Waschburger *et al. Mobile DNA* (2025) 16:48
https://doi.org/10.1186/s13100-025-00388-3

Mobile DNA

RESEARCH Open Access

Transposable elements: a key piece in the genomic evolution and adaptation of Myrtaceae species

Edgar Luis Waschburger[1†], João Pedro Carmo Filgueiras[1†], Henrique da Rocha Moreira Antoniolli[1], Maríndia Deprá[1], Romain Guyot[2†] and Andreia Carina Turchetto-Zolet[1*†]

Abstract

Background Myrtaceae is a family of woody trees with over 5,800 species, representing the sixth most diverse plant family. It includes many economically important members distributed throughout East Asia, Oceania, and the Americas, including but not limited to *Eucalyptus grandis* and *Syzygium aromaticum*. Most available Myrtaceae genome assemblies are arranged in 11 chromosomes, and possess large variability in genome sizes, sometimes over triple the size. Although coding sequences add to this disparity, transposable elements (TEs) are the main contributors to genome size variation.

Results In our research, we have characterized the landscape of TEs in 18 species of Myrtaceae. Our results showed that LTR Class I elements are the main contributors to genome size variations in Myrtaceae. Furthermore, specific lineages among the Gypsy and Copia superfamilies are linked to historical events of transposon activity amongst Myrtaceae tribes. Extracted climatic and distribution data were in correlation with TE profiles, indicating possible lineages more sensitive to climatic conditions. A gene ontology over-representation analysis revealed shared biological processes influenced by TEs, and exclusive ones linked to different environmental responses. Lastly, we identified high-identity sequences among many species, and performed phylogenies for horizontal transposable element transfer (HTT) events analysis. A positive HTT of a Copia/Ivana TE among Syzygieae and Myrteae tribes could affect the regulation of proximal microorganism defense response genes.

Conclusions Our findings suggest that TEs may influence the genetic diversity present in Myrtaceae, where TE lineages contribute asymmetrically to their genomic profiles. More importantly, specific lineages are correlated with climatic variables possibly by their influence on proximal genes, a balance between genetic variation and fitness influence. Lastly, the impact of TEs on microorganism defense response genes appears to be a key element in the adaptation process of Myrtaceae species.

†Edgar Luis Waschburger, João Pedro Carmo Filgueiras, Romain Guyot and Andreia Carina Turchetto-Zolet contributed equally to this work.

*Correspondence:
Andreia Carina Turchetto-Zolet
carina.turchetto@ufrgs.br

Full list of author information is available at the end of the article

Keywords Transposable elements, Myrtaceae, Genomics, Evolution, Adaptation, Microorganism, Defense response, Climate

Background

Myrtaceae is a family of woody trees with 17 tribes and more than 5,800 species, representing the sixth most diverse plant family [1]. It is suggested that the earliest common ancestor of the family may have originated 90 million years ago (Mya) in Gondwana, before the split of the southern continents [2]. Naturally, it is proposed that speciation events followed the shifting of the tectonic plates, becoming the foundation of the huge diversity within this family. Currently, members are dispersed along the tropical and subtropical regions in all continents, especially in the southern hemisphere, where subsequent diversification events are postulated to have taken place (Fig. 1). For example, the *Syzygium* genus - the most varied woody tree genus - reached continental Asia with a history of migration and speciation events starting from Australia and New Guinea [3]. The *Eugenia* genus, on the other hand, ascended South America and acclimated to the humid tropical forests where it thrived and rapidly expanded. Currently, it is the most prevalent genus in the Amazon Rain Forest and the most diverse tree genus in the Atlantic Forest. It also has been dispersed to the Caribbean and Pacific regions, overseas to Asia [4], and every continent due to its high adaptability.

In Myrtaceae, a plethora of molecular markers have been used to explore their diversity, from phylogenetic relationships to population genetics, in both the Neotropics and Oceania [5, 6]. Efforts have also been made to assemble the chloroplastid and nuclear genomes of different members of this family, with a primary focus on the *Eucalyptus* genus due to its high economical interest, being the first Myrtaceae genome published [7]. A total of 78 genomes are publicly available in NCBI Genomes as of April 2025, 45 being from *Eucalyptus* species, a genus native to Australia. However, this represents only 1% of the diversity of the family, exemplifying the potential for future studies in Myrtaceae. In general, the available genomic sequences of Myrtaceae species are arranged in 11 chromosomes with high levels of synteny for coding sequences compared to the genome sequence of *Eucalyptus grandis* [8, 9]. Yet, this similarity in genetic arrangement contrasts with their variability in genome size and non-coding regions, ranging from around 270 Mbp in *Melaleuca quinquenervia* to 930 Mbp in *Rhodomyrtus psidioides*, over triple the size. Although gene tandem repeats and gene family expansions add to this disparity, the vast majority are likely to be transposable elements, since they are the main contributors to genome size variation [10].

Transposable elements (TEs) are DNA sequences able to transverse the genome through self excision (or self-copying) and insertion. They are divided into two main classes, based on the molecule used for the transposition process: Class I includes TEs with an intermediary RNA molecule for transposition, which allows the reverse transcription of multiple self-copies, while Class II TEs are excised from their residing locus in the genome and transposed to another locus [11, 12]. The former is notably prevalent in plants, making up to 85% of genome sequence composition in *Zea mays* [13], especially those of the Long Terminal Repeat (LTR) family. TEs insertions are not always random or dispersed. Instead, there are some loci insertion tendencies that depend on the family of TEs, host tissue, or even the TE life cycle. This is likely due to both TE-specific adaptations and a heterogeneous purifying force acting in the genome [14]. Furthermore, they provide a tool for generating genetic diversity, if the host is able to balance deleterious mutations and diversifying forces.

TE genome activity results in genome size increase and may cause nucleotypic effects, which influence biological characteristics such as life cycle, minimum generation time, and growth rates [15]. TEs have an extensive repertoire of case studies with gene neofunctionalization, environmental adaptation, and have even been reported to be key players in the coevolution of plants and microorganisms [16, 17]. A TE characterization study in Rosaceae identified variety-specific expression of TEs that may influence gene expression between *Malus domestica* cvs. 'Gala' and 'GDDH13' [18]. A study described that the pericentromeric TE-rich regions of *Cucumis melo* harbor many melon-specific genes, while distal chromosome arm regions aggregate genes shared with closely related taxa, such as *Cucumis sativus*, thus acting as prone regions to the formation of novel genes [19]. Another study on maize identified a TE responsible for repressing the expression of a gene related to light signaling, allowing the adaptation of the crop to the longer days of higher latitudes [20]. A recent study on *Zymoseptoria tritici*, a wheat pathogen, identified multiple copy number variations derived from TE activity in its genome [21]. The authors postulate that the genetic variation introduced by these elements served as the basis for their adaptation process across different global temperature ranges. Similarly, variations in palm species genome size appear to be related to environmental aridity, where specific TE lineages are differentially abundant depending on the plant's niche [22]. Suffice it to say, TEs are one of the main

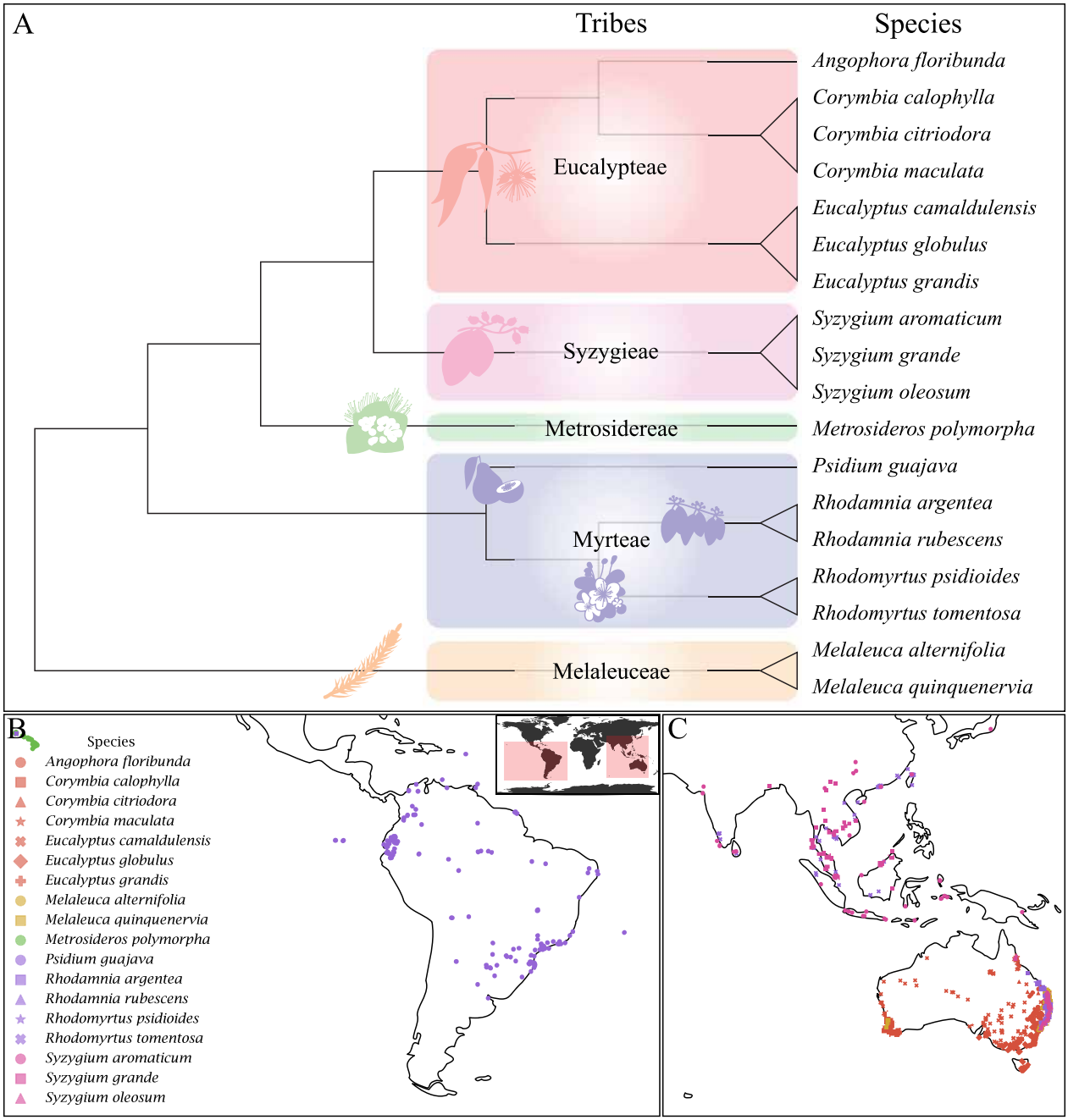

Fig. 1 Phylogenetic relationships of the 18 Myrtaceae species present in this study (**A**). Proximate Myrtaceae distribution along their native regions (**B** and **C**)

sources of variability and directly influence diversity and adaptation.

