universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

Transposable element dynamics during the adaptive radiation of *Diospyros* from New Caledonia

verfasst von / submitted by

Edith Mayra Briceño Huayta

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023/Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 220
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Joint-Masterstudium Evolutionary Systems Biology
Betreut von / Supervisor:	Assoz. Prof. Dipl.-Ing. Dr. Ovidiu Paun

Acknowledgments

This MSc thesis was accomplished with guidance, expertise, and contributions from several people, which allowed me to complete my master's degree. I want to thank all who were part of this academic journey.

I would like to start by expressing my sincere gratitude to my supervisor, Ovidiu Paun, for all his support, advice, and patience during this thesis work. I am grateful for the opportunity to learn from his expertise during our lots of meetings that were very insightful to address and complete this work, but also on my future career.

My sincere thanks to the Life Science Computer Cluster of the University of Vienna (Lisc) team for their indispensable technical support regarding softwares/pipelines installations, testing, and job time extensions. I especially want to thank Professor Thomas Rattei and Johann Dorn. To Katherine Emilianova, PhD, of the Plant Ecological Genomics group, I would like to express my gratitude for her willingness to share some of her invaluable work on performing and developing TEs libraries, which were essential for this thesis. Similarly, I would like to express my gratitude to Teerna Khastgir, MSc., for sharing important files from her PhD dissertation. I would like to express my gratitude to both for being willing to guide me when I was analysing the genomic data. I want to thank Daniela Paun for her crucial work on the library preparation; without her work, it wouldn't be possible to get high-quality accessions. I would like to thank Florian Brueck for his valuable suggestions on the German version of the abstract of this thesis. I would also like to thank all members of the research group for their valuable feedback on the beginning of my thesis work.

Ultimately, without the love and support of my family—especially my parents, brother, and sister—none of this would have been possible. My mom never dissuaded me from choosing to study overseas, and my dad was my spiritual support to achieve this goal. I must thank my friends and colleagues, Irina Vakhitova, Ronit Chakraborty, Anushe Zhand-Mansoori, Ximena Acosta, Clessio Gomes, and Katharina Ostertag, for believing in me and encouraging me in reaching my goals. I owe Sebastian Allram and his family a debt of gratitude for their love and support during the ups and downs of the thesis journey.

Abstract

The rapid diversification of biological forms into a wide range of ecological roles is known as adaptive radiation. New ecological opportunities and intrinsic genomic properties are expected to be the primary drivers for the rapid adaptive radiation. Transposable elements (TEs) are the major source of de novo mutations and adaptive phenotypic variants. In New Caledonia, one lineage of *Diospyros* (Ebenaceae) radiated into more than 30 plant species in the last 10 My. The aim of this MSc thesis was to gain a preliminary understanding of TE dynamics in *Diospyros* from New Caledonia to get first insights into TEs as possible drivers of adaptive radiation. To find TE insertions, the most recent computational methods and paired-end short next-generation sequencing reads from 29 species of *Diospyros* were employed. The results showed a noticeable TE expansion in the radiating clade compared to the non-radiating species. LTRs, LINEs, and DNA transposons made up the majority of the TE composition, with the biggest TE expansion seen in the radiating clade for the LTR superfamily Gypsy, and, to a lesser extent, Copia. Furthermore, the variant diversity in one of the Gypsy families revealed a high amount of SNPs, a few indels, and terminal deletions. These findings suggested that LTR retrotransposons Gypsy, the most active and abundant TEs in the *Diospyros* radiating clade, had undergone a recent and massive TE expansion.

Zusammenfassung

Die rasche Diversifizierung biologischer Formen in ein breites Spektrum ökologischer Rollen ist als adaptive Radiation bekannt. Es wird davon ausgegangen, dass neue ökologische Möglichkeiten und intrinsische genomische Eigenschaften die treibenden Faktoren für schnelle adaptive Radiation sind. Transponierbare Elemente (TEs) sind die Hauptquelle für De-novo-Mutationen und adaptive phänotypische Varianten. In Neukaledonien hat eine *Diospyros*-Linie (Ebenaceae) in den letzten 10 Millionen Jahren mehr als 30 Pflanzenarten hervorgebracht. Ziel dieser Masterarbeit war es, ein vorläufiges Verständnis der TE-Dynamik in *Diospyros* aus Neukaledonien zu erlangen, um erste Einblicke in TEs als mögliche Treiber adaptiver Radiation zu erhalten. Um TE-Insertionen zu finden, wurden die neuesten Berechnungsmethoden und kurze paired-end reads der Next-Generation-Sequenzierung von 29 *Diospyros*-Arten verwendet. Die Ergebnisse zeigten eine deutliche TE-Expansion in der Radiationsgruppe im Vergleich zu den nicht strahlenden Arten. LTRs, LINEs und DNA-Transposonen machten den Großteil der TE-Zusammensetzung aus, wobei die größte TE-Expansion in der strahlenden Gruppe bei der LTR-Superfamilie Gypsy und in geringerem Maße bei Copia zu beobachten war. Außerdem ergab die Variantenvielfalt in einer der Gypsy-Familien eine hohe Anzahl von SNPs, einige Indels und terminale Deletionen. Diese Ergebnisse deuten darauf hin, dass die LTR-Retrotransposonen Gypsy, die aktivsten und die am häufigsten vorkommenden TEs in der *Diospyros*-Radiationsgruppe gewesen sind und in jüngster Zeit eine massive TE-Expansion erfahren haben.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SUPPLEMENTARY

1. Introduction

Adaptive radiations are extremely diverse clades, resulting within a short evolutionary time from an interplay between adaptation to distinct ecological niches, speciation, and extinction (Ronco et al., 2021). Darwin's finches in the Galapagos archipelago are probably the most striking example of adaptive radiation, which led to emergence of 14 species during the last 2.3 million years (Sato et al., 2001), with diverse shapes and sizes of beaks adapted to particular ecological niches across the islands (Grant & Grant, 2002; Tebbich et al., 2010). Alongside the finches, the mockingbirds (*Nesomimus sp.*), which share ecological niches with the finches, diversified into only four species (Tebbich et al., 2010). This suggests that adaptive radiation is triggered not only by new ecological opportunities but also by certain genome configurations that can generate novel and potentially adaptive genomic variants (Schrader & Schmitz, 2018).

Adaptive radiations begin often with a long-distance dispersal event and likely with a small founder population (Levin et al., 2009; Paun et al., 2016; Schenk, 2021; Schrader and Schmitz, 2018;). The genetic bottleneck events followed by new ecological conditions may induce genomic stress conditions in the founders, which can trigger TEs bursts. Low effective population size (Ne) and reduced strength of selection can favour TE invasion across the genome (Schrader & Schmitz, 2018), especially in low recombination genomic regions (Kent et al., 2017). Through different processes (e.g., domestication, exaptation, retrogenes, genome plasticity, etc.) TE are co-opted and can generate novel and adaptive variants (Schrader & Schmitz, 2018). TE can have a direct or indirect effect on genes and genomes influencing (i) genome size; (ii) chromosomal rearrangements by ectopic recombination; (iii) gene architectures by point mutations, insertions, and deletions; (iv) novel genes; (v) gene movement; (vi) gene expression (transcriptional and posttranscriptional levels, including alternative splicing); (vii) genetic networks (TE wiring); and (viii) epigenetic regulation (Bennetzen & Wang, 2014; Oliver & Greene, 2009). A remarkable example of a TE insertion as a driver of adaptive evolution is the well-known "industrial melanism" of the peppered moth, where an intronic insertion of a DNA transposon increased the transcript abundance of the Cortex gene, generating an adaptive wing colour phenotype during the heavy pollution of the industrial revolution period (Schrader & Schmitz, 2018; Hof et al., 2016). Likewise, paralogous TE copies can promote erroneous transposition and ectopic recombination, leading to chromosome rearrangement and genomic plasticity (Oliver & Greene, 2009). Despite TE activity being known for its often-deleterious effects, in adaptive evolution, TE accumulation has been reported to be the source of beneficial variations (Bennetzen & Wang, 2014; Fedoroff, 2012), and it can confer genome plasticity (Schrader & Schmitz, 2018).

The enormous proportion of TEs in eukaryotes (e.g., in plants, about 80% of the genome) and their high diversity made it necessary to have a unified hierarchical classification (Wicker et al., 2007). According to the TE classification proposed by Wicker et al. (2007), TEs were grouped by their transposition intermediate: RNA (Class I or retrotransposons) or DNA (Class II or transposons) and enzymatic characteristics. Retrotransposons use the 'copy and paste' mechanism, where the RNA transcript is reverse transcribed to a cDNA and re-integrated in the genome by an integrase or endonuclease. DNA transposons use the 'cut and paste' or "peel and paste' (i.e., for Helitron, where the intermediate is a circular DNA) with a transposase that excises the DNA. The next hierarchy is based on the TE enzymatic principles: Class I elements make the copy themselves, such as the orders: LTRs (long terminal repeats), DIRS (*Dictyostelium* intermediate repeat sequence), LINEs (long interspersed nuclear element), SINEs (short interspersed nuclear elements), etc., and Class II elements go elsewhere to reintegrate in the genome, such as the orders TIR (terminal inverted repeats) and Helitron. Within each order, the superfamilies (Gypsy, Copia, L1, hAT, etc.) are grouped by their replication criteria, protein-coding or non-coding domain, and target site duplication (TSD) length. Finally, TE families are grouped by their DNA conservation, which is limited to highly conserved coding areas, thus making the TE families so diverse in the genome (Wicker et al., 2007). Beside them, another group of abundant TEs are the so-called non-autonomous TEs, which are derivative deletions of autonomous TEs (with mutations and coding regions) (Wicker et al., 2007). Non-autonomous TEs have degenerative or lack coding regions, such as MITE (miniature inverted transposable elements), which are valuable to get insight into adaptive evolution (Bennetzen & Wang, 2014; Bourque et al., 2018; Wicker et al., 2007). LTRs and MITEs constitute the most abundant TEs in plants (Dubin et al., 2018); e.g., *Arabidopsis* complete genome has ~1200 MITEs, maize comprises 50-80% retrotransposons, and barley >70%. In 67 plant genomes, 49.8 million TE insertions were reported in a TE plant database (http://apte.cp.utfpr.edu.br/). Of these, 62.85% were LTRs and 37.15% were DNA transposons (Pedro et al., 2021). Likewise, the large genome sizes in several plant species are due to retrotransposition activity. In 14 *Lilium* species, the amplification of LTRs might be responsible for their large genomes (30-45 Gb), and in maize, LTRs bursts had doubled their genome size within the last 6 Mya (Feschotte et al., 2002).

Currently, there are several pipelines for discovering TE insertions, which differ in the type of TE insertion to be identified, i.e., matching a reference insertion or identifying a novel (non-reference) insertion, and the sequencing method (e.g., whole genome sequencing WGS, Pool-Seq data) of the query sequences. Here we used McClintock 2 (Chen et al., 2023), a meta-pipeline to identify TE insertions from Illumina data with up to 12 software packages. Two of them, Ngs_te_mapper2 (Han et al., 2021) and PopoolationTE2 (Kofler et al., 2016)

were chosen for this MSc project as they have the best CPU time efficiency among the others and have the most recently updated methods. Ngs_te_mapper2 algorithm aims to detect TE insertions in WGS by using two approaches for non-reference TE and reference TE (see Figure 1) (Han et al., 2021; Linheiro & Bergman, 2012). In contrast, PopoolationTE2 can detect TE frequencies in a population. However, this approach also identifies the presence and absence of TE insertions in single individuals (see Figure 2) (Kofler et al., 2016). In addition to the discovery and localization of TE insertions in the genomes offered by the previous algorithms, a TE dynamic visualization by DeviaTE was also performed to understand genetic variation among (SNPs, insertions, and deletions) and within the species and get better insights about the evolutionary dynamics of bursts of TE (Weilguny & Kofler, 2019).

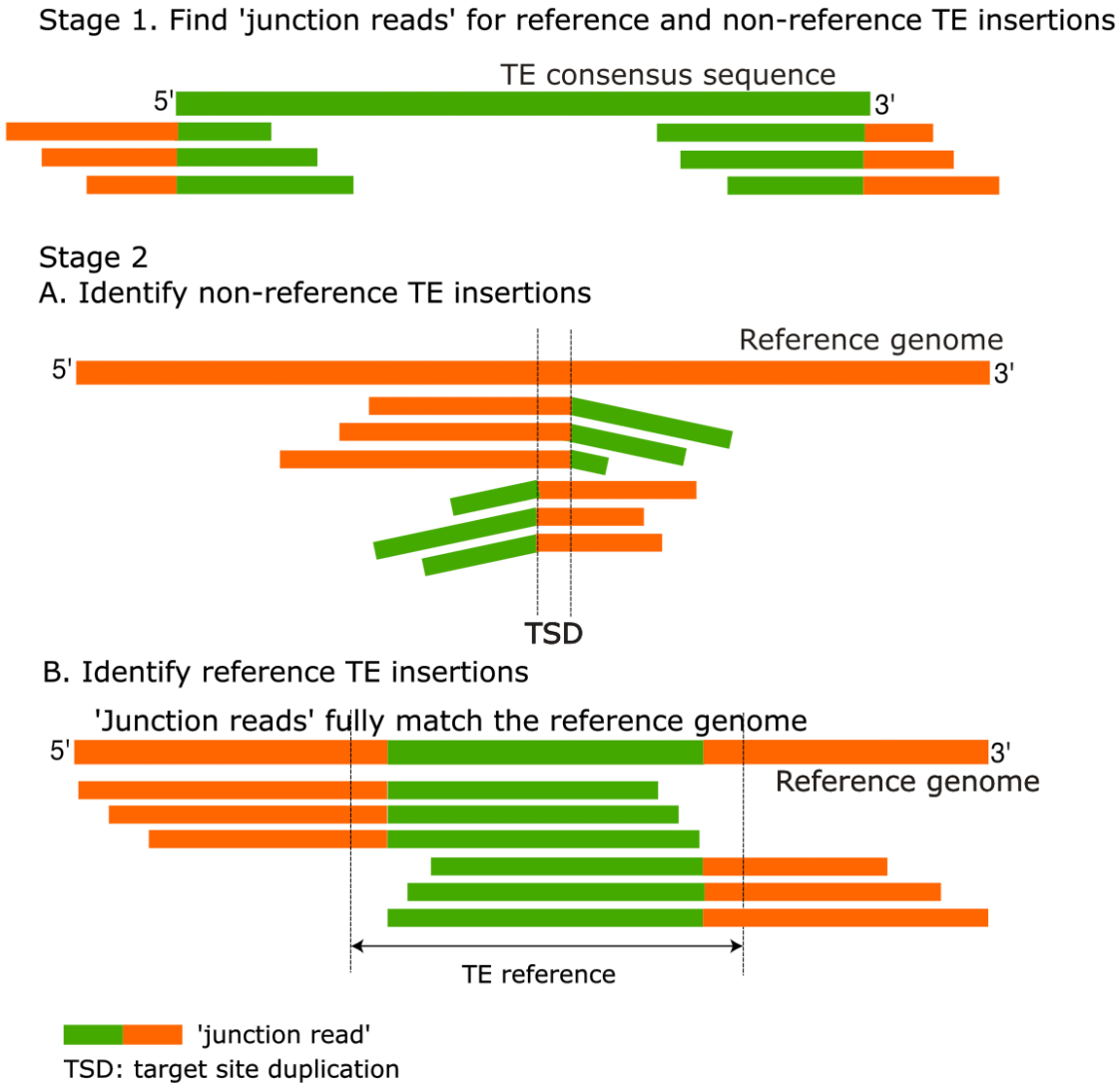

Figure 1. Detection of TE insertions by Ngs_te_mapper2. In Stage 1, the query sequences are aligned to a consensus TE library to find and retain “junction reads” that flank the start and end of TE sequences. This stage is identical for both reference and non-reference TE detection. In Stage 2, A., the portion of “junction reads” outside of the TE sequence is aligned against the reference genome. The intervals between overlapping ‘junction reads’ define the position of the TSD and the presence of non-reference TE. Alternatively, for discovering reference TEs in Stage 2, B., ‘junction reads’ are aligned to the reference genome, but in contrast to the non-reference TEs, here the ‘junction reads’ will fully align to the reference genome. Then, these ‘junction reads’ are clustered to define the location of the reference TE (Han et al., 2021). The scheme was modified from the Ngs_te_mapper2 repository in Github (Han & Bergman, 2020) (https://github.com/bergmanlab/ngs_te_mapper2).

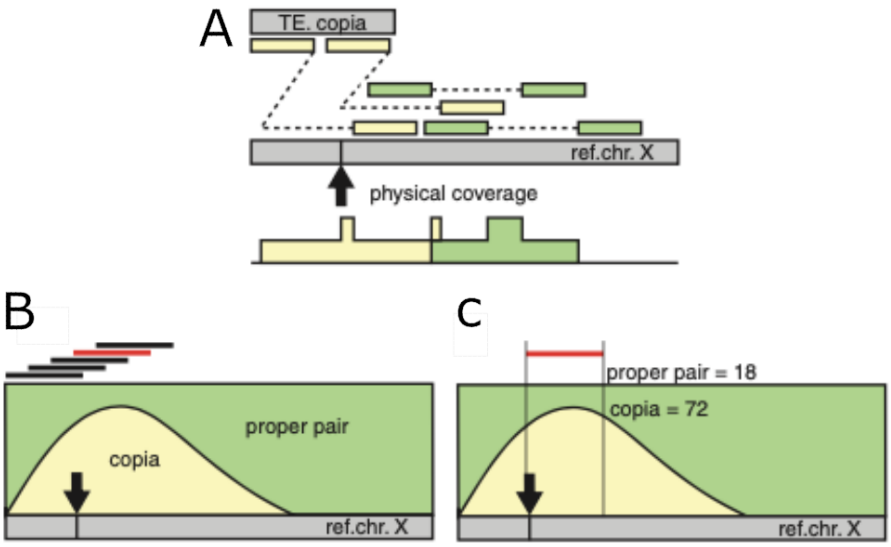

Figure 2. Detection of TE insertions by PopoolationTE2. A. The yellow rectangles show ‘discordant’ paired-end reads in which one read maps to a TE and the other to the reference genome, which suggest the presence of a non-reference TE. The green rectangles show properly mapped paired-end reads to the reference genome, suggesting the absence of a non-reference TE. The span region between the paired-end reads is called physical coverage, and for discordant pairs the median distance of proper pairs across the genome is used. B. A sliding window approach across regions with discordant pairs is used to identify the region with the highest coverage that supports the TE insertion (shown with a red line). C. TE frequency is estimated as the ratio of the physical coverage supporting the TE insertion divided by the total physical coverage. The scheme was modified from Kofler et al. (2016).

Adaptive radiation studies have been largely focused on animal systems, but several evolutionary plant groups have also investigated this phenomenon; some salient cases are Hawaiian silverswords, Hawaiian lobeliads, and columbines (Schenk, 2021). This MSc project is focused on the New Caledonian radiation of *Diospyros* (persimmons), the largest genus in the Ebenaceae family. Four ancestral lineages of *Diospyros* conquered New Caledonia via long-distance dispersal at different times (*ca.* 2M, 1M, 850K, and 700K generations ago) (O. Paun, personal communication, February, 2023; Turner et al., 2013a). Only one of these events resulted in a clade of more than 30 closely related but ecologically divergent species, while the other three only resulted in a few (up to 5) species (Paun et al., 2016; Samuel et al., 2019). New Caledonia is located on the southwestern Pacific (ca. 1300 km east of Australia), and it was separated from Gondwana and submerged ca. 80 Mya and re-emerged ca. 37 Mya (Pillon, 2012). While it was submerged, oceanic layers of heavy metal-rich sediments accumulated that define much of the present soil composition on the archipelago (Morat et al., 2012). *Diospyros* took advantage of the multiple ecological niches and is widely distributed on the archipelago, except in the mangroves (Samuel et al., 2019; Turner et al., 2013b). Divergence within the early phase of the radiation was comparatively slow and correlated with (macro)climatic conditions (mountain range on Grande Terre E-W), whereas during the most recent periods divergence was faster and appears to have been associated with specialization to edaphic conditions (Paun et al., 2016), like other plant groups on the archipelago (e.g., Barrabé et al., 2018). Regarding the genome size, the radiating group has almost a twofold size in comparison to the non-radiating groups and compared to the closest related outgroup, *D. sandwicensis*, but with no increase in ploidy (O. Paun, personal communication, February, 2023). A previous study has shown that the enlarged genomes of the radiating group are notably enriched with TEs where distinct families have proliferated (Turner et al., 2013a).

