A NEW AND ROBUST STRATEGY FOR CHARACTERIZING TRANSPOSABLE ELEMENT ABUNDANCES AND DYNAMICS, DEMONSTRATED IN TWO PLANT FAMILIES, THE BRASSICACEAE AND ROSACEAE

by

HONGYE ZHOU

(Under the Direction of Jeffrey Bennetzen)

ABSTRACT

Flowering plants (angiosperms) are the most diverse group of land plants, with >260,000 current species classified into 64 orders and 416 families. Angiosperm families Brassicaceae and Rosaceae have been extensively studied and include numerous crops and the model plant Arabidopsis. Genome sizes differ by >2000-fold in angiosperms, while gene numbers and genic colinearity are more highly conserved. Transposable elements (TEs) make up 15% to 60% of the nuclear DNA in most sequenced angiosperm genomes but can comprise >85% of the DNA in diploid genomes that are >2.5 Gb in size. TE contents are dynamic in genomes because of their amplification and their removal by such mechanisms as unequal homologous recombination and illegitimate recombination. Many TEs, including Helitrons and LTR-retrotransposons, can acquire gene fragments, and sometimes entire genes, which adds to the plasticity for gene creation. Horizontal transfers of TEs are rare but can even occur between distantly related species.

Comparative inter-genome analysis of TEs will shed light on TE dynamics during evolution, including its influence on genome size. Unfortunately, most current TE research is limited to intra-species analyses. Rare inter-species comparisons suffer from heterogenous

methods and standards for TE annotation in each published genome. My study quantifies TE abundance to the superfamily level using raw DNA sequencing reads in 17 Brassicaceae species and 12 Rosaceae species. These results indicated that raw read analysis provides a more accurate estimation of TE content. Moreover, the patterns of TE accumulation indicated that a great number of different TE superfamilies can become predominant in specific genomes, and that the patterns of their amplification show a little phylogenetic signal. Hence, massive TE amplifications that influence genome size appear to be rare but random, at least at this level of analysis.

INDEX WORDS: Angiosperms, Genome annotation, Genome composition, LTR-retrotransposons, Shotgun sequence analysis, Transposable element amplification

A NEW AND ROBUST STRATEGY FOR CHARACTERIZING TRANSPOSABLE ELEMENT ABUNDANCES AND DYNAMICS, DEMONSTRATED IN TWO PLANT FAMILES, THE BRASSICACEAE AND ROSACEAE

by

HONGYE ZHOU

B. S., Wuhan University of Technology, China, 2013

B.A., Huazhong University of Science and Technology, China, 2013

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2020

A NEW AND ROBUST STRATEGY FOR CHARACTERIZING TRANSPOSABLE ELEMENT ABUNDANCES AND DYNAMICS, DEMONSTRATED IN TWO PLANT FAMILIES, THE BRASSICACEAE AND ROSACEAE

By

HONGYE ZHOU

Major Professor:	Jeffrey Bennetzen
Committee:	Jessica Kissinger
	Katrien Devos
	Liang Liu

Electronic Version Approved:

Ron Walcott

Interim Dean of the Graduate School

The University of Georgia

May 2020

To the memory of my mother, Zengying Gao

ACKNOWLEDGEMENTS

I would like first to thank my supervisor, Dr. Jeff Bennetzen, for the last 5 years of taking me as his student and nurturing me with great patience. The most I learnt from Dr. Bennetzen is his passion towards science and the attitude he has for his career which will have a life-time influence on me. I appreciate my great experience spending my mid to late twenties in the Bennetzen lab where I grew stronger, more mature and professional. I still remember Jeff told me that "credibility is the thing that, once you lose it, it is hard to have it back", which always alarms me when I am faced with my work.

During the five years, all Bennetzen lab members gave me a lot of support and advice on my research and presentation skills either in lab meetings or private talks. Especially, I want to thank Dr. Minkyu Park, Dr. Hao Wang, Dr. Xuewen Wang and Ms. Aye Htun.

All my committee members, Dr. Katrien Devos, Dr. Jessica Kissinger and Dr. Liang Liu, thank you for either having given me advice on my research directly or having spent time browsing my dissertation.

I would like to thank GACRC staff Shan-Ho Tsai. Ms. Tsai helped me a lot when I had problems on GACRC clusters, even on weekends or late nights.

I also owe thanks to Dr. Elizabeth Trippe and Mr. Debkanta Chakraborty for giving me a lot of advice on my comprehensive exam presentation.

Finally, I would like to end the acknowledgements by quoting the theme of my favorite manga, JoJo's bizarre adventure, that is, "human praise is the praise of bravery".

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction and Literature Review

Angiosperms

Angiosperms (flowering plants) are estimated to have >260,000 species in ~13,000 known genera. Despite their relatively recent radiation starting in the early Cretaceous, angiosperms have come to dominate most terrestrial plant communities. Eudicots comprise nearly 75% of extant angiosperm species. Core eudicots include 97% of eudicot species diversity, and the rosid clade is one of the three distinct groups within the core eudicots. One group of rosids (eurosids) is divided into two groups, Fabids and Malvids, which Rosaceae and Brassicaceae belong to, respectively.

Lineages within the Brassicaceae have remarkable speciation rates (such as in the genera *Arabis* (1) and *Draba* (2)), the highest among reported plant groups. The number of Brassicaceae species has increased greatly over the past 30 Myr.

The Rosaceae are characterized by their remarkable diversity in morphological traits. They mostly reside in temperate regions of the Northern hemisphere and have adapted to a wide variety of ecosystems (from tropical to tundra). The tremendous agricultural and economic importance of Rosaceae is mainly associated with their fruit.

Though the genus number of Rosaceae (~100) is much smaller than that of Brassicaceae (~400), they share similar species numbers of ~3,000 to 4,000. Brassicaceae and Rosaceae mostly have smaller genome sizes (<1000 Mb) than other angiosperms (which average >5000 Mb) (3).

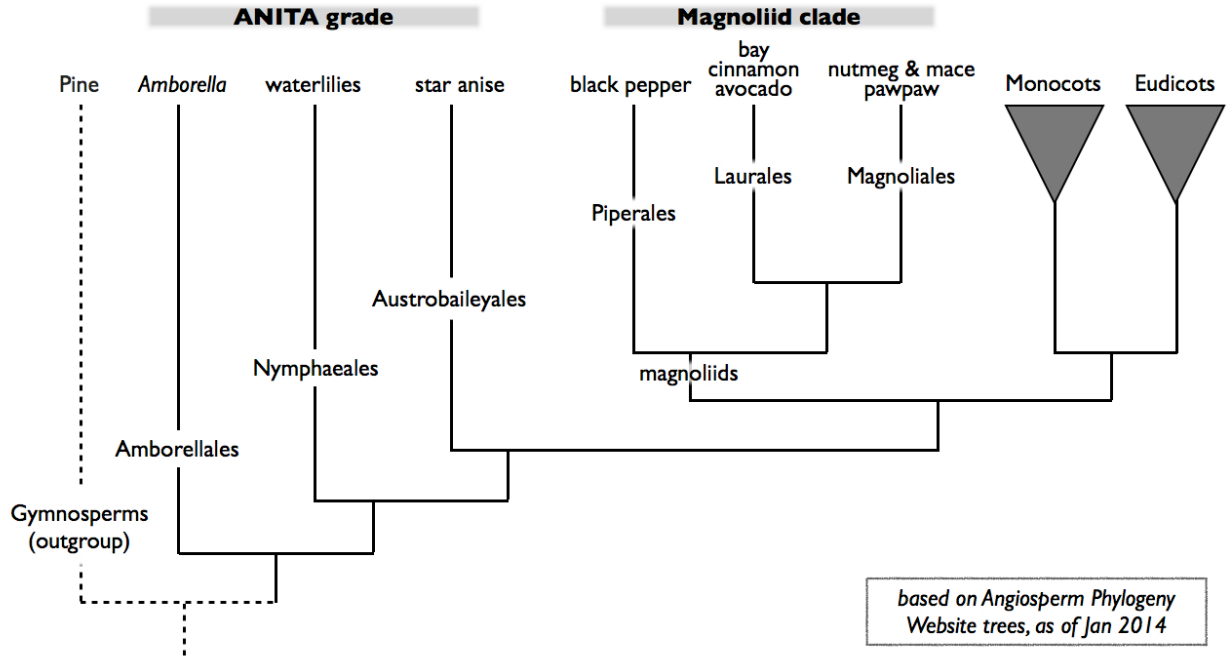

Figure 1.1: **Angiosperm phylogeny**

Figure 1.1 was accessed on Mar.9.2020 from https://botanistinthekitchen.blog/

Transposable Elements and Angiosperm Genome Evolution

The genome sizes of angiosperms differ greatly in both inter-species and intra-species comparisons (4). Angiosperm genomes can be as tiny as 63 Mb (1C) in *Genlisea aurea* and as gigantic as 130,000 Mb (1C) in *Paris japonica*. Although there is such a difference in genome sizes, gene number and collinearity are much more conserved among angiosperms. Coinciding with genome size variability, TE abundance is reported to vary greatly from 13% to 95% (in *Arabidopsis thaliana* and wheat, respectively). Genome size variation is believed to be mostly caused by ancestral polyploidy and transposon amplification, plus a small contribution of segmental duplications. Most genome size variability in angiosperms could be explained primarily by difference in repetitive DNA content, particularly for Long Terminal Repeat Retrotransposons (LTR-

RTs).

TEs can copy or cut themselves out before being inserted into new locations of a genome. Every time that a TE transposes, it can add a new copy to the genome and change previous architectures. According to the way TEs transpose and their structural features, TEs are classified into either Class I or Class II (5). For class I TEs, RNA is transcribed from the original copy and is then reverse transcribed by reverse transcriptase (rt) to generate new copies. The most abundant class I TEs in plants are LTR-RTs, whose structural features including direct long-terminal repeats (LTRs) at both ends, a PBS (primer binding site), a PPT (poly-purine track), flanking 3-5 bp TSDs (target site duplications) and (usually) canonical terminal dinucleotides 5′ -TG…CA-3′. Class II TEs use a cut-and-paste mechanism for most families and a rolling-circle replication pathway for Helitrons. Cut-and-paste TEs usually have terminal inverted repeat (TIR) sequences, so they can also be called TIR TEs. A TIR TE can generate a net increase in TE number if the TE transposes right after it is replicated in S phase, and then inserts in a location that has not yet been replicated (6). Also, repair of a TIR TE excision site often uses the sister chromatid as template, such that the excised sequence is copied back into the excision site (7). Therefore, like retroelements, TIR TEs can rapidly increase their copy numbers. Class I TEs are further classified into orders (LTR-RT and non-LTR-RT), subclasses (e.g., Copia and Gypsy in LTR-RTs) and families (based on degree of similarities of LTRs of LTR-RTs). Non-LTR-RTs include LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements). The generally accepted structures and phylogenetic relationships of retrotransposons based on their *rt* sequence is presented below (8). The deep branch of non-LTR-RTs shown in Figure 1.2 suggests that non-LTR-RTs are the earliest

retrotransposons and the current diverse structure of various retrotransposons (e.g. non-LTR-RTs and LTR-RTs) could be explained as the gain and loss of functions from the core progenitor elements (9). Some LINEs have intact ORFs that specify transposition functions, but most have accumulated deletions (especially 5′ deletions) or other rearrangements. SINEs, in contrast, never encode transposition functions but rather contain structures related to tRNA genes or other small RNA genes, plus a short stretch of poly A at the 3′ end. SINEs depend on the proteins encoded by autonomous LINEs for retrotransposition (10). Based on phylogenetic analysis of reverse transcriptase, LINEs are grouped as L1 (LINE-1), RTE, I, R2 and Jockey, which can be further sub-grouped into 28 clades in Brassicaceae (11). Plant genomes mainly have L1 and a few RTE clades (12).

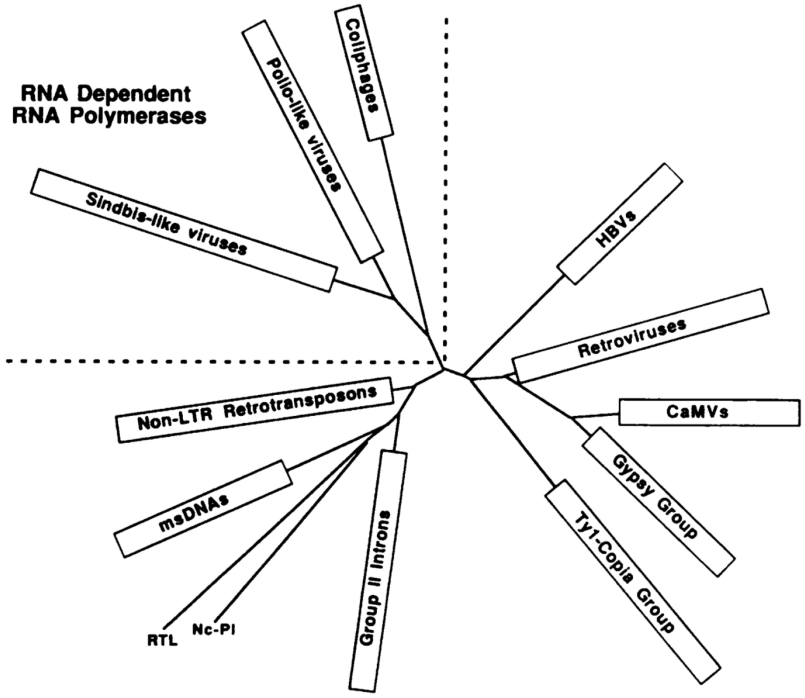

Figure 1.2: **An unrooted phylogeny of rt and RNA polymerase sequences (Xiong, and Eickbus,1990)**

The unrooted phylogenetic tree was built with NJ method using 82 reverse transcriptases of retroelements from animals, plants, protozoans and bacteria and 15 RNA polymerases from various plus-strand RNA viruses. Plus-strand RNA viruses were within the dotted line regions. The length of the boxes corresponds to the most divergent element within that box.

Class II TEs are grouped into TIR elements and Helitrons based on their structural features, which are indicative of their transposition mechanisms. Among TIR TEs, autonomous TIR TEs are primarily classified into 6 superfamilies in plants (Tc1- Mariner, hAT, Mutator, P, PIF-harbinger and CACTA), mainly based on the phylogeny of their transposase, and alternatively on their TIR sequences and having a shared length of their TSD. Meanwhile, MITEs (miniature inverted-repeat transposable elements) are small (<500 bp) TEs flanked by target site duplications and are considered as variably truncated derivatives of autonomous TIR TEs (13). MITEs are non-autonomous and thus depend on the transposases encoded by some autonomous transposons of the same family for a trans-acting transposase protein because transposases of a family can only recognize and mobilize the DNA between the specific TIRs of the same family. MITEs can be further grouped into Stowaway-like, Tourist-like families and others according to the transposase responsible for their transposition. Helitrons are the first transposons discovered to transpose by rolling-circle replication (14). Unlike TIR TEs, Helitrons do not have terminal inverted repeats and do not cause targets site duplication during transposition. This lack of a well-defined structure delayed Helitron discovery to 2001 (14). Helitrons have been found to be abundant in maize, and these TEs often acquire and sometimes express gene fragments inside them, which creates an opportunity for gene creation (15).

In plant genomes, TEs are widely distributed but are functionally suppressed, presumably to avoid excessive mutagenesis. This TE silencing is by a combination of autonomous element mutation (16) and epigenetic suppression (17). TEs are the most

variable part of the genome and can provide both the raw material and rearrangement mechanisms for gene and genome evolution. Overall genome architecture, gene creation and gene regulation are greatly related to TE activity. Additionally, TE activities are known to be activated by some biotic or abiotic stresses such as abrupt temperature change, hybridization, cell culture, viral infection, and tissue damage (18-22). For example, TE mPing was observed to amplify to hundreds of copies in rice after cell culture and, but different levels of amplification were observed between japonica and indica cultivars (23). Earlier, Tos10, Tos17 and Tos19 (families of LTR-RT) were observed to transpose in rice cell culture, and Tos17 copy numbers increased with prolonged cell culture (24).

When activated by tissue culture, mPing was found to preferentially insert into genic regions and at least 4 out 8 insertions of Tos17 were at coding regions. It was found that several families in the LTR-RT Copia subclass had heat-responsive transcriptional regulators in their LTRs, causing some TE transcriptional activation under heat stress (25). TEs' roles in adding regulatory plasticity for plants to confront environmental and biological challenges was emphasized by McClintock in her speech "The Significance of Responses of the Genome to Challenges" (26), although her initial name "controlling elements" was indicative of her belief of a routine regulatory role for TEs in general plant gene expression. The hypothetical model that TEs might help plants survive lethal stresses that might otherwise lead to their extinction was first put forward by McClintock (26) and there is a strong link between some TE expression and steps in defense gene pathways under stress (27). Though TEs can amplify to hundreds or thousands of copies in a short time, most new TE insertion have a mild impact on the genome because they usually avoid inserting into the coding portions of genes, but promoter-preferred insertions have

led some of them to either upregulate or downregulate gene expression, as demonstrated by studies of active mPing in rice (28).

Genome rearrangements like inversion, duplication, and deletion generated by events such as chromosome breakage, ectopic recombination and aborted transposition have been extensively studied in humans (28). The genome architectures present in plants today are the outcome of polyploidization and numerous rearrangements. The rapid amplification and removal of TEs itself is the main source of genome rearrangements in non-genic regions while other types of genome rearrangements such as inversion, deletion and duplication can be caused by TEs through chromosome breakage, aborted transposition or ectopic recombination between homologous TEs at different chromosomal locations (29). Also, the mobility of TEs can reallocate genes in the genome. Intact new gene copies can yield functional loci if retained by natural selection. Many cases of possible gene creation in angiosperms were reported to be associated with gene fragment capture and transposition. *Pack_MULEs*, which are Mutator-like elements (*MULEs*), can capture genomic fragments to form chimeric elements (30). In rice, *Pack_MULEs* have captured DNA fragments from multiple loci, and some are expressed as chimeric transcripts, demonstrating their potential to create novel genes through fusion of diverse genomic sequences. Helitrons can capture host gene fragments from many genes and possibly create novel chimeric genes. In maize, Helitrons were seen to have captured from 0 to 9 gene fragments, and most elements in two major subfamilies were inserted within the last 1 million years, indicating their recent and robust role in genome evolution (31). In maize, most captured gene fragments were found to be undergoing random drift while about 4% were undergoing purifying selection and another 4% were undergoing adaptive selection (32).

Besides, the insertion of Helitrons is biased to genic regions and near other Helitrons.

Evidence from several species has shown that TEs are harboring potentially functional genes but the functionality of the new copies has not been tested yet. A CACTA element in the wheat A genome harboring an acireductone dioxygenase-like protein (*ALP*) gene has been argued to have accelerated the duplication of the ALP family 20-fold (33). Retrogenes are candidate genes whose mRNA-coding domains are at least partly found within a retroelement. There are hundreds or thousands of such genes in many plant genomes, including a case where many leucine-rich repeat protein genes associated with plant disease resistance were retroduplicated in pepper plants (34). The majority of retrogenes in rice and sorghum are argued to have appeared after the split of rice and sorghum lineages (~50 million years ago) because some could be only found in one of the two species (35). DNA removal in plants is so rapid that sequences in the Poaceae without necessary function are retained for less than 5-10 million years, so older retrogenes (generated >10 MYA) must have a role that has been selected either by natural forces or human intervention (e.g., by farmers) (36-38).

TE insertion into the coding portion of a locus usually causes gene inactivation but occasionally produces novel functional alleles (39-42). If inserted into a gene, mRNA transcripts often will be terminated earlier or later than expected and expressed proteins will be truncated or elongated. TE insertion could also influence protein function by frame shifting or reshuffling (43, 44). If inserted near a gene, a TE can rewire the regulatory networks of nearby genes to be depressed or enhanced, generally or in a specific tissue, for gene expression (45). Because TEs often have their own promoters or other regulatory regions, an inserted TE could promote expression of downstream genes. For example, in maize, either Tourist or Hopscotch or both insertions about 60 kb far from the *tb1* ORF

acted as an enhancer and partially explained the apical dominance of maize compared to its progenitor, teosinte (46). The insertion was estimated to predate maize domestication by at least 10,000 years, so this case indicates that selection by the early farmers was acting on standing variation instead of recent mutation.

One interesting question is whether most genes' regulatory regions are derived from past TE insertions followed by subsequent sequence drift and selection. TE-rich regions such as heterochromatin and peri-centromeric regions have hypermethylation to depress TE activity. Multiple lines of evidence suggest that methylation inhibits TE transcription, and global demethylation of genomic DNA will reactivate TE transcription strongly (47). There is extensive evidence in numerous plant species that all cases of gene regulation by epigenetic factors such as DNA methylation are derived from a TE insertion history (17).

Meanwhile, TEs can bring karyotypic changes though transposition-induced activities such as chromosomal breakage or unequal homologous recombination (48). Accumulated rearrangements can play roles in driving plant speciation, for instance by reducing the fertility of heterozygous hybrids (49). Beyond promoting angiosperm speciation and diversification, TEs may also sometimes act as guardians for species integrity. Plant hybrids can have massive transposition bursts caused by transient loss of epigenetic TE suppression, and this could reduce hybrid fitness and thereby prevent gene flow between divergent but close species. TE activation triggered by hybridization has been found in many different species for many different TE families (50). If hybrids could survive DNA combination from different species, and the following TE burst events that rewire the regulatory gene networks, rearrange overall structures and reshape epigenetic environments, it is a big step closer to speciation. The diversification rate of dicot

angiosperms is much higher than for monocot angiosperms by the criteria of morphological diversity, species numbers, ecological distribution, etc. Intra-species and inter-species TE abundance differences and the plasticity that TEs give to gene creation and genome structure could explain a great deal about angiosperm diversification. On one hand, new copies of TEs are generated in the genome by transposition. On the other hand, TEs are also being removed by various deletion processes, especially illegitimate recombination and unequal homologous recombination (37, 51, 52). A particularly important type of unequal recombination is between the directly repeated LTRs of an LTR-RT to generate a solo LTR. This process can attenuate genome growth from LTR-RT amplification, but does not fully reverse genome size increases. Because most TEs will not be retained by natural selection, they are expected to undergo random drift and eventual removal.

TE activity in plants is dynamic but highly repressed by epigenetic processes. Closely related species or phenotypically identical cultivars could have dramatically different TE landscapes for just one or several TE families. TE amplification outcomes will be impacted by such phenomena as insertion site preferences and epigenetic surveillance. Few TEs older than 10 million years can be found in plant genomes, indicating that TEs are being erased at a great rate. Hence, older TEs are usually too truncated or degenerate after going through unequal homologous or illegitimate recombination to be recognizable. Degraded and truncated TEs are no longer active because key structural features and coding potential are erased.

The most abundant TEs regarding DNA amounts are LTR-RTs for almost all documented angiosperms, except those with tiny genome sizes. In most angiosperm genomes, LTR-RTs make up >80% of all TEs by the criterion of genome mass. Overall

TE proportions in the Poaceae family are 85% or more in average size genomes such as those of diploid wheat or maize, but the largest diploid genomes have not yet been carefully investigated. As for non-LTR-RTs, LINEs are often mildly repetitive in plant genomes, with highly heterogeneous and truncated copies. But Del2 is extremely abundant in the large genomes of some *Lilium* species, with ~250,000 copies (52). SINEs are moderately to highly repetitive non-LTR TE in plant genomes, but are usually less than 300 bp long, so they contribute little to genome size variation. Alu, a SINE element in mammals, has up to a million copies in some mammalian genomes and is good at hijacking reverse transcriptase from autonomous LINEs for its amplification (53), but no Alu SINEs have been found in plants. In tobacco, the TS family of SINEs has as many as 50,000 copies. DNA transposon Casper (CACTA) has at least 3,000 copies in *T. monococcum* (54), a diploid ancestor of domesticated polyploid wheat. TE proportions at the class and order levels are drastically different among angiosperms. In a monocot example, rice has ~20% and ~13% while maize has ~75% and ~8% of class I and class II TEs, respectively, measured as the percent of the nuclear genomes that they account for. In a dicot example, *Solanum lycopersicum* has ~63% and ~1% while melon has ~15% and ~5% class I and class II TEs, respectively.

Diverse TE content in angiosperms is partially due to the number and timing of amplification burst events at the TE family level because TE amplification could expand plant genomes in a short term just through a single or a few TE families. For example, the amplification of a few LTR-retrotransposons doubled the genome size of *Zea mays* in as short as 3 Myr (55). *Zea luxurians*, which has twice the genome size of *Zea mays*, had multiple different LTR-RT family amplifications that explain its rapid gain in genome size over the last 140,000 years when the divergence was estimated to have happened

(56), gaining more than 2.5 Gb of new LTR-RTs in this short time frame. Likewise, by studying 3 diploid members of *Gossypium*, it was found that two of the *Gossypium* species had undergone a three-fold increase in genome size primarily because one group of Gypsy-like retrotransposons has undergone massive proliferation over the last 5-10 Myr (57). pvCACTA1, a new family found in the CACTA superfamily that has ~12,000 copies in common bean and other *Phaseolus* species, was not detected in soybean (*Glycine max*) even though *Phaseolus* and *Glycine* shared a common ancestor 8 to 20 MYA (58). On the other hand, very different rates of TE removal can also help explain TE abundance variation in angiosperms (59). For example, deletion by illegitimate recombination appears to be more active in Arabidopsis than in tobacco (60).

Problems in Current Comparative TE Analysis

Since the advent of the whole genome shotgun (WGS) strategy for sequencing genomes, more and more plant genomes have been assembled, annotated and published. Next generation sequencing (NGS) with a lower price, and thus the opportunity for higher coverage, revolutionized how we obtain genetic information. However, the assembly of repetitive elements still remains a confounding issue that cannot be solved by increases in coverage alone (61). The repetitive nature of TEs adds extreme difficulty to genome assembly. Genome assembling algorithms such as the overlap graph tries to find the shortest common substring, which will simply collapse repeats. Algorithms like the De Bruijn graph could reconstruct TE copies to some extent, but it still reshuffles TE sequences in a way that is greatly dependent on the k-mer employed. Compressed alignments are produced when multiple copies of transposons are collapsed to one location, which leads to mis-assembly and highly underestimated TE contents when highly repetitive TE are usually collapsed and reduced to a minimum number of copies.

The accuracy and completeness of full genome assembly has been investigated by comparing genomic sequences in an assembly with fosmids assembled from Sanger sequencing in date palm, and it was found that genes were recovered at a much higher frequency than repeats in the assembly (62). In addition, sometimes genome assemblers do not know where to insert TEs from distantly separated portions of the genome but mistakenly switch them around in genome assemblies. These assembly problems mean that "full genome assemblies" usually (perhaps always) do not represent fully the real TE copy numbers or TE locations. What make this situation particularly problematic are how many TEs there are in the genome. Most angiosperms have >50% of their nuclear genome in TEs and some >90%.

New software packages such as McClintock and RelocaTE can identify either polymorphic or shared TE insertions between a reference and unassembled NGS (63, 64), but these are of limited value if the initial reference genome assembly is deficient, as all or virtually all are. Also, a substantial number of reads are routinely left unassembled, especially repeats.

Because TE annotations on plant genomes were completed individually with somewhat different approaches, results are not easily or convincingly comparable. Finally, it should be noted that TE prediction softwares have high false-positive rates so that they require manual inspection that is almost never performed in published genome sequencing studies. Thus, a large-scale and accurate comparative TE study among different plant species is nearly impossible because of the lack of TE content accuracy in the published assemblies. Long-read sequencing technologies (65), which generate longer read that can fully cover most TE copies, provides the potential for better assemblies, but still require a manual TE annotation component if they are to provide any chance for

accurate TE discovery and annotation.

Short reads of next generation sequencing have been used to quantify TEs without a need for genome assembly and were found to be effective in maize, date palm and a few other species. In date palm, a homology-based TE search on WGS data was able to find 50-fold and 25-fold more LTR-RTs of Gypsy and Copia superfamilies, respectively, than in the assembly (62). Homology-based TE finding program like RepeatMasker uses a search engine such as Basic Local Alignment Search Tools (BLAST) with a TE library to search across provided sequences for TE contents. Thus, a good TE library is required before one does homology-based searches. Researchers usually take the best-known library from closely related taxa and use it for TE masking in a new genome assembly. The volatile dynamics of both TE family and sequence divergence renders the TE landscapes of even closely related species greatly different (4), so this strategy will miss many rare or novel TEs. That is, just taking the TE library from closely related species and searching by homology is insufficient to accurately quantify TE content. Besides, the sensitivity and specificity of TE detection solely through homology-based methods decrease with increasing phylogenetic distance between query species and reference species. Thus, if we want to quantify TEs more precisely in each species, a species-specific TE library for each individual genome is optimal.

Different TEs have canonical structures that are necessary for their transposition. The algorithms for structure-based TE discovery methods were designed based on our prior knowledge about the common structural features of TEs. Unlike homology-based methods, structure-based methods are less biased by differing degrees of similarity to a provided TE library. Meanwhile, a structure-based method is more specific regarding the definition of TE architecture than de novo methods that mainly look for fingerprints of

the transposition process (e.g., dispersed repeats with similar boundaries). Thus, structure-based methods can discover TEs even if they have low copy numbers, an essential skill because many TE families are likely to exist in low copy numbers, such as the copy number of intact LTR-RTs in maize, which is a median of one copy per family per genome (64).

Selecting TEs by specific structural requirements at the beginning will remove many false positives that do not pass the structural criteria we set. For example, the canonical start and end sequences of LTRs are TG and CA. However, non-canonical motifs in LTRs that are not palindromic are present and were usually ignored for LTR-RT mining. For example, Tos17, which is activated in rice tissue culture, has non-canonical 5′-TG…GA-3′ motifs (64). TARE1, which has highly amplified in the tomato genome has 5′-TA…CA-3′ motifs (66). Seven types of non-canonical LTRs from 42 out of 50 genomes were identified recently, and the majority of them are Copia elements (67). LTR_finder and LTR_harvest are two structure-based softwares to find LTR-RTs at the genome scale and both of them suffered from a high false positive rate, more or less 50% although LTR_finder is more accurate than LTR-harvest (68). Thus, post-software processing such as manual inspection or inspection by home-made scripts is required to remove false positives. LTR_retriever is a newly published package that first merges the output from either or all of LTR_finder, LTR_harvest and MGEScan-LTR, and then removes false positives with the most crucial two steps including the removal of LTR-RTs having majority sequences of tandem repeats, exclusion of LTR-RTs having extended alignments beyond LTR regions. The outcome of false negative depends greatly on degrees of stringency of the parameters, but false negatives won't influence the masking results unless it is single copy LTR-RT in the genome. LTR-retriever was

claimed to achieve 91% sensitivity, 96% specificity and 95% accuracy of LTR-RT detection in rice (69). The software MITE_hunter uses structure-based methods (TIR and TSD) to find MITEs at the first step and removes false positives by MSA (Multiple Sequence Alignments).

Some programs undertake a de novo search for TEs that is based on an intrinsic repetitive nature of TEs. As mentioned above, this assumption is often incorrect, but this strategy does identify the most abundant TEs that are expected to be most important for genome structural variation. Repeat Modeler is software to de novo find TEs using assembled sequence data and Repeat Explorer is software that uses shotgun reads (69). The first step in de novo discovery with these programs is a genome-wide self-self-comparison, which takes a lot of computational space and time. In this approach, pairs of repeats will be clustered into repeats families and later classified into specific TE categories. De novo methods are not only specific to TEs, because they also find tandem repeats, satellites and segmental duplications. These de novo methods are not able to find TEs with copy numbers of one, two or whatever repetition threshold that the user sets.

Genomes have many TE relics that have been scrambled by mutation, recombination and nested insertions. For example, in *A. thaliana*, only 10% of TE copies were detected to be at least 95% similar to a consensus library (70). This suggests limited recent activity and the prevalence of old elements. However, most of the above methods are ineffective in finding highly degraded repeats. On the one hand, intergenic regions such as TEs are diverging rapidly, so, even if we could find the recently inserted elements of one TE family and use them as the library, the degraded older elements of the same TE family may no longer be recognizable by sequence similarity. On the other hand, some TE families may only be fossil remains in a genome (with no modern relatives in their TE

family), so only active/new TE families would be detected. Thus, those degraded TE families are no longer active and have had no recent amplification, which makes it even harder to find those degraded elements without recent copies in the genome to put into the TE library. False negatives caused by TE degradation may be predicted by taking known TEs that are structurally degraded, using a context-sensitive evolutionary model, but this has not been shown to be effective. Or we can infer the ancestral states of TEs and use them as consensus element to search for older TEs from the same family. Fortunately, it has been found that consensus repeats from foreign species were even more effective and accurate to find ancestral repeats in *A. thaliana* because of shared ancestry despite the long decay of those repeats (71). This may be explained because repeat decay and/or disappearance may have created less degenerate orthologous repeats in one species compared to another. Hence, the more species you include, the bigger chance you have to find any specific repeat in a still-identifiable structure. Thus, a pan-species TE library should be able to find recent as well as ancestral repeats.

The abundances and degrees of TE type heterogeneity are dynamic at both inter-species and intra-species levels. One of the justifications for my project is that most genome sequencing projects did a poor job of annotating their repetitive DNAs, such that over half of the published genome sequencing manuscripts appeared to be dramatically incorrect in calculating TE and other repeat contents (see below). This is often because they used incomplete or inaccurate methods and is also because most annotate only the assembled DNA but ignore the unassembled proportion that is often most enriched with repeats. In order to fully understand the role TEs play in angiosperm evolution, the first step is to quantify the TEs in each genome accurately. Afterwards, TE abundance will be describable at the TE family level in each species from major angiosperm families.

Because TE activation and repression occur at the family level, not at the level of subclass or superfamily (66, 72, 73), understanding individual family dynamics will allow a much more detailed understanding of patterns in the history of TE activity. In this way, I will be able to draw a map of TE abundance and dynamics during angiosperm diversification with higher resolution.

We chose 2 angiosperm families for this analysis because they are the most studied and have the most available WGS data and genome assemblies covering the largest number of clades in each plant family phylogeny. These two families are the Brassicaceae, including well-studied cruciferous vegetables and scientific models, and the Rosaceae, including domesticated species such as apple, pear, and peach that do not vary greatly in genome size. Brassicaceae and Rosaceae anchor the low range of angiosperm genome sizes compared to families such as *Poaceae* and *Liliaceae*.

Objective and Overview of the Dissertation

TE abundances go through ups and downs because of different frequencies/rates of proliferation and removal. This phenomenon is called TE dynamics and influences the evolution of genomes greatly. For now, even with the availability of the sequenced genomes of many plant species across various genera and families, there is seldom any comparative and homogeneous TE analysis going beyond the plant genus level. Scientists have concentrated more on TE population studies or other intra-genera studies. In the beginning, my study was planned to fill the missing gap in inter-genus comparative TE analysis in some plant families, here for Rosaceae and Brassicaceae. We chose Rosaceae and Brassicaceae for two main reasons. First, they both have small average genome sizes, which made repetitive element analysis relatively simple. Second, currently sequenced genomes within the two plant families

span the family phylogeny better than for other plant families. For example, the currently sequenced species in *Solanaceae* are mainly centered on two genera, *Capsicum* and *Solanum*. Though we sampled 12 and 17 species in Rosaceae and Brassicaceae, they belong to 10 and 12 genera, respectively and those genera spanned evenly the phylogeny of those two plant families, including early divergent lineages.

On the way to achieve this goal, we found that current TE analysis seldom went beyond looking at the abundance and content at the TE superfamily and family levels. TE scientists have developed a profound and hierarchical classification system based mainly on TE structural signatures and how they transpose. TEs can be classified, from high to low order, to class, subclass, superfamily and family based both on their sequence features (e.g., TIR or TSD size) and degree of sequence homology.

TEs are active on an individual or family basis, so the activity of a TE superfamily is the averaged outcome of the combined activities of numerous TE individuals or families, and thus not informative relative to any single family. Observing just the TE population effects would not allow us to look in detail at how TEs progressively influence the genome. Therefore, we decided to dig deeper to a lower hierarchical classification level, the TE family level, especially for the most dominant TE types, LTR-RTs. We knew in advance that using genome assemblies to find TEs would be essential, but that use of the assembly as a sole source for TE discovery would add a huge bias to TE content determinations (51). So, we utilized NGS raw reads, a more random and accurate representation of the genome, instead of the genome assembly, to finalize our conclusions regarding TE types and abundances. Most importantly, we applied consistent analysis standards to all sampled species, so that the results could be compared across plant species and families. To conclude, the purpose of these dissertations

studies was to use a more accurate way to quantify TEs, especially LTR-RTs, down to the TE family level in numerous currently available plant genome sequences with homogenous discovery and quantitation methods in each of two plant families, Brassicaceae and Rosaceae. Upon finishing the project, I had discovered, quantitated and described TE types and abundance for 29 sampled species in Brassicaceae and Rosaceae. Moreover, linear models were fitted to investigate any possible relationships between the abundance of some LTR-RT superfamilies or families with such characteristics as genome size, karyotype, self-compatibility and life cycle. This analysis is the first attempt to look into TE abundances at plant family levels. I also believe that this analysis will provide insights into the appropriate design and goals of future larger-scale research that will investigate TE dynamics.