In this study, our objective was to identify and characterize TE in Myrtaceae, while analyzing the classes that contribute most to genome size variation. We have also classified them into lineages and investigated their activity profile to explain the variable contributions among 18 Myrtaceae. Furthermore, we constructed a robust approach for horizontal TE transfer (HTT) analysis and examined the TE landscape for high identity sequences. Lastly, by comparing publicly available climatic data with TE profiles, some LTR lineage expansions appear to be correlated with the ecological distributions of Myrtaceae. By investigating TE proximal regions, we have identified candidate gene families that may be under direct influence of TE activity and other diversifying forces.

Methods

Genome quality and annotation

A total of 22 species from the Myrtaceae family had their genomic sequences downloaded from NCBI's GenBank (Supplementary Material S1). Only up to three individual species from each genus were considered for further analyses to prevent over-representation. Genomes passed through a BUSCO [23] filtering step, where sequences with completeness lower than 90% were excluded. The resulting genomes had their metrics measured by QUAST [24] and annotated for transposable elements using a *de novo* approach with the EDTA (v2.0.0) pipeline [25]. The resulting libraries were reannotated using TEsorter [26] to identify the TE lineages and reduce unknown sequence numbers. Genome sequences were softmasked using the predicted TE libraries with RepeatMasker [27] and, together with protein evidence from *Eucalyptus grandis* and *Arabidopsis thaliana* (downloaded from Phytozomev13, phytozome-next.jgi.doe.gov), were used as input for gene prediction using Braker3 [28]. Linear models of TE composition and genome assembly metrics were performed by normalizing phylogenetic relationships using the phylogenetic generalized least squares (PGLS) method available in the R package caper [28].

Horizontal TE transfer identification

Identified BUSCO genes for each species were pairwise compared using the BLASTn algorithm [29]. Similar to Aubin et al. 2023 [30], the single top hit for each BUSCO gene was kept to create species-pairwise frequency distribution plots, which were used to calculate a threshold value for statistically deviating sequence identities ($\alpha \leq 0.05$ & $\sigma \geq 2.00$). Additionally, TE libraries were pairwise compared using BLASTn, and hits with smaller sequence coverage greater than 80% and sequence identity over the calculated threshold value for species-specific pairwise comparison were kept for further horizontal transfer analysis. Finally, to validate candidate sequences as horizontal transfers, each candidate was used as a query to retrieve sequences from a merged database of all Myrtaceae TE sequences identified previously. After filtering for duplicates, sequences were used as input for phylogenetic analysis and branch selection.

Phylogenetic and evolutionary selection analysis

Possible HTT sequences were aligned using MAFFT [31]. Aligned regions corresponding to conserved domains, as identified by NCBI's CDD (www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi), were used as input for model selection and maximum-likelihood phylogenetic reconstruction under IQTREE2 software [32]. For the species tree provided to the PGLS model, 100 BUSCO genes shared by all species analyzed were randomly selected, and their protein sequences were aligned using MAFFT. Amino acid sequences were translated back to DNA, and gaps were replaced by missing data. The resulting alignments were concatenated and used as input for phylogenetic reconstruction in IQTREE2. Specified parameters included the gene partition file and flags -m MF+MERGE and -rcluster 10. Branch support was done using 1000 replicates of UFBootstrap [33], and the resulting trees were visualized using the FigTree software (tree.bio.ed.ac.uk/). To further support the HTT hypothesis, an analysis of branch selection was done using the phylogenetic tree and in-frame sequence alignment on the EasyCodeML software [34]. In the case of a branch analysis with positive selection, an HTT hypothesis is more parsimonious, for signals of purifying selection may indicate a domestication by the host's genome of the TE sequence. In other words, the TE gained functional relevancy within the host's genome, and is now being conserved along its lineage, explaining the phylogenetic discrepancy.

Estimation of TE divergence times

TE lineages annotated by TEsorter were retrieved and aligned using MAFFT [31]. Distance matrices were calculated for every LTR lineage using the Kimura-2 substitution matrix [35] available in the EMBOSS package [36]. Each distance value (K) was used to calculate the expected insertion time (T) using the following formula: T = K/2r, where r is a constant mutation rate of 1.5x10-8. This same mutation rate was previously employed for the estimation of TE insertion times in other Myrtaceae species [9, 37]. Not all species are included in insertion time plots for some lineages, since we were unable to retrieve their 3'/5' LTR regions. This also resulted in a decrease in the expected number of comparisons for species with high amounts of some LTR lineages.

Climatic variables and GO over-representation

Climatic variables (list available in Supplementary Material 6) were downloaded from WorldClim [38] (worldclim.org), and species occurrence data were retrieved from GBIF [39] (gbif.org) using the Python package pygbif. Native distributions for each species were checked on the Royal Botanical Garden WCVP database [40] (powo.science.kew.org), and only points from close regions were processed. Principal Component Analysis (PCA) clustering and bi-plot were done in python by standardizing each variable to have a mean and unit variance equal to zero. Gene ontology over-representation analysis was conducted by extracting genes present in a 2 kbp region upstream or downstream of annotated TEs. The aminoacid sequences encoded by these genes were used as query for a BLASTp [29] search against the *E. grandis* ENSEMBL proteome (e-value = 1e-20), and only the top hit for each gene was considered. Different species hits

were merged if they were from the same LTR lineage to increase statistical power. The resulting list of genes served as input to the GO over-representation analysis in the Panther database (www.pantherdb.org) using default parameters and *p*-value equal or less than 0.05 [41].

Results

Myrtaceae class I TEs greatly influence genome size variation

Out of 22 Myrtaceae species retrieved from the NCBI genome database, 19 had BUSCO completeness over 90% and were used for TE annotation (with the removal of *Campomaneisa xanthocarpa*, *Eugenia klotzchiana*, and *Psidium friedrichsthalianum*). Many reached over 95% completeness, an indication of high-quality assemblies. Assembly sizes varied considerably, from 269 Mbp (*Melaleuca quinquenervia*) to 931 Mbp (*Rhodomyrtus psidioides*), as did their N50 values, 28 Kbp (*Eugenia uniflora*) to 58 Mbp (*Eucalyptus grandis*). Unfortunately, there is a lack of flow cytometry studies analyzing DNA content and estimated genome size for most species in our study, as such we cannot affirm the overall completeness of retrieved assemblies.

The genome assemblies passed through the EDTA pipeline, and the resulting libraries were re-annotated by the TEsorter software in order to reduce the number of unknown sequences and annotate LTR lineages. *E. uniflora* was excluded due to its high fragmentation, so the subsequent analyses were carried out with 18 species (Table 1).

Among the Myrtaceae species analyzed, the percentage values of genomic repeats varied according to assembly size, with the assembly of *Melaleuca quinquenervia* (269 Mbp) harboring 25.42% of repeats and *Rhodomyrtus psidioides* (931 Mbp) with 93.59%, over 870 Mbp of repeated sequences (Fig. 2A). The other species ranged mainly from 35–60% of genomic repeats, with a mean of 47.96% for the entire dataset. Overall, a total of 17 different TE types were found, plus regions related to pararetrovirus sequences. The majority of TEs, which belonged to the Class I LTRs, such as LTR/Gypsy and LTR/Copia, ranging from as little as 3.94% and 6.77% in the assembly of *Melaleuca quinquenervia* to 32.05% and 24.81% in the assembly of *Rhodomyrtus psidioides*, for Gypsy and Copia super-families respectively. The other fractions of the genome are mainly composed of DNA/Helitron and LINEs. Five species (*Corymbia citriodora*, *Eucalyptus grandis*, *Metrosideros polymorpha*, *Rhodamnia argentea* and *Syzygium oleosum*) presented a very specific TE profile of Class II elements, including TIR/CACTA, TIR/Mutator, TIR/PIF Harbinger, and TIR/hAT (Fig. 2A), which is not found on other species.

