance and their variability. In this way, we can link transposition to the establishment of genetic variability within the same species and between closely related species.

The choice of the proper model to study the relationship between transpositional dynamics and genetic variability will need to take some aspects into account. The point is to have the possibility to build a fair description of the TD that can be straightforwardly linked with genetic variability, and this basically behoves us to build an annotation of TEs that can be easily characterised, on a species which genetics is known and relatively easy to study. Among the most common crops, the species that belong to the Prunus genus can represent a good model for different reasons. The Prunus genus encompasses stone-fruit trees and shrubs of the Rosaceae family, namely Plums, Cherries, Apricots, Peaches and Almonds (Shi et al. 2013). These species share a diploid genome constituted by 8 chromosomes which size ranges roughly around 250 MB (Verde et al. 2013). Among them, Peach (Prunus persica) and Almond (Prunus dulcis) represent a case of particular interest for their economical relevance and for being sister species, close enough that they can be crossed together to generate fertile hybrids. However, whilst the most Almond varieties is out-crossing, Peach is auto-fertile. The mechanism preventing self-pollination in the hermaphrodite flower of Prunus monoicous species is conserved genus-wide, and the auto-fertile condition of Peach is deemed to have arisen secondarily (Akagi et al., 2016.). The different reproductive strategy of the two species is probably one of the main causes of the differences in terms of intraspecific variability that are observed between Peach and Almond, as nucleotide diversity is roughly seven-fold higher in this latter (Velasco et al. 2016).

The fruits we use to know as Nectarines are actually Peaches that have loss the capability to grow trichomes over the skin of the fruit (Tijskens et al. 2007). This somatic mutation, that has been positively bred for its positive outcome in terms of consumer acceptance, is actually caused by an insertion of a Class I Long terminal Repeat Retro- Transposon (LTR-RT) within the 3rd exon of the MYB25 gene (Vendramin et al. 2014). We recall this case as one of the many proofs of the usually high activity of LTR-RTs in plants (DeFraia and Slotkin 2014). As already introduced, these retro-elements are provided with a conserved genomic structure that allows the structural identification (Gremme, Steinbiss, and Kurtz 2013) and the insertion time determination (SanMiguel et al. 1998). LTR-RTs are hence quite suitable for studies aimed at the determination of TD.

If on one side we can easily insight the TD of LTR-RTs for their structure, they are also very active in plants and proven to trigger well described mutations in Peach. Peach and Almond are two major crops, diploids, with a genome that is relatively small in size, close enough to be crossed together but with some relevant differences at genetic level. These considerations enable the study of the impact of LTR-RTs on Peach and Almond genetics as a good framework to study the relationship between TD and genetic variability.2.2 Scope of the chapter

Within a collective effort to annotate and characterise the Almond (Prunus dulcis) genome, we have annotated the TEs of all classes in this species and updated the already available transposon annotation in Peach (Verde et al. 2013). Then, a workflow to obtain a more selected and refined annotation of LTR-RTs has been designed and implemented. The resulting LTR-RTs annotation from Peach and Almond underwent three distinct analyses to characterise the TD. First, we have compared the two distributions to assess their interspecific variability. Second, the insertion time has been determined for each LTR-RT copy to assess the evolutionary dynamics in the two species. Finally, the intraspecific insertional variability has been described by comparing the reference genomes to 35 DNA-seq samples from 35 Peach and Almond varieties.

2.3 Materials and Methods

2.3.1 Peach and Almond reference genomes

The de novo annotation of TEs was performed on the assembled genomes of Peach (Prunus persica) and Almond (Prunus dulcis). Peach genome (Prunus persica v. 2.0, released on January 2015) is sequenced from the double haploid genotype of the peach cv. Lovell (Verde et al. 2013). Almond genome (Prunus dulcis Texas Genome v2.0, released on October 2018) is sequenced from the cv. Texas.

2.3.2 DNA-seq data from Peach and Almond varieties

Whole Genome Sequencing (WGS) pair-end read datasets from 16 Peach and 19 Almond varieties were generated within a CRAG- wide collaboration. Technical details, provenience and data availability are reported in Table 2.1.

2.3.3 De novo identification of Transposable Elements

To identify and characterise the total amount of mobile DNA in Peach and Almond genomes, we used the REPET package v 2.5 (Flutre et al. 2011). The package consist of two distinct piplines. The first one to be ran, TEdenovo (https://urgi.versailles.inra.fr/Tools/REPET/TEdenovo-tuto), computes consensus sequences out of highly repetitive DNA and filters those ones that have sequence similarity with known transposons consensus available in the REPBASE (https://www.girinst.org/repbase/) database. Obtained putative TEs consensus sequences serves as input of the second pipeline, TEannot (https://urgi.versailles.inra.fr/Tools/REPET/TEannot-tuto), that is aimed at annotating, strcutrually defining and classifiyng the single insertions over the genome. The pipeline was ran within the PiraTE virtual machine (Berthelier et al. 2018).

2.3.4 Fine annotation of LTR-RT

To better focus on LTR-RTs, we concentrated our efforts on a refined annotation of LTR-RTs in Peach and Almond. Our workflow starts with the structural annotation of putative LTR-RT that is performed by the package LTR harvest 3.8 (Gremme, Steinbiss, and Kurtz 2013), an LTR-RTs structural identification tool that is part of Genome Tools 1.5.9 (http://genometools.org/index.html).

LTR harvest results are then filtered in reason of their quality. Each identified insertion undergoes to three distinct filtering steps:

Anonymous nucleotides content. The number of non deteremined or anonymous nucleotides (“N”) is determined. Sequences are kept for downstram analyses only if they have a proportion that is < 10% of anonymous nucleotides .

Tandem Repeats content. Tandem Repeats (TRs) within the putative LTR-RTs sequences are identified by the package TRF 4.09 (https://tandem.bu.edu/trf/trf.html) that is ran on its locally downloadable version (Benson 1999). Only the insertions which sequence is covered by TRs for less than its 50% are kept for further analyses.

Presence of transposon protein coding regions. The program hmmscan (https://www.ebi.ac.uk/Tools/hmmer/search/hmmscan) from the HMMer suite (Potter et al. 2018)is used to identify the coding regions on each putative LTR-RT. The whole trasposon sequence is translated into a the amminoacidic code on its 6 possible reading frames by the EMBOSS package’s transeq (Rice, Longden, and Bleasby 2000), and each reading frame is submitted to hmmscan. Hmmscan aligns the protein sequences to hidden markov models (Eddy 1998) computed on known transposon proteins profiles collection available at the Gypsy Database (http://gydb.org/). Only those putative LTR-RTs that are found bearing at least one transposon protein putative coding region are kept in the dataset.

The resulting sequences constitute the refined annotation of LTR-RTs we will refer to in the rest of the chapter.

2.3.5 Classification of LTR-RTs

The filtered putative LTR-RTs are then assigned, whenever possible, to the two major superfamilies (Ty3-Gypsy-like or “gypsy” and Ty1-copia-like or “copia”) in the light of the disposition of their domains. More specifically, we relied on the well conserved HMM profile of the integrase (PF02022) and of the reverse transcriptase (PF00078). We have classified as “copia” those elements having the integrase upstream to the retrotranscriptase, as “gypsy” those elements having the retrotrancriptase upstream to the integrase, and as “unclassified” those elements who had some TEs protein coding regions but was missing of either or both of retrotranscriptase and integrase.

After having aligned all vs all the identified sequences from both Peach and Almond, we used the package SiLiX (http://lbbe.univ-lyon1.fr/-SiLiX-?lang=en) to cluster the sequences that share the 80% of sequence identity over the 80% of sequence length (Miele, Penel, and Duret 2011).

2.3.6 Insertion time determination

The 5′- and 3′-LTR end of each LTR-RT have been pairwise aligned to estimate the insertion time of LTR-RTs (SanMiguel et al. 1998) with the program MUSCLE v. 3.8.5 (Edgar 2004). The age of the insertion was based on the nucleotide differences between the two LTR using the Kimura two-parameter method, with an average substitution rate (r) of 2 × 10– 9 substitutions per synonymous site per year. The insertion time (T) was obtained as T = k/2r (Kimura 1980).

2.3.7 Structural variability detection in Peach and Almond DNA-seq data.

Each DNA-seq paired-end sample from cultivated and wild peach and almond varieties that is reported in Table 2.1 was aligned to the Peach or Almond reference genome by BWA-align (Li and Durbin 2009) and compressed in BAM file using the SAMTOOLS package (Li et al. 2009). Bam files were later submitted to the packcage PINDEL (Ye et al. 2009) to identify deletions in samples and to the package Jitterbug to detect insertions in samples (Hénaff et al. 2015).

Table 2.1 - List of sequenced varieties available as pair-end reads WGS leaf samples. (CRAG- IRTA)

A. Cultivated Peach

Variety Name	Species	Origin
Armking	*Prunus persica*	USA
Belbinette	*Prunus persica*	France
Bigtop	*Prunus persica*	Spain
Blanvio	*Prunus persica*	Spain
Cakereine	*Prunus persica*	USA
Flatmoon	*Prunus persica*	USA
Ghiaccio	*Prunus persica*	Italy
Nectalady	*Prunus persica*	USA
Nectaross	*Prunus persica*	France
Nectatop	*Prunus persica*	France
Platurno	*Prunus persica*	Spain
PN251	*Prunus persica*	Spain
Sangui	*Prunus persica*	France
Sweetdream	*Prunus persica*	USA
Tifany	*Prunus persica*	USA
UFO	*Prunus persica*	France

B. Cultivated Almond varieties

A la dame	*Prunus dulcis*	France
Atocha	*Prunus dulcis*	Spain
Bartre	*Prunus dulcis*	France
Doree	*Prunus dulcis*	France
Falsa Barese	*Prunus dulcis*	Italy
Ferragnes	*Prunus dulcis*	France
Ferrastar	*Prunus dulcis*	France
Gabais	*Prunus dulcis*	France
Garfi	*Prunus dulcis*	Italy
Guara	*Prunus dulcis*	Spain
Johnstons	*Prunus dulcis*	USA
Marinada	*Prunus dulcis*	Spain
Pointued dlAurellie	*Prunus dulcis*	France
Primoski	*Prunus dulcis*	Russia
Princesse JR	*Prunus dulcis*	France
R2345	*Prunus dulcis*	Spain
Ripon	*Prunus dulcis*	USA
Strouds	*Prunus dulcis*	USA
Vialfas Cita	*Prunus dulcis*	Spain

2.4 Results

2.4.1 Class I and II TEs annotation in Peach and Almond

As already introduced, a good annotation of TEs represents a step that is crucial for the characterization of a newly assembled genome. This is why our first efforts have been concentrated on the annotation of TEs in Almond, as part of the Almond genome assembly and annotation project. Although a TE annotation in Peach was already available along with the genome data (Verde et al. 2013), we decided to update it applying the same workflow as Almond to allow a comparison of TE content in the two species.

Our workflow relied on the pipeline REPET (Flutre et al. 2011)that is designed to first identify repetitive DNA and then to classify as Class I or II transposons those repetitive elements that show high sequence similarity with TEs consensuses coming from the RepBase databank (Bao, Kojima, and Kohany 2015). The software is provided with two distinct pipelines. TEdenovo is the first to be ran, as it begins with the identification of repetitive DNA from the whole genome sequence, and the computation of consensus sequences. The consensuses that show sequence similarity with TEs consensus from the RepBase databank are selected and classified as TEs consensus. These consensuses represent the final output of TEdenovo and serve as input to the TEannot pipeline, that identifies complete and partial copies along the genome sequences, returning the final TEs annotation. We have run the two REPET pipelines within the PIRATE (Berthelier et al. 2018) virtual machine, starting from the publicly available Peach genome v 2.0 (Verde et al. 2013) and the Almond genome draft generated in the framework of the almond genome project, P.dulcis26 (Alioto et al, 2019).

In Table 2.2 we can appreciate that about the 40% of the genome of the two species is covered by TEs, a result that is very similar in the two species. In both species, the Class I elements (retroelements) cover about the 24% of the whole genome, whilst the Class II (DNA) elements cover about the 14%. Most of these latter are represented by Tandem Inverted Repeats (TIR), that cover the 13% in Almond and the 14.3% in Peach. Among the Class I retroelements, LTR-RTs are the most abundant family, covering the 22% in Almond and the 21.5 % in Peach. Other superfamilies from Class I and Class II are present in minor percentage, but their relative quantity is comparable between Peach and Almond.

In Figure 2.1, The gene and TE coverage is plotted for each one of the 8 chromosomes that constitute the Peach and Almond genome. The concentration of TEs along the pseudomolecules is comparable and displays no particular region of differential accumulation between the two species, confirming that they share a very similar composition and distribution of mobile DNA.

The TE annotation shows that the two species share a comparable amount of TEs, that cover roughly the 40% of the whole genome sequence. The classification of TEs into classes and types and their respective coverage indicates a very similar composition of TEs.

Almond *Prunus dulcis*		Peach Prunus persica
38.6	**Total TEs**	**37.6**
0.3	DIRS	0.4
1	LARD	1.6
2	LINE	2.2
22	LTR	21.5
0.3	SINE	0.4
24.1	**Total Class I**	**22.8**
1.4	Helitron	0.9
1	MITE	1.3
0.4	Maverick	0.2
13	TIR	14.2
14.5	**Total Classi II**	**14.4**

Table 2.2 TEs genome coverage in Peach and Almond (REPET)

Comparison between TE content in Peach and Almond genomes after annotation with the REPET pipeline. For each superfamily and for each class, the coverage is indicated as percentage of the whole genome length. Note that the sum of the coverage values of single superfamilies within a class slightly exceeds the indicated global coverage for the class due to some overlapping / nested insertions / ambiguous classifications.

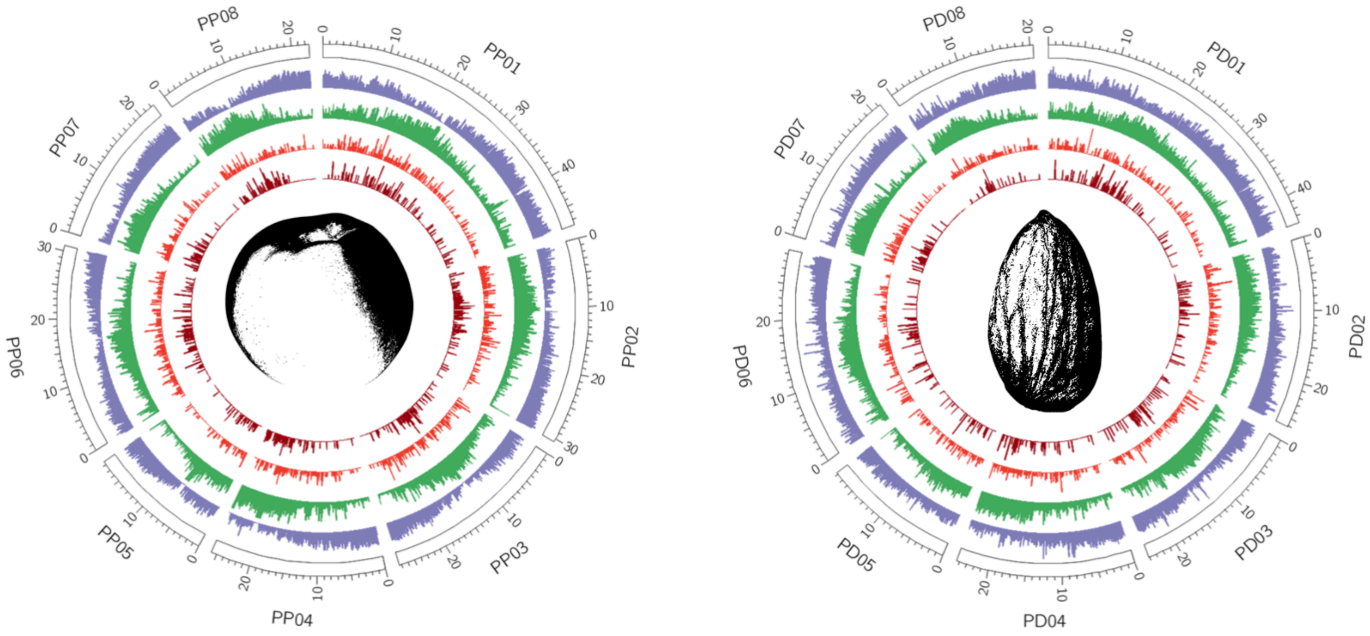


Figure 2.1 - TEs genome coverage in Peach and Almond. Coverage per chromosome (REPET)

The coverage per chromosome of *de novo* identified TEs is reported for the eight chromosomes in Peach (left) and Almond (right). **Legend: Genes (violet), Transposons Class I and II (green), Copia (red), Gypsy (dark red).**

2.4.2 LTR-RTs annotation in Peach and Almond

LTR-RTs cover a little more than the 20% of Peach and Almond genomes. Such a high coverage indicates that these retro-transposons have been active during the evolution of these two species. We have already discussed that LTR-RTs share some particular features that ease the characterization of their dynamics, from the possibility to rely on their conserved structure to annotate them de novo to the possibility to infer their insertion time starting from the sequence identity between the two long terminal repeats (SanMiguel et al. 1998).

The software LTR Harvest (Gremme, Steinbiss, and Kurtz 2013) is a part of the Genome Tools suite (http://genometools.org/index.html) and is designed to retrieve putative LTR-RTs from assembled genomes in fasta format. We used LTR-Harvest, to annotate putative LTR-RTs. These annotations underwent several filtering steps, and each element has been classified according to the spatial organization of the coding region. Details of annotation protocols are discussed in material and methods (¶ 2.3.4).

The result of LTR-RTs annotation is reported in Figure 2.2. Peach and Almond share a similar number of detected LTR-RTs, 2148 in Almond and 2215 in Peach. Copia elements constitute roughly 45% of total elements in both species (964 in Almond and 1040 in Peach), whilst the gypsy constitute about the 20% of total Peach and Almond LTR-RTs (392 in Almond and 517 in Peach). The elements that are labeled as "unclassified" have been found of transposon-related domains, but lack of one or both the integrase and retrotranscriptase, and we couldn't classify them. They constitute the 37% of Almond elements (792) and the 30% of Peach elements (658). As for the partial elements, we decided to sort them into solo-LTR and truncated elements based on the region of similarity they showed with putative complete elements. As we can appreciate in the Figure 2.2, their number compares in the two species.