This MSc thesis aimed to contribute to getting better insights into the TE dynamics during adaptive evolution in plants as one of the sources of tremendous phenotypic variation during the adaptation to new ecological conditions. To accomplish it, this research thesis tackled the TE dynamic along with the adaptive radiation of *Diospyros* from New Caledonia by characterising the TE composition and their patterns of mobilisation between the radiating and non-radiating species. The aims here are: 1) detect TE insertions within the genomes of selected representatives of the radiating and non-radiating species; 2) characterise TE activity across the radiating and non-radiating species; and 3) identify TE families with higher activity between the radiating and non-radiating species that could drive the genome size increase and the explosive radiation of *Diospyros.*

2. Material and Methods

Accessions

This study focused on 29 accessions of *Diospyros*, of which 23 were part of the radiating clade in New Caledonia; four accessions belonged to the non-radiating species in New Caledonia; and two outgroups for the radiating clade. In the radiating clade, ten accessions belonged to unique species, and 13 corresponded to six species, where two accessions were used for each species, and only one species had three accessions. For the non-radiating group, the following were included: *D. macrocarpa*, *D. outbatchensis*, *D. olen*, and *D. fasciculosa*. As outgroups of the radiating clade (Samuel et al. 2019; Turner et al. 2013a;), *D. ferrea is* a widely distributed tropical species in Africa and the Indian Ocean, and *D. sandwicensis* is a species distributed only in the Hawaiian Islands (see Table 1).

Genome sequencing

The genome sequences were provided by the lab group; however, here we will briefly describe the techniques used for DNA library, extraction, and sequencing. The DNA library was prepared with the NEBNext Ultra II DNA PCR-free Library Prep Kit for Illumina (New England Biolabs E7410), using the adapters from the NEBNext Multiplex Oligos for Illumina (Unique Dual Index UMI DNA Set 1 - E7395). DNA was extracted with the CTAB protocol, adding a Sorbitol wash before the extraction. The samples were sequenced with whole genome sequencing on the NovaSeq S4 Illumina platform at the Vienna BioCenter Core Facilities. The length of the paired-end reads was 150 bp. The number of paired-end reads per accession was between 57 and 236 million (see Table 1). The paired-end reads were quality trimmed simultaneously while removing Illumina adapters with trim-galore-0.6.7 (Quality Phred score cutoff: 20) (Krueger et al., 2021).

Reference genome and consensus TE library

A reference genome of *D. impolita* v.1.0 ('impolitaG') consisting of 1,470 contigs was used in this research (Paun et al. unpublished). The consensus TE library was in Fasta format, and it was performed in RepeatModeler v. 2.0.4-5.36.1-1.0.6 (Flynn et al., 2019). It contained 2,368 TE families. In addition, a second consensus TE library in Fasta format was generated by EDTA (Ou et al., 2019). This file comprised 10,147 TE families, and it was used for DeviaTE analyses. Both consensus TE libraries were produced by a member of the Plant Ecological Genomics group at the University of Vienna, PhD Katherine Emilanova.

Table 1. Species, accession names, read numbers, genome sizes, and coverage estimation of the 29 accessions. The non-radiating species are shown with green rows, whereas the outgroups with light blue, and the radiating clade species with rosa.

N°	Species	Accession	Raw pair reads (millions)	Genome size (Mb)*	Coverage**
1	*D. fasciculosa*	fasciculosaBT013	95.80	1105	25.43
2	*D. macrocarpa*	macrocarpaBT047	80.20	NA	NA
3	*D. olen*	olenBT184	122.60	841	42.77
4	*D. outbatchensis*	outbatchensisBT163	90.90	NA	NA
5	*D. ferrea*	ferreaSD2012	82.70	NA	NA
6	*D. sandwicensis*	sandwiceG	57.40	NA	NA
7	*D. calciphila*	calciphilaBT313	139.2	1936	21.09
8	*D. cherrierii_1*	cherrieriBT262	126.1	1575	23.50
9	*D. cherrierii_2*	cherrieriBT293	150.90	1575	28.12
10	*D. erudita*	eruditaBT275	150.90	2015	21.98
11	*D. flavocarpa*	flavocarpaBT128	149.20	NA	NA
12	*D. glans_1*	glansBT093	124.90	1985	18.46
13	*D. glans_2*	glansBT094	167.00	1985	24.68
14	*D. labillardierei*	labillardiereiBT123	144.90	NA	NA
15	*D. minimifolia_1*	minimifoliaBT131	150.20	1535	28.70
16	*D. minimifolia_2*	minimifoliaD4G	67.40	1535	12.88
17	*D. parviflora_1*	parvifloraBT060	122.00	2112	16.94
18	*D. parviflora_2*	parvifloraBT288	132.30	2112	18.38
19	*D. pustulata*	pustulataBT137	157.30	1506	30.64
20	*D. revolutissima*	revolutissimaBT219	161.70	2005	23.66
21	*D. tridentata*	tridentataBT206	109.40	2103	15.27
22	*D. trisulca*	trisulcaBT192	171.30	NA	NA
23	*D. umbrosa_1*	umbrosaBT251	138.40	1526	26.62
24	*D. umbrosa_2*	umbrosaBT255	130.40	1526	25.08
25	*D. veillonii*	veilloniiBT227	157.50	NA	NA
26	*D. vieillardii_1*	vieillardiiBT024	236.00	1526	45.38
27	*D. vieillardii_2*	vieillardiiBT286	119.30	1526	22.94
28	*D. vieillardii_3*	vieillardiiBT023	98.60	1526	18.96
29	*D. yahouensis*	yahouensisBT238	143.50	NA	NA

*Genome size was taken from Turner et al. (2013a). 1C=1pg=978 Mb. If more than one estimate was available per species, the average was calculated.
**Coverage= number of reads * length of the paired-reads / genome size in base pairs.

Mapping reads with BWA-MEM

To get a first glimpse of the presence of TEs in our samples, we mapped the trimmed reads against the consensus TE library using the Burrows-Wheeler Aligner v. 0.7.17 (BWA-MEM) (Li, 2013). The output sam files were converted to bam format with samtools v. 1.17 (samtools view -Sb) (Li et al., 2009). The mapping counts were summarised with FeatureCounts (Liao et al., 2013a) under Subread v. 2.0.6 (Liao et al., 2013b). The CPM (counts per million normalisation) method was applied to make the counting comparable among the species. A posterior heat map was produced with heatmap.2 by gplots v. 3.1.3 in RStudio v. 2022.2.2.485 (RStudio Team, 2022) with R v. 4.2.0 (R Core Team, 2022). The normalised counts were log-converted prior to plotting the heatmap. Then a chi-square test was calculated to find out the significant enrichment of TE families among the main TE clades in the heatmap. Unknown TE families, simple repeats, and rRNAs were filtered out of this test.

TE detection

The TE insertions were detected with McClintock 2 (Chen et al., 2023), which is a meta-pipeline to detect TEs along with 12 methods. Firstly, we tested this package to run all methods with a small reference genome of only 3 contigs, and it was completed in about 10 days. However, when it included the complete reference genome of *D. impolita*, it took more than 20 days for one accession. For that reason, the best time-efficient and updated methods were picked up to run within McClintock 2: Ngs_te_mapper2 (Han et al., 2021) (Figure 1) and PopoolationTE2 (Kofler et al., 2016) (Figure 2). Both methods had a similar running time (~ 3 days) for our reference genome and no failures. McClintock 2 required three input files: a reference genome, a consensus TE Fasta file, and the paired-end reads. The execution of McClintock 2 was done in two phases to avoid redundancy and improve efficiency. In phase 1 (--make annotation), a consensus TE annotation GFF (General Feature Format) and TE taxonomy TSV (Tab-separated values) file were generated. In phase 2, the previous output directory (--resume) was used to discover the TE insertions. Every species was run separately. McClintock 2 outputs provided filtered and unfiltered TE insertions; however, to have an overview of all TE detected, we opted for the unfiltered TE insertions.

TE activity across the phylogenetic tree

From the TE insertions detected by Popoolationte2, the ten TE families with the highest TE insertion numbers were selected for each accession and annotated with the help of a phylogenetic tree of *Diospyros*. The phylogeny was prepared by Teerna Khastgir, MSc. The phylogenetic tree was pruned and annotated in iTOL (https://itol.embl.de/) (Letunic & Bork, 2006) to see TE activity across the phylogeny.

Visualising TE activity within species

We used DeviaTE (Weilguny & Kofler, 2019) to analyse and visualise the dynamics of TEs within species. It provided a tabular report and graph for each TE family, showing coverage (ambiguous and unambiguous), SNP (fixed or polymorphic sites), indels, and internal and terminal deletions at each position in the genome. The input files for DeviaTE were prepared as follows: 1) To reduce sequence redundancy and increase the alignment performance, the consensus reference TE was clustered to 80% identity using CD-HIT-EST (-c 0.8) (Li & Godzik, 2006). 2) Pair-end reads were aligned to the consensus reference TE with Bowtie2 v. 2.5.1 (Langmead & Salzberg, 2012) aligner to speed up this phase on DeviaTE. The output Sam files were converted to Bam format using Samtools view (-h -S -b) (Li et al., 2009). 3) To compare the TE diversity between species, a single-gene copy normalisation was carried out in DeviaTE. The genes were annotated using Busco-5.4.7 (Manni et al., 2021), and five single-copy genes (see Supplementary 2) were included in the consensus reference TE. 4) This new consensus file TE was indexed with Bowtie2 v. 2.5.1 (Langmead & Salzberg, 2012) as it was required by DeviaTE. Finally, DeviaTE was run with the aligned bam file (--input_bam), single-copy gene normalization (--single_copy_genes), the consensus TE fasta file (--library), and the selected TE families (--families). Similar analyses were run for the EDTA-derived consensus reference TE, when the same steps for clustering, alignment, and single-copy genes were followed prior to run Deviate.

The bioinformatics work was performed in the Life Science Compute Cluster (LiSC) of the University of Vienna. The scripts used to perform the softwares/pipelines and plots are available in the Appendix, Supplementary 1.

3. Results

Mapping TEs to a TE database with BWA-MEM

This approach included 12 selected species of *Diospyros*, among them members of the radiating clade, outgroup species to the radiation, and non-radiating species (Figure 3). The mapping rate towards the *D. impolita* TE consensus library was markedly higher (50-63%) in the radiating group than in the non-radiating species (7-11%), according to Samtools flagstat (see Supplementary 3). The ten TE families with the highest mapping counts were led by an unknown TE family (TE148) with 239,339.5 reads (after CPM normalisation) in *D. labillardierei*, a member of the radiating clade. Followed by four Copia families (TE1765, TE554, TE513, TE304), registered in *D. macrocarpa*, a member of the non-radiating group; then three Gypsy families (TE1687, TE687, TE1666) in both *D. sandwicensis* and *D. labillardierei*; a second unknown TE family (TE1810) in *D. macrocarpa*; and one rRNA (TE1920) recorded in fifth place of mapping counts. In contrast, the highest ten averaged mapping counts across the 12 species were led by the previously mentioned unknown TE families, followed by one of the Copia TEs (TE1765), six Gypsy families (TE1687, TE268, TE687, TE1666, TE1354, TE2313), and finally the mentioned rRNA (see Supplementary 4 and 5).

A graphical representation of the mapping counts is shown as heatmaps (Figures 3 and 4) and provide an overview of TE activity dynamics over the *Diospyros* lineages. The first heatmap included all accessions (Figure 3) where many TE families are shown to have raised their activity in the radiating clade in comparison to the non-radiating species (e.g., TE families in TE clade 5). In contrast, other TE families had a converse pattern, where the TE activity was reduced in the radiating members (e.g., TE clade 3a). Only a few TE families (17) exhibited relatively consistent TE activity across the accessions (< 1 SD). To test whether TE families were affected by particular types of activity in particular clades of accessions, chi-squared tests have been performed. The TE dendrogram of the heatmap was divided into five main TE clades (Figure 3). The TE ratio (number of TE families per superfamily/total number of TE families) was calculated for every TE superfamily in both the TE clades and in the annotated consensus TE file. A crosstab (2x2) was used to contrast the expected and observed mapped counts for each TE superfamily and estimate the p-values to find out whether any TE had a significant overrepresentation or underrepresentation in the respective clade. As a result, three retrotransposon families (LTR/Copia, LTR/Gypsy and LINE/L1) and one DNA transposon family (TIR/CMC-EnSpm) exhibited significant activity across several TE clades. Gypsy was overrepresented (p-value 0.0005) in TE clade 1 that shows high copy number in the radiating clade but underrepresented in TE clades 3 (p-value 0.018) and 4 (p-

value 0.001), which show much lower copy numbers. In turn, Copia was significantly underrepresented in both TE clades 1 (p-value 0.004) and 5 (p-value 0.003) that show high copy numbers in the radiation but was overrepresented in TE clades 3 (particularly in TE clade 3; p-value 0.01) and 4 (p-value 0.0005) that showed low copy number for the radiating clade. LINE 1 was overrepresented (p-value 0.04) in TE clade 3, and especially in TE clade 3c, that showed average copy numbers in the radiating species, but in general lower in the non-radiating ones, including outgroups. In addition, CMC-EnSpm was also underrepresented (p-value 0.04) in TE clade 1 and significantly overrepresented in TE clade 5 (p-value 0.03). In TE clade 2, no TE family was significantly over or underrepresented. This clade showed fairly low, constant activity across all samples. Consistent with this, Copia was also overrepresented in TE clades 3a and 3b (p-values 0.006 and 0.009 respectively) especially in the non-radiating species, and correspondingly in the radiating species and *D. ferrea*. Interesting, LINE1 was overrepresented (p-value 0.011) in the clade 3c, especially with high copy numbers in the outgroups and the non-radiating members.

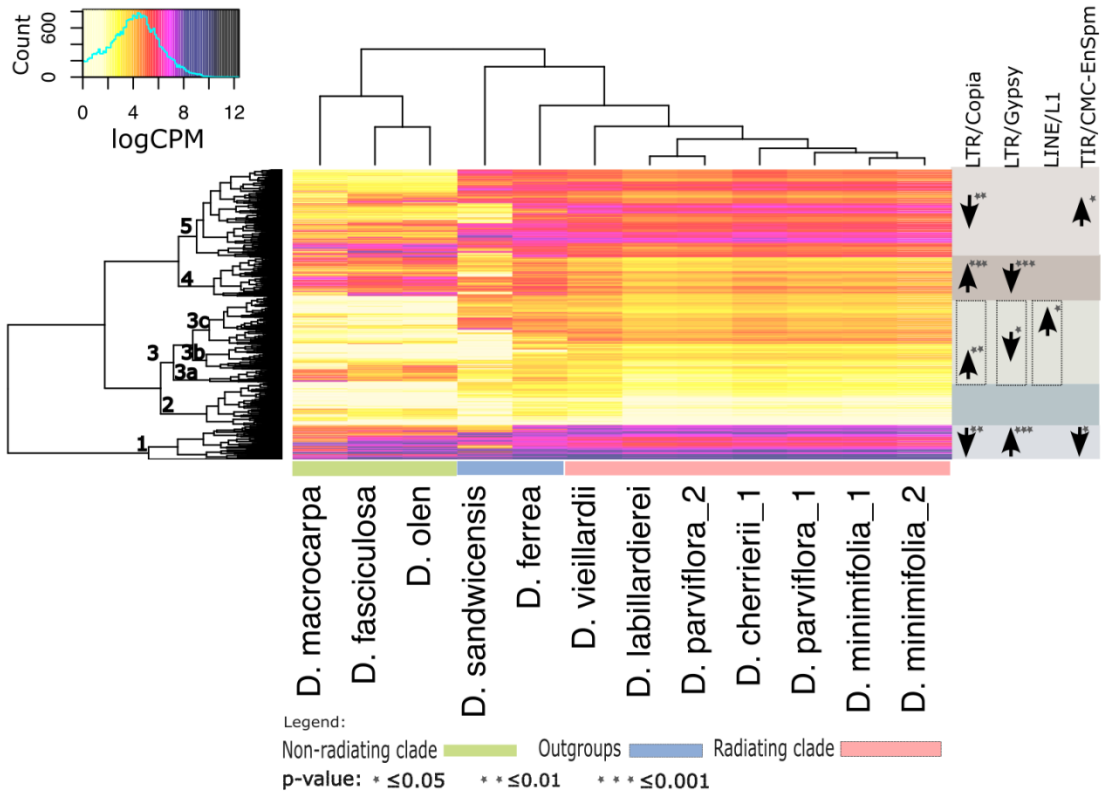

Figure 3. Heatmap of the library size-normalised mapping counts for selected Diospyros species, against a consensus reference TE library based on *D. impolita*. The TE dendrogram (left side) shows the TE clades 1 to 5. The histogram in upper left corner shows a density curve for the TE counts, where the x-axis represents the logCPM (log10 from CPM), from lowest (white) to highest (black) numbers, and the y-axis indicates the density counts for each logCPM value. On the right side, the TE families are shown with significant overrepresentation or underrepresentation in each TE clade (square colours). The arrows pointing up symbolize overrepresentation and the arrows pointing down symbolize underrepresentation. The black stars indicate the significance of the p-values according to the legend. The colored bar (green, blue, and rosa) under the heatmap indicate the species groups. The arrows within the dashed lines show the overrepresented TE families for both clades 3, 3a, 3b and 3c; except the arrow in the middle, which represents only clade 3.

A second heatmap (Figure 4) included the radiating species and the outgroup *D. ferrea*. Here, the TE activity appeared homogeneous among the radiating clade species compared to *D. ferrea*. To look for specific TE activity in the TE clades, the TE dendrogram was grouped into four TE clades (a-d). Only TE clades a, b, and c were significantly enriched for some TE families, while clade d was significantly depleted for Gypsy. TE clade a was enriched in LTR/Gypsy (p-value 0.0005), with mostly homogeneous and very high copy number in all samples. Copia and TIR/CMC-EnSpm (p-values 0.002 and 0.01, respectively) were underrepresented in this TE clade. TE clade b was only enriched in Copia (p-value 0.0005), but in contrast to the previous clade, Gypsy, and TIR/hAT-Ac were underrepresented (p-values 0.009 and 0.05, respectively) with very low copy number in most of the TE clade. TE clade c was enriched in DNA transposons of the families TIR/hAT-Tag1 (p-value 0.032) and Helitron (p-value 0.024). TE clades c and d were significantly underrepresented in Copia and Gypsy respectively (p-values 0.005 and 0.024). In TE clade c, the activity was quite homogenous and high, so that the low copy numbers of Copia in this TE clade were characteristic for all the samples. In TE clade d, Gypsy showed low copy numbers especially in the radiating species (see enrichment data in Supplementary 6).