As the assemblies were retrieved from NCBI's genome database, we investigated whether any ulterior information could be associated with the results obtained. Available metadata on genome sequencing methods was not correlated with overall TE sequence numbers (data not shown). Furthermore, we calculated whether assembly fragmentation could have impacted TE identification, and apart from polintons, no other TE sequence numbers were found to be associated with assembly fragmentation

Table 1 Myrtaceae species genome assembly metrics

	BUSCO					Metrics			
Species	**C**	**S**	**D**	**F**	**M**	**N50**	**Scaffolds**	**≥ 50kbp**	**Size**
Angophora floribunda	96.9	92.9	4.0	1.3	1.8	36.024	37	25	388.348
Corymbia calophylla	97.8	93.7	4.1	0.9	1.3	39.866	35	17	394.859
Corymbia citriodora	98.4	92.8	5.6	0.8	0.8	31.549	14887	569	537.870
Corymbia maculata	97.3	95.0	2.3	1.0	1.7	40.548	25	15	403.964
Eucalyptus camaldulensis	96.8	92.3	4.5	1.3	1.9	52.653	18	14	558.569
Eucalyptus globulus	96.7	91.8	4.9	1.3	2.0	51.386	48	43	545.015
Eucalyptus grandis	96.2	90.2	6.0	1.6	2.2	58.486	37	34	616.359
Melaleuca alternifolia	98.4	91.3	7.1	0.5	1.1	1.894	3128	504	362.022
Melaleuca quinquenervia	98.5	96.8	1.7	0.6	0.9	22.766	196	92	269.216
Metrosideros polymorpha	97.7	95.7	2.0	1.0	1.3	26.760	11	11	274.748
Psidium guajava	98.4	96.3	2.1	0.6	1.0	40.370	44	35	443.753
Rhodamnia argentea	98.1	94.8	3.3	0.7	1.2	32.337	75	35	346.711
Rhodamnia rubescens	96.9	91.2	5.7	1.3	1.8	3.644	174	165	353.992
Rhodomyrtus psidioides	98.6	87.5	11.1	0.4	1.0	5.669	828	380	931.146
Rhodomyrtus tomentosa	98.1	96.0	2.1	0.7	1.2	43.802	11	11	470.350
Syzygium aromaticum	98.1	95.9	2.2	0.9	1.0	35.418	24	22	370.222
Syzygium grande	94.5	91.9	2.6	1.3	4.2	39.560	174	111	405.097
Syzygium oleosum	98.1	92.9	5.2	0.7	1.2	11.740	72	72	407.067

BUSCO genes are divided into percentages of complete (C), single (S), duplicated (D), fragmented (F) and missing (M). BUSCO annotation was done using the eudicots_odb10 library with 2326 genes. Metrics are given in Mbp

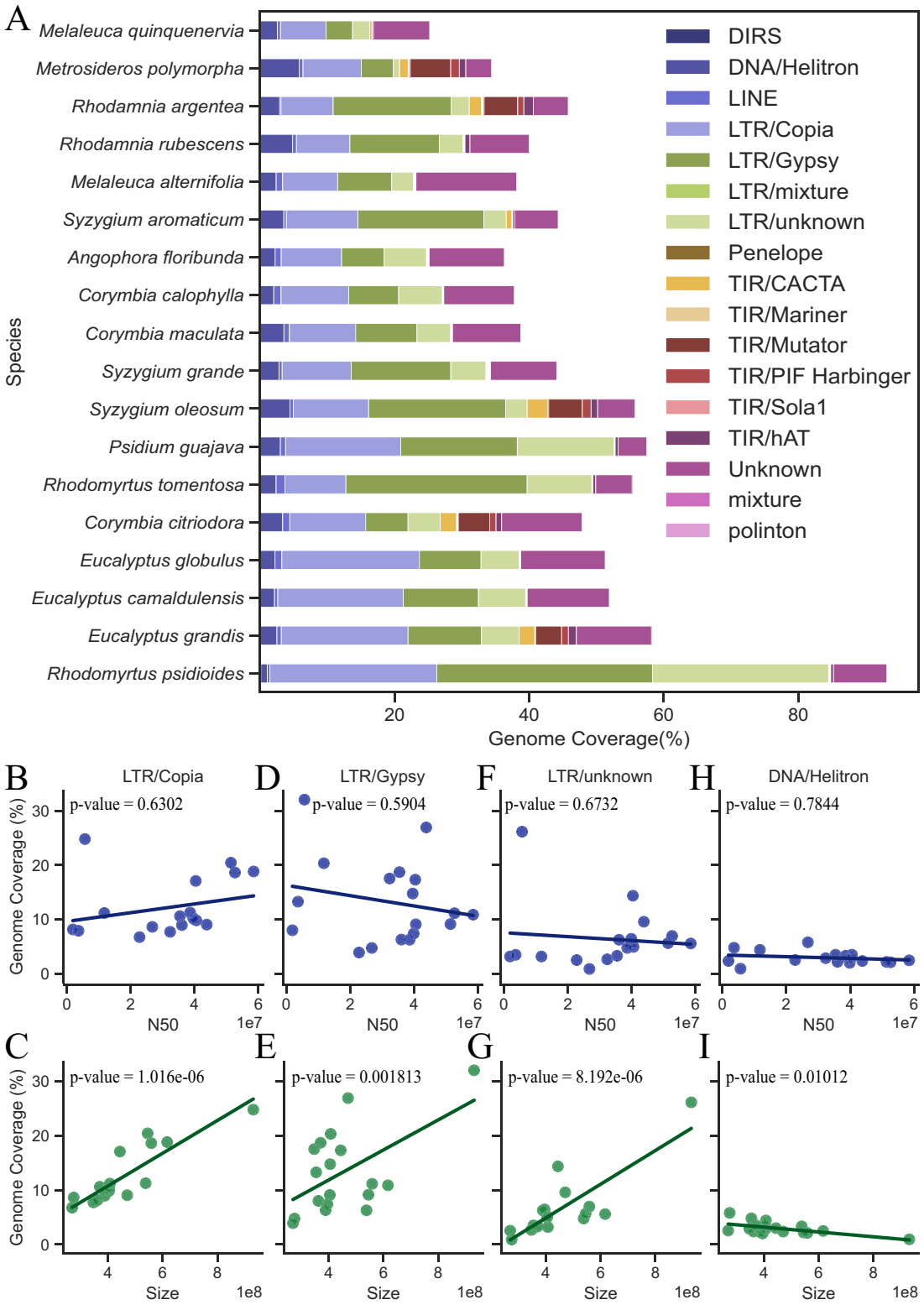

Fig. 2 Myrtaceae TE composition and correlation with genome metrics. TE composition of 18 Myrtaceae species ordered by assembly size. The X axis represents the total percentage every element type occupies in the assembly (**A**). Linear correlation plots of N50 (**B**, **D**, **F** and **H**) and assembly size (**C**, **E**, **G** and **I**) values from the Myrtaceae assemblies and assembly coverage percentages of LTR/Copia (**B** and **C**), LTR/Gypsy (**D** and **E**), LTR/Unknown (**F** and **G**) and DNA/Helitron (H and I) elements

values (N50) (Fig. 2B, D, F, H) (Supplementary Figure 1 and Supplementary Material 2).

Considering that TEs are the main players in genome size variations, we sought to identify the main lineages with the highest impacts. To achieve this, we plotted linear correlations among assembly sizes and TE genome coverage values, calculating the *p*-value by normalizing for phylogenetic relationships for each plot (Supplementary Figure 2). A total of eight TEs were found to be statistically correlated with assembly size. Half of which (Penelope, DIRS, TIR/Sola1, and LTR/mixture) had values proximal to zero. The other half comprises the Class II DNA/Helitron (Fig. 2I), negatively correlated with assembly size, and the major division of LTR elements (LTR/Unknown, LTR/Gypsy, and LTR/Copia) (Fig. 2C, E, G), all positively correlated with assembly size.

LTR/Copia and Gypsy are taxon-specifically enriched

As Gypsy and Copia are TE super-families, we sought to annotate the identified LTR sequences into their different lineages. Fifteen lineages of LTRs were identified with over 100 members in at least one out of the 18 species, including the nine Copia lineages (Ale, Alesia, Angela, Bianca, Ikeros, Ivana, SIRE, TAR, and Tork) and six Gypsy lineages (Athila, CRM, Galadriel, Ogre, Tekay, and non-chromo-outgroup).

Similarly, the copy number of each LTR lineage was plotted against the genomic N50 values to identify potential biases in their identification, but no *p*-value was significant (Supplementary Figure 3 and Supplementary Material 3). On the contrary, when comparing copy numbers against genome size, all of the LTR lineages resulted in statistically significant *p*-values, indicating that all lineages play a role in genome size variations (Supplementary Figure 4). Furthermore, by exploring the relative abundance of LTR lineages among Myrtaceae (Fig. 3A), we can identify taxon-specific clusters. For example, the lineages of Copia elements, Ikeros and SIRE, were found to be enriched specifically on the *Eucalyptus* genus, deviating from others in its tribe. The Gypsy/Tekay elements are the ones that contribute the most to the overall TE libraries of Myrtaceae species, especially in the Syzygieae tribe. *Rhodomyrtus tomentosa* represents a special case with inflation of Gypsy/Tekay numbers, the highest number identified in all species (11615 sequences, representing 41% of the total LTRs identified in *R. tomentosa*). This event is likely very recent in its genus diversification, and thus species-specific, as no similar profiles have been identified in Myrteae. Lastly, the lineage of Gypsy/Ogre TEs is another significant contributor to genome size variation in the Myrteae tribe species. Interestingly, *Rhodamnia argentea* possesses almost 30% of the total