To further inspect this aspect, we have clustered all complete LTR-RTs from peach and almond into clusters defined by 80% of sequence identity over 80% of the sequence length using the software SiLiX (Miele et al. 2011). Results are reported in Table 2.3. We obtained 353 clusters and 78% of them (277) are mixted clusters, being composed by at least one element identified in Peach and at least one element

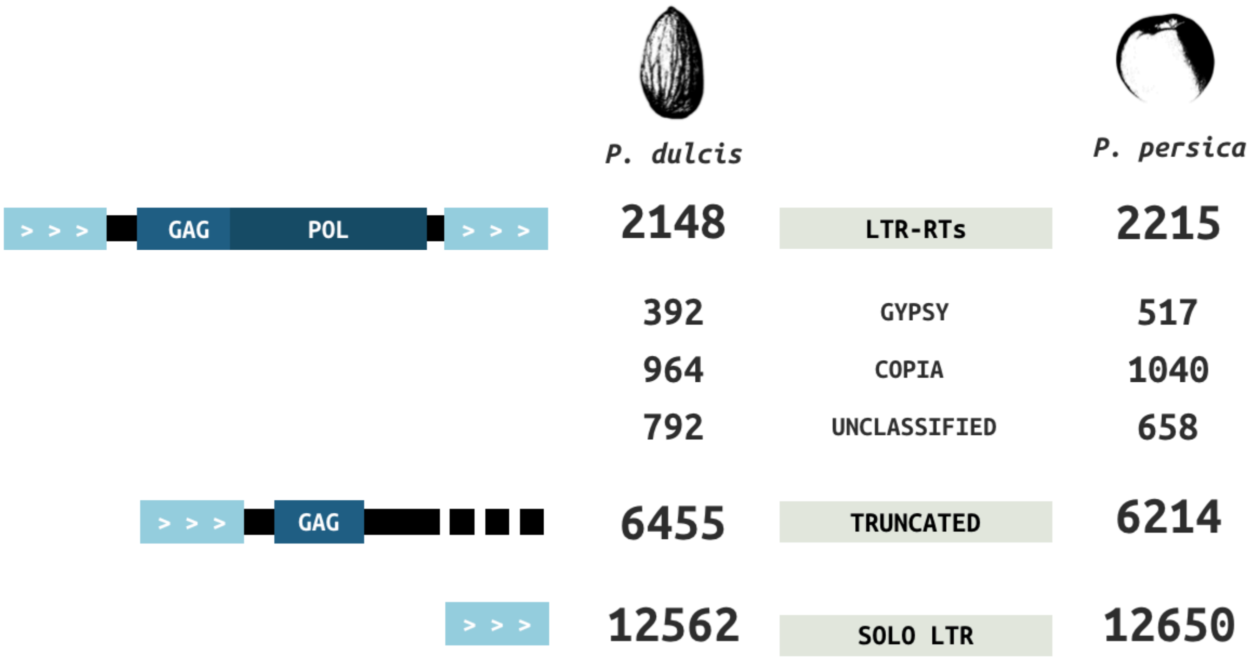


Figure 2.2 - LTR-RTs annotation in Almond and Peach. Single insertions.

Results of the refined annotation of LTR-RT in Peach and Almond. Collecting entire and partial identified elements. These latter are divided into “truncated” elements and “solo LTR”.

	Total	Singletons	Almond- only	Mixed	Peach - only
Families	353		46	277	30
Elements	4363	1500	109	2666	88

Table 2.3 - Conservation of LTR-RTs between Peach and Almond. Family statistics.

Families are sets of LTR-RTs that mutually share the 80% of sequence identity over the 80% of the sequence length and are composed of at least two elements annotated in Peach or Almond. The table compares the number of families that are composed of elements coming from one of the two species only or from both species (mixed families), and the elements that are collected in species-specific and mixed families.

identified in Almond. Out of the 4363 elements we have identified in the two species (2148 in Almond and 2215 in Peach), 2863 elments can be clustered (65%), and 2666 belong to a mixed cluster. This means that Peach and Almond share the 78% of the LTR-RT clusters which include 93% of the clustered elements.

All these results indicate that Peach and Almond share the same LTR-RT families, that have been probably inherited from their common ancestor. However, LTR-RTs have continued to transpose after the spolit of the two species, providing some specifitity to the peach and almond LTR-RT content.

2.4.3 LTR-RTs insertion time determination

Some further insights about the evolutionary history of the mobile DNA in these two species can be gathered by comparing their insertion time distribution for LTR-RTs, which can be calculated based on the number of nucleotyde differences between the two LTR of each element (¶ 2.3.6). The results are plotted in Figure 2.3 as density (Figure 2.3 - A) and count (Figure 2.3 - B) of single elements over the time expressed in Millions of Years Ago (MYA). The pots for almond and peach LTR-RTs are different and the difference between the two curves is confirmed as significant by a two-samples Kolmogorov- Smirnov test (p = 2.2 e-16). This result suggests that, although the two species share most of the LTR-RTs families, Peach and Almond seem to have accumulated their insertions in a different period. If Almond has more elements that are older than 5 MYA, Peach elements seem to be younger.

In spite of the genome similarity and the shared TE composition, the difference we find in terms of insertion time suggest a different LTR-RTs transpositional dynamics. This difference can be further appreciated in Figure 2.4. We have selected the 12 most populated families in our distribution. All of them are composed by elements in both Peach and Almond genomes. As the histogram distribution in Figure 2.4 shows, the elements in Peach tend to be younger even in the same family.

2.4.4 LTR-RTs insertional variability in Peach and Almond

The intra-specific variability of LTR-RTs can be useful to get further insights into their transpositional dynamics in Peach and Almond. In this study, we based our analysis on the WGS data produced within a CRAG- wide collaboration with the research groups led by Txosse Aranzana and Pere Arus. We analyzed 16 Peach and 19 Almond DNA-seq samples, each one obtained from DNA extracted from leaf tissue of a single tree that belongs to a given variety.

To quantify the structural variability associated to LTR-RTs, we searched both for insertions in the sample genomes not present in their corresponding reference genomes and for the absence in the sample genomes of insertions present in their respective reference genomes. The first step is carried on by the software Jitterbug ,. The program requires as input both the bam file obtained by maping of the samples paired sequences to the reference genome and the TEs annotation, scanning the paired-end reads for discordant reads and soft-clipped reads as proof of the presence of a TE (Hénaff et al. 2015). The result is an

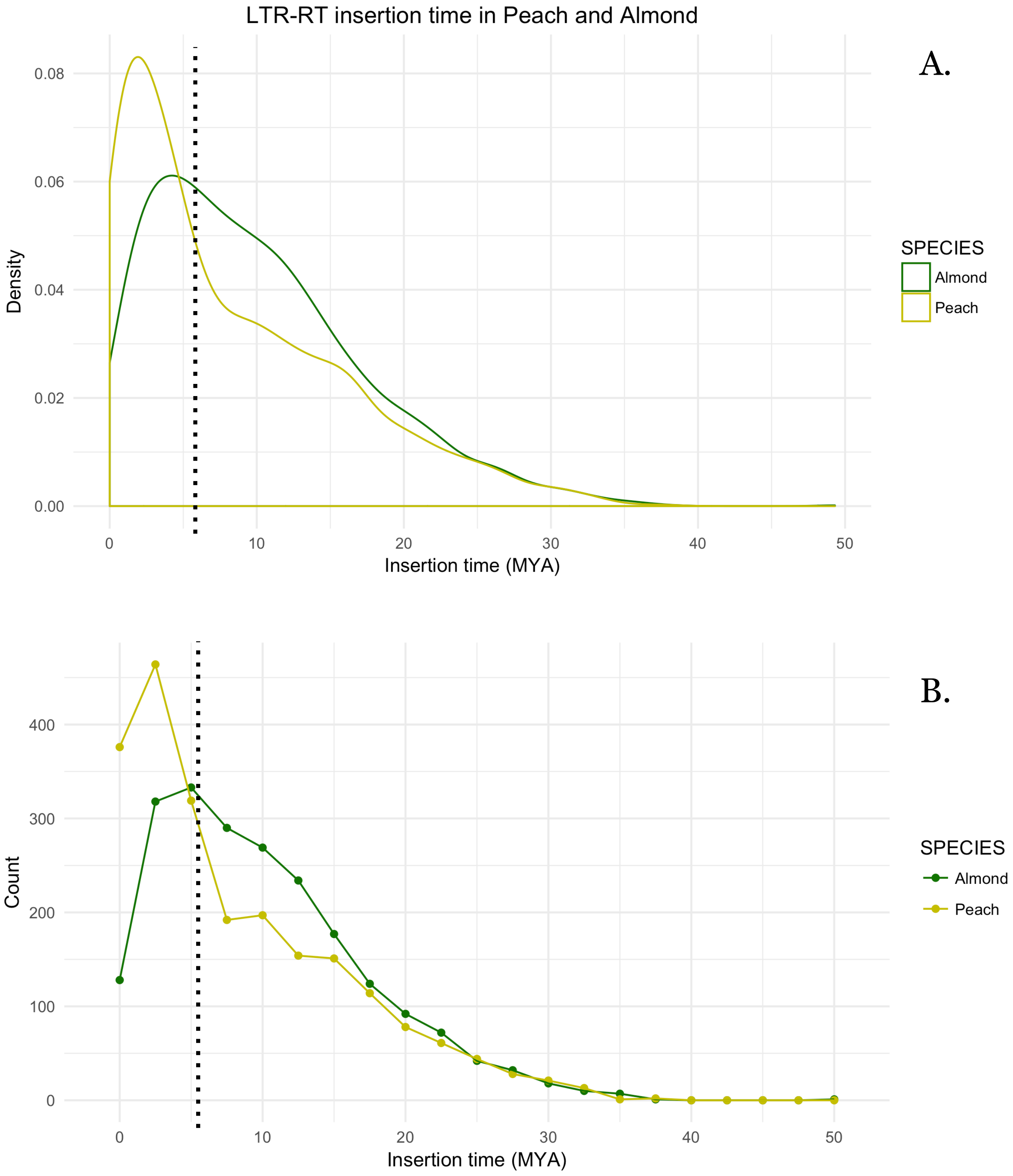


Figure 2.3 - Insertion time distribution in Peach and Almond.

The distribution of Insertion Time between Almond (**Green**) and Peach (**Yellow**) is plotted as density (A) and count (B) of single elements and expressed in Millions Years Ago (MYA).

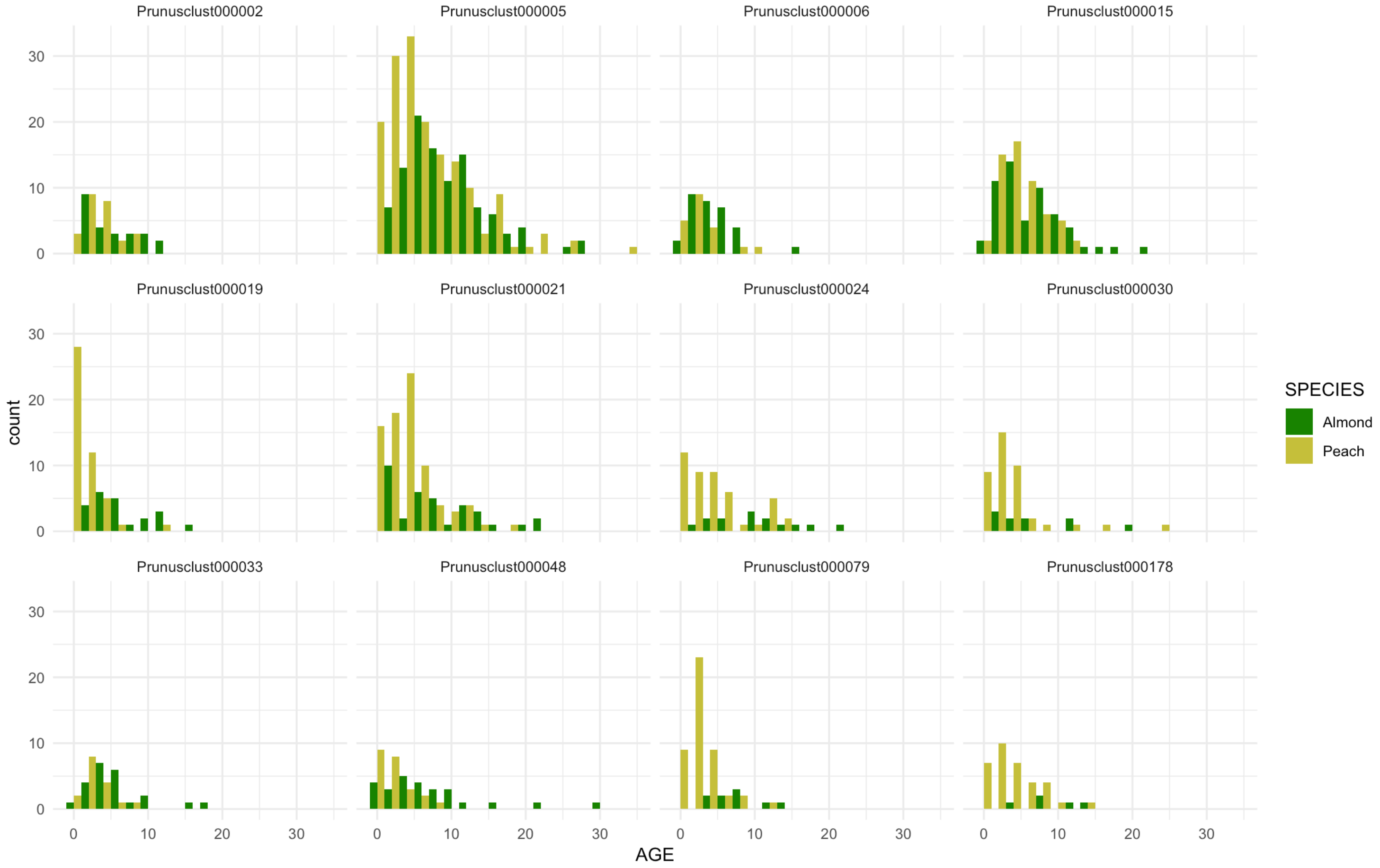


Figure 2.4 - Insertion time distribution in Peach and Almond.

Per each of the 12 most populated families, the distribution of Insertion Time between Almond (**Green**) and Peach (**Yellow**) is shown as histogram (bin = 2 MYA). Elements in Peach are generally younger than their orthologs in Almond.

annotation of positions on the reference genome that correspond to putative TEs insertions in the sample that are not annotated on the reference. The second step, consisting in detecting polymorphisms associated to absence in the sample of TEs that are annotated on the reference genome has been performed using the software PINDEL (Ye et al. 2009). This package was not specifically designed to detect TE polymorphisms, but it is a general-purpose software aimed at the detection of structural variation from DNA-Seq data, and more specifically insertions, tandem duplications, inversions and deletions. However, it can be used to detect LTR-RT polymorphisms by selecting the PINDEL deletions in that overlap to LTR-RTs annotation.

The results of this analysis are reported in Table 2.4 and Figure 2.5. For the 16 Peach and 19 Almond varieties, the number of mutations detected by PINDEL are indicated as “Reference TEs polymorphisms”, and the number of mutations detected by Jitterbug are indicated by “Non-reference TEs polymorphisms”. The comparison of the values for single varieties and of their density between species in Figure 2.5 highlight that Almond has a higher variability than Peach.

2.4.5 LTR-RTs transpositional dynamics in Peach and Almond

In the introduction we remarked that a fair description of the Transpositional Dynamics can be set up through the combination of two variables: the insertion time distribution and the insertional variability. This task is particularly tricky to be performed with Jitterbug, due to the difficulty to date non-reference insertions. Conversely, PINDEL describes the polimorphicity of elements that have been annotated on the reference and calculated for their insertion time. Hence, we focused our attention to this latter distribution of polymorphisms to link the distribution of insertion time and the variability within scanned population.

Figure 2.6 displays the number of fixed and polymorphic elements along with their insertion time distribution. Almond roughly doubles the number of polymorphic elements identified in Peach. Whilst in this latter only 16% of identified LTR-RTs in the reference genome is polymorphic, around the 35% of almond reference elements overlaps with some variation in the sampled population.

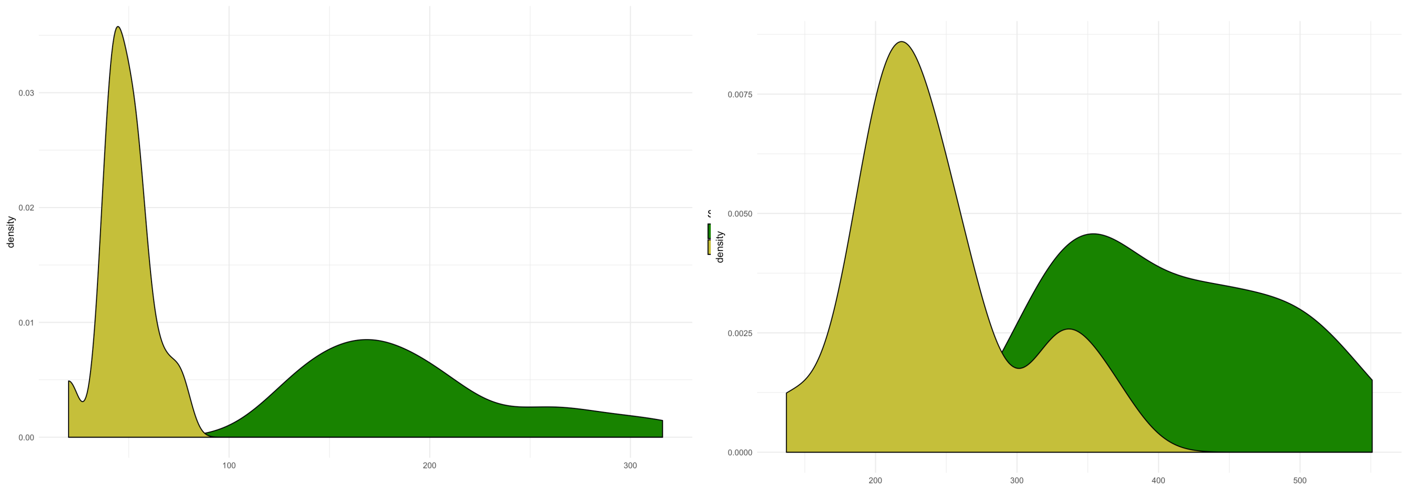
As we plot the insertion time for fixed and polymorphic elements, we see that the pattern of the insertion time distribution curves differs in two specific regions. If we split the distribution into two main timeframes, before and after the proposed speciation (5.8 MYA) as done in Figure 2.7, we see that the accumulation of insertions between Peach and Almond in these two periods is different.

Before the speciation (Figure 2.7-A), if fixed elements are comparable in number in the two species (1024 in Peach and 1080 in Almond), polymorphic elements differ considerably (496 elements in Almond and 105 in Peach, a ratio that approximates the 5 folds). After the speciation (Figure 2.7-B), the situation is inversed. Peach and Almond have a comparable number of polymorphic elements (267 and 230 respectively), while Peach almost doubles Almond in the number of fixed elements (763 and 398

Almond Varieties	Ref. TEs polymorphisms	Non-Ref TEs polymorphisms	Peach Varieties	Ref. TEs polymorphisms	Non-Ref TEs polymorphisms
A la dame	133	354	Armking	40	193
Atocha	163	423	Belbinette	44	229
Bartre	169	303	Bigtop	46	201
Doree	194	431	Blanvio	55	330
FalsaBarese	138	353	Cakereine	59	267
Ferragnes	146	338	Flatmoon	41	263
Ferrastar	167	403	Ghiaccio	67	331
GABAIS	177	311	Nectalady	20	137
GUARA	316	509	Nectaross	37	193
Garfi	256	448	Nectatop	42	250
Johnstons	217	551	PN251	49	205
Marcona	260	505	Platurno	76	366
Marinada	189	376	Sangui	52	231
Primoski	202	344	Sweetdream	43	223
P. d'Aureille	152	481	T1E425	55	228
PrincesseJR	210	372	Tifany	51	212
Ripon	128	331			
Strouds	174	448			
Vialfas Cita	288	504			

Table 2.4 - Number of LTR-RTs polymorphisms per variety.