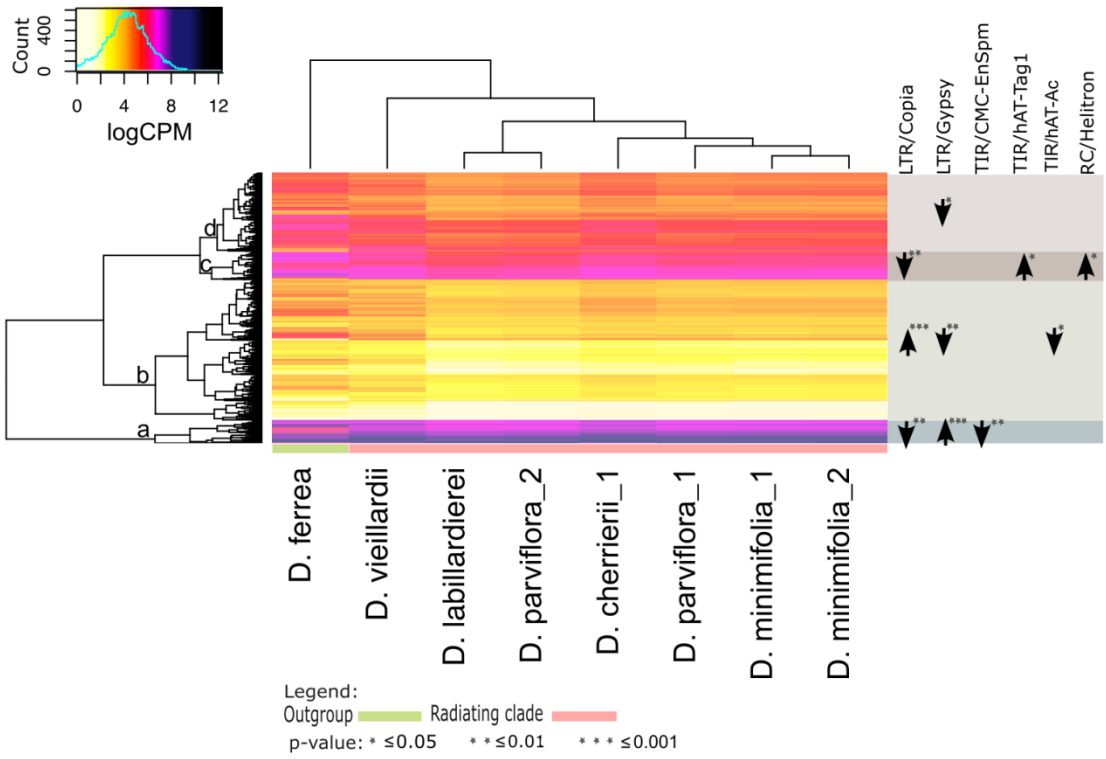

Figure 4. Heatmap of the library size-normalised mapping counts for selected radiating species and the outgroup *D. ferrea*. The TE family dendrogram (left side) is grouped into four clades: a to d. The histogram graph in the upper left corner shows the same as described for Figure 3. On the right side, TE families significantly over- or underrepresented are shown for each clade. The arrow directions, the black stars, and the bar colours under the heatmap are as described for Figure 3.

TE insertions detected by McClintock2

The purpose to use McClintock2 (Chen et al., 2023) was to detect TE insertions alongside 12 packages; however, due to long CPU time efficiency, we restrict the full analysis for two pipelines: Ngs_te_mapper2 (Han et al., 2021) and PopoolationTE2 (Kofler et al., 2016). In total, for the 28 accessions, PopoolationTE2 (ahead referred to as Po.TE2) discovered 10,194,620 TE insertions, and Ngs_te_mapper2 (ahead referred to as Ngs.TE2) listed 213,554 (Figure 5). Po.TE2 did not make any differentiation between reference and non-reference TE, but Ngs.TE2 did; nonetheless we decided to use the total TE insertions reported because of the intent of this study.

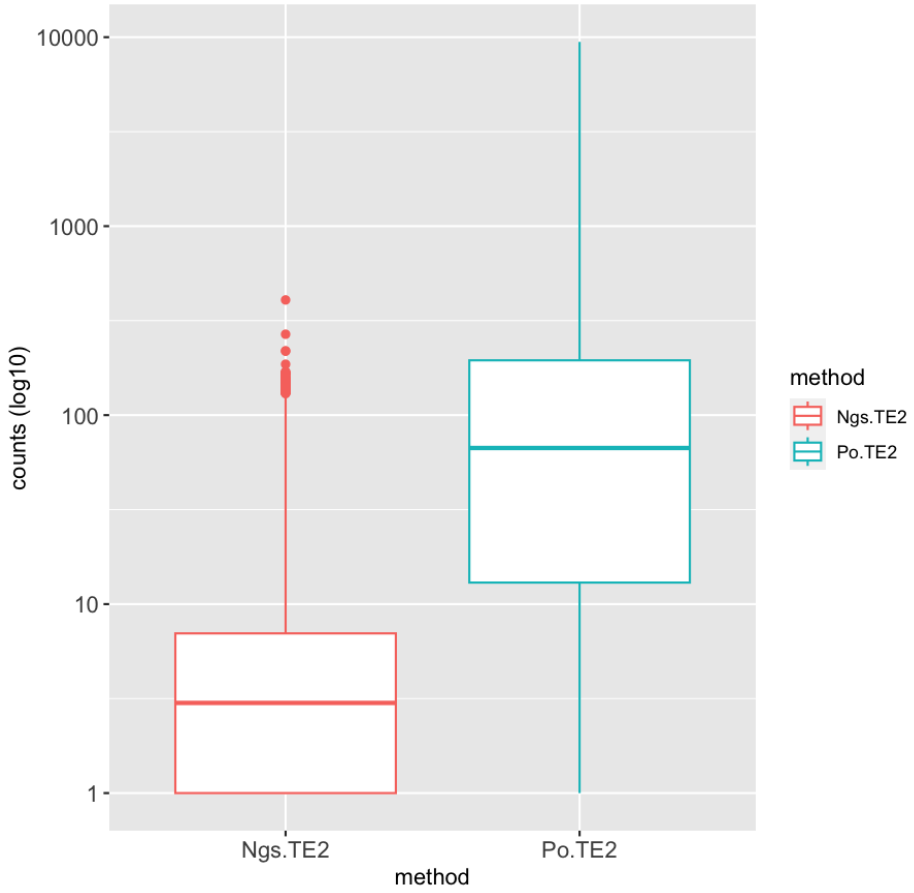

Figure 5. Box plot of the TE insertions counts per accession for Ngs_te_mapper2 and PopoolationTE2. The counts in the y-axis are on a log 10 scale.

PopoolationTE2

A beanplot (Figure 6) illustrates TE insertions per family throughout species and their density distribution as identified with Po.TE2. As it was expected, the four New Caledonian non-radiating species (mean of TE insertions 35,176.5 and mean of TE insertions per family 33.4) exhibited fewer TE insertions than the radiating clade (mean of TE insertions 439,252.1 and mean of TE insertions per family 212.3). In the non-radiating group, most of the TE insertion counts were distributed under a value of ten on the y-axis (purple shape surrounding the green dots). In contrast to the mapping heatmap results, the TE insertions in the outgroup *D. ferrea* were lower than in *D. sandwicensis* (70,524 and 319,844, respectively). TE counts per family for *D. ferrea* were slightly higher than the non-radiating species but lower than *D. sandwicensis*. Its distribution shape was above the non-radiating species, between 10 (log y-axis) and the overall mean (48.3 on the log scale). On the other hand, TE insertions of *D. sandwicensis* were concentrated very close to the overall mean, highlighting an apparent similar TE activity to the radiating species. The radiating species (mean 60.8 in the log scale) exhibited very similar density shapes, most of them with a mean above the overall mean. The

distribution of the TE insertions discovered by Po.TE2 over all the *Diospyros* species demonstrated a persistent TE expansion in all radiating species (see TE copy number per accession in Supplementary 7).

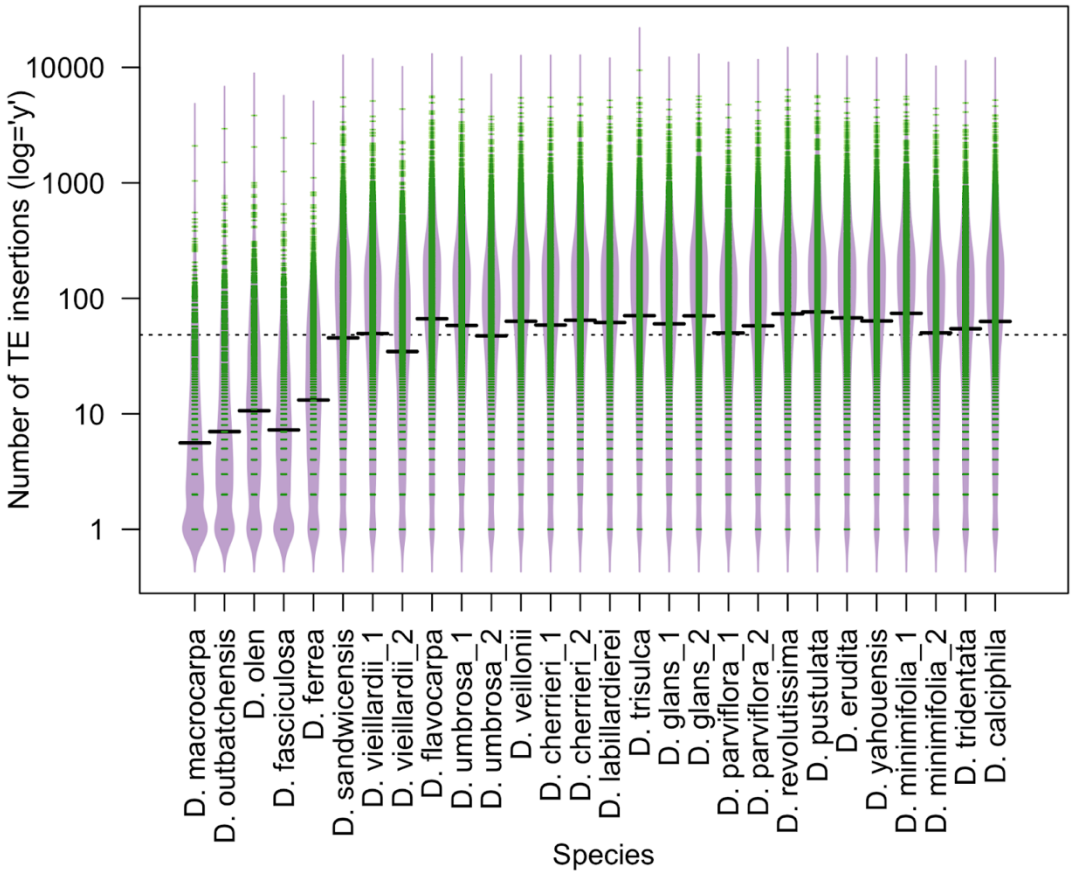

Figure 6. Beanplot for TE insertion counts for every TE family identified with Po.TE2. Green lines show TE insertion count for each family on a log scale.

The violet shape surrounding the green lines shows the density distribution of TE insertions. Thick black lines are the mean for each accession. The dashed line represents the overall mean. Po.TE2 was developed with the purpose of estimating TE frequencies for multiple samples from a population (i.e., Pool-Seq data). However, here, Po.TE2 was performed individually with whole genome sequencing samples (WGS). Therefore, TE 'frequencies' account for the homozygosity/heterozygosity of TE haploid insertions, but could also be influenced by paralogues loci, especially if they are not duplicated in the reference genome, but they have recently duplicate in the respective accession. Figure 7 shows histograms of TE frequency across the species. Within the non-radiating species (first row in the Figure 7), the TE frequency distribution was above 0.5 for at least 50% of the TE insertions. where *D. macrocarpa* had the highest TE frequency with a median of 0.86. In the outgroups, surprisingly, the histogram for *D. ferrea* (median of 1) suggested that the vast majority of TE

insertions were likely homozygous in this species. Within the radiating species (from *D. vieillardii_1* to *D. calciphila*), the TE frequency values differ a lot among them. An interesting case was *D. vieillardii_2* (median was 0.869), which had the highest number of homozygous TE insertions across all samples. Although there was a slightly higher proportion of heterozygosity in the radiating clade, potentially suggesting a more recent expansion, the pattern is not very clear to associate with bursts of TE insertions.

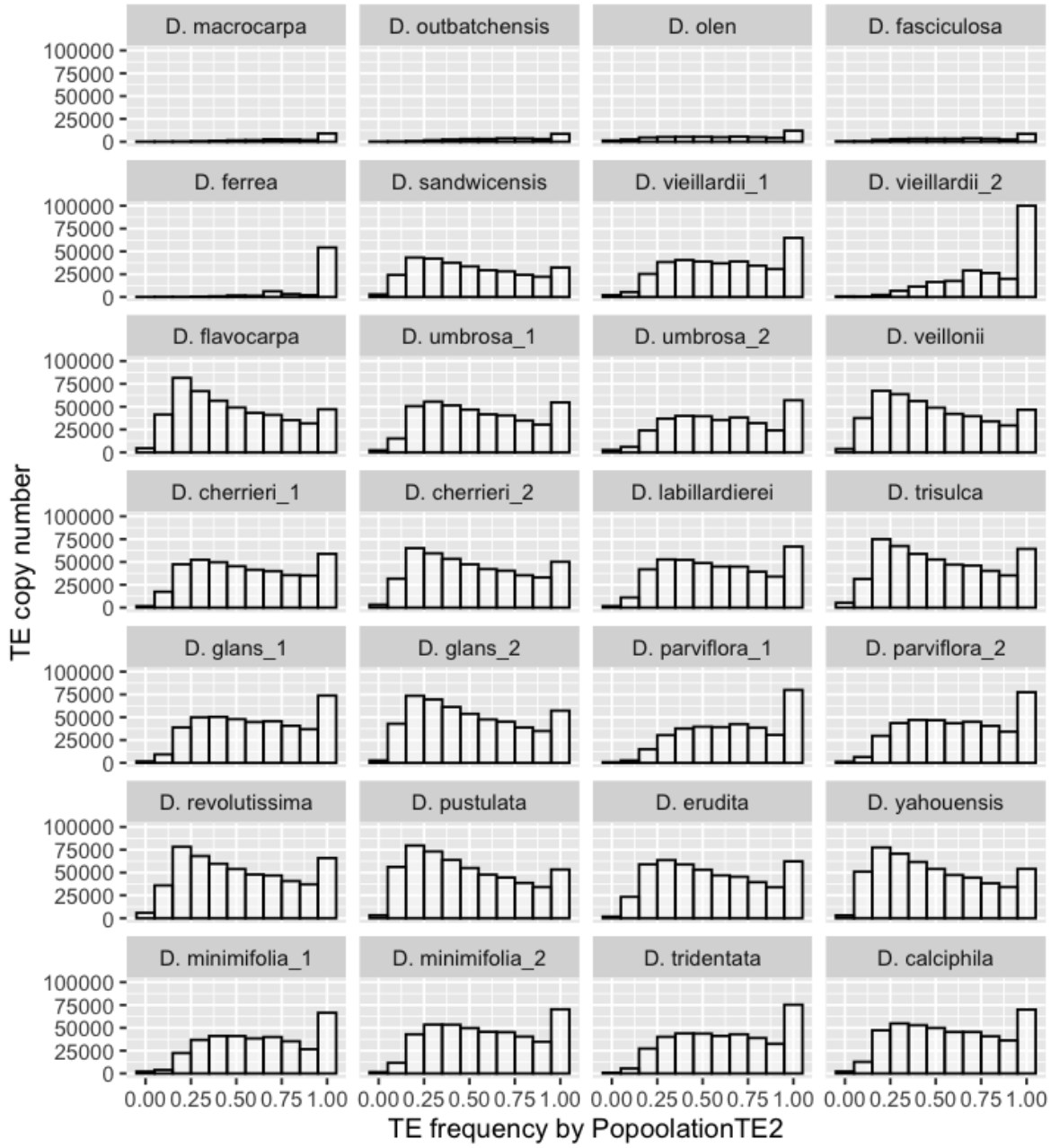

Figure 7. Histograms of TE frequencies per accession provided by Po.TE2. TE frequencies with more than three zero decimals (< 0.001) were rounded as zero.

Ngs_te_mapper2

Even though there was a big disparity in the TE counts and TE density distribution by Ngs.TE2 with respect to Po.TE2, the beanplot (Figure 8) of the TE insertions per family exhibited a similar pattern to the Po.TE2 results (Figure 6). The non-radiating species (mean of TE insertions 444.3 and mean of TE insertions per family 2.69) showed lower TE counts than the radiating clade (mean of TE insertions 9,323.4 and mean of TE insertions per family 7.23). The outgroup *D. ferrea* (2,053) showed lower TE insertions than the outgroup *D. sandwicensis* (4,609). The mean of TE insertions per family (black lines) of the outgroups *D. ferrea* and *D. sandwicensis* were placed between the two clades, where the first species got a higher TE insertion than the non-radiating species but lower than *D. sandwicensis*. The mean of TE counts of *D. sandwicensis* were close to the overall mean (3.46 in the log scale, see dashed line), in agreement with previous results by Po.TE2. Regarding the density distribution, the Ngs.TE density distribution was highly concentrated under the overall mean for all species. Nevertheless, the TE density shape was very different among the non-radiating and radiating species; the non-radiating species showed most of their TE distribution with low counts, while in the radiating species, the density shape was narrow beneath the overall mean and equally distributed above the overall mean until reaching high TE counts (see TE copy number per accession in Supplementary 8).

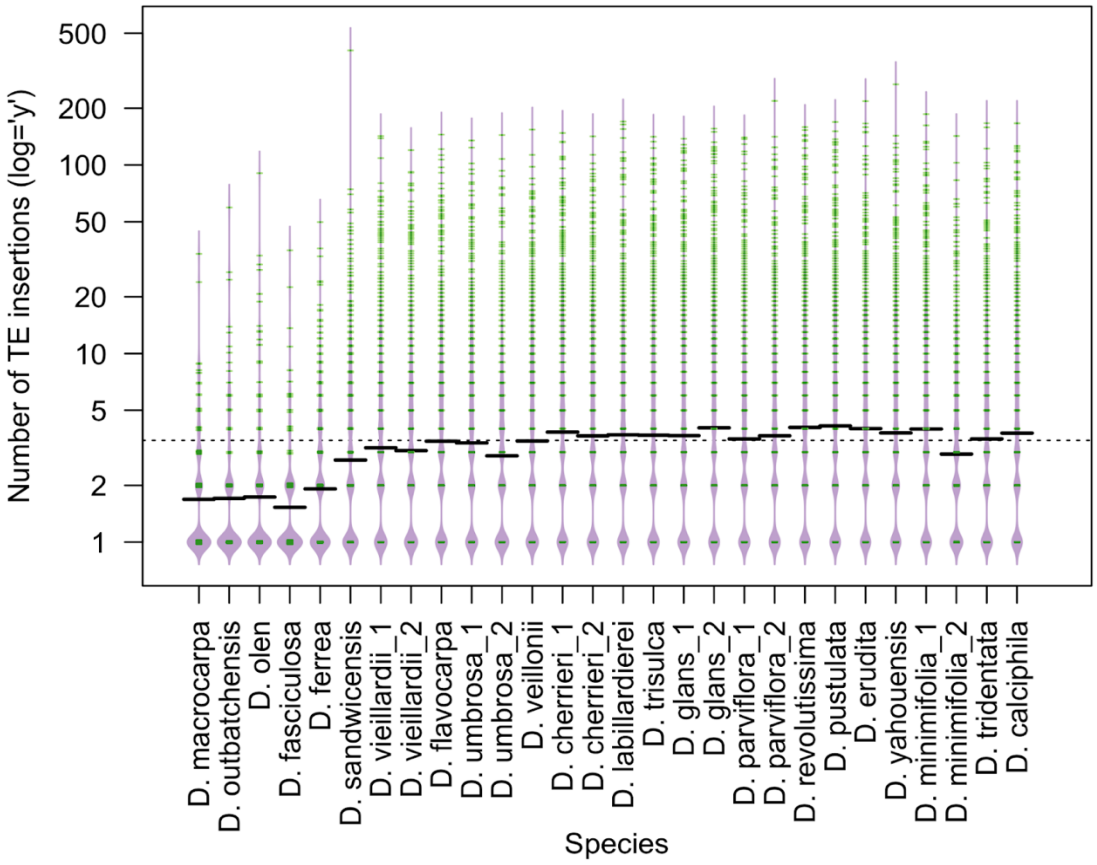

Figure 8. Beanplot of TE insertion counts for every TE family identified with Ngs.TE2. Green lines show TE insertion numbers on a log scale, wrapped in violet shapes that represent the density distribution of the TE insertions. The black lines are the average of TE insertions per accession. The dashed line shows the overall mean.