Significance

Transposable elements have tremendous effects on genome structure and the evolution of gene function. Through activities such as transposition, insertion, excision, chromosome breakage and unequal recombination, transposable elements can rearrange genome, alter gene structures, regulate gene expressions, move genes and even determine gene numbers. This genome plasticity can provide the raw material for natural selection to improve plant fitness and adaption to novel environments, including new climates. Increased infertility rates in progeny from wide crosses may be linked to TE-related rearrangements. Also, closely related plant species with different mating systems may have very distinct TE landscapes. TEs of all major classes are present in almost all or all plant species. In this dissertation, I investigated NGS raw reads as well as the genome assembly to quantify TE contents and abundance in every sampled species because raw reads should be a random, and thus more accurate, representation of the genome than the genome assembly. Almost all of the current studies of

plant TE populations stopped at high hierarchical levels, namely TE subclasses or superfamilies. Yet, from numerous observations on the timing of TE activity bursts (66, 72, 73) or the relationships between transposases and their targeted TIRs (74), it is obvious that TEs amplification is regulated at the family level. No case has been observed where all (or even a large number) of TE families suddenly activate into amplification. Therefore, there are an insufficient number of TE studies at the TE family level. In this dissertation, instead of staying at Copia or Gypsy subclasses or some higher classification level like class I versus class II, LTR-RTs were classified into superfamilies and families.

Previous comparative TE analyses, especially those published in reference genome papers, suffered from using TE annotations from the compared species that were analyzed by different methods and standards. Also, very few efforts were made to remove false TE predictions from the original output of TE-mining softwares to build a satisfactory TE library. In contrast, I built very high-standard TE libraries, especially for LTR-RTs, by having those elements screened for necessary structural signatures (for example, intact LTR-RTs were required to be flanked by TSDs, and have both PPT and PBS). Also, this dissertation applied a homogeneous method of TE quantification on all sampled species, which excludes the bias brought by TE analysis on different standards of genome assembly by different operators. Previous comparative TE studies were mainly intra-species or intra-genus investigations (56, 75, 76) to infer the evolutionary outcome of TE activities, with few inter-species (77, 78) and even fewer inter-genus studies (73). My study provides an inter-genus comparative TE analysis within two plant families that were chosen to cover a broad range of the family phylogeny. Having explored two families permits comparison of TE dynamics at the plant family level.

I found that a pan-plant family TE library greatly enhanced the TE mining for each

species. Last but not least, all NGS data, nuclear genome assemblies, and chloroplast genomes of all 29 species were downloaded from available databases, so this study emphasizes the re-usability of existing data to make novel findings without spending a penny on sequencing. Overall, this dissertation provides the first detailed exploration of the TE types, abundances and dynamics in raw reads by first building a pan-species TE library from genome assemblies of many species. This not only provides insights into TE dynamics and genome evolution on a plant family level (ranging across tens of millions of years) but also the direction of future research to target those potentially active TE families to understand the transposition process and its outcomes

Chapter 2

Materials and Methods

The table below (Table 2.1) describes the sources of sequence data used in this study. Some properties of the relevant genomes (nuclear genome size and ploidy level) are also indicated.

Genomes Analyzed

Table 2.1: **Species information**

Family	Species Abbr.	Raw Reads	Nuclear Genome	Plastid Genome	Ploidy level	Genome Size (Mb)
R	Pper	*Prunus persica*			2	265[79]
	Pmum	*Prunus mume*			2	280[80]
	Pavi	*Prunus avium*			2	353[81]
	Mdom	*Malus domestica*			2	742[82]
	Pbre	*Pyrus brestchneideri*		*Pyrus pyrifolia*	2	527[83]
	Ptri	*Purshia tridentata*			??	215[1]
	Ddru	*Dryas drummondii*			2	253[84]
	Fves	*Fragaria vesca*			2	241[85]
	Pmic	*Potentilla micrachtha*			2	406[86]
	Rchi	*Rosa chinensis*		*Rosa praelucens*	2	533[87]
	Rell	*Rubus ellipticus*	*Rubus occidentalis*	*Rubus takesimensis*	2	338[2]
	Gurb	*Geum urbanum*		*Geum rupestre*	2	1475[2]
outgroup	Zjuj	*Ziziphus jujuba*				
B	Brap	*Brassica rapa*			2	485[88]
	Bole	*Brassica oleracea*			2	603[89]
	Rsat	*Raphanus sativus*			2	534[2]
	Bnig	*Brassica nigra*			2	591[90]
	Siri	*Sisymbrium irio*			4	262[2]
	Spar	*Schrenkiella parvula*			2	140[91]
	Esal	*Eutrema salsugineum*			2	260[92]
	Tarv	*Thlaspi arvense*			2	539[93]
	Aalp	*Arabis alpina*			2	370[94]
	Atha	*Arabidopsis thaliana*			2	157[95]
	Alyr	*Arabidopsis lyrata*			2	245[96]
	Csat	*Camelina sativa*			6	261[97]
	Crub	*Capsella rubella*			2	219[98]

	Bstr	*Boechera stricta*			2	264[99]
	Chir	*Cardamine hirsuta*			2	225[95]
	Esyr	*Euclidium syriacum*			2	262[94]
	Aara	*Aethionema arabicum*			2	240[100]
outgroup	Chas	*Cleome hassleriana*				

[??] Unknown
[1] Estimated from K-mer counting with Jellyfish
[2] Values from Kew Gardens database

The table shows the species sources for raw reads, nuclear genome assemblies, plastid genomes, ploidy level and estimated genome sizes. The diagonal line indicates the sources of the information described in that cell is identical to its left cell. Raw reads were always from the left-column species, but assemblies used for initial TE definition or for chloroplast assembly are listed in the "Nuclear Genome" or "Plastid Genome" columns, respectively. In the "Family" column, B stands for Brassicaceae and R stands for Rosaceae. Almost all sampled species are diploids. Names in the brackets are abbreviated taxonomic name combining the first letter of the genus and first 3 letters of the species. The numbers in superscripted parentheses represent reference numbers for the publications on genome sizes.

Data Preprocessing

Each species in this analysis was required to have a whole genome assembly, NGS raw reads and a plastid genome assembly. Species in each family were selected on the criteria that we have as many species as possible in each plant family and that the species best cover the phylogeny of that plant family. With these considerations in mind, 17 Brassicaceae species in 14 genera and 12 Rosaceae species in 10 genera were collected. Almost all of the collected species are diploids, except *Sisymbrium irio* and *Camelina sativa* in Brassicaceae, which are a tetraploid and a hexaploid, respectively. Monoploid genome sizes (1 Cx value) for each species were mainly sought and retrieved from the Kew Gardens genome size database (https://cvalues.science.kew.org/). Genome assemblies were mainly downloaded from the NCBI genome database (https://www.ncbi.nlm.nih.gov/genome). Some genome assemblies that could not be found at NCBI were downloaded from http://brassicadb.org/brad/ and

https://www.rosaceae.org. NGS raw reads were all downloaded from the NCBI SRA archive (https://www.ncbi.nlm.nih.gov/sra), where Illumina HiSeq was the sequencing instrument and random fragment selection for sequencing was prioritized. The sources of genome assemblies, raw reads and chloroplast assemblies of each analyzed species were provided in Table A.1.

SRA files were first extracted to fastq files by SRA-Toolkit 2.9.1 and then TrimGalore 0.4.5 was used to improve overall read quality, including adapter trimming and low quality read removal with a minimum mean quality score of 30. BLAT was used to search the raw reads against mitochondrial and chloroplast genomes in order to retain raw reads from the nuclear genome for further analysis. Finally, each raw read was trimmed to 100 base pairs and saved to the fasta format to ensure a homogenous and consistent analysis across these species.

Phylogenetic Tree Building

The majority of the chloroplast genomes were obtained from the NCBI organellar database (https://www.ncbi.nlm.nih.gov/genome/organelle/). For those species without plastome assemblies in the database, the plastome of another species which lies in the same genus was downloaded and was used for phylogenetic placement for that species. For those species that did not have the plastome assembly from another species in the same genus, I performed a de novo plastome assembly using Illumina paired end reads from NCBI SRA. The plastomes that I thus assembled were for species *Prunus avium*, *Purshia tridentata* and *Potentilla micrachtha*. The sequencing library insertion sizes (if not provided) were estimated by mapping the reads back to genome assemblies using bowtie2 (101), samtools (102) and picard (103). From the paired-end raw reads and estimated insertion sizes, plastomes were assembled using NOVOPlasty2.6.4. (104) Chloroplast protein-encoding genes of *Arabidopsis thaliana* (from NCBI) were used to search against each genome using BLAT to identify as

many chloroplast genes as possible in each plastome. Each chloroplast gene shared by all 21 species was aligned using Prank and ambiguous alignment regions were trimmed with Gblocks0.91b (105) with parameter (-t =c) after removing start and stop codons. All 45 refined PCG alignments were then concatenated into a super DNA matrix and partitioned with PartitionFinder2 (106) to determine the best partition strategy. The best partition scheme by AICc standards from PartitionFinder2 was used as partition strategy and GTR+G model was used for RAxML ML analysis (107).

LTR-RT Identification and Classification

The program tRNAscan was used to search for tRNAs in the genome to help LTR_finder to locate PBS sequences in LTR-RTs. Intact LTR-RTs were found by two softwares, LTR_finder and LTR harvest, and filtered by LTR_retriever to remove likely false positives. The input parameters used for LTR_finder were -F 11111000000 -d 100 -D 20000- p 20 -C -M 0.8 -l 100 -L 7000 and the input parameters for LTR_harvest were -minlenltr 100 -maxlenltr 7000 -mindistltr 200 -maxdisltr 27000 -mintsd 5 -maxtsd 5 -motif TGCA -motiftmis 0. The parameter settings for LTR_finder and LTR_harvest were adjusted to have the same standards in each search component for each program (e.g., the maximum length of the LTRs was set to be 7000 bp in both LTR_finder and LTR_harvest). Intact LTR-RTs are defined as having long-terminal repeats at both ends, PBS just 3′ to the 5′ LTR, a PPT just upstream of the 3′ LTR, a TSD and canonical terminal dinucleotides 5′-TG…CA-3′.

There are superfamilies in both Copia and Gypsy subclasses of LTR-RTs. The major superfamilies in the Copia subclass are Ale/retrofit, Angela/*tork*, *bianca*, Ivana/*oryco*, Maximus/*sire* and TAR/*tork* while the major superfamilies in the Gypsy subclass are CR/*crm*, Tekay/*del*, *galadriel*, *reina* and *tat/athila*. In this study, we call the level one step higher than

family the superfamily and the level one step higher than the superfamily the subclass, which is a little bit different from the nomenclature system usually employed (108, 109). To clarify with examples, we call Copia a subclass and we call *tork* a superfamily. The rt (reverse transcriptase) hmmer profiles grouped by LTR-RT superfamilies were downloaded from Gydb2 where there are 52 LTR-RT superfamilies in total. However, the superfamily *bianca* is missing from Gydb2 (110). The rt of the LTR-RT superfamily *bianca* was retrieved from NCBI and a *bianca* rt hmmer profile was constructed and added to the larger rt hmmer profile pool. Intact LTR-RTs were translated and searched against the rt hmmer profile using hmmsearch with the parameter "-domE 0.01". The superfamily of the best hit that has the highest bit score to the LTR-RT can be taken and assigned as the superfamily of that LTR-RT. If no best hit was provided as an output, then I concluded that there is no reverse transcriptase within that element. Such elements were named RLX, where X stands for unknown and RL stands for LTR-RT. For each intact LTR-RT, by looking at the internal regions, they could be assigned to either a specific LTR-RT superfamily or RLX. The 5′ LTRs of each LTR-RTs within a plant family were clustered by similarity and length with CD-HIT/4.6.8 using parameters -c (sequence identity threshold) 0.8, -aL (alignment coverage for the longer sequence) 0.75, and -aS (alignment overage for the shorter sequence) 0.8. LTR-RTs that had their 5′ LTR clustered together were defined as belonging to the same LTR-RT family, following the 80-80 criteria. If the 5′ LTR of an LTR-RT could not be clustered to that of another LTR-RT, we call this LTR-RT a single LTR-RT. When a single LTR-RT is an RLX, we call it SRLX. When an RLX is not a single RLX, which means that its 5′ LTR could be clustered to at least one other RLX (and hence of the same family), I call them MRLXs.

An LTR-RT family is defined as a group of mobile DNAs that share at least 80%

nucleotide similarities (5), not including indel variation. I'll describe LTR-RT classification as an example to clarify how we classify it to subclass, superfamily and family. The LTR-RT family is classified by the "80-80 rule" (5) by clustering their 5′ LTRs. If intact LTR-RTs have their 5′ LTR clustered together spanning 80% of the length with 80% or more similarity, they are then classified to the same family. The superfamily of an LTR-RT is decided by how much their *rt* resembles those classified *rt* to the collection of superfamilies in the gydb2 database. If none of LTR-RTs within a family could be assigned a known superfamily name by their *rt* (an outcome occurring only when there is no clear *rt* in any family member), then I have chosen to call this a member of an RLX family. I've also created the names MRLX (Multiple RLX) or SRLX (Single RLX) depending on the number of intact copies in the genome of that LTR-RT family. If at least one intact member of an LTR-RT family could be assigned to a known superfamily, then the family name for an LTR-RT is named after that superfamily. Within an LTR-RT family that I discovered and defined with the 80-80 rule, it rarely but sometimes occurs that different *rt* sequences are found that suggest membership in two different superfamilies, but most LTR-RTs within a family do not exhibit this issue (Table 3.2 and Table 4.2). If there is a conflict caused by different internal *rt* sequences, the superfamily name of a family is decided by the majority rule. For example, if I found five elements in a family of which, according to their *rt* regions, three were assigned to *bianca*, one was assigned to *tork* and one had a missing *rt,* then I would name the family *bianca_x* where the "x" term indicates the index of the family in the superfamily *bianca.*

Pan-species TE Library Construction and TE Quantification in NGS Raw Reads

LTR-RTs found in each genome assembly were classified to the TE family level.

Transposable elements other than LTR-RTs were identified initially by similarities to MIPS TE database (111) or with similarities to the TEs in two TE libraries made for two specific species (TAIR10 at https://www.arabidopsis.org for *A. thaliana* and the *F. vesca* TE library provided by Dr Hao Wang (85). These LINEs, SINEs, TIR TEs and Helitrons were, at most, classified to the sub-class level and this constituted my "non LTR-RT TE" library. All of the transposable elements from the same plant family (Brassicaceae and Rosaceae) were combined and called the pan-species TE databases for Rosaceae or for Brassicaceae.

For each species, 50 Mb of 100-bp high-quality raw reads were compared to the TE pan-species database of that plant family with RepeatMasker, and this was repeated 3 times to confirm the consistency of the output. Hits that were > 50 bp with 80% or more similarity in sequence were kept for further analysis and were called "filtered" results. Outputs from RepeatMasker default settings but without filtering, which required sequence alignment to have a matching score 255, are called 'unfiltered', 'raw' or 'default' result.

Table 2.2: **Identity and length quantiles of masked raw reads with RepeatMasker default settings (score cutoff of 255) in *Arabidopsis thaliana***

	1%	5%	10%	50%	90%	99%
Identity (%)	73.5	76.8	78.8	89.4	90	99
Length (bp)	20	47	65	100	100	100

Table 2.2 shows the identity and length quantiles of 'unfiltered'/'raw' output of masked raw reads in *A. thaliana*. In order to pass the cutoff score of 255, masked raw reads should have a decent length or have very high similarity to TEs in the pan-species library. In *A. thaliana*, only 5% of the raw reads of all masked raw reads in 50 Mb were masked less than 47 bp out of 100 bp. Likewise, only 1% of raw reads of all masked raw reads in 50Mb

were masked with less than 74% identity.

We decided to compare our results to the standard masking procedure and databases in order to make up for possible under-estimation of LTR-RT contents, perhaps due to missing some LTR-RTs caused by the absence of an intact LTR-RT family member in any of the studied species or technical limitations. After the 50 Mb of raw read data of each analyzed species was masked by the pan-species library I built, the remaining reads were again masked with RepeatMasker against the MIPs LTR-RT library plus LTR-RTs from *A. thaliana* and *F. vesca*. The LTR-RT library combining LTR-RTs from MIPs, *A. thaliana* and *F. vesca* is about four times larger than the pan-species library I built. In this way, we can investigate how many more homologies of LTR-RTs in the raw reads might be present beyond those found by my pan-species TE libraries.

Data Visualization

Data were mainly manipulated with Python scripts that I wrote and plotted using RStudio. The R package software employed was ggtree, dplyr, ggstance and ggplot2. Colors in each plot were mainly chosen with color brewer (http://colorbrewer2.org).

Chapter 3

TE Abundances and Dynamics in Brassicaceae

Brassicaceae Overview

The Brassicaceae (Cruciferae) is a monophyletic flowering plant family that currently includes 372 genera and 4060 accepted species. Most of them are herbaceous plants. They are adapted to various environments, but are particularly common in tropical regions. The top genera regarding species number are *Draba* (440 species), *Erysimum* (261 species), *Lepidium* (234 species), *Cardamine* (233 species) and *Alyssum* (207 species). Phylogenetically, Brassicaceae was estimated to have split from its sister family Cleomaceae 54-60 MYA (112). There are 5 major clades of core Brassicaceae, and Aethionemeae is the sixth clade that is sister to the core Brassicaceae. Major clades A to E in core Brassicaceae diverged from their common ancestors in an E to A temporal order, with the E lineage diverging ~27-34 MYA (112). A, B and C split the last around 23 to 28 MYA from their respective closest common ancestors (112). Some representative genera of this study in core Brassicaceae major clades are listed here: *Arabidopsis*, *Erysimum* and *Cardamine* in clade A; *Raphanus*, *Brassica*, *Sisymbrium* , *Schrenkiella* , *Thlaspi*, *Eutrema* and *Arabis* in clade B; *Iberia*, *Lunaria* and *Biscutella* in clade C; *Alyssum* and *Berteroa* in clade D; *Euclidieae*, *Clausia* and *Bunias* in clade E. Clade E separated from the rest of the core Brassicaceae early but diverged significantly later than those in clade B and C. Some tribes in clades B and C generated many new radiations

after their ancient divergence from other tribes, generating many terminal branches in a 3 MY time span (112). In clade A, *Macropodium* diverged early, followed by other tribes in clade A.

Aethionemeae is estimated to have diverged 37 to 42 MYA and is classified to clade F (112), while core Brassicaceae lineages are estimated to have originated 27 to 34 MYA. *A. arabicum* is found distributed mostly in the Middle East and east Mediterranean regions. Recently, *A. arabicum* was used as a novel model plant to study the light control of seed germination given its many accessions with natural variation in light responses (113).

The genus *Euclidieae* is in clade E. *E. syriacum,* with the common name Syrian mustard, is an annual herb. *E. syriacum*, as a pseudo-halophyte, excludes salt from the root in order to withstand moderate soil salinity, and can therefore be used for degraded land rehabilitation (114).

Arabis alpina (alpine rockcress) in clade B is a perennial model plant that requires vernalization for flowering. Comparisons between *A. thaliana* and *A. alpina* have provided insight into the molecular basis of perenniality (115). *Thlaspi arvense* (field pennycress) is a potential biodiesel feedstock. *T. arvense* is able to endure harsh winters in the Midwestern United States and Canada. As a winter cover crop, *T. arvense* provides an important ecosystem adding nutrition to barren winter grounds (116, 117).

Halophytic plants can tolerate high concentrations of salt in soil. *Eutrema salsugineum* is a halophytic crucifer that can naturally tolerate multiple abiotic stresses such as coldness and salinity (118). Therefore, the study of *E. salsugineum*'s adaption to salinity could benefit the bioengineering of other species to have halophytic traits and thereby adapt to some marginal lands. Like *E. salsugineum*, *Schrenkiella parvula* is an

extremophyte, and it shares extensive synteny with *A. thaliana*. *S. parvula* is well adapted to high levels of multiple ions, like Na+ and Mg^{2+}, at levels that are lethal to Arabidopsis (119). The genome size of *S. parvula* equals that of *A. thaliana,* while *E. salsugineum* is about twice as large. Studies among *A. thaliana*, *E. salsugineum* and *S. parvula* will shed light on the molecular mechanisms of abiotic stress tolerance.

Sisymbrium irio, known as London rocket, is a tetraploid with a monoploid genome size of 260 Mb. *Raphanus sativa* (Radish) is a popular edible root vegetable having various colors and shapes.

Brassica crops are consumed as daily nutrition by humans. They also serve as a model system for the study of genome evolution, partly because they have gone through a complex history of polyploidy events. Brassica crops *B. nigra*, *B. rapa* and *B. oleracea* are diploids of respective B, A and C genomes (120).

In clade A, the most well-known species is *A. thaliana,* which is the annual model organism for plant research due to its small stature, small genome size, short life cycle, ease of genetic transformation and the availability of extensive genetic diversity and genetically characterized populations, including mutagenized stocks. In modern biology, numerous fundamental questions regarding plant physiology and biochemistry were first investigated and resolved in Arabidopsi*s* (121). *Arabidopsis lyrata*, a self-incompatible perennial that diverged from the self-compatible *A. thaliana* lineage ~10 MYA (122), has double the genome size of *A. thaliana*, although they are both diploids. The larger genome of *A. lyrata* can be explained by less DNA loss from small deletions in noncoding DNA and transposons, and by a greater expansion of LTR-RTs than in *A. thaliana* (4). *Camelina sativa*, as an emerging oilseed crop, has properties that suggest it will be excellent for biofuel production under low water and low fertilizer regimens (123). *Capsella rubella* is

another well-studied species that is a close relative of Arabidopsis. *C. rubella* is an excellent model to study the evolution of self-fertilization due to its estimated recent separation (0.3-0.5MYA) from a self-incompatible relative, *C. grandiflora* (124).

Boechera stricta has the common name Drummond's rockcress, and is a perennial subalpine herb that is native to the Rocky Mountains of the western US. Due to its population distributions at different elevations, a case study on trait plasticity in the context of climate change was conducted on *B. stricta*. It was found that genetic plasticity could buffer the fitness declines in plants that are predicted to be caused by climate change (125). *Cardamine hirsuta* is an annual or biennial species that is common in moist environments. Because of its diverse leaf forms compared to its close relative, *A. thaliana*, *C. hirsuta* is used to study developmental processes that determine leaf morphology (126).

Previous Characterizations of Transposable Elements in Brassicaceae

More than a dozen Brassicaceae species have had their genomes sequenced and published, including *Arabidopsis thaliana* (121), *Arabidopsis lyrata* (4), *Capsella rubella* (98), *Camelina sativa* (97), *Cardamine hirsuta* (127), *Brassica rapa* (88, 128), *Brassica oleracea* (89), *Raphanus sativus* (129), *Eutrema salsugineum* (130), *Thlaspi arvense* (93), *Arabis alpine* (131), and *Aethionema arabicum* (132). *Boechera stricta*, *Euclidium syriacum*, *Brassica nigra*, *Sisymbrium irio*, and *Schrenkiella parvula* do not have their sequenced genomes published yet, but their genome assemblies have been made available online (sources are shown in Table A.1). The initial publication of the *A. thaliana* genome sequence indicated that at least 10% of the total genome was comprised of TEs (121). The most recent detailed publication on Arabidopsis TE content has stated that ~18.5% of the nuclear genome consists of TEs (including 6.6% known LTR-RTs), while the comparable values are 29.7% total TEs and 13.1% known LTR-RTs in *A. lyrata* (4). The annual *A.*

thaliana has a very small genome size (~130 Mb), but perennial *A. lyrata* has a genome size > 200 Mb. This could be largely explained by the reduced activity of transposable elements and a more efficient TE elimination in *A. thaliana,* mainly by the accumulation of small deletions through illegitimate recombination (133). Unlike most flowering plant genomes, the *A. thaliana* nucleus contains more class II TEs than class I TEs.

The last common ancestor between *Arabidopsis* and *Brassica* is about 15 to 20 MYA. The ~3X larger genome size of *B. oleracea* than *A. thaliana* is mainly caused by differences in the degree of amplification of both class I and class II TEs, though class I TEs are the most abundant in both species (76). *Capsella rubella* is a Brassicaceae diploid with an estimated genome size of 129Mb. *C. grandiflora* and *C. rubella* separated 30,000 to 50,000 years ago and *C. rubella* lost self-incompatibility almost the same time (124). *C. rubella* is a great system to study how phenotypic variation is achieved to adapt to changing environments as *C. rubella* has a wider distribution than its congeneric species and TEs are regarded as an important source for genetic variation accounting for the high phenotypic diversity in *C. rubella* (134). Previous study suggested that flowering time variation was a key trait correlated to adaption and fitness in novel environments and climate (135). TE insertions to *FLC (flowering locus C)* could somehow explain the variation in flowering time partially in *C. rubella* (134, 136). TE study by short-reads WGS in three recently diverged *Capsella* species of different mating systems, self-incompatible *C. grandiflora*, self-compatible *C. rubella* and *C. orientalis*, showed very different TE dynamics in those three species, where C. *grandiflora* has the highest abundance (136). Though *C. rubella* is closer to *A. lyrata* regarding the total amount of nuclear DNA, the TE abundance and density of *C. rubella* resembles more those of *A. thaliana* (98). The C value of a homozygous doubled haploid line of hexaploid *C. sativa*

was estimated to be 750 Mb (137) and therefore the monoploid genome size of *C. sativa* is about 250 Mb. There are 19% RTs and 3% DNA transposons in *C. sativa* genome assembly (97). *Cardamine hirsuta* is a diploid with an inflated estimated genome size of 230 Mb compared to *A. thaliana*. The inflation of the genome of *C. hirsuta* compared to *A. thaliana* could be mainly explained by having long centromeric and pericentromeric regions enriched with LTR-retrotransposons while *A. thaliana* has about only 14 Mb centromeric regions (127). However, there is a difference that the average age of LTR-RTs in *C. hirsuta* (~ 4.8 Myr) (127) is older than that of *A. lyrata* (~0.8 Myr) (4). *Brassica rapa*, *Brassica nigra* and *Brassica oleracea* are diploid species representing A, B and C genomes respectively, which further hybridize to give rise to several allopolyploid species. *B. rapa* was the first to be sequenced among *Brassica* species whose genome assembly later facilitated the assemblies of other *Brassica* species. However, the first version of the *B. rapa* genome assembly only covered 58% of the estimated genome size (485 Mb) (88). The genome size difference between *A. thaliana* and *B. rapa* was also largely due to the expansion of LTR-RTs and it is estimated that there are 39% transposon-related sequence in the genome, 27% of which are RTs and about 3% are DNA transposons in the early version (V1.5) of *B. rapa* genome assembly (88). A newer version of the *B. rapa* genome assembly (V2.0) was generated by deeper Illumina sequencing accompanied by PacBio sequencing. It covered 80% of the *B. rapa* genome. The authors claimed that there were 32% TEs in the V2.0 genome compared to 25% in the V1.5 genome (138). Besides, according to the age distribution of LTR-RTs in *B. rapa* and *B. oleracea*, they were able to identify a LTR-RT expansion event ~6.5 MYA in *B. rapa* (138). *Brassica oleracea* (603 Mb) has a very similar genome size as *Brassica nigra* (591 Mb), but larger than *Brassica rapa* (485 Mb), all shown in Table 2.1. B. *oleracea* has about 22% class I TEs and 16% class II TEs, making a total of 28% TEs in the genome assembly. It was also found that the intact

LTR-RT of *B. oleracea* amplified continuously over the last 4 MY while a great majority (~68%) of the intact LTR-RT of *B. rapa* came into burst within the last 1 MY (139). Repetitive contents have been analyzed in those Brassica species, but a great proportion of unknown elements made the analysis less citable (90). In the genome of *Raphanus sativus*, about 32% of the genome assembly Rs1.0 was identified as repetitive elements with the Repbase TE library (140). *Eutrema salsugineum* is native to the saline soils in eastern China and is widely used as a model plant to study the salt-tolerance of halophytic plants (118). Because stress levels in halophytes are continuously higher than those in glycophytes, halophytes may have a more efficient capacity to the tolerance of excessive TE activities induced by environmental stress (141). The genome size of *E. salsugineum* is 100 Mb larger than that of *A. thaliana* and there is a dramatic expansion of pericentromeric heterochromatin in *E. salsugineum* probably due to high level of stress (92). Another two montane perennial *Eutrema* species, *E. heterophyllum* and *E. yunnanense*, recently diverged with contrast altitude preference as *E. heterophyllum* lives in the high-altitude Qinghai-Tibet plateau and *E. yunnanense* lives in lowland. The estimated genome sizes for *E. heterophyllum* and *E. yunnanense* are 405 and 423 respectively, both higher than that of *E. salsugineum* (~260 Mb). There is about 70% repetitive DNA and LTR-RTs are the most abundant sub-class of TEs in both *E. heterophyllum* and *E. yunnanense* (142). *Thlaspi arvense* (Pennycress) is a self-compatible winter annual diploid with a 1C DNA content of 539 Mb and it is used as a winter cover crop to survive the extreme harshness common to the Canadian Plains and Midwest United States (116). There are about 21% RTs and 2% DNA transposons in the genome assembly of *T. arvense* (93). *E. salsugineum* is close to *T. arvense* tough they have very different genome sizes but the same karyotype (n=7) (93). Perennial *Arabis alpina* (estimated genome size 370 Mb) has

a large fraction of middle-aged TEs (85% to 95% similarity) but a reduced amount of very young elements, especially for the most abundant Gypsy subclass, which suggests a burst event of Gypsy LTR-RT continuously over some time ago with a mild removal rate. Two gypsy LTR-RT family, accounted for more than one fifth of all Gypsy LTR-RT in the *A. alpina* genome, which could largely explain the Gypsy burst events (131). The co-occurrence of TE expansion and deficiency of DNA methylation maintenance in *A. alpine* may probably reveal a causal relationship (131). *Aethionema arabicum* is a small annual diploid (estimated genome size 240 Mb) with a short life cycle starting from germination in Spring to the reproduction and end of life cycle before heat strikes in summer (143). Seldom any TE analysis has been done in *A. arabicum*. As we can conclude, few inter-species or inter-genus comparative TE analyses have been conducted in Brassicaceae and most of the comparative studies are limited to the comparisons to the TEs of *A. thaliana* with genome assemblies at most to the sub-class TE level (such as Copia and Gypsy subclasses) but not further to the TE superfamily or family level.

Results and Discussion

As seen in Table 3.1 and plotted in Figure 3.1, many (11,819) intact LTR-RTs were found in genome assemblies of these 17 Brassicaceae species. Defined by the presence of intact termini but without requiring fully complete interior components, intact LTR-RTs in these Brassicaceae species could be assigned to a known LTR-RT superfamily by their rt sequences ~80% of the time. The remaining 20% of intact LTR-RTs could not be assigned to some known LTR-RT superfamilies because of their lacking or having heavily degenerated rt sequences. These were called RLXs (unknown LTR-RTs). About 35% of RLXs did not have 5′ LTRs that clustered with any other 5′ LTR, and thus were named SRLX (Single Unknown LTR-RTs), while the remaining ~65%

RLXs whose 5′ LTR could be clustered with at least one other 5′ LTR of another intact LTR-RT were named MRLX (Multiple Unknown LTR-RTs). SRLX and MRLX are both superfamily level designations, and MRLXs can be further classified to family levels based on the clusters made with their 5′ LTRs.

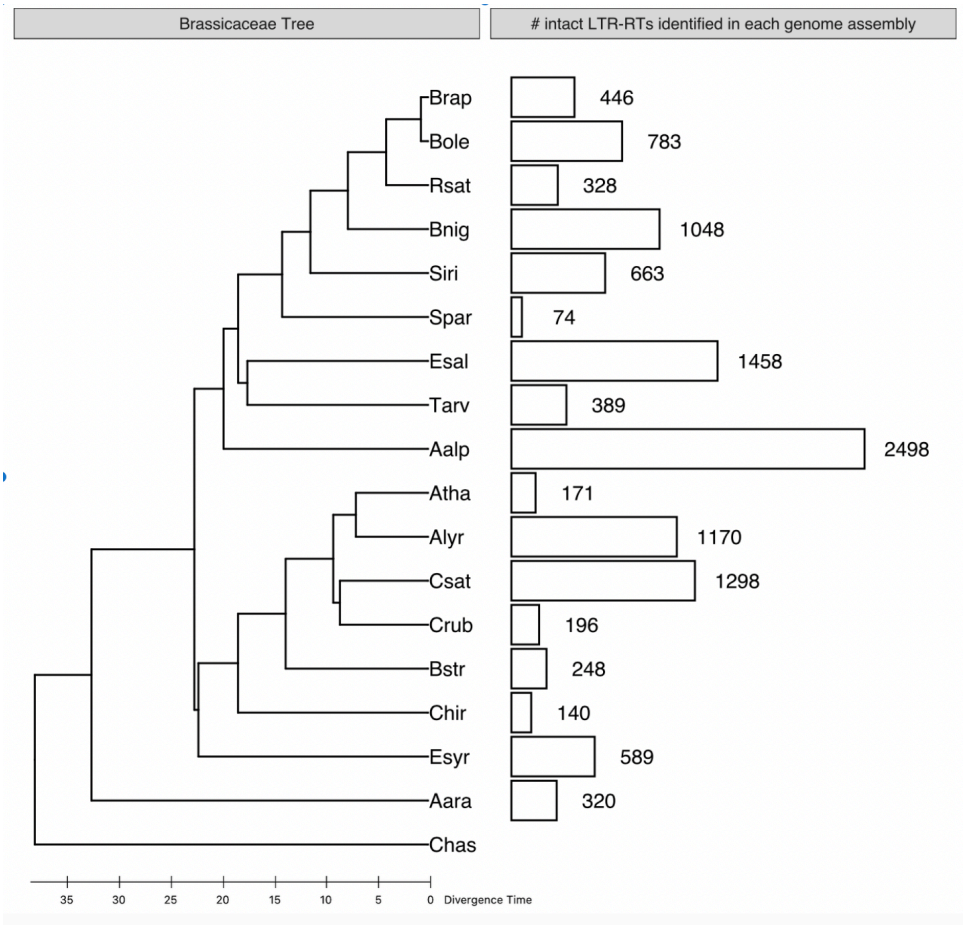

Figure 3.1: **Total number of intact LTR-RTs in Brassicaceae genome assemblies**

Table 3.1: **Number of intact LTR-RTs discovered in the studied Brassicaceae genome assemblies**

Total #	11,819
Named	9,526 (80%)
MRLX	1,476 (13%)
SRLX	817 (7%)

Table 3.2 shows the summary statistics for clusters made with 5′ LTR for all intact

LTR-RTs in Brassicaceae. Within the 11,819 intact LTR-RTs, 2,521 (21%) of them could not be put into any cluster by their 5′ LTRs and the remaining 9298 (79%) could be placed into 1,737 clusters (i.e., LTR-RT families) solely by their 5′ LTRs. Hence, 4,258 families were defined, of which 2,521 families had only one intact family member. Within the 1,737 clusters, 337 (19%) clusters were all RLXs, 1,174 (68%) clusters have their members from only one named superfamily within each cluster, and remaining 226 (13%) have their members within each cluster from more than 1 named LTR-RT superfamily by 5′ LTR sequence characteristics alone. However, a great majority (80%) of intact LTR-RTs could be assigned to a single superfamily by their rt, indicating that the rt sequence is a more definitive phylogenetic signal than is the LTR sequence itself. The majority (87%) of non-singleton LTR-RT families identified by clustering 5′ LTRs have all their members from one superfamily by rt sequence criteria, which confirmed the effectiveness and consistency of these clustering methods.