The table reports the number of polymorphisms per variety in Peach and Almond. Reference polymorphisms are calculated on PINDEL (Ye et al 2009) result. The number of non-reference insertions is calculated on the base of Jitterbug (Henaff et al. 2015) result.



A. Reference
LTR-RTs polymorphisms

B. Non-reference
LTR-RTs polymorphisms

Figure 2.5 - Number of LTR-RTs polymorphisms per variety.

Density distributions of reference (A) and non-reference (B) polymorphisms per variety in Almond (**Green**) and Peach (**Yellow**).

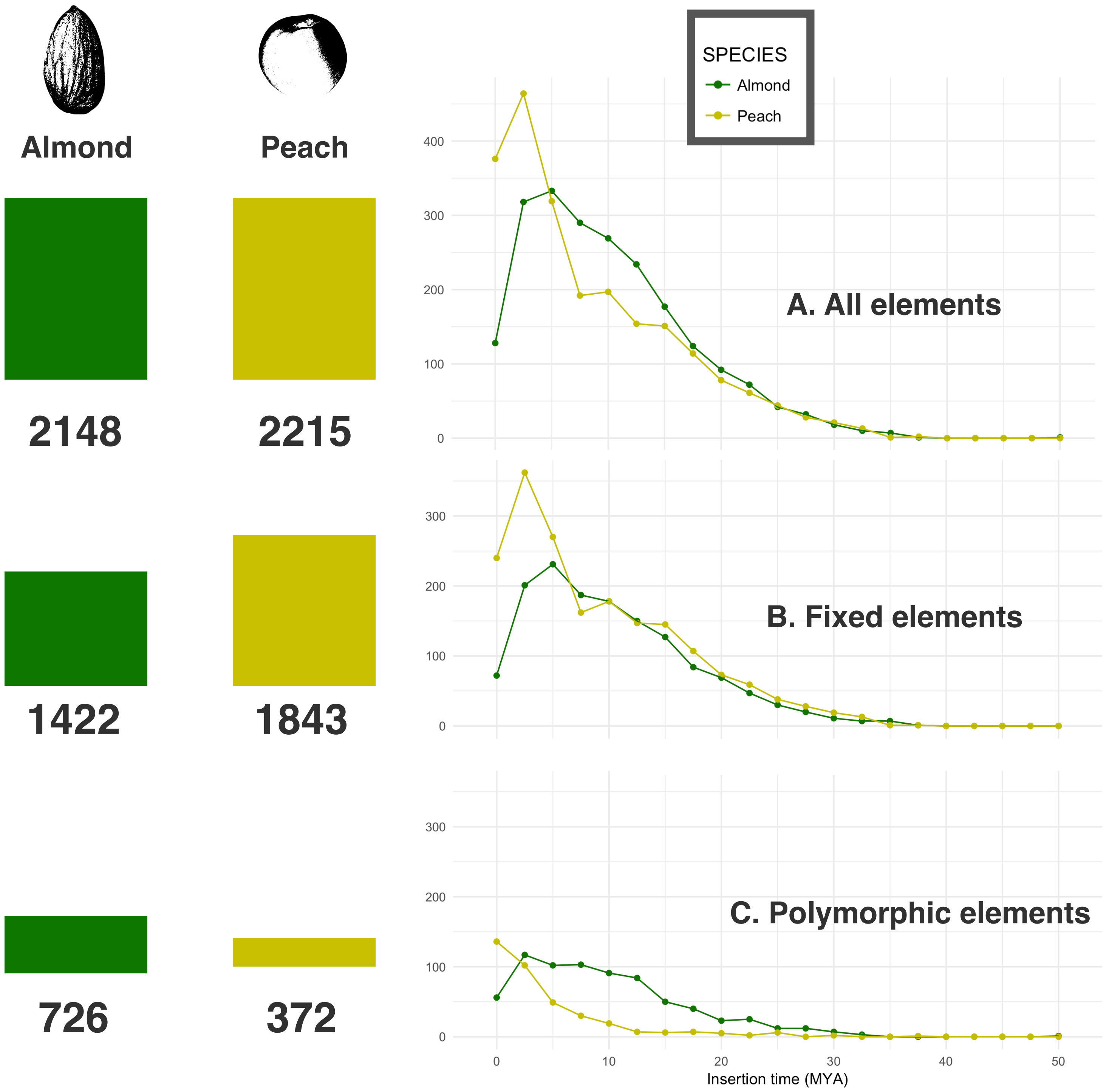


Figure 2.6 LTR-RTs polymorphisms and insertion time distirbution in Peach and Almond.

Number and insertion time distribution of total (A), fixed (B) and polymorphic (C) LTR-RTs in Almond (**Green**) and Peach (**Yellow**).

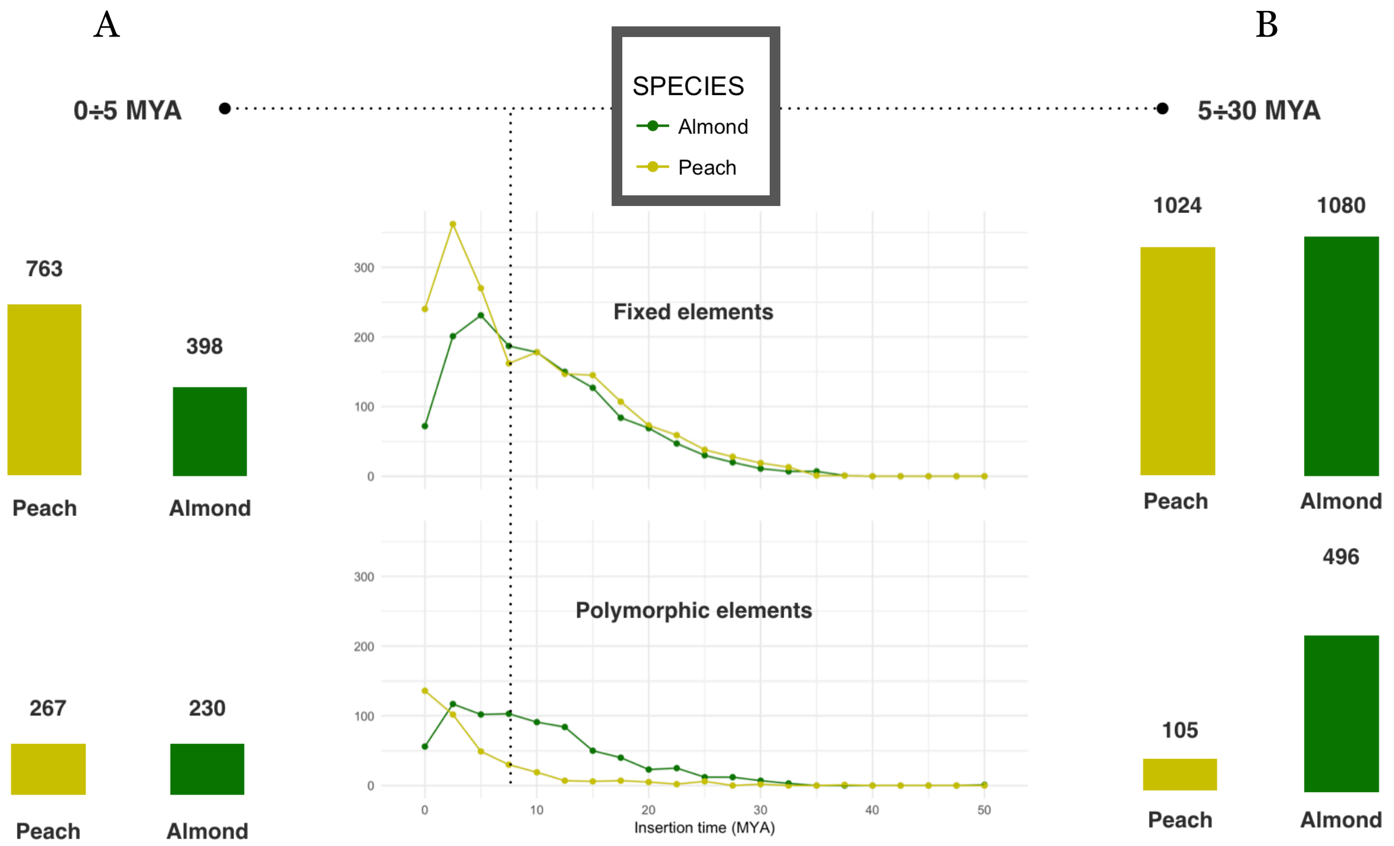


Figure 2.7 Comparison between the proportion of fixed and polymorphic LTR-RTs at different timeframes.

The ITD spectrum has been divided into two major timeframes to compare the count of fixed and polymorphic elements in Almond (**Green**) and Peach (**Yellow**).

A. In the 0÷5 MYA timeframe, Peach and Almond share the same number of polymorphic elements, but Peach has two times more fixed elements than Almond.
B. In the 5÷30 MYA timeframe, Peach and Almond share the same number of fixed elements, but polymorphic elements in Almond are 5 folds more than in Peach.

respectively). Summarizing, if Almond has more polymorphic elements that have inserted before speciation, Peach has more fixed elements that have inserted after speciation.

2.4.6 TEs elimination in Peach and Almond.

A thorough analysis of the PINDEL result can return further information on the LTR-RTs TDs in Peach and Almond. Until now, we have treated the PINDEL output in two ways. First, we have counted how many deletions overlapping an LTR-RTs are present for each species (Figure 2.5 and Table 2.4), and then we have labeled as "polymorphic" all those LTR-RTs that overlap with at least one PINDEL deletion in at least one variety (Figures 2.6 and 2.7). This software calls as "deletion" those regions that are defined by at least two split reads that serve as evidence for the absence of genetic material in that specific region (Ye et al. 2009). However, the differences that PINDEL detects as "deletions" in the samples can be either the effect of a real deletion in the sampled variety, or the consequence of an insertion in the reference variety.

To sort this off, we may compare the size of the PINDEL – defined "deletion" to the size of the element that overlaps. If their size is comparable,, the size of the deletion and the size of its overlapping LTR-RTs in the reference approximate 1, this means that an insertion occurred in the reference variety. Instead, if the size of the deletion is shorter than the size of the element, this can imply the loss of part of the element, and suggest an event of partial elimination in the sample.

For each PINDEL deletions found in Peach and Almond varieties, we plotted its ratio with the respective overlapping LTR-RTs in the reference in Figure 2.7. It shows that while in peach most of PINDEL deletions approximate the size of the element they overlap and are probably due to insertions in the reference, in Almond we detect 1102 deletions whith a ratio lower than 0.8. These cases represent the 23% on a total of 4849 deletions in the 19 Almond varieties. In Peach we have scanned 16 varieties finding an overall of 1205 mutations. Out of them, only 52 (4.3%) have a ratio with the element they overlap in the reference that is lower than 0.8. The distribution of deletion/element ratios is shown in Figure 2.7-A. In Almond, these 1102 deletions overlap with 259 polymorphic elements which rather uniform distribution on the insertion time is plotted in Figure 2.7 -B.

There are two main known mechanisms of LTR-RTs elimination: the recombination of the LTR ends that results in the elimination of the whole element except for one LTR end that is usually defined as solo-LTR, and the deletion of part of the element that leaves a truncate LTR-RT in the genome. From our data, we can't realistically assign the deletions we identified through PINDEL in Peach and Almond varieties to one of these cases, but their disproportion in the two species indicates that Almond may be more efficient in eliminating TEs than Peach and suggests that this difference could add to the TD divergence between the two species.

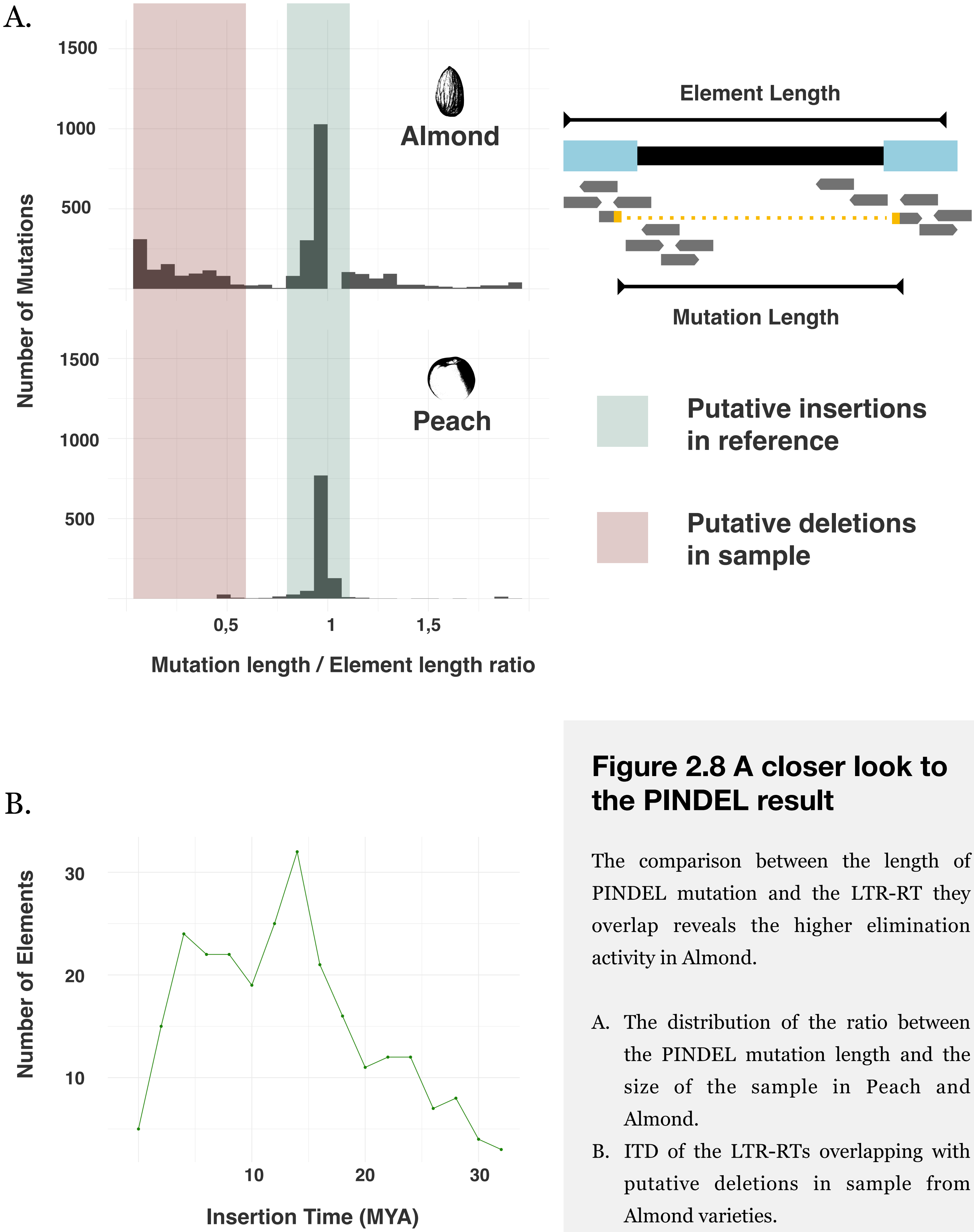


Figure 2.8 A closer look to the PINDEL result

The comparison between the length of PINDEL mutation and the LTR-RT they overlap reveals the higher elimination activity in Almond.

A. The distribution of the ratio between the PINDEL mutation length and the size of the sample in Peach and Almond.
B. ITD of the LTR-RTs overlapping with putative deletions in sample from Almond varieties.

2.5 Discussion

2.5.1 Peach and Almond share the same LTR-RTs

The first relevant result of this work is that Peach and Almond share a very similar composition of TEs. As already illustrated, we made use of two different approaches to annotate the TEs from assembled genomes. REPET is designed to assess the global transposon load of entire genomes (Flutre et al. 2015). We deployed this software to estimate the global transposon load in Peach and Almond as ranging around the 38% of whole genome coverage. This result, which is roughly 25% higher than the most recent estimation in Peach (Verde et al. 2013), is comparable with Apple's (Malus domestica) TE load, (42.4% coverage of a 750 MB genome), lower than Pear (Pyrus pyrus, 52% coverage of a 450 MB genome), but higher than strawberry (Fragaria vesca, 22% of a 800 MB genome). With respect to crops that have a big (>1GB) genome, the TE load is about one half (Sorghum and Wheat share an estimate 80% TEs load) (Vitte et al. 2014).

The workflow we based on LTR harvest is limited to LTR-RTs and is quite strict in identifying only the copies that have a good structural conservation, but it returns a set of well curated and classified LTR-RTs. Besides their differences, the comparison of both annotations in the two species indicates that Peach and Almond share the same composition of TEs. LTR-RTs cover about the 20% of the genome in both species, and the 93% of the families that we characterized are shared between them.

The similarity between the TE load and composition between the two species is consistent with their evolutionary proximity, and a first look

2.5.2 Transpositional Dynamics diverges in Peach and Almond

The population dynamics that preceded the separation between Peach and Almond may explain the higher number of older polymorphic elements in Almond. Several studies point out that the separation between Peach and Almond coincided with a reduction in size of Peach population. Paleogenomics data (Velasco et al. 2016) suggest that the raise of the Tibetean Plateau, that followed the collision between the Asian and the Indian Plaques 10 MYA (Molnar et al. 1975), might have been the trigger for the separation between Peach and Almond. The lower number of polymorphisms in Peach is consistent with the steady but reduction in size of its population during speciation (Velasco et al. 2016).

After the speciation, the higher number of fixed elements in Peach suggests that the transpositional activity in this species was higher than in Almond. Even though we do not have data proving a significant difference between the two species in terms of transpositional activity, it is true that the Tibetan plateau uplift (10-2.5 MYA), concomitant with the separation between Peach and Almond, caused drastic local climate changes on the eastern side, where fossils of early peaches have been found (Zheng, Crawford, and Chen 2014). A strong climate variation can suffice to explain both Peach reduction in population size and diversity, and an increased transpositional activity that is usually associated with several types od abiotic stresses (Makarevitch et al. 2015; Rey et al. 2016).

Although the reasons causing the differences in the temporal accumulation of TEs and in the fluctuation of their diversity in Peach and Almond need further deepening, our results strongly indicate that the TDs in these two species differs significantly.