TEs hierarchical composition

Po.TE2 reported a total of 2,222 unique TE families for all 28 Diospyros species, of which 76.2% were TEs with no family name. Within the rest, the most representative TE orders were 17.8% LTRs, 4.3% TIRs (DNA transposons), 0.99% LINEs, and others 0.68%. On the other hand, Ngs.TE2 discovered 1,628 TE families over all 28 species, of which 76.1% were unknown TE families, 16.9% LTRs, 5% TIRs, and 1.2% LINEs.

The TE superfamily composition per species by Po.TE2 (Figure 9) consisted of 21 TE superfamilies, of which the non-radiated species had 19 superfamilies each of them, and every radiated species and outgroups had 21 superfamilies. However, the number of TE families for each superfamily had a moderate difference among the clades. The most predominant superfamilies were LTR_Copia and LTR_Gypsy, with the highest SD among the others, 12.73 and 32.44, respectively. LTR_Copia ranged from 110 to 146 in the non-radiating and from 152 to 161 TE families in the radiating and outgroups. On the other hand, LTR_Gypsy spanned 85 to 153 TE families in the non-radiating clade and 178 to 208 TE families in the radiating clade (see Supplementary 9).

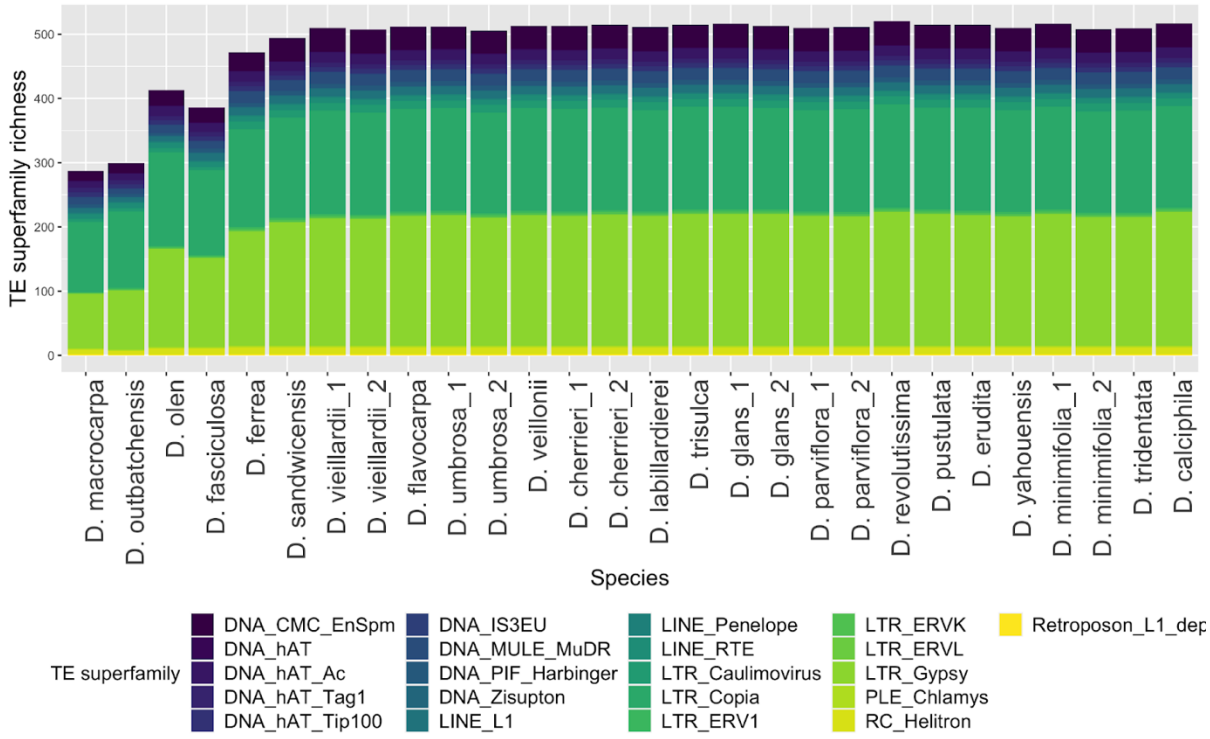

Figure 9. TE superfamily composition for the Diospyros species by Po.TE2. The y-axis depicts the number of TE families for each TE superfamily. Unknown TE families, repeat elements and rRNA were removed for better visual representation.

The TE superfamily composition by Ngs.TE2 comprised a total of 21 TE superfamilies (Figure 10), but it differed among the radiating and non-radiating species and outgroups. The non-radiating species, *D. macrocarpa* (8), *D. outbachensis* (8), *D. olen* (10), and *D. fasciculosa* (11), showed a less diverse TE superfamily composition than the outgroups *D. ferrea* and *D. sandwicensis (*18 and 19), respectively. The radiating species had the most diverse TE superfamilies, ranging from 18 to 21 TE superfamilies. As was the case in Po.TE2, here the most diverse TE superfamilies were also Copia and Gypsy; and the most variable between the non-divergence and the divergence clade (SD = 33.64 and 46.39, respectively). The TE superfamily richness for LTR_Copia and LTR_Gypsy was dramatically higher in the radiating clade than in the outgroups and the non-radiating species. LTR_Copia in the non-radiating group ranged from 4 to 21, while *D. ferrea* and *D. sandwicensis* had 56 and 77 TE superfamilies*,* and in the radiating clade, it ranged from 86 to 105. LTR_Gypsy in the no-radiating clade, from 11 to 16, in the outgroups 56 and 60 in the same order as above, and in the radiating clade, it was between 121 and 144. In the radiating clade, *D. minimifolia_2* and *D. umbrosa_2* had the lowest TE superfamily richness, but it did not occur in the replicated accession from the same species, so these results could be associated with their library quality (see Supplementary 10).

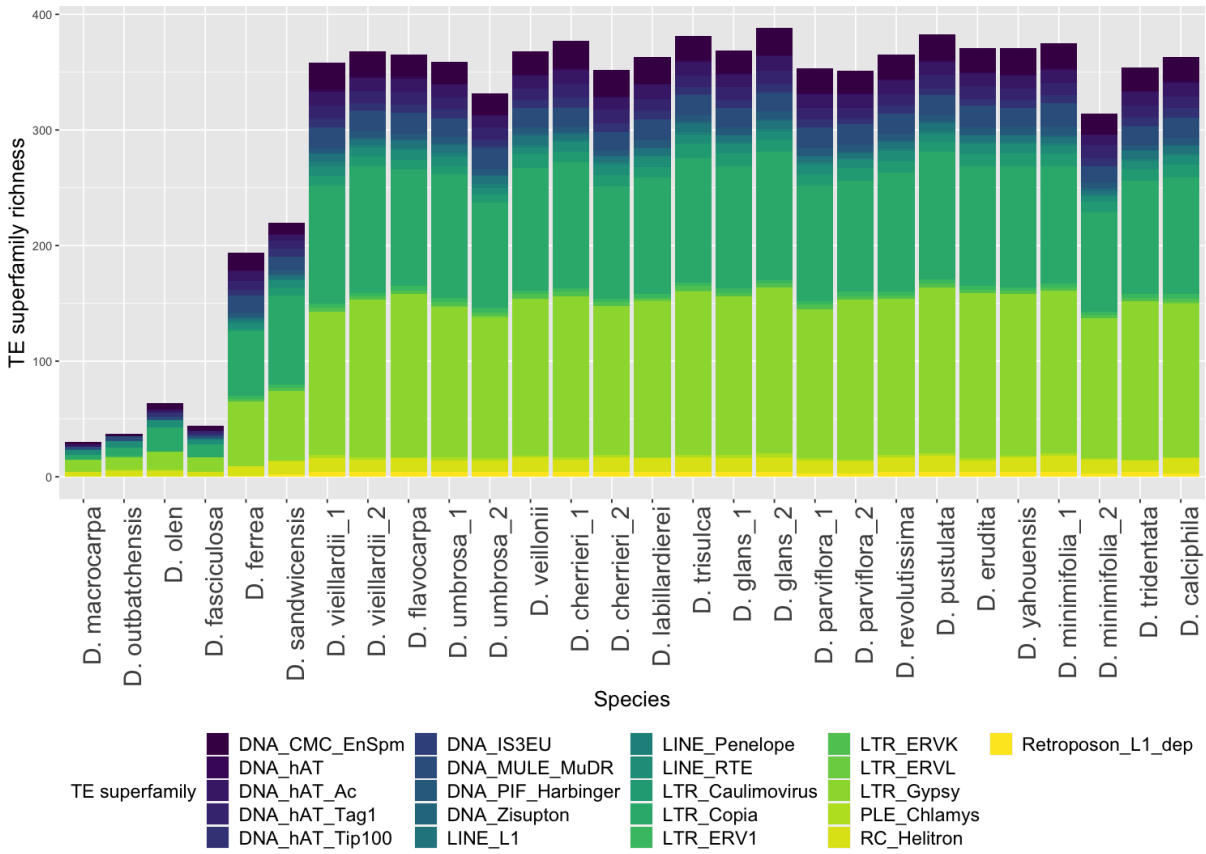

Figure 10. TE superfamily composition for the Diospyros species by Ngs.TE2. The y-axis depicts the number of TE families for each TE superfamily. Unknown TE families, repeat elements and rRNA were removed for better visualization.

TE activity in the phylogenetic tree of Diospyros

Because of the large number of TE families, we then selected a subset of the ten TE families with the highest number of TE insertions for each of the 28 species. Altogether this resulted in the retention of a total of 35 TE families, and their copy counts have been plotted on a phylogenetic tree (Figure 11). The TE families included 18 retrotransposons (Class 1) and 17 unknown TEs. Within the retrotransposons, six were LINEs (Long interspersed nuclear elements) and 12 were LTRs (Long terminal repeats). Within the LTRs, there were Gypsy (ten families) and Copia (two families). Within these 35 families were also included the ten TE families (indicated with red stars) with the most variation across the phylogeny based on the SD of their counts. It comprised seven Gypsy families, one Copia, and two unknown families. These results were in line with previous bar plots, which included all TEs discovered, but in contrast, here the unknown families were included.

Regarding the activity of the TE families, some LINE family members were present in the non-radiating species (LINE_RTE1-2,5-6), but they were absent in the other species. Only LINE_RTE4 showed steady growth towards the radiating clade. Conversely, Gypsy families were not uncovered in the non-radiating species, but a few of them were reported in the outgroups, *D. ferrea* and *D. sandwicensis* (LTR_Gypsy5 and LTR_Gypsy6). Some Gypsy families were present all over the radiating clade (e.g., LTR_Gypsy6), while other Gypsy families were absent in several species (e.g., LTR_Gypsy3) across the phylogeny. There were only two Copia families included in this set, where LTR_Copia1 exhibited increased TE activity across all species of the radiating clade; in contrast, LTR_Copia2 was reported only in *D. sandwicensis*. Likewise, the unknown TE families presented a high TE activity among the clades and outgroups, where some of them were only turned up in the non-radiated group and disappeared in the radiated clade, and vice versa. The TE activity was clearly noticeable across the species; it was higher in the radiating clade than in the non-radiating group.

Some species in the phylogeny had two accessions to determine whether there were differences within species. Two of these species, *D. vieillardii* and *D. umbrosa* had variable TE activity among the replicates. In contrast, *D. cherrierii_1* and *D. cherrierii_2*, and *D. minimifolia_1* and *D. minimifolia_2,* had more homogeneous TE activity. Furthermore, the TE activity was rather similar between the outgroups; aside from the TE copy number, *D. sandwicensis* had two LINEs, two Copia and two Gypsy, while *D. ferrea*, three LINEs, one Gypsy and one Copia.

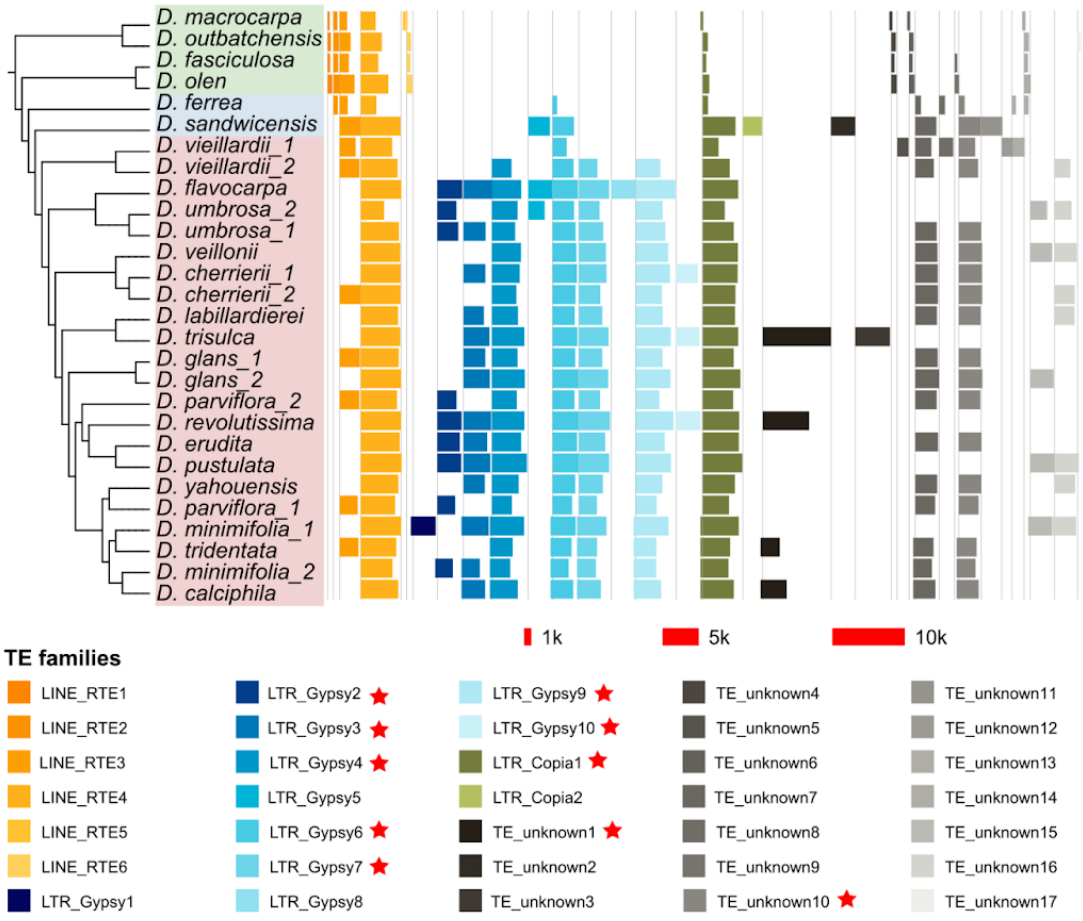

Figure 11. Phylogenetic tree of the Diospyros species and the top ten TE families in terms of the number of TE insertions for every species reported by Po.TE2. The clades in the phylogeny are grouped by colour: green represents the non-radiating species, blue shows the outgroups, and rose represents the radiating clade. The red bar sizes of 1k, 5k, and 10k represent 1,000, 5,000, and 10,000 TE insertion counts, respectively. The red stars represent the ten families with the highest SD (see the original names of the TE families in Supplementary 11).

Visualization of the variant activity within TE families

To reduce sequence redundancy, the consensus reference TE was clustered with cd-hit-est, resulting in a 40% reduction where 1,576 remained from 2,368 TE families. Of the ten most variable TE families from Po.TE2 (Figure 11), some representatives were further investigated with DeviaTE. Here we present the variant diversity for one of the Gypsy families (LTR_Gypsy2) (Figure 12) with 6,763 base pairs. DeviaTE also calculated the number of TE insertions, which was consistent with previous results: an increase of more than 10-fold in TE insertions in the radiating clade compared to the non-radiating New Caledonian species and

outgroups (Table 2). Variant diversity was presented for only three species, one member for each group (Figure 12). Coverage was unevenly distributed and had multiple peaks, which may be associated with the alignment method. For *D. macrocarpa*, there was a discrepancy in coverage between the ambiguous and unambiguous alignment reads, whereas in *D. ferrea* and *D. cherrierii_2* it was mostly unambiguous except in both endings. With respect to the variant diversity, all three species had a large number of SNPs: *D. macrocarpa*, 880 SNPs; *D. ferrea*, 1383 SNPs; and *D. cherrierii_2,* 1494 SNPs. In relation to fixed differences to the reference nucleotide, *D. macrocarpa* and *D. ferrea* had a small number of fixed variants (48 and 52, respectively), and *D. cherrierii_2.* had no fixed variant. Compared to *D. macrocarpa*, *D. ferrea* and *D. cherrierii_2* showed a few terminal deletions (58 and 1, accordingly), along with a very small number of insertions (8 and 4, respectively) and deletions (8 and 3, respectively).

Table 2. Number of TE insertions for LTR_Gypsy2 by DeviaTE. *hq=high coverage.

Accession	N° of TE insertions	N° TE insertions (hq* > 20)
D. macrocarpa	5.24	1.86
D. fasciculosa	4.97	1.68
D. ferrea	14.97	13.43
D. vieillardii_1	24.61	10.25
D. cherrierii_2	14.54	13.28
D. parviflora_1	49.83	18.04
D. minimifolia_1	55.51	16.73
D. minimifolia_2	52.97	16.84

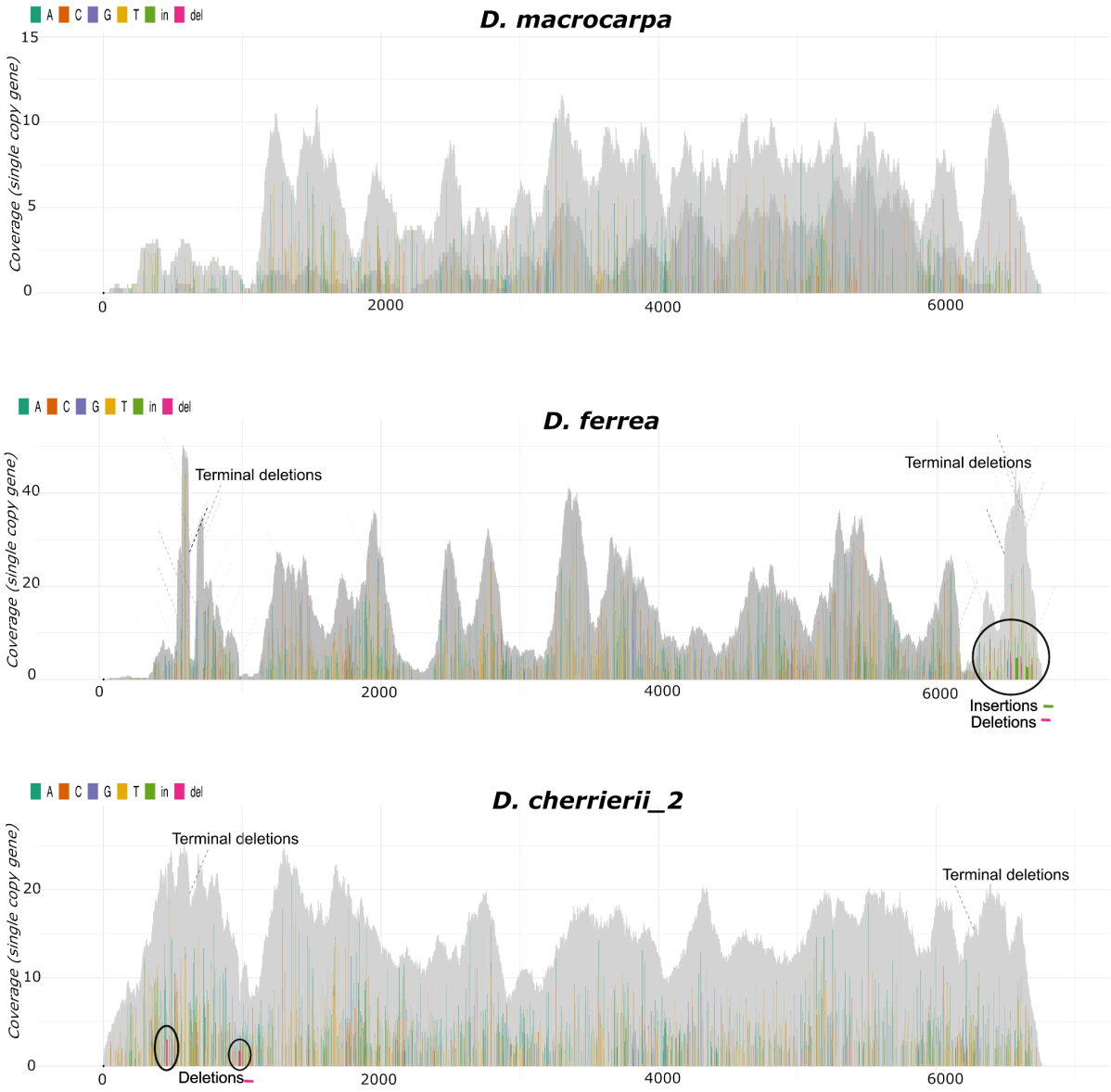

Figure 12. Variant diversity within a Gypsy family (LTR_Gypsy2) by using the consensus TE by RepeatModeller. The Y axis shows the coverage normalized with the single-copy gene method. The light grey colour shows the ambiguous coverage (defined by multiple mapping reads) and the dark grey colour the unambiguous coverage (defined by uniquely mapping reads). The lines with several colours are SNPs. The fixed variants are shown by lines of one unique colour that goes up to total coverage. The dashed lines represent terminal deletions. The black circles highlight insertions and deletions in *D. ferrea* and *D. cherrierii_2.*

On the other hand, clustering the EDTA consensus TE family resulted in a 16.65% redundancy reduction, where 8,458 remained of 10,147 TE families. One representative of Copia (10644 base pairs) and one of Gypsy (9811 base pairs) were selected for this analysis (Figure 13). The coverage of Copia and Gypsy illustrated a disproportion for the alignment, as was also

shown previously when DeviaTE was analysed with the consensus TE by RepeatModeller. The coverage was lower than in the previous results (Figure 11), despite the same alignment and clustering methods applied to both files. Here, the variant diversity difference relies on the terminal deletions occurring within the genome (Copia, 75, and Gypsy, 44) and internal deletions: one reported in Copia and one in Gypsy supported by the 26.18% and 13.74% of the reads at that particular position (Figure 13). Internal deletions were not present in the previous variant diversity (Figure 12). Nevertheless, SNPs, fixed variants, insertions, and deletions were in similar proportion to the results observed by using the RepeatModeller consensus TE.