Table 3.2: **Numbers and types of clusters of LTR-RTs in the studied Brassicaceae genome assemblies**

<table>
<tr><td colspan="4"># of clusters</td></tr>
<tr><td rowspan="3">With multiple elements</td><td rowspan="3">1,737</td><td>With all RLXs within it</td><td>337 (19%)</td></tr>
<tr><td>With elements classified into only one named superfamily</td><td>1,174 (68%)</td></tr>
<tr><td>With elements classified into more than 1 named superfamily</td><td>226 (13%)</td></tr>
<tr><td>With single elements</td><td colspan="3">2,521</td></tr>
</table>

Figure 3.1 shows the distribution of the number of intact LTR-RTs found in Brassicaceae genome assemblies. These numbers are partly an outcome of TE abundance and intactness differences, but are also a technical outcome of the quality of the genome assemblies, because a poorly-assembled genome will preferentially omit repetitive DNAs

of most or all types, particularly longish repeat sequences like those observed for most intact LTR-RTs.

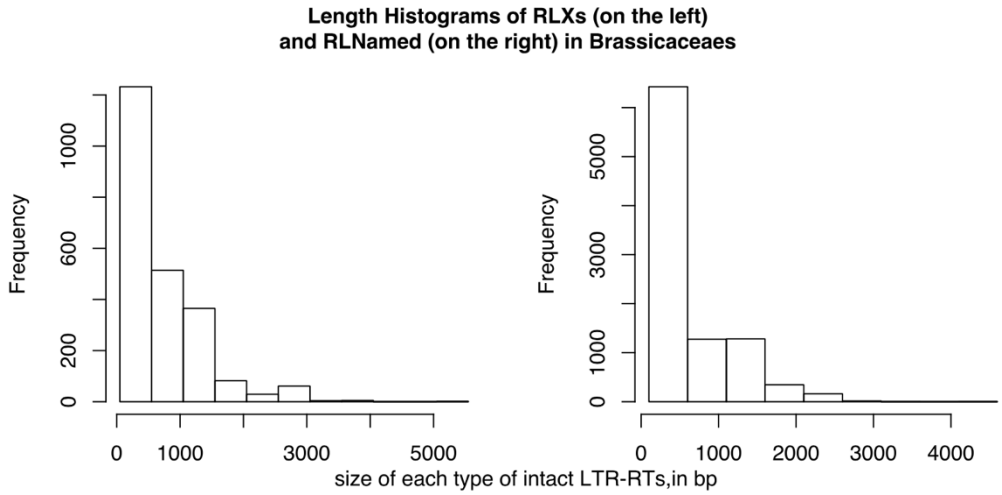

Figure 3.2: **Length distributions of different types of LTR-RT families in Brassicaceae**

The X axis denotes the size of each intact LTR-RT of each type, in bp, and the Y axis indicates the number of intact LTR-RTs of that size in the total set of Brassicaceae genomes studied.

Length distribution of unknown (RLX) and known (RLNamed) intact LTR-RTs in Brassicaceae are shown in Figure 3.2. The majority of intact LTR-RTs are under 2000 bp in length. The size distributions of known intact LTR-RTs are shown with the Gypsy subclass on the left and Copia subclass on the right in Figure 3.3. Gypsy LTR-RTs are longer, on average, than those of Copia. The lengths of intact known LTR-RTs vary greatly within each LTR-RT superfamily, which suggests a mixture of autonomous and non-autonomous LTR-RTs within each superfamily.

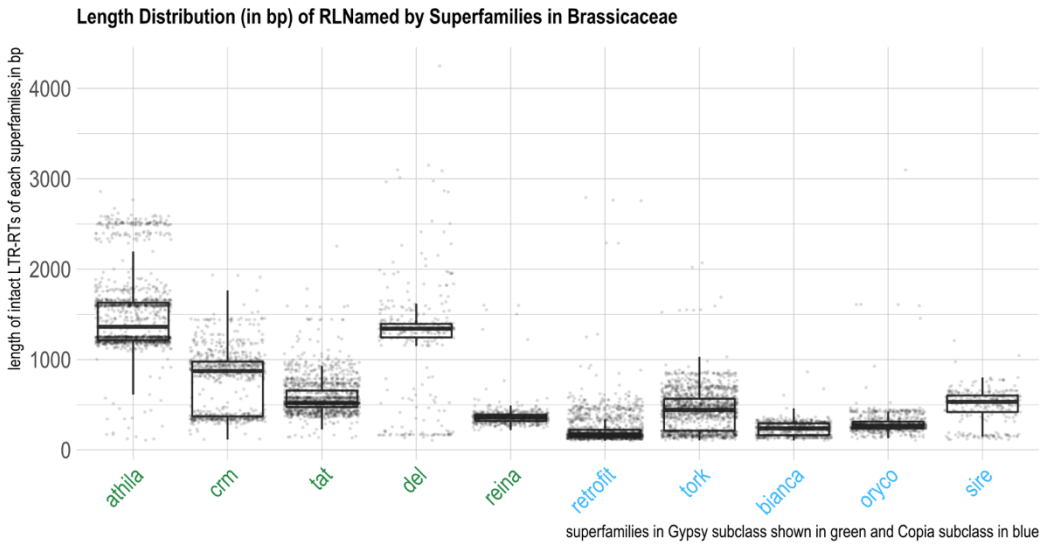

Figure 3.3: **Length distribution of named superfamilies of LTR-RTs in Brassicaceae**

In Figure 3.4, the percentages of identified TEs in raw reads are displayed, with the Brassicaceae aligned on the phylogenetic tree that I constructed. As I only kept the masked reads that were at least 50 bp long and with 80% or more identity to our pan-species Brassicaceae TE library to calculate the number of hits, I provide a stringent estimate of TE amounts in Brassicaceae species that I call the "filtered" result. Table 3.3 and Table 4.3 also provide the percentage of "raw" masked reads that did not go through length (50 bp) and 80% identity filter, which should provide a better estimate of older TEs, but will also be sensitive to a higher level of false positives. Of the studied species, *Thlaspi arvense* has the highest TE percentage (48%) while *Capsella rubella* has the lowest (8%). TE abundances in nucleotides in this analysis were calculated by multiplying the estimated monoploid genome sizes times the estimated TE percentages in raw reads. Estimated monoploid genome sizes were retrieved from publications or the Kew Gardens database where those contributors estimated genome sizes with either k-mer frequency analysis or flow cytometry. Given the fact that the accuracy of genome size measurement is probably not much better than +/- 20% according to different

choice of k size in k-mer estimation, sample preparation to decrease the effect from interfering cytosolic compound and the pronounced differences in estimated genome sizes with these two methods, we are expected to see a +/- 20% bias in our estimated number of Mb of TE DNA in each species.

Because genome sizes vary greatly among these Brassicaceae species, the TE percentage abundances and nuclear Mb can be quite different . For example, *A. thaliana* has 26% TEs, but that 26% only makes up 40 Mb given its tiny genome (~157 Mb). *B. stricta* has 17% TEs in its estimated 264 Mb genome, which contributes 44 Mb. For a better understanding of the number of Mb of TEs in each genome, see Figure 3.6.

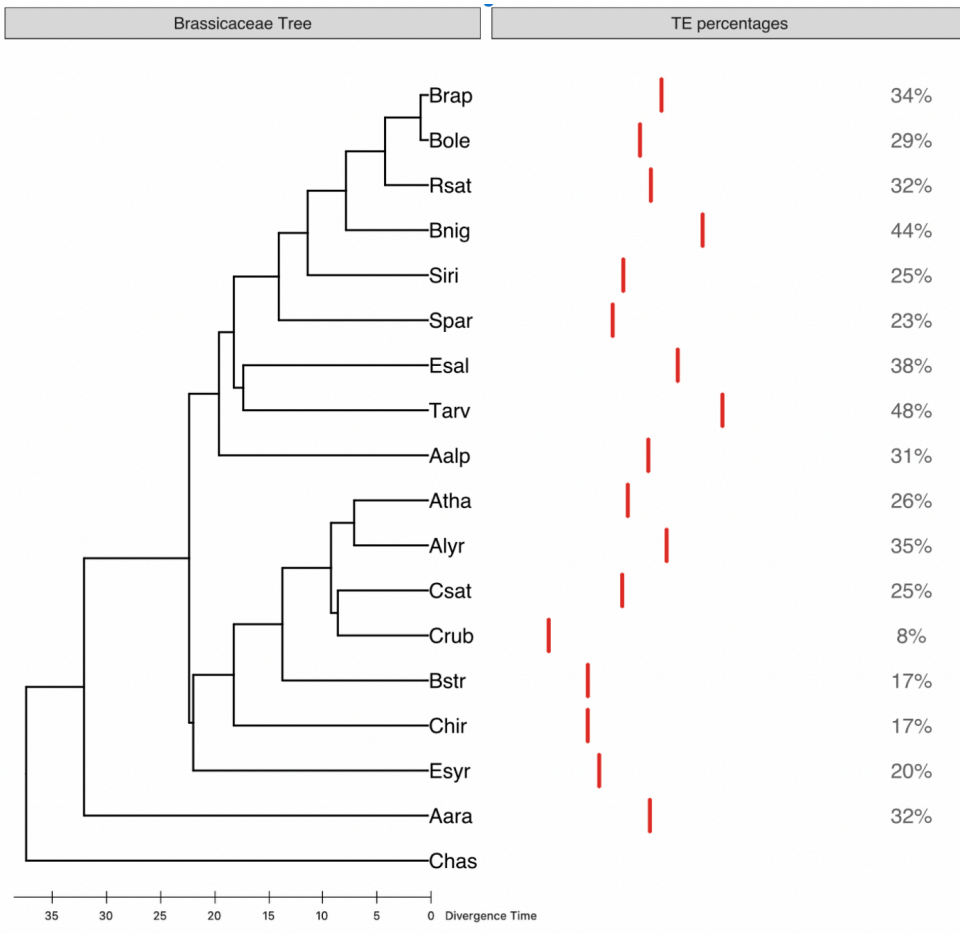

Figure 3.4: **Box plot of the percentages of TEs in Brassicaceae**

In Figure 3.4, Each screening was repeated 3 times with a separately chosen 50 Mb of raw data, and each repetition resulted in the same results (+/- 0.05%). TE percentages are shown with what appears to be a red line, because the boxes are very thin. Numbers to the right are percentages of the total nuclear genome occupied by identified TE sequences.

Table 3.3: **'Filtered' and 'unfiltered' TE percentages in Brassicaceae**

Species Abbr.	Species	TE percentage (default) (%)	TE percentage, after filter (%)	% of "old" TEs	Ratio	% more LTR-RTs[1]
Brap	*Brassica rapa*	40	34	6	0.85	2.6
Bole	*Brassica oleracea*	36	29	7	0.81	0.9
Rsat	*Raphanus sativus*	37	32	5	0.86	1.6
Bnig	*Brassica nigra*	53	44	9	0.83	3.3
Siri	*Sisymbrium irio*	35	25	10	0.71	0.9
Spar	*Schrenkiella parvula*	26	23	3	0.88	0.7
Esal	*Eutrema salsugineum*	43	38	5	0.88	1.2
Tarv	*Thlaspi arvense*	60	48	12	0.80	0.9
Aalp	*Arabis alpina*	36	31	5	0.86	1.3
Atha	*Arabidopsis thaliana*	31	26	5	0.84	3.1
Alyr	*Arabidopsis lyrata*	44	35	9	0.80	3.7
Csat	*Camelina sativa*	32	25	7	0.78	3.9
Crub	*Capsella rubella*	11	8	3	0.73	2.1
Bstr	*Boechera stricta*	23	17	6	0.74	1.6
Chir	*Cardamine hirsuta*	24	17	7	0.71	1.8
Esyr	*Euclidium syriacum*	25	20	5	0.80	2.4
Aara	*Aethionema arabicum*	37	32	5	0.86	0.9
Chas	*Cleome hassleriana*					

TE percentages from RepeatMasker outputs in 50 Mb raw reads with ('filtered') or without (default) length (50bp or longer) and similarity (80% or more) filter were shown for analyzed species. '% of "old" TEs' is the difference between the default TE percentage and filtered TE percentage. 'Ratio' is 'TE percentage after filter' divided by 'TE percentage (default)' where larger values in 'Ratio' indicate higher percentages of recently inserted TEs. They are argued to be more "recent" because of their higher identity and intactness. [1]The column denotes the percentages of LTR-RTs in the previously masked 50 Mb raw read data further masked by other LTR-RT database (with LTR-RTs from MIPs, *A. thaliana* and *F. vesca*). Species highlighted in yellow are those having the "ratio" smaller than 0.8.

The proportions of LTR-RTs in three LTR-RT types (named LTR-RTs, SRLX and MRLX) in raw reads are shown in Figure 3.5. This result is also based on the filter of masked reads to be 50 bp or longer that have 80% or more identity to the Brassicaceae pan-species LTR-RT library. The vast majority of raw reads were masked by LTR-RTs assigned to named

superfamilies and a small proportion of raw reads were masked by LTR-RTs assigned to RLX (SRLX or MRLX) in most of the 17 analyzed Brassicaceae species. One exception to this observation is in *S. parvula,* where the majority of the LTR-RTs are SRLXs. A second exception is in *A. arabicum,* with a majority being MRLXs. Of the 337 MRX families found in total across the Brassicaceae, 147 MRX families were in *A. arabicum*, and two of these were very highly abundant in that genome. A high proportion of MRLX or SRLX transposons may at least partly indicate a poor genome assembly that did not allow intact LTR-RTs to often be discovered.

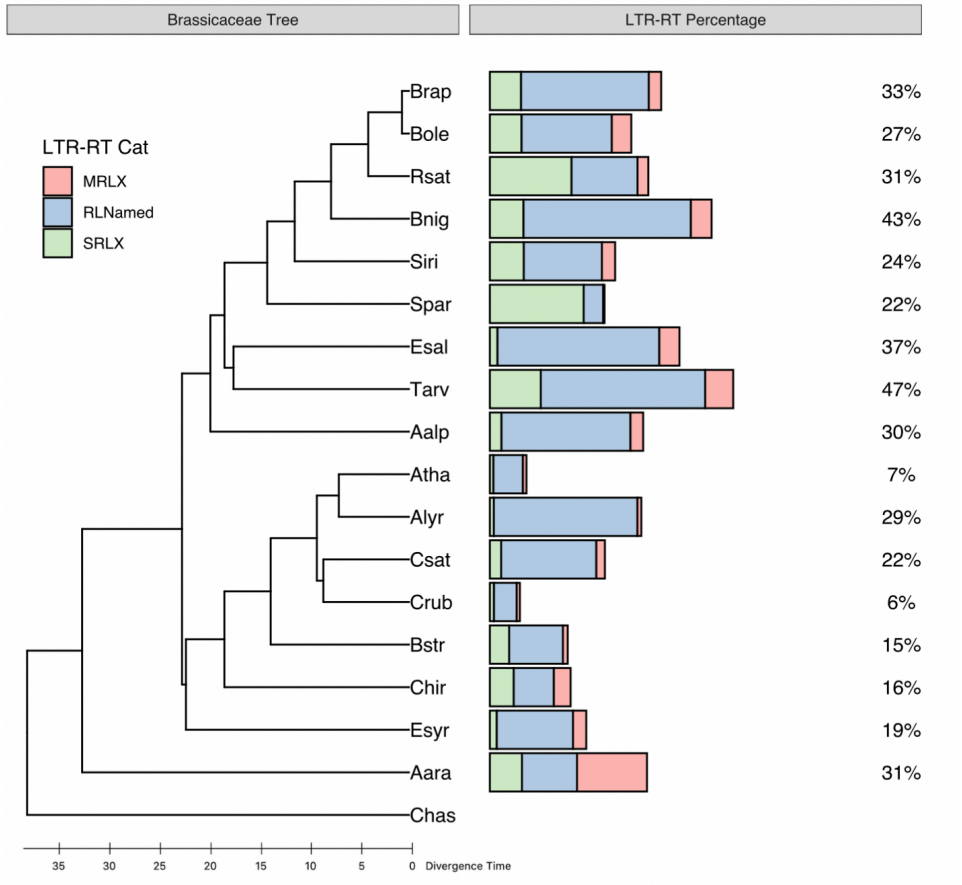

Figure 3.5: **Stacked bar plot of the percentages of LTR-RTs in Brassicaceae** The values are the total percentages of different categories of LTR-RTs in

Brassicaceae, RLX (MRLX and SRLX) and RLNamed, quantified in 50Mb of raw read data.

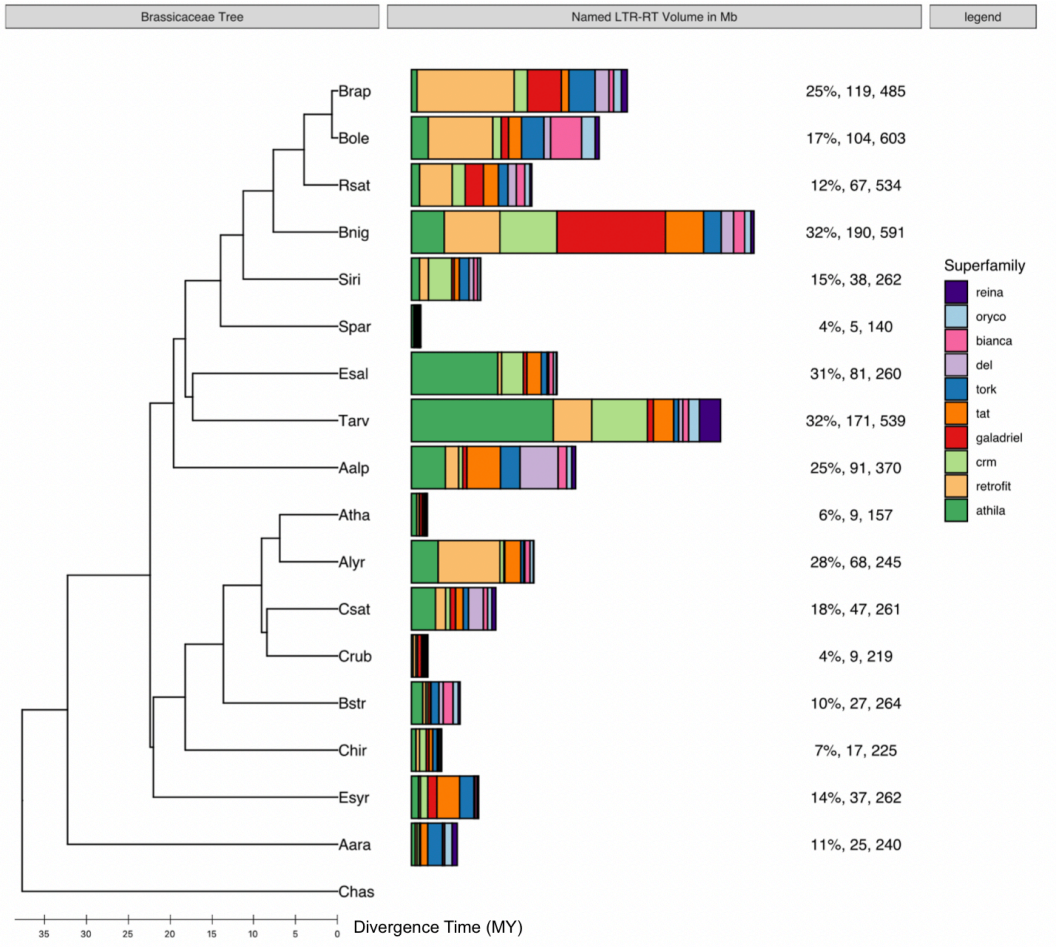

Figure 3.6: **Stacked bar plot of LTR-RT superfamilies in Brassicaceae**

From left to right, the values next to the bars denote the percentages of total Named superfamilies in the genome, the size of total Named superfamilies (Mb) and genome sizes (Mb).

The most abundant known LTR-RT superfamilies vary by species. In *B. rapa*, *B. oleracea* and *R. sativus,* the most abundant superfamily is *retrofit*. In their close relative *B. nigra*, the most abundant superfamily is *galadriel* which apparently became dominant after the divergence between of the B genome (*B. nigra*) from these other lineages 6-7 MYA (144, 145). Some superfamilies show major increases in abundances that appear to

be on a terminal branch, suggesting very recent amplification. These include *bianca* in *B. oleracea* and *retrofit* in *A. lyrata.* The most abundant superfamilies in *E. salsugineum* and *T. arvense* are *athila* and *crm,* while *reina* and *retrofit* are abundant in *T. arvense* but not in *E. salsugineum*. Increased abundances of these four superfamilies are largely responsible for the genome expansion of 90 Mb in *T. arvense* compared to *E. salsugineum*. Two *Arabidopsis* species, *A. thaliana* and *A. lyrata*, diverged about 6 MYA and have a genome size difference of 100 Mb. We found that the most abundant LTR-RT superfamily in *A. thaliana* is *athila* but retrofit in *A. lyrata*. About 60 Mb of the 100-Mb genome size difference between *A. thaliana* and *A. lyrata* can be explained by the enrichment of *retrofit*, *tat* and *athila* in *A. lyrata*. Several other specific LTR-RT amplification results explain other genome size differences (Figure 3.6).

Phylogenetic signals were tested to see whether the nucleotide amounts of LTR-RT superfamilies distribute randomly or have some phylogenetic correlations. In Table A.23, *athila* is seen to be, on average, the most abundant LTR-RT superfamily. It is also the only LTR-RT superfamily showing a strong phylogenetic correlation in these 17 analyzed Brassicaceae species.

Figure 3.7 is an alternative representation of Figure 3.6 to show the known LTR-RT superfamily abundances across the Brassicaceae. The top 10 known LTR-RT superfamilies in Brassicaceae are *athila*, *galadriel*, *crm, tat*, *del*, *reina*, *retrofit*, *tork*, *bianca* and *oryco*. The first 6 are Gypsy elements and the last 4 are Copia. Sometimes Copia dominates and other times Gypsy dominates. The two species with the most total known LTR-RTs, *B. nigra* and *T. arvense*, have more Gypsy than Copia LTR-RTs. The genome size expansion in *A. lyrata* can be explained mainly by Gypsy accumulation, especially the *retrofit* superfamily.

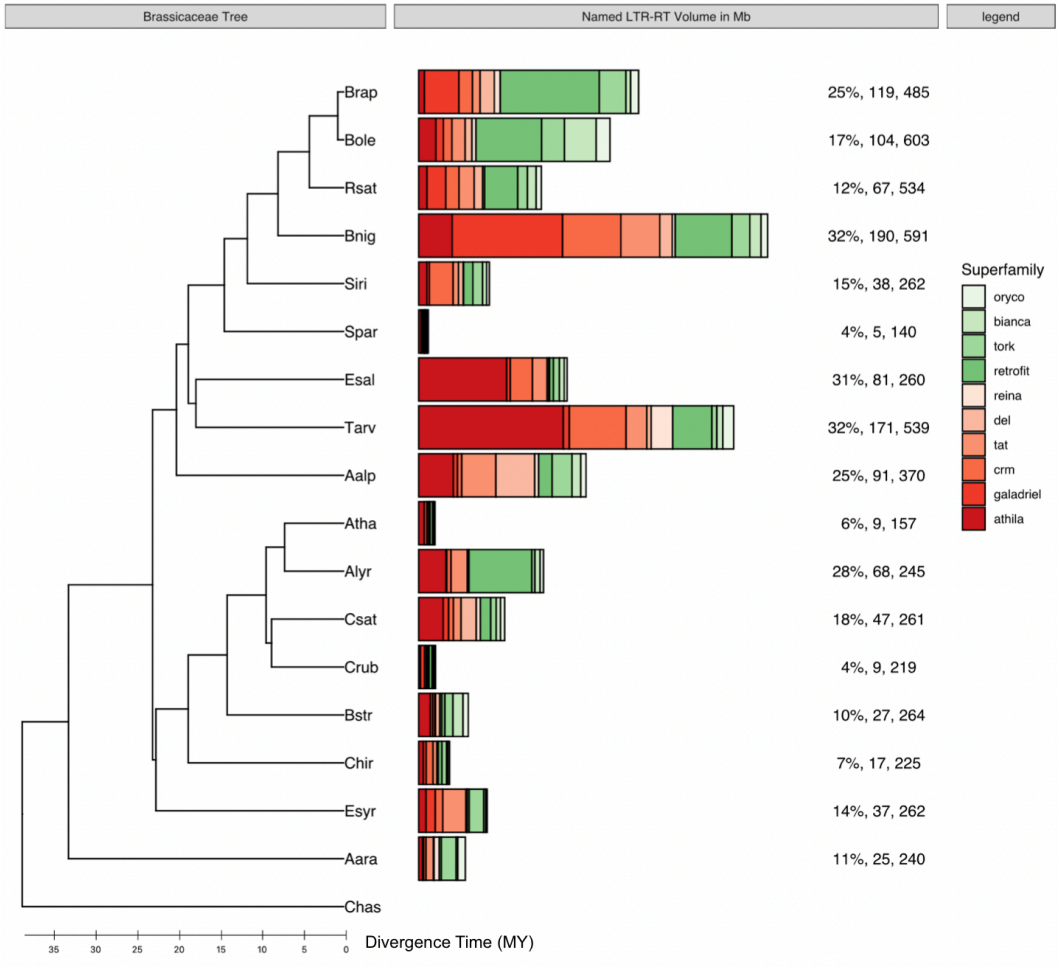

Figure 3.7: **Stacked bar plot of LTR-RT subclasses in Brassicaceae**

Superfamilies of Copia subclass were shown in green and of Gypsy subclass in red. The lighter the color, the fewer the amounts. From left to right, the values next to the bars denote percentages of total Named superfamilies, the sizes of total Named superfamilies (Mb) and genome sizes (Mb).

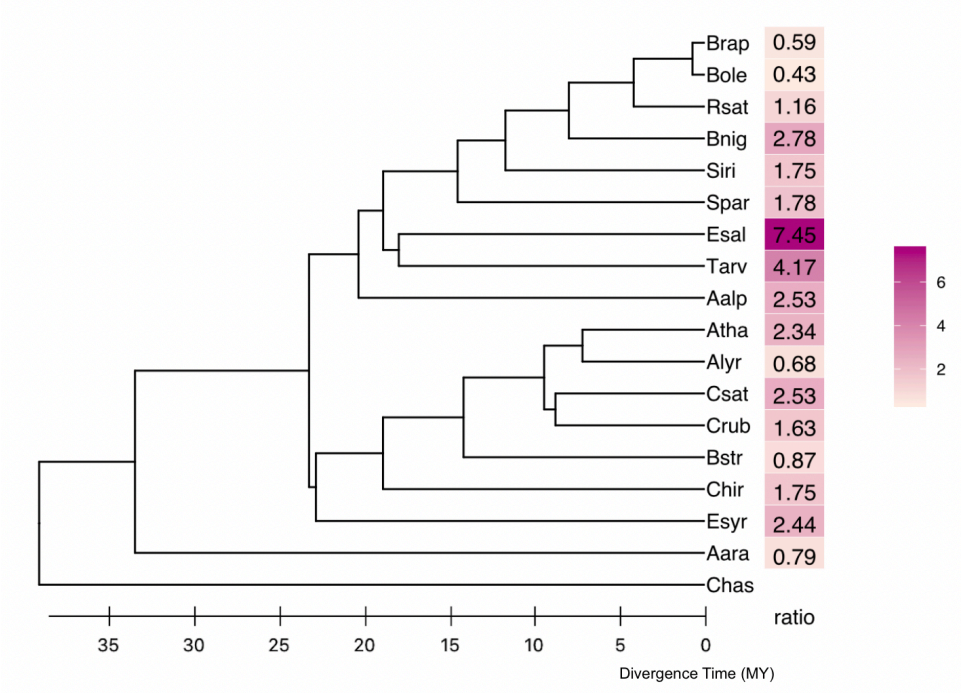

Figure 3.8: **Gypsy-to-Copia ratios in Brassicaceae**

The ratios of Gypsy to Copia quantities (in Mb) are shown in Figure 3.8, which is a further representation of Gypsy-vs-Copia dominance in Brassicaceae. The sister branches, *E. salsugineum* and *T. arvense,* both have a very high Gypsy-to-Copia ratio. Since *E. salsugineum* and *T. arvense* diverged more than 10 MYA, the shared pattern of Gypsy abundance must be a convergent (independent amplification) process, because LTR-RTs are randomly deleted in just a few MY (51, 70, 72, 133, 146). As *B. nigra* diverged from *B. rapa* and *B oleracea* about 6 to 7 MYA, it is interesting that *B. nigra* has Gypsy dominant while *B. rapa* and *B. oleracea* have Copia dominant. The basal lineage in Brassicaceae, *A. arabicum*, has more Copia but the basal lineage in clade A, *E. syriacum* has more Gypsy. Those cases provide evidence that Copia-vs-Gypsy dominance among Brassicaceae species can switch quickly even between closely related lineages.

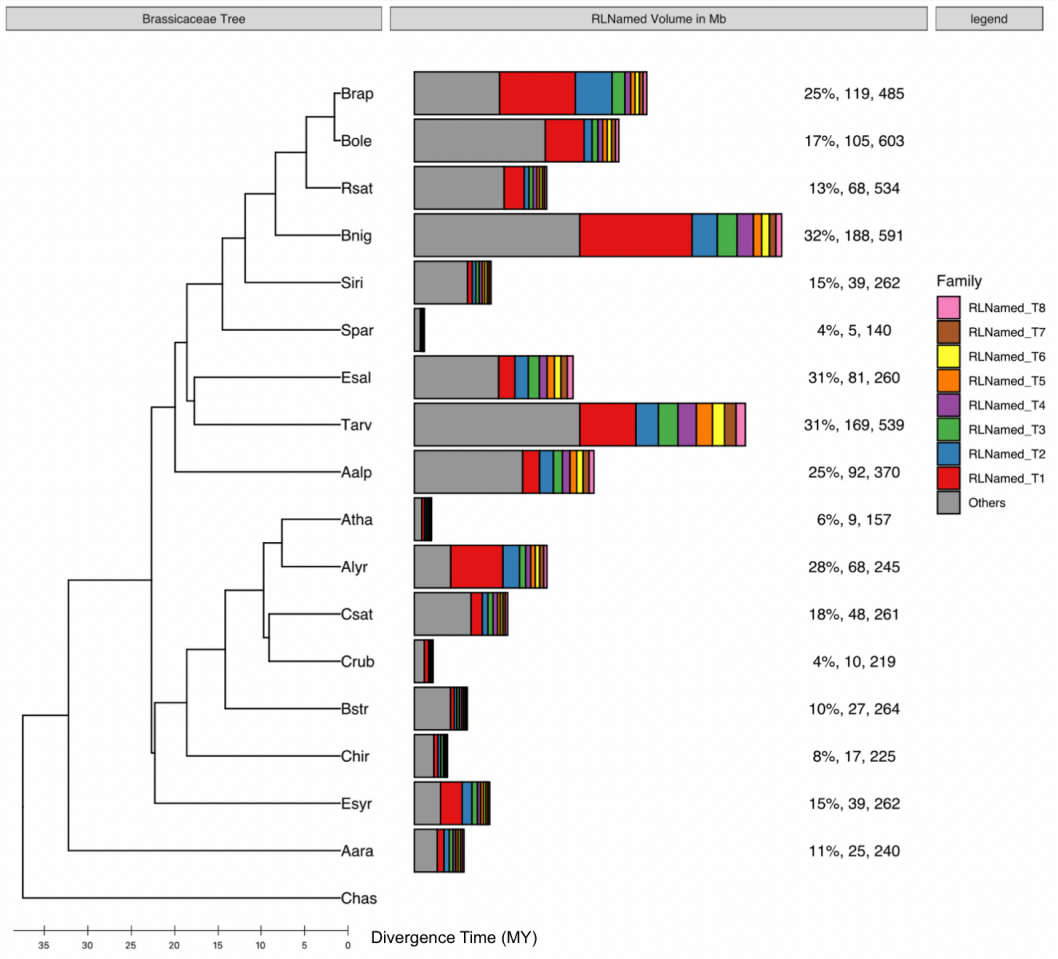

Figure 3.9: **Stacked bar plot of the most abundant Named LTR-RT families in Brassicaceae**

Named LTR-RT family contents were quantified in 50 Mb raw reads. 'Others' (gray fill) indicates the combined amounts of less abundant LTR-RT families. From left to right, the values next to the bars denote percentages of top named LTR-RT families, the size of top named LTR-RT families (Mb) and genome sizes (Mb).

Figure 3.9 was plotted based on the abundance of LTR-RT families that could be given some family name rather than RLXs, and it depicts the abundances of the top 8 LTR-RT Named families. These top 8 known LTR-RT families together can make up as much as 55% of all known LTR-RTs in some species. In some genomes, abundances of the top known LTR-RT families look uniformly distributed, such as those in *E. salsugineum* and *A. alpina*. For

example, in *E. salsugineum*, each top known LTR-RT family makes up about 1% to 2% of the genome. More common, however, are cases where a few LTR-RT families are much more abundant than all others. In species like *A. lyrata*, *B. nigra* and *B. rapa*, the top 3 known families make up more than 1/3 of all known LTR-RTs. In *B. nigra*, the most abundant known family (RLNamed_T1), comprises more than 10% of the genome. The maximum and minimum percentages of the most abundant known LTR-RT family in each Brassicaceae species are 10.88% and 0.55% in *A. lyrata* and *S. parvula,* respectively (in Table A.2). The most abundant known LTR-RT families have mean, minimum and maximum genome representations of 3.38%, 0.56% and 10.88% while the second most abundant known LTR-RT families have mean, minimum and maximum genome representations of 1.45%, 0.31% and 3.87%, respectively, across the 17 analyzed Brassicaceae species (in Table A.3).

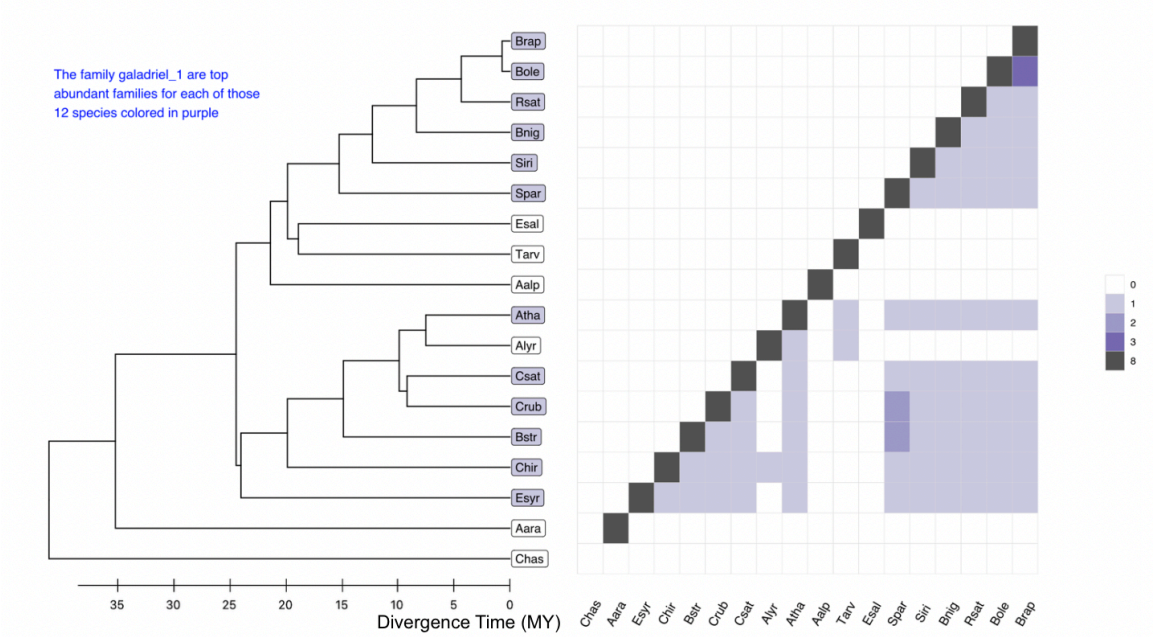

Figure 3.10: **Heatmap of the number of top (most abundant) 8 Named LTR-RT families shared by Brassicaceae**

We were able to find only 4 intact LTR-RTs from the superfamily *galadriel* in genome assemblies of 17 Brassicaceae species. One from *B. nigra*, one from *C. rubella* and the other two from *B. rapa*. Three intact *galadriel* LTR-RTs could be clustered into a family by their 5′ LTRs and this family was named *galadriel_1*. One intact *galadriel* from *C. rubella* was left as a single LTR-RT (SRLX) because its 5′ LTR couldn't be clustered to 5′ LTRs of any other LTR-RTs in the Brassicaceae pan-species TE library. Interestingly, the *galadriel_1* family is among the top (most abundant) 8 known LTR-RT families in 12 Brassicaceae species (indicated in purple in Figure 3.10) out of the 17 analyzed Brassicaceae species. The same analysis done in another plant family (Rosaceae) indicates that it is very rare to share top families among distantly related species. More common, only closely related species such as sister branches in a phylogenetic tree shared top families.