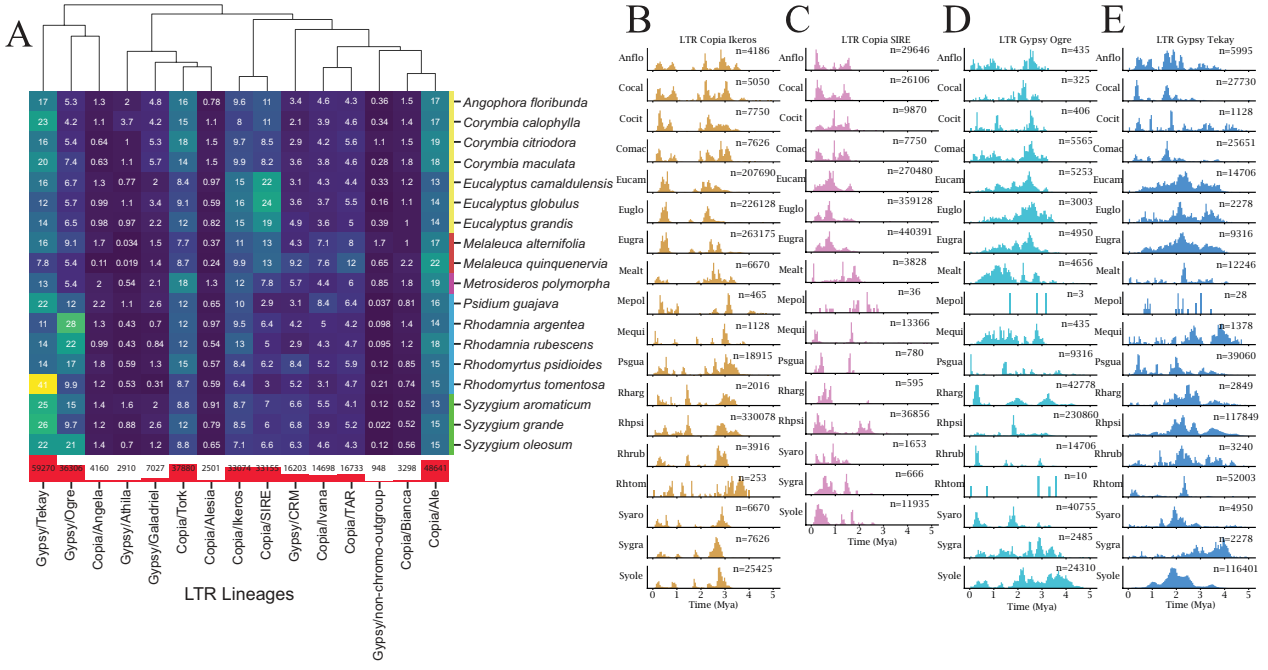

Fig. 3 LTR TE lineage representation among Myrtaceae species and insertion times. Clustermap of LTR TE library percentages (**A**). Every number indicates how much the LTR lineage contributes to the total LTR TE library size of that species. All lines add up to roughly 100% - meaning total LTR TE library size. Species are ordered alphabetically and tribes are depicted in different colored bars: yellow (Eucalipteae), red (Melaleuceae), pink (Metrosidereae), blue (Myrteae) and green (Syzygieae). At the bottom, the total number of LTR lineage copies found in Myrtaceae is shown. LTR insertion times for the four most prevalent lineages: Ikeros (**B**), SIRE (**C**), Ogre (**D**), and Tekay (**E**) are shown on the right. Every species has the total number of comparisons (n) to the right of the curve. Species are identified by the two letters of their genus followed by the three letters of their species

size of the library as Ogre elements, much more than any other species. Considering *Syzygium oleosum* has the highest assembly size of the analyzed Syzygieae species, it also appears to harbor the highest amount of Gypsy/Ogre elements (21% of total LTRs, compared to only 15% and 9.7% for *Syzygium aromaticum* and *Syzygium grande* respectively).

In order to shed light on the ancestry of such lineage expansions and LTR activity, every LTR lineage had its divergence times calculated and analyzed by species (Fig. 3B-E) (Supplementary Figure 5 and 6). Overall, divergence times fluctuate between recent times and 3.5 Mya, with very few LTRs over the 4 Mya mark. These values represent a very recent time window compared to the speciation and distribution of the tribes within the Myrtaceae family, which occurred around 50 million years ago (Mya). Consequently, any inference about ancestral expansion events (those that happened more than 10 Mya) is limited due to these sequences' half-life but can be extrapolated by lineage representation among closely related species. After all, vertical inheritance from a common ancestor is more evolutionary parsimonious than many distinct events of TE replication. LTR lineages show very contrasting peak profiles when comparing LTR divergence times among the Myrtaceae family (Fig. 3B-E) (Supplementary Figure 5 and 6). The Copia/Ikeros lineage possesses two peaks ($\approx$ 3 Mya. and $\approx$ 0.5 Mya) (Fig. 3A). On the other hand, The Copia/SIRE lineage profile is more complex (Fig. 3B), with peak numbers varying among species. The *Eucalyptus* genus, where it is more prevalent, has a main peak just short of 1 Mya. and shares more recent activities with other members of its tribe. Lineages Gypsy/Ogre and Gypsy/Tekay also display many peaks, with more recent events of duplication notable in separate species in the family: the genus *Rhodamnia*, *Psidium guajava* and *Syzygium aromaticum* for Gypsy/Ogre, and *Rhodomyrtus tomentosa*, *Psidium guajava* and *Corymbia calophyla* for Gypsy/Tekay (Fig. 3D and E). Another information to note is that the genus *Syzygiumn* - with proportionally the most contribution of Gypsy/Tekay elements - has had little to no activity detected in the last 1 My. Looking at the lineages that contribute the most to genome size variation, we find that Gypsy/Tekay elements are comparatively the most prevalent lineage in Myrtaceae, composing roughly 20% of the TE repertoire in Myrteae and Syzygieae tribes. Additionally, majority of these elements are likely to be vertically inherited from a common ancestor across all Myrtaceae, purified from the genome ever since. This is corroborated by element divergence times, where species with higher copy numbers, compared to their mean tribe value, are found to have insertion peaks below 0.5 Mya, likely from recent transpositional activity events.

TEs influence diversity and adaptation mechanisms

Knowing that TEs are drivers of genetic variability and have previously been correlated with species adaptation, we sought to identify whether the species-specific TE landscape could, to an extent, be linked to the real-world distribution of Myrtaceae species. For this, we gathered the TE data previously generated, and climatic data from the natural distributions of each Myrtaceae species (Fig. 4A). Species distribution was divided into three main groups: the center-most group composed of Myrtaceae species with east-australian origins, the left group denotes species encompassing regions in Australia not limited to the east coast, and the right group includes species originating from east and south-east Asia. Ungroupped dots are *P. guajava*, with South American origin, and *M. polymorpha*, native to the islands of Hawaii.

By plotting the PCA features, we are able to interpret which variables best explain the dataset's distribution (Fig. 4B, C, and D). As expected, most variables with the highest magnitudes represent climatic features, while TEs explain each less than 1.5% of total variance. However, some TEs appear to contribute more than others to PCA loadings, such as Copia/SIRE, Copia/Bianca and Gypsy/Tekay (Fig. 4C). To evaluate potential TE correlations to climatic variables, a Generalized Additive Model (GAM) was performed among every pair of features using their z-scores relative to the overall species distribution. Two comparisons obtained *p*-values below 0.05 and are shown in Fig. 4E and F. In general, both interactions indicate an increase in TE numbers as the climatic variables deviate from the group mean, which could be a sign of TE activity in more extreme abiotic conditions.

Furthermore, we performed a Gene Ontology over-representation analysis including genes within 2 kbp of detected TEs, and separated them by lineages (Fig. 5). Overall, the genes that are under the most influence participate in processes related to signaling, cell communication, and defense responses, present in almost all lineages. Some biological processes refer to different facets of the same process. These include: toxin regulation (signalling, synthesis, metabolism, xenobiotic export, and xenobiotic detoxification), carbohydrate pathways (metabolism, and transport), response to bacteria (defense response, and unspecified response), reactive species detoxification (hydrogen peroxide metabolism, catabolism, and detoxification), and lipid pathways (metabolism, catabolism, and biosynthesis). Other biological processes are related to plant secondary metabolism and were found to be specific to some LTR lineages, for example Gypsy/Tekay elements are near terpenoid, diterpenoid, and isoprenoid biosynthetic genes, while Gypsy/Ogre elements were one of the few TEs enriched in the vicinity of lipid biosynthetic genes and carbohydrate transporters. This is