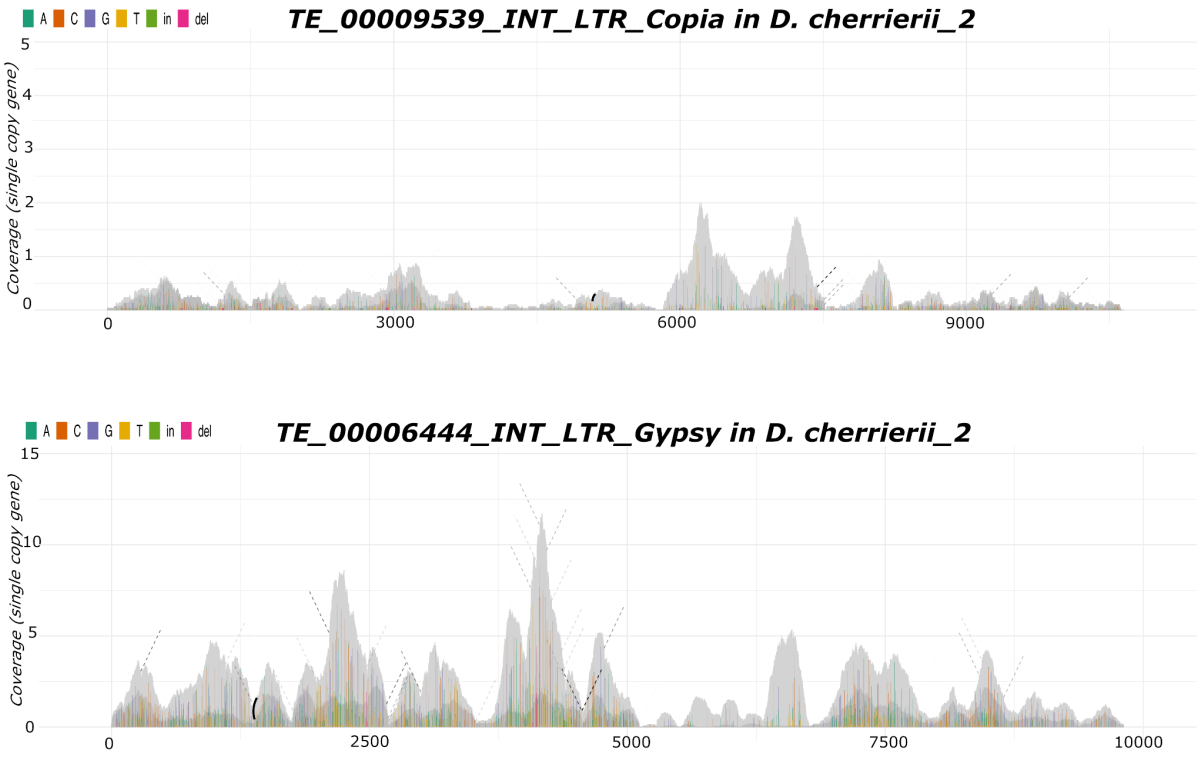

Figure 13. Variant diversity within one Copia and one Gypsy TE families by using the consensus TE by EDTA. The legend is the same as in Figure 11. Internal deletions are represented by the black arcs.

4. Discussion

Discovering TE insertions in Diospyros

To understand the TE dynamic in *Diospyros*, we used four different approaches with short read WGS data: a simple mapping and counting approach, Ngs_te_mapper2, PopoolationTE2, and DeviaTE. Despite the number of TE insertions found differed between the methods (in particular between Ngs.TE2 and Po.TE2), they agreed that there was an expansion of TE insertions in the radiating clade compared to the non-radiating species (Figures 6 and 7), which is also consistent with a previous study by Samuel et al. (2019) that found based on genome skimming nearly twice the proportion of repeatable elements in the radiating clade compared to non-radiating species. However, the extent of TE accumulation in the *D. ferrea* and *D. sandwicensis* outgroups was not entirely clear, as they showed contrasting results. Previous phylogenetic results by Samuel at al. (2019) and Turner et al. (2013a) have shown that *D. sandwicensis* is the closest relative of the largest *Diospyros* clade in New Caledonia, therefore, it was no surprise that Ngs.TE2 and PoTE2 predicted the TE insertions of this species parallel with the insertions in the radiating clade. On the contrary, the alignment method using BWA-MEM showed higher similarity between *D. ferrea* and the radiating clade. Nevertheless, the BWA-MEM algorithm was not developed with the aim of detecting TE insertions, and it was used here as a first approach to obtain an initial overview of the dynamics of TEs.

One of the reasons for the large difference (Figures 5, 6 and 7) in the number of TE insertions between the methods could be due to the differences in the pipeline approaches. Ngs.TE2 aims to detect TSDs as potential non-reference TE insertions (Han et al., 2021), and it requires the full alignment of TE sequence for reference TEs. TSDs are good predictors of TE superfamilies (e.g., Copia and Gypsy) because they have conserved regions and fixed sizes in LTRs and TIRs. In contrast, Po.TE2 aims to detect TE insertions by using a discordant alignment of paired-end reads (Kofler et al., 2016). Moreover, it is important to emphasise that each species was run separately. Thus, when TE insertions were compared across species, there may be a bias introduced by the library size (Table 1), which was variable and somewhat lower for the non-radiating species. On the contrary, DeviaTE included a normalisation method to estimate TE insertions counts by using the number of reads mapping the gene sequences which occur only once in the genome (Weilguny & Kofler, 2019). Despite the TE differences in the number of insertions between Ngs.TE2, Po.TE2 and DeviaTE, they agreed on the presence of a burst of TEs in the divergence species (Figures 7-8, and Table 2).

Characterisation of TE composition

Even though the difference in number of TE insertions among the methods was quite large, the taxonomic composition of orders (e.g., LTRs, LINEs, DNA transposons, etc.) and superfamily level (e.g., Gypsy) was consistent between Ngs.TE2 and Po.TE2 pipelines. However, the TE family abundance varied among these approaches (Figures 9 and 10). The retrotransposons LTRs were the dominant order across the *Diospyros* species in both methods, Ngs.TE2 and Po.TE2, consistent with previous TE studies in other groups of plants such as *Arabidopsis thaliana* (Quesneville, 2020), *Populus trichocarpa* (Ramakrishnan et al., 2022), *Vitis vinifera* (Ramakrishnan et al., 2022), etc.

In LTRs, the superfamilies Gypsy and Copia were the most abundant in both methods, but only Ngs.TE2 exhibited an evident expansion of these TE superfamilies in the radiating clade. The increased copy number of Gypsy and Copia could be the reason for the increased genome size in that clade (Table 1). These superfamilies were also the most abundant among the non-radiating species. The abundance of the LTRs has been already associated with increased genome sizes, reported in several groups of plants (Ramakrishnan et al., 2022), including *Diospyros* (Samuel et al., 2019).

TE composition families also lie in the consensus reference TE. Unfortunately, this file comprised 76.8% of unknown TE families due to the lack of TEs, which have a direct impact on the TE composition and limit the interpretation of TE activity. The consensus TE had a lack of MITE elements, which is one of the most abundant TE superfamilies in plants (Feschotte et al., 2002). MITEs are challenging to identify because of their small size and palindromic structure (Flynn et al., 2019). In contrast, the consensus TE produced by EDTA comprised 21.5% of unknown TE and it included MITEs. EDTA was not included in the main analysis of this thesis, but it suggests that the TE composition would have differed with this consensus TE reference.

Due to the lack of high-quality reference genomes for non-model species, a unique reference genome of *D. impolita* and a consensus TE reference of the same accession were used during this thesis, which could influence the data results, especially for more divergent species such as the outgroups and the non-radiating species.

TE dynamic during the adaptive radiation

TE family enrichment in the heatmaps (Figures 3 and 4) showed some insights into TE activity after the adaptive radiation. It found significant enrichment for retrotransposons (LTRs and L1) and DNA transposons. Copia and Gypsy showed a contrasting activity, Copia was

overrepresented in the species groups with less mapping counts in both heatmaps (TE clades 3, 4 and b), indicating a low activity and frequency in these genomes; while Gypsy was overrepresented in the TE clades with the highest counts (TE clades 1 and a) and underrepresented in low mapping TE clades. Copia and Gypsy were neither overrepresented or underrepresented in the same TE clade. DNA transposons (TIR/hAT-Tag1 and RC/Helitron) were more significantly active together in the TE clade c, which comprised a very high number of mapping counts. These outcomes were in line with Po.TE2, where Copia and Gypsy were the most variable and frequent TEs by comparing the non-radiating to the radiating species, after the unknown TEs (Figure 11).

The results of the intragenomic TE diversity by DeviaTE in one of the Gypsy (LTR_Gypsy2) superfamilies displayed a lot of polymorphic sites but a few fixed variants, as well as a few indels and terminal deletions. Although here it was shown for only one TE family, similar results were observed with other TE families. Perhaps one of the causes of the low coverage observed in both DeviaTE results (Figures 12 and 13) is due to the alignment method, which was performed separately with Bowtie2. Nevertheless, the high abundance of SNPs indicates a recent TE invasion in the radiating clade, but it is not conclusive due to the uneven distribution of the coverage and differences in the unique and multiple mapped reads.

5. Conclusions

Compared to the non-radiating species and outgroups, the radiating clade had, on average, 12-fold TEs (Po.TE2) and 21-fold TEs (Ngs.TE2), according to the TE insertions discovered in the *Diospyros* chosen species. The nearest relative of the radiating species, *D. sandwicensis*, displayed a copy number of TEs parallel to those of the radiating species, although it had a significantly smaller genome size. In contrast *D. ferrea*, showed TE insertion counts similar to those of the non-radiating species. The TE families were composed of DNA transposons, LINEs, and LTRs. The last retrotransposons were the most prevalent TEs, where Copia and Gypsy had the greatest number of TE families and the most variation. In terms of TE activity, by comparing the non-radiating and radiating species, Gypsy was the TE with the highest activity. Furthermore, Gypsy was significantly enriched in the TE clades with the highest mapping counts, while Copia was significantly enriched in the TE clades with the lowest mapping counts. The TE dynamics in the phylogenetic tree showed large differences in the number of TE insertions per TE family between species groups and among the most related species within the radiating clade. The variant diversity within *D. macrocapa*, *D. ferrea* and *D. cherrierii_2* exhibited a high number of SNPs in contrast to non-polymorphic sites, indels and terminal deletions.

6. References

Barrabé, L., Lavergne, S., Karnadi-Abdelkader, G., Drew, B. T., Birnbaum, P., & Gâteblé, G. (2018). Changing Ecological Opportunities Facilitated the Explosive Diversification of New Caledonian Oxera (Lamiaceae). *Systematic Biology*, *68*(3), 460–481. https://doi.org/10.1093/sysbio/syy070

Bennetzen, J. L., & Wang, H. (2014). The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology*, *65*(1), 505-530. https://doi.org/10.1146/annurev-arplant-050213-035811

Bourque, G., Burns, K. H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H. L., Macfarlan, T. S., Mager, D. L., & Feschotte, C. (2018). Ten things you should know about transposable elements. *Genome Biology*, *19*(1). https://doi.org/10.1186/s13059-018-1577-z

Chen, J., Basting, P. J., Han, S., Garfinkel, D. J., & Bergman, C. M. (2023). Reproducible evaluation of transposable element detectors with McClintock 2 guides accurate inference of Ty insertion patterns in yeast. *Mobile DNA*, *14*(1). https://doi.org/10.1186/s13100-023-00296-4

Dubin, M. J., Mittelsten Scheid, O., & Becker, C. (2018). Transposons: A blessing curse. *Current Opinion in Plant Biology*, *42*, 23-29. https://doi.org/10.1016/j.pbi.2018.01.003

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

Fedoroff, N. V. (2012). Transposable elements, epigenetics, and genome evolution. *Science*, *338*(6108), 758-767. https://doi.org/10.1126/science.338.6108.758

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C., & Smit, A. F. (2019). RepeatModeler2: Automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A. 2020 Apr 28;117(17)*, 9451-9457. https://doi.org/10.1101/856591

Grant, P., & Grant, B. R. (2002). Adaptive radiation of Darwin's finches. *American Scientist*, *90*(2), 130-139. https://doi.org/10.1511/2002.2.130

Han, S., Basting, P. J., Dias, G. B., Luhur, A., Zelhof, A. C., & Bergman, C. M. (2021). Transposable element profiles reveal cell line identity and loss of heterozygosity in *Drosophila* cell culture. *Genetics*, *219*(2). https://doi.org/10.1093/genetics/iyab113

Han, S. & Bergman, C. M. (2020, December 01). *Ngs_te_mapper2: A program to identify transposable element insertions using next generation sequencing data* v. 2. Github repository. https://github.com/bergmanlab/ngs_te_mapper2

Hof, A. E., Campagne, P., Rigden, D. J., Yung, C. J., Lingley, J., Quail, M. A., Hall, N., Darby, A. C., & Saccheri, I. J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. *Nature*, *534*(7605), 102-105. https://doi.org/10.1038/nature17951

Kent, T. V., Uzunović, J., & Wright, S. I. (2017). Coevolution between transposable elements and recombination. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1736), 20160458. https://doi.org/10.1098/rstb.2016.0458

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

Krueger, F., James, F., Ewels, P., Afyounian, E. & Schuster-Boeckler, B. (2021). FelixKrueger/TrimGalore: v0.6.7 - DOI via Zenodo (0.6.7). Zenodo. https://doi.org/10.5281/zenodo.5127899

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

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: An online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, *44*(W1), W242-W245. https://doi.org/10.1093/nar/gkw290

Levin, S. A., Carpenter, S. R., Godfray, H. C., Kinzig, A. P., Loreau, M., Losos, J. B., Walker, B., & Wilcove, D. S. (2009). Adaptive radiation. In *The Princeton guide to ecology* (pp. 143-152). Princeton University Press.

Li H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997v2 [q-bio.GN]

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The sequence alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078-2079. https://doi.org/10.1093/bioinformatics/btp352

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

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

Liao, Y., Smyth, G. K., & Shi, W. (2013a). FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923-930. https://doi.org/10.1093/bioinformatics/btt656

Liao, Y., Smyth, G. K., & Shi, W. (2013b). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, *41*(10), e108-e108. https://doi.org/10.1093/nar/gkt214

Manni, M., Berkeley, M. R., Seppey, M., & Zdobnov, E. M. (2021). BUSCO: Assessing Genomic data quality and beyond. *Current Protocols*, *1*, e323. https://doi.org/10.1002/cpz1.323

Morat, P., Jaffré, T., Tronchet, F., Munzinger, J., Pillon, Y., Veillon, J., Chalopin, M., Birnbaum, P., Rigault, F., Dagostini, G., Tinel, J., & Lowry, P. P. (2012). Le référentiel taxonomique Florical et Les caractéristiques de la flore vasculaire indigene de la nouvelle-calédonie. *Adansonia*, *34*(2), 179-221. https://doi.org/10.5252/a2012n2a1

Oliver, K. R., & Greene, W. K. (2009). Transposable elements: powerful facilitators of evolution. *BioEssays*, *31*(7), 703–714. https://doi.org/10.1002/bies.200800219

Ou, S., Su, W., Liao, Y., Chougule, K., Ware, D., Peterson, T., Jiang, N., Hirsch, C. N., & Hufford, M. B. (2019). Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol*, *20*, 275. https://doi.org/10.1101/657890

Paun, O., Turner, B., Trucchi, E., Munzinger, J., Chase, M. W., & Samuel, R. (2016). Processes Driving the Adaptive Radiation of a Tropical Tree (Diospyros, Ebenaceae) in New Caledonia, a Biodiversity Hotspot. *Systematic Biology*, *65*(2), 212–227. https://doi.org/10.1093/sysbio/syv076

Pedro D.L.F., Amorim T.S., Varani A., Guyot, R., Domingues, D.S. & Paschoal, A.R. (2021). An Atlas of Plant Transposable Elements. F1000Res. Nov 24; 10:1194. Doi: 10.12688/f1000research.74524.1. PMID: 35035898; PMCID: PMC8729191.