Family *galadriel_1* is the most abundant known LTR-RT family in 5 species, the 2nd most abundant known LTR-RT family in 6 species and the 5th most abundant known LTR-RT family in 1 species in Brassicaceae. Interestingly, although *galadriel_1* is a very repetitive family in Brassicaceae, it was rarely assembled in any of these genomes (probably because of its high copy number) and that is probably why we found only 3 intact copies. Fortunately, raw read analysis allowed us to determine that this family is so abundant in the analyzed Brassicaceae species.

It is interesting that 12 Brassicaceae species, with tens of millions of years of independent descent, share *galadriel_1* as an abundant LTR-RT. This observation motivated me to take a closer look at the 4 intact *galadriel* copies. In Table A.4, I summarized the 5′ LTR length, full length, rt length and identities between the two LTRs of each intact *galadriel*. The four *galadriel* elements are about 4,000 to 5,000 bp with a conserved 5′ LTR length of 461 to

484 bp. Within three individuals in the family *galadriel_1*, only *galadriel_1_1* has an intact *rt* while *galadriel_1_2* lacks *rt* completely and *galadriel_1_3* has a degenerated and fragmented *rt*. The element *galadriel_1_1* encodes a full-length rt though it is only 67% identical to the rt in gydb database, which indicates that a hmmer search is a sensitive tool to find rt even under condition of low similarity. The element *galadriel_S* also encodes a full length rt and it was aligned with those rt from *galadriel_1_1*, *galadriel_S* and the *galadriel* in gydb (Figure B.3). It is clear that many species did not contribute to the intact *galadriel* LTR-RTs in the pan-species intact LTR-RT library but actually have a lot of *galadriel* LTR-RTs in their genome. With better assemblies, it will be interesting to see if the intact *galadriel* TEs were actually missing or just not assembled. The frequency of intact versus degraded LTR-RTs (e.g., solo LTRs and fragmented LTR-RTs) have been shown to be quite variable across species (59, 133, 147). Analysis of short raw reads cannot discern highly intact LTR-RT from smaller fragments, thus making it impossible to differentiate between a paucity of intact LTR-RT due to biological or technical reasons. Of the four intact *Galadriel* elements, the divergence of their paired LTRs indicates that they inserted between 0 MYA and 2 MYA (Table A.4) (36). Hence, these LTR-RTs are much younger than the divergence dates of the lineages investigated in the Brassicaceae in this study. This result agrees with the observed rapid removal of TE sequences (51, 70, 133, 146), which predicts that any abundant LTR-RT family that is shared by terminal lineages would have to have been an independent amplification unless those lineages diverged very recently.

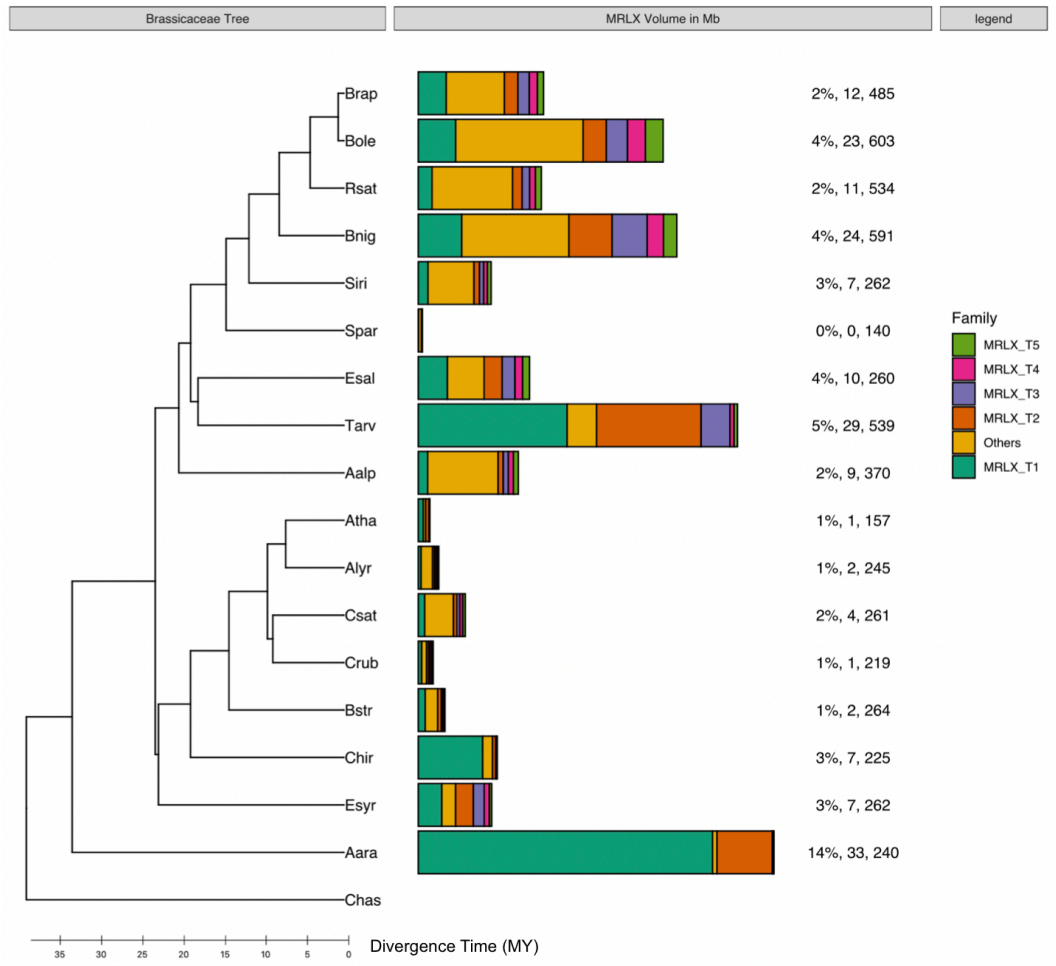

Figure 3.11: **Stacked bar plot of top (most abundant) MRLX families in Brassicaceae**

MRLX family contents were quantified in 50 Mb of raw reads and ordered by family amounts (from top most abundant to fifth most abundant). 'Others' (yellow fill) indicates the combined amounts of less abundant MRLX families. From left to right, the values next to the bars denote percentages of top MRLX families, the sizes of top MRLX families (Mb) and genome sizes (Mb)

RLXs (unknown LTR-RTs) were not given a superfamily name because they lacked an rt sequence signal. RLXs make up a small proportion of most Brassicaceae genomes. An MRLX family is an RLX family where several elements could be clustered by 5′ LTR similarities. Similar to what we have observed in the known LTR-RT families, the top 5 most abundant MRLX families make up usually <50% of the total MRLXs in those Brassicaceae species. Exceptions are the most and 2nd most abundant MRLX families in *T. arvense*, *A.*

arabicum and *C. hirsuta* that make up 80% to 95% of all MRLX contents in those genomes. In *A. arabicum*, the first and second most abundant MRLX families make up ~95% of all total MRLX contents in that species (33 Mb, ~14% of the genome). The top 1 and 2 most abundant MRLX family comprises ~4% (29 Mb) of the genome in *T. arvense*. The absence of any detected *rt* from these families suggest that they are all non-autonomous LTR-RTs in the genome, so their mobilization to such abundance is surprising.

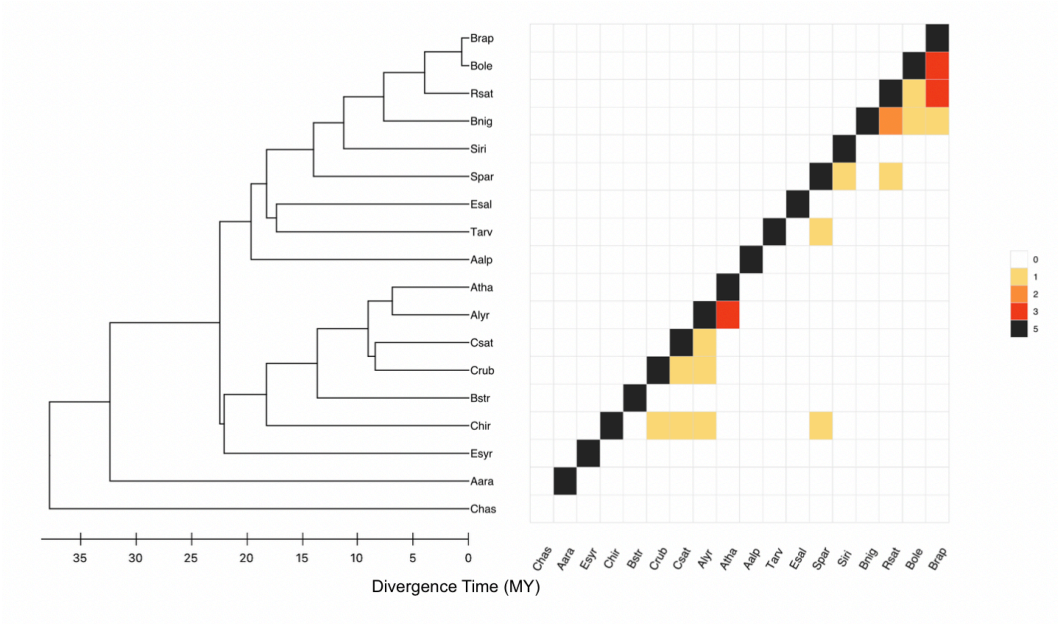

Figure 3.12: **Heatmap of the number of top (most abundant) 5 MRLX families shared by Brassicaceae**

In Figure 3.12, the values represented by colors in each cubic cell denote a pairwise relationship between two species, where the color in the cell indicates the number of top MRLX families shared by the corresponding two species in the tree. There are few shared top MRLX families between any two Brassicaceae species of a distant relationship in the phylogeny. Most colored cells were found on diagonal lines in the heatmap because only closely related species shared top MRLX families. An unusual case of shared top 5 MRLX families was observed

between Chir (*Cardamine hirsuta*) and Spar (*Schrenkiella parvula*), where *Schrenkiella parvula* is placed in clade B and *Cardamine* is placed in clade A of the Brassicaceae phylogeny. Clade A and clade B shared a common ancestor ~33-27 MYA (112).

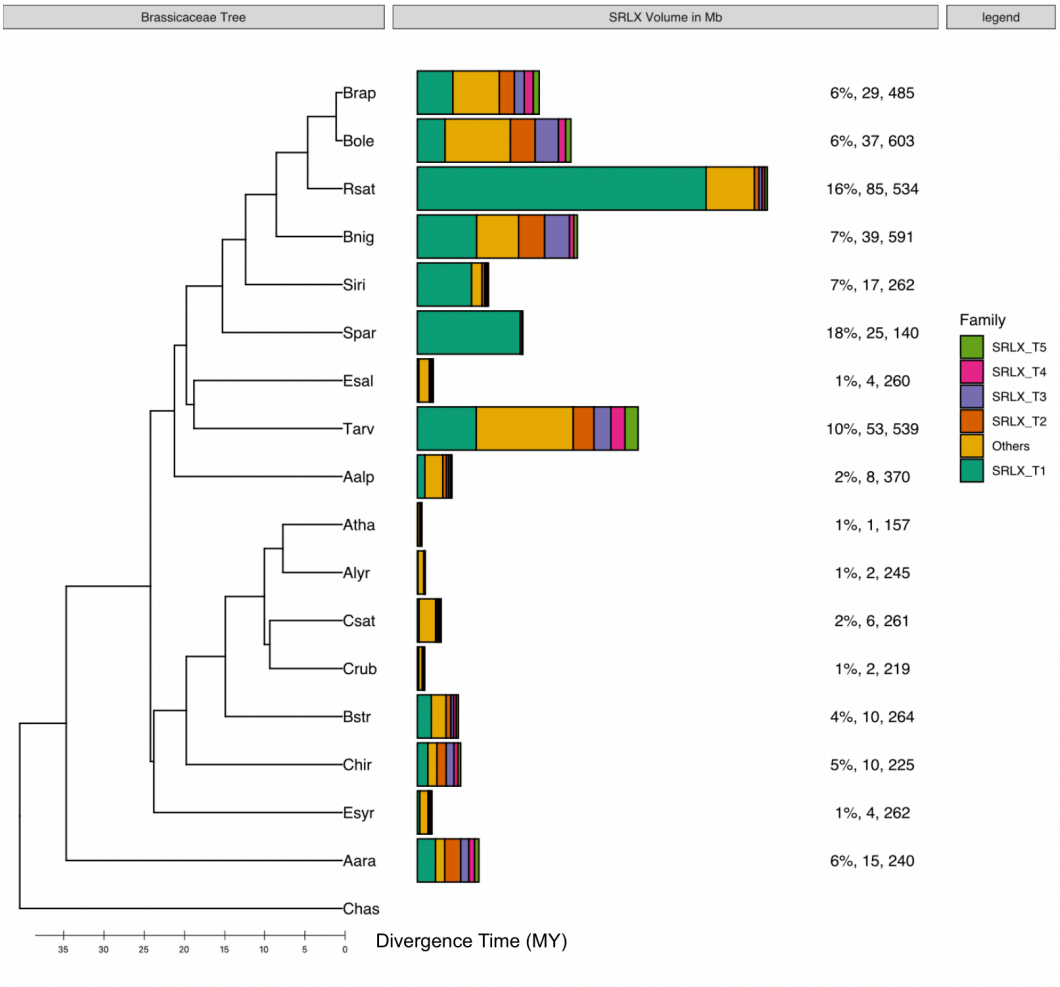

Figure 3.13: **Stacked bar plot of the top (most abundant) SRLX families in Brassicaceae**

SRLX family contents were quantified in 50 Mb raw reads and ordered by family amounts (from top most abundant to fifth most abundant). 'Others' (yellow fill) indicates the combined amounts of less abundant SRLX families. From left to right, the values next to the bars denote percentages of top SRLX families, the size of top SRLX families (Mb) and genome sizes (Mb).

SRLXs are intact RLXs that could not be clustered with any other intact copy in the Brassicaceae pan-species TE library using 5′ LTR similarity. The abundance of top SRLX families are similar to those of MRLXs but it is interesting that there is a greater combined SRLX content than combined MRLX content across Brassicaceae. This suggests, as previously published (66, 72), that there are many-fold more low-copy-number families than there are high-copy-number families in any angiosperm genome. It is also common that highly abundant LTR-RTs quantified in the raw reads have only 1 intact copy in the genome assemblies, and these would be called SRLXs. The top 5 most abundant SRLX families make up, on average, ~50% of the total SRLX contents, a similar result to that observed with the MRLX elements.

In *S. parvula*, *S. irio* and *R. sativus*, the most abundant RLX families make up 80% to 95% of all SRLX in their genomes. A single SRLX in *S. parvula* makes up ~17% (in Figure 3.13) of its genome (24 Mb) while known LTR-RTs and MRLX in total comprise <5% of its genome (Figures 3.9 and 3.11). Therefore, the super abundant LTR-RT family in *S. parvula* is a non-autonomous LTR-RT, that has only a single intact copy in its assembly. In *R. sativus*, the top 8 known LTR-RT family comprise ~13% of the genome while the combined RLX category (MRLX and SRLX) are ~18% of the genome, and a single SRLX makes up ~13% of the genome. This once suggests a tremendous abundance of non-autonomous LTR-RTs. The ratios of RLX (unknown LTR-RTs) to RLNamed (known LTR-RTs) in Brassicaceae are shown in Figure B.5 where *S. parvula*, *A. arabicum* and *R. sativus* all had a high proportion of RLXs (with RLX-to-RLNamed ratios 4.31, 1.82 and 1.38, respectively), which is quite unusual compared to the rest of the analyzed Brassicaceae species.

Top families of SRLX are not shared among distant species (see Figure 3.14), as also seen for most cases of top MRLX families (in Figure 3.12). The most phylogenetically distant case

of shared top SRLX families is observed between Alyr (*Arabidopsis lyrata*) and Bstr (*Boechera stricta*) where *Arabidopsis lyrata* and *Boechera stricta* were estimated to have a common ancestor ~17-14 MYA (112).

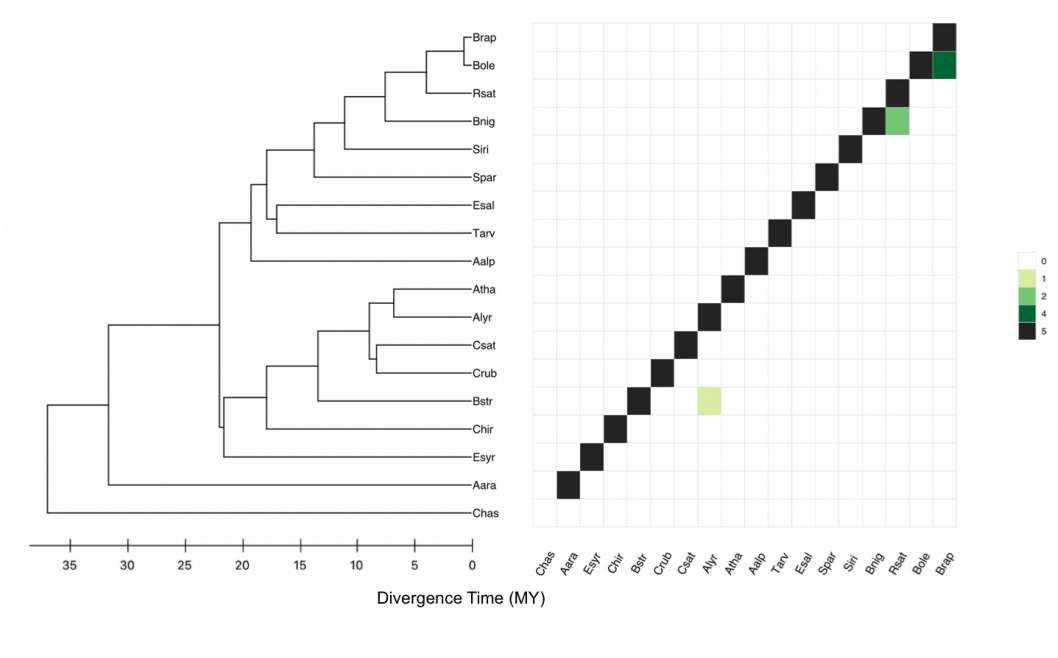

Figure 3.14: **Heatmap of the top (most abundant) 5 SRLX families shared by Brassicaceae**

Chapter 4

TE Abundances and Dynamics in Rosaceae

Rosaceae Overview

Rosaceae is a mid-sized angiosperm family, with >3000 species. Many Rosaceae have distinct fleshy and dried fruits or ornamental flowers with significant economic value. Rosaceae is further classified into 3 subfamilies, Rosoideae, Amygdaloideae and Dryadoideae (148). Rosoideae and Amygdaloideae are large subfamilies with a respective ~2,000 and ~1,000 species, while Dryadoideae has 30 species. According to the phylogeny constructed with nuclear gene sequences, the crown Rosaceae arose ~101 MYA with the separation of Dryadoideae and immediate divergence of the other two subfamilies at ~100 MYA. Hence, these three subfamilies diverged within a very short time.

Fruit are common in angiosperms. Fruit protect the seed development and encourage dispersal of seeds, especially by animals (149). Seeds have a variety of morphological traits that assist dispersal. For example, coconuts can travel along ocean currents to new coasts because they are resistant to immersion in sea water (150), while some fruits have appendages that let them be transported easily by wind or animals. Popular Rosaceae fruits consumed daily by humans include peach, apple, pear, plum, cherries, and berries. Rosaceae are different from other families of angiosperms in that Rosaceae species have distinct fruits while species within other families (e.g. Brassicaceae and Poaceae) produce quite similar fruits (151). For example, the seeds of

apple and pear have a soft core containing multiple seeds while those of peach and plum have a hard-centered single shell. Therefore, Rosaceae is an excellent system to study the evolved mechanisms of fruit development. For the clades included in this study, Amygdaloideae contributed the genera *Malus*, *Pyrus* and *Prunus*, Rosoideae provided *Potentilla*, *Rosa*, *Geum* and *Rubus*, and Dryadoideae gave *Dryas* and *Purshia*.

In Dryadoideae, it is interesting that the *Dryas* genus contains nodulating (e.g., *D. drummondii*) (152) and non-nodulating (e.g., *D. octopetala*) (153) species. *Dryas drummondii* is a slow-growing woody perennial up to three feet tall, usually found in gravel-rich soils. Root nodules and atmospheric nitrogen fixation were observed in *D. drummondii*, a non-leguminous plant, in a few locations (153). Therefore, *Dryas* is a novel system to study the genetic basis for nodulation. *Purshia tridentata* (antelope bitterbrush) is a bushy perennial that grows up to 8 feet tall and adapts well to desert life and provides high quality fodder for domestic livestock, antelope, deer and elk. *P. tridentata* grows slowly, can be deeply rooted (https://plants.usda.gov/plantguide/pdf/pg_putr2.pdf) and shows wide ecotypic variation. *P. tridentata* also bears root nodules (154), and the nodule structure resembles that of *D. drummondii* (155).

In *Rosoideae*, *Fragaria vesca* (wild strawberry) is a herbaceous diploid with a perennial growth habit. *F. vesca* and *P. micrantha* diverged from a common ancestor ~24 MYA (86). *Potentilla micrantha* is a perennial herb that does not develop accessary berries but shares morphological and ecological characteristics with *F. vesca*. Roses are the most important ornamental plants grown worldwide. The selection focus of roses is mainly based on aesthetic effects such as flower color, scent and architecture. *Rosa chinensis* (Chinese rose) is a perennial shrub that can reach 1 to 2 meters in height. *Rubus* is a large and diverse genus having various edible berries such as raspberries and blackberries

(156). *Geum urbanum* (wood avens) is a perennial herb that grows in shady places like forests (157). *G. urbanums*' root infusions have been used to reduce the bleeding and inflammation of gums, with the main responsible compound apparently belonging to the ellagitannins (158).

In Amygdaloideae, *Prunus* is a genus of shrubs and trees with more than 400 species that include such popular fruits as peaches, plums, cherries, apricots and almonds. Many fruits in *Prunus* have a history of human domestication for fruit size, edibility and decorative purposes. Domesticated *Prunus persica* var. *persica* is a deciduous tree native to China and quite popular for its juicy sweet fruit and beautiful blossoms. Peach selection and domestication were estimated to have begun ~7500 years ago, mainly through cloning (79). *Prunus mume* (Chinese plum) is a perennial tree. Plum blossoms (Meihua) flower from late winter to early spring and have a symbolic meaning in Chinese culture standing for purity and the spirit that never gives in to life's hardships. *Prunus avium* (wild cherries/sweet cherries) is a flowering perennial tree native to Europe and Asia. *Malus domestica* (domesticated apple) is the main fruit crop of temperate regions of the world. Apple has been selected for different fruit colors, sizes, flavors, aromas and cooking quality. It was found that an LTR-RT insertion upstream of MdMYB1, an activator of anthocyanin biosynthesis, is associated with the red skin color of apple (159). Pear is one of the oldest fruits, cultivated in six continents and mainly produced in China (160). Pear is an edible interspecific hybrid fruit. *Pyrus brestchneideri* (Chinese white pear) contains high water contents and tastes crisp.

Previous Characterizations of Transposable Elements in Rosaceae

Rosaceae species with the most agricultural, economic and ornamental importance such as strawberry, apple, peach, Chinese pear, sweet cherry, black raspberry and Chinese

plum, have all had their genomes sequenced, annotated and published, including TE annotation (79-81, 83, 85, 161, 162). Less studied Rosaceae species, such as *Potentilla micrantha* and *Rosa chinensis* (86, 87), also have had their genomes sequenced and published.

Woodland strawberry, *Fragaria vesca*, is a herbaceous diploid with a small genome (~ 240 Mb), which is annotated as having about 22% TEs and 16% LTR-RTs in total. However, the genome of *Fragaria vesca* was annotated as lacking highly abundant TE families, and this may be true because it has a small genome size, as do other species in Rosaceae (85). The TE percentage is as low as 1.3% in *Fragaria vesca* (85) if masked by the Repbase TE library. Therefore, a species-specific TE library is required for an accurate analysis of TE contents of that species.

Rosa chinensis is a diploid Rosaceae species with an estimated genome size of 560 Mb. *Rosa chinensis* was reported to have 35% TEs in the genome, where LTR-RTs made up 28% of its genome (87). One Gypsy tat-like family in *R. chinensis* is very abundant and its copies make up ~3% of its genome (87). *Rosa* and *Fragaria* both belong to the Rosoideae subfamily and diverged around 50 MYA (163). A comparative TE analysis was done between *R. chinensis* and *F. vesca,* indicating that there were two-fold more TEs in *R. chinensis* than in *F. vesca,* mainly accounted for by differences in the amounts of LTR-RTs. This difference largely explained the large genome of *R. chinensis* (87).

Potentilla micrantha contains a diploid genome with an estimated 1 Cx (monoploid genome) size of 406 Mb. *Potentilla micrantha* and *Fragaria vesca* diverged from their common ancestor about 24 MYA (164). In *Potentilla micrantha*, 44% of intact LTR-RTs belong to Copia. Superfamily *bianca* makes up ~51% of all Copia LTR-RTs. In *P. micrantha*, slightly over half of all intact LTR-RTs belong to Gypsy, and the *tat*

superfamily contributes the largest proportion (86).

Peach (*Prunus persica*), sweet cherry (*Prunus avium*) and Chinese plum (*Prunus mume*) all have small genomes (from 265 to 353 Mb). The peach genome is comprised of ~10% Gypsy LTR-RTs and ~9% Copia LTR-RTs. The phylogenetic placement of *rt* of LTR-RT families of *F. vesca* and *P. persica* indicated that LTR-RT superfamily diversification predated species diversification (79), as has been observed in every other investigation of this question. Hence, these TE superfamilies were all present in the common ancestors of these Rosaceae lineages. About 43% of *Prunus mume*'s genome was annotated as TEs (80), but additional detail has not been reported on *Prunus* TEs than that described herein.

Pyrus pyrifolia is reported as having 42% LTR-RTs, most of which inserted <2.5 MYA (165). It was reported that the highly abundant families in *P. pyrifolia* were rarely found in *Malus* or *Prunus*. The retroelement rt sequences were found to be highly heterogenous in intact LTR-RTs in *Pyrus* (165). The study of LTR-RTs in two other pear species, *P. communis* and *P. bretschneideri*, provided support to a previous finding (36) that the nucleotide substitution rate of LTR-RTs is at least two-fold faster than those of most gene-encoding sequences (166).

Apple is a diploid tree with an estimated genome size ranging from ~650 to ~750 Mb for different varieties. The apple BioNano genome assembly has ~57% TEs, of which most (>60%) are LTR-RTs comprising ~37% of the nuclear genome. In *Malus domestica*, a 9716 bp consensus LTR-RT family which makes up ~3.6% of the genome was identified (161). This highly repetitive LTR-RT, with more than 500 full-length copies, is called a LARD (Large Retrotransposon Derivative), a category of LTR-RTs that are notable both for their unusual size and lack of coding domains (161).

Purifying selection is responsible for removing deleterious mutation to maintain the long-term stability of biological structures and processes. Protein-encoding genes of a LTR-RT such as reverse transcriptase and integrase are required for the transposition of LTR-RTs, and thus the majority of them should undergo purifying selection to guarantee their normal functions. Reverse transcriptase genes of taxonomically widely separated species maintain good conservation (9), which indicates the existence of a strong purifying selection regimes on LTR-RT genes. For example, in rice, integrase genes of the Gypsy subclass exhibited purifying selection more than 99% of the time and positive selection less than 1% of the time (72). Meanwhile, high levels of purifying selection on reverse transcriptase genes were reported in various plants. The predicted rt protein sequences of different LTR-RT superfamilies were compared phylogenetically in some mulberry species, leading to the subclassification of Gypsy and Copia elements to the superfamily level (79). This study also concluded that the rt of 3 LTR-RT families in Copia underwent positive selection, while most Copia and all families in Gypsy underwent purifying selection (167).

Rubus occidentalis (black raspberry) is a diploid fruit crop with an estimated genome size of 293 Mb. The genome of *R. occidentalis* was reported to have ~11% Copia and ~12% Gypsy LTR-RTs (162).

No TE analysis has been done in any species in the *Dryas* or *Purshia* genera. However, raw reads and genome assemblies are available through NCBI.

Regarding TE research in Rosaceae, it is clear that there were few comparative TE analyses made beyond an intraspecies level. Most of the comparative interspecies TE analyses in Rosaceae were to compare TE contents of a newly-sequenced species to *Fragaria vesca*. Overall, Rosaceae TE studies provide some insights such as that non-

autonomous LTR-RT families can be highly abundant in Rosaceae species (161) and that the most repetitive LTR-RT family usually makes up less than 5% of the genome (87). There is a clear need for more studies of Rosaceae TEs, and these could best be performed with data that are analyzed in the same manner across the genera investigated.

Results and Discussion

Figure 4.1 depicts the number of intact LTR-RTs discovered by the techniques described in Materials. As can be seen, the numbers vary dramatically across the genomes investigated (>40 fold). This is likely to mainly be a function of the quality of the genome assemblies investigated, and not primarily an indication of any major differences in the genome properties of the investigated species.

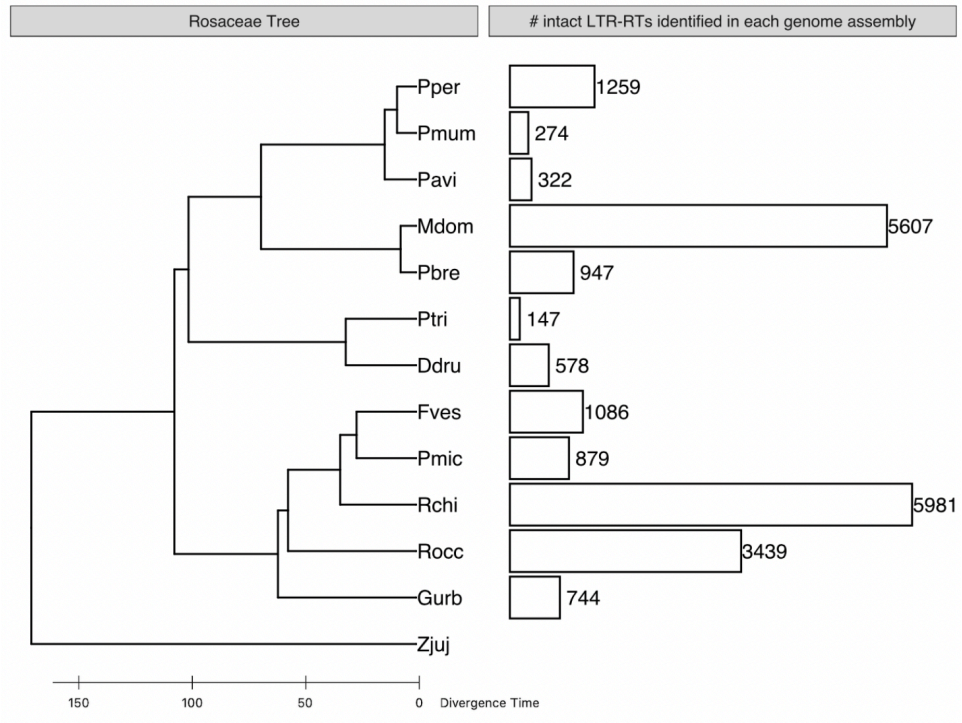

Figure 4.1: **Total number of intact LTR-RTs in Rosaceae genome assemblies**

Table 4.1: **Numbers of intact LTR-RTs discovered in the studied Rosaceae genome assemblies**

Total #	21,263
Named	16,367 (77%)
MRLX	3,945 (19%)
SRLX	951 (4%)

As seen in Table 4.1 and plotted in Figure 4.1, many (21,263) intact LTR-RTs were found in genome assemblies of these 12 Rosaceae species. Defined by the presence of intact termini but without requiring fully intact interior components, intact LTR-RTs in Rosaceae could be assigned to a known superfamily by their rt sequences ~77% of the time. The remaining ~23% of intact LTR-RTs could not be assigned to some known LTR-RT superfamilies because they lacked or had a heavily degenerated/fragmented *rt* sequence. These were called RLXs (unknown LTR-RTs). About 19% of RLXs did not have 5′ LTRs that clustered with any other 5′ LTRs and thus were named SRLX (Single Unknown LTR-RTs), while the remaining ~81% of RLXs whose 5′ LTR could be clustered with at least one other 5′ LTR of another intact LTR-RT were named MRLX (Multiple Unknown LTR-RTs). SRLX and MRLX are both superfamily level designations, but MRLXs can be further classified to family levels based on the clusters made with their 5′ LTRs.

Table 4.2: **Numbers and types of clusters of LTR-RTs in the studied Rosaceae genome assemblies**

<table>
<tr><td colspan="4"># of clusters</td></tr>
<tr><td rowspan="3">With multiple elements</td><td rowspan="3">1,912</td><td>With all RLXs within it</td><td>553 (29%)</td></tr>
<tr><td>With elements classified into only one named superfamily</td><td>1,118 (58%)</td></tr>
<tr><td>With elements classified into more than 1 named superfamily</td><td>241 (13%)</td></tr>
<tr><td>With single elements</td><td colspan="3">2,024</td></tr>
</table>

Table 4.2 shows the summary statistics of clusters made with 5′ LTRs for all intact LTR-RTs in Rosaceae. Within the 21,263 intact LTR-RTs, 2,042 (9.6%) of them could not be put into any clusters (LTR-RT families) by their 5′ LTRs and the remaining 19,221 (90.4%) could be placed into 1,912 clusters (i.e., LTR-RT families) solely by their 5′ LTRs. Hence, 3,936 families were defined, of which 2,024 families had only one intact family member within the studied Rosaceae. Of the 1,912 clusters, 553 (29%) were RLXs, 1,118 (58%) clusters have their members from only one named superfamily within each cluster, and the remaining 241 (13%) have all their members within each cluster from more than 1 named superfamily by 5′ LTR sequence characteristics alone. However, the majority of intact LTR-RTs (77%) could be assigned to a single superfamily by their rt, indicating that the rt sequence is a more definitive phylogenetic signal that is the LTR sequence itself. The majority (87%) of those clusters (LTR-RT families) identified by clustering 5′ LTRs have all their members from one superfamily by rt sequence criteria, which confirmed the effectiveness and consistency of these clustering methods.

Figure 4.1 shows the distribution of the number of intact LTR-RTs found in Rosaceae genome assemblies. These numbers are partly an outcome of TE abundance and intactness differences, but are also a technical outcome of the quality of the genome assemblies, because a poorly-assembled genome will preferentially omit repetitive DNAs of most or all types, particularly longish repeat sequences like those observed for most intact LTR-RTs.

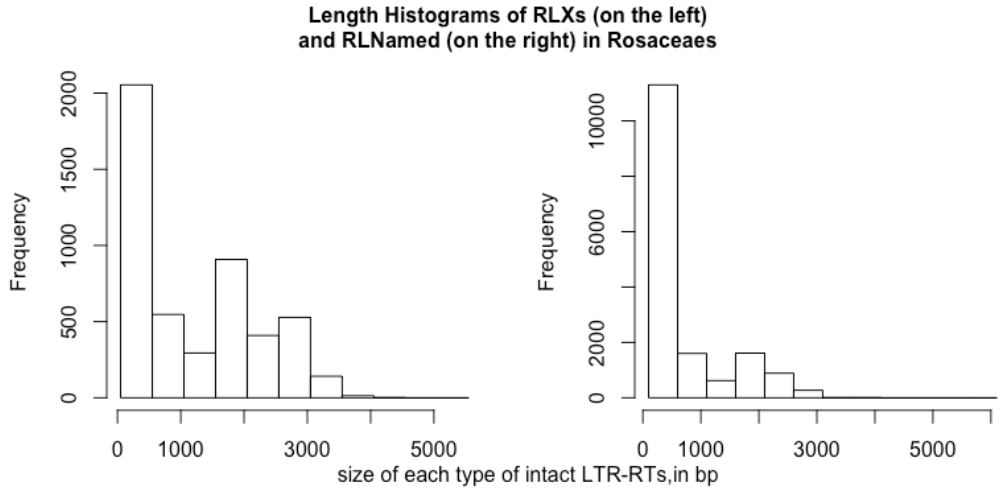

Figure 4.2: **Length distributions of different types of LTR-RTs in Rosaceae**

The X axis denotes the size of each intact LTR-RT of each type, in bp, and the Y axis indicates the number of intact LTR-RTs of that size in the total set of Rosaceae genomes studied.