Such a high similarity makes the comparison of TD in the LTR-RTs identified in the two species quite intriguing. We have produced three results that are relevant to the analysis of the transpositional dynamics of LTR-RTs. First, the insertion time distribution differs between the two species, with Peach LTR-RTs being relatively younger than Almond elements (Fig. 2.3 and 2.4). Second, the insertional variability is higher in Almond than Peach (Table 2.4 and Fig. 2.5) and, third, the proportion of partial deletions over putative new insertions in reference is higher in Almond (Fig. 2.8).

The loss of nucleotide variability that has been described in Peach as concomitant to its separation from Almond (Velasco et al. 2016) is observed in LTR-RTs insertional variability as well. The higher content of old insertions in Almond coincides with a major load of polymorphic elements in this timeframe (Figure 2.6), that are likely to have been lost in Peach during the population size reduction concomitant to the speciation (Velasco et al. 2016). After the inferred separation date, Peach elements had what seems to be a transpositional burst. Furthermore, the higher number and proportion of partial deletions on Almond LTR-RTs, suggests a difference in the efficiency of elimination in the two species.

TD diverges neatly between the two species even though they share the same LTR-RTs families. What is interesting under an evolutionary point of view is that besides the same LTR-RTs family have been active in both genomes, the single copies have spread independently in the two species, hitting different loci, and having a potentially different impact over the genetic variability. Even though this impact is subject to current and future investigations, the results we have presented here point to a possibly relevant impact of the different TD in Peach and Almond to the establishment of their differences at genetic level.

2.5.3 Insights on Peach and Almond evolution

The reduction of variability in Peach occurs in concomitance to the raise of Tibetan plateau, that is placed to begin at 10 MYA, to later reach an intensive uplift at 2.6 MYA (Yu et al. 2018). In our data, these time intervals coincide with the loss of variability (10-12 MYA) and with the peak of new insertions (2.5 MYA) in Peach (Figure 2.5 – 2.6). Therefore, the distribution of diversity over the insertion time distribution fits quite well with the available knowledge on the evolutionary dynamics in Peach and Almond (Yu et al. 2018; Velasco et al. 2016).

Chapter 3. LTR-RT expression in Peach (*Prunus persica*) leaves after *Xantomonas arboricola* infection.

3.1 Introduction.

Whereas the DNA- Seq allows detecting the traces that transpositional activity has left on the genome, transcriptomic data can be informative on what elements are potentially active. Transposable Elements are usually transcribed and transcription is an intrinsic part of the transpositional mechanism in Class I retrotransposons (Wicker et al., 2009). Hence, we may argue that the quantitation of TE-associate transcripts can give some approximation of the transpositional activity of LTR-RT, provided that we keep in mind that the mRNA synthesis is just the first of a long chain of events that constitute the transpositional cycle of retrotransposons. Each step is associated with a non-null probability of failure, and it is subject to silencing mechanisms deployed by the host genome to reduce TEs proliferation (Chuong et al., 2016). The proportion of retrotransposon-associated transcripts that successfully cause a new insertion is hardly inferable from the mere quantitation of transcription. Anyways, even though transcriptional activation of retrotransposons is a non-sufficient condition, it is still necessary to initiate the transposition and can be informative on what retrotransposons are more prone to generate new insertions.

Although RNA-Seq represents a viable solution to perform a genome-wide quantitation of the transcription of TEs, the repetitive nature of these elements renders this task particularly tricky. TEs are present in several copies in the genome. In our annotation on Peach and Almond, for instance, we computed TEs families on a sequence similarity criterion (see paragraph 2.4.2). The biggest family we have identified is composed by 268 elements (106 in Almond and 162 in Peach, see Figure 2.1). Such a high repetitiveness renders the quantitation of the expression of each single TE insertion very hard to be measured. The short reads (50-150 bp) that are used in most of RNA-seq experiments have the tendency to map to different loci, making difficult to say which copies are actually expressing. Also, RNA-seq datasets suffer from the presence of background noise that might be mapped to TEs, biasing the transcription quantitation towards an over-estimation of transcriptional level of bigger TE families. For this reason, the software solutions that are aimed at the quantitation of transcription of whole TEs famiies. TEtools, for instance, is designed to quantitate the expression starting from TEs consensus sequences (Lerat et al. 2014), whilst TEtranscripts is designed to normalize the quantitation of the transcription of a single family on its size, in order to reduce the size-bias.

The activation of retrotransposons has been often linked to several stress conditions (Makarevitch et al., 2014, Chuong et al., 2016), among which viral and bacterial infections can coincide with a change of transposition levels in plants (Horváth et al., 2017). Data from whole transcriptome experiments of bacterial infection can, therefore, represent a proper context to start analyzing the possibility to properly quantify the transcription level of peach retrotransposons that is potentially associated with their transposition.

The *Xanthomonas arboricola pv. Pruni* (Xap) represents a very challenging problem for peach growers. It's able to infect all the green tissues of the plant, and the first symptoms are represented by dark scares – usually on leaves – that grow rapidly into greasy bacterial oozes (Boudon et al., 2005). The infection is spread plant-wide by rain splashes, and it can bring to premature defoliation that is associated with severe yield loss in fruit production (Lamichhane et al., 2014). Xap is considered as a major threat in Prunus crops production, and if the understanding of resistance mechanisms is a point of big agricultural interest, the analysis of the transcriptomic changes that occur during Xap infection can give some useful insights on the level of activity of LTR-RTs in Peach and how it variates between several cultivars and condition.

The study entitled "Transcriptome reprogramming of resistant and susceptible peach genotypes during Xanthomonas arboricola pv. pruni early leaf infection" (doi: https://doi.org/10.1371/journal.pone.0196590) was published in early 2018 and makes a good case to investigate the behaviour of LTR-RTs during a Xap infection. Peach cultivars can be sorted by their susceptibility to Xap infection into susceptible, moderately susceptible, moderately resistant and highly resistant (Gervasi et al., 2018). This study compares the transcriptome changes between the moderately susceptible cultivar "JH Hale" and the highly resistant cultivar "Redkist". Leaf samples were extracted 30 minutes, 1 hour and 3 hours post infection (hpi) and the transcriptome changes were investigated through RNA-Seq analysis. The aim was to identify candidate genes that are potentially connected with Xap resistance in Peach.

3.2 Scope of the chapter

The main purpose of the work described in this chapter is to perform a quantitation of the transcriptional level of LTR-RTs in Peach. To do so, we will analyze the publicly available RNA-Seq data relative to this experiment (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388577) in order to quantify the transcriptional level associated with LTR-RT transpositional activity. Peach cultivars JH Hale and Redkist differ for their sensitivity to *Xantomonas aroboicola var prunii* (Xap) infection, with JH Hale being more susceptible than the highly resistant Redkist. Gervasi et al. have extracted and sequenced transcriptomic RNA at 0.5, 1 and 3 hours post infection (h.p.i.). We will make use of the publicly available databases coming from this work to quantitate the expression of LTR-RTs in Peach during the progression of a pathogen infection.

This chapter will hence start from a concrete example to provide an overview of the potentiality and limitations of RNA-Seq analysis to quantitate the expression associated to transposable elements. Also, we will select the differentially expressed genes that lay next to a LTR-RTs insertion to assess the potential impact of transposition on gene regulation in Peach.

3.3 Materials and Methods

3.3.1 Data origin and background experiment.

The article entitled “Transcriptome reprogramming of resistant and susceptible peach genotypes during Xanthomonas arboricola pv. pruni early leaf infection” (Gervasi et al., 2018) was published by the end of April 2018 with the intent to show how the leaf transcriptome reprogramming after infection with the proteobacteria Xanthomonas arboricola (Xai) variates between the variety JH Hale, that is more susceptible to infection, and the variety Redkist, that is more resistant. RNA extraction is performed as size-fractionation in a time course fashion, with sampling at 0.5, 1 and 3 hours-post-infection (h.p.i.). Per each condition, at least three technical replicates have been done. Data are available online at the National Center for Biotechnology Information - Short Reads Archive (NCBI-SRA) database (https://www.ncbi.nlm.nih.gov/sra) under project accession PRJNA388577 as Illumina HiSeq 2000 short reads fastq files.

3.3.2 Quality check, data trimming and reads alignment

The quality of each fastq distribution has been checked with the software fastqc v. 0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Fastq have been subsequently trimmed by the software trimmomatic v. 1.0 (Bolger et al., 2014) to remove the adapter and potential ribosomal RNA (rRNA) contamination. Short reads have been then aligned to the Peach Genome v. 2.0 (Verde et al., 2013) with the program STAR v. 2.6 (Dobin et al., 2015), allowing the alignment of multi-mapping reads aligning to up to 100 positions, as suggested by TEtranscripts developers to detect transposition (Jin et al., 2015).

3.3.3 Transcripts quantitation and differential expression

We then submitted ran the package TE transcripts providing our LTR-RT annotation along with the gene annotation released with the Peach Genome v. 2.0 and available online (https://www.rosaceae.org/organism/Prunus/persica). TE transcripts integrate the reads quantitation, the normalization and the calculation of differential expression by DESeq2 (Jin et al, 2015, McDermaid et al., 2012). The pipeline has been run with standard parameters and allowing the count of multi-mapping reads.

3.3.4 Background noise calculation and filtering

Genomic contamination and random non-specific transcription can alter a transcriptomic RNA sample, generating background noise in downstream calculations (Lin et al., 2016). Although this problem is often underestimated in RNA-Seq analyses, the presence of the background may increase the number of false positives. For these reasons, we have masked the peach genome to exclude all the nucleotides annotated as genes or TEs (REPET annotation), to obtain an annotation of genome windows that are realistically non-coding regions. This annotation was submitted to TE transcripts as gene annotation. The normalized result (baseMean) for each window has then been plotted to view its distribution. A value corresponding to the 99 percentile of the distribution (baseMean = 350) has been selected as a minimum threshold to accept an expressed feature in downstream analyses.

3.4 Results

3.4.1 Experiment and dataset selection

As said, this chapter is thought to discuss the potential of NGS- based transcriptome analysis in a context that matches with the thesis workframe, i.e. the transpositional dynamics in crop species belonging to the Prunus genus. At present time (January 2019), the laboratory is working on RNA extraction from peach and almond samples to be sequenced by an RNA-seq strategy and, at the moment, our efforts on the computational side are centered on the development of a robust workflow to quantitate TE-associated transcription from deep sequencing data. To do so, we decided to rely on publicly available datasets and to select those ones that fulfilled our requirements.

We scanned the NCBI-SRA (https://www.ncbi.nlm.nih.gov/sra) database to select peach or almond transcriptomic RNA-Seq datasets that responded to some criteria, such as the presence of biological replicates, a fair coverage and a minimum quality of the reads.The dataset relative to the article “Transcriptome reprogramming of resistant and susceptible peach genotypes during Xanthomonas arboricola pv. pruni early leaf infection” (Gervasi et al., 2018) — associated with the NCBI project accession PRJNA388577 — is particularly fit to what we needed to implement our workflow.

As Table 3.1 shows, the project PRJNA388577 is available on NCBI-SRA as a dataset of 32 distinct files, each one relative to a single run. The susceptible variety JH Hale and the resistant Redkist have been infected with Xantohomonas arboricola var prunii (Xap), and leaf tissue was extracted at 0.5, 1 and 3 hours post infection (h.p.i.). Each condition has four biological replicates (Gervasi et al., 2018). Table 3.1 indicates the size of each sample in column 4 in millions reads, indicating that the samples have a high coverage (34 millions reads on the average). If the presence of biological replicates is needed to assess the statistical significance of differential expression (McDermaid et al., 2012), a high coverage is fundamental to successfully detect transcription associated to mobile elements. TEs usually have a low transcription level, and a good quality and depth of the sampling is needed to quantitate it (Jin et al, 2015). This dataset thus satisfies the technical requirements to permit the quantitation of LTR-RT transcription, and it is generated from samples of leaves during a bacterial infection course, a condition that is usually associated with an increase of transpositional activity (Horváth et al., 2017). It is hence a good example to investigate LTR-RT transcriptional activity in Peach.

3.4.2 TEs transcription beyond background noise

RNA-seq samples can suffer from background noise coming from either non-specific transcription and the sample contamination from genomic DNA (Lin et al., 2016). Although the effect of this noise is practically negligible during downstream analyses that are aimed at determining gene expression, quantifying TEs

Variety Name	Sampling Condition	FASTQ file size (millions reads)	Run ID (NCBI -SRA)
JH Hale	Control	34.6	SRR5631978
JH Hale	Control	49.4	SRR5631977
JH Hale	Control	27.7	SRR5631984
JH Hale	Control	38.3	SRR5631983
JH Hale	0.5 h.p.i.	21.0	SRR5631957
JH Hale	0.5 h.p.i.	46.6	SRR5631958
JH Hale	0.5 h.p.i.	20.1	SRR5631959
JH Hale	0.5 h.p.i.	38.5	SRR5631960
JH Hale	1 h.p.i	29.3	SRR5631961
JH Hale	1 h.p.i	34.0	SRR5631962
JH Hale	1 h.p.i	39.7	SRR5631963
JH Hale	1 h.p.i	46.4	SRR5631964
JH Hale	3 h.p.i	36.6	SRR5631955
JH Hale	3 h.p.i	47.7	SRR5631956
JH Hale	3 h.p.i	25.6	SRR5631986
JH Hale	3 h.p.i	40.2	SRR5631985
Redkist	Control	31.3	SRR5631974
Redkist	Control	34.8	SRR5631973
Redkist	Control	29.8	SRR5631972
Redkist	Control	35.8	SRR5631971
Redkist	0.5 h.p.i.	29.6	SRR5631970
Redkist	0.5 h.p.i.	41.1	SRR5631969
Redkist	0.5 h.p.i.	20.7	SRR5631968
Redkist	0.5 h.p.i.	37.6	SRR5631967
Redkist	1 h.p.i	18.7	SRR5631966
Redkist	1 h.p.i	47.0	SRR5631965
Redkist	1 h.p.i	16.5	SRR5631980
Redkist	1 h.p.i	40.6	SRR5631979
Redkist	3 h.p.i	37.3	SRR5631982
Redkist	3 h.p.i	52.3	SRR5631981
Redkist	3 h.p.i	27.9	SRR5631976
Redkist	3 h.p.i	35.9	SRR5631975

Table 3.1 - Gervasi et al. experiment on *Xanthomonas arboricola*. Dataset features and availability.

Details of the samples published along with the "*Transcriptome reprogramming of resistant and susceptible peach genotypes during Xanthomonas arboricola pv. pruni early leaf infection*" article (Gervasi et al., 2018). The *Xanthomonas arboricola* infection course experiment is repeated over two varieties (*JH Hale* and *Redkist*) at three different timings (0.5, 1 and 3 h.p.i.). RNA is extracted in 4 biological replicates per condition including the control. Data are available at *https://www.ncbi.nlm.nih.gov/bioproject/PRJNA388577*.

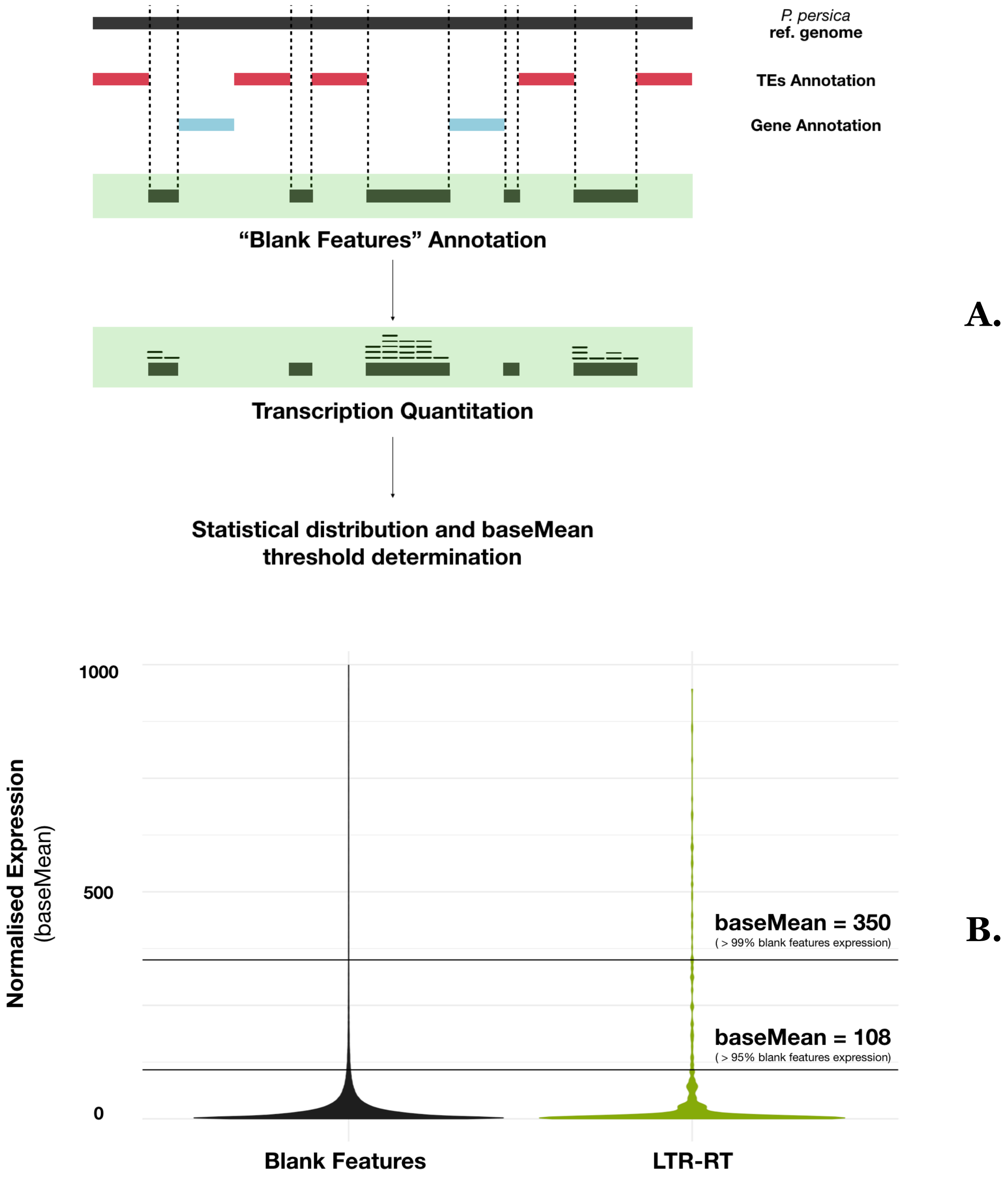


Figure 3.1 Background noise calculation and baseMean threshold determination.

A. The workflow to quantify the background noise in RNA-seq data begins with the creation of a "Blank Features" annotation that collects all the genomic regions that are not transposons or genes. To this annotation, we linked the quantified and normalised transcription from *JH Hale* control samples (SRR5631977, SRR5631978, SRR5631983, SRR5631984). The distribution of average baseMean values per features is then used to calculate a filtering threshold.