Pillon, Y. (2012). Time and tempo of diversification in the flora of New Caledonia. *Botanical Journal of the Linnean Society*, *170*(3), 288-298. https://doi.org/10.1111/j.1095-8339.2012.01274.x

Quesneville, H. (2020). Twenty years of transposable element analysis in the arabidopsis thaliana genome. *Mobile DNA*, *11*(1). https://doi.org/10.1186/s13100-020-00223-x

Ramakrishnan, M., Satish, L., Sharma, A., Kurungara Vinod, K., Emamverdian, A., Zhou, M., & Wei, Q. (2022). Transposable elements in plants: Recent advancements, tools and prospects. *Plant Molecular Biology Reporter*, *40*(4), 628-645. https://doi.org/10.1007/s11105-022-01342-w

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

RStudio Team (2022). RStudio: Integrated Development Environment for R. RStudio, PBC, Boston, MA. http://www.rstudio.com/

Ronco, F., Matschiner, M., Böhne, A., Boila, A., Büscher, H. H., El Taher, A., Indermaur, A., Malinsky, M., Ricci, V., Kahmen, A., Jentoft, S., & Salzburger, W. (2021). Drivers and dynamics of a massive adaptive radiation in cichlid fishes. *Nature*, *589*(7840), 76–81. https://doi.org/10.1038/s41586-020-2930-4

Samuel, R., Turner, B., Duangjai, S., Munzinger, J., Paun, O., Barfuss, M. H. J., & Chase, M. W. (2019). Systematics and evolution of the Old World Ebenaceae, a review with emphasis on the large genusDiospyrosand its radiation in New Caledonia. *Botanical Journal of the Linnean Society*, *189*(2), 99–114. https://doi.org/10.1093/botlinnean/boy081

Schrader, L., & Schmitz, J. (2018). The impact of transposable elements in adaptive evolution. *Molecular Ecology*, *28*(6), 1537–1549. https://doi.org/10.1111/mec.14794

Schenk, J. J. (2021). The Next Generation of Adaptive Radiation Studies in Plants. *International Journal of Plant Sciences*, *182*(4), 245–262. https://doi.org/10.1086/713445

Sato, A., Tichy, H., O'hUigin, C., Grant, P. R., Grant, B. R., & Klein, J. (2001). On the Origin of Darwin's Finches. *Molecular Biology and Evolution*, *18*(3), 299–311. https://doi.org/10.1093/oxfordjournals.molbev.a003806

Tebbich, S., Sterelny, K., & Teschke, I. (2010). The tale of the finch: adaptive radiation and behavioural flexibility. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1543), 1099–1109. https://doi.org/10.1098/rstb.2009.0291

Turner, B., Munzinger, J., Duangjai, S., Temsch, E. M., Stockenhuber, R., Barfuss, M. H., . . . Samuel, R. (2013a). Molecular phylogenetics of New Caledonian Diospyros (Ebenaceae) using plastid and nuclear markers. *Molecular Phylogenetics and Evolution*, *69*(3), 740–763. https://doi.org/10.1016/j.ympev.2013.07.002

Turner, B., Paun, O., Munzinger, J., Duangjai, S., Chase, M. W., & Samuel, R. (2013b). Analyses of amplified fragment length polymorphisms (AFLP) indicate rapid radiation of Diospyros species (Ebenaceae) endemic to New Caledonia. *BMC Evolutionary Biology*, *13*(1), 269. https://doi.org/10.1186/1471-2148-13-269

Weilguny, L., & Kofler, R. (2019). DeviaTE: Assembly-free analysis and visualization of mobile genetic element composition. *Molecular Ecology Resources*, *19*(5), 1346-1354. https://doi.org/10.1111/1755-0998.13030

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973-982. https://doi.org/10.1038/nrg2165

Appendix

Supplementary 1. Scripting

```
"""Bash scripting"""

# The scripts below contain some examples of the options used to run each
software/pipeline
```

```
# Trimming with Trimgalore
```

```
#!/bin/bash
#
#SBATCH --job-name=array_trimgalore
#SBATCH --constraint=array-8core
#SBATCH --array=0-12
#SBATCH --cpus-per-task=4
#SBATCH --mem=1000M
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --output=log/%x-%j-%N.err
#SBATCH --partition=basic

cd /scratch/botany/mayra/diospyros/new_accessions_bam
samples=(calciphilaBT313 eruditaBT275 flavocarpaBT128 glansBT093 glansBT094
pustulataBT137 revolutissimaBT219_merge13 tridentataBT206 trisBT192_trisulcaBT192
umbrosaBT251 umbrosaBT255 veilloniiBT227 yahouensisBT238)
module load conda
conda activate trim-galore-0.6.10
trim_galore --paired --fastqc -o
/scratch/botany/mayra/diospyros/new_accessions_bam_trim ${samples[
$SLURM_ARRAY_TASK_ID ]}_R1.fq.gz ${samples[ $SLURM_ARRAY_TASK_ID ]}_R2.fq.gz
```

```
# Mapping with BWA-MEM
```

```
#!/bin/bash
#
#SBATCH --job-name=bwa
#SBATCH --cpus-per-task=1
#SBATCH --time=1-00:00:00
#SBATCH --mem=16000MB
#SBATCH --mail-type=ALL
#SBATCH --output=/scratch/botany/mayra/diospyros/bwa.out
#SBATCH --output=/scratch/botany/mayra/diospyros/bwa.err
#SBATCH --partition=basic

module load bwa
module load samtools

cd /scratch/botany/mayra/diospyros/fastqc_trim
samples=(cherrieriBT262 fasciculosaBT013 labillardiereiBT123 macrocarpaBT047
minimifoliaBT131 minimifoliaD4G olenBT184 parvifloraBT060 parvifloraBT288
vieillardiiBT023)
bwa index /scratch/botany/mayra/diospyros/impolita-families.fa
bwa mem /scratch/botany/mayra/diospyros/impolita-families.fa -t 16
ferreaSD2012.1_val_1.fq.gz ferreaSD2012.2_val_2.fq.gz | samtools view -Sb |
samtools sort -@16 -o ferreaSD2012_sort.bam
```

```
# Mapping stats with Samtools flagstat
```

```
#!/bin/bash
#
#SBATCH --job-name=flagstat10
#SBATCH --time=02-00:00:00
```

```
#SBATCH --array=0-10
#SBATCH --cpus-per-task=4
#SBATCH --mem=3000M
#SBATCH --mail-type=ALL
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

module load samtools
cd /scratch/botany/mayra/diospyros/bowtie2_outputs
samples=(cherrieriBT262 fasciculosaBT013 labillardiereiBT123 macrocarpaBT047
minimifoliaBT131 minimifoliaD4G olenBT184 parvifloraBT060 parvifloraBT288
vieillardiiBT024 cherrierBT293)

for i in /scratch/botany/mayra/diospyros/bowtie2_outputs/*cdhit4_aligned.bam
do
samtools flagstat -O tsv $i > $i.txt
done
```

```
# Summarising the mapping reads with FeatureCounts
```

```
#!/bin/bash
#
#SBATCH --job-name=featurecounts
#SBATCH --cpus-per-task=8
#SBATCH --mem=16GB
#SBATCH --time=00-12:00:00
#SBATCH --mail-type=ALL
#SBATCH --output=/scratch/botany/mayra/diospyros/featurecounts.out
#SBATCH --output=/scratch/botany/mayra/diospyros/featurecounts.err
#SBATCH --partition=basic
module load subread

cd /scratch/botany/mayra/diospyros

#samples=(cherrieriBT262 fasciculosaBT013 labillardiereiBT123 macrocarpaBT047
minimifoliaBT131 minimifoliaD4G olenBT184 parvifloraBT060 parvifloraBT288
vieillardiiBT023)

featureCounts -a /scratch/botany/mayra/diospyros/impolita_consensus.gff -o
./counts_impolita2.txt -g ID -t TE -p -M --fraction -T 4
/scratch/botany/mayra/diospyros/fastqc_trim_new/*bam
```

```
# Detecting TEs with McClintock 2
# Phase 1: McClintock 2 generated the consensus TE in GFF and the TE taxonomy in
TSV formats
```

```
#!/bin/bash
#
#SBATCH --job-name=mcclintock_diospyros_16
#SBATCH --time=12:00:00
#SBATCH --cpus-per-task=16
#SBATCH --mem=20000
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

module load conda
conda activate mcclintock-2.0.0
pushd $TMPDIR
mcclintock.py -r
/scratch/botany/katie/assembled_genomes/assemblies/impolita/working/impolita.fasta
-c /scratch/botany/mayra/diospyros/impolita-families_mcclintok7.fa  -p 4 -o
mcclintock_annotations_16 --make_annotations
popd
mv $TMPDIR/mcclintock_annotations_16 .
```

```
# Detecting TEs with McClintock 2
# Phase 2: McClintock 2 used the output files from Phase 1 to detect TEs
```

```
#!/bin/bash
#
#SBATCH --job-name=mccl_newaccessions
#SBATCH --time=10-00:00:00
#SBATCH --constraint=array-8core
#SBATCH --array=0-4
#SBATCH --cpus-per-task=8
#SBATCH --mem=50000M
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

module load conda
conda activate mcclintock-2.0.0
data_folder=/scratch/botany/mayra/diospyros
samples=(eruditaBT275 glansBT094 pustulataBT137 revolutissimaBT219_merge13
trisBT192_trisulcaBT192)
pushd $TMPDIR
rsync -a /scratch/botany/mayra/diospyros/mcclintock_annotations_new/
mcclintock_annotations_${samples[ $SLURM_ARRAY_TASK_ID ]}/
mcclintock.py -r $data_folder/impolita.fasta -c $data_folder/impolita-
families_mcclintok7.fa -m ngs_te_mapper2,popoolationte2 -1
$data_folder/new_accessions_bam_trim/${samples[ $SLURM_ARRAY_TASK_ID
]}_R1_val_1.fq.gz -2 $data_folder/new_accessions_bam_trim/${samples[
$SLURM_ARRAY_TASK_ID ]}_R2_val_2.fq.gz -p 8 -o mcclintock_annotations_${samples[
$SLURM_ARRAY_TASK_ID ]} --resume
popd
mv $TMPDIR/mcclintock_annotations_${samples[ $SLURM_ARRAY_TASK_ID ]} .
```

```
# Data preparation to execute it in DEVIATE
```

```
# Cluster with CD-HIT-EST

#!/bin/bash
#
#SBATCH --job-name=cd-hit
#SBATCH --time=00-05:00:00
#SBATCH --cpus-per-task=1
#SBATCH --mem=1000M
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

module load cdhit
cd /scratch/botany/mayra/diospyros

cd-hit-est -i /scratch/botany/mayra/diospyros/impolita-
families_mcclintok9_mod.EDTA.TElib.fa -o impolita-
families_mcclintok9_mod.EDTA.TElib.cdhit_08_n5 -c 0.8 -n 5

#Alignment with Bowtie2
#!/bin/bash
#
#SBATCH --job-name=bowtie4_macrocarpa_local
#SBATCH --time=02-00:00:00
#SBATCH --cpus-per-task=2
#SBATCH --mem=5000M
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

#samples=(cherrieriBT262 fasciculosaBT013 labillardiereiBT123 macrocarpaBT047
minimifoliaBT131 minimifoliaD4G olenBT184 parvifloraBT060 parvifloraBT288)
data_folder=/scratch/botany/mayra/diospyros
```

```
module load bowtie2
bowtie2 -p 4 -q -1
$data_folder/accessions/fastqc_trim/macrocarpaBT047_R1_val_1.fq.gz -2
$data_folder/accessions/fastqc_trim/macrocarpaBT047_R2_val_2.fq.gz -x
$data_folder/impolita.fasta.mod.EDTA.TElib_cdhit5_08_n5_sed_cat_seqkt50 -S
$data_folder/bowtie2_outputs_new/macrocarpa_cdhit5_edta_aligned.sam --local --very-
sensitive-local
module load samtools
samtools view -h -S -b
$data_folder/bowtie2_outputs_new/macrocarpa_cdhit5_edta_aligned.sam >
$data_folder/bowtie2_outputs_new/macrocarpa_cdhit5_local_edta_aligned.bam
rm $data_folder/bowtie2_outputs_new/macrocarpa_cdhit5_local_edta_aligned.sam
```

```
#DEVIATE
```

```
#!/bin/bash
#
#SBATCH --job-name=deviate_macrocarpa_bwa
#SBATCH --time=01-00:00:00
#SBATCH --cpus-per-task=2
#SBATCH --mem=5000M
#SBATCH --output=log/%x-%j-%N.out
#SBATCH --error=log/%x-%j-%N.err
#SBATCH --mail-type=ALL

data_folder=/scratch/botany/mayra/diospyros
module load conda
conda activate deviaTE_env
module load samtools

deviaTE --input_bam
$data_folder/bowtie2_outputs_new/macrocarpa_cdhit5_local_edta_aligned.bam --
families
TE_00005612_INT_LTR_Copia,TE_00001282_INT_LTR_Copia,TE_00009539_INT_LTR_Copia,TE_00
008076_INT_LTR_Copia,TE_00003705_INT_LTR_Copia,TE_00009883_INT_LTR_Gypsy,TE_0000644
4_INT_LTR_Gypsy,TE_00002613_INT_LTR_Gypsy,TE_00005857_INT_LTR_Gypsy,TE_00000640_INT
_LTR_Gypsy --single_copy_genes Gene1,Gene2,Gene3,Gene4,Gene5 --library
$data_folder/impolita.fasta.mod.EDTA.TElib_cdhit5_08_n5_sed_cat_seqkt50.fa
```

```
"""R scripting"""
```

```
# Plotting heatmaps
```

```
# Read the file
counts <- as.matrix(read.table("counts.txt"))
head(counts)
#Select only the radiating species and D. ferrea
counts_select=counts[,c(1,3,5:6,8:11)]
head(counts_select)

# Plot heatmap with log values
par(mar=c(5,7,10,5))
png(filename="counts_heatmap.png", width=850, height=750)
pdf(file="counts_heatmap.pdf", width=7, height=6)
heatmap.2(log(counts+1), trace="none", cexRow=0.8,cexCol = 1.5,
key.par=list(mar=c(3,3,5,3)), labRow = NA, key.ylab="Count", margins=c(15,5),
main="CPM TE content",key.xlab = "logCPM", keysize = 1.5 , key=TRUE, labCol = c("D.
cherrierii_1","D. fasciculosa","D. labillardierei", "D. macrocarpa","D.
minimifolia_1", "D. minimifolia_2","D. olen","D. parviflora_1", "D. parviflora_2",
"D. vieillardii", "D. ferrea","D. sandwicensis"), col=
colorRampPalette(c("lemonchiffon", "lemonchiffon", "yellow","orange", "red",
"magenta", "midnightblue", "midnightblue", "black", "black "))(100))
graphics.off()
dev.off()
```

```
# Plot heatmap with log values for the selected species
png(filename="counts_select_heatmap.png", width=850, height=750)
pdf(file=" counts_select _heatmap.pdf", width=7, height=6)
heatmap.2(log(counts_select+1), trace="none", cexRow=0.8,cexCol = 1.5,
key.par=list(mar=c(3,3,5,3)), labRow = NA, key.ylab="Count", margins=c(15,5),
main="CPM TE content",key.xlab = "logCPM", keysize = 1.5, key=TRUE, labCol = c("D.
cherrierii_1", "D. labillardierei", "D. minimifolia_1", "D. minimifolia_2","D.
parviflora_1", "D. parviflora_2", "D. vieillardii", "D. ferrea"), col=
colorRampPalette(c("lemonchiffon", "lemonchiffon", "yellow","orange", "red",
"magenta", "midnightblue", "midnightblue", "black", "black "))(100))
graphics.off()
dev.off()
```

```
# Plotting boxplots for all TE insertions detected per method
```

```
# Creating a column for the TE method names
method=rep(c("Po.TE2"), times=53034)
te_insertions_popte2=cbind(te_insertions_all, method)
str(te_insertions_popte2)

method2=rep(c("Ngs.TE2"), each=30439)
te_insertions_ngs.te2=cbind(te_insertions_ngs_te_mapper2,method2)
str(te_insertions_ngs.te2)

# Filtering the columns needed
library(tidyverse)
te_insertions_ngs.te2_select=te_insertions_ngs.te2 %>%
select(count,accession,method2)
head(te_insertions_ngs.te2_select)
colnames(te_insertions_ngs.te2_select)=c("count","accession","method")

te_insertions_popte2_select=te_insertions_popte2 %>% select(count,accession,method)
head(te_insertions_popte2_select)

# Join files by rows
te_insertions=rbind(te_insertions_popte2_select,te_insertions_ngs.te2_select)
head(te_insertions)

# Plotting the boxplot
library(ggplot2)
ggplot_te_all=ggplot(te_insertions, aes(x = method, y = count, color = method))
ggplot_te_all_boxplot=ggplot_te_all + geom_boxplot() + scale_y_log10() + labs(y =
"counts (log10)") + theme(axis.text = element_text(size = 11))
show(ggplot_te_all_boxplot)
```

```
# Plotting TE frequency by PopoolationTE2
```

```
# Transform number as numeric

teinsertions_frequency_all_numeric=transform(teinsertions_frequency_all, te_freq =
as.numeric(te_freq))

# Convert to factor and change the order of levels
new_order=c("macrocarpaBT047","outbatchensisBT163","olenBT184","fasciculosaBT013","
ferreaSD2012","sandwiceG","vieillardiiBT286","vieillardiiBT024","flavocarpaBT128","
umbrosaBT251","umbrosaBT255","veilloniiBT227","cherrierBT293","cherrieriBT262","lab
illardiereiBT123","trisulcaBT192","glansBT093","glansBT094","parvifloraBT060","parv
ifloraBT288","revolutissimaBT219","pustulataBT137","eruditaBT275","minimifoliaBT131
","minimifoliaD4G","yahouensisBT238","tridentataBT206","calciphilaBT313")
teinsertions_frequency_all_numeric$accession=factor(teinsertions_frequency_all_nume
ric$accession, levels=new_order)
# Change the name of levels
levels(teinsertions_frequency_all_numeric$accession)=c("D. macrocarpa","D.
outbatchensis","D. olen","D. fasciculosa","D. ferrea","D. sandwicensis","D.
vieillardii_1","D. vieillardii_2","D. flavocarpa","D. umbrosa_1","D. umbrosa_2","D.
veillonii","D. cherrieri_1","D. cherrieri_2","D. labillardierei","D. trisulca","D.
```

```
glans_1","D. glans_2","D. parviflora_1","D. parviflora_2","D. revolutissima","D.
pustulata","D. erudita","D. yahouensis","D. minimifolia_1","D. minimifolia_2","D.
tridentata","D. calciphila")
# Plot the histogram
library(ggplot2)
te_freq_wrap=ggplot(teinsertions_frequency_all_numeric, aes(x=te_freq)) +
geom_histogram(aes(y=..count..),binwidth=0.1, colour="black", fill="white",
alpha=.5, position="identity") + facet_wrap(vars(accession), ncol=4) +  labs(x =
"TE frequency by PopoolationTE2", y="TE abundance")
show(te_freq_wrap)
```

```
# Plotting beanplots
```

```
# For Ngs_te_mapper2
# Convert a factor and change the order of the levels
te_insertions_ngs_te_mapper2$accession=factor(te_insertions_ngs_te_mapper2$accessio
n ,
levels=c("macrocarpaBT047","outbatchensisBT163","olenBT184","fasciculosaBT013","fer
reaSD2012","sandwiceG","vieillardiiBT286","vieillardiiBT024","flavocarpaBT128","umb
rosaBT251","umbrosaBT255","veilloniiBT227","cherrieriBT293","cherrieriBT262","labil
lardiereiBT123","trisulcaBT192","glansBT093","glansBT094","parvifloraBT060","parvif
loraBT288","revolutissimaBT219","pustulataBT137","eruditaBT275","yahouensisBT238","
minimifoliaBT131","minimifoliaD4G","tridentataBT206","calciphilaBT313"))
# Change the label of the levels
label=c("D. macrocarpa","D. outbatchensis","D. olen","D. fasciculosa","D.
ferrea","D. sandwicensis","D. vieillardii_1","D. vieillardii_2","D. flavocarpa","D.
umbrosa_1","D. umbrosa_2","D. veillonii","D. cherrieri_1","D. cherrieri_2","D.
labillardierei","D. trisulca","D. glans_1","D. glans_2","D. parviflora_1","D.
parviflora_2","D. revolutissima","D. pustulata","D. erudita","D. yahouensis","D.
minimifolia_1","D. minimifolia_2","D. tridentata","D. calciphila")
# Plot beanplot with beanlines and overalline as mean
library(beanplot)
beanplot_ngs=beanplot(count ~
te_insertions_ngs_te_mapper2$accession,data=te_insertions_ngs_te_mapper2, col =
c("#CAB2D6", "#33A02C", "#B2DF8A"), border = "#CAB2D6", boxwex = 1, at = 1:28 +
0.2, log="y", method="jitter", overallline = "mean", beanlines = "mean",las=2,
cex.axis = 1, cex.lab=1, ylab="Number of TE insertions (log='y')", names=label,
par(mar = c(12, 5, 4, 2)+ 0.1), xlab=xaxis, main="TE insertions by Ngs.TE2
(overalline and beanlines are mean)")
# Add the x-axis label
xaxis=mtext("Species", side = 1, line = 7)
# For PopoolationTE2
# Convert a factor and change the order of the levels
te_insertions_all$accession=factor(te_insertions_all$accession ,
levels=c("macrocarpaBT047","outbatchensisBT163","olenBT184","fasciculosaBT013","fer
reaSD2012","sandwiceG","vieillardiiBT286","vieillardiiBT024","flavocarpaBT128","umb
rosaBT251","umbrosaBT255","veilloniiBT227","cherrieriBT293","cherrieriBT262","labil
lardiereiBT123","trisulcaBT192","glansBT093","glansBT094","parvifloraBT060","parvif
loraBT288","revolutissimaBT219","pustulataBT137","eruditaBT275","yahouensisBT238","
minimifoliaBT131","minimifoliaD4G","tridentataBT206","calciphilaBT313"))
# Change the label of the levels
label=c("D. macrocarpa","D. outbatchensis","D. olen","D. fasciculosa","D.
ferrea","D. sandwicensis","D. vieillardii_1","D. vieillardii_2","D. flavocarpa","D.
umbrosa_1","D. umbrosa_2","D. veillonii","D. cherrieri_1","D. cherrieri_2","D.
labillardierei","D. trisulca","D. glans_1","D. glans_2","D. parviflora_1","D.
parviflora_2","D. revolutissima","D. pustulata","D. erudita","D. yahouensis","D.
minimifolia_1","D. minimifolia_2","D. tridentata","D. calciphila")
library(beanplot)
# Plot beanplot with beanlines and overalline as mean
beanplot=beanplot(count ~ te_insertions_all$accession,data=te_insertions_all, col =
c("#CAB2D6", "#33A02C", "#B2DF8A"), border = "#CAB2D6", boxwex = 1, at = 1:28 +
0.2, log="y", method="jitter", overalline="mean", beanlines="mean",las=2, cex.axis
= 1, cex.lab=1, ylab="Number of TE insertions (log='y')", names=label, par(mar =
c(12, 5, 4, 2)+ 0.1), xlab=xaxis, main="TE insertions by Po.TE2 (overalline and
beanlines are mean)")
# Add the x-axis label
xaxis=mtext("Species", side = 1, line = 7)
```

```
# Plotting barplots with TE insertions per species
```

```
# Ngs_te_mapper2
# Read file and add colnames
ngst2_geocol=read.table("ngste2_geocol_copy.txt")
colnames(ngst2_geocol)=c("counts", "te_order", "accession")
# Change the labels of x-axis
label=c("D. macrocarpa","D. outbatchensis","D. olen","D. fasciculosa","D.
ferrea","D. sandwicensis","D. vieillardii_1","D. vieillardii_2","D. flavocarpa","D.
umbrosa_1","D. umbrosa_2","D. veillonii","D. cherrieri_1","D. cherrieri_2","D.
labillardierei","D. trisulca","D. glans_1","D. glans_2","D. parviflora_1","D.
parviflora_2","D. revolutissima","D. pustulata","D. erudita","D. yahouensis","D.
minimifolia_1","D. minimifolia_2","D. tridentata","D. calciphila")
# Plot the graph
library(ggplot2)
ngste2_orders=ggplot(ngst2_geocol, aes(x=accession, y=counts))
geocol_ngste2=ngste2_orders + geom_col(aes(fill=te_order)) + theme(axis.text.x =
element_text(angle = 90, hjust = 1, size=18), axis.title = element_text(size=16),
legend.position = "bottom", legend.text = element_text(size=14), legend.title =
element_text(size = 14), plot.margin=unit(c(.1,.1,.1,.1), "cm")) +
scale_x_discrete(limits =
c("macrocarpaBT047","outbatchensisBT163","olenBT184","fasciculosaBT013","ferreaSD20
12","sandwiceG","vieillardiiBT286","vieillardiiBT024","flavocarpaBT128","umbrosaBT2
51","umbrosaBT255","veilloniiBT227","cherrieriBT293","cherrieriBT262","labillardier
eiBT123","trisulcaBT192","glansBT093","glansBT094","parvifloraBT060","parvifloraBT2
88","revolutissimaBT219","pustulataBT137","eruditaBT275","yahouensisBT238","minimif
oliaBT131","minimifoliaD4G","tridentataBT206","calciphilaBT313"), labels=label)
# Change the colour palette
library(viridis)
geocol_viridis_ngste2=geocol_ngste2 + scale_fill_viridis(discrete=TRUE, option =
"D") + labs(x="Species", y="TE superfamily richness", fill="TE superfamily")
#geo_col_scale=geo_col + scale_fill_manual(values = cm.colors(21))
show(geocol_viridis_ngste2)

# PopoolationTE2
# Read file and add colnames
te_popte2_geocol=read.table("te_popte2_geocol_copy.txt")
colnames(te_popte2_geocol)=c("counts", "te_order", "accession")
# Change the labels of the species
label=c("D. macrocarpa","D. outbatchensis","D. olen","D. fasciculosa","D.
ferrea","D. sandwicensis","D. vieillardii_1","D. vieillardii_2","D. flavocarpa","D.
umbrosa_1","D. umbrosa_2","D. veillonii","D. cherrieri_1","D. cherrieri_2","D.
labillardierei","D. trisulca","D. glans_1","D. glans_2","D. parviflora_1","D.
parviflora_2","D. revolutissima","D. pustulata","D. erudita","D. yahouensis","D.
minimifolia_1","D. minimifolia_2","D. tridentata","D. calciphila")
# Plot the graph
library(ggplot2)
te_orders=ggplot(te_popte2_geocol, aes(x=accession, y=counts))
geo_col=te_orders + geom_col(aes(fill=te_order)) + theme(axis.text.x =
element_text(angle = 90, hjust = 1, size=18), axis.title = element_text(size=16),
legend.position = "bottom", legend.text = element_text(size=14), legend.title =
element_text(size = 14), plot.margin=unit(c(.1,.1,.1,.1), "cm")) +
scale_x_discrete(limits =
c("macrocarpaBT047","outbatchensisBT163","olenBT184","fasciculosaBT013","ferreaSD20
12","sandwiceG","vieillardiiBT286","vieillardiiBT024","flavocarpaBT128","umbrosaBT2
51","umbrosaBT255","veilloniiBT227","cherrieriBT293","cherrieriBT262","labillardier
eiBT123","trisulcaBT192","glansBT093","glansBT094","parvifloraBT060","parvifloraBT2
88","revolutissimaBT219","pustulataBT137","eruditaBT275","yahouensisBT238","minimif
oliaBT131","minimifoliaD4G","tridentataBT206","calciphilaBT313"), labels=label) +
labs(x="Species", y="TE superfamily richness", fill="TE superfamily")
# Change the colour palette
library(viridis)
geo_col_viridis=geo_col + scale_fill_viridis(discrete=TRUE, option = "D")
show(geo_col_viridis)
```