Length distributions of unknown (RLX) and known (RLNamed) intact LTR-RTs in Rosaceae are shown in Figure 4.2. The majority of intact LTR-RTs are under 2,000 bp in length. The size distributions of known intact LTR-RTs are shown with the Gypsy subclass on the left and the Copia subclass on the right in Figure 4.3. Gypsy LTR-RTs are longer, on average, than those of Copia. The lengths of intact known LTR-RTs vary greatly within each superfamily, which suggests a mixture of autonomous and non-autonomous elements within each LTR-RT superfamily.

On average, *athila* and *del* are the superfamilies with the largest elements in the Gypsy subclass and *sire* and *tork* are the longest superfamilies in Copia. The lengths of *del* in Rosaceae vary more than those in Brassicaceae. Besides, a good proportion of RLX in Rosaceae are 1,500 to 3,500 bp long while there are far fewer RLXs of this size in Brassicaceae.

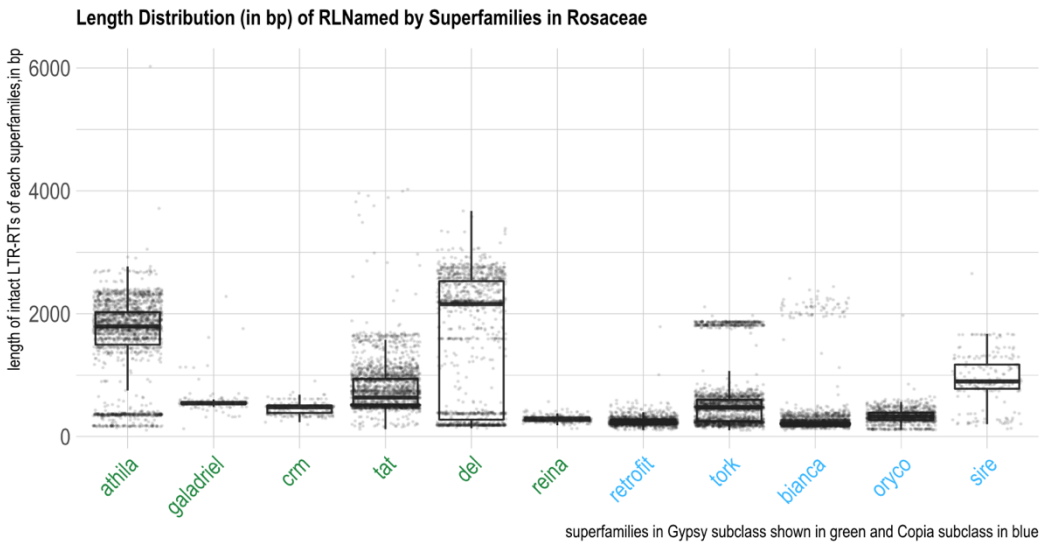

Figure 4.3: **Length distribution of named superfamilies of LTR-RTs in Rosaceae**

In Figure 4.4, the percentages of identified TEs in raw reads are displayed, with Rosaceae aligned on the phylogenetic tree that I constructed. As I only kept the masked reads that were at least 50bp long and with 80% or more identity to our pan-species Rosaceae TE library to calculate the number of hits, I provided a stringent estimate of TE amounts in Rosaceae species that I call the "filtered" result. Table 4.3 also provide the percentage of "raw" masked reads that did not go through length (50 bp) and 80% identity filter, which should provide a better estimate of older TEs, but will also be sensitive to a higher level of false positives. Of the studied species, *Malus domestica* has the highest TE percentages (44%) while *Dryas drummondii* has the lowest (16%). However, because genomes sizes vary greatly among these Rosaceae species, the TE abundances in nucleotide amount is another story. For example, *G. urbanum* has 35% TEs, but those TEs make up 516 Mb given its relatively big genome (1,475 Mb). *M. domestica* has 44% TEs in its ~742 Mb genome, which contributes 326 Mb. For a better understanding of the number of Mb of TEs in each genome, see Figure 4.6.

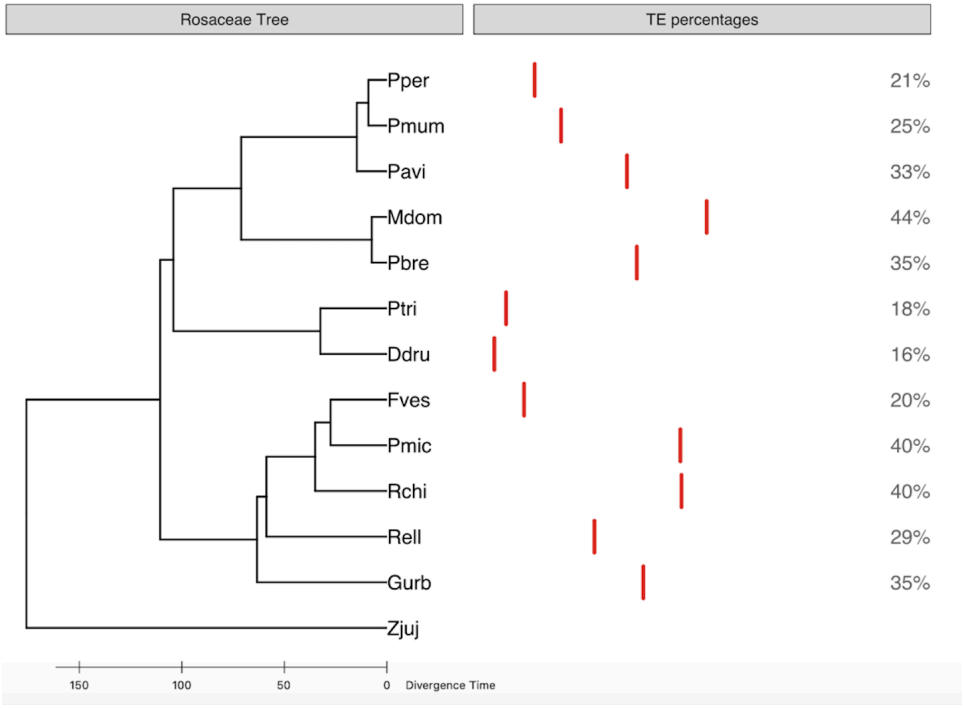

Figure 4.4: **Box plot of the percentages of TEs in Rosaceae**

In Figure 4.4, each screening was repeated with a separately chosen 50 Mb of raw read data, and each repetition resulted in the same results (+/- 0.05%). TE percentages are shown with what appears to be a red line, because the boxes are very thin. Numbers to the right are percentages of the total nuclear genome occupied by identified TE sequences.

Table 4.3: **'Filtered' and 'unfiltered' TE percentages in Rosaceae**

Species Abbr.	Species	TE percentage (default) (%)	TE percentage, after filter (%)	% of "old" TEs	Ratio	% more LTR-RTs[1]
Pper	*Prunus persica*	25	21	4	0.84	4.5
Pmum	*Prunus mume*	29	25	4	0.86	1.7
Pavi	*Prunus avium*	39	33	6	0.85	1.7
Mdom	*Malus domestica*	48	44	4	0.92	1.3
Pbre	*Pyrus brestchneideri*	39	35	4	0.90	1.3
Ptri	*Purshia tridentata*	21	18	3	0.86	1.9
Ddru	*Dryas drummondii*	19	16	3	0.84	2.2
Fves	*Fragaria vesca*	23	20	3	0.87	3.1
Pmic	*Potentilla micrachtha*	45	40	5	0.89	1.8
Rchi	*Rosa chinensis*	44	40	4	0.91	1.2
Rell	*Rubus ellipticus*	34	29	5	0.85	1.3
Gurb	*Geum urbanum*	43	35	8	0.81	0.9
Zjuj	*Ziziphus jujuba*					

TE percentages from RepeatMasker outputs in 50 Mb raw reads with ("filtered") or without ("default") length (50bp or longer) and similarity (80% or more) filter were shown for analyzed species. '% of "old" TEs' is the difference between the default TE percentage and filtered TE percentage. 'Ratio' is 'TE percentage after filter' divided by 'TE percentage (default)' where larger values in 'Ratio' indicate higher percentages of recently inserted TEs. They are argued to be more "recent" because of their higher identity and intactness. [1]The column denotes the percentages of LTR-RTs in the previously masked 50 Mb raw read data further masked by other LTR-RT database (with LTR-RTs from MIPs, *A. thaliana* and *F. vesca*). Species highlighted in blue are those having the "ratio" larger than 0.9.

The proportions of LTR-RTs in three LTR-RT types (Named LTR-RT, SRLX and MRLX) in raw reads are shown in Figure 4.5. This result is also based on the filter of masked reads to be 50 bp or longer that have 80% or more identity to our Rosaceae pan-species LTR-RT library. The vast majority of raw reads were masked by LTR-RTs assigned to a known superfamily and a small proportion of raw reads were masked by LTR-RTs assigned to RLX (SRLX or MRLX) in most of the 12 analyzed Rosaceae species. Exception to this observation are *P. tridentata* with a majority of SRLXs and *P. brestchneideri* with a majority of SRLXs and MRLXs. A high proportion of MRLX or SRLX may at least partly indicate a poor genome assembly that did not allow intact LTR-RTs to often be discovered.

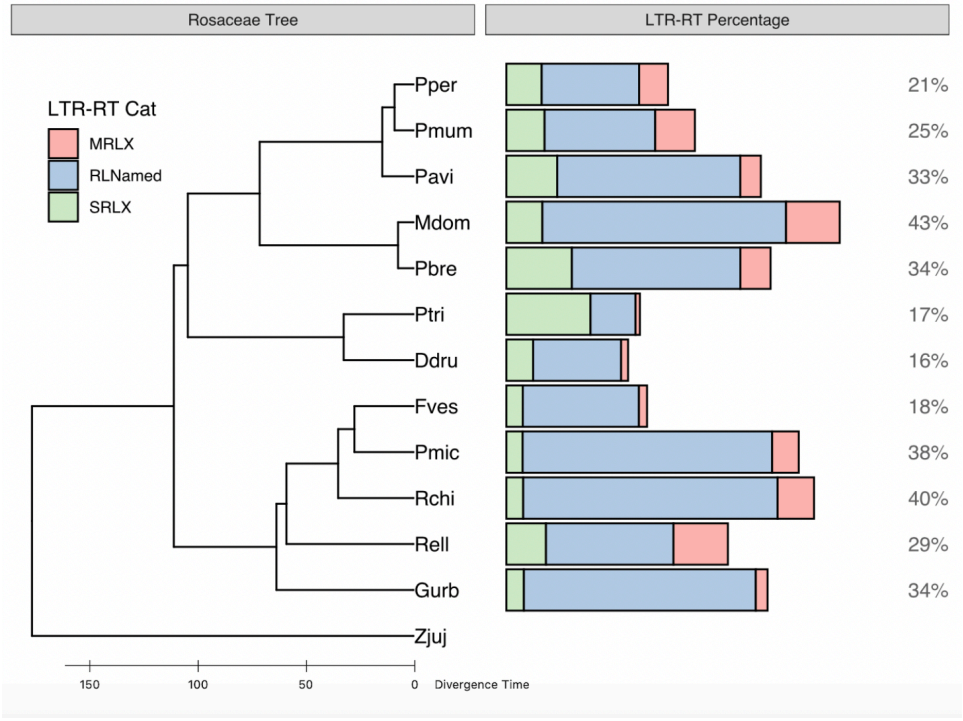

Figure 4.5: **Stacked bar plot of the percentages of LTR-RTs in Rosaceae**

The values are the total percentages of different categories of LTR-RTs in Rosaceae, RLX (MRLX and SRLX) and RLNamed, quantified in 50Mb of raw read data.

The LTR-RT known (i.e., named) superfamily abundance in each genome is plotted in Rosaceae phylogenetic order in Figure 4.6. These abundances are plotted as box volumes proportional to the number of Mb per 1 Cx (monoploid genome) nuclear genome, and labeled by color. I observed that *G. urbanum* and *M. domestica* have the most abundant named LTR-RTs in total, 433 Mb and 231 Mb, respectively. *P. tridentata* and *D. drummondii* have the least abundant named LTR-RTs in total, each contributing ~12 Mb and ~27 Mb. Although the genome size of *G. urbanum* (~1,475 Mb) is close to 7 times that of *P. tridentata* (215 Mb), the contents of total named LTR-RTs in *G. urbanum* is 36 times those of *P. tridentata*.

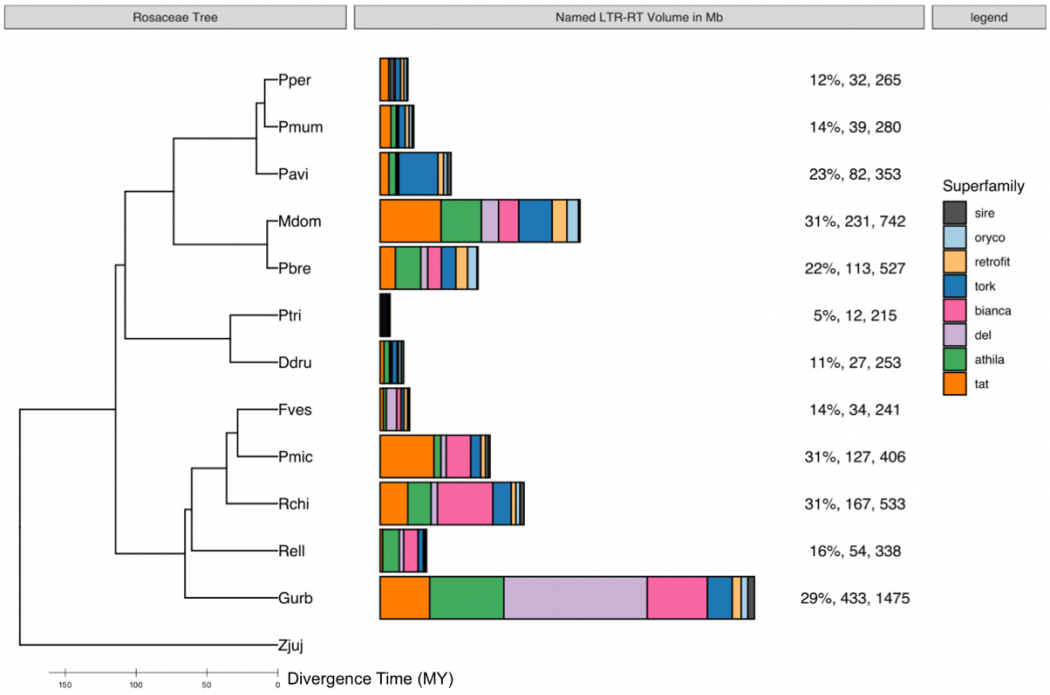

Figure 4.6: **Stacked bar plot of LTR-RT superfamilies in Rosaceae**

From left to right, the values next to the bars denote percentages of total Named superfamilies in the genome, the size of total Named superfamilies (Mb) and genome sizes (Mb).

The most abundant known LTR-RT superfamilies vary by species. The *del* superfamily in *G. urbanum* amplified extensively such that the amount of *del* in *G. urbanum* is nearly the genome size of *M. domestica*. But the activity of the *del* superfamily was low in other analyzed Rosaceae species. In addition, superfamilies *athila*, *bianca* and *tat* partly explain the larger genome of *G. urbanum*.

Fragaria and *Potentilla* lineages diverged from a common ancestor about 24 MYA (164), and the ~165 Mb larger genome of *P. micrantha* compared to *F. vesca* can be explained mostly by amplifications of *tat* and *bianca*, which together make up more than 18% of the *P. micrantha* genome but are almost absent from the genome of *F. vesca*. *Purshia* and *Dryas* diverged from a common ancestor ~30 MYA (163) and they both have

a small genome size and low TE content.

The common ancestors of *Prunus* diverged about 22 MYA (163). In *P. avium*, *tork* amplified and made a 70-Mb expansion of the genome, but this superfamily is less abundant in the other two *Prunus* species. The common ancestor of *Malus* and *Pyrus* began to diverge ~ 5-20 Mya, and the genome size of *M. domestica* is about 1.4 times that of *P. brestchneideri*. The large genome size of *M. domestica* can be explained by the higher abundance of 3 LTR-RT superfamilies, namely *tat*, *athila* and *tork*. Several other specific LTR-RT superfamily abundance differences explain other genome size differences (Figure 4.6).

Phylogenetic signals were tested to see whether the genome quantities of LTR-RT superfamilies distribute randomly or have some phylogenetic correlations. In Table A.23, *del* is seen to be, on average, the third most abundant LTR-RT superfamily in these 12 analyzed Rosaceae species but it is the only superfamily of the top 3 LTR-RT superfamilies showing strong phylogenetic correlations. The two overall most abundant superfamilies showed no phylogenetic signals at all. Some other LTR-RT superfamilies such as *retrofit* and *oryco* also showed strong phylogenetic correlation but they make up a small proportion of the total LTR-RT amounts and are thus not the key player in genome dynamics. The phylogenetic signal tests of LTR-RT superfamily contents indicate a volatile and dynamic pattern of LTR-RT amounts (Mb) on superfamily levels among these analyzed Rosaceae species.

A previous study claimed that the *sire* superfamily had been lost in *P. brestchneideri*, using the TE annotations from its genome assembly (166). However, my analysis (in Figure 4.6) indicates that there are *sire* homologies in *P. brestchneideri* by raw-read quantification. This confirmed again the better sensitivity of raw-read analysis

in finding TEs, rather than genome assemblies.

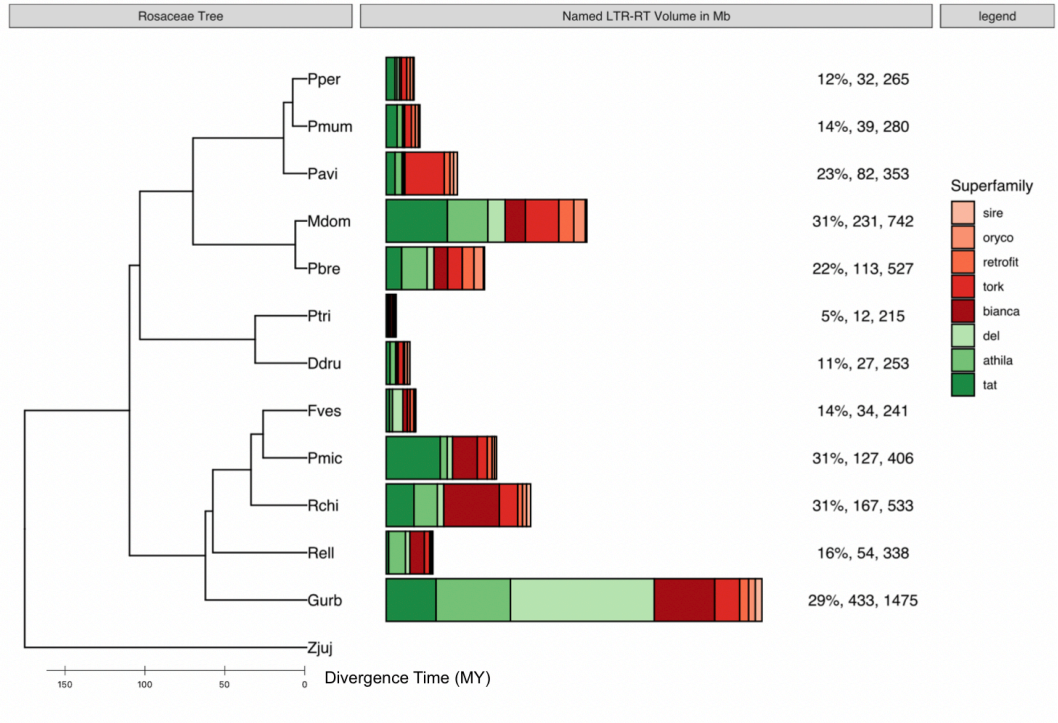

Figure 4.7: **Stacked bar plot of LTR-RT subclasses in Rosaceae**

Superfamilies of Copia subclass are shown in green and of Gypsy subclass in red. The lighter the color, the less the amount. From left to right, the values next to the bars denote percentages of total Named superfamilies, the sizes of total Named superfamilies (Mb) and genome sizes (Mb).

Figure 4.7 is an alternative representation of Figure 4.6 to show the known LTR-RT subclass abundances across *Rosacea*, with Copia in green and Gypsy in red. The top 8 known LTR-RT superfamilies in Rosaceae are *tat*, *athila*, *del*, *bianca*, *tork*, *retrofit*, *oryco* and *sire*. The first 3 are Gypsy elements and the last 5 are Copia. Sometimes Copia dominates and other times Gypsy dominates. In *Prunus*, Copia wins by a narrow margin in *P. persica* and *P. mume* but outscores in *P. avium*. In those 12 analyzed Rosaceae species, it seems that a species with a large genome size such as *M. domestica* and *G. urbanum* see Gypsy as the primary agent of genome expansion.

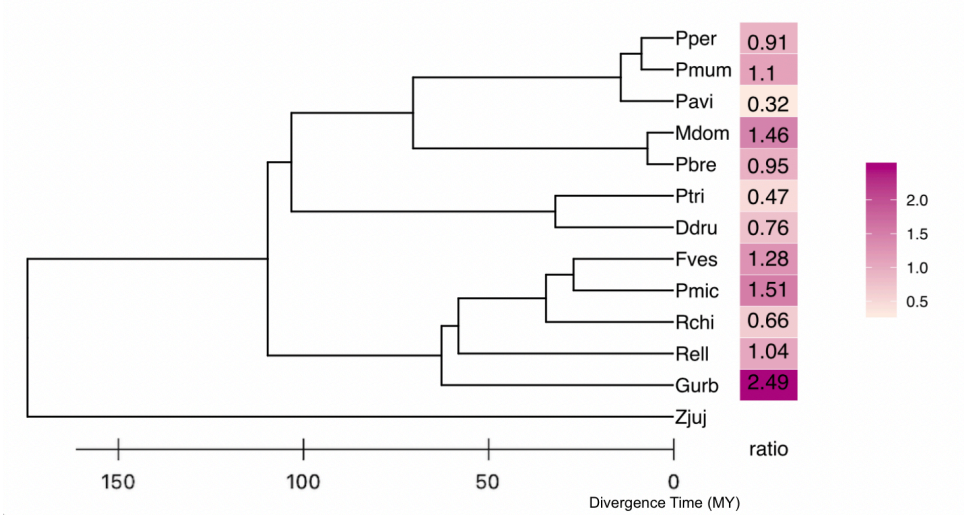

Figure 4.8: **Gypsy-to-Copia ratios in Rosaceae**

The ratios of Gypsy to Copia quantities (in Mb) are shown in Figure 4.8, which is a further representation of Gypsy-vs-Copia dominance in Rosaceae. The Gypsy-to-Copia ratios do not vary a lot across the analyzed Rosaceae species. Species lying in sister branches have close *ratios* such as *P. persica* (0.91) and *P. mume* (1.1), *P. tridentata* (0.47) and *D. drummondii* (0.76), or *F. vesca* (1.28) and *P. micrantha* (1.51).

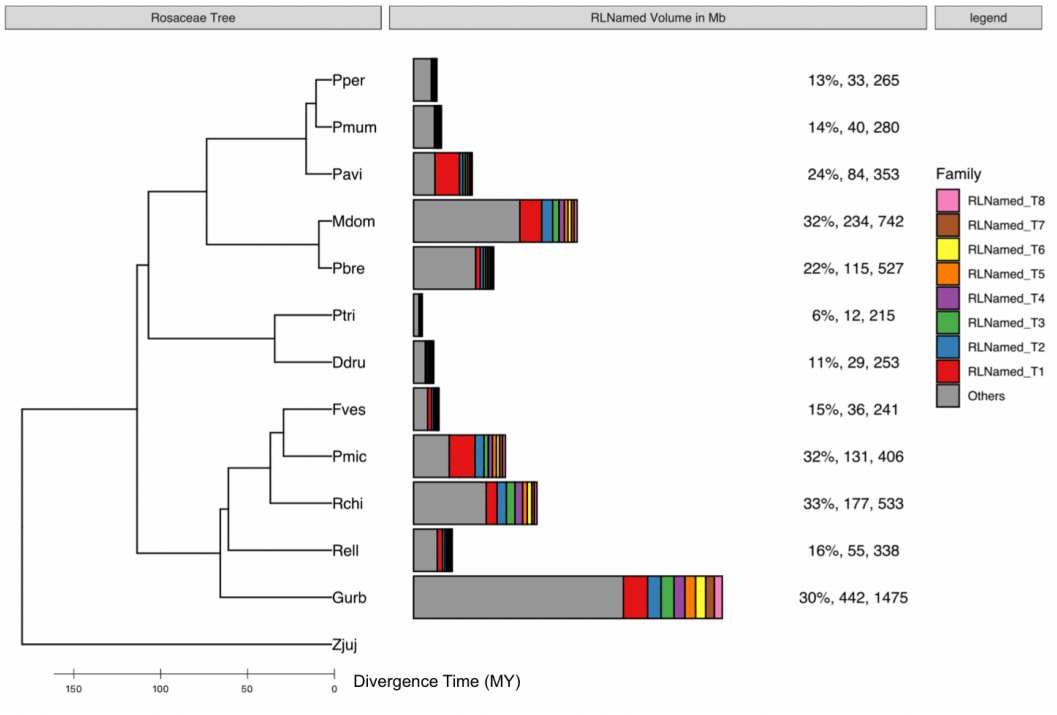

Figure 4.9: **Stacked bar plot of the most abundant Named LTR-RT families in Rosaceae**

Named LTR-RT family contents were quantified in 50 Mb raw reads. 'Others' (gray fill) indicates the combined amounts of less abundant LTR-RT families. From left to right, the values next to the bars denote percentages of top named LTR-RT families, the size of top named LTR-RT families (Mb) and genome sizes (Mb).

Figure 4.9 was plotted based on the abundance of LTR-RT families that could be given some family name rather than RLXs, and it depicts the abundances of the top 8 LTR-RT Named families. These top 8 known LTR-RT families together can make up as much as 40% of all known LTR-RTs in some species. In some genomes, abundances of top known LTR-RT families look uniformly distributed, such as those in *R. chinensis*, in Figure 4.9. The most abundant known LTR-RT families have mean, minimum and maximum genome representations of 3.10%, 0.56% and 9.95% while the second most abundant known LTR-RT families have mean, minimum and maximum genome representations of 1.33%, 0.32% and 3.16%, respectively, across the 12 analyzed

Rosaceae species (Figure A.3).

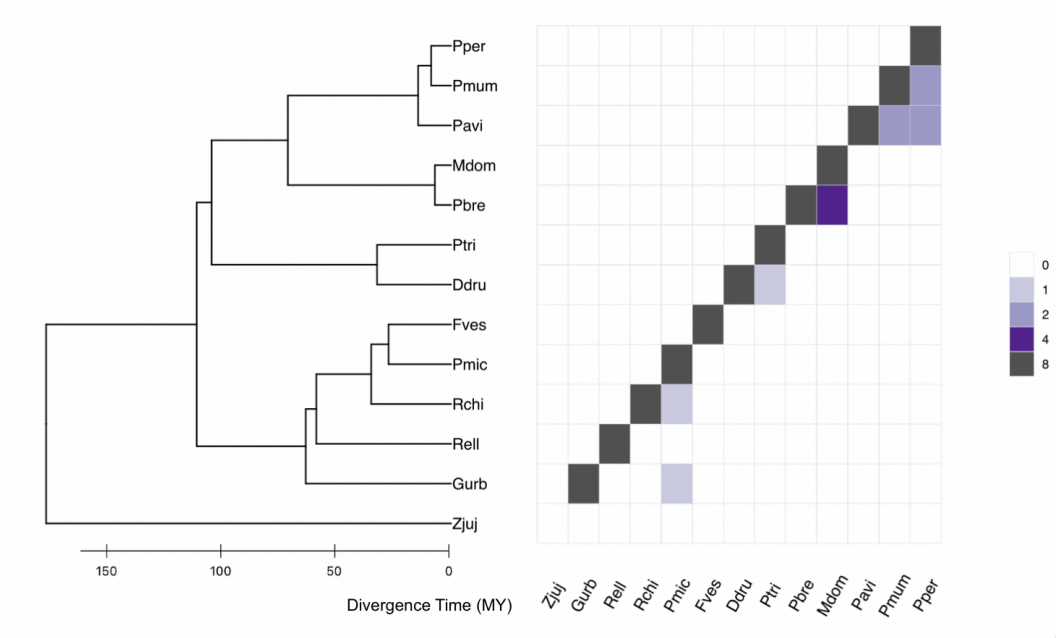

Figure 4.10: **Heatmap of the number of top (most abundant) 8 Named LTR-RT families shared by Rosaceae**

From Figure 4.10, I observed that at most 1 or 2 top LTR-RT families were shared between closely related species, but not among distantly related species. For example, *M. domestica* and *P. brestchneideri* share 4 top known LTR-RT families out of their top 8 known LTR-RT families (in Figure 4.10). An outstanding case of shared top MRLX families was observed between Gurb (*Geum urbanum*) and Pmic (*Potentilla micrachtha*) where they were estimated to have a common ancestor ~100MYA (163).

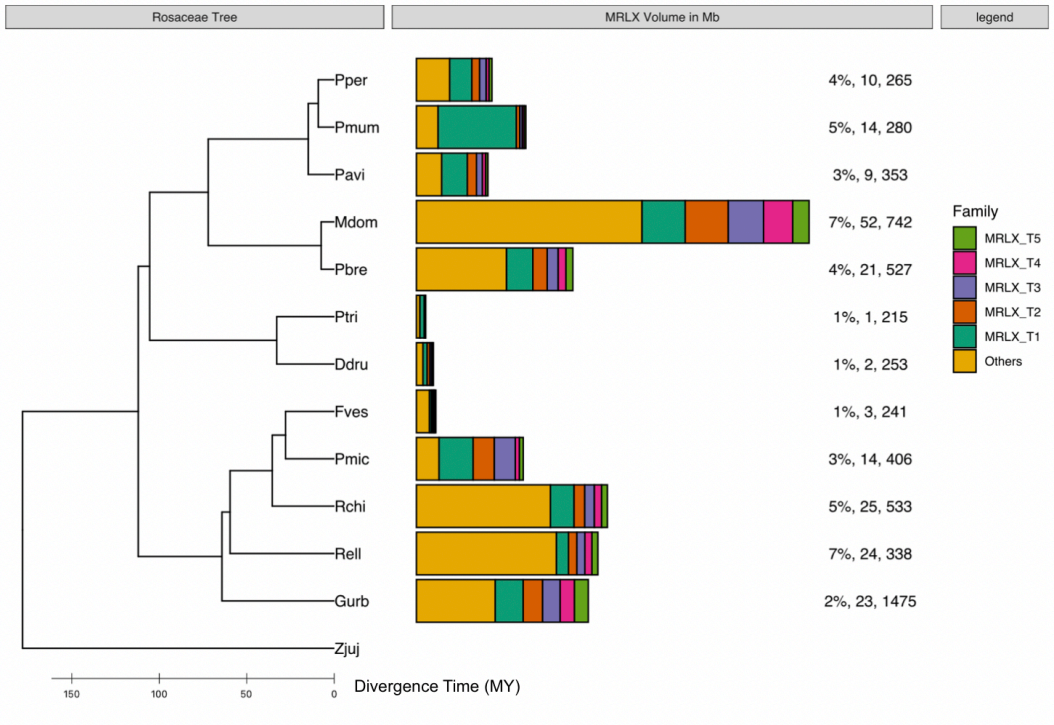

Figure 4.11: **Stacked bar plot of top (most abundant) MRLX families in Rosaceae**

MRLX family contents were quantified in 50 Mb of raw reads and ordered by family amounts (from top most abundant to the fifth most abundant). 'Others' (yellow fill) indicates the combined amounts of less abundant MRLX families. From left to right, the values next to the bars denote percentages of top MRLX families, the sizes of top MRLX families (Mb) and genome sizes (Mb).

From Figure 4.11 and Figure 4.13, we can see RLX (MLRX and SRLX) makes up >10% of the genome in most of the 12 analyzed Rosaceae species. The top 5 most abundant MRLX families in Rosaceae make up ~30% to ~85% of all MRLX families in analyzed Rosaceae. In *P. micrantha*, the top 3 most abundant MRLX family make up ~80% of all MRLX in its genome, which is similar to what is observed in *P. mume* and *P. avium*. In *P. mume*, a MRLX family makes up 3% to 4% of the genome, which is the average percentage (3.1%) of the most abundant known LTR-RT families across the 12 analyzed Rosaceae species (Figure 4.11 and Table A.3). Each of the top 5 most abundant MRLX families contributes quite equally to the genome size of *M. domestica*, with each

of them comprising ~1.5% of the genome, but altogether accounting for 52 Mb due to its large genome.

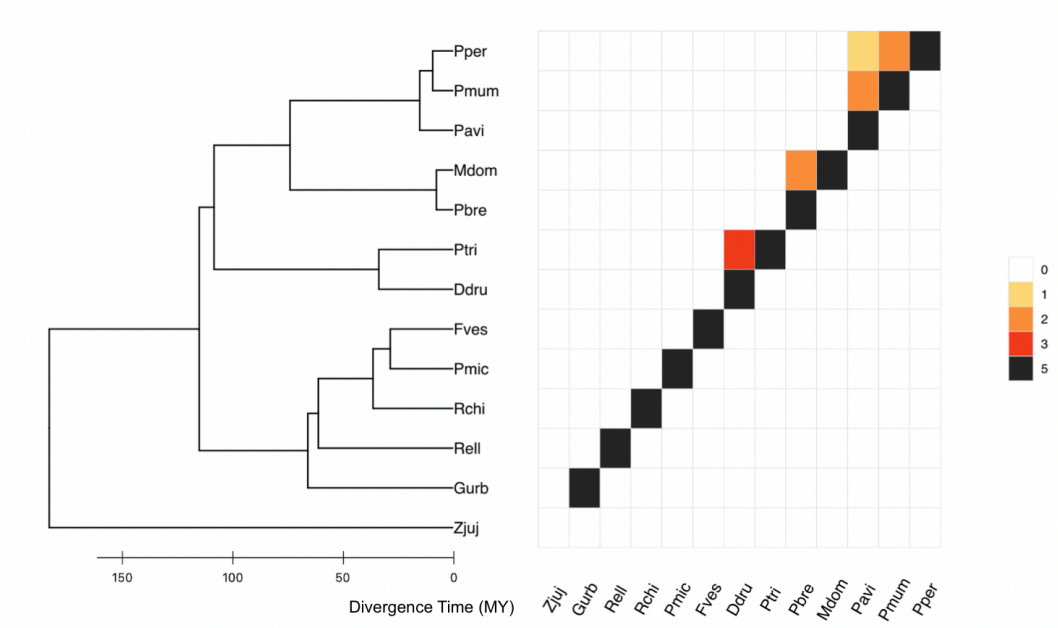

Figure 4.12: **Heatmap of the number of top (most abundant) 5 MRLX families shared by Rosaceae**

In Figure 4.12, the top 5 MRLX families are shared only between *P. tridentata* and *D. drummondii* (with a common ancestor ~30-40 MYA (163)) plus *M. domestica* and *P. bretschneideri* (with a common ancestor ~ 20-30 MYA (163)) and among the 3 studied *Prunus* species (shared a common ancestor ~15-10 MYA (163)). Figure 4.12 confirms again that top MRLX families are seldom shared by distantly related species, similar to what we have observed in the known LTR-RT families in those Rosaceae species.