B. The results are plotted as violin plot on normalised expression in BaseMean (normalised reads count for feature length). We traced two BaseMean thresholds that include respectively the 95% (BM=108) and 99% (BM=350) of blank features.

transcription requires us to put greater attention in it. TEs are usually transcribed at a very low level and, in many cases, determining whether the transcription associated to a specific family is the effect of active transcription at low frequency, or background noise might not be straightforward (Jin et al., 2015).

To tackle this, we implemented a protocol to quantify transcriptional background noise from our data. As described in the graph in Figure 3.1- A, our first concern is to build up an annotation of all sequences that are possibly not liable to active transcription. These "blank" genomic features are obtained by rallying all sequences that are not annotated as genes or as transposable elements. We have merged our annotation of TEs from all Classes (REPET annotation) with the publicly available Prunus persica gene annotation (ftp://ftp.bioinfo.wsu.edu/species/Prunus_persica/Prunus_persica-genome.v2.0.a1/genes/), to then obtain an annotation with all those genomic intervals that are not covered by either transposons or genes. Then, we quantitated and normalised the expression linked to these blank features in the JH Hale control samples (SRR5631977, SRR5631978, SRR5631983, SRR5631984), and compared its distribution with the one that is relative to the LTR-RTs expression. The violin plot in Figure 3.1- B explains why our concerns on the possible effect of background noise over TEs expression downstream analyses are well justified. The chart compares the relative distribution of baseMean value per feature between blank features and LTR-RTs, that show a very similar distribution. The 95% of the TEs and blank features have a baseMean value that is lower than 108, and the 99% of them is below 350. To minimise the possibility of having false positives, we decided to consider a minimum threshold of 350 baseMean to accept a TE family as expressed.

3.4.3 Eleven LTR-RT families are expressed in JH Hale and Redkist

TE transcripts is one of the few publicly available software solutions that has been designed to return the expression of TE families along with gene expression in RNA-Seq analysis (Jin et al., 2015). The repetitive nature of TEs makes the quantitation of transcription rather difficult for two reasons. First, the high sequence similarity between different insertions causes the increase of multi -mapping reads, jeopardizing any possibility to determine univocally which insertion is transcribed. Young insertions are the ones that are expected to be transcriptionally more active, but at the same time are the ones with higher sequence similarity with other insertions, and this means that they are more prone to align multi-mapping reads. Hence, a quantitation of expression for the single insertions would be hardly realistic, despite some very recent efforts will be mentioned in the discussion to this chapter. We can still work on the quantitation of transcription at family-level, but this takes us to the second big problem that the repetitiveness of TEs will give us. As we have seen in the previous paragraph, the background noise in RNA-seq data is enough to be on par with the transcription level associated to most of our LTR-RTs. TEs can be linked with reads that are generated by background noise, and their impact on the final quantitation at the family level is likely to be proportional to the size of the family. A family with more insertions will accumulate more background noise than a smaller family, and the resulting quantitation would be the artifactual effect of the size of the family rather than the result of actual transcriptional activity. TE transcripts tackle this inconvenience by re-proportioning the quantitation of TE families over their size by the application of the Expectation-Maximization (EM) algorithm (Jin et al., 2015). The result is a inference of the transcriptional activity at family-level which effect of the family size is minimized.

We have run TE transcripts by submitting the gene annotation and our LTR-RT annotation along with the bam files. The pipeline quantitates the transcription on genes and transposons separately, applying the EM correction on these latter. Then, the inference of differential expression is done by submitting the count tables to the package DEseq2 (McDermaid et al., 2012). The resulting TE Expressed Families (EF) are filtered to remove the above-mentioned background noise threshold (baseMean > 350). The filtering returns all those families that display a baseMean value > 350 in at least one condition, including control, and in at least one variety.

Figure 3.2 and Table 3.2 show some information on the 11 EFs that are found active in the two varieties. There are 11 families which expression is quantitated as above the threshold. Seven of them (Pclust-6, Pclust-21, Pclust-33, Pclust-48, Pclust-87, Pclust-144, Pclust-155) are classified as Copia-like families, three are Gypsy-like families (Pclust-3, Pclust-76 and Pclust-131), whilst only the Pclust-202 family is marked as Unclassified. Four families are expressed differentially after infection. In Figure 2.3 - A the expression is plotted as heatmap, and show that the three Copia families Pclust-33, Pclust-21 and Pclust-6 and the Gypsy family Pclust-3 are differentially expressed (f.c. > 1.5 in at least one sample) after infection. Expression level and fold-change per sample are reported in Table 2.3. Among these 4 families, Pclust-21 and Pclust-s3 show an expression pattern that is different between JH-Hale and Redkist. Pclust-21 results downregulated in Redkist during infection, whilst its expression level in JH Hale doesn't seem to change considerably after Xai inoculation. Similarly, Pclust-33 results downregulated at 0.5 and 1 h.p.i. in JH Hale but meets no particular variation in Redkist during infection.

The absolute value of expression level that is reported in Table 2.3 clearly shows that the expression associated to family Pclust-21 outstands the expression of the other families, reaching a baseMean=17562 in JH Hale and a baseMean=22961 in Redkist. In Figure 23 B, the distribution of the elements in Peach and Almond is shown in relation with the insertion time. In general, most of the insertions in Peach are relatively young, particularly in the family Pclust-21, that shows a peak of recent insertions.

3.4.4 Expression of genes that are flanked by LTR-RTs.

Peach and Almond share a comparable number of predicted genes, that are roughly around 25.000 units in each species (Verde et al. 2013, Alioto et al. 2019). In order to perform an initial assessment of the potential impact of transposition on gene regulation, we counted those genes that lay next to a TE, at a

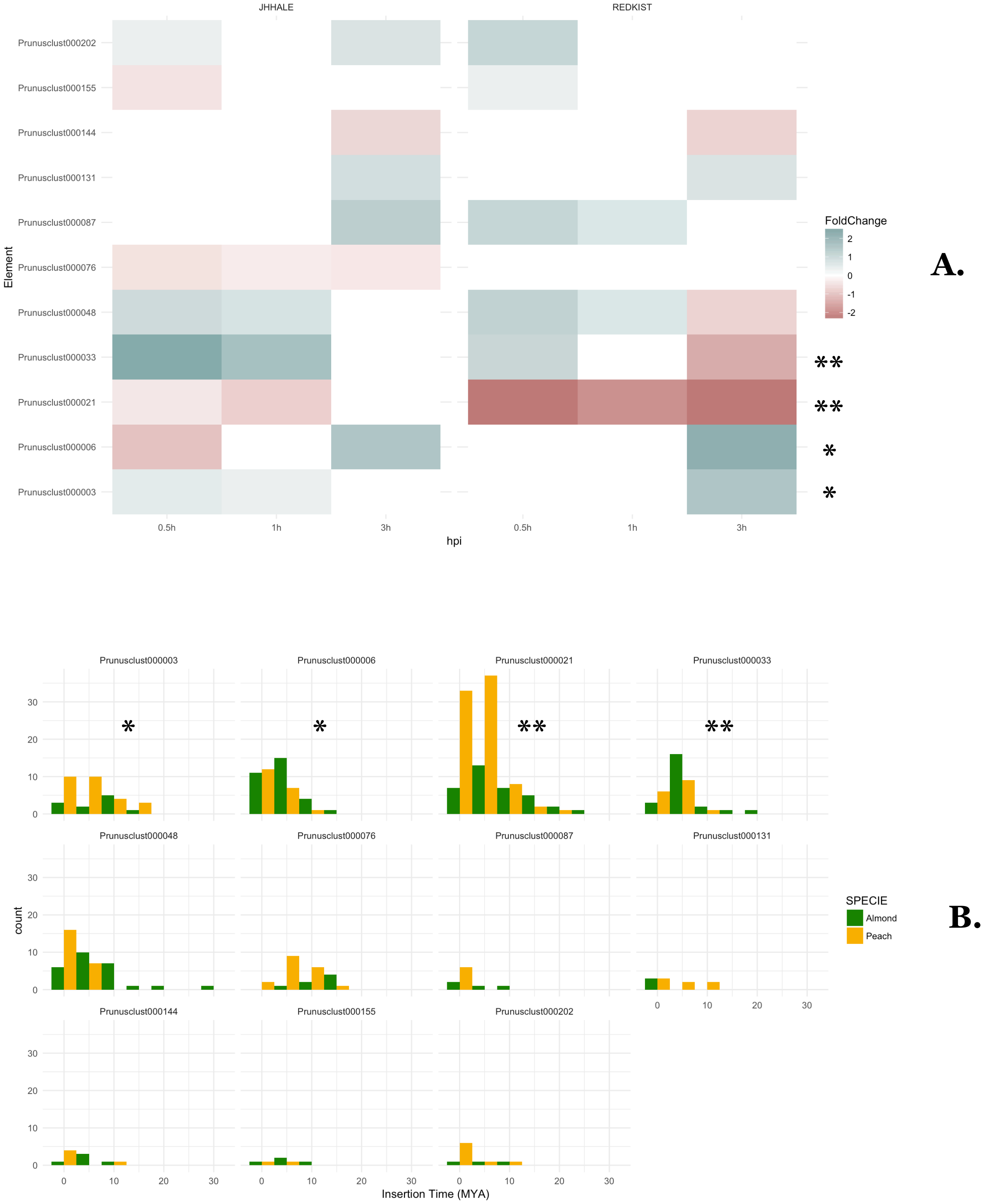


Figure 3.2 LTR-RT expressed family in *JH Hale* and *Redkist* varieties at different timings of *X. arboricola* infection.

A. Differentially expressed LTR-RT families in JH Hale and Redkist at 0.5, 1 and 3 hours post infection (hpi).

B. Insertion Time distribution of the elements belonging to expressing families in Peach (yellow) and Almond (green).

* Elements that are differentially expressed in at least one of the samples with fold change > 1.5.

** Elements which expression pattern differs in the two varieties

		JH Hale				Redkist			
LTR-RT Family	**Classification**	**Ctrl (base Mean)**	**0.5 h.p.i. (log$_2$ FC)**	**1 h.p.i. (log$_2$ FC)**	**3 h.p.i. (log$_2$ FC)**	**Ctrl (baseMean)**	**0.5 h.p.i. (log$_2$ FC)**	**1 h.p.i. (log$_2$ FC)**	**3 h.p.i. (log$_2$ FC)**
Pclust-3	**Gypsy**	1072	0.50	0.40	NotSig	882	NotSig	NotSig	1.59
Pclust-6	**Copia**	816	-1.03	NotSig	1.57	596	NotSig	NotSig	2.24
Pclust-21	**Copia**	17562	-0.40	-0.80	NotSig	22961	-2.38	-1.94	-2.39
Pclust-33	**Copia**	451	2.43	1.74	NotSig	222	1.07	NotSig	-1.46
Pclust-48	**Copia**	1510	0.96	0.78	NotSig	1125	1.26	0.68	-0.72
Pclust-76	**Gypsy**	381	-0.48	-0.31	-0.37	405	NotSig	NotSig	NotSig
Pclust-87	**Copia**	1743	NotSig	NotSig	1.36	821	1.11	0.66	NotSig
Pclust-131	**Gypsy**	551	NotSig	NotSig	0.89	698	NotSig	NotSig	0.72
Pclust-144	**Copia**	1065	NotSig	NotSig	-0.67	559	NotSig	NotSig	-0.72
Pclust-155	**Copia**	463	-0.44	NotSig	NotSig	389	0.38	NotSig	NotSig
Pclust-202	**Unclassified**	337	0.37	NotSig	0.74	432	1.12	NotSig	NotSig

Table 3.2 Expression of LRT-RT families

The 11 LTR-RT families that are found expressing in JH Hale and Redkist peach varieties in uninfected control (ctrl) and at 0.5, 1 and 3 h.p.i..

distance that is lower than 1 Kb. In Table 3.4 A, we report the numbers, distinguishing between those ones having an upstream LTR-RTs insertion, that is potentially involved in cis-gene regulation, and those ones having an LTR-RT downstream, that could hence influence the epigenetic regulation. In Almond, we find 561 genes flanked by an LTR retrotransposon, 291 upstream and 270 downstream. In Peach, there are 614 genes that are flanked by an LTR-RT. Out of them, 318 have an upstream LTR-RT insertion, and 296 a downstream insertion.

The experiment exposed in this chapter allows us to count the genes which regulation is potentially affected by the presence of a flanking LTR-RT and that are differentially expressed during an infection. We have hence selected the genes that are significantly differentially expressed during infection with a baseMean > 150. Then, we sorted the resulting genes into "small variation" (-1.5 < log2FC < 1.5), "upregulated" (log2FC > 1.5),or "downregulated" (log2FC < -1.5), Finally, we counted those genes that lay +/- 1000 bp from a LTR-RTs. Results are reported in Table 3.4. In general, our analysis confirms what Gervasi et al. proposed in their publication. Redkist has a higher number of differentially expressed genes during infection. JH Hale counts 73 genes differentially expressed, 50 upregulated and 23 downregulated. Redkist displays 136 differentially expressed genes (96 upregulated and 40 downregulated), which is almost 2 fold the number of genes that meet strong expression variation in JH Hale. As for the genes expressed with small variations, we notice that the ratio is inverted in the two species. JH Hale counts for 12557 expressed genes with -1.5 < log2FC < 1.5, whilst Redkist counts for 9666.

The two varieties also differ in terms of number of differentially expressed genes that are flanked by a LTR-RT. In JH Hale, the 1.46% of genes expressed with small variation and the 24.7% of differentially expressed genes are flanked by a LTR-RTs. In redkist, these number rise to respectively the 11% and 50%.

A. Total number of genes that are flanked by TEs in Peach and Almond genomes

	Almond	Peach
Close |d| < 1000 bp	561	614
Close Upstream d < -1000 bp	291	318
Close Downstream d > 1000 bp	270	296

B. Genes flanking a LTR-retrotransposon that are expressed (baseMean > 150) during *Sai infection in* JH Hale and Redkist.

	Peach var. JH Hale			Peach var. Redkist		
Classification	Small Variation -1.5< log$_2$FC <1.5	Upreg log$_2$FC > 1.5	Downreg log$_2$FC < -1.5	Small Variation -1.5< log$_2$FC <1.5	Upreg log$_2$FC > 1.5	Downreg log$_2$FC < -1.5
Close |d| < 1000 bp	180	14	4	106	48	20
Close Up d < -1000 bp	89	6	2	49	4	0
Close Down d > 1000 bp	91	8	2	57	12	3
Far |d| > 1000 bp	12197	22	15	9454	32	17
Total	12557	50	23	9666	96	40
% Close	**1.43%**	**24.7%**		**11 %**	**50 %**	

Table 3.4 Differentially expressed genes that flank an LTR-RTs insertion in Peach

In **A**, we report the total number of genes in Peach and Almond reference genomes that lay at a distance that is < 1000 bp (upstream or downstream) from a LTR-RTs insertion. In **B**, we focus on the genes that are found expressed and flank an LTR-RTs. We apply the limit of 150 baseMean to consider a gene as "expressing", a log2FC > 1.5 threshold for up-regulated and a log2FC < -1.5 for down-regulated genes.

3.4 Discussion

3.4.1 RLC-21 is an active family that has contributed to the recent increase of LTR-RTs insertions in Peach

The evolutionary dynamics of TEs that is presented in Chapter 2 highlights how LTR-RTs have had a rapid increase in new insertions in the most recent evolutionary time-frame (0-5 MYA) in Peach that is not confirmed in Almond. Despite this burst doesn't look as the consequence of the activation of one or few specific families, but a more generalized increase of the fixation rate in Peach, some families have been particularly active in recent times, and as the Figure 3.2 A. demonstrates, identified LTR-RTs EFs have had recent transpositional activity.

We must not forget that the mere transcription quantitation does not constitute proof of transposition in RTs. As already mentioned, there are several steps in the transpositional cycle that are liable to failure or to genome counteracting mechanisms, and there is no linearity in the relationship between transcriptional level and frequency of new insertions. A new RT insertion can't be generated without the transcriptional activation of an existing RT copy, and we do recognize transcription as a necessary but non-sufficient condition for transposition. Even if we can assume the event of a new insertion as highly sporadic, transcription is a si ne qua non condition for this to happen. Here's why the results we get in this study can provide us with some data that are consistent to what the insertion time distribution already told us.

In this, our attention goes on the RLC-21 family because of its high expression level, that is 10 fold higher than the second in rank, RLC-87 (Table 3.2). RLC-21 is a family that collects 35 insertions in Almond and 81 in Peach. As Figure 3.2 shows, the difference between the two species comes from a recent increase of the insertions in Peach. In this species, there are 58 RLC-21 insertions that are younger than 5 MYA, of which 16 are younger than our limit of detection (0.25 MYA). The high expression of this family is thus consistent to what we had found in our annotation. RLC-21 is a family that has accumulated recent insertions in Peach, and that is found constitutively expressed and downregulated in Peach variety Redkist during Xai infection.

3.4.2 The potential impact of transposition on immune response in Peach

A very interesting consequence of this PhD project is the possibility to assess the impact of transposition on gene regulation. To do so, we need to deepen our knowledge of Peach and Almond genetics (Zhang 2016). At the moment, our initial exploration of RNA-seq data is able to provide a fair outlook of the potential impact of transposition on the regulation of genes that are potentially involved in immune response to Xai infection.

This outlook is based on the widely accepted assumption that the proximity of a TE insertion to a gene can affect the regulation of this latter (Hollister and Gaut 2009, Yadav et al 2018). If upstream insertion can potentially provide novel regulatory regions, downstream insertions can change gene regulation by modifying the local epigenetic status(Bennentzen and Wang 2014). We have chosen a threshold of 1 kb to include both of these possibilities.

The result is that about ¼ of differntially expressed genes in JH Hale during Xai infection are flanked by TEs. In the resistant variety Redkist, this level reaches the 50%. However explorational, these results constitute a strong indication of a possible impact of transposition on the evolution of immune response in Peach.

Unfortunately, this analysis is based on the annotation on the Peach reference genome, and we have no information about the polimorphicity of these elements in the two analyzed varieties, which assessment could lead to a better understanding of the actual effect of LTR-RTs transposition on the regulation of Xai infection-induced or repressed genes.

3.4.3 Optimising the workflow for LTR-RTs RNA-Seq downstream analysis

The laboratory is now working on extracting transcriptome data from Peach and Almond to be analysed in the next future, and this is the reason why our priority at the moment is to implement a protocol for the quantitation of transcription that is associated with RTs and eventually to transposition. The advantage of choosing this study on transcriptome reprogramming in Peach during Xanthomonas arboricola pv. Prunii is that we could build our workflow on one of the species we aim to study and to obtain some information on TEs dynamics in Peach that is relevant to this thesis. But the same construction of a protocol that relies on state-of-the-art algorithms to tackle a not so easy task as the quantitation of TEs transcription and its linkage to gene activity, represented a high priority during this work.