Supplementary 2. Genes selected for single-copy gene normalisation in Deviate

N°	*Gene name*	*Contig name and position in the reference genome*
1	9at4069.fna	impolita_tig00006770:4197350-4224032
2	9993at4069.fna	impolita_tig00006668:148633-152252
3	9974at4069.fna	impolita_tig00007128:2507349-2509157
4	9942at4069.fna	impolita_tig00006272:5184850-5240922
5	9932at4069.fna	impolita_tig00006505:872675-874087

Supplementary 3. Samtools flagstat of mapping TEs to a consensus TE with BWA-MEM. All QC-Failed were 0.

Flag field	*D. cherrierii_1*	*D. fasciculosa*	*D. ferrea*	*D. labillardierei*	*D. macrocarpa*	*D. minimifolia_1*	*D. minimifolia_2*	*D. olen*	*D. parviflora_1*	*D. parviflora_2*	*D. sandwicensis*	*D. vieillardii_3*
	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed	Passed
total	256,604,719	192,196,270	166,004,179	296,157,518	160,039,349	305,339,740	136,966,539	245,569,657	247,378,135	268,931,027	276,322,123	198,663,518
primary	252,021,656	191,283,806	164,767,770	289,477,530	159,300,916	299,814,164	134,651,148	244,649,734	243,071,062	263,879,830	273,821,740	195,567,714
mapped	138,266,286	20,276,020	57,329,052	185,478,979	10,738,785	170,972,468	75,743,237	21,350,284	137,178,541	160,770,253	78,095,336	99,277,607
mapped %	53.88%	10.55%	34.53%	62.63%	6.71%	55.99%	55.30%	8.69%	55.45%	59.78%	28.26%	49.97%
paired in sequencing	252,021,656	191,283,806	164,767,770	289,477,530	159,300,916	299,814,164	134,651,148	244,649,734	243,071,062	263,879,830	273,821,740	195,567,714
properly paired	97,828,038	14,228,190	52,776,742	134,259,458	7,845,266	117,133,076	53,028,654	14,304,860	107,534,156	119,470,866	52,747,870	88,383,412
properly paired %	38.82%	7.44%	32.03%	46.38%	4.92%	39.07%	39.38%	5.85%	44.24%	45.27%	19.26%	45.19%
with itself and mate mapped	118,271,506	15,057,522	54,224,238	160,628,662	8,204,210	144,252,498	64,723,310	15,065,694	122,363,724	140,258,464	59,636,258	93,670,958
singletons	15,411,717	4,306,034	1,868,405	18,170,329	1,796,142	21,194,394	8,704,536	5,364,667	10,507,744	15,460,592	15,958,695	2,510,845
singletons %	6.12%	2.25%	1.13%	6.28%	1.13%	7.07%	6.46%	2.19%	4.32%	5.86%	5.83%	1.28%

Supplementary 4. Record of the highest ten mapping counts (normalised with CPM). MAX is the maximum value in the row.

TE ID	*TE NAMES*	*D. cherrierii_1*	*D. fasciculosa*	*D. ferrea*	*D. labillardierei*	*D. macrocarpa*	*D. minimifolia_1*	*D. minimifolia_2*
TE148	rnd-1_family-14#Unknown	30,875.39	5,280.59	2,331.35	239,339.53	7,518.20	99,409.88	86,136.78
TE1765	rnd-6_family-152#LTR/Copia	8,330.42	22,836.99	12,808.67	12,302.63	184,783.53	8,916.31	7,909.19
TE1810	rnd-6_family-5012#Unknown	18,720.55	64,466.01	15,290.31	10,986.57	93,262.47	15,120.98	11,945.58
TE554	rnd-1_family-100#LTR/Copia	2,719.54	4,541.97	3,027.75	3,925.55	57,597.37	2,873.39	2,493.12
TE1920	rnd-6_family-7227#rRNA	5,367.69	36,255.84	9,320.70	3,978.91	43,731.79	3,992.28	4,093.41
TE1687	rnd-6_family-1011#LTR/Gypsy	11,927.62	11,975.83	12,571.01	7,921.26	3,055.71	10,615.81	10,321.52
TE687	rnd-1_family-4#LTR/Gypsy	4,075.12	845.93	824.7	36,235.10	1,199.40	15,865.67	16,503.51
TE513	rnd-1_family-214#LTR/Copia	1,855.92	4,554.06	2,623.87	972.01	31,399.90	1,399.42	1,475.33
TE304	rnd-1_family-192#LTR/Copia	822.42	960.56	1,256.47	477.72	28,507.87	697.21	741.06
TE1666	rnd-6_family-3088#LTR/Gypsy	11,986.04	3,908.28	11,005.78	7,014.89	975.11	9,314.50	9,334.85

(continue)

TE ID	*TE NAMES*	*D. olen*	*D. parviflora_1*	*D. parviflora_2*	*D. sandwicensis*	*D. vieillardii_3*	*MAX*
TE148	rnd-1_family-14#Unknown	4,567.45	154,091.68	210,611.23	2,429.91	32,791.30	239,339.53
TE1765	rnd-6_family-152#LTR/Copia	24,401.37	9,795.61	11,730.96	444.62	15,482.68	184,783.53
TE1810	rnd-6_family-5012#Unknown	63,054.90	11,727.46	8,154.78	6,174.71	23,453.13	93,262.47
TE554	rnd-1_family-100#LTR/Copia	4,332.68	2,982.64	3,689.22	151.84	4,788.04	57,597.37
TE1920	rnd-6_family-7227#rRNA	49,821.87	4,451.51	3,149.54	20,315.81	2,442.38	49,821.87
TE1687	rnd-6_family-1011#LTR/Gypsy	5,748.06	9,747.19	8,193.73	40,053.35	11,725.48	40,053.35
TE687	rnd-1_family-4#LTR/Gypsy	772.92	17,235.31	26,924.35	1,281.19	8,658.20	36,235.10
TE513	rnd-1_family-214#LTR/Copia	2,855.24	1,239.51	1,119.64	435.06	1,057.15	31,399.90
TE304	rnd-1_family-192#LTR/Copia	1,062.44	659.04	574.88	111.81	751.76	28,507.87
TE1666	rnd-6_family-3088#LTR/Gypsy	1,325.22	8,770.73	7,757.03	27,645.31	10,817.98	27,645.31

Supplementary 5. Record of the highest ten mean mapping counts calculated for each TE over all species. MEAN was calculated by row.

TE ID	*TE NAMES*	*D. cherrierii_1*	*D. fasciculosa*	*D. ferrea*	*D. labillardierei*	*D. macrocarpa*	*D. minimifolia_1*	*D. minimifolia_2*
TE148	rnd-1_family-14#Unknown	30,875.39	5,280.59	2,331.35	239,339.53	7,518.20	99,409.88	86,136.78
TE1810	rnd-6_family-5012#Unknown	18,720.55	64,466.01	15,290.31	10,986.57	93,262.47	15,120.98	11,945.58
TE1765	rnd-6_family-152#LTR/Copia	8,330.42	22,836.99	12,808.67	12,302.63	184,783.53	8,916.31	7,909.19
TE1920	rnd-6_family-7227#rRNA	5,367.69	36,255.84	9,320.70	3,978.91	43,731.79	3,992.28	4,093.41
TE1687	rnd-6_family-1011#LTR/Gypsy	11,927.62	11,975.83	12,571.01	7,921.26	3,055.71	10,615.81	10,321.52
TE268	rnd-1_family-24#LTR/Gypsy	20,254.04	1,042.96	5,966.82	19,042.92	1,067.92	19,547.17	18,219.09
TE687	rnd-1_family-4#LTR/Gypsy	4,075.12	845.93	824.7	36,235.10	1,199.40	15,865.67	16,503.51
TE1666	rnd-6_family-3088#LTR/Gypsy	11,986.04	3,908.28	11,005.78	7,014.89	975.11	9,314.50	9,334.85
TE1354	rnd-4_family-47#LTR/Gypsy	14,029.82	2,543.30	10,674.95	8,013.01	1,000.68	10,619.36	10,774.39
TE2313	rnd-6_family-8663#LTR/Gypsy	14,520.95	8,228.64	8,853.19	8,402.13	824.69	11,610.93	12,153.22

(continue)

TE ID	*TE NAMES*	*D. olen*	*D. parviflora_1*	*D. parviflora_2*	*D. sandwicensis*	*D. vieillardii_3*	*MEAN*
TE148	rnd-1_family-14#Unknown	4,567.45	154,091.68	210,611.23	2,429.91	32,791.30	67,270.25
TE1810	rnd-6_family-5012#Unknown	63,054.90	11,727.46	8,154.78	6,174.71	23,453.13	32,827.50
TE1765	rnd-6_family-152#LTR/Copia	24,401.37	9,795.61	11,730.96	444.62	15,482.68	36,841.11
TE1920	rnd-6_family-7227#rRNA	49,821.87	4,451.51	3,149.54	20,315.81	2,442.38	15,248.66
TE1687	rnd-6_family-1011#LTR/Gypsy	5,748.06	9,747.19	8,193.73	40,053.35	11,725.48	9,769.82
TE268	rnd-1_family-24#LTR/Gypsy	1,712.74	16,110.64	17,261.10	1,985.64	10,066.90	12,162.99
TE687	rnd-1_family-4#LTR/Gypsy	772.92	17,235.31	26,924.35	1,281.19	8,658.20	10,792.78
TE1666	rnd-6_family-3088#LTR/Gypsy	1,325.22	8,770.73	7,757.03	27,645.31	10,817.98	7,648.49
TE1354	rnd-4_family-47#LTR/Gypsy	1,401.14	9,942.30	8,907.56	19,637.80	12,111.17	8,236.50
TE2313	rnd-6_family-8663#LTR/Gypsy	7,000.75	10,656.81	9,721.10	841.33	12,394.25	9,227.68

Supplementary 6. Calculation of the TE enrichment in the heatmaps. P-values calculated by chi-square test. NS: no significant, NA: no applies because it's absent. Over means overrepresented, under means underrepresented.

N°	*Consensus TE library*			*Clade 1*				*Clade 2*			
	Consensus TE families	N° TE families	Ratio	N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment
1	DNA/CMC-EnSpm	41	0.08	4	0.03	0.04	under	7	0.13	0.29	NS
2	DNA/hAT	1	0	NA	NA	NA	NA	NA	NA	NA	NA
3	DNA/hAT-Ac	16	0.03	5	0.03	1.00	NS	NA	NA	NA	NA
4	DNA/hAT-Tag1	7	0.01	4	0.03	0.27	NS	NA	NA	NA	NA
5	DNA/hAT-Tip100	7	0.01	2	0.01	1.00	NS	NA	NA	NA	NA
6	DNA/IS3EU	1	0	NA	NA	NA	NA	NA	NA	NA	NA
7	DNA/MULE-MuDR	18	0.03	6	0.04	0.82	NS	NA	NA	NA	NA
8	DNA/PIF-Harbinger	7	0.01	NA	NA	NA	NA	NA	NA	NA	NA
9	DNA/Zisupton	1	0	NA	NA	NA	NA	NA	NA	NA	NA
10	LINE/L1	12	0.02	NA	NA	NA	NA	2	0.04	0.64	NS
11	LINE/Penelope	2	0	NA	NA	NA	NA	NA	NA	NA	NA
12	LINE/RTE-BovB	10	0.02	2	0.01	0.74	NS	2	0.04	0.61	NS
13	LTR/Caulimovirus	12	0.02	NA	NA	NA	NA	NA	NA	NA	NA
14	LTR/Copia	167	0.31	27	0.18	0.004	under	24	0.43	0.06	NS
15	LTR/ERV1	3	0.01	1	0.01	1.00	NS	NA	NA	NA	NA
16	LTR/ERVK	3	0.01	1	0.01	1.00	NS	NA	NA	NA	NA
17	LTR/ERVL	1	0	NA	NA	NA	NA	NA	NA	NA	NA
18	LTR/Gypsy	215	0.4	90	0.61	0.0005	over	20	0.36	0.57	NS
19	PLE/Chlamys	3	0.01	NA	NA	NA	NA	1	0.02	0.32	NS
20	RC/Helitron	10	0.02	5	0.03	0.35	NS	NA	NA	NA	NA
21	Retroposon/L1-dep	2	0	NA	NA	NA	NA	NA	NA	NA	NA
	Total	539		147				56			

(Continue)

Clade 3				Clade 4				Clade 5			
N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment
5	0.06	0.68	NS	4	0.04	0.38	NS	21	0.13	0.030	over
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.40	NS
NA	NA	NA	NA	4	0.04	0.51	NS	7	0.04	0.44	NS
NA	NA	NA	NA	NA	NA	NA	NA	3	0.02	0.71	NS
2	0.02	0.61	NS	1	0.01	1.00	NS	2	0.01	1.00	NS
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.42	NS
2	0.02	0.74	NS	3	0.03	1.00	NS	7	0.04	0.61	NS
2	0.02	0.63	NS	1	0.01	1.00	NS	4	0.03	0.45	NS
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.40	NS
6	0.07	0.04	over	NA	NA	NA	NA	4	0.03	1.00	NS
NA	NA	NA	NA	NA	NA	NA	NA	2	0.01	0.21	NS
1	0.01	0.71	NS	4	0.04	0.13	NS	1	0.01	0.34	NS
3	0.03	0.71	NS	1	0.01	0.71	NS	8	0.05	0.09	NS
40	0.45	0.01	over	46	0.52	0.0005	over	30	0.19	0.003	under
NA	NA	NA	NA	NA	NA	NA	NA	2	0.01	0.60	NS
2	0.02	0.16	NS	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.41	NS
23	0.26	0.018	under	20	0.22	0.001	under	62	0.39	0.86	NS
1	0.01	1.00	NS	1	0.01	1.00	NS	NA	NA	NA	NA
1	0.01	0.73	NS	4	0.04	0.12	NS	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	2	0.01	0.23	NS
88				89				159			