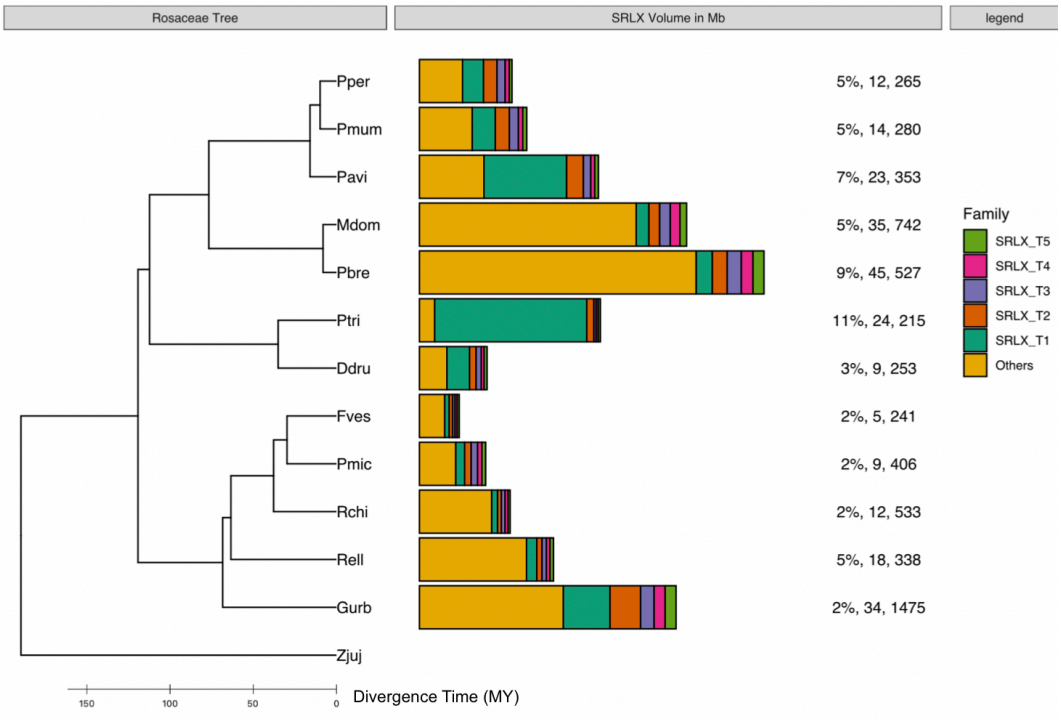

Figure 4.13: **Stacked bar plot of the top (most abundant) SRLX families in Rosaceae**

SRLX family contents were quantified in 50 Mb raw reads and ordered by family amounts (from top most abundant to the fifth most abundant). 'Others' (yellow fill) indicates the combined amounts of less abundant SRLX families. From left to right, the values next to the bars denote percentages of top SRLX families, the sizes of top SRLX families (Mb) and genome sizes (Mb).

P. tridentata contains a few MRLXs (~1% and ~1Mb) and Named LTR-RTs (~6% and 12Mb) (in Figures 4.11 and Figure 4.13). However, a SRLX family in *P. tridentata* comprises about 10% of its genome (Figure 4.13).The LTR composition in *P. tridentata* (6% Named LTR-RTs and 12% RLXs) indicates the abundance and dominance of RLXs over Named LTR-RTs in in *P. tridentata*, which can be further supported by the ratio of RLX to RLNamed (2.16 for *P. tridentata*) shown in Figure B.6. A SRLX family in *P. avium* makes up ~4% of its genome, nearly equaling the total amount of MRLXs in that genome. The abundance of a SRLX in any analyzed Rosaceae species indicates again the failure of genome assembly to place intact LTR-RTs in the genome, even when highly

abundant.

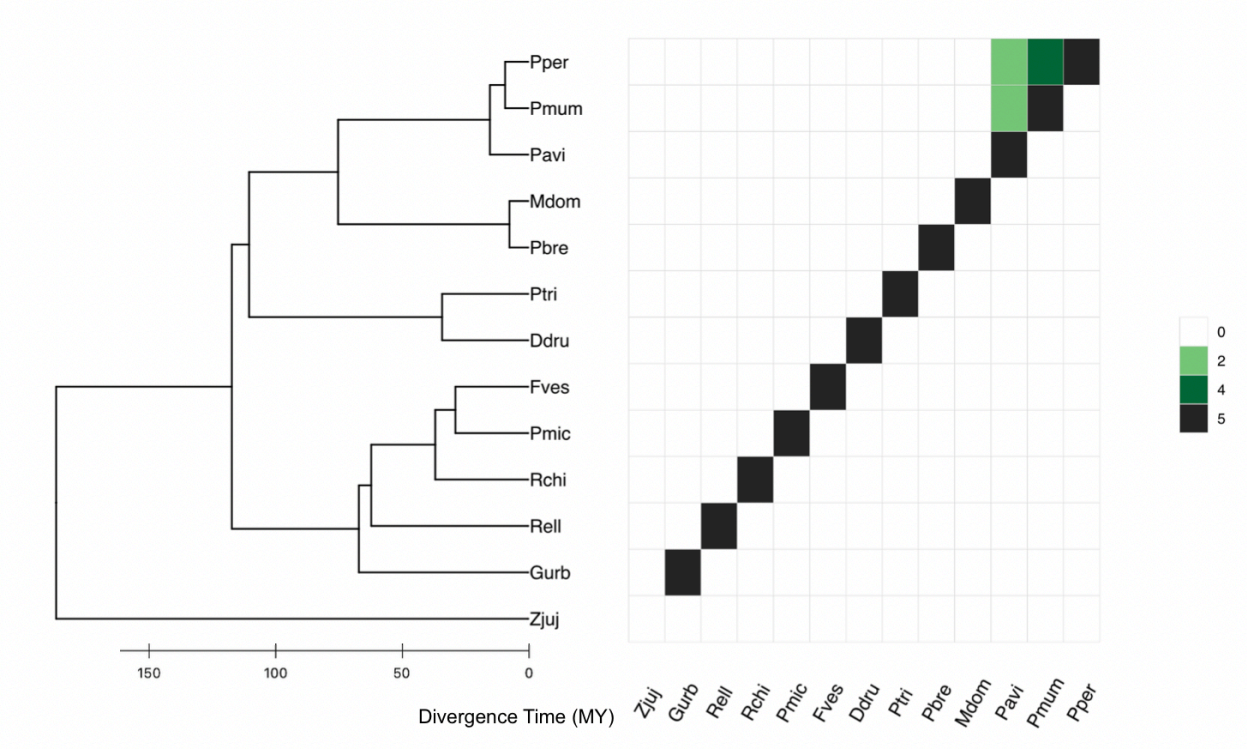

Figure 4.14: **Heatmap of the number of top (most abundant) 5 SRLX families shared by Rosaceae**

In Figure 4.14, the top 5 SRLX families are shared only among the 3 studied *Prunus* species (with a common ancestor ~10-15 MYA (163)). The results in Figure 4.14 confirm again that the top 5 SRLX families are seldom shared by distantly related species, similar to what we have observed in the known LTR-RT and MRLX families in those Rosaceae species.

Chapter 5

Conclusions, Discussion and Future Research

Conclusions and Discussion

We believe that quantifying TEs in raw reads with a TE library that is deep and plant family-specific gets us closer to the real amount of TE contents in the 2 analyzed plant families than previous studies. This should be true for both class I (RNA elements) and class II (DNA elements) TEs, but our project was primarily focused on the most important genome component at the level of genome size, the LTR-RTs. Our discovery process for DNA elements and RNA elements other than LTR-RTs was not terribly sophisticated, relying only on similarities to known TEs of these types in the Munich Information Center for Protein Sequences (MIPs) repeat database (111), TAIR and *F. vesca* TE databases. Hence, we expect that our DNA TE analysis is an under-estimate of true amounts. However, because most DNA TEs and non-LTR retroelements are much smaller than the average LTR-RT, we believe (and others have reported (121) that their representation in genome assemblies is a fairly accurate estimate of their true amounts. In this regard, when we compared DNA element composition by our analysis to previous analyses, we found a great deal less variation between our results and the published results than we did when we made this comparison for LTR-RTs (see Tables 5.1 and 5.2 below).

In general, LTR-RT percentages quantified with raw reads are higher than the values in corresponding published records that are based on genome assemblies (Tables 5.1 and 5.2). We believe this is both a function of our superior TE libraries and of the investigation of raw

reads rather than assemblies.

Table 5.1: **Brassicaceae LTR-RT percentages in published records compared to this analysis**

Species	Reported LTR-RT percentage	LTR-RT percentage in this analysis[1]
B. rapa	<27[(88)]	>38
B. oleracea	32[(89)]	>32
R. sativus	24[(168)]	>35
S. irio	11[(100)]	>33
S. parvula	4[(100)]	>24
E. salsugineum	27[(100)]	>41
T. arvense	<21[(93)]	>59
A. thaliana	<8[(169)]	>9
A. lyrata	<16[(169)]	>34
C. rubella	<5[(100)]	>7
A. arabicum	<10[(100)]	>35

Numbers in superscript parentheses indicate the publications that provided these values.
[1]Given the techniques employed, wherein there may be no intact elements for some families even in the full set of species, and because of the fact that highly degenerated TEs will be missed, these values should all be taken as minima.

Table 5.2: **Rosaceae LTR-RT percentages in published records compared to this analysis**

Species	Reported LTR-RT percentage	LTR-RT percentage in this analysis[1]
P. persica	19[(79)]	>24
P. mume	27[(80)]	>29
M. domestica	31[(82)]	>48
D. drummondii	13[(84)]	>19
F. vesca	16[(85)]	>21
P. brestchneideri	42[(83)]	>40
P. micrantha	24[(86)]	>43
R. chinensis	28[(87)]	>43

Numbers in superscript parentheses indicate the publications that provided these values.
[1]Given the technique employed, wherein there may be no intact elements for some families even in the full set of species, and because of the fact that highly degenerated TEs will be missed, these values should all be taken as minima.

It is interesting to note that some published LTR-RT content results are very close to

our own. These include Arabidopsis, which is well known for the quality of its genome annotation. Some values differ by more than 2-fold in the Brassicaceae, and this correlates mostly with the quality of genome assembly and how TEs are identified in that species. Overall, our values agree more closely with the Rosaceae published values, which may be because labs had different level of expertise of annotating TEs. For example, TE annotations of *Fragaria vesca* and *Pyrus brestchneideri* were done by Dr Hao Wang from Dr Bennetzen's lab and my reported findings of those two species were very close to what he had reported.

Figure 5.1: **Scatterplot of genome sizes (Mb) vs LTR-RT amount (Mb)**

Genome sizes (Mb) were plotted against "filtered" LTR-RT amounts (Mb). A regression line (in blue) was fitted against the data and the gray shadow covers the 95% C.I. for the regression.

In a few cases, published annotations indicated more LTR-RTs than were calculated in our studies. This is probably an outcome of different stringencies applied, which again emphasizes the value of performing comparative analyses only when the annotation styles are identical. In this regard, with consistent annotation styles, one can now plot genome size (in Mb) versus LTR-RT content (in Mb) for all species in each of these two families (Figure 5.1). This result shows that genome size differences strongly correlate with LTR-RT content with a

R^2 score of 0.92 (in Table A.16), as previously published (59, 170).

Outcomes of this analysis are the result of tradeoffs between false positives and false negatives. Given that LTR-retriever was claimed to be a decent software to remove false positives with multiple steps which had been described in methods, this analysis should not suffer over-estimation of LTR-RTs in raw reads. After 50 Mb of raw read data was masked by the pan-species library of that plant family, the masked 50 Mb of raw read data was taken and re-masked with LTR-RTs mainly from MIPs database, there are an average of 1-2 % more LTR-RTs in 50 Mb of raw read data that can be further masked by LTR-RTs from MIPs database (shown in Table 3.3 and Table 4.3) though the LTR-LTR quantity in the MIPs database is about 4 times larger than the LTR-RTs we have in our pan-species LTR-RT library. Therefore, this study suffers less false negative predictions than do previous studies. LTR-RT measurement of percentages in this study is a deep analysis. As we have discussed earlier, on the way to transform LTR-RT percentages to nucleotide amount, we took the estimated genome sizes of these species, which can render +/- 20% bias in LTR-RT nucleotide amounts. In Table A.22, percentages of Class II TEs are presented and most of them had a reasonably low amount. We took class II TEs from MIPs, *A. thaliana* and *F.vesca* to be the library and tried to mask them in raw read data. Finally, we were able to find a much higher percentage of Class II TEs in raw read data of *A. thaliana* and *F. vesca* than other species. Likewise, generally more Class II TEs in raw reads were quantified if that species was closer to *A. thalian* or *F. vesca*. Therefore, as for the analysis of LTR-RTs in raw read data, Class II TE measurement is also species-dependent. We expect more false negatives than false positives in Class II TE analysis in this study.

It is worth noting that 90% of the analyses described above were done after filtering the

raw reads to be 50 bp or longer and to have 80% or more identity to what we have in the plant-family-specific pan-species TE library, which is a stringent but safe estimate. Therefore, we predict that we will miss >3% of the TE homologies in raw reads, because the homology was shorter than 50 bp and had <80% identify to any TE in the corresponding TE library. Possible missing TE percentages are shown in Table 3.3 and Table 4.3. Some masked raw reads that couldn't pass the filter are likely from old TEs, because old TEs will be more divergent, and some will be where <50bp of the end of a TE were present in a read.

In a few cases, I found a major LTR-RT family that only had one or a few intact LTR-RTs in the entire plant family TE database, so that means there could have been cases with zero intact LTR-RTs for the entire plant family. So, members of such families, even if highly repetitive, would not have been recorded as LTR-RTs in my analysis. This was one of the reasons that a pan-species TE library was of value, in the hope that at least one of these species would have at least one copy of an intact LTR-RT in its assembly.

It is interesting to see whether this divergence property (filtered versus default (with RepeatMasker default settings[#])) can be used to detect old TEs. From Table 3.3 and Table 4.3, the ratios suggest that many species like *S. irio* and *C. hirsuta* have a high proportion of old TEs, suggesting either slow DNA removal and/or a lack of recent TE activity.

I tried to make use of those masked reads that pass RepeatMasker default settings but didn't pass the length or identity filter I used to test the effectiveness of pan-species library screening, based on the assumption that these "unfiltered reads" are mainly from old TEs. The "unfiltered" output employed the default parameters (having the RepeatMasker cutoff score 225), which was a routine strategy in published genome annotations. The TE family of each

[#] Repeat masker of default settings outputs local alignments of DNA sequences to TEs with a cutoff score 225.

bp/piece of a raw read was decided in my study by which TE it was masked against in the pan-species library, with the best match scores to each intact LTR-RT in the library routinely from its original species source.

In order to study the effectiveness of a pan-species TE library, I worked on raw reads masked by LTR-RTs from a pan-species LTR-RT library that are 50 bp or longer because we know the species source of each intact LTR-RT. Then, I designated raw reads RepeatMasker masked with <80% identity as raw reads from old LTR-RTs, those masked with 80% to 95% identity as raw reads from mid-aged LTR-RTs, and those with 95% or more identity as raw reads from young LTR-RTs, all calculated in nucleotide amount (bp). These designations could be investigated for how much LTR-RT sequence they masked from their own species and from all other species (shown in Tables A.7 to A.14). It was expected, of course, that within-species designations would be highly enriched for the youngest category. This result was observed (Table A.8 and Table A.12). Regarding masked raw reads from mid-aged LTR-RTs of a species, LTR-RTs in the pan-species library of the same species are marginally more abundant than those from other species. However, for masked raw reads from "old" LTR-RTs of a species, LTR-RTs in the library from some closely related species were more abundant than those from the same species by a great deal, summarized in Table A.19 and A.20.

The proportions of raw reads masked by LTR-RTs in its pan-species library with 95% or more identities are extremely high for some specific species, suggesting a very high level of recent LTR-RT amplification. For example, *E. salsugineum* has 59% of masked reads with 95% or more identity to its LTR-RT library, 97% of which were masked by intact LTR-RTs from its own genome assemblies and only 3% by intact LTR-RTs from other genome assemblies in the pan-species library.

Three species have more than 10% of their masked raw reads with 95% or more identity to LTR-RTs in the library from other species. They are *P. mume*, *A. thaliana* and *C. rubella* with proportions of 11%, 10% and 14%, respectively (those were highlighted in blue in Table A.6). These are potential cases for horizontal transfer of LTR-RTs (171). By checking Table A.10 and Table A.14 for analyzed Brassicaceae and Rosaceae species, respectively, those species having their intact LTR-RTs contribute greatly to the masking of raw reads from recent LTR-RTs of another species include *B. rapa* for *R. sativus* (40%)*, C. rubella* (42%), *S. parvula* (37%), *C. hirsuta* (32%) and *P. persica* for *P. mume* (49%).

On average, the most abundant LTR-RT superfamilies in Brassicaceae are *athila* and *retrofit* and those in Rosaceae are *tat* and *athila/bianca*. One from Gypsy and the other from Copia makes up the top 2 most abundant known LTR-RT families in both Rosaceae and Brassicaceae. Some LTR-RT superfamilies appear to have stayed fairly silent in most species but amplified tremendously in one lineage. Many of these amplifications by one or a few families seems to have at least doubled genome sizes, such as *bianca* in *B. oleracea*, *del* in *A. alpina*, *galadriel* in *B. nigra* , *retrofit* in *A. lyrata, tork* in *P. avium* and *del* in *G. urbanum*.

The ratios of Gypsy to Copia vary more in analyzed Brassicaceae species than in analyzed Rosaceae species. Some Rosaceae species within a sister branch but with a common ancestor >20 MYA shared similar Gypsy-to-Copia ratios (0.47 in *P. tridentata* vs 0.76 in *D. drummondii,* 1.28 in *F. vesca* vs 1.51 in *P. micrantha*), as shown in Figure 4.8. Other closely related Brassicaceae species (with a common ancestor <10 MYA) have very distinct Gypsy-to-Copia ratios (7.45 in *E. salsugineum* vs 4.17 in *T. arvense,* 2.34 in *A. thaliana* vs 0.68 in *A. lyrata*) in Figure 3.8.

A two-sample t-test of Gypsy-to-Copia ratios in Table A.15 indicates that we can accept

the hypothesis that the mean ratio of Gypsy-to-Copia in analyzed Brassicaceae species is higher than that in analyzed Rosaceae species. This is because, as shown by another two-sample t-test that the mean Copia content in Table A.15 is higher (in Mb) in Rosaceae than in Brassicaceae.

In Table 3.3 and Table 4.3, the percentages of mid-aged and young TEs in each sampled species are shown. On average, analyzed Rosaceae species have larger proportions of ancient TEs than those in analyzed Brassicaceae species. From Table A.15, we can reject the null hypothesis and accept the alternative hypothesis that the mean percentage of mid-aged and young TEs in analyzed Rosaceae species is larger than that in analyzed Brassicaceae species at a 0.01 significance level. Five analyzed Brassicaceae species have 22% to 29% old TEs while none of them has less than 10% old TEs of all TEs. Three analyzed Rosacea species have 8% to 10% old TEs, while no studied Rosacea has >19% old TEs (Table 3.3, Table 4.3 Figure B.1 and Figure B.2). For those 3 species having a small proportion of old TEs, two of them lie in a sister branch (*M. domestica* and *P. brestchneideri)*. Similarly, for the 5 analyzed Brassicaceae species with >20% old TEs, four of them lie side by side in clade A of their phylogenetic tree. These results suggest a generally higher level of recent LTR-RT activity in the Rosaceae than in the Brassicaceae, and also suggest that low levels or high levels of activity are traits shared by sister lineages.

As I stated earlier, TE analysis of a species of your interest totally based on a TE library taken from another species may not be satisfactory because there may be species-specific TEs, at least as defined by any homology criterion. Hence, species-specific TE libraries are necessary to conduct a reliable TE analysis. My dissertation performed a well-designed quantification of LTR-RTs in the analyzed Brassicaceae and Rosaceae species. I found intact LTR-RTs in each individual genome assembly and eventually constructed two pan-species

libraries, one for Brassicaceae and one for Rosaceae pan-species. However, the TEs other than LTR-RTs in the TE library are taken from the MIPS repetitive element database (111), *A. thaliana* TEs (used for Brassicaceae pan-species library only) and an *F. vesca* TE database (used for Rosaceae pan-species library only). Therefore, the quantifications of those TEs other than LTR-RTs is less precise and more biased than the quantifications of LTR-RTs, partly because the more closely-related the species in Brassicaceae/Rosaceae is to *A. thaliana*/*F. vesca*, the more non LTR-RT TEs will be found.

A few of the most abundant known LTR-RT families (RLNamed) plus the most abundant unknown LTR-RT (RLX) families make up somewhere near 50% of the LTR-RTs in the analyzed Brassicaceae and Rosaceae species. Then, the hundreds or thousands of the less abundant LTR-RT contents (RLNamed and RLX families) make up the other ~50%. Hence, I conclude that these top (most abundant) few families of LTR-RTs (RLNamed and RLX) are the major factors in genome size variation in the analyzed Brassicaceae and Rosaceae species.

It was shown repeatedly that the top known (RLNamed) or unknown (MRLX or SRLX) LTR-RT families are seldom shared by different species. A pair of close species lying in a sister branch with a common ancestor from several to a dozen MYA will sometimes share 1 or 2 top LTR-RT families. This indicates two things: (1) LTR-RTs are evolving fast and (2) that there is no robust pattern for which LTR-RT families will be the major determinants of genome size in any broad lineage of these two plant families. The majority of LTR-RTs classified to the same known LTR-RT superfamily no longer share similarity in their LTRs. However, there is an interesting case in analyzed Brassicaceae species that the LTR-RT family *galadriel_1*, which belongs to the superfamily *galadriel*, is one of the top 8 known LTR-RT families in 12 out of 17 Brassicaceae species. We were only able to find 3 intact LTR-RTs in the family

*galadriel*_1 and 4 intact LTR-RTs in the superfamily *galadriel* across the 17 analyzed Brassicaceae species, indicating how poor the traditional genome assemblies can be for finding even the most abundant repeats.

Future Research

I evaluated mating system, genome size, karyotype, and life cycle (collected in Table A.17) for their possible association with total TE contents or contents of specific TEs. Other factors such as winter hardiness and deciduousness are hard to acquire remotely because of either uncertainty or unavailability of the data. We fitted a linear model (in Table A.16) to predict the total LTR-RT amount (in Mb) with a high R^2 score where genome size and Gypsy-to-Copia ratio are significant predictors. Similar analysis did not show any strong associations with any other factors (life cycle, karyotype and self-compatibility). Future studies to correlate lineage biological properties with TE abundances and/or amplification history are warranted.

It was well-known that the amounts of LTR-RTs or TEs are highly correlated with genome sizes, therefore it is not a surprise that genome size is a significant predictor to predict LTR-RT contents (in Mb) in those species (and see Figure 5.1). But I was able to discover, at least in these two families, that the Gypsy-to-Copia ratio is also a significant predictor. I was able to reject the simpler model having only the genome size as a single predictor at the significance level 0.1 (in Table A.17), which makes sense because Gypsy LTR-RTs are, more often than not, both longer (in these species) and more abundant than Copia LTR-RTs.

Brassicaceae has 4,060 accepted species in 272 genera and Rosaceae has 4,828 accepted species in 91 genera. In this study, we analyzed 17 species of 12 genera in Brassicaceae and 12 species of 10 genera in Rosaceae. The analyzed species in this study had their genomes sequenced primarily because of their scientific and economic importance to human beings, so

the sampling was biased from the start. We were not able to sample species randomly from the phylogeny but had to select species which have genome assemblies available. Anyone interested in exploring the correlation between TE contents and plant characteristics would be best served by using a phylogenetic and/or lifestyle criteria as the major determinant for selecting the species to be studied (172, 173).

REFERENCES

1. Karl R, Koch MA. A world-wide perspective on crucifer speciation and evolution: phylogenetics, biogeography and trait evolution in tribe Arabideae. Annals of Botany. 2013;112(6):983-1001.

2. Jordon-Thaden IE, Al-Shehbaz IA, Koch MA. Species richness of the globally distributed, arctic–alpine genus *Draba L.*(Brassicaceae). Alpine Botany. 2013;123(2):97-106.

3. Dodsworth S, Leitch AR, Leitch IJ. Genome size diversity in angiosperms and its influence on gene space. Current opinion in genetics & development. 2015;35:73-8.

4. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 2011;43(5):476-81.

5. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics. 2007;8(12):973.

6. Greenblatt IM. A chromosome replication pattern deduced from pericarp phenotypes resulting from movements of the transposable element, Modulator, in maize. Genetics. 1984;108(2):471-85.

7. Chen J, Greenblatt IM, Dellaporta SL. Transposition of Ac from the P locus of maize into unreplicated chromosomal sites. Genetics. 1987;117(1):109-16.

8. Eickbush TH, Jamburuthugoda VK. The diversity of retrotransposons and the properties of their reverse transcriptases. Virus research. 2008;134(1-2):221-34.

9. Xiong Y, Eickbush TH. Origin and evolution of retroelements based upon their reverse transcriptase sequences. The EMBO journal. 1990;9(10):3353-62.

10. Schmidt T. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant molecular biology. 1999;40(6):903-10.

11. Nouroz F, Noreen S, Khan MF, Ahmed S, Heslop-Harrison JP. Identification and characterization of mobile genetic elements LINEs from *Brassica* genome. Gene. 2017;627:94-105.

12. Wenke T, Holtgräwe D, Horn AV, Weisshaar B, Schmidt T. An abundant and heavily truncated non-LTR retrotransposon (LINE) family in *Beta vulgaris*. Plant molecular biology. 2009;71(6):585-97.

13. Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. Nature Reviews Genetics. 2002;3(5):329.

14. Kapitonov VV, Jurka J. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences. 2001;98(15):8714-9.

15. Yang L, Bennetzen JL. Structure-based discovery and description of plant and animal Helitrons. Proceedings of the National Academy of Sciences. 2009;106(31):12832-7.

16. Robillard É, Le Rouzic A, Zhang Z, Capy P, Hua-Van A. Experimental evolution reveals hyperparasitic interactions among transposable elements. Proceedings of the National Academy of Sciences. 2016;113(51):14763-8.

17. Lisch D, Bennetzen JL. Transposable element origins of epigenetic gene regulation. Current opinion in plant biology. 2011;14(2):156-61.

18. Grandbastien M-A, Audeon C, Bonnivard E, Casacuberta J, Chalhoub B, Costa A-P, et al. Stress activation and genomic impact of Tnt1 retrotransposons in Solanaceae. Cytogenetic and genome research. 2005;110(1-4):229-41.

19. Jardim SS, Schuch AP, Pereira CM, Loreto ELS. Effects of heat and UV radiation on the mobilization of transposon *mariner-Mos1*. Cell Stress and Chaperones. 2015;20(5):843-51.

20. Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, et al. Transposable elements contribute to activation of maize genes in response to abiotic stress. PLoS genetics. 2015;11(1):e1004915.

21. Liu B, Wendel JF. Retrotransposon activation followed by rapid repression in introgressed rice plants. Genome. 2000;43(5):874-80.

22. Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH. Genome evolution of wild barley (*Hordeum spontaneum*) by *BARE-1* retrotransposon dynamics in response to sharp microclimatic divergence. Proceedings of the National Academy of Sciences. 2000;97(12):6603-7.

23. Jiang N, Bao Z, Zhang X, Hirochika H, Eddy SR, McCouch SR, et al. An active DNA transposon family in rice. Nature. 2003;421(6919):163.

24. Hirochika H, Sugimoto K, Otsuki Y, Tsugawa H, Kanda M. Retrotransposons of rice involved in mutations induced by tissue culture. Proceedings of the National Academy of Sciences.

1996;93(15):7783-8.

25. Pietzenuk B, Markus C, Gaubert H, Bagwan N, Merotto A, Bucher E, et al. Recurrent evolution of heat-responsiveness in Brassicaceae *COPIA* elements. Genome biology. 2016;17(1):209.

26. McClintock B. The significance of responses of the genome to challenge. 1983.

27. Grandbastien M-A, Lucas H, Morel J-B, Mhiri C, Vernhettes S, Casacuberta JM. The expression of the tobacco Tnt1 retrotransposon is linked to plant defense responses. Evolution and Impact of Transposable Elements: Springer; 1997. p. 241-52.

28. Naito K, Cho E, Yang G, Campbell MA, Yano K, Okumoto Y, et al. Dramatic amplification of a rice transposable element during recent domestication. Proceedings of the National Academy of Sciences. 2006;103(47):17620-5.

29. Bennetzen JL. Transposable elements, gene creation and genome rearrangement in flowering plants. Current opinion in genetics & development. 2005;15(6):621-7.

30. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431(7008):569.

31. Du C, Fefelova N, Caronna J, He L, Dooner HK. The polychromatic Helitron landscape of the maize genome. Proceedings of the National Academy of Sciences. 2009;106(47):19916-21.

32. Yang L, Bennetzen JL. Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proceedings of the National Academy of Sciences. 2009;106(47):19922-7.

33. Akhunov ED, Akhunova AR, Dvorak J. Mechanisms and rates of birth and death of dispersed duplicated genes during the evolution of a multigene family in diploid and tetraploid wheats. Molecular biology and evolution. 2006;24(2):539-50.

34. Kim S, Park J, Yeom S-I, Kim Y-M, Seo E, Kim K-T, et al. New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. Genome biology. 2017;18(1):210.

35. Jiang S-Y, Ramachandran S. Genome-wide survey and comparative analysis of LTR retrotransposons and their captured genes in rice and sorghum. PLoS One. 2013;8(7):e71118.

36. Ma J, Bennetzen JL. Rapid recent growth and divergence of rice nuclear genomes. Proceedings of the National Academy of Sciences. 2004;101(34):12404-10.

37. Hawkins JS, Proulx SR, Rapp RA, Wendel JF. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proceedings of the National Academy of Sciences. 2009;106(42):17811-6.

38. Vitte C, Panaud O, Quesneville H. LTR retrotransposons in rice (*Oryza sativa,* L.): recent burst amplifications followed by rapid DNA loss. BMC genomics. 2007;8(1):218.

39. Gazzani S, Gendall AR, Lister C, Dean C. Analysis of the molecular basis of flowering time variation in Arabidopsis accessions. Plant physiology. 2003;132(2):1107-14.

40. Kobayashi S, Goto-Yamamoto N, Hirochika H. Retrotransposon-induced mutations in grape skin color. Science. 2004;304(5673):982-.

41. Lockton S, Gaut BS. The contribution of transposable elements to expressed coding sequence in *Arabidopsis thaliana*. Journal of molecular evolution. 2009;68(1):80-9.

42. Muterko A, Balashova I, Cockram J, Kalendar R, Sivolap Y. The new wheat vernalization response allele *Vrn-D1s* is caused by DNA transposon insertion in the first intron. Plant Molecular Biology Reporter. 2015;33(2):294-303.

43. Kim H-Y, Schiefelbein JW, Raboy V, Furtek DB, Nelson OE. RNA splicing permits expression of a maize gene with a defective Suppressor-mutator transposable element insertion in an exon. Proceedings of the National Academy of Sciences. 1987;84(16):5863-7.

44. Giroux MJ, Clancy M, Baier J, Ingham L, McCarty D, Hannah LC. *De novo* synthesis of an intron by the maize transposable element *Dissociation*. Proceedings of the National Academy of Sciences. 1994;91(25):12150-4.

45. Chuong EB, Elde NC, Feschotte C. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics. 2017;18(2):71.

46. Studer A, Zhao Q, Ross-Ibarra J, Doebley J. Identification of a functional transposon insertion in the maize domestication gene *tb1*. Nature genetics. 2011;43(11):1160.

47. Miura A, Yonebayashi S, Watanabe K, Toyama T, Shimada H, Kakutani T. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. Nature. 2001;411(6834):212.

48. McClintock B. The origin and behavior of mutable loci in maize. Proceedings of the National Academy of Sciences. 1950;36(6):344-55.

49. Grant V. Plant speciation. Columbia University Press; 1981.

50. Shan X, Liu Z, Dong Z, Wang Y, Chen Y, Lin X, et al. Mobilization of the active MITE transposons mPing and Pong in rice by introgression from wild rice (*Zizania latifolia* Griseb.). Molecular Biology and Evolution. 2005;22(4):976-90.

51. Shirasu K, Schulman AH, Lahaye T, Schulze-Lefert P. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. Genome Research. 2000;10(7):908-15.

52. Lim JK, Simmons MJ. Gross chromosome rearrangements mediated by transposable elements in *Drosophila melanogaster*. Bioessays. 1994;16(4):269-75.

53. Ogiwara I, Miya M, Ohshima K, Okada N. Retropositional parasitism of SINEs on LINEs: identification of SINEs and LINEs in elasmobranchs. Molecular biology and evolution. 1999;16(9):1238-50.

54. Wicker T, Guyot R, Yahiaoui N, Keller B. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. Plant Physiology. 2003;132(1):52-63.

55. Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, et al. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome research. 2006;16(10):1262-9.

56. Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J. Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome biology and evolution. 2011;3:219-29.

57. Hawkins JS, Kim H, Nason JD, Wing RA, Wendel JF. Differential lineage-specific amplification of transposable elements is responsible for genome size variation in *Gossypium*. Genome research. 2006;16(10):1252-61.

58. Gao D, Zhao D, Abernathy B, Iwata-Otsubo A, Herrera-Estrella A, Jiang N, et al. Dynamics of a novel highly repetitive CACTA family in common bean (*Phaseolus vulgaris*). G3: Genes, Genomes, Genetics. 2016;6(7):2091-101.

59. Vitte C, Bennetzen JL. Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proceedings of the National Academy of Sciences. 2006;103(47):17638-43.

60. Orel N, Puchta H. Differences in the processing of DNA ends in *Arabidopsis thaliana* and tobacco: possible implications for genome evolution. Plant molecular biology. 2003;51(4):523-31.

61. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends in genetics. 2008;24(3):142-9.

62. Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, et al. De novo genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). Nature biotechnology. 2011;29(6):521.

63. Nelson MG, Linheiro RS, Bergman CM. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. G3: Genes, Genomes, Genetics. 2017;7(8):2763-78.

64. Robb SM, Lu L, Valencia E, Burnette JM, Okumoto Y, Wessler SR, et al. The use of RelocaTE and unassembled short reads to produce high-resolution snapshots of transposable element generated diversity in rice. G3: Genes, Genomes, Genetics. 2013;3(6):949-57.

65. van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The third revolution in sequencing technology. Trends in Genetics. 2018;34(9):666-81.

66. Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon J-M, et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS genetics. 2009;5(11):e1000732.

67. Yin H, Liu J, Xu Y, Liu X, Zhang S, Ma J, et al. *TARE1*, a mutated *Copia*-like LTR retrotransposon followed by recent massive amplification in tomato. PloS one. 2013;8(7):e68587.

68. Lerat E. Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity. 2010;104(6):520-33.

69. Ou S, Jiang N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant physiology. 2018;176(2):1410-22.

70. SanMiguel P, Tikhonov A, Jin Y-K, Motchoulskaia N, Zakharov D, Melake-Berhan A, et al. Nested retrotransposons in the intergenic regions of the maize genome. Science. 1996;274(5288):765-8.

71. Maumus F, Quesneville H. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. Nature communications. 2014;5:4104.

72. Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Research. 2009;19(2):243-54.

73. Estep M, DeBarry J, Bennetzen J. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. Heredity. 2013;110(2):194.

74. McClintock B. The suppressor-mutator system of control of gene action in maize. The suppressor-mutator system of control of gene action in maize. 1957.

75. Lockton S, Gaut BS. The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC evolutionary biology. 2010;10(1):10.

76. Zhang X, Wessler SR. Genome-wide comparative analysis of the transposable elements in the related species *Arabidopsis thaliana* and *Brassica oleracea*. Proceedings of the National Academy of Sciences. 2004;101(15):5589-94.

77. Belyayev A, Kalendar R, Brodsky L, Nevo E, Schulman AH, Raskina O. Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. Mobile DnA. 2010;1(1):6.

78. Huang X, Lu G, Zhao Q, Liu X, Han B. Genome-wide analysis of transposon insertion polymorphisms reveals intraspecific variation in cultivated rice. Plant physiology. 2008;148(1):25-40.

79. International Peach Genome I, Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, et al. The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet. 2013;45(5):487-94.

80. Zhang Q, Chen W, Sun L, Zhao F, Huang B, Yang W, et al. The genome of *Prunus mume*. Nat Commun. 2012;3:1318.

81. Shirasawa K, Isuzugawa K, Ikenaga M, Saito Y, Yamamoto T, Hirakawa H, et al. The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. DNA Res. 2017;24(5):499-508.

82. Velasco R, Zharkikh A, Affourtit J, Dhingra A, Cestaro A, Kalyanaraman A, et al. The genome of the domesticated apple (*Malus x domestica* Borkh.). Nat Genet. 2010;42(10):833-9.

83. Wu J, Wang Z, Shi Z, Zhang S, Ming R, Zhu S, et al. The genome of the pear (*Pyrus*

bretschneideri Rehd.). Genome Res. 2013;23(2):396-408.

84. Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook MB, et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science. 2018;361(6398).

85. Shulaev V, Sargent DJ, Crowhurst RN, Mockler TC, Folkerts O, Delcher AL, et al. The genome of woodland strawberry (*Fragaria vesca*). Nat Genet. 2011;43(2):109-16.

86. Buti M, Moretto M, Barghini E, Mascagni F, Natali L, Brilli M, et al. The genome sequence and transcriptome of *Potentilla micrantha* and their comparison to *Fragaria vesca* (the woodland strawberry). Gigascience. 2018;7(4):1-14.