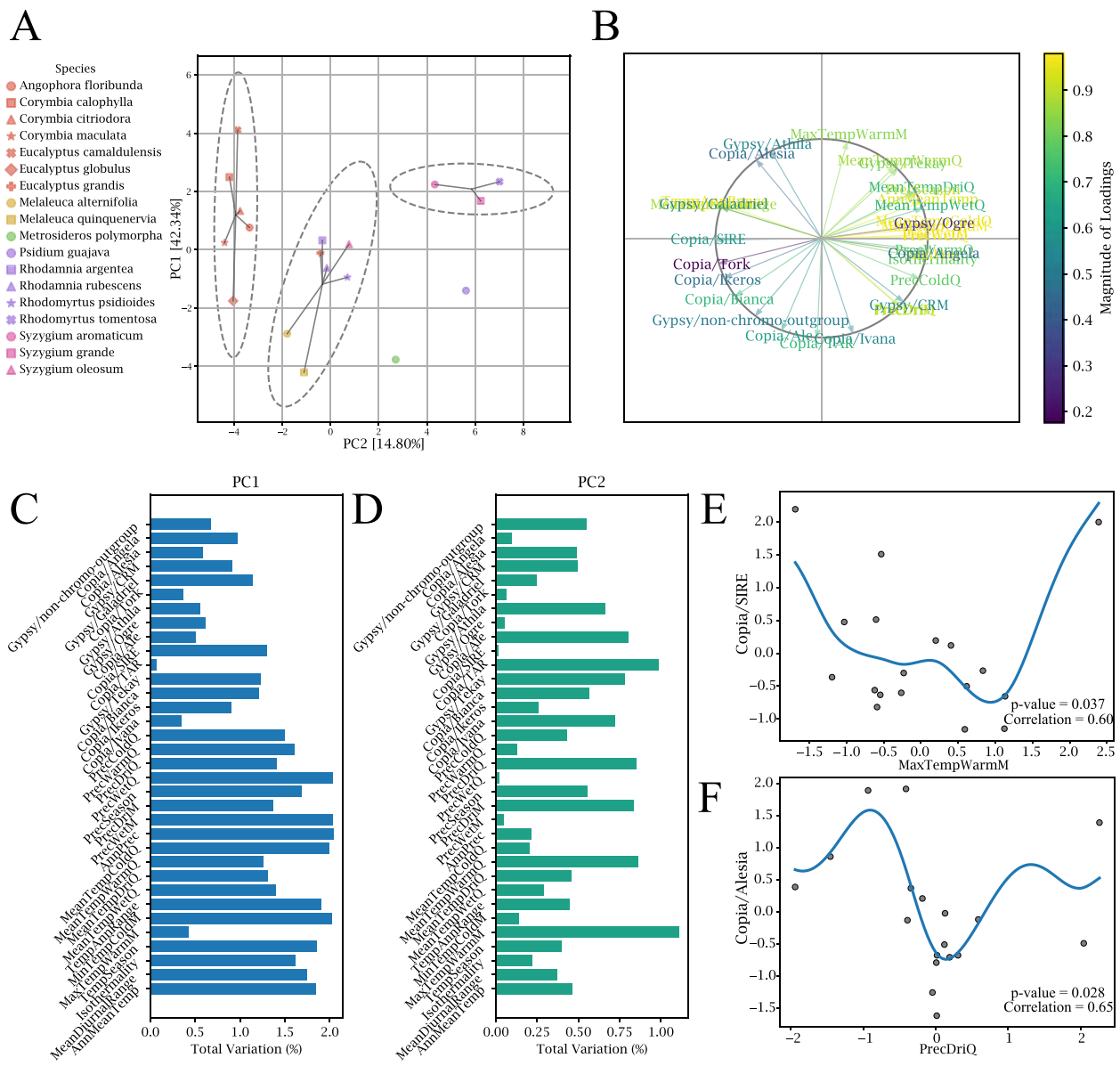

Fig. 4 Correlation of TE presence and climatic variables. PCA plot (**A**) and bi-plot (**B**). Ellipses correspond to 95% confidence regions. Contribution of each variable to the total variance of PC1 (**C**) and PC2 (**D**). Generalized Additive Models (GAM) of TEs and Climatic Variables (**E** and **F**)

interesting because these TEs could influence the diversity of secondary metabolites in Myrtaceae and subsequently affect adaptation mechanisms.

Myrtaceae TE composition faces ecosystem interaction

In order to identify possible HTT events occuring in Myrtaceae, we retrieved high identity hits between species. For each pair comparison among Myrtaceae species, hits above a calculated identity threshold (see Methods "Horizontal TE transfer identification" section) were retrieved and are available in the Supplementary Figure 7 and 8.

Many hits were from *E. grandis* and the tribes Myrteae and Syzygieae. Some of the hits provide alignment bit-scores above 6000, indicating not only high levels of identity, but along most of the sequence (Supplementary Figure 8). Many of these hits belonged to the Copia and Gypsy super-families, and recovered the species phylogenetic relationships. Although phylogenetic analyses indicate they do not represent HTT events, these are likely many ancestral domestication events that have been vertically inherited. However, we retrieved one phylogenetic relationship of a Copia/Ivana TE suggesting a positive HTT event (Fig. 6A). To determine its likelihood and to dispose of a TE domestication hypothesis, we used both the tree topology and codon-alingment of the available overlapping domains, restricted to only the *gag* domain, in a selective pressure analysis. By selecting

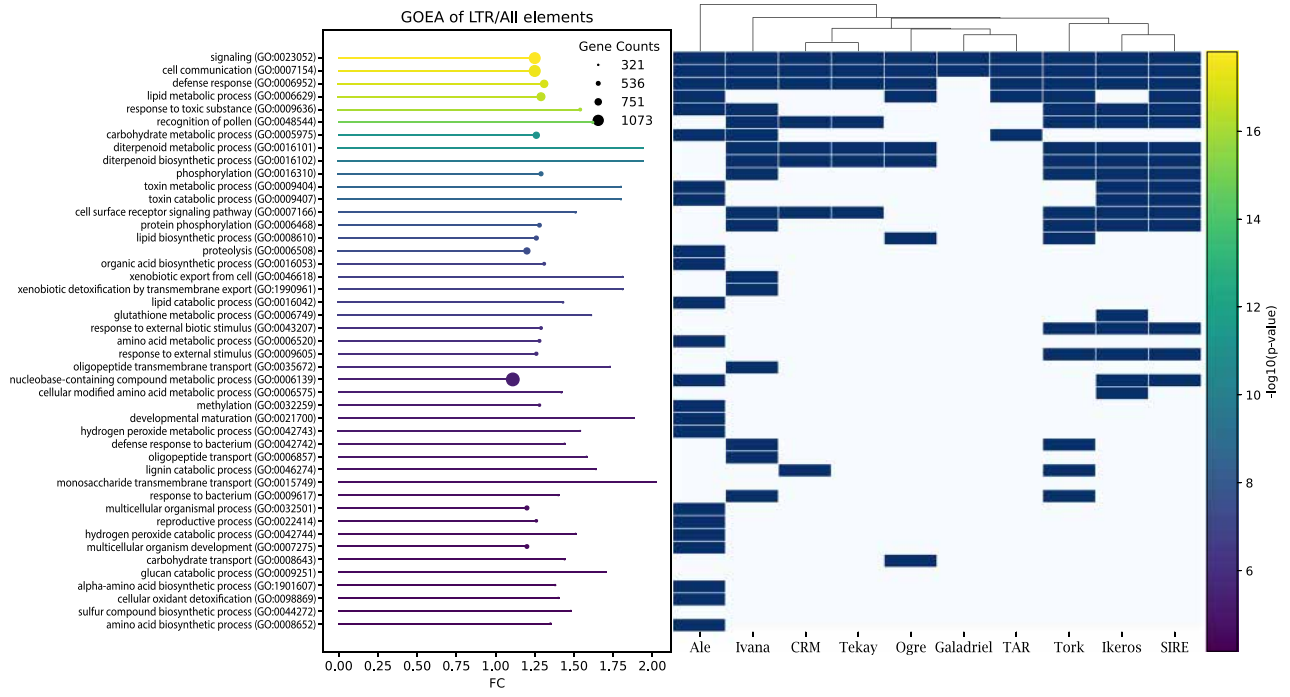

Fig. 5 Characterization of genetic networks nearby TEs. Gene Ontology over-representation analysis with a presence-absence clustermap of each TE lineage. Colors on the graph indicate *p*-values, while dot sizes represent the number of genes identified. The placement on the axis denotes how many times the GO term is over-represented in the dataset

the tree branches related to the HTT + Syzygium clade and its main subbranches, the analysis resulted in a non-statistically significant *p*-value (*p*-value = 0.05208). Thus, the model that best explains the selective evolutionary forces among the branches of the tree is a contiguous and homogeneous negative pressure/neutral evolution, favoring synonymous substitutions over non-synonymous, i.e. $\delta N/\delta S$ (ω) ≤ 1. Furthermore, even if the M1 model reached statistical significance, given the proximity of the *p*-value, the ω value calculated for the branch of HTT *R. argentea* sequences tends to infinity ($\omega = 999$). This indicates diversifying selection, adding more support to the HTT hypothesis if the *p*-value reaches significance. In light of this, the phylogenetical incongruence is likely a HTT event of a Copia/Ivana TE between *R. argentea* and *S. grande* (Fig. 6B). The direction of the transposition is not clear, but given the distance values of the branches, it is more parsimonious that a Copia/Ivana TE jumped from *S. grande* onto *R. argentea*, and has slowly been purified by the host since entering. Although *R. argentea* plants have not been reported in Malaysa nor the Philippines, *S. grande* has a wider global distribution, including Australia, which could explain the necessary contact for an HTT event. Furthermore, *S. grande* seeds are dispersed by flying migratory animals, such as bats, and are also capable of being dispersed by water, increasing the genetic flow between island regions.

Discussion

Myrtaceae species are part of one of the most diverse families in the plant kingdom. This diversity is reflected in their repertoire of secondary metabolites, adaptation capabilities, and genome organization. The publishing of the first Myrtaceae genome was a huge stepping stone, allowing a more thorough study of complex gene networks and regulatory sequences to this day [7]. However, only 1% of the family has sequenced genomes, with great bias between its tribes. Nonetheless, in our study, we analyzed available genomic data of Myrtaceae to understand the genetic mechanisms related to TEs that drive their diversity and adaptation. Of the 22 publicly available genome assemblies chosen, 18 were suitable for downstream analyses, a couple with published data on TE identification. When comparing studies, many discrepancies in TE identification were detected, which are inherent to tool usage, and given that the majority of studies choose mainly RepeatModeler, we expect that a robust approach using EDTA, a pipeline that incorporates RepeatModeler and other five programs [25] (together with TEsorter for TE reannotation), will contribute to better annotated libraries.