As already discussed, the repetitive nature of TEs renders the quantitation of their transcription rather difficult because of the impossibility to unambiguously link NGS short reads to a single TE locus (Jin et al., 2015). Also, the linkage of transcriptional background noise (Lin et al., 2016) to the loci of the same family has the undesired effect to considerably raise the level of transcription in big families (Jin et al., 2015) and to generate false positives. We decided to deal with these drawbacks by applying a background noise filter and by relying on the TEtrasncripts package that is thought to compensate for the family size during transcriptional quantitation, obtaining a realistic result at family-level. Despite some recent efforts to implement a solution to detect transcriptionally active TEs loci were published very recently (Valdebenito-Maturana and Riadi, 2018), the quantitation at family-level still represents the most widely accepted and reliable approximation.

If this quantitation gave us the opportunity to confirm that active families collect recent insertions, the linkage with gene expression has the potential to provide more insights on the impact of transposition over biological processes such as immune defense. In conclusion, even though we didn’t perform an actual benchmarking on TEtranscripts, the result is enforced by its consistency with our DNA-seq based data on LTR-RTs activity in Peach.

General discussion and final remarks

Part A. The bioinformatics of mobile DNA

During the development of this project, we needed to tackle three bioinformatics problems that are very common in the computational investigation of mobile DNA. The first is the de novo annotation of TEs from assembled genomes in fasta format. The second is the annotation of TEs polymorphisms from population data and the third is the quantitation of transcription of retrotransposons from RNA-seq. All these problems, however different, are made particularly challenging by some specific features of the TEs that are briefly discussed.

The first feature is indeed the repetitiveness. TEs are present in multiple copies, and some TEs families can be made of hundreds of almost identical elements that are spread throughout the genome. This has a dramatic effect on the NGS analysis since the 100-150 bp short reads that come from repetitive elements end up mapping in different loci, and it is very challenging to precisely measure, for instance, the expression level associated to a single TE copy. Repetitiveness has also an effect on the quality of the genome assembly. A very explanatory reading about how repetitive regions can complicate a genome assembly was published in 2018 by Peona et al.. Using a nice allegory, the authors compare the repetitive regions in a genome assembly to those puzzle pieces that occur several times in a puzzle game, and that are particularly difficult to place. They contain ambiguous information on their exact position, and this renders the correct assembly of repetitive regions quite difficult (Peona et al 2018). If several repetitive sequences are not even assembled, regions that have a high concentration of TEs can be affected by wrong assemblings and, in general, these regions may result collapsed in the final assembly (Amemiya, Kundaje and Boyle 2019).

A second problem is given by TEs fragmentation. The most of transposon-derived sequences is constituted by elements that are mutated, inserted or partially deleted (Bao et al. 2015). Some TEs, like the LTR retroelements, have the tendency to insert each other and form nested insertions and chimeric elements (Lexa et al. 2018). A TEs annotation needs hence to be able to distinguish between entire elements, that are potentially functional and interesting for the analysis of polymorphisms and the expression quantitation, and partial elements that are more difficult to characterise and might confuse the downstream analyses.

Poor sequence conservation in TEs non-coding regions. The annotation of TEs relies on two different strategies. Given a list of consensus sequences, available in public databases such as RepBase (https://www.girinst.org/repbase/), one strategy is to scan the genome for regions that show some sequence similarity with a TEs consensus. Although this strategy is viable and actually implemented in some major annotation pipelines, it is still flawed by the fact that TEs non-coding regions have poor sequence similarity. Alignments are generally rooted in the coding region, that is instead usually conserved, but the result might not provide a clear definition of element's boundaries (Bourque et al. 2018), and require the researcher to perform an annotation that is based on structural features (Ellinghaus et al 2008).

The project discussed in this manuscript required us to tackle these three sources of error as we performed the reference annotation, the detection of polymorphisms and the quantitation of transcription via RNA-Seq.

I. Annotation of TEs from reference genomes

In Chapter 2, the annotation of TEs from Peach and Almond reference genomes is reported as a double round annotation in which we have first identified all the Class I and Class II TEs by REPET (Flutre et al. 2011) to assess the global TE coverage in the two genomes and then we performed a more refined annotation of LTR-RTs starting from the LTR Harvest pipeline (Ellinghaus et al 2008) output. The difference between the two approaches is that REPET performs the annotation of single TE copies based on the sequence similarity with TEs consensus sequences that are computed by LTR harvest, instead, is designed to identify the structural features that are typical of a LTR-RT insertion, or else the long terminal repeats (LTRs) and the tandem site duplications (TSD) that characterize these elements (Ellinghaus et al 2008). This latter annotation was then filtered for its quality (i.e. excluding those elements covered widely by tandem repetitions and anonymous nucleotides) and for the presence of putative coding regions identified via HMMscan (Luciani et al. 2018).

The difference between the two approaches can be somehow described as a trade-off between the sensitivity and the accuracy of the prediction. REPET allows the user to obtain a genome-wide annotation of Class I and Class II TEs but, to our experience, lack of the capability to accurately define the boundaries of single elements and to distinguish between partial and complete TEs. The refined annotation is more accurate in this sense, but it is, of course, limited to LTR-RTs.

This latter annotation is more reliable for further analyses of insertional variability and transcriptional quantitation. We chose to focus on LTR-RTs for their relatively high presence and activity in plant genomes (Casacuberta and Santiago 2003), and because they include transcription in their transpositional cycle (Eickbush 1992). LTR-RTs transcription level, quantitated from RNA-seq analysis, can be used as an indication of possible transpositional activity. This is why, summarizing, we performed two separate annotations with different strategies.

II. TEs polymorphism detection from WGS data

The technical insights on the detection of transpositional polymorphisms from DNA-seq data constitute the topic of the benchmarking project exposed in Chapter 1, and there is no much need for further discussion on the technical side. Anyways, readers might have noticed an incongruence between the results of the benchmarking and the procedure followed to obtain Peach and Almond population polymorphisms in Chapter 2. The benchmarking project indicates Teflon as the best overall performing package in the detection of both reference and non-reference insertions. A reasonable consequence of this finding would have been to apply this package to the detection of polymorphisms in Peach and Almond cultivated varieties (Table 2.1). Instead, the results are based on PINDEL for the polymorphisms relative to the reference insertions, and on Jitterbug for the non-reference insertions.

In all honesty, this choice is merely consequent to the temporal organization of the project. We designed the protocol to assess the intraspecific variability in Peach and Almond from WGS samples of cultivated species in 2016, whilst the benchmarking project started in the summer of 2018. So, even though a repetition of the insertional variability analysis with a different combination of software would have been probably beneficial, we decided to keep the old results for a matter of time requirements imposed by the almond genome project deadlines. Anyways, however probably not optimized, we consider the presented results as very solid. PINDEL, that was used to assess the reference insertions polymorphisms, has proven a high performance on detection of deletions that are smaller than 20.000 kb (Mu et al. 2015) and potentially linked with the presence or absence of TEs in the sample. PINDEL was not included in our benchmarking because it is not designed to detect polymorphisms that are due to transposition, but is a general-purpose structural variability detection software, and hence poorly comparable to those packages that are designed specifically for transposition. Anyways, since we have applied the result is still acceptable. Our benchmark highlights that Jitterbug has a lower sensitivity respect to other programs, but a high level of sensitivity (Figure 1.4). As long as the two species are compared with the same tool, and with the same underestimation of the actual insertions but with a sufficient number of insertions, the result can be considered a representative subset of the real amount of insertions.

III. The quantitation of LTR-RTs transcription

If the mapping of transpositional polymorphisms from population data provides us with an outlook on the consequences of transpositional activity, the detection of transcription can be an indication of the activity of retrotransposons. For this reason, there is a growing interest in implementing computational methods that are able to analyze transposons' transcription starting from NGS RNA-seq data. This evaluation is made rather complicated by two main factors: the usually low transcriptional level that is associated with transposon activity (Qui and Ungerer 2018) and the repetitiveness of these elements.

TEs are highly epigenetically silenced sequences which are rarely transcribed and hence usually associated with low transcriptional levels (Qui and Ungerer 2018). The detection of poorly transcribed mRNAs always implies a problem of normalisation. Sample contamination, random transcription and assembly defects can generate a background noise in the RNA-seq distribution (Dillies et al. 2012). The presence of this background noise imposes us a question: as we associate a low transcriptional level to a given genomic feature, how can we tell if this is the product of real transcription or background noise? Figure 3.1 from Chapter 3 explains how this problem can affect the quantitation of the poorly transcribed LTR-RTs in Gervasi et al. Peach RNA-seq dataset.

Once we decided for a way to eliminate the background noise, it was the turn to cope with the repetitiveness of the elements. RNA-seq distributions obtained by NGS technologies are constituted by short readings of transcriptomic RNA. If retrotransposition starts from an element which is part of a family of recent amplification, the resulting mRNA will be almost indistinguishable from the one that would be generated by other elements from the same family. This implies that most of the TE-associated reads in an RNA-seq distribution map to different loci, and renders the identification of the expressing copy very difficult. When we were working on it, the available packages for TE transcription quantitation only returned the expression value for family and not for single copy (Jin et al. 2015; Lerat et al. 2016), even though very recent efforts have been made to determine the expression level at single-copy resolution (Valdebenito-Maturana and Riadi 2018). Among the explored packages, the approach of TEtranscripts (Jin et al. 2015) tackles the problems associated with the quantitation of TEs family expression with a rather effective approach. After having quantitated the expression level for a family, the program selects the multi-mapping reads and submits them to an iterative normalisation that tries to reduce the effect of large families (Expectation-Maximization "EM" algorithm). Large families accumulate more multi-mapping reads than small families, and their expression value can be overestimated. The EM algorithm normalises the number of multi mapping reads for the size of the family, returning a relative expression value that is incorporated with the count of single-mapping reads.

The outcome of this program using the data we have scanned (Gervasi et al. 2018) is an expression profile at family level that contains no evident incongruencies with what one would expect from the transcriptional behaviour of these elements. The families found expressing in Redkist and JH Hale Peach varieties (Figure 3.2 and Table 3.2) include several recent insertions that are likely to be potentially active, and even though the result includes some numerous families (e.g. RLC-21), the most numerous families have been found not significantly expressed (e.g. RLC-5).

Even though the quantitation of TEs transcription from RNA-seq probably requires further implementation efforts, TEtranscripts proved to be based on a very solid approach able to tackle the quantitation of repetitive elements in a reasonable and effective way.

Part B. The impact of transposition in Peach and Almond evolution.

The biological connotations of this thesis work are all based on the idea that transposition has a relevant impact on the generation of genetic variability. The creation of new insertions can lead to the increase of intraspecific variability that is particularly evident, in crops, as the new trait that is introduced has some agronomical relevance, as in the case of pink grapes, glabrous peaches (nectarines) or when the genes for cold stress response are rewired in rice (Contreras et al. 2015). At interspecific level, very close species can share the same TEs family, but their dynamics might have been diverged dramatically during the species differentiation. Traces of a high transpositional activity have been found to be concomitant to different speciation events. In 2018, Morata et al. identified the expansion of the pericentromeric region in several Melon (Cucumis melo) chromosomes that occurred after its separation from Cucumber (Cucumis sativus) (Morata et al. 2018). The rate of speciation has been associated to the transpositional activity in Mammals. A recent work that compared the ratio of number of TEs on genome length (density of insertions) with the relative rate of speciation among the 29 Mammalia taxonomical families, found a correlation between the transpositional activity and the tendency to differentiate new species (Ricci et al. 2018). A review published in 2014 by Alexander Belyayev proposes the burst of transposable elements as a major evolutionary driving force (Belyayev 2014).

In this work, we annotated TEs from two very close species and found a comparable composition and spatial distribution of these elements. Differently form Melon and Cucumber, Peach and Almond do not display a macroscopic difference in the accumulation of TEs along the chromosome. Both the global TEs annotation made with REPET and the refined LTR-RTs annotation made starting by the LTR-Harvest structural prediction confirmed that these two very close species shared the same TEs. Relevant differences emerged as we considered two variables associated with the LTR-RTs, or else the insertion time and the intraspecific variability of these elements. This comparison resulted to be very insightful for the reconstruction of many aspects of the evolutionary history of these two species. It’s a case in which the reconstruction of transpositional dynamics allows us to recapitulate the evolutionary history of two species.

I. Climate change, population reduction and transpositional burst at the base of Peach differentiation.

The evolutionary history of Peach (Prunus persica) and Almond (Prunus dulcis) has a narratively intriguing connection with the rise of Himalaya, that has been extensively discussed by Yu et al. on Nature Communications in 2018. The rise of the Tibetan plateau took started about 15 million years ago (MYA), in consequence of the impact between the Indian and the Eurasian plaques, and arrested about 5 MYA (Tapponier et al. 2001). The population of Peach's ancestor present in the region that corresponds to modern Southwestern China underwent insulation and serious population reduction (Velasco et al. 2014). In Figure 2.7 from Chapter 2, we can appreciate how peach and almond differ in terms of polymorphic elements in the time-frame 5-15 MYA, a result that confirms the population reduction already proposed in previous works (Velasco et al 2014; Yu et al 2018).

In this by-the-book case of allopatric speciation, the climate changes associated with the dramatic geological movements imposed by the rise of Himalaya caused a deep change in the local ecology. The main point of Yu et al. article in 2018 is that the presence of more frugivore species favoured the evolution of a fleshy fruit. In all cases, the early times of the Peach ancestor were characterised by the presence of several potential sources of abiotic and biotic stress, along with a reduce population size. All these factors are compatible with the transposon burst we observe starting from 5 MYA in Peach (Figure 2.6 - A). If the increase of transpositional activity can be consequent to climate and ecological change, the rate of fixation is probably due to the small size of the population. LTR-RTs are generally elements that insert the genome randomly and are eventually selected by purifying selection (Baucom et al. 2009). The size of the population has been correlated with the selection efficiency (Lynch and Conrey 2003), and this might explain the higher fixed/polymorphic rate that is observed in the last 5 MY in Peach (Figure 2.6 - A).

II. The impact of transposition on Peach and Almond genetics

Our results indicate that the transpositional dynamics diverged in Peach and Almond after their separation, and this is a potentially important source of genetic variability. As we have reviewed extensively in the introduction, the effects of transposition on gene activity are usually divided into disruptive mutations and flanking insertions that change gene regulation by providing new regulatory elements or changing the epigenetic status of the surrounding region (Contreras et al. 2015). After having highlighted a significant difference between Peach and Almond LTR-RTs dynamics, the very next question is how this difference affected the genetics of these two species. The answer to this question is all but simple. To assess how much transposable elements have impressed the genetic variation that we observe between Peach and Almond, we need to rely on solid knowledge about which genetic networks control the most important phenotypes in these two species. Even though many interesting contributions have been provided in recent years (e.g. Lopez-girona et al 2017; Mora et al. 2017), this theme remains very open.

Further contributions will help to better unravel the genetics of these two crop species, and the involvement of TEs will become more clear with the time.

In the meanwhile, we can affirm that this work indicates transpositional dynamics as a major difference between these two important crops at genomic level. If the extent of the impact of this asymmetric distribution of TEs between Peach and Almond on genetic variability will be clarified by further and following studies, this doctoral thesis allows us to indicate transposition as one of the most important differences between Peach and Almond, and as a potential genomic mechanism that impressed the differentiation between them.

III. Potential impact of this work on crop breeding

Finally, it's worthwhile spending a few lines to highlight the potential applications of the results presented in this work. The annotation of transposable elements and their characterisation provide novel genetic information that can be beneficial to determine the genetics that underlie agronomically important traits. This work outputs the work frame idea that TEs movement has been fundamental in the establishment of the genetic differences between peach and almond, being transpositional dynamics a macroscopic difference between these two very close genomes. If this suggests that several agronomic traits can be a consequence of TEs movement, the annotation and the further dynamics characterization provide the tools to investigate this in detail.

The results from Chapter 2 contributed to the almond genome paper authored by Alioto et al. and included in the supplementary material. In this paper, we report the case of the sweet kernel phenotype in almond. It has been recently shown that this phenotype is due to reduced expression of the genes that encode two cytochrome P450 enzymes catalysing the first steps of the biosynthesis of the amygdalin, a cyanogenic diglucoside (Thodberg et al. 2018). After the comparison of the sequence of one of these almond genes, CYP71AN24, with its homologs in peach, sweet cherry and P. mume, we have found that it is flanked by several almond-specific highly methylated TE insertions (Alioto et al. 2019), that might have a possible impact on the expression level of this gene and on the sweet almond phenotype.

This is just an example of how this work can potentially contribute with further knowledge to the improvement of genetic breeding protocols and trait selection.

Supplementary material

Published article: Transposons played a major role in the diversification between the closely related almond and peach genomes: Results from the almond genome sequence

ABSTRACT

We sequenced the genome of the highly heterozygous almond Prunus dulcis cv. Texas combining short and long-read sequencing. We obtained a genome assembly totaling 227.6 Mb of the estimated 238 Mb almond genome size, of which 91% is anchored to eight pseudomolecules corresponding to its haploid chromosome complement, and annotated 27,969 protein-coding genes and 6,747 non-coding transcripts. By phylogenomic comparison with the genomes of 16 additional close and distant species we estimated that almond and peach (P. persica) diverged around 5.88 Mya. These two genomes are highly syntenic and show a high degree of sequence conservation (20 nucleotide substitutions/kb). However, they also exhibit a high number of presence/absence variants, many attributable to the movement of transposable elements (TEs). TEs have generated an important number of presence/absence variants between almond and peach, and we show that the recent history of TE movement seems markedly different between them. TEs may also be at the origin of important phenotypic differences between both species, and in particular, for the sweet kernel phenotype, a key agronomic and domestication character for almond. Here we show that in sweet almond cultivars, highly methylated TE insertions surround a gene involved in the biosynthesis of amygdalin, whose reduced expression has been correlated with the sweet almond phenotype. Altogether, our results suggest a key role of TEs in the recent history and diversification of almond and its close relative peach.

AUTHORS

Tyler Alioto, Konstantinos G. Alexiou, Amélie Bardil, **Fabio Barteri**, Raúl Castanera, Fernando Cruz, Amit Dhingra, Henri Duval, Ángel Fernández i Martí, Leonor Frias, Beatriz Galán, José L. Garcia, Werner Howad, Jèssica Gómez-Garrido, Marta Gut, Irene Julca, Jordi Morata, Pere Puigdomènech, Paolo Ribeca, María José Rubio Cabetas, Anna Vlasova, Michelle Wirthensohn, Jordi Garcia-Mas, Toni Gabaldón, Josep M. Casacuberta, Pere Arús

JOURNAL : Plant Journal (**https://doi.org/10.1111/tpj.14538**)

Article in Press: LTR-TEs abundance, timing and mobility in *S. commersonii* and S. tuberosum genomes following cold stress conditions

ABSTRACT

From an evolutionary perspective, long-terminal repeat retrotransposons (LTR-RT) activity during stress may be viewed as a mean by which organisms can keep up rates of genetic adaptation to changing conditions. Potato is one of the most important crop consumed worldwide, but studies on LTR-RT characterization are still lacking. Here, we assessed the abundance, insertion time and activity of LTR-RTs in both cultivated Solanum tuberosum and its cold tolerant wild relative S. commersonii genomes. Gypsy elements were more abundant than Copia ones, suggesting that the former was somehow more successful in colonizing potato genomes. However, Copia elements, and in particular the Ale lineage, are younger than Gypsy ones, since their insertion time was in average ~ 2 Mya. Due to the ability of LTR-RTs to be circularized by the host DNA repair mechanisms, we identified via mobilome-seq a Copia/Ale element (called nightshade, informal name used for potato family) active in S. tuberosum genome. Our analyses represent a valuable resource for comparative genomics within the Solanaceae, transposon-tagging and for the design of cultivar-specific molecular markers in potato.