87. Hibrand Saint-Oyant L, Ruttink T, Hamama L, Kirov I, Lakhwani D, Zhou NN, et al. A high-quality genome sequence of *Rosa chinensis* to elucidate ornamental traits. Nat Plants. 2018;4(7):473-84.

88. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet. 2011;43(10):1035-9.

89. Sun D, Wang C, Zhang X, Zhang W, Jiang H, Yao X, et al. Draft genome sequence of cauliflower (*Brassica oleracea* L. var. *botrytis*) provides new insights into the C genome in *Brassica* species. Hortic Res. 2019;6:82.

90. Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet. 2016;48(10):1225-32.

91. Dassanayake M, Oh DH, Haas JS, Hernandez A, Hong H, Ali S, et al. The genome of the extremophile crucifer *Thellungiella parvula*. Nat Genet. 2011;43(9):913-8.

92. Wu H-J, Zhang Z, Wang J-Y, Oh D-H, Dassanayake M, Liu B, et al. Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proceedings of the National Academy of Sciences. 2012;109(30):12219-24.

93. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. A draft genome of field pennycress (*Thlaspi arvense*) provides tools for the domestication of a new winter biofuel crop. DNA Res. 2015;22(2):121-31.

94. Jiao W-B, Accinelli GG, Hartwig B, Kiefer C, Baker D, Severing E, et al. Improving and correcting the contiguity of long-read genome assemblies of three plant species using optical mapping and chromosome conformation capture data. Genome research. 2017;27(5):778-86.

95. Johnston JS, Pepper AE, Hall AE, Chen ZJ, Hodnett G, Drabek J, et al. Evolution of genome size in Brassicaceae. Ann Bot. 2005;95(1):229-35.

96. Lysak MA, Koch MA, Beaulieu JM, Meister A, Leitch IJ. The dynamic ups and downs of genome size evolution in Brassicaceae. Mol Biol Evol. 2009;26(1):85-98.

97. Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, et al. The emerging biofuel crop *Camelina sativa* retains a highly undifferentiated hexaploid genome structure. Nat Commun. 2014;5:3706.

98. Slotte T, Hazzouri KM, Agren JA, Koenig D, Maumus F, Guo YL, et al. The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat Genet. 2013;45(7):831-5.

99. Schranz ME, Kantama L, de Jong H, Mitchell-Olds T. Asexual reproduction in a close relative of *Arabidopsis*: a genetic investigation of apomixis in *Boechera* (Brassicaceae). New Phytol. 2006;171(2):425-38.

100. Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, et al. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nature genetics. 2013;45(8):891.

101. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods. 2012;9(4):357.

102. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

103. Toolkit P. Broad institute, GitHub repository. 2016.

104. Dierckxsens N, Mardulyn P, Smits G. NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic acids research. 2017;45(4):e18-e.

105. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. 2000;17(4):540-52.

106. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Molecular biology and evolution. 2017;34(3):772-3.

107. Stamatakis A, Ludwig T, Meier H. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 2005;21(4):456-63.

108. Suguiyama VF, Vasconcelos LAB, Rossi MM, Biondo C, de Setta N. The population genetic structure approach adds new insights into the evolution of plant LTR retrotransposon lineages. Plos One. 2019;14(5):e0214542.

109. Du J, Tian Z, Hans CS, Laten HM, Cannon SB, Jackson SA, et al. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. 2010;63(4):584-98.

110. Llorens C, Futami R, Covelli L, Dominguez-Escriba L, Viu JM, Tamarit D, et al. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 2011;39(Database issue):D70-4.

111. Mewes H-W, Amid C, Arnold R, Frishman D, Güldener U, Mannhaupt G, et al. MIPS: analysis and annotation of proteins from whole genomes. Nucleic acids research. 2004;32(suppl_1):D41-D4.

112. Huang C-H, Sun R, Hu Y, Zeng L, Zhang N, Cai L, et al. Resolution of Brassicaceae phylogeny using nuclear genes uncovers nested radiations and supports convergent morphological evolution. Molecular biology and evolution. 2016;33(2):394-412.

113. Mérai Z, Graeber K, Wilhelmsson P, Ullrich KK, Arshad W, Grosche C, et al. *Aethionema arabicum*: a novel model plant to study the light control of seed germination. Journal of experimental botany. 2019;70(12):3313-28.

114. Wucherer W, Veste M, Herrera Bonilla O, Breckle S-W, editors. Halophytes as useful tools for rehabilitation of degraded lands and soil protection. Proceedings of the first international forum on ecological construction of the Western Beijing, Beijing; 2005.

115. Wang R, Farrona S, Vincent C, Joecker A, Schoof H, Turck F, et al. PEP1 regulates perennial flowering in *Arabis alpina*. Nature. 2009;459(7245):423-7.

116. Fan J, Shonnard DR, Kalnes TN, Johnsen PB, Rao S. A life cycle assessment of pennycress (*Thlaspi arvense* L.)-derived jet fuel and diesel. Biomass and Bioenergy. 2013;55:87-100.

117. Moser BR, Knothe G, Vaughn SF, Isbell TA. Production and evaluation of biodiesel from field pennycress (*Thlaspi arvense* L.) oil. Energy & Fuels. 2009;23(8):4149-55.

118. Amtmann A, Bohnert HJ, Bressan RA. Abiotic stress and plant genome evolution. Search for new models. Am Soc Plant Biol; 2005.

119. Oh D-H, Hong H, Lee SY, Yun D-J, Bohnert HJ, Dassanayake M. Genome structures and transcriptomes signify niche adaptation for the multiple-ion-tolerant extremophyte *Schrenkiella parvula*. Plant Physiology. 2014;164(4):2123-38.

120. Nagaharu U, Nagaharu N. Genome analysis in *Brassica* with special reference to the experimental formation of *B. napus* and peculiar mode of fertilization. 1935.

121. Arabidopsis Genome I. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature. 2000;408(6814):796-815.

122. Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, et al. The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. Science. 2010;327(5961):92-4.

123. Zubr J. Oil-seed crop: *Camelina sativa*. Industrial crops and products. 1997;6(2):113-9.

124. Guo Y-L, Bechsgaard JS, Slotte T, Neuffer B, Lascoux M, Weigel D, et al. Recent speciation of *Capsella rubella* from *Capsella grandiflora*, associated with loss of self-incompatibility and an extreme bottleneck. Proceedings of the National Academy of Sciences. 2009;106(13):5246-51.

125. Anderson JT, Gezon ZJ. Plasticity in functional traits in the context of climate change: a case study of the subalpine forb *Boechera stricta* (Brassicaceae). Global Change Biology. 2015;21(4):1689-703.

126. Hay A, Tsiantis M. The genetic basis for differences in leaf form between *Arabidopsis thaliana* and its wild relative *Cardamine hirsuta*. Nature genetics. 2006;38(8):942-7.

127. Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio RD, et al. The *Cardamine hirsuta* genome offers insight into the evolution of morphological diversity. Nat Plants. 2016;2(11):16167.

128. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, et al. Improved *Brassica rapa* reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res. 2018;5:50.

129. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, et al. Draft sequences of the radish (*Raphanus sativus* L.) genome. DNA Res. 2014;21(5):481-90.

130. Yang R, Jarvis DE, CheA draft genome of field pennycressn H, Beilstein MA, Grimwood J, Jenkins J, et al. The Reference Genome of the Halophytic Plant *Eutrema salsugineum*. Front Plant Sci.

2013;4:46.

131. Willing EM, Rawat V, Mandakova T, Maumus F, James GV, Nordstrom KJ, et al. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. Nat Plants. 2015;1:14023.

132. Nguyen TP, Muhlich C, Mohammadin S, van den Bergh E, Platts AE, Haas FB, et al. Genome Improvement and Genetic Map Construction for *Aethionema arabicum*, the First Divergent Branch in the Brassicaceae Family. G3 (Bethesda). 2019;9(11):3521-30.

133. Devos KM, Brown JK, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. Genome research. 2002;12(7):1075-9.

134. Niu X-M, Xu Y-C, Li Z-W, Bian Y-T, Hou X-H, Chen J-F, et al. Transposable elements drive rapid phenotypic variation in *Capsella rubella*. Proceedings of the National Academy of Sciences. 2019;116(14):6908-13.

135. Anderson JT, Inouye DW, McKinney AM, Colautti RI, Mitchell-Olds T. Phenotypic plasticity and adaptive evolution contribute to advancing flowering phenology in response to climate change. Proceedings of the Royal Society B: Biological Sciences. 2012;279(1743):3843-52.

136. Ågren JA, Wang W, Koenig D, Neuffer B, Weigel D, Wright SI. Mating system shifts and transposable element evolution in the plant genus *Capsella*. BMC genomics. 2014;15(1):602.

137. Hutcheon C, Ditt RF, Beilstein M, Comai L, Schroeder J, Goldstein E, et al. Polyploid genome of *Camelina sativa* revealed by isolation of fatty acid synthesis genes. BMC plant biology. 2010;10(1):233.

138. Cai C, Wang X, Liu B, Wu J, Liang J, Cui Y, et al. *Brassica rapa* genome 2.0: a reference upgrade through sequence re-assembly and gene re-annotation. Molecular plant. 2017;10(4):649-51.

139. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin IA, et al. The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. Nat Commun. 2014;5:3930.

140. Yu HJ, Baek S, Lee YJ, Cho A, Mun JH. The radish genome database (RadishGD): an integrated information resource for radish genomics. Database (Oxford). 2019;2019.

141. Wessler SR. Turned on by stress. Plant retrotransposons. Curr Biol. 1996;6(8):959-61.

142. Guo X, Hu Q, Hao G, Wang X, Zhang D, Ma T, et al. The genomes of two *Eutrema* species provide insight into plant adaptation to high altitudes. DNA Research. 2018;25(3):307-15.

143. Bibalani GH. Investigation on flowering phenology of Brassicaceae in the Shanjan region Shabestar district, NW Iran (usage for honeybees). Annals of Biological Research. 2012;6:1958-68.

144. Navabi Z-K, Huebert T, Sharpe AG, O'Neill CM, Bancroft I, Parkin IA. Conserved microstructure of the *Brassica* B Genome of *Brassica nigra* in relation to homologous regions of *Arabidopsis thaliana*, *B. rapa* and *B. oleracea*. BMC genomics. 2013;14(1):250.

145. Lysak MA, Koch MA, Pecinka A, Schubert I. Chromosome triplication found across the tribe Brassiceae. Genome research. 2005;15(4):516-25.

146. Wicker T, Yahiaoui N, Guyot R, Schlagenhauf E, Liu Z-D, Dubcovsky J, et al. Rapid genome divergence at orthologous low molecular weight glutenin loci of the A and Am genomes of wheat. The Plant Cell. 2003;15(5):1186-97.

147. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. Genome research. 2004;14(5):860-9.

148. Potter D, Eriksson T. Evans, RC. Oh, S Smedmark, J Morgan, DR Kerr, M Robertson, KR Arsenault, M Dickinson, TA and Campbell, C. 2007:5-43.

149. Ridley HN. The dispersal of plants throughout the world: L. Reeve & Company, Limited; 1930.

150. Beccari O. The origin and dispersal of *Cocos nucifera*: Phil. Journ Sci. 1917;12:1.

151. Committee FE. Flora of North America: North of Mexico Volume 9: Magnoliophyta: Picramniaceae to Rosaceae. 2015.

152. Newcomb W. Fine structure of the root nodules of *Dryas drummondii* Richards (Rosaceae). Canadian journal of botany. 1981;59(12):2500-14.

153. Billault-Penneteau B, Sandré A, Parniske M, Pawlowski K. *Dryas* as a Model for Studying the Root Symbioses of the Rosaceae. Frontiers in Plant Science. 2019;10:661.

154. Bond G. Observations on the root nodules of *Purshia tridentata*. Proceedings of the Royal Society of London Series B Biological Sciences. 1976;193(1111):127-35.

155. Kohls SJ, Thimmapuram J, Buschena CA, Paschke MW, Dawson JO. Nodulation patterns of actinorhizal plants in the family Rosaceae. Plant and Soil. 1994;162(2):229-39.

156. Deighton N, Brennan R, Finn C, Davies HV. Antioxidant properties of domesticated and wild *Rubus* species. Journal of the Science of Food and Agriculture. 2000;80(9):1307-13.

157. Taylor K. Biological flora of the British Isles, no. 197. *Geum urbanum* L. Journal of Ecology. 1997;85(5):705-20.

158. Al-Snafi AE. Constituents and pharmacology of *Geum urbanum*-A review. IOSR Journal of pharmacy. 2019;9(5):28-33.

159. Takos AM, Jaffé FW, Jacob SR, Bogs J, Robinson SP, Walker AR. Light-induced expression of a MYB gene regulates anthocyanin biosynthesis in red apples. Plant physiology. 2006;142(3):1216-32.

160. Bell RL. Pears (Pyrus). Genetic Resources of Temperate Fruit and Nut Crops 290. 1991:657-700.

161. Daccord N, Celton JM, Linsmith G, Becker C, Choisne N, Schijlen E, et al. High-quality *de novo* assembly of the apple genome and methylome dynamics of early fruit development. Nat Genet. 2017;49(7):1099-106.

162. VanBuren R, Bryant D, Bushakra JM, Vining KJ, Edger PP, Rowley ER, et al. The genome of black raspberry (*Rubus occidentalis*). The Plant Journal. 2016;87(6):535-47.

163. Xiang Y, Huang C-H, Hu Y, Wen J, Li S, Yi T, et al. Evolution of Rosaceae fruit types based on nuclear phylogeny in the context of geological times and genome duplication. Molecular biology and evolution. 2016;34(2):262-81.

164. Njuguna W, Liston A, Cronn R, Ashman T-L, Bassil N. Insights into phylogeny, sex function and age of *Fragaria* based on whole chloroplast genome sequencing. Molecular Phylogenetics and Evolution. 2013;66(1):17-29.

165. Jiang S, Cai D, Sun Y, Teng Y. Isolation and characterization of putative functional long terminal repeat retrotransposons in the *Pyrus* genome. Mob DNA. 2016;7:1.

166. Yin H, Du J, Wu J, Wei S, Xu Y, Tao S, et al. Genome-wide annotation and comparative analysis of long terminal repeat retrotransposons between pear species of *P. bretschneideri* and *P. communis*. Scientific reports. 2015;5:17644.

167. Ma B, Kuang L, Xin Y, He N. New Insights into Long Terminal Repeat Retrotransposons in *Mulberry* Species. Genes. 2019;10(4):285.

168. Zhang X, Yue Z, Mei S, Qiu Y, Yang X, Chen X, et al. A de novo genome of a Chinese radish cultivar. Hortic Plant J. 2015;1(3):155-64.

169. Ragupathy R, You FM, Cloutier S. Arguments for standardizing transposable element annotation

in plant genomes. Trends in plant science. 2013;18(7):367-76.

170. Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. Cytogenetic and genome research. 2005;110(1-4):91-107.

171. El Baidouri M, Carpentier M-C, Cooke R, Gao D, Lasserre E, Llauro C, et al. Widespread and frequent horizontal transfers of transposable elements in plants. Genome research. 2014;24(5):831-8.

172. Bennetzen JL, Kellogg EA. Do plants have a one-way ticket to genomic obesity? The Plant Cell. 1997;9(9):1509.

173. Soltis DE, Soltis PS, Bennett MD, Leitch IJ. Evolution of genome size in the angiosperms. American Journal of Botany. 2003;90(11):1596-603.

174. Guo X, Liu J, Hao G, Zhang L, Mao K, Wang X, et al. Plastome phylogeny and early diversification of Brassicaceae. BMC genomics. 2017;18(1):176.

APPENDICES

A Tables

Table A.1: **Sources of genome assemblies, chloroplast genomes and raw reads**

Family	Species Abbr.	Species	Raw reads	Genome assembly	Plastid genome*
R	Pper	*Prunus persica*	SRR3238135	NCBI	NCBI
	Pmum	*Prunus mume*	SRR654705	NCBI	NCBI
	Pavi	*Prunus avium*	SRR4280456	NCBI	Assembled
	Mdom	*Malus domestica*	SRR2996119	NCBI	NCBI
	Pbre	*Pyrus brestchneideri*	SRR609905, SRR609906	NCBI	NCBI
	Ptri	*Purshia tridentata*	SRR5314001	NCBI	Assembled
	Ddru	*Dryas drummondii*	SRR5313979	NCBI	NCBI
	Fves	*Fragaria vesca*	SRR5275202	NCBI	NCBI
	Pmic	*Potentilla micrachtha*	ERR2008847	NCBI	Assembled
	Rchi	*Rosa chinensis*	SRR7077020	NCBI	NCBI
	Rell	*Rubus ellipticus*	SRR7121773	Robert VanBuren, Donald Danforth	NCBI
	Gurb	*Geum urbanum*	ERR2187925	NCBI	NCBI
B	Brap	*Brassica rapa*	SRR385942	NCBI	NCBI
	Bole	*Brassica oleracea*	SRR4289360	NCBI	NCBI
	Rsat	*Raphanus sativus*	DRR014098	NCBI	NCBI
	Bnig	*Brassica nigra*	SRR2054766	NCBI	NCBI
	Siri	*Sisymbrium irio*	http://mustang.biol.mcgill	NCBI	Xinyi Guo[174]
	Spar	*Schrenkiella parvula*	SRR074858	NCBI	NCBI
	Esal	*Eutrema salsugineum*	Xinyi Guo, Sichuan Univ	NCBI	NCBI
	Tarv	*Thlaspi arvense*	SRR1801299	NCBI	NCBI
	Aalp	*Arabis alpina*	ERR233318	NCBI	NCBI
	Atha	*Arabidopsis thaliana*	SRR1946225	NCBI	NCBI
	Alyr	*Arabidopsis lyrata*	DRR013374	NCBI	NCBI
	Csat	*Camelina sativa*	SRR1171875	NCBI	NCBI
	Crub	*Capsella rubella*	SRR3923591-SRR3923593	NCBI	NCBI
	Bstr	*Boechera stricta*	SRR1592624	NCBI	Xinyi Guo[174]
	Chir	*Cardamine hirsuta*	SRR3506501	NCBI	Xinyi Guo[174]
	Esyr	*Euclidium syriacum*	SRR1801309	NCBI	Xinyi Guo[174]
	Aara	*Aethionema arabicum*	SRR4417656	NCBI	NCBI

Table A.1 shows the sources of genome assemblies, chloroplast genomes and raw reads for all analyzed species in Brassicaceae and Rosaceae. Raw reads were mainly downloaded from NCBI SRA and run numbers were provides starting with SRR, ERR or DRR. Raw reads of a species that were not from NCBI SRA were marked with provider's name and institute or website. Most of the chloroplast genome assemblies were downloaded from NCBI organellar genome database. *A few were assemblies by me, indicated as 'Assembled' in the column 'Plastid genome'. Chloroplast protein-encoding genes of 4 Brassicaceae species were provided by Dr. Guo for me to construct the Brassicaceae phylogenetic tree.

Table A.2: **Percentage of top 2 most abundant known LTR-RT families**.

Family	Species Abbr.	Species	Top known LTR-RT families			
			1st	Per	2nd	Per
R	Pper	*Prunus persica*	*tor*	0.56	*tat*	0.41
	Pmum	*Prunus mume*	*tat*	0.67	*tat*	0.58
	Pavi	*Prunus avium*	*tor*	9.95	*ath*	1.31
	Mdom	*Malus domestica*	*tat*	4.21	*tor*	2.11
	Pbre	*Pyrus brestchneideri*	*bia*	1.22	*bia*	0.85
	Ptri	*Purshia tridentata*	*bia*	0.59	*sir*	0.32
	Ddru	*Dryas drummondii*	*sir*	1.01	*ath*	0.99
	Fves	*Fragaria vesca*	*del*	2.39	*del*	1.32
	Pmic	*Potentilla micrachtha*	*tat*	9.06	*bia*	3.16
	Rchi	*Rosa chinensis*	*bia*	2.89	*bia*	2.52
	Rell	*Rubus ellipticus*	*bia*	2.24	*bia*	1.04
	Gurb	*Geum urbanum*	*bia*	2.35	*del*	1.31
B	Brap	*Brassica rapa*	*ret*	7.99	*gal*	3.87
	Bole	*Brassica oleracea*	*ret*	3.28	*gal*	0.67
	Rsat	*Raphanus sativus*	*gal*	1.93	*tat*	0.45
	Bnig	*Brassica nigra*	*gal*	9.72	*crm*	2.18
	Siri	*Sisymbrium irio*	*ath*	0.89	*tor*	0.71
	Spar	*Schrenkiella parvula*	*gal*	0.55	*ath*	0.38
	Esal	*Eutrema salsugineum*	*ath*	3.19	*crm*	2.62
	Tarv	*Thlaspi arvense*	*ath*	5.31	*ret*	2.14
	Aalp	*Arabis alpina*	*ath*	2.32	*del*	1.92
	Atha	*Arabidopsis thaliana*	*gal*	0.94	*ath*	0.61
	Alyr	*Arabidopsis lyrata*	*ret*	10.88	*ath*	3.46
	Csat	*Camelina sativa*	*ath*	2.18	*gal*	1.12
	Crub	*Capsella rubella*	*gal*	1.01	*ath*	0.31
	Bstr	*Boechera stricta*	*bia*	0.69	*gal*	0.50
	Chir	*Cardamine hirsuta*	*crm*	0.91	*gal*	0.65
	Esyr	*Euclidium syriacum*	*tat*	4.23	*gal*	1.91
	Aara	*Aethionema arabicum*	*ory*	1.42	*tat*	1.12

Superfamilies of the top known LTR-RT families are shown in Table A.2. For example, if *bia_1* and *bia_2* are the top 2 most abundant families, then their

corresponding superfamilies are *bia* and *bia*, respectively. 'Per' column indicates the percentages of the nuclear genome occupied by the top known LTR-RT families in analyzed species.

Table A.3: **Percentage statistics of the top 2 most abundant known LTR-RT families**

Family	Family rank	Min.	1st Qu	Median	Mean	3rd Qu.	Max.
Rosaceae	1st	0.560	0.925	2.295	3.095	3.220	9.950
	2nd	0.320	0.783	1.175	1.327	1.152	3.160
Brassicaceae	1st	0.550	0.940	2.180	3.379	4.230	10.880
	2nd	0.310	0.610	1.120	1.448	2.140	3.870

Table A.4: **Four *galadriel* LTR-RTs in Brassicaceae**

Name	Species abbr.	5′ LTR length (bp)	full length (bp)	rt length	Cov (%)	rt Per Ident (%)	2 LTR indent (%)
galadriel_1_1	Bnig	469	5573	225	100	67	96
galadriel_1_2	Brap	474	3801	missing			100
galadriel_1_3	Brap	484	5134	3 small fragments	15	44 to 60	100
galadriel_S	Crub	461	5317	225	100	63	95

BLAT searches were conducted between rt from those 4 LTR-RTs to the *galadriel* rt in the gydb database. 'Cov' stands for coverage of the full rt amino acid sequence of gydb in the alignment, shown in percentages. 'rt Per Ident' indicates the similarity between the rt of each of the 4 galadriel LTR-RTs to the galadriel rt from the gydb database, shown in percentages'. '2 LTR ident' measures the similarity between 2 LTRs of each LTR-RT, shown in percentages.

Table A.5: **Identities between 5′ LTRs of LTR-RTs in *galadriel_1* family**

	galadriel_1_1	*galadriel_1_2*
galadriel_1_2	82%	
galadriel_1_3	80%	82%

The family galadriel_1 has 3 members and the pairwise similarities(%) of their 5′ LTR are shown.

Table A.6: **Proportions of young, mid-aged and old LTR-RTs in raw reads**

Family	Species	Young	Mid	Old	LTR-RT Perc.	Perc of recent elements from other species
R	*Prunus persica*	36	51	12	24	3
	Prunus mume	18	68	15	29	11
	Prunus avium	34	50	16	39	4
	Malus domestica	54	38	9	47	1
	Pyrus brestchneideri	41	47	12	39	6
	Purshia tridentata	19	64	16	21	0
	Dryas drummondii	43	42	15	18	0
	Fragaria vesca	48	38	14	21	0
	Potentilla micrachtha	40	52	8	41	1
	Rosa chinensis	57	37	6	43	0
	Rubus ellipticus	12	73	15	33	0
	Geum urbanum	20	64	17	40	0
B	*Brassica rapa*	27	60	13	38	5
	Brassica oleracea	21	66	13	31	9
	Raphanus sativus	7	81	12	35	5
	Brassica nigra	26	56	18	51	3
	Sisymbrium irio	23	50	27	33	1
	Schrenkiella parvula	4	88	7	24	2
	Eutrema salsugineum	59	33	8	40	1
	Thlaspi arvense	10	71	20	58	1
	Arabis alpina	21	51	41	32	1
	Arabidopsis thaliana	34	45	21	9	10
	Arabidopsis lyrata	41	45	14	34	1
	Camelina sativa	31	57	12	25	4
	Capsella rubella	36	47	17	7	14
	Boechera stricta	18	60	22	19	4
	Cardamine hirsuta	7	67	26	21	3
	Euclidium syriacum	43	41	16	22	7
	Aethionema arabicum	33	53	14	35	1

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked pieces >50 bp to the LTR-RT library. The percentage of read nucleotide masked by LTR-RT library using the pan-species LTR-RT library either by those from its own species or from the remaining other species. Reads masked with 95% or more identity are regarded as reads from young elements, with 80% to 95% from mid-aged elements and with 80% or less from old elements. Perc = percentage.

Table A.7: **Percentages of nucleotides masked in Brassicaceae**

	Br	Bo	Rs	Bn	Si	Sp	Esa	Ta	Aal	At	Al	Cs	Cr	Bs	Ch	Esy	Aar
Br	50	22	1	19	2		1		2								
Bo	27	44	2	19	1		1		2		1	1					
Rs	10	13	46	24	2		1		2								
Bn	27	7	1	60	1		1		1								
Si	2	1		2	87		2		2		1	1					
Sp		1		1		80	3	1	2		2	2					
Esa	2			1			90	1	1								
Ta	1			1	2		3	83	2		3	1					
Aal	2			1			1		91			1					
At	13	1		1	1		1	1	3	45	27	4	1	1			
Al	1							1	91			2	1				
Cs	5	1		1	1		1	1	2	1	3	83	1				
Cr	17	1		1	1		2	1	5	3	9	13	44	1	1	1	1
Bs	4	1		2	2	1	7	1	6		6	6	1	58	1	1	2
Ch	4	1		2	2		4	2	5	2	8	8	2		57	1	
Esy	10	1		1	1		2	1	4		2	2				75	
Aar	1	1		1	1		2		2		1	1					89

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.8: **Percentages of nucleotides in raw reads masked with 95% or more identity in Brassicaceae**

	Br	Bo	Rs	Bn	Si	Sp	Esa	Ta	Aal	At	Al	Cs	Cr	Bs	Ch	Esy	Aar
Br	82	9		6					3								
Bo	34	56	1						1								
Rs	40	6	27	22	1						3						
Bn	10	1		88					1								
Si					94				1								
Sp	37				1	56	1					1					
Esa	2						98										
Ta	8				1			87	1		3						
Aal	3								97								
At	25								3	70	2						
Al	1										98						
Cs	11								1			88					
Cr	42								4	1	1		61				
Bs	11								2					79		1	5
Ch	32								4		1	1			62		
Esy	15															83	
Aar	2																98

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.9: P**ercentages of nucleotides masked with 80% to 95% identity in Brassicaceae**

	Br	Bo	Rs	Bn	Si	Sp	Esa	Ta	Aal	At	Al	Cs	Cr	Bs	Ch	Esy	Aar
Br	45	27	2	20	2		1		1		1						1
Bo	28	43		20	1		1		1		1	1					
Rs	7	11	54	22	2		1		1			1					
Bn	32	8	1	54	1		1		1		1						
Si	1	1		2	86		2	1	2		1	1					
Sp	1	1		1	2	86	2	1			1	1					
Esa	3	1		1	2		86		2		1	2				1	
Ta	1			1	2		3	86	2		3	1					
Aal	2			1	1		1		89		1	1				1	
At	5	1		1	1		1	1	4	43	33	6	2		1	1	
Al	1						1	1	1	1	89	3	1	1			
Cs	2				1		1		2	1	3	87	1				
Cr	7	1		1	1		2	1	6	4	13	19	41	1	1	1	
Bs	2	1		1	1	1	7	2	6	1	7	6	1	62	1	1	1
Ch	2	1		1	2		3	1	4	1	6	6	2	2	67	1	
Esy	6	1		1	1		2	1	4		2	2				78	
Aar		1			1		2		1		1	2					90

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.10: P**ercentages of nucleotides masked with 80% or less identity in Brassicaceae**

	Br	Bo	Rs	Bn	Si	Sp	Esa	Ta	Aal	At	Al	Cs	Cr	Bs	Ch	Esy	Aar
Br	9	28	3	36	4		4	2	5	1	3	2				1	1
Bo	11	29	4	31	3		5		5	1	1	3	1	1	1	1	1
Rs	12	23	9	34	5		4	1	3		3	2				1	1
Bn	35	10	2	38			4	1	4		2	1				2	
Si	1	1		2	82		4	1	3		1	1				1	
Sp	2	4	2	6	13	17	11	6	10	2	8	7	2	2	1	2	2
Esa	3	3	1	3	9	1	55	3	6	1	4	5	1	1	1	1	1
Ta	1	1		2	3	1	6	73	3	1	3	3	1	1		1	1
Aal	5	3	1	6	3		6	2	53	1	7	5	1	2	1	2	1
At	10	2	1	1	1		2	2	3	8	56	7	2	1	1	1	1
Al	1	1		1	1		1	1	2	2	79	5	1	1	1	1	
Cs	6	2	1	3	3	1	4	2	5	2	8	54	4	1	1	1	1
Cr	11	2	1	2	3		5	3	7	5	16	23	14	3	1	3	2
Bs	4	3	1	4	4	1	12	4	8	2	10	11	2	30	1	1	2
Ch	2	3	1	3	4	1	8	4	6	3	13	14	3	3	29	2	1
Esy	7	4	2	4	3		7	3	8	1	6	6	2	2	1	43	1
Aar	2	2		3	4		6	1	5	1	3	4	1	1		2	64

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.11: **Percentages of nucleotides masked in Rosaceae**

	Pp	Pmu	Pa	Md	Pb	Pt	Dd	Fv	Pmi	Rc	Re	Gu
Pp	56	20	12	4	1		1	1		3	1	
Pmu	33	41	11	5	2		2	1		4	1	
Pa	25	12	54	3	1		1			2	1	
Md	1		1	86	10					1		
Pb	1		1	39	57					1		
Pt	2	1	1	4	1	77	10	1		3	1	
Dd	2	1				4	80	1		4	2	
Fv	1			3	1			79	2	12	2	1
Pmi				1				4	88	6	2	1
Rc				1				1		96	1	
Re				1				1		7	88	
Gu				1				3	2	5	2	88

Table A.12: **Percentages of nucleotides masked with 95% or more identity in Rosaceae**

	Pp	Pmu	Pa	Md	Pb	Pt	Dd	Fv	Pmi	Rc	Re	Gu
Pp	92	4	3	1								
Pmu	49	38	8	2	1		1					
Pa	8	2	88	1								
Md				98	2							
Pb				15	85							
Pt						98	1					
Dd							99					
Fv								99		1		
Pmi									99	1		
Rc										100		
Re										1	98	
Gu												99

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.13: **Percentages of nucleotides masked with 80% to 95% identity in Rosaceae**

	Pp	Pmu	Pa	Md	Pb	Pt	Dd	Fv	Pmi	Rc	Re	Gu
Pp	38	33	16	4	1		1	1	1	3	1	
Pmu	30	47	12	4	2		1			2	1	
Pa	35	13	43	4	1		1	1		2	1	
Md	1	1	1	76	19					1		
Pb	1	1	1	55	39					1		
Pt	1	1	1	3		80	9	1		3		
Dd	3	1	1	4	1	4	76	1		6	2	
Fv	1		1	3				68	3	20	3	1
Pmi				1				5	83	8	2	1
Rc				1				1	1	94	2	
Re				1						6	91	
Gu				1				2	2	3	1	90

Table A.14: **Percentages of nucleotides masked with 80% or less identity in Rosaceae**

	Pp	Pmu	Pa	Md	Pb	Pt	Dd	Fv	Pmi	Rc	Re	Gu
Pp	25	17	19	12	3	1	5	3	1	9	5	1
Pmu	26	18	14	11	4	1	5	2	1	14	4	1
Pa	33	31	12	6	2	2	3	1		6	3	
Md	3	2	2	58	25		2	1		5	2	1
Pb	3	2	3	56	26		2	1		4	2	1
Pt	4	2	2	10	2	41	23	1	1	10	3	1
Dd	5	3	3	12	3	13	39	2	1	13	6	1
Fv	2	1	2	9	2		1	37	5	33	6	2
Pmi	1	1	1	5	1		1	13	38	28	8	2
Rc	1	1	1	6	2		1	4	2	74	7	1
Re	1	1	1	4	1	1	1	1	2	20	67	1
Gu	1			3	1		1	6	3	15	6	64

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are indicated by the first letter of the genus plus the first 1 or 2 letters of the species names and are presented in phylogenetic order. Values in cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked homologies > 50bp to the LTR-RT library.

Table A.15: **Two sample t-test of metrics between Brassicaceae and Rosaceae**

	Test of Equal Variance, p-value	Group	Mean (Mb)	Two-tailed p-value	One-tailed p-value
unknown LTR-RT	0.467	B	31.11	0.598	
		R	36.50		
known LTR-RT	0.004**	B	65.88	0.210	
		R	115.82		
Copia	0.053	B	24.56	0.044*	0.022*
		R	49.08		
Gypsy	0.009**	B	40.30	0.402	
		R	63.46		
total LTR-RT	0.031*	B	97.00	0.229	
		R	152.31		
ratio[3]	0.001**	B	2.207	0.016*	0.008**
		R	1.079		
ratio[4]	0.061	B	3.564	0.708	
		R	3.102		
percentage[5]	0.062	B	0.808	0.003**	0.001**
		R	0.867		

Table A.16: **A linear model to predict LTR-RT amount with plant characteristics**

Response Variable	Models	Equation	R^2
Total LTR-RTs (Mb)	Model 1	LTR-RT (Mb) = -36.41** + 0.40** × $genome\ size\ (Mb)$	0.924
	Model 2	$LTR-RTs(Mb) = -48.15^{**} + 0.40^{**} \times genome\ size + 6.99^{\circ} \times Gypsy-to-Copia\ ratio$	0.933

** Significant at 0.01
* Significant at 0.05
[3] The ratios of Gypsy to Copia
[4] The ratio of known LTR-RT(RLNamed) to unknown LTR-RT(RLX)
[5] The percentage of middle-aged(80% - 95%) and young (>95%) TEs
° significant at 0.1

Table A.17: **Brassicaceae and Rosaceae characteristics**

Family	Species Abbr.	Species	Life cycle	karyotype	self-compatibility	Genome Size (Mb)
R	Pper	*Prunus persica*	P(4)	8	Y	265[79]
	Pmum	*Prunus mume*	P(4)	8	Y	280[80]
	Pavi	*Prunus avium*	P(4)	8	Y	353[81]
	Mdom	*Malus domestica*	P(4)	17	N	742[82]
	Pbre	*Pyrus brestchneideri*	P(4)	17	Y	527[83]
	Ptri	*Purshia tridentata*	P(4)	9	N	215[1]
	Ddru	*Dryas drummondii*	P(4)	9	Y	253[84]
	Fves	*Fragaria vesca*	P(4)	7	Y	241[85]
	Pmic	*Potentilla micrachtha*	P(4)	7	Y	406[86]
	Rchi	*Rosa chinensis*	P(4)	7	Y	533[87]
	Rell	*Rubus ellipticus*	P(4)	7	Y	338[2]
	Gurb	*Geum urbanum*	P(4)	21	Y	1475[2]
B	Brap	*Brassica rapa*	A-B(2)	8	N	485[88]
	Bole	*Brassica oleracea*	P(4)	9	N	603[89]
	Rsat	*Raphanus sativus*	A-B(2)	9	N	534[2]
	Bnig	*Brassica nigra*	A(1)	8	N	591[90]
	Siri	*Sisymbrium irio*	A(1)	7	Y	262[2]
	Spar	*Schrenkiella parvula*	A(1)	7	Y	140[91]
	Esal	*Eutrema salsugineum*	A(1)	7	Y	260[92]
	Tarv	*Thlaspi arvense*	A(1)	7	Y	539[93]
	Aalp	*Arabis alpina*	P(4)	8	Y	370[94]
	Atha	*Arabidopsis thaliana*	A(1)	5	Y	157[95]
	Alyr	*Arabidopsis lyrata*	P(4)	8	N	245[96]
	Csat	*Camelina sativa*	A-B(2)	20	Y	261[97]
	Crub	*Capsella rubella*	A(1)	8	Y	219[98]
	Bstr	*Boechera stricta*	B-P(3)	7	Y	264[99]
	Chir	*Cardamine hirsuta*	A(1)	8	Y	225[95]
	Esyr	*Euclidium syriacum*	A(1)	7	Y[?]	262[94]
	Aara	*Aethionema arabicum*	A(1)	20	Y	240[100]

[?] Missing value imputed with self-compatibility
[1] Estimated from K-mer counting with Jellyfish
2 Values from the Kew Gardens database

Life cycle, karyotype, self-compatibility and genome size are collected here for the analyzed species. There are 4 ranks for life cycle: 1 for annual, 2 for annual to biannual, 3 for biannual to perennial and 4 for perennials. The numbers in superscripted parentheses in the 'Genome Size' column represent reference numbers for the publications on genome sizes. In the 'Family' column, R stands for Rosaceae and B stands for Brassicaceae.