In addition to Eudicots and Myrtales genome duplication events [42], TEs appear to be the main players in genome size variation within Myrtaceae. Specifically, the LTR super-families Gypsy and Copia were found to be positively correlated with an increase in assembly

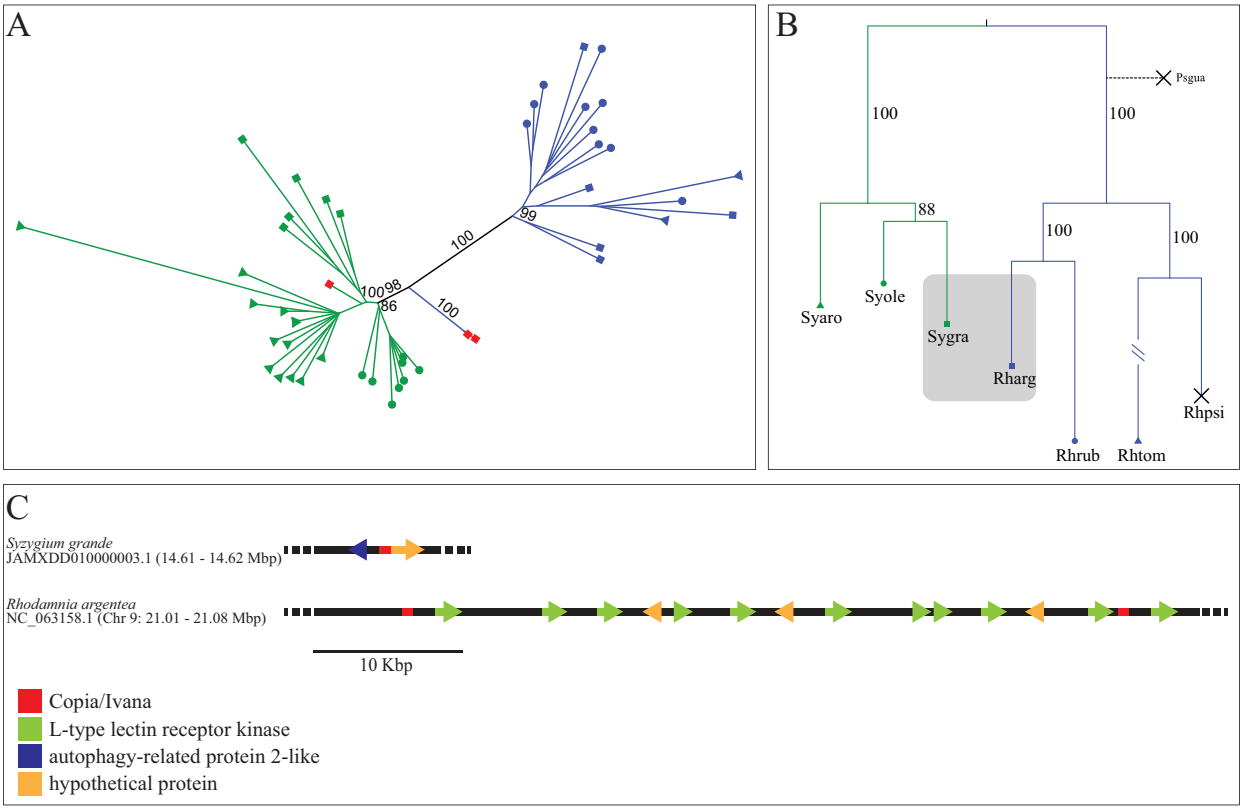

Fig. 6 Case study of HTT in Myrtaceae. Phylogenetic reconstruction of the relationships among Copia/Ivana TEs (**A**). High identity sequences are depicted in red. Phylogenetic relationship among the Syzygieae and Myrteae clade calculated by aligning 100 BUSCO genes (**B**). Likely HTT event represented by a gray box. Myrteae species are colored in blue: *R. argentea* (square), *R. rubescens* (circle), *R. tomentosa* (triangle). *Syzygium* species are colored in green: *S. aromaticum* (triange), *S. grande* (square) and *S. oleosum* (circle). Genetic organization near HTT sequences (**C**)

size. Additionally, these super-families are divided into many different lineages, and nearly all were found to be positively correlated with assembly size. Specific lineage contributions can be seen in Fig. 3A, together with TE divergence times for the most prevalent ones. Overall, lineage divergences appear distinct, with Gypsy/Tekay losing activity in some species, while Copia/SIRE elements seem more active in recent times, and finally Copia/Ikeros with peaks both in recent and more distant time windows. In general, it is a considerable limitation of any study to seek correlations between TE profiles and climate, because the latter is only available for the past 100 years, while TE profiles result from a cumulative history of climate interactions for the past millions of years, together with each species biological history. Nevertheless, we sought to identify whether there is any correlation and, if so, to what extent each impacts each other. As such, most TE profiles do not correlate with current climate variables, with the exception of Copia/Alesia and Copia/SIRE elements (Fig. 4E and F). The former appears to correlate with precipitation values and the latter with maximum temperature, both with an R^2 close to 60%. As the values deviate from the mean, a TE increase is expected. Both lineages also appear to have the most activity in recent times, which may help explain why they better correlate with climatic variables (Supplementary Figures 5 and 6).

Copia/Ivana elements were not correlated with any variables, although they were increased *Psidium guajava* and in the *Melaleuca* genus (Fig. 3). Its major activity peak stands around 2 to 3 Mya and displays recent activity across all species. The most prevalent genetic networks influenced by Copia/Ivana elements seem to act in microorganism detection and signaling, such as response to bacterium, defense response to bacterium, xenobiotic export from cell, and xenobiotic detoxification by transmembrane export. Curiously, these TEs were also found to have partaken in a HTT event between *Syzygium grande* and *Rhodamnia argentea* (Fig. 6). HTTs have been detected across the plant kingdom in a widely distributed fashion [43], indicating an influence of the ecosystem onto the species genome (since they jump the reproductive barrier), and as many of the Myrtaceae species here analyzed share a common incidence region, they are likely subjected to similar evolutionary influences. While the majority of other TEs analyzed displayed genomic patterns more akin to domestication, a single Copia/Ivana element likely jumped from *S. grande* - in a

genomic region near an autophagy-related protein2-like gene - onto a genomic locus in *R. argentea* rich in L-type Lectin Receptor Kinases (LRKs) over 800 thousand years ago. These L-type LRKs compose a gene family that orchestrates plant immunity by providing broad-spectrum disease resistance. By acting as cell membrane receptors, these genes were also found to respond to plant-hormones and consequently abiotic stresses [44]. As such, HTT events are likely related to adaptation processes in Myrtaceae, in this case the transfer of a Copia/Ivana element from *S. grande* onto a genomic region in *R. argentea*, where it duplicated in the last 50 thousand years, and both copies possibly influence L-type LRKs in their role in plant immune responses and abiotic stresses. Lastly, it is uncertain whether there are more HTT events among Myrtaceae, since our approach of BUSCO identity distributions may have proven to be too strict (once it did not allow comparisons among the same genera). This may also have been due to HTT events being intrinsically challenging to detect in this family, either because they are rare or efficiently purified from the genome. Other studies that analyzed HTT events across distantly related plant taxa have also included *Eucalyptus grandis* in their analyses, but did not report any positive cases [43, 45].

By checking genes that were recovered in all species (Supplementary Material 4), disregarding the LTR lineage, many presented the Toll/Interleukin-1 receptor (TIR) domain, with 239 TIR domain genes found in the GO over-representation gene list. This domain is commonly found in a subgroup of nucleotide binding sites and leucine-rich repeat (NBS-LRR) proteins, responsible for monitoring pathogen infection and mediating plant immune responses [46], one of the over-represented biological processes recovered. NBS-LRR genes are one of the biggest gene superfamilies in plants, representing over 1% of coding genes. 1215 putative genes were identified in *E. grandis*, with Myrtaceae species sharing an increase in NBS-LRR TIR-domain containing (TNL) subgroup numbers in both *E.grandis* and *Melaleuca quinquenervia* [47, 48]. This incredible diversity is explained by repeated tandem duplications of genes in uneven crossing-over events, yet this mechanism is not sufficient to explain their distribution across chromosomes, and mixed subfamily gene clusters [46, 49]. Furthermore, TNLs in *Lactuca sativa* were able to be divided in Type-I, and Type-II based on a number of characteristics that mainly relate to different rates of evolution [50]. Type-I sequences were enriched in the center of gene clusters and presented 3' intron identities higher than flanking coding sequences, which could be a case of LTR TE mediated template switching [51].

Other enriched gene families are the terpene synthase (TPS) and terpene cyclase genes. These genes encode for a family of enzymes responsible for the production of a class of fragrant oils that are heavily sough after in the industry, especially in Myrtaceae [52]. The phylogenetic relationships among the TPS genes subdivide them into seven clades, with superposition of chemical products [53]. Subfamilies a, b and g are mainly involved in the secondary metabolism of plants, while subfamilies c, e and f are specialized in the primary metabolism [54]. A previous genome-wide study analyzed this gene family in 50 plant genomes, and found lineage-specific family expansions events, especially in *E. grandis* [55]. The *Eucalyptus* genera possesses around 100 TPS genes arranged in genetic clusters with many events of pseudogenization [56], similar to the NBS-LRR family. Furthermore, 30% of TPS genes were found to be flanked by TEs in maize, and although tandem duplication plays a major role in the increase of gene copy numbers, TEs were also found to contribute to the diversification of the TPS gene family in some plant lineages [55]. This indicates that the great diversity of TPS genes in Myrtaceae has likely been influenced by transpositional activity.

Lastly, many genes presented the 2-oxoglutarate/Fe(II)-dependent dioxygenase (2OGD) domain, catalyzing chemical reactions including, but not limited to, hydroxylations, demethylations, desaturations, ring closure and ring cleavage [57]. These genes are part of the 2OGD gene superfamily, with 130 members identified in *Arabidopsis thaliana* and 271 members in *Brassica napus* [58]. Its biggest subfamily - DOXC - partakes in the biosynthetic process of many plant hormones and secondary metabolites [59], and comprises genes present in the GO over-representation list, such as *antocyanidin synthase* and *flavonol synthase* related to antocyanidin biosynthesis [60]. Other DOXC members have been found to increase drought tolerance when silenced [61], and the knockdown mutant of *downy-mildew resistant 6* was found to also grant disease resistance to necrotrophic pathogens [62] (*downy-mildew resistant 6* was also found in our GO over-representation list). Overall, TEs seem to influence the majority of plant biology aspects; this is likely done by influencing the evolution and neofunctionalization of gene superfamilies, which in turn are responsible for core mechanisms that orchestrate the organism's biology and adaptation to the environment.