AUTHORS

Salvaore Esposito, **Fabio Barteri**, Josep M. Casacuberta, Marie Mirouze, Domenico Carputo, Riccardo Aversano.

JOURNAL: Planta (**DOI: 10.1007/s00425-019-03283-3**)

Acknowledgments

First of all, I'd like to thank my thesis advisor **Dr Josep Casacuberta** for allowing me to pursue this PhD within his lab and under his careful scientific guidance, for his commitment in lab direction and his patience. Likewise, I thank **Dr Jordi Morata** for thoroughly supervising the implementation of my project and for his fundamental technical, scientific and personal support, and **Dr Carlos Vicient** for his important help during the last part of my internship. Along with them, I'd like to thank all the researchers who coincided with me in the lab, and which support has been fundamental in many occasions, **Amelie**, **Pol**, **Raul**, **Carlos**, **Marc**, **Pedro**, **Roger**, **Beatriz** and **Cristina**.

Most of this work has been carried on in collaboration with other groups and within a shared effort to sequence and annotate the Almond genome. In this, I feel to express my gratitude to **Dr Pere Arús** for leading the project, to **Dr Konstantinos Alexiou** who supervised the DNA sequencing and to the scientists who supervised important parts of this project, such as **Dr Jordi García-Mas** and **Dr Toni Gabaldón**.

The Centre for Research in Agricultural Genomics (CRAG) hosted my whole PhD project, and I'd like to thank both the researchers and the support personnel who strive for the maintenance and the betterment of this center. In particular, I thank **Dr Martí Bernardo**, **Dr Gonzalo Vera** and **Dr Víctor Manuel González** from the bioinformatics support team for the help on the technical side, for the high-quality training and the amazing scientific discussion. Also, I thank the HR for their help during my settlement in Barcelona.

A PhD thesis is a complex and though life experience that requires the student to lean on the people who are willing to share their experience, knowledge and wisdom to help you. Under an academic point of view, the members of the Thesis Panel (Commisió de seguiment) **Dr Amparo Monfort**, **Dr Maria José Aranzana** and **Dr Josep Alluè** have been of great help revising my progress every year and sharing suggestions and I'd like to thank them for their work. But on a more personal point of view, I'd like to express my particular gratitude to **Dr Pere Puigdomènech** for helping me during the hardest time of my thesis, and **Dr David Caparros** for being of constant support.

Special thanks go to **Dr Molly Gale Hammell** and all her group at the Cold Spring Harbor Laboratory (CSHL) for kindly hosting me as visiting student in 2018.

Of course, nothing would have been possible without the support of my friends and family. I'd like to thank my **parents** and my **siblings** for their unconditional love and support, my **friends of a lifetime** for having understood my decision to leave Rome, and my **new friends** here in Barcelona for having made me feel home again. Thanks to **Lello** for hosting me in the USA during the visiting and, most of all, thanks to **Miriam.** For having been the most beautiful thing in this PhD.

Bibliography

Adrion, J. R., Song, M. J., Schrider, D. R., Hahn, M. W., & Schaack, S. (2017). Genome-Wide Estimates of Transposable Element Insertion and Deletion Rates in Drosophila Melanogaster. Genome Biology and Evolution, 9(5), 1329–1340. https://doi.org/10.1093/gbe/evx050

Akagi, T., Henry, I. M., Morimoto, T., & Tao, R. (n.d.). Insights into the Prunus-Specific S-RNase-Based Self-Incompatibility System from a Genome-Wide Analysis of the Evolutionary Radiation of S Locus-Related F-box Genes. https://doi.org/10.1093/pcp/pcw077

Alioto, T., Alexiou, K. G., Bardil, A., Barteri, F., Castanera, R., Cruz, F., ... Arús, P. (2019). 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, tpj.14538. https://doi.org/10.1111/tpj.14538

Alkan, C., Sajjadian, S., & Eichler, E. E. (2011). Limitations of next-generation genome sequence assembly. Nature Methods, 8(1), 61–65. https://doi.org/10.1038/nmeth.1527

Bao, W., Kojima, K. K., & Kohany, O. (2015). Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA, 6(1), 11. https://doi.org/10.1186/s13100-015-0041-9

Bardil, A., Tayalé, A., & Parisod, C. (2015). Evolutionary dynamics of retrotransposons following autopolyploidy in the Buckler Mustard species complex. The Plant Journal, 82(4), 621–631. https://doi.org/10.1111/tpj.12837

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Research, 27(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Berthelier, J., Casse, N., Daccord, N., Jamilloux, V., Saint-Jean, B., & Carrier, G. (2018). A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga Tisochrysis lutea. BMC Genomics, 19(1), 378. https://doi.org/10.1186/s12864-018-4763-1

Bolger, Anthony M., et al. "Trimmomatic: a Flexible Trimmer for Illumina Sequence Data." Bioinformatics, vol. 30, no. 15, 2014, pp. 2114–2120., doi:10.1093/bioinformatics/btu170.

Bortiri, E., Oh, S.-H., Jiang, J., Baggett, S., Granger, A., Weeks, C., ... Parfitt, D. E. (n.d.). Phylogeny and Systematics of Prunus (Rosaceae) as Determined by Sequence Analysis of ITS and the Chloroplast trnL-trnF Spacer DNA. Systematic Botany. American Society of Plant Taxonomists. https://doi.org/10.2307/3093861

Boudon, Sylvain, et al. “Structure and Origin of Xanthomonas Arboricola Pv. Pruni Populations Causing Bacterial Spot of Stone Fruit Trees in Western Europe.” Phytopathology, vol. 95, no. 9, 2005, pp. 1081–1088., doi:10.1094/phyto-95-1081.

Brown, T. A. (2002). The Human Genome. Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK21134/

Bureau, T. E., & Wessler, S. R. (1992). Tourist: a large family of small inverted repeat elements frequently associated with maize genes. The Plant Cell, 4(10), 1283–1294. https://doi.org/10.1105/tpc.4.10.1283

Cao, K., Zheng, Z., Wang, L., Liu, X., Zhu, G., Fang, W., … Wang, J. (2014). Comparative population genomics reveals the domestication history of the peach, Prunus persica, and human influences on perennial fruit crops. Genome Biology, 15(7), 415. https://doi.org/10.1186/S13059-014-0415-1

Capella-Gutierrez, S., Silla-Martinez, J. M., & Gabaldon, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics, 25(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Carpentier, M.-C., Manfroi, E., Wei, F.-J., Wu, H.-P., Lasserre, E., Llauro, C., … Panaud, O. (2019). Retrotranspositional landscape of Asian rice revealed by 3000 genomes. Nature Communications, 10(1), 24. https://doi.org/10.1038/s41467-018-07974-5

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. https://doi.org/10.1016/S0378-1119(03)00557-2

Chanda, Bidisha, et al. “Glycerol-3-Phosphate Is a Critical Mobile Inducer of Systemic Immunity in Plants.” Nature Genetics, vol. 43, no. 5, 2011, pp. 421–427., doi:10.1038/ng.798.

Chen, J., Hu, Q., Zhang, Y., Lu, C., & Kuang, H. (2014). P-MITE: a database for plant miniature inverted-repeat transposable elements. Nucleic Acids Research, 42(D1), D1176–D1181. https://doi.org/10.1093/nar/gkt1000

Cicconi, A., Micheli, E., Vernì, F., Jackson, A., Gradilla, A. C., Cipressa, F., … Raffa, G. D. (2017). The Drosophila telomere-capping protein Verrocchio binds single-stranded DNA and protects telomeres from DNA damage response. Nucleic Acids Research, 45(6), 3068–3085. https://doi.org/10.1093/nar/gkw1244

Cirulli, E. T., & Goldstein, D. B. (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nature Reviews Genetics, 11(6), 415–425. https://doi.org/10.1038/nrg2779

Curcio, M. J., & Derbyshire, K. M. (2003). The outs and ins of transposition: from Mu to Kangaroo. Nature Reviews Molecular Cell Biology, 4(11), 865–877. https://doi.org/10.1038/nrm1241

Das, B., Ahmed, N., & Singh, P. (2011). Prunus diversity-early and present development: A review. International Journal of Biodiversity and Conservation, 3(14), 721–734. https://doi.org/10.5897/IJBCX11.003

DeFraia, C., & Slotkin, R. K. (2014). Analysis of Retrotransposon Activity in Plants. In Methods in molecular biology (Clifton, N.J.) (Vol. 1112, pp. 195–210). https://doi.org/10.1007/978-1-62703-773-0_13

Dirlewanger, E., Cosson, P., Howad, W., Capdeville, G., Bosselut, N., Claverie, M., … Esmenjaud, D. (2004). Microsatellite genetic linkage maps of myrobalan plum and an almond-peach hybrid?location of root-knot nematode resistance genes. Theoretical and Applied Genetics, 109(4), 827–838. https://doi.org/10.1007/s00122-004-1694-9

Disdero, E., & Filée, J. (2017). LoRTE: Detecting transposon-induced genomic variants using low coverage PacBio long read sequences. Mobile DNA, 8(1), 5. https://doi.org/10.1186/s13100-017-0088-x

Dobin, Alexander, and Thomas R. Gingeras. “Mapping RNA-Seq Reads with STAR.” Current Protocols in Bioinformatics, 2015, doi:10.1002/0471250953.bi1114s51.

Dubey, Namo, and Kunal Singh. “Role of NBS-LRR Proteins in Plant Defense.” Molecular Aspects of Plant-Pathogen Interaction, 2018, pp. 115–138., doi:10.1007/978-981-10-7371-7_5.

Eckardt, N. A. (2000). Sequencing the rice genome. The Plant Cell, 12(11), 2011–2017. https://doi.org/10.1105/TPC.12.11.2011

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics, 14(9), 755–763. https://doi.org/10.1093/bioinformatics/14.9.755

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics, 9(1), 18. https://doi.org/10.1186/1471-2105-9-18

Eriksson, T., Hibbs, M. S., Yoder, A. D., Delwiche, C. F., & Donoghue, M. J. (2003). The Phylogeny of Rosoideae (Rosaceae) Based on Sequences of the Internal Transcribed Spacers (ITS) of Nuclear Ribosomal DNA and the trnL/F Region of Chloroplast DNA. International Journal of Plant Sciences, 164(2), 197–211. https://doi.org/10.1086/346163

Estep, M. C., DeBarry, J. D., & Bennetzen, J. L. (2013). The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity, 110(2), 194–204. https://doi.org/10.1038/hdy.2012.99

Ewing, A. D. (2015). Transposable element detection from whole genome sequence data. Mobile DNA, 6(1), 24. https://doi.org/10.1186/s13100-015-0055-3

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

Figueiredo, Joana, et al. “Subtilisin-like Proteases in Plant Defence: the Past, the Present and Beyond.” Molecular Plant Pathology, vol. 19, no. 4, 2017, pp. 1017–1028., doi:10.1111/mpp.12567.

Finnegan DJ, 1992. Transposable elements. In: The genome of Drosophila melanogaster (Lindsley DL and Zimm GG, eds). San Diego: Academic Press; 1096–1107.

Fiston-Lavier, A.-S., Barrón, M. G., Petrov, D. A., & González, J. (2015). T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. Nucleic Acids Research, 43(4), e22. https://doi.org/10.1093/nar/gku1250

Flutre, T., Duprat, E., Feuillet, C., & Quesneville, H. (2011). Considering Transposable Element Diversification in De Novo Annotation Approaches. PLoS ONE, 6(1), e16526. https://doi.org/10.1371/journal.pone.0016526

Gajewski, W. (1959). Evolution in the Genus Geum. Evolution, 13(3), 378. https://doi.org/10.2307/2406114

Gardner, E. J., Lam, V. K., Harris, D. N., Chuang, N. T., Scott, E. C., Pittard, W. S., … Devine, S. E. (2017). The Mobile Element Locator Tool (MELT): population-scale mobile element discovery and biology. Genome Research, 27(11), 1916. https://doi.org/10.1101/GR.218032.116

George, A. P., & Nissen, R. J. (1992). Effects of water stress, nitrogen and paclobutrazol on flowering, yield and fruit quality of the low-chill peach cultivar, ‘Flordaprince.’ Scientia Horticulturae, 49(3–4), 197–209. https://doi.org/10.1016/0304-4238(92)90157-8

Gervasi, Fabio, et al. “Transcriptome Reprogramming of Resistant and Susceptible Peach Genotypes during Xanthomonas Arboricola Pv. Pruni Early Leaf Infection.” Plos One, vol. 13, no. 4, 2018, doi: 10.1371/journal.pone.0196590.

Goerner-Potvin, P., & Bourque, G. (2018). Computational tools to unmask transposable elements. Nature Reviews Genetics, 19(11), 688–704. https://doi.org/10.1038/s41576-018-0050-x

Goodier, J. L., & Kazazian, H. H. (2008). Retrotransposons Revisited: The Restraint and Rehabilitation of Parasites. Cell, 135(1), 23–35. https://doi.org/10.1016/j.cell.2008.09.022

Goodwin, S., McPherson, J. D., & McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics, 17(6), 333–351. https://doi.org/10.1038/nrg.2016.49

Gradziel, T. M. (2003). INTERSPECIFIC HYBRIDIZATIONS AND SUBSEQUENT GENE INTROGRESSION WITHIN PRUNUS SUBGENUS AMYGDALUS. Acta Horticulturae, (622), 249–255. https://doi.org/10.17660/ActaHortic.2003.622.22

Gradziel, T. M. (2009). Almond (Prunus dulcis) Breeding. In Breeding Plantation Tree Crops: Temperate Species (pp. 1–31). New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-71203-1_1

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 10(3), 645–656. https://doi.org/10.1109/TCBB.2013.68

Han, Y., & Wessler, S. R. (2010). MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Research, 38(22), e199–e199. https://doi.org/10.1093/nar/gkq862

Handsaker, R. E., Korn, J. M., Nemesh, J., & McCarroll, S. A. (2011). Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. Nature Genetics, 43(3), 269–276. https://doi.org/10.1038/ng.768

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

Hoen, D. R., & Bureau, T. E. (2012). Transposable Element Exaptation in Plants (pp. 219–251). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-31842-9_12

Hoen, D. R., Hickey, G., Bourque, G., Casacuberta, J., Cordaux, R., Feschotte, C., … Blanchette, M. (2015). A call for benchmarking transposable element annotation methods. Mobile DNA, 6(1), 13. https://doi.org/10.1186/s13100-015-0044-6

Horváth, Vivien, et al. “Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response.” Trends in Genetics, vol. 33, no. 11, 2017, pp. 832–841., doi:10.1016/j.tig.2017.08.007.

Hufford, L. (1992). Rosidae and Their Relationships to Other Nonmagnoliid Dicotyledons: A Phylogenetic Analysis Using Morphological and Chemical Data. Annals of the Missouri Botanical Garden, 79(2), 218. https://doi.org/10.2307/2399767

Hénaff, E., Zapata, L., Casacuberta, J. M., & Ossowski, S. (2015). Jitterbug: somatic and germline transposon insertion detection at single-nucleotide resolution. BMC Genomics, 16(1), 768. https://doi.org/10.1186/s12864-015-1975-5

Jackson, S. A. (2016). Rice: The First Crop Genome. Rice, 9(1), 14. https://doi.org/10.1186/s12284-016-0087-4

Jiang, C., Chen, C., Huang, Z., Liu, R., & Verdier, J. (2015). ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. BMC Bioinformatics, 16(1), 72. https://doi.org/10.1186/s12859-015-0507-2

Jin, Ying, et al. "TEtranscripts: a Package for Including Transposable Elements in Differential Expression Analysis of RNA-Seq Datasets." Bioinformatics, vol. 31, no. 22, 2015, pp. 3593–3599., doi:10.1093/bioinformatics/btv422.

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. https://doi.org/10.1007/s001220000511

KALKMAN, & C. (1965). The Old Worlds species of Prunus subgen. Laurocerasus including those formerly referred to Pygeum. Blumea, 13, 1–115. Retrieved from https://ci.nii.ac.jp/naid/20001207257/

Kader, S.A; Proebsting, E. L. (1992). Journal of the American Society for Horticultural Science. American Society for Horticultural Science (USA). [American Society for Horticultural Science]. Retrieved from http://agris.fao.org/agris-search/search.do?recordID=US9602406

Kaessmann, H. (2010). Origins, evolution, and phenotypic impact of new genes. Genome Research, 20(10), 1313–1326. https://doi.org/10.1101/gr.101386.109

Kang, Hong-Gu, et al. "CRT1 Is a Nuclear-Translocated MORC Endonuclease That Participates in Multiple Levels of Plant Immunity." Nature Communications, vol. 3, no. 1, 2012, doi:10.1038/ncomms2279.

Katoh, K., & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Molecular Biology and Evolution, 30(4), 772–780. https://doi.org/10.1093/molbev/mst010

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, 6(1), 4. https://doi.org/10.1186/1939-8433-6-4

Keane, T. M., Wong, K., & Adams, D. J. (2013). RetroSeq: transposable element discovery from next-generation sequencing data. Bioinformatics, 29(3), 389–390. https://doi.org/10.1093/bioinformatics/bts697

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution, 16(2), 111–120. https://doi.org/10.1007/BF01731581

Klein, S. J., & O'Neill, R. J. (2018). Transposable elements: genome innovation, chromosome diversity, and centromere conflict. Chromosome Research, 26(1–2), 5–23. https://doi.org/10.1007/s10577-017-9569-5

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. https://doi.org/10.1093/molbev/msw137

Ladizinsky, G. (1999). On the Origin of Almond. Genetic Resources and Crop Evolution, 46(2), 143–147. https://doi.org/10.1023/A:1008690409554

Lamichhane, Jay Ram. "Xanthomonas ArboricolaDiseases of Stone Fruit, Almond, and Walnut Trees: Progress Toward Understanding and Management." Plant Disease, vol. 98, no. 12, 2014, pp. 1600–1610., doi:10.1094/pdis-08-14-0831-fe.

Le, Q. H., Wright, S., Yu, Z., & Bureau, T. (2000). Transposon diversity in Arabidopsis thaliana. Proceedings of the National Academy of Sciences of the United States of America, 97(13), 7376–7381. https://doi.org/10.1073/pnas.97.13.7376

Lee, S.-I., & Kim, N.-S. (2014). Transposable elements and genome size variations in plants. Genomics & Informatics, 12(3), 87–97. https://doi.org/10.5808/GI.2014.12.3.87

Leinonen, R., Sugawara, H., Shumway, M., & International Nucleotide Sequence Database Collaboration. (2011). The sequence read archive. Nucleic Acids Research, 39(Database issue), D19-21. https://doi.org/10.1093/nar/gkq1019

Lerat, Emmanuelle, et al. "TEtools Facilitates Big Data Expression Analysis of Transposable Elements and Reveals an Antagonism between Their Activity and That of PiRNA Genes." Nucleic Acids Research, 2016, doi:10.1093/nar/gkw953.