Table A.18: **The superfamilies of the top 8 most abundant known LTR-RT families**

Family	Species	Superfamilies of the top abundant known LTR-RT families								Superfamily	
		1st	2nd	3rd	4th	5th	6th	7th	8th	1st	2nd
R	*Prunus persica*	tor	tat	tat	tat	del	tat	tor	tor	tat	tor
	Prunus mume	tat	tat	tor	sir	ath	tor	ath	tor	tat	tor
	Prunus avium	tor	ath	tat	sir	tor	tor	ory	ret	tor	tat
	Malus domestica	tat	tor	bia	bia	tat	ath	ath	ath	tat	ath
	Pyrus brestchneideri	bia	bia	tat	tat	ath	ath	del	del	ath	tat
	Purshia tridentata	bia	sir	tat	ret	ath	tat	del	ret	tor	tat
	Dryas drummondii	sir	ath	ath	tat	bia	tat	ory	ath	ath	tor
	Fragaria vesca	del	del	ret	tat	bia	bia	tat	bia	del	bia
	Potentilla micrachtha	tat	bia	tat	bia	tat	bia	tat	bia	tat	bia
	Rosa chinensis	bia	bia	tat	bia	bia	ath	ath	tat	bia	tat
	Rubus ellipticus	bia	bia	del	ath	ath	ath	ath	ath	ath	bia
	Geum urbanum	bia	del	ath	ath	tat	del	bia	ath	del	ath
B	*Brassica rapa*	ret	gal	crm	ret	tor	del	rei	ret	ret	gal
	Brassica oleracea	ret	gal	ory	ath	ath	bia	ret	bia	ret	bia
	Raphanus sativus	gal	tat	del	ath	ret	tat	ret	ret	ret	gal
	Brassica nigra	gal	crm	tat	crm	crm	ath	ath	ret	gal	crm
	Sisymbrium irio	ath	tor	crm	crm	gal	ath	crm	crm	crm	tor
	Schrenkiella parvula	gal	ath	ath	del	tat	crm	ath	tor	ath	crm
	Eutrema salsugineum	ath	crm	ath	ath	ath	ath	ath	ath	ath	crm
	Thlaspi arvense	ath	ret	crm	rei	ath	ath	ath	tat	ath	crm
	Arabis alpina	ath	del	ath	tat	ath	tat	tat	del	ath	bia
	Arabidopsis thaliana	gal	ath	ret	ath	ath	ath	ath	bia	ath	gal
	Arabidopsis lyrata	ret	ath	tat	ath	ath	tat	tat	ret	ret	ath
	Camelina sativa	ath	gal	del	ath	ath	del	ath	ath	ret	gal
	Capsella rubella	gal	ath	ret	tat	sir	ory	crm	crm	gal	ret
	Boechera stricta	bia	gal	del	ath	ory	ath	ory	bia	ath	bia
	Cardamine hirsuta	crm	gal	tat	crm	tor	sir	sir	bia	crm	ath
	Euclidium syriacum	tat	gal	tor	ath	crm	tor	ath	sir	tat	tor
	Aethionema arabicum	ory	tat	tor	rei	tor	rei	tat	tor	tor	ory

Table A.19: **Decomposition of nucleotides masked by LTR-RTs in Brassicaceae**

	Young		Middle-aged		Old		All	
	Self	Other	Self	Other	Self	Other	Self	Other
Brap	82.1	17.9	44.6	55.4	9.1	90.9	50.2	49.8
Bole	55.7	44.3	42.7	57.3	29.1	70.9	43.6	56.4
Rsat	27.1	72.9	53.7	46.3	8.7	91.3	46.2	53.8
Bnig	88.3	11.7	54.5	45.5	38.3	61.7	60.5	39.5
Siri	94.3	5.7	85.9	14.1	82.4	17.6	86.9	13.1
Spar	56.2	43.8	86.3	13.7	16.9	83.1	79.8	20.2
Esal	97.6	2.4	85.7	14.3	55.0	45	90.1	9.9
Tarv	87.5	12.5	85.9	14.1	73.0	27	83.5	16.5
Aalp	97.1	2.9	89.4	10.6	53.5	46.5	90.6	9.4
Atha	69.8	30.2	43.3	56.7	7.9	92.1	44.9	55.1
Alyr	98.2	1.8	88.5	11.5	79.4	20.6	91.2	8.8
Csat	87.6	12.4	87.5	12.5	54.1	45.9	83.5	16.5
Crub	60.8	39.2	41.1	58.9	13.7	86.3	43.7	56.3
Bstr	79.5	20.5	61.7	38.3	30.2	69.8	57.8	42.2
Chir	61.8	38.2	67.2	32.8	28.9	71.1	57.0	43
Esyr	83.1	16.9	78.3	21.7	43.0	57	74.9	25.1
Aara	97.8	2.2	90.1	9.9	63.6	36.4	89.1	10.9

Table A.20: **Decomposition of nucleotides masked by LTR-RTs in Rosaceae**

	Young		Middle-aged		Old		All	
	Self	Other	Self	Other	Self	Other	Self	Other
Pper	92.2	7.8	37.8	62.2	25.2	74.8	56.1	43.9
Pmum	37.5	62.5	46.9	53.1	17.5	82.5	40.0	60
Pavi	88.3	11.7	42.7	57.3	12.1	87.9	53.6	46.4
Mdom	97.9	2.1	76.1	23.9	58.3	41.7	86.2	13.8
Pbre	85.1	14.9	39.4	60.6	26.2	73.8	56.5	43.5
Ptri	97.9	2.1	79.8	20.2	41.0	59	76.9	23.1
Ddru	99.2	0.8	76.0	24	38.6	61.4	80.3	19.7
Fves	99.0	1	67.7	32.3	36.9	63.1	78.6	21.4
Pmic	98.7	1.3	83.2	16.8	37.8	62.2	85.8	14.2
Rchi	99.9	0.1	94.2	5.8	73.7	26.3	96.2	3.8
Rell	98.1	1.9	91.1	8.9	66.9	33.1	88.5	11.5
Gurb	99.1	0.9	90.5	9.5	64.3	35.7	87.8	12.2

Read nucleotides were RepeatMasker masked with the LTR-RT library. Rows are query species and columns are species having intact LTR-RTs contributing to the LTR-RT library. Species are presented in phylogenetic order. Values in

cells indicate the percentage of nucleotides of query species masked by LTR-RTs from library species. This analysis only includes those masked pieces >50 bp to the LTR-RT library. The percentage of read nucleotide masked by LTR-RT library using the pan-species LTR-RT library either by those from its own species or from the remaining other species. Reads masked with 95% or more identity are regarded as reads from young elements, with 80% to 95% from mid-aged elements and with 80% or less from old elements. Perc = percentage.

Table A.21: **Genome assembly completeness**

Family	Species Abbr.	Species	Genome size (Mb)	Genome assembly size (Mb)	Genome completeness(%)
R	Pper	*Prunus persica*	265[79]	227	86
	Pmum	*Prunus mume*	280[80]	234	84
	Pavi	*Prunus avium*	353[81]	373	106
	Mdom	*Malus domestica*	742[82]	709	96
	Pbre	*Pyrus brestchneideri*	527[83]	509	97
	Ptri	*Purshia tridentata*	215[1]	176	82
	Ddru	*Dryas drummondii*	253[84]	226	89
	Fves	*Fragaria vesca*	241[85]	220	91
	Pmic	*Potentilla micrachtha*	406[86]	330	81
	Rchi	*Rosa chinensis*	533[87]	516	97
	Rell	*Rubus ellipticus*	338[2]		
	Gurb	*Geum urbanum*	1475[2]	1217	83
B	Brap	*Brassica rapa*	485[88]	284	59
	Bole	*Brassica oleracea*	603[89]	489	81
	Rsat	*Raphanus sativus*	534[2]	402	75
	Bnig	*Brassica nigra*	591[90]	402	68
	Siri	*Sisymbrium irio*	262[2]	240	92
	Spar	*Schrenkiella parvula*	140[91]	137	98
	Esal	*Eutrema salsugineum*	260[92]	238	92
	Tarv	*Thlaspi arvense*	539[93]	325	60
	Aalp	*Arabis alpina*	370[94]	301	81
	Atha	*Arabidopsis thaliana*	157[95]	119	76
	Alyr	*Arabidopsis lyrata*	245[96]	184	75
	Csat	*Camelina sativa*	261[97]	199	76
	Crub	*Capsella rubella*	219[98]	130	59
	Bstr	*Boechera stricta*	264[99]	165	63
	Chir	*Cardamine hirsuta*	225[95]	191	85
	Esyr	*Euclidium syriacum*	262[94]	226	86
	Aara	*Aethionema arabicum*	240[100]	170	71

[1] Estimated from K-mer counting with Jellyfish
[2] Values from the Kew Gardens database

The value for ‘Genome completeness’ equals the assembly size (Mb) divided by the estimated genome size (Mb).

Table A.22: **Percentages of Class II TEs in 50 Mb of raw read data**

Family	Species Abbr.	Species	% non LTR-RT TEs
R	Pper	*Prunus persica*	0.2
	Pmum	*Prunus mume*	0.1
	Pavi	*Prunus avium*	0.1
	Mdom	*Malus domestica*	0.1
	Pbre	*Pyrus brestchneideri*	0.1
	Ptri	*Purshia tridentata*	0.2
	Ddru	*Dryas drummondii*	0.2
	Fves	*Fragaria vesca*	1
	Pmic	*Potentilla micrachtha*	2
	Rchi	*Rosa chinensis*	0.1
	Rell	*Rubus ellipticus*	0.1
	Gurb	*Geum urbanum*	1
B	Brap	*Brassica rapa*	0.8
	Bole	*Brassica oleracea*	1
	Rsat	*Raphanus sativus*	0.8
	Bnig	*Brassica nigra*	0.6
	Siri	*Sisymbrium irio*	1
	Spar	*Schrenkiella parvula*	0.6
	Esal	*Eutrema salsugineum*	1
	Tarv	*Thlaspi arvense*	1
	Aalp	*Arabis alpina*	1
	Atha	*Arabidopsis thaliana*	18
	Alyr	*Arabidopsis lyrata*	5
	Csat	*Camelina sativa*	2
	Crub	*Capsella rubella*	2
	Bstr	*Boechera stricta*	2
	Chir	*Cardamine hirsuta*	1
	Esyr	*Euclidium syriacum*	1
	Aara	*Aethionema arabicum*	1

The values are the total percentages of Class II TEs in Rosaceae and Brassicaceae , quantified in 50Mb of raw read data.

Table A.23: **Tests of phylogenetic signals of quantities of top Named LTR-RTs**

Family	LTR-RT superfamilies	$Pagel's$	p-value ($\lambda's\ test$)	$Blomberg's\ K$	p-value ($K's\ test$)
R	tat	0.00	1	0.24	0.41
	athila	0.936	0.077	0.88	0.014*
	del	0.99	0.034*	1.40	0.033*
	bianca	0.99	0.09	0.86	0.025*
	tork	0.00	1	0.17	0.679
	retrofit	0.94	0.009**	0.75	0.02*
	oryco	0.96	0.009**	0.80	0.019*
	sire	0.00	1	0.45	0.149
B	athila	0.98	0.049*	0.83	0.102
	retrofit	0.32	0.594	0.44	0.468
	crm	0.99	0.308	0.75	0.117
	galadriel	0.00	1	0.41	0.698
	tat	0.00	1	0.55	0.263
	tork	0.76	0.090	0.69	0.084
	del	0.86	1	0.61	0.282
	bianca	0.00	1	0.08	0.915
	oryco	0.24	0.834	0.30	0.687
	reina	0.98	0.431	0.69	0.239

R stands for Rosaceae and B for Brassicaceae. Pagel's lambda and Blomberg's K are two indicators for phylogenetic signal and both were calculated above. Larger values of Pagel's lambda and Blomberg's K than 0.8 indicates strong phylogenic signals. Amounts in nucleotides of each LTR-RT superfamily in each plant family were used for the calculation of phylogenetic signals. Small p-values of test statistics can let us to reject the null hypothesis of randomness without phylogenetic signals. The statistical tests of Pagel's lambda and Blomberg's K don't agree with each other sometimes.

B Figures

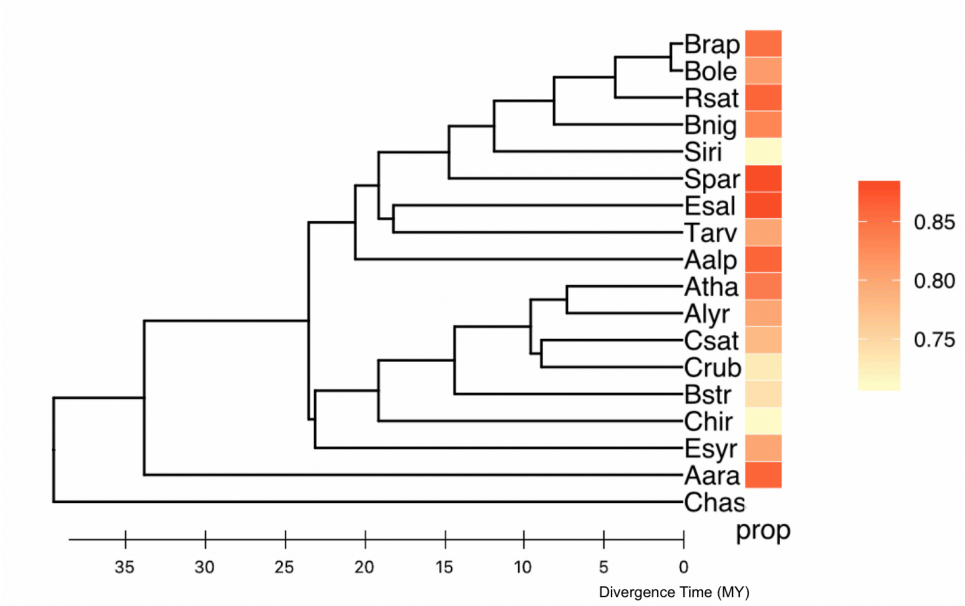

Figure B.1: **Proportions of mid-aged and young LTR-RTs in Brassicaceae**

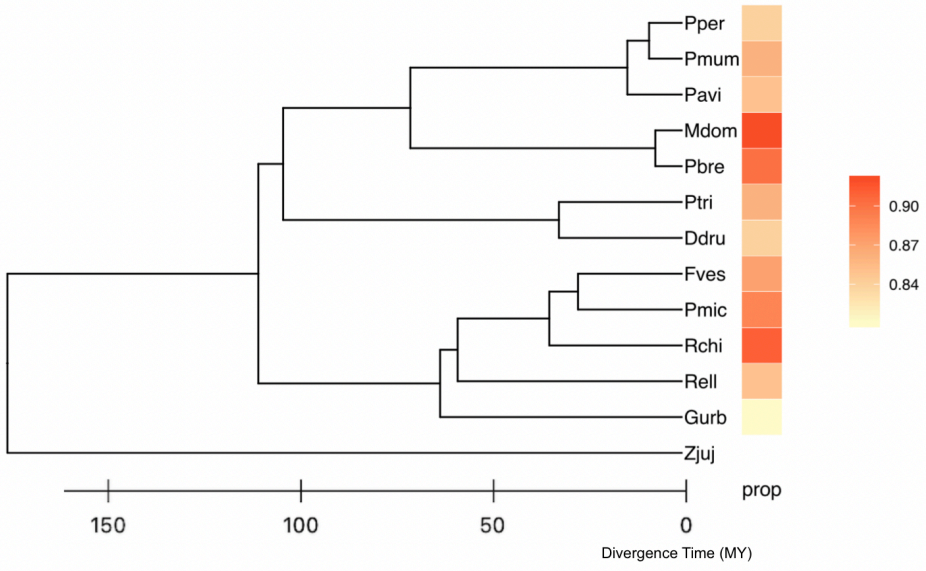

Figure B.2: **Proportions of mid-aged and young LTR-RTs in Rosaceae**

```
CLUSTAL O(1.2.4) multiple sequence alignment

RT_GALADRIEL          LDAGLIRSSKLAPIGAPVLFQKKQDGSLRLCVDYRALNKVTVKNKYPIPLVADLFDRLGK      60
RT_GALADRIEL_1_1      LSTGKLRPS-KAPYGAPVLFQKKHDGSLRMCVDYRALNKVTIKNRYPIPLIADLFDRLGG      59
RT_GALADRIEL_S        LSTRILRPS-KAPYGAPVLFQNKHDGSLRMCVHYRALNMVTIKNCYPIPLIADLFDRLGC      59
                      *.:  :* *  ** *******:*:*****:**.***** **:** *****:********

RT_GALADRIEL          AKVFTKLDLRTGYWQVRIAEGDEAKITCVTRYGSYEFLVMPFGLTNAPATFCTLMNQVLF      120
RT_GALADRIEL_1_1      AKVYTKMDLQKGYYQVRIAEGDEPKTACITRYGSFEWLVMPFGLTNAPATFCTLMNQILQ      119
RT_GALADRIEL_S        AKVYTKMDLQKGYYQVRIAKGDEPKTACVTRYGSFEWLVMPFGLTNAPATFCTLMNQILQ      119
                      ***:**:**:.**:*****:*** * :*:*****:*:********************:*

RT_GALADRIEL          EYLIDEFVVVYLDDIVVYSRTLEEHVEHLRLVLARLREYTLYVKLEKCSFAQKEIKFLGH      180
RT_GALADRIEL_1_1      PY-LDRFVVAYLDDIVIYSNSMEEHVEHLQTVFRVLRENELFVKRVKCTFATEEVQFLGH      178
RT_GALADRIEL_S        PY-LDRFVVAYLDDIVIYSNSKEEHVEHLKIVFQVLRENELYVKREKCTFATEEVQFLGH      178
                       * :*.***.******:**.: *******: *:  ***  *:**  **:** :*::****

RT_GALADRIEL          VIEEGRIKMDKKKVQAIVEWRAPKDVKELRSFLGLANYYRKFIAGYS    227
RT_GALADRIEL_1_1      VIGHGKLKMDEPKVKAIIEWEVPTKVTEMRSFLGLVNFYRRFIEGYS    225
RT_GALADRIEL_S        IIGRGKLRMENAKVSAI-TSEAPNKIHELRFFLDLVNYYQHFIEGYS    224
                      :* .*::::*:: **.**   ..*..: *:* **.*.*:*::** ***
```

Figure B.3: **Multiple sequence alignment of *rt* encoded by 3 *galadriel* LTR-RTs**

RT_GALADRIEL_1_1 is the rt of *galadriel_1_1* in *B. nigra*, RT_GALADRIEL is the *galadriel* rt from the gydb database and RT_GALADRIEL is the rt from *galadriel_S* in *C. rubella*

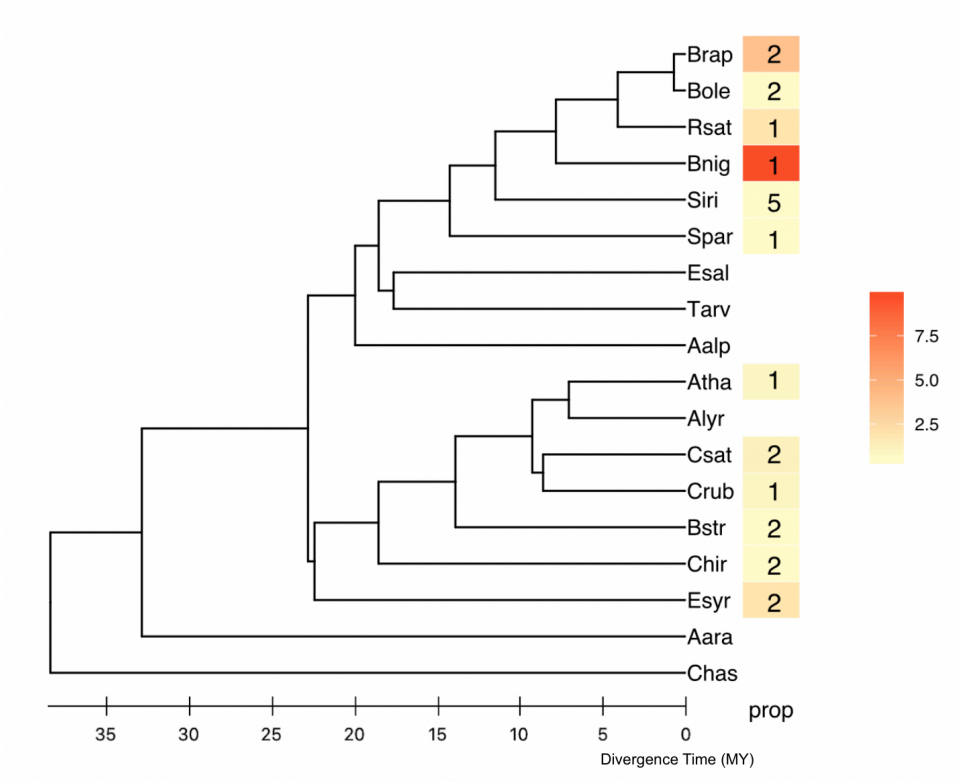

Figure B.4: **Ranks and percentages of *galadriel_1* in 12 Brassicaceae species**

As indicated, *galadriel_1* is a most abundant (top) LTR-RT family in 12 out of 17 analyzed Brassicaceae species. The *warmer* the color, the higher

percentage of *galadriel_1* in that species. The values in each cell indicate the ranks of *galadriel_1* among the top abundant Named LTR-RT families in each species. For example, *galadriel_1* is the most abundant (ranked 1st) Named LTR-RT family in the species *B. nigra* ,and makes up ~10% of *B. nigra*'s genome.

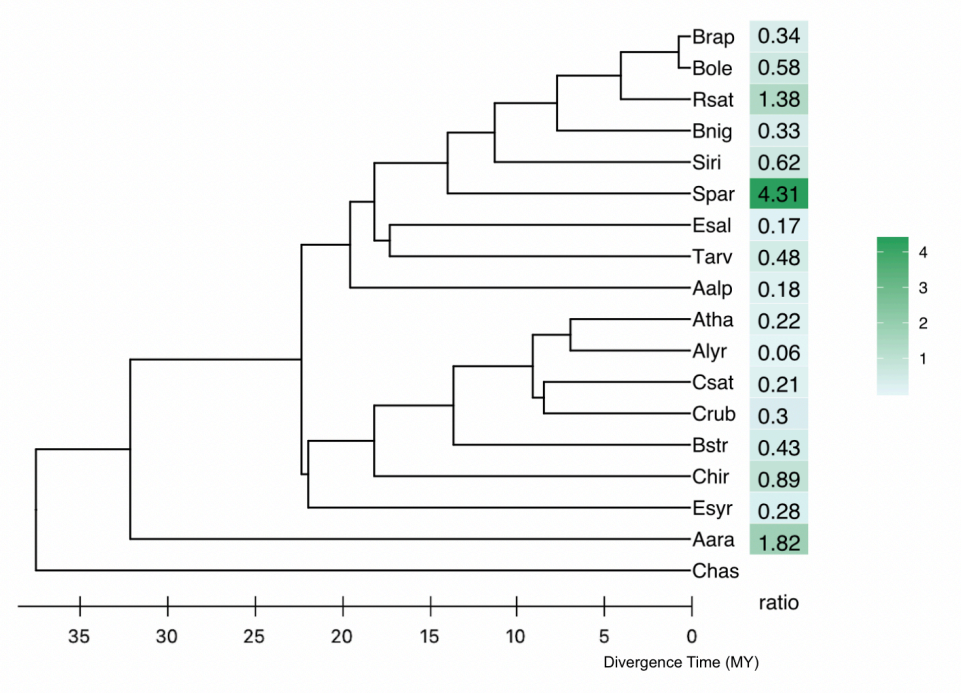

Figure B.5: **The ratios of RLX to RLNamed in Brassicaceae**

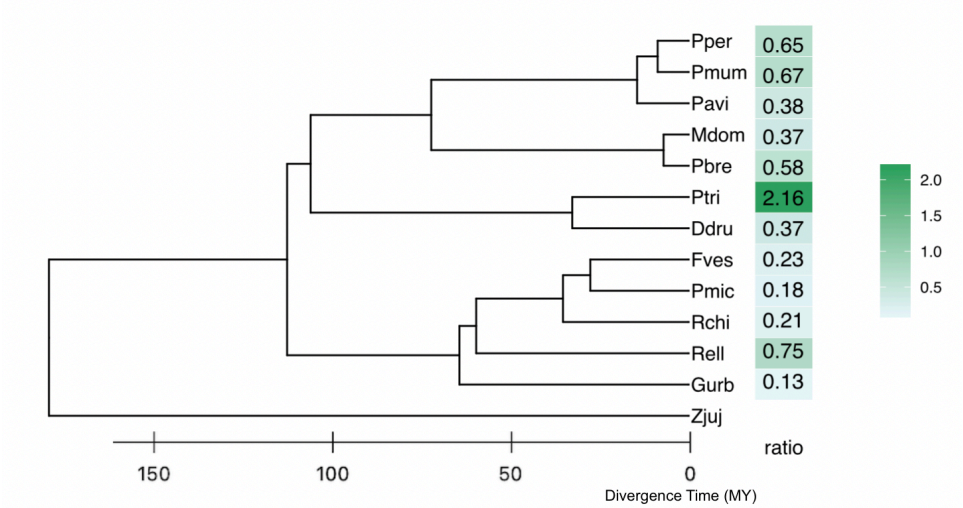

Figure B.6: **The ratios of RLX to RLNamed in Rosaceae**

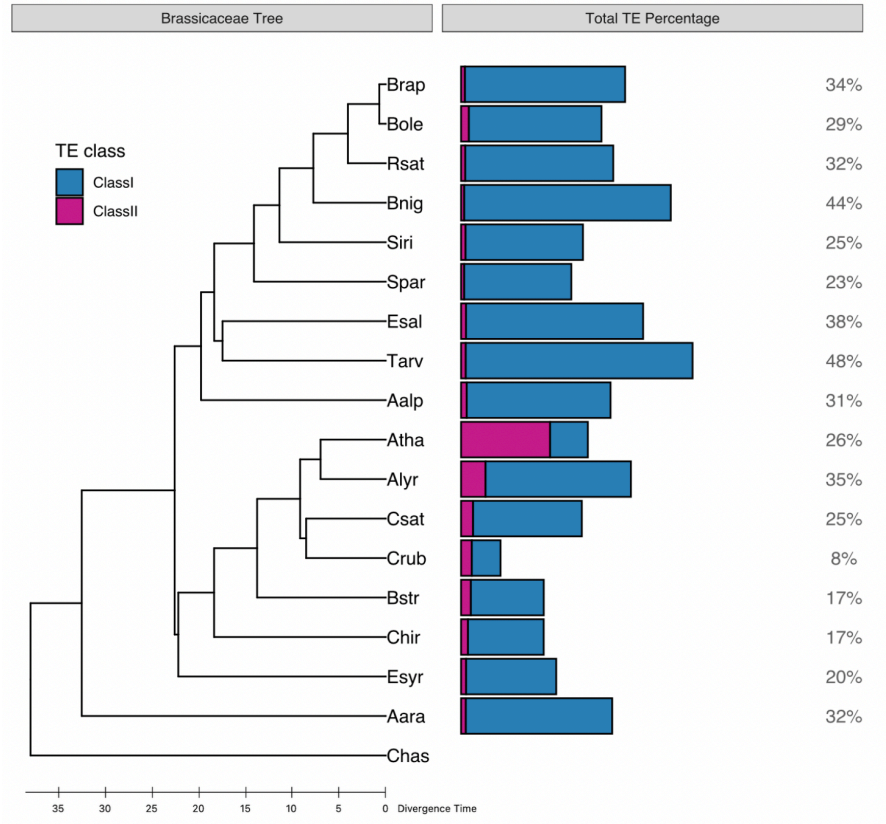

Figure B.7: **Total TE percentages by class in Brassicaceae**

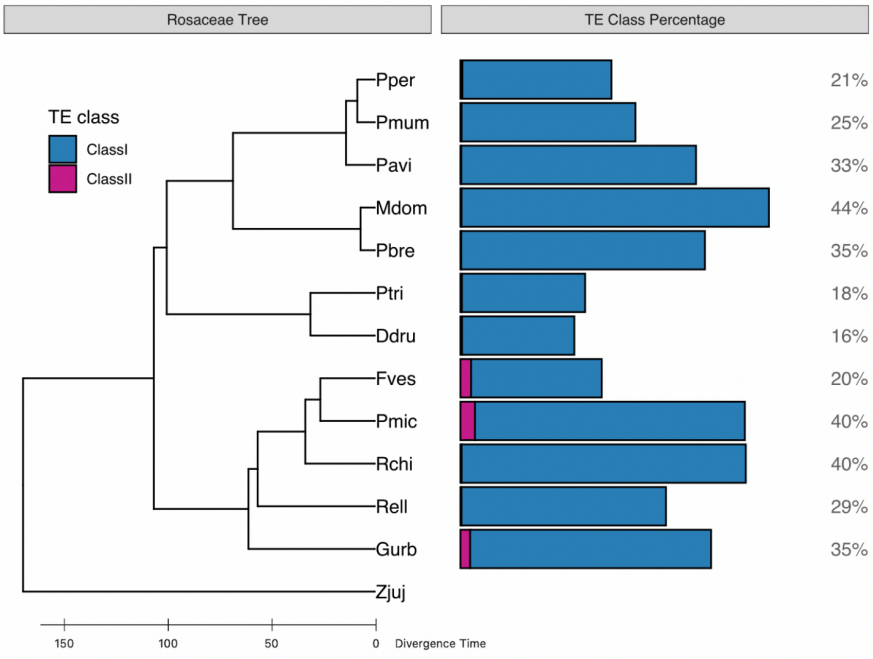

Figure B.8: **Total TE percentages by class in Rosaceae**

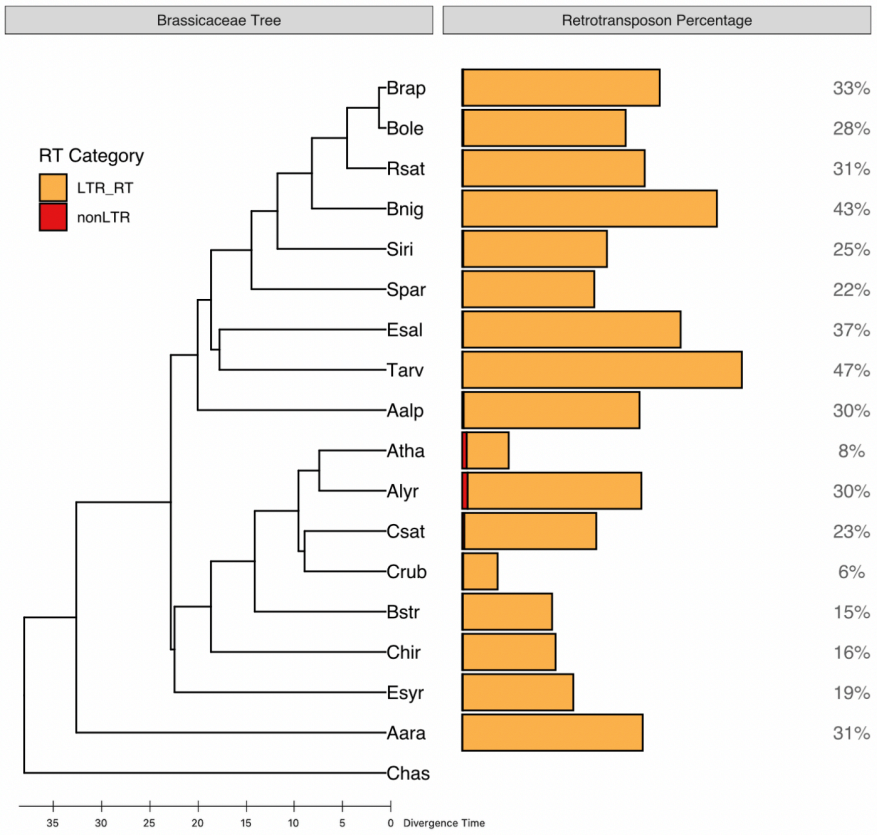

Figure B.9: **Total retrotransposon percentages in Brassicaceae**

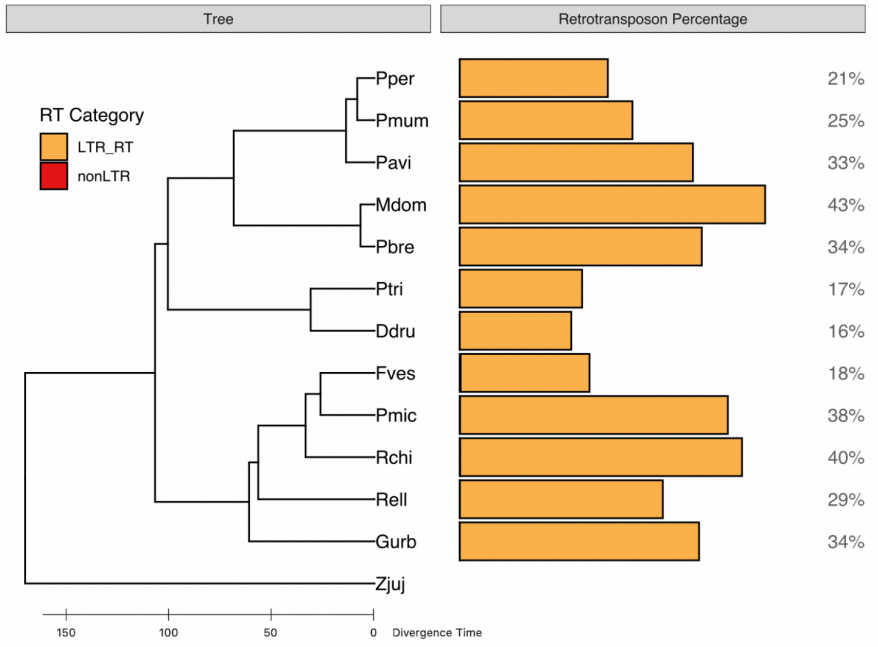

Figure B.10: **Total retrotransposon percentages in Rosaceae**

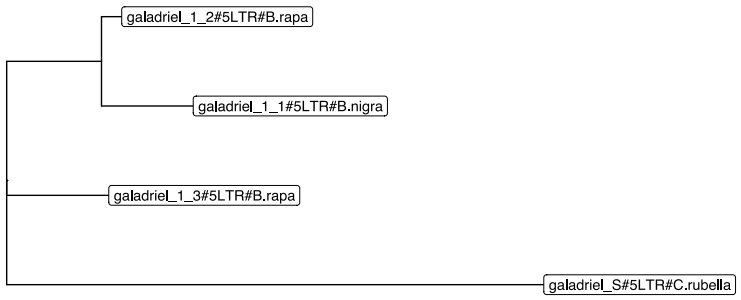

Figure B.11: **The phylogeny of 4 *galadriel* LTR-RTs in Brassicaceae**

C Abbreviations

TE: Transposable Element

nTE: Non-autonomous TE

RT: Retrotransposon

LTR-RT: LTR-retrotransposon

RLX: Unknown LTR-retrotransposon

RLNamed: Known LTR-retrotransposon

SRLX: Unknown LTR-RT with one copy in a family

MRLX: Unknown LTR-RT with more than 1 copy in a family

rt : Reverse transcriptase

rt: Reverse transcriptase gene

MYA: Million years ago

MY: million years