Conclusion

Myrtaceae is one of the most diverse families in plants, with economically important species due to their wood, spices, fleshy fruits and essential oils. This plant family includes already declared invasive members with worldwide distributions, exemplifying their high adaptability to different environments. Other members are more known for their resistance to stress, ability to survive in nutrient poor soils as well as water deficient environments [63]. It is well known that TEs influence gene expression and

may lead to gene duplications, thus acting as diversifying forces to adaptation processes. In our study, we characterized the TE landscape in 18 Myrtaceae species, identifying Class I retrotransposons as the main contributors to genome size variations encountered in these species. More specifically, every LTR lineage of the retrotransposons analyzed contributed to genome variation. Moreover, some lineages are more enriched in some taxons than others, leading to characteristic profiles within the subdivisions of this plant family that may, or may not, have similar transpositional activity histories. By exploring species distribution and climatic data, we were able to correlate the presence of certain LTR lineages with abiotic factors. Together with a GO over-representation analysis of proximal genes, we identified the main biological processes these same TEs may influence and consequently aid the adaptation process of some species within the Myrtaceae plant family. The vast repertoire of influenced genes were found to act in defense responses to other organisms, especially the TNS subgroup of NBS-LRR proteins. This indicates a deep interplay between TEs and gene superfamilies to the molecular communication with microorganisms, be they pathogens or symbiotic. Furthermore, it also suggests that microorganisms interactions may play a bigger role in the adaptation process than previously thought. An adaptation process that is not dependent on only a singular species. As our HTT analysis suggests, TEs may jump the horizontal barrier amongst Myrtaceae species, and insert themselves in gene-rich regions and influence proximal genes. In the case of *Rhodamnia argentea*, a Copia/Ivana TE likely jumped from *Syzygium grande* onto an L-type LRK gene rich region, where it may influence the regulation of interspecies communication with microorganisms. Additionally, TPS genes, responsible for the production of terpenes and essential oils (a huge factor in the economical relevancy of Myrtaceae species) also seem to be affected by TEs. Recently, a study detected HTT events between TPS genes amongst plants and microorganisms [64], which together with our results, may point to TEs as a possible explanation to these events. It would be interesting to evaluate how much these interspecies interactions influenced by defense response genes facilitate HTT events and the genetic diversity of plant secondary metabolites.

Abbreviations

2OGD	2-oxoglutarate/Fe(II)-dependent dioxygenase
GO	Gene ontology
HTT	Horizontal transposable element transfer
LRK	Lectin receptor kinase
LTR	Long terminal repeats
L-type LRK	Legume-type LRK
NBS-LRR	Nucleotide binding sites and leucine-rich repeat
PCA	Principal component analysis
TE	Transposable element
TIR	Toll/Interleukin-1 receptor
TNL	NBS-LRR TIR-domain containing
TPS	Terpene synthase

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13100-025-00388-3.

Supplementary Material 1: `material_s1.txt` – URLs for utilized genome assemblies. `material_s2.txt` – TEs PGLS results. `material_s3.txt` – LTR lineages PGLS results. `material_s4.txt` – Complete gene enrichment list. `material_s5.txt` – BUSCO table of annotated genomes. `sup_figures.pdf` – Supplementary Figures 1 through 10.

Acknowledgements

We are grateful to the Institut Français de Bioinformatique and the IFB Core Cluster platform, financed under the Programme d'Investissements d'Avenir funded by the Agence Nationale de la Recherche (RENABI-IFB ANR-11-INBS-0013 and MUDIS4LS ANR-21-ESRE-0048), for providing help and/or computing and/or storage resources. We would also like to thank l'Ambassade de France au Brésil and Campus France for enabling the collaboration among our research groups by awarding ELW a 4-month mobility grant in Montpellier, France.

Authors' contributions

ELW performed the main writing and data analysis, JPCF contributed with the HTT analysis and manuscript revision. HdRMA, MD, and RG reviewed the manuscript and contributed with relevant methodological suggestions. RG further aided the development of the project by facilitating the use of a shared computing environment. ACTZ promoted the conceptualization of the project, together with manuscript revision, and financial support.

Funding

This work was financially supported by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq; grant numbers: 313949/2023-9 and 404419/2024-0), Coordenação de Aperfeicoamento de Pessoal de Nível Superior (CAPES), and Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (grant number 24/2551-0001325-3).

Data availability

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors consent to the publication of the article.

Competing interests

The authors declare no competing interests.

Author details

[1]Programa de Pós-Graduação em Genética e Biologia Molecular (PPGBM), Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalvez, 9500, Porto Alegre 91540-000, Rio Grande do Sul, Brazil
[2]UMR DIADE, IRD, CIRAD, Université de Montpellier, Av. Agropolis, 911, Montpellier 34394, Occitania, France

Received: 16 April 2025 / Accepted: 18 November 2025
Published online: 20 December 2025

References

1. Wilson PG, O'Brien MM, Heslewood MM, Quinn CJ. Relationships within Myrtaceae sensu lato based on a matK phylogeny. Plant Syst Evol. 2005;251(1):3–19. https://doi.org/10.1007/s00606-004-0162-y.
2. Thornhill AH, Ho SYW, Külheim C, Crisp MD. Interpreting the modern distribution of Myrtaceae using a dated molecular phylogeny. Mol Phylogenet Evol. 2015;93:29–43. https://doi.org/10.1016/j.ympev.2015.07.007.
3. Low YW, Rajaraman S, Tomlin CM, Ahmad JA, Ardi WH, Armstrong K, et al. Genomic insights into rapid speciation within the world's largest tree genus Syzygium. Nat Commun. 2022;13(1):5031. https://doi.org/10.1038/s41467-022-32637-x.
4. Mazine FF, Faria JEQ, Giaretta A, Vasconcelos T, Forest F, Lucas E. Phylogeny and biogeography of the hyper–diverse genus Eugenia (Myrtaceae: Myrteae), with emphasis on E. sect. Umbellatae, the most unmanageable clade. TAXON. 2018;67(4):752–69. https://doi.org/10.12705/674.5.
5. Grattapaglia D, Vaillancourt RE, Shepherd M, Thumma BR, Foley W, Külheim C, et al. Progress in Myrtaceae genetics and genomics: Eucalyptus as the pivotal genus. Tree Genet Genomes. 2012;8(3):463–508. https://doi.org/10.1007/s11295-012-0491-x.
6. De Souza Neto JD, Dos Santos EK, Lucas E, Vetö NM, Barrientos-Diaz O, Staggemeier VG, et al. Advances and perspectives on the evolutionary history and diversification of Neotropical Myrteae (Myrtaceae). Bot J Linn Soc. 2022;199(1):173–95. https://doi.org/10.1093/botlinnean/boab095.
7. Myburg AA, Grattapaglia D, Tuskan GA, Hellsten U, Hayes RD, Grimwood J, et al. The genome of Eucalyptus grandis. Nature. 2014;510(7505):356–62. https://doi.org/10.1038/nature13308.
8. Voelker J, Shepherd M, Mauleon R. A high-quality draft genome for Melaleuca alternifolia (tea tree): a new platform for evolutionary genomics of myrtaceous terpene-rich species. GigaByte. 2021;2021:28. https://doi.org/10.46471/gigabyte.28.
9. Ouadi S, Sierro N, Goepfert S, Bovet L, Glauser G, Vallat A, et al. The clove (Syzygium aromaticum) genome provides insights into the eugenol biosynthesis pathway. Commun Biol. 2022;5(1):1–13. https://doi.org/10.1038/s42003-022-03618-z.
10. Bennetzen JL, Wang H. The Contributions of Transposable Elements to the Structure, Function, and Evolution of Plant Genomes. Annu Rev Plant Biol. 2014;65(1):505–30. https://doi.org/10.1146/annurev-arplant-050213-035811.
11. Finnegan DJ. Eukaryotic transposable elements and genome evolution. Trends Genet. 1989;5:103–7. https://doi.org/10.1016/0168-9525(89)90039-5.
12. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82. https://doi.org/10.1038/nrg2165.
13. Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. The genomic ecosystem of transposable elements in maize. PLoS Genet. 2021;17(10):e1009768. https://doi.org/10.1371/journal.pgen.1009768.
14. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19(1):199. https://doi.org/10.1186/s13059-018-1577-z.
15. Bennett MD. Variation in Genomic Form in Plants and Its Ecological Implications. New Phytol. 1987;106:177–200. https://doi.org/10.1111/j.1469-8137.1987.tb04689.x.
16. Casacuberta E, González J. The impact of transposable elements in environmental adaptation. Mol Ecol. 2013;22(6):1503–17. https://doi.org/10.1111/mec.12170.
17. Delaux PM, Schornack S. Plant evolution driven by interactions with symbiotic and pathogenic microbes. Science. 2021;371(6531):6605. https://doi.org/10.1126/science.aba6605.
18. Yu Z, Li J, Wang H, Ping B, Li X, Liu Z, et al. Transposable elements in Rosaceae: insights into genome evolution, expression dynamics, and syntenic gene regulation. Hortic Res. 2024;11(6):uhae118. https://doi.org/10.1093/hr/uhae118.
19. Morata J, Tormo M, Alexiou KG, Vives C, Ramos-Onsins SE, Garcia-Mas J, et al. The Evolutionary Consequences of Transposon-Related Pericentromer Expansion in Melon. Genome Biol Evol. 2018;10(6):1584. https://doi.org/10.1093/gbe/evy115.
20. Huang C, Sun H, Xu D, Chen Q, Liang Y, Wang X, et al. ZmCCT9 enhances maize adaptation to higher latitudes. Proc Natl Acad Sci. 2018;115(2):E334–41. https://doi.org/10.1073/pnas.1718058115.
21. Tralamazza SM, Gluck-Thaler E, Feurtey A, Croll D. Copy number variation introduced by a massive mobile element facilitates global thermal adaptation in a fungal wheat pathogen. Nat Commun. 2024;15:5728. https://doi.org/10.1038/s41467-024-49913-7.
22. Schley RJ, Pellicer J, Ge XJ, Barrett C, Bellot S, Guignard MS, et al. The ecology of palm genomes: repeat-associated genome size expansion is constrained by aridity. New Phytol. 2022;236(2):433–46. https://doi.org/10.1111/nph.18323.
23. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38(10):4647–54. https://doi.org/10.1093/molbev/msab199.
24. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018;34(13):142–50. https://doi.org/10.1093/bioinformatics/bty266.
25. Ou S, Su W, Liao Y, Chougule K, Agda JRA, Hellinga AJ, et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. Genome Biol. 2019;20(1):275. https://doi.org/10.1186/s13059-019-1905-y.
26. Zhang RG, Li GY, Wang XL, Dainat J, Wang ZX, Ou S, et al. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic Res. 2022;9:uhac017. https://doi.org/10.1093/hr/uhac017.
27. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci USA. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.
28. Gabriel L, Brůna T, Hoff KJ, Ebel M, Lomsadze A, Borodovsky M, et al. BRAKER3: Fully Automated Genome Annotation Using RNA-Seq and Protein Evidence with GeneMark-ETP, AUGUSTUS and TSEBRA. bioRxiv. 2023. p. 2023.06.10.544449. https://doi.org/10.1101/2023.06.10.544449.
29. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.
30. Aubin E, Llauro C, Garrigue J, Mirouze M, Panaud O, El Baidouri M. Genome-wide analysis of horizontal transfer in non-model wild species from a natural ecosystem reveals new insights into genetic exchange in plants. PLOS Genet. 2023;19(10):1–26. https://doi.org/10.1371/journal.pgen.1010964.
31. Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform. 2019;20(4):1160–6. https://doi.org/10.1093/bib/bbx108.
32. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530–4. https://doi.org/10.1093/molbev/msaa015.
33. Minh BQ, Nguyen MAT, von Haeseler A. Ultrafast Approximation for Phylogenetic Bootstrap. Mol Biol Evol. 2013;30(5):1188–95. https://doi.org/10.1093/molbev/mst024.
34. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW. EasyCodeML: A visual tool for analysis of selection using CodeML. Ecol Evol. 2019;9(7):3891–8. https://doi.org/10.1002/ece3.5015.
35. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980;16:111–20.
36. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16(6):276–7. https://doi.org/10.1016/S0168-9525(00)02024-2.
37. Marcon HS, Domingues DS, Silva JC, Borges RJ, Matioli FF, de Mattos Fontes MR, et al. Transcriptionally active LTR retrotransposons in Eucalyptus genus are differentially expressed and insertionally polymorphic. BMC Plant Biol. 2015;15(1):198. https://doi.org/10.1186/s12870-015-0550-1.
38. Fick SE, Hijmans RJ. WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. Int J Climatol. 2017;37(12):4302–15. https://doi.org/10.1002/joc.5086.
39. GBIF.Org User. Derived Dataset. The Global Biodiversity Information Facility; 2024. https://doi.org/10.15468/dd.ktwpe4.
40. Govaerts R, Nic Lughadha E, Black N, Turner R, Paton A. The World Checklist of Vascular Plants, a continuously updated resource for exploring global plant diversity. Sci Data. 2021;8(1):215. https://doi.org/10.1038/s41597-021-00997-6.
41. Mi H, Muruganujan A, Huang X, Ebert D, Mills C, Guo X, et al. Protocol Update for large-scale genome and gene function analysis with the PANTHER classification system (v.14.0). Nat Protoc. 2019;14(3):703–21. https://doi.org/10.1038/s41596-019-0128-8.
42. Hao Y, Zhou YZ, Chen B, Chen GZ, Wen ZY, Zhang D, et al. The Melastoma dodecandrum genome and the evolution of Myrtales. J Genet Genomics. 2022;49(2):120–31. https://doi.org/10.1016/j.jgg.2021.10.004.