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

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

Lin, Yanzhu, et al. “Comparison of Normalization and Differential Expression Analyses Using RNA-Seq Data from 726 Individual Drosophila Melanogaster.” BMC Genomics, vol. 17, no. 1, 2016, doi:10.1186/s12864-015-2353-z.

Linheiro, R. S., & Bergman, C. M. (2012). Whole genome resequencing reveals natural target site preferences of transposable elements in Drosophila melanogaster. PloS One, 7(2), e30008. https://doi.org/10.1371/journal.pone.0030008

Lyu, H., He, Z., Wu, C.-I., & Shi, S. (2018). Convergent adaptive evolution in marginal environments: unloading transposable elements as a common strategy among mangrove genomes. New Phytologist, 217(1), 428–438. https://doi.org/10.1111/nph.14784

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J., & Springer, N. M. (2015). Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. PLoS Genetics, 11(1), e1004915. https://doi.org/10.1371/journal.pgen.1004915

Mandal, Mihir K., et al. “Glycerol-3-Phosphate and Systemic Immunity.” Plant Signaling &amp; Behavior, vol. 6, no. 11, 2011, pp. 1871–1874., doi:10.4161/psb.6.11.17901.

Mardis, E. R., & Wilson, R. K. (2009). Cancer genome sequencing: a review. Human Molecular Genetics, 18(R2), R163–R168. https://doi.org/10.1093/hmg/ddp396

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. https://doi.org/10.1073/pnas.36.6.344

Mcdermaid, Adam, et al. “ViDGER: An R Package for Integrative Interpretation of Differential Gene Expression Results of RNA-Seq Data.” 2018, doi:10.1101/268896.

Miele, V., Penel, S., & Duret, L. (2011). Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics, 12(1), 116. https://doi.org/10.1186/1471-2105-12-116

Miller, A. J., & Gross, B. L. (2011). From forest to field: Perennial fruit crop domestication. American Journal of Botany, 98(9), 1389–1414. https://doi.org/10.3732/ajb.1000522

Mills, R. E., Bennett, E. A., Iskow, R. C., & Devine, S. E. (2007). Which transposable elements are active in the human genome? Trends in Genetics, 23(4), 183–191. https://doi.org/10.1016/J.TIG.2007.02.006

Molnar, P., Tapponnier, P., KIDD, W. S. F., & YIN, A. (1975). Cenozoic Tectonics of Asia: Effects of a Continental Collision: Features of recent continental tectonics in Asia can be interpreted as results of the

India-Eurasia collision. Science (New York, N.Y.), 189(4201), 419–426. https://doi.org/10.1126/science.189.4201.419

Morata, J., Marín, F., Payet, J., & Casacuberta, J. M. (2018). Plant Lineage-Specific Amplification of Transcription Factor Binding Motifs by Miniature Inverted-Repeat Transposable Elements (MITEs). Genome Biology and Evolution, 10(5), 1210–1220. https://doi.org/10.1093/gbe/evy073

Morgan, D. R., Soltis, D. E., & Robertson, K. R. (1994). Systematic and Evolutionary Implications of rbcL Sequence Variation in Rosaceae. American Journal of Botany, 81(7), 890. https://doi.org/10.2307/2445770

Naito, K. (n.d.). mPing: The bursting transposon. https://doi.org/10.1270/jsbbs.64.109

Ohta, S., Katsuki, T., Tanaka, T., Hayashi, T., Sato, Y., & Yamamoto, T. (2005). Genetic Variation in Flowering Cherries (Prunus subgenus Cerasus) Characterized by SSR Markers. Breeding Science, 55(4), 415–424. Retrieved from https://ci.nii.ac.jp/naid/10016876452/

Oliver, Keith R., and Wayne K. Greene. "Transposable Elements: Powerful Facilitators of Evolution." BioEssays, vol. 31, no. 7, 2009, pp. 703–714., doi:10.1002/bies.200800219.

Orgel, L. E., & Crick, F. H. C. (1980). Selfish DNA: the ultimate parasite. Nature, 284(5757), 604–607. https://doi.org/10.1038/284604a0

Pascal, T., Kervella, J., Pfeiffer, F. G., Sauge, M. H., & Esmenjaud, D. (1998). EVALUATION OF THE INTERSPECIFIC PROGENY PRUNUS PERSICA CV SUMMERGRAND X PRUNUS DAVIDIANA FOR DISEASE RESISTANCE AND SOME AGRONOMIC FEATURES. Acta Horticulturae, (465), 185–192. https://doi.org/10.17660/ActaHortic.1998.465.21

Peona, V., Weissensteiner, M. H., & Suh, A. (2018). How complete are "complete" genome assemblies?-An avian perspective. Molecular Ecology Resources, 18(6), 1188–1195. https://doi.org/10.1111/1755-0998.12933

Platt, R. N., Blanco-Berdugo, L., & Ray, D. A. (2016). Accurate Transposable Element Annotation Is Vital When Analyzing New Genome Assemblies. Genome Biology and Evolution, 8(2), 403–410. https://doi.org/10.1093/gbe/evw009

Platzer, A., Nizhynska, V., & Long, Q. (2012). TE-Locate: A Tool to Locate and Group Transposable Element Occurrences Using Paired-End Next-Generation Sequencing Data. Biology, 1(2), 395–410. https://doi.org/10.3390/biology1020395

Potter, D., Eriksson, T., Evans, R. C., Oh, S., Smedmark, J. E. E., Morgan, D. R., ... Campbell, C. S. (2007). Phylogeny and classification of Rosaceae. Plant Systematics and Evolution, 266(1–2), 5–43. https://doi.org/10.1007/s00606-007-0539-9

Potter, S. C., Luciani, A., Eddy, S. R., Park, Y., Lopez, R., & Finn, R. D. (2018). HMMER web server: 2018 update. Nucleic Acids Research, 46(W1), W200–W204. https://doi.org/10.1093/nar/gky448

Powers, D. M. W., & Ailab. (2011). EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS &amp; CORRELATION, 2(1), 37–63. Retrieved from http://www.bioinfo.in/contents.php?id=51

Pritham, E. J., Putliwala, T., & Feschotte, C. (2007). Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene, 390(1–2), 3–17. https://doi.org/10.1016/J.GENE.2006.08.008

Pundir, Sangya, et al. “UniProt Tools.” Current Protocols in Bioinformatics, 2016, doi: 10.1002/0471250953.bi0129s53.

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

Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., & Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. Nature Genetics, 23(4), 437–441. https://doi.org/10.1038/70550

Rey, O., Danchin, E., Mirouze, M., Loot, C., & Blanchet, S. (2016). Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. Trends in Ecology & Evolution, 31(7), 514–526. https://doi.org/10.1016/J.TREE.2016.03.013

Rice, P. A., & Baker, T. A. (2001). Comparative architecture of transposase and integrase complexes. Nature Structural Biology, 8(4), 302–307. https://doi.org/10.1038/86166

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: the European Molecular Biology Open Software Suite. Trends in Genetics: TIG, 16(6), 276–277. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10827456

Rishishwar, L., Mariño-Ramírez, L., & Jordan, I. K. (2016). Benchmarking computational tools for polymorphic transposable element detection. Briefings in Bioinformatics, 18(6), bbw072. https://doi.org/10.1093/bib/bbw072

Robb, S. M. C., Lu, L., Valencia, E., Burnette, J. M., Okumoto, Y., Wessler, S. R., & Stajich, J. E. (2013). The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3 (Bethesda, Md.), 3(6), 949–957. https://doi.org/10.1534/g3.112.005348

Rollins, R. C. (n.d.). The End of a Generation of Harvard Botanists. Taxon. Wiley. https://doi.org/10.2307/1216929

Sabot, François, et al. "Transpositional Landscape of the Rice Genome Revealed by Paired-End Mapping of High-Throughput Re-Sequencing Data." The Plant Journal, vol. 66, no. 2, 2011, pp. 241–246., doi:10.1111/j.1365-313x.2011.04492.x.

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

Sanger, F., & Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology, 94(3), 441–448. https://doi.org/10.1016/0022-2836(75)90213-2

Schadt, E. E., Turner, S., & Kasarskis, A. (2010). A window into third-generation sequencing. Human Molecular Genetics, 19(R2), R227–R240. https://doi.org/10.1093/hmg/ddq416

Schuster, S. C. (2008). Next-generation sequencing transforms today's biology. Nature Methods, 5(1), 16–18. https://doi.org/10.1038/nmeth1156

Seberg, Ole, and Gitte Petersen. "A Unified Classification System for Eukaryotic Transposable Elements Should Reflect Their Phylogeny." Nature Reviews Genetics, vol. 10, no. 4, 2009, pp. 276–276., doi: 10.1038/nrg2165-c3.

Seki, S., Takata, A., Nakamura, T., Akiyama, K., & Watanabe, S. (1993). A possible cause of heterogeneity of mammalian apurinic/apyrimidinic endonuclease. The International Journal of Biochemistry, 25(1), 53–59. https://doi.org/10.1016/0020-711x(93)90489-2

Shi, S., Li, J., Sun, J., Yu, J., & Zhou, S. (2013). Phylogeny and Classification of Prunus sensu lato (Rosaceae). Journal of Integrative Plant Biology, 55(11), 1069–1079. https://doi.org/10.1111/jipb.12095

Shimada, T., Hayama, H., Nishimura, K., Yamaguchi, M., & Yoshida, M. (2001). The genetic diversities of 4 species of subg. Lithocerasus (Prunus, Rosaceae) revealed by RAPD analysis. Euphytica, 117(1), 85–90. https://doi.org/10.1023/A:1004193327542

Siva, N. (2008). 1000 Genomes project. Nature Biotechnology, 26(3), 256–256. https://doi.org/10.1038/nbt0308-256b

Song, X., & Cao, X. (2017). Transposon-mediated epigenetic regulation contributes to phenotypic diversity and environmental adaptation in rice. Current Opinion in Plant Biology, 36, 111–118. https://doi.org/10.1016/J.PBI.2017.02.004

Soundararajan, P., Won, S. Y., & Kim, J. S. (2019). Insight on Rosaceae Family with Genome Sequencing and Functional Genomics Perspective. BioMed Research International, 2019, 1–12. https://doi.org/10.1155/2019/7519687

Steele, S. J., Levin, H. L., Craigie, R., & Sandmeyer, S. B. (1998). A map of interactions between the proteins of a retrotransposon. Journal of Virology, 72(11), 9318–9322. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9765482

Stewart, C., Kural, D., Strömberg, M. P., Walker, J. A., Konkel, M. K., Stütz, A. M., … 1000 Genomes Project, 1000 Genomes. (2011). A comprehensive map of mobile element insertion polymorphisms in humans. PLoS Genetics, 7(8), e1002236. https://doi.org/10.1371/journal.pgen.1002236

Tao, R., Watari, A., Hanada, T., Habu, T., Yaegaki, H., Yamaguchi, M., & Yamane, H. (2006). Self-compatible peach (Prunus persica) has mutant versions of the S haplotypes found in self-incompatible Prunus species. Plant Molecular Biology, 63(1), 109–123. https://doi.org/10.1007/s11103-006-9076-0

Temin, H. M., & Baltimore, D. (1972). RNA-Directed DNA Synthesis and RNA Tumor Viruses. Advances in Virus Research, 17, 129–186. https://doi.org/10.1016/S0065-3527(08)60749-6

This, P., Lacombe, T., Cadle-Davidson, M., & Owens, C. L. (2007). Wine grape (Vitis vinifera L.) color associates with allelic variation in the domestication gene VvmybA1. Theoretical and Applied Genetics, 114(4), 723–730. https://doi.org/10.1007/s00122-006-0472-2

Thodberg, S., Cueto, J. Del, Mazzeo, R., Pavan, S., Lotti, C., Dicenta, F., Jakobsen Neilson, E.H., Møller, B.L. and Sánchez-Pérez, R. (2018) Elucidation of the Amygdalin Pathway Reveals the Metabolic Basis of Bitter and Sweet Almonds (Prunus dulcis). Plant Physiol., 178, 1096–1111.

Tijskens, L. M. M., Zerbini, P. E., Schouten, R. E., Vanoli, M., Jacob, S., Grassi, M., … Torricelli, A. (2007). Assessing harvest maturity in nectarines. Postharvest Biology and Technology, 45(2), 204–213. https://doi.org/10.1016/J.POSTHARVBIO.2007.01.014

Treangen, T. J., & Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics, 13(1), 36–46. https://doi.org/10.1038/nrg3117

Valdebenito-Maturana, Braulio, and Gonzalo Riadi. “TEcandidates: Prediction of Genomic Origin of Expressed Transposable Elements Using RNA-Seq Data.” Bioinformatics, vol. 34, no. 22, 2018, pp. 3915–3916., doi:10.1093/bioinformatics/bty423.

Velasco, D., Hough, J., Aradhya, M., & Ross-Ibarra, J. (2016). Evolutionary Genomics of Peach and Almond Domestication. G3 (Bethesda, Md.), 6(12), 3985–3993. https://doi.org/10.1534/g3.116.032672

Velasco, R., Zharkikh, A., Affourtit, J., Dhingra, A., Cestaro, A., Kalyanaraman, A., … Viola, R. (2010). The genome of the domesticated apple (Malus × domestica Borkh.). Nature Genetics, 42(10), 833–839. https://doi.org/10.1038/ng.654

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. https://doi.org/10.1371/journal.pone.0090574

Verde, I., Abbott, A. G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., … 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. https://doi.org/10.1038/ng.2586

Verde, Ignazio, et al. “The High-Quality Draft Genome of Peach (Prunus Persica) Identifies Unique Patterns of Genetic Diversity, Domestication and Genome Evolution.” Nature Genetics, vol. 45, no. 5, 2013, pp. 487–494., doi:10.1038/ng.2586.

Vitte, C., et al. “The Bright Side of Transposons in Crop Evolution.” Briefings in Functional Genomics, vol. 13, no. 4, 2014, pp. 276–295., doi:10.1093/bfgp/elu002.

WAGER, L. R. (1933). The Rise of the Himalaya. Nature, 132(3322), 28–28. https://doi.org/10.1038/132028a0

Wagstaff, B. J., Hedges, D. J., Derbes, R. S., Campos Sanchez, R., Chiaromonte, F., Makova, K. D., & Roy-Engel, A. M. (2012). Rescuing Alu: Recovery of New Inserts Shows LINE-1 Preserves Alu Activity through A-Tail Expansion. PLoS Genetics, 8(8), e1002842. https://doi.org/10.1371/journal.pgen.1002842

Wang, Xi, et al. “Transposon Variants and Their Effects on Gene Expression in Arabidopsis.” PLoS Genetics, vol. 9, no. 2, 2013, doi:10.1371/journal.pgen.1003255.

Wessler, S. R., Bureau, T. E., & White, S. E. (1995). LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Current Opinion in Genetics & Development, 5(6), 814–821. https://doi.org/10.1016/0959-437x(95)80016-x

Witte, C. P., Le, Q. H., Bureau, T., & Kumar, A. (2001). Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. Proceedings of the National Academy of Sciences of the United States of America, 98(24), 13778–13783. https://doi.org/10.1073/pnas.241341898

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

Yang, Y., Fang, X., Galy, A., Jin, Z., Wu, F., Yang, R., … Gao, S. (2016). Plateau uplift forcing climate change around 8.6 Ma on the northeastern Tibetan Plateau: Evidence from an integrated sedimentary Sr record. Palaeogeography, Palaeoclimatology, Palaeoecology, 461, 418–431. https://doi.org/10.1016/J.PALAEO.2016.09.002

Yasuda, K., Ito, M., Sugita, T., Tsukiyama, T., Saito, H., Naito, K., … Okumoto, Y. (2013). Utilization of transposable element mPing as a novel genetic tool for modification of the stress response in rice. Molecular Breeding, 32(3), 505–516. https://doi.org/10.1007/s11032-013-9885-1

Yazbek, M. M., & Al-Zein, M. S. (2014). Wild almonds gone wild: revisiting Darwin’s statement on the origin of peaches. Genetic Resources and Crop Evolution, 61(7), 1319–1328. https://doi.org/10.1007/s10722-014-0113-6

Yazbek, M., & Oh, S.-H. (2013). Peaches and almonds: phylogeny of Prunus subg. Amygdalus (Rosaceae) based on DNA sequences and morphology. Plant Systematics and Evolution, 299(8), 1403–1418. https://doi.org/10.1007/s00606-013-0802-1

Ye, J., McGinnis, S., & Madden, T. L. (2006). BLAST: improvements for better sequence analysis. Nucleic Acids Research, 34(Web Server), W6–W9. https://doi.org/10.1093/nar/gkl164

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25(21), 2865–2871. https://doi.org/10.1093/bioinformatics/btp394

Yu, Y., Fu, J., Xu, Y., Zhang, J., Ren, F., Zhao, H., … Xie, H. (2018). Genome re-sequencing reveals the evolutionary history of peach fruit edibility. Nature Communications, 9(1), 5404. https://doi.org/10.1038/s41467-018-07744-3

Zhang, J., Chen, L.-L., Sun, S., Kudrna, D., Copetti, D., Li, W., ... Zhang, Q. (2016). Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. Scientific Data, 3, 160076. https://doi.org/10.1038/sdata.2016.76

Zhang, S.-D., Jin, J.-J., Chen, S.-Y., Chase, M. W., Soltis, D. E., Li, H.-T., ... Yi, T.-S. (2017). Diversification of Rosaceae since the Late Cretaceous based on plastid phylogenomics. New Phytologist, 214, 1355–1367. https://doi.org/10.1111/nph.14461

Zhang, Xue, et al. “Predicting Essential Genes and Proteins Based on Machine Learning and Network Topological Features: A Comprehensive Review.” Frontiers in Physiology, vol. 7, 2016, doi:10.3389/fphys.2016.00075

Zheng, Y., Crawford, G. W., & Chen, X. (2014). Archaeological Evidence for Peach (Prunus persica) Cultivation and Domestication in China. PLoS ONE, 9(9), e106595. https://doi.org/10.1371/journal.pone.0106595

Zhisheng, A., Kutzbach, J. E., Prell, W. L., & Porter, S. C. (2001). Evolution of Asian monsoons and phased uplift of the Himalaya–Tibetan plateau since Late Miocene times. Nature, 411(6833), 62–66. https://doi.org/10.1038/35075035

Zhuang, J., Wang, J., Theurkauf, W., & Weng, Z. (2014). TEMP: a computational method for analyzing transposable element polymorphism in populations. Nucleic Acids Research, 42(11), 6826. https://doi.org/10.1093/NAR/GKU323

van Dijk, E. L., Jaszczyszyn, Y., Naquin, D., & Thermes, C. (2018). The Third Revolution in Sequencing Technology. Trends in Genetics, 34(9), 666–681. https://doi.org/10.1016/J.TIG.2018.05.008