(Continue)

Clade 3a				Clade 3b				Clade 3c			
N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment
1	0.03	0.50	NS	2	0.09	1.00	NS	2	0.06	0.77	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.03	0.37	NS	NA	NA	NA	NA	1	0.03	0.41	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	1	0.04	1.00	NS	1	0.03	1.00	NS
NA	NA	NA	NA	NA	NA	NA	NA	2	0.06	0.09	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	2	0.09	0.11	NS	4	0.12	0.011	over
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.03	1.00	NS	NA	NA	NA	NA	NA	NA	NA	NA
1	0.03	1.00	NS	1	0.04	1.00	NS	1	0.03	1.00	NS
18	0.58	0.006	over	13	0.57	0.009	over	9	0.26	0.70	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.03	0.19	NS	NA	NA	NA	NA	1	0.03	0.22	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7	0.23	0.06	NS	4	0.17	0.06	NS	12	0.35	0.74	NS
1	0.03	0.22	NS	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	1	0.03	1.00	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
31				23				34			

(Continue)

Clade a				Clade b				Clade c			
N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment	N° TE families	Ratio	p-value	Enrichment
1	0.01	0.010	under	15	0.08	1.00	NS	7	0.09	0.83	NS
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3	0.03	1.00	NS	1	0.01	0.05	under	4	0.05	0.49	NS
NA	NA	NA	NA	NA	NA	NA	NA	4	0.05	0.037	over
NA	NA	NA	NA	1	0.01	0.46	NS	2	0.03	0.63	NS
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.24	NS
5	0.04	0.78	NS	2	0.01	0.13	NS	6	0.08	0.10	NS
NA	NA	NA	NA	3	0.02	1.00	NS	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
1	0.01	0.47	NS	8	0.04	0.20	NS	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	7	0.04	0.28	NS	NA	NA	NA	NA
NA	NA	NA	NA	4	0.02	1.00	NS	3	0.04	0.41	NS
19	0.16	0.002	under	92	0.46	0.0005	over	12	0.16	0.005	under
1	0.01	1.00	NS	NA	NA	NA	NA	1	0.01	1.00	NS
NA	NA	NA	NA	2	0.01	0.62	NS	1	0.01	1.00	NS
NA	NA	NA	NA	NA	NA	NA	NA	1	0.01	0.23	NS
84	0.72	0.0005	over	58	0.29	0.009	under	30	0.39	0.89	NS
NA	NA	NA	NA	3	0.02	0.35	NS	NA	NA	NA	NA
NA	NA	NA	NA	3	0.02	0.77	NS	5	0.06	0.024	over
2	0.02	0.15	NS	NA	NA	NA	NA	NA	NA	NA	NA
116				199				77			

(Continue)

Clade d			
N° TE families	Ratio	p-value	Enrichment
18	0.12	0.10	NS
1	0.01	0.37	NS
8	0.05	0.21	NS
3	0.02	0.70	NS
4	0.03	0.26	NS
NA	NA	NA	NA
5	0.03	1.00	NS
4	0.03	0.25	NS
1	0.01	0.38	NS
3	0.02	1.00	NS
2	0.01	0.19	NS
3	0.02	1.00	NS
5	0.03	0.56	NS
44	0.3	0.83	NS
1	0.01	1.00	NS
NA	NA	NA	NA
NA	NA	NA	NA
43	0.29	0.024	under
NA	NA	NA	NA
2	0.01	0.76	NS
NA	NA	NA	NA
147			

Supplementary 7. TE insertions stats calculated for PopoolationTE2. Min=minimum TE counts in a TE family. Max=maximum TE counts in a TE family. Mean=mean of TE counts. Median=median of TE counts. SD=standard deviation of the TE counts.

N°	*Species*	*TE insertion counts*	*Min*	*Max*	*Mean*	*Median*	*SD*
1	*D. macrocarpa*	21375	1	2088	25.94	4	97.25
2	*D. outbatchensis*	29953	1	2941	35.41	5	136.87
3	*D. olen*	56073	1	3832	44.54	10	153.51
4	*D. fasciculosa*	33305	1	2451	27.94	7	97.19
5	*D. ferrea*	70524	1	2192	41.46	14	95.42
6	*D. sandwicensis*	319844	1	5499	177.99	61	381.08
7	*D. vieillardii_2*	356090	1	5112	175.15	66	347.93
8	*D. vieillardii_1*	231118	1	4359	118.16	43	239.33
9	*D. flavocarpa*	498177	1	5633	239.16	90	481.2
10	*D. umbrosa_1*	423248	1	5286	206.87	78	411.71
11	*D. umbrosa_2*	335206	1	3749	163.59	62	325.9
12	*D. veillonii*	468827	1	5451	226.27	85	444.84
13	*D. cherrierii_2*	424759	1	5477	208.01	79	404.38
14	*D. cherrierii_1*	461390	1	5501	223.54	85	439.74
15	*D. labillardierei*	438407	1	5177	213.44	83	408.67
16	*D. trisulca*	523620	1	9442	249.34	96	521.59
17	*D. glans_1*	439507	1	5269	211.4	82	404.7
18	*D. glans_2*	527276	1	5607	250.25	99	479.72
19	*D. parviflora_1*	356696	1	4753	174.42	69	337.17
20	*D. parviflora_2*	415368	1	5027	201.34	80	383.49
21	*D. revolutissima*	539823	1	6399	256.94	100	512.58
22	*D. pustulata*	548920	1	5660	261.27	108	496.48
23	*D. erudita*	487485	1	5390	234.59	95	444.69
24	*D. yahouensis*	448904	1	5218	217.81	87	413.65
25	*D. minimifolia_1*	535639	1	5582	255.43	103	489.5
26	*D. minimifolia_2*	353854	1	4405	173.37	68	337.14
27	*D. tridentata*	391791	1	4919	190.28	74	366.9
28	*D. calciphila*	457441	1	5203	219.5	87	423.26

Supplementary 8. TE insertions stats calculated for Ngs_te_mapper2. Min=minimum TE counts in a TE family. Max=maximum TE counts in a TE family. Mean=mean of TE counts. Median=median of TE counts. SD=standard deviation of the TE counts.

N°	*Species*	*TE insertion counts*	*Min*	*Max*	*Mean*	*Median*	*SD*
1	*D. macrocarpa*	326	1	34	2.51	1	3.85
2	*D. outbatchensis*	427	1	60	2.9	1	5.96
3	*D. olen*	350	1	90	3.12	1	7.37
4	*D. fasciculosa*	674	1	36	2.22	1	3.64
5	*D. ferrea*	2053	1	50	2.83	2	3.85
6	*D. sandwicensis*	4609	1	407	5.38	2	15.73
7	*D. vieillardii_2*	7692	1	142	6.11	3	10.89
8	*D. vieillardii_1*	7514	1	120	5.82	3	9.91
9	*D. flavocarpa*	8536	1	145	6.65	3	11.67
10	*D. umbrosa_1*	8188	1	135	6.42	3	11.09
11	*D. umbrosa_2*	6069	1	144	5.39	3	10.22
12	*D. veillonii*	8681	1	154	6.82	3	11.53
13	*D. cherrierii_2*	10301	1	148	7.69	4	13.12
14	*D. cherrierii_1*	9066	1	142	7.13	3	11.84
15	*D. labillardierei*	9698	1	170	7.56	3	14.48
16	*D. trisulca*	10214	1	141	7.56	3	13.43
17	*D. glans_1*	10143	1	138	7.64	3	13.94
18	*D. glans_2*	11864	1	156	8.51	4	15.45
19	*D. parviflora_1*	8753	1	140	7	3	13.04
20	*D. parviflora_2*	9425	1	219	7.43	3	14.16
21	*D. revolutissima*	11267	1	159	8.48	4	15.67
22	*D. pustulata*	11871	1	169	8.76	4	15.72
23	*D. erudita*	11038	1	218	8.38	4	16.33
24	*D. yahouensis*	10099	1	268	7.7	3	15.03
25	*D. minimifolia_1*	10588	1	186	8.17	4	14.71
26	*D. minimifolia_2*	5740	1	142	5.3	3	9.22
27	*D. tridentata*	8304	1	167	6.84	3	12.59
28	*D. calciphila*	10064	1	167	7.7	3	14.23

Supplementary 9. TE superfamilies abundances for every species by PopoolationTE2.

Species	*DNA_CMC_EnSpm*	*DNA_IS3EU*	*DNA_MULE_MuDR*	*DNA_PIF_Harbinger*	*DNA_Zisupton*	*DNA_hAT*	*DNA_hAT_Ac*	*DNA_hAT_Tag1*	*DNA_hAT_Tip100*	*LINE_L1*
D. calciphila	37	1	18	7	1	1	16	7	7	12
D. cherrierii_1	37	1	18	7	1	1	16	7	7	12
D. cherrierii_2	37	1	18	7	1	1	16	7	7	12
D. erudita	37	1	18	7	1	1	16	7	7	12
D. fasciculosa	23	0	12	6	1	1	14	7	7	12
D. ferrea	29	1	18	6	1	1	16	7	7	12
D. flavocarpa	36	1	18	7	1	1	16	7	7	12
D. glans_1	37	1	18	7	1	1	16	7	7	12
D. glans_2	36	1	18	7	1	1	16	7	7	12
D. labillardierei	37	1	18	7	1	1	16	7	7	12
D. macrocarpa	15	0	12	5	1	1	10	7	7	7
D. minimifolia_1	37	1	18	7	1	1	16	7	7	12
D. minimifolia_2	36	1	18	7	1	1	16	7	7	12
D. olen	24	0	13	4	1	1	15	7	7	8
D. outbatchensis	15	0	9	4	1	1	10	7	6	7
D. parviflora_1	36	1	18	7	1	1	16	7	7	12
D. parviflora_2	36	1	18	7	1	1	16	7	7	12
D. pustulata	37	1	18	7	1	1	16	7	7	12
D. revolutissima	37	1	18	7	1	1	16	7	7	12
D. sandwicensis	36	1	16	7	1	1	14	7	7	12
D. tridentata	36	1	18	7	1	1	16	7	7	12
D. trisulca	36	1	18	7	1	1	16	7	7	12
D. umbrosa_1	35	1	18	7	1	1	16	7	7	12
D. umbrosa_2	36	1	18	7	1	1	16	7	7	12
D. veillonii	36	1	18	7	1	1	16	7	7	12
D. vieillardii_1	37	1	18	7	1	1	16	7	7	12
D. vieillardii_2	37	1	18	7	1	1	16	7	7	12
D. yahouensis	36	1	18	7	1	1	16	7	7	12

(continue)

Species	LINE_Penelope	*LINE_RTE*	*LTR_Caulimovirus*	*LTR_Copia*	*LTR_ERV1*	*LTR_ERVK*	*LTR_ERVL*	*LTR_Gypsy*	*PLE_Chlamys*	*RC_Helitron*	*Retroposon_L1_dep*
D. calciphila	2	8	12	158	3	3	1	208	3	10	2
D. cherrierii_1	2	8	12	160	3	3	1	204	3	10	2
D. cherrierii_2	2	8	12	160	3	3	1	202	3	10	2
D. erudita	2	8	12	161	3	3	1	203	3	10	2
D. fasciculosa	2	8	5	132	2	2	0	139	2	10	1
D. ferrea	2	8	12	152	3	3	1	178	3	10	2
D. flavocarpa	2	8	12	160	3	3	1	202	3	10	2
D. glans_1	2	8	12	160	3	3	1	205	3	10	2
D. glans_2	2	8	12	158	3	3	1	205	3	10	2
D. labillardierei	2	8	12	158	3	3	1	202	3	10	2
D. macrocarpa	2	8	4	110	1	1	0	85	2	7	2
D. minimifolia_1	2	8	12	160	3	3	1	205	3	10	2
D. minimifolia_2	2	8	12	158	3	3	1	200	3	10	2
D. olen	2	8	7	146	2	2	0	153	2	10	1
D. outbatchensis	2	8	5	119	2	2	0	92	2	6	1
D. parviflora_1	2	8	12	158	3	3	1	202	3	10	2
D. parviflora_2	2	8	12	160	3	3	1	201	3	10	2
D. pustulata	2	8	12	159	3	3	1	205	3	10	2
D. revolutissima	2	8	12	161	3	3	1	208	3	10	2
D. sandwicensis	2	8	12	156	3	3	1	192	3	10	2
D. tridentata	2	8	12	159	3	3	1	200	3	10	2
D. trisulca	2	8	12	160	3	3	1	205	3	10	2
D. umbrosa_1	2	8	12	160	3	3	1	203	3	10	2
D. umbrosa_2	2	8	12	157	3	3	1	199	3	10	2
D. veillonii	2	8	12	160	3	3	1	203	3	10	2
D. vieillardii_1	2	8	12	159	3	3	1	197	3	10	2
D. vieillardii_2	2	8	12	161	3	3	1	198	3	10	2
D. yahouensis	2	8	12	159	3	3	1	201	3	10	2

Supplementary 10. TE superfamilies abundances for every species by Ngs_te_mapper2.

Species	DNA_CMC_EnSpm	*DNA_IS3EU*	*DNA_MULE_MuDR*	*DNA_PIF_Harbinger*	*DNA_Zisupton*	*DNA_hAT*	*DNA_hAT_Ac*	*DNA_hAT_Tag1*	*DNA_hAT_Tip100*	*LINE_L1*
D. calciphila	21	0	18	6	1	1	12	10	8	7
D. cherrierii_1	23	1	15	5	1	1	10	12	7	5
D. cherrierii_2	24	1	16	5	1	1	12	12	8	5
D. erudita	21	0	19	6	1	1	11	11	6	5
D. fasciculosa	4	0	2	1	0	0	1	2	1	1
D. ferrea	15	0	15	4	1	1	9	7	5	2
D. flavocarpa	19	0	18	5	1	2	11	11	7	6
D. glans_1	20	0	17	6	1	1	11	11	7	5
D. glans_2	23	1	16	7	1	1	13	11	7	7
D. labillardierei	23	0	17	7	1	1	12	10	8	6
D. macrocarpa	2	0	2	1	0	0	2	0	0	0
D. minimifolia_1	22	0	20	8	1	1	11	11	7	6
D. minimifolia_2	18	0	14	5	2	0	9	11	7	4
D. olen	6	0	3	0	0	0	2	1	3	0
D. outbatchensis	2	0	4	0	0	0	0	0	0	0
D. parviflora_1	21	0	17	7	1	1	12	10	7	5
D. parviflora_2	19	0	18	6	1	1	12	9	5	4
D. pustulata	23	0	17	6	1	1	11	12	6	7
D. revolutissima	21	0	17	6	1	1	12	10	7	6
D. sandwicensis	10	0	11	4	1	1	5	7	7	3
D. tridentata	20	1	16	4	1	1	12	10	7	7
D. trisulca	21	1	17	6	1	1	12	10	6	8
D. umbrosa_1	19	0	16	6	1	1	11	10	8	7
D. umbrosa_2	19	2	18	5	1	0	11	10	6	7
D. veillonii	20	0	17	5	2	1	11	10	7	8
D. vieillardii_1	22	1	17	6	1	1	11	11	6	5
D. vieillardii_2	23	0	18	4	1	2	12	11	8	7
D. yahouensis	23	0	16	7	1	1	11	10	7	5

(continue)

Species	LINE_Penelope	*LINE_RTE*	*LTR_Caulimovirus*	*LTR_Copia*	*LTR_ERV1*	*LTR_ERVK*	*LTR_ERVL*	*LTR_Gypsy*	*PLE_Chlamys*	*RC_Helitron*	*Retroposon_L1_dep*
D. calciphila	1	8	11	101	4	3	1	133	1	13	3
D. cherrierii_1	2	9	10	97	3	3	0	129	2	13	4
D. cherrierii_2	2	9	9	109	3	4	0	139	2	11	4
D. erudita	2	9	10	104	3	3	0	143	2	10	4
D. fasciculosa	0	4	0	11	0	0	0	13	0	4	0
D. ferrea	2	5	2	56	3	2	0	56	0	8	1
D. flavocarpa	2	9	8	101	4	3	0	141	1	12	4
D. glans_1	2	8	11	106	4	3	0	137	3	12	4
D. glans_2	2	8	10	111	3	3	0	144	3	13	4
D. labillardierei	1	9	9	101	3	2	1	135	1	12	4
D. macrocarpa	0	4	0	4	0	0	0	11	0	4	0
D. minimifolia_1	2	7	10	102	3	3	0	141	2	14	4
D. minimifolia_2	1	5	9	86	3	3	0	121	1	12	3
D. olen	0	6	0	21	0	0	0	16	1	5	0
D. outbatchensis	0	6	0	7	1	0	0	11	1	5	0
D. parviflora_1	2	8	10	100	3	4	0	129	2	11	3
D. parviflora_2	2	7	11	96	4	3	0	138	2	10	3
D. pustulata	2	7	9	110	4	3	0	144	2	14	4
D. revolutissima	2	9	10	103	3	3	0	135	2	13	4
D. sandwicensis	1	6	7	77	3	3	0	60	1	11	2
D. tridentata	2	7	10	98	3	3	0	137	1	10	4
D. trisulca	2	8	12	108	3	4	1	141	2	13	4
D. umbrosa_1	3	8	7	107	4	3	1	130	3	10	4
D. umbrosa_2	3	6	7	91	4	3	1	122	2	10	4
D. veillonii	2	6	12	106	3	4	0	136	1	13	4
D. vieillardii_1	2	8	8	110	3	3	0	136	2	11	4
D. vieillardii_2	3	9	8	102	4	3	0	124	3	12	4
D. yahouensis	2	8	11	105	3	3	0	140	1	13	4

Supplementary 11. Legend of TE family names in the consensus TE by RepeatModeller (see Figure 11)

TE ID	*Consensus TE name*
TE_unknown1	rnd_1_family_1_Unknown
TE_unknown2	rnd_1_family_14_Unknown
TE_unknown3	rnd_1_family_263_Unknown
TE_unknown4	rnd_1_family_749_Unknown
TE_unknown5	rnd_1_family_898_Unknown
TE_unknown6	rnd_1_family_986_Unknown
TE_unknown7	rnd_3_family_201_Unknown
TE_unknown8	rnd_3_family_251_Unknown
TE_unknown9	rnd_3_family_59_Unknown
TE_unknown10	rnd_4_family_256_Unknown
TE_unknown11	rnd_4_family_340_Unknown
TE_unknown12	rnd_4_family_383_Unknown
TE_unknown13	rnd_5_family_181_Unknown
TE_unknown14	rnd_5_family_916_Unknown
TE_unknown15	rnd_5_family_989_Unknown
TE_unknown16	rnd_6_family_1113_Unknown
TE_unknown17	rnd_6_family_1492_Unknown
LTR_Gypsy1	rnd_1_family_20_LTR_Gypsy
LTR_Gypsy2	rnd_1_family_54_LTR_Gypsy
LTR_Gypsy3	rnd_4_family_195_LTR_Gypsy
LTR_Gypsy4	rnd_4_family_342_LTR_Gypsy
LTR_Gypsy5	rnd_4_family_797_LTR_Gypsy
LTR_Gypsy6	rnd_5_family_1298_LTR_Gypsy
LTR_Gypsy7	rnd_5_family_1744_LTR_Gypsy
LTR_Gypsy8	rnd_5_family_918_LTR_Gypsy
LTR_Gypsy9	rnd_6_family_3088_LTR_Gypsy
LTR_Gypsy10	rnd_6_family_8663_LTR_Gypsy
LTR_Copia1	rnd_1_family_334_LTR_Copia
LTR_Copia2	rnd_5_family_4562_LTR_Copia
LINE_RTE1	rnd_3_family_363_LINE_RTE_BovB
LINE_RTE2	rnd_4_family_520_LINE_RTE_BovB
LINE_RTE3	rnd_4_family_694_LINE_RTE_BovB
LINE_RTE4	rnd_5_family_1219_LINE_RTE_BovB
LINE_RTE5	rnd_5_family_333_LINE_RTE_BovB
LINE_RTE6	rnd_6_family_26_LINE_RTE_BovB