43. Baidouri ME, Carpentier MC, Cooke R, Gao D, Lasserre E, Llauro C, et al. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. 2014;24(5):831–8. https://doi.org/10.1101/gr.164400.113.
44. Wang Y, Bouwmeester K. L-type lectin receptor kinases: New forces in plant immunity. PLoS Pathog. 2017;13(8):e1006433. https://doi.org/10.1371/journal.ppat.1006433.
45. Orozco-Arias S, Dupeyron M, Gutiérrez-Duque D, Tabares-Soto R, Guyot R. High nucleotide similarity of three Copia lineage LTR retrotransposons among plant genomes. Genome. 2023;66(3):51–61. https://doi.org/10.1139/gen-2022-0026.
46. McHale L, Tan X, Koehl P, Michelmore RW. Plant NBS-LRR proteins: adaptable guards. Genome Biol. 2006;7(4):212. https://doi.org/10.1186/gb-2006-7-4-212.
47. Christie N, Tobias PA, Naidoo S, Külheim C. The Eucalyptus grandis NBS-LRR Gene Family: Physical Clustering and Expression Hotspots. Front Plant Sci. 2016;6. https://doi.org/10.3389/fpls.2015.01238.
48. Chen SH, Martino AM, Luo Z, Schwessinger B, Jones A, Tolessa T, et al. A high-quality pseudo-phased genome for *Melaleuca quinquenervia* shows allelic diversity of NLR-type resistance genes. GigaScience. 2023;12:giad102. https://doi.org/10.1093/gigascience/giad102.
49. Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW. Genome-Wide Analysis of NBS-LRR–Encoding Genes in Arabidopsis[W]. Plant Cell. 2003;15(4):809–34. https://doi.org/10.1105/tpc.009308.
50. Kuang H, Woo SS, Meyers BC, Nevo E, Michelmore RW. Multiple Genetic Processes Result in Heterogeneous Rates of Evolution within the Major Cluster Disease Resistance Genes in Lettuce. Plant Cell. 2004;16(11):2870. https://doi.org/10.1105/tpc.104.025502.
51. Ma H, Wang M, Zhang YE, Tan S. The power of "controllers": Transposon-mediated duplicated genes evolve towards neofunctionalization. J Genet Genomics. 2023;50(7):462–72. https://doi.org/10.1016/j.jgg.2023.04.003.
52. Padovan A, Keszei A, Külheim C, Foley WJ. The evolution of foliar terpene diversity in Myrtaceae. Phytochem Rev. 2014;13(3):695–716. https://doi.org/10.1007/s11101-013-9331-3.
53. Chen F, Tholl D, Bohlmann J, Pichersky E. The family of terpene synthases in plants: a mid-size family of genes for specialized metabolism that is highly diversified throughout the kingdom. Plant J. 2011;66(1):212–29. https://doi.org/10.1111/j.1365-313X.2011.04520.x.
54. Jia Q, Brown R, Köllner TG, Fu J, Chen X, Wong GKS, et al. Origin and early evolution of the plant terpene synthase family. Proc Natl Acad Sci USA. 2022;119(15):e2100361119. https://doi.org/10.1073/pnas.2100361119.
55. Jiang SY, Jin J, Sarojam R, Ramachandran S. A Comprehensive Survey on the Terpene Synthase Gene Family Provides New Insight into Its Evolutionary Patterns. Genome Biol Evol. 2019;11(8):2078. https://doi.org/10.1093/gbe/evz142.
56. Külheim C, Padovan A, Hefer C, Krause ST, Köllner TG, Myburg AA, et al. The Eucalyptus terpene synthase gene family. BMC Genomics. 2015;16(1):450. https://doi.org/10.1186/s12864-015-1598-x.
57. Farrow SC, Facchini PJ. Functional diversity of 2-oxoglutarate/Fe(II)-dependent dioxygenases in plant metabolism. Front Plant Sci. 2014;5:524. https://doi.org/10.3389/fpls.2014.00524.
58. Jiang D, Li G, Chen G, Lei J, Cao B, Chen C. Genome-Wide Identification and Expression Profiling of 2OGD Superfamily Genes from Three Brassica Plants. Genes. 2021;12(9):1399. https://doi.org/10.3390/genes12091399.
59. Nadi R, Mateo-Bonmatí E, Juan-Vicente L, Micol JL. The 2OGD Superfamily: Emerging Functions in Plant Epigenetics and Hormone Metabolism. Mol Plant. 2018;11(10):1222–4. https://doi.org/10.1016/j.molp.2018.09.002.
60. Wang Y, Shi Y, Li K, Yang D, Liu N, Zhang L, et al. Roles of the 2-Oxoglutarate-Dependent Dioxygenase Superfamily in the Flavonoid Pathway: A Review of the Functional Diversity of F3H, FNS I, FLS, and LDOX/ANS. Molecules. 2021;26(21):6745. https://doi.org/10.3390/molecules26216745.
61. Zhang R, Chen X, Wang Y, Hu X, Zhu Q, Yang L, et al. Genome-wide identification of hormone biosynthetic and metabolism genes in the 2OGD family of tobacco and JOX genes silencing enhances drought tolerance in plants. Int J Biol Macromol. 2024;280:135731. https://doi.org/10.1016/j.ijbiomac.2024.135731.
62. Karlsson M, Kieu NP, Lenman M, Marttila S, Resjö S, Zahid MA, et al. CRISPR/Cas9 genome editing of potato StDMR6-1 results in plants less affected by different stress conditions. Hortic Res. 2024;11(7):uhae130. https://doi.org/10.1093/hr/uhae130.
63. Healey AL, Shepherd M, King GJ, Butler JB, Freeman JS, Lee DJ, et al. Pests, diseases, and aridity have shaped the genome of Corymbia citriodora. Commun Biol. 2021;4:537. https://doi.org/10.1038/s42003-021-02009-0.
64. Yan XM, Zhou SS, Liu H, Zhao SW, Tian XC, Shi TL, et al. Unraveling the evolutionary dynamics of the TPS gene family in land plants. Front Plant Sci. 2023;14. Publisher: Frontiers. https://doi.org/10.3389/fpls.2023.1273648.

Publisher's Note