Genome
An International Journal
for Genetics and Genomics

Article

High nucleotide similarity of three *Copia* lineage LTR retrotransposons among plant genomes

Simon Orozco-Arias[a,b], **Mathilde Dupeyron**[c], **David Gutiérrez-Duque**[d], **Reinel Tabares-Soto**[b,e], **and Romain Guyot**[c,e]

[a]Department of Computer Sciences, Universidad Autónoma de Manizales, Colombia; [b]Department of Systems and Informatics, Universidad de Caldas, Colombia; [c]Institut de Recherche pour le Développement, IRD, CIRAD, Université de Montpellier, France; [d]School of Biology, Universidad de Caldas, Manizales, Colombia; [e]Department of Electronics and Automatization, Universidad Autónoma de Manizales, Colombia

Corresponding author: **Romain Guyot** (email: romain.guyot@ird.fr)

Abstract

Transposable elements (TEs) are mobile elements found in the majority of eukaryotic genomes. TEs deeply impact the structure and evolution of chromosomes and can induce mutations affecting coding genes. In plants, the major group of TEs is long terminal repeat retrotransposons (LTR-RTs). They are classified into superfamilies (*Gypsy*, *Copia*) and subclassified into lineages. Horizontal transfer (HT), defined as the nonsexual transmission of genetic material between species, is a process allowing LTR-RTs to invade a new genome. Although this phenomenon was considered rare, recent studies demonstrate numerous transfers of LTR-RTs. This study aims to determine which LTR-RT lineages are shared with high similarity among 69 plant genomes. We identified and classified 88 450 LTR-RTs and determined 143 cases of high similarities between pairs of genomes. Most of them involved three *Copia* lineages (*Oryco*/*Ivana*, *Retrofit*/*Ale*, and *Tork*/*Tar*/*Ikeros*). A detailed analysis of three cases of high similarities involving *Tork*/*Tar*/*Ikeros* group shows an uneven distribution in the phylogeny of the elements and incongruence with between phylogenetic trees topologies, indicating they could be originated from HTs. Overall, our results suggest that LTR-RT *Copia* lineages share outstanding similarity between distant species and may likely be involved in HT mechanisms more frequent than initially estimated.

Key words: LTR retrotransposons, plant genomes, horizontal transfers, bioinformatics, genomics

Introduction

Transposable elements (TEs) are mobile elements able to move from one locus to another. Their mobility can induce mutations, introduce phenotypic novelties, play key roles in environmental adaptation (Li et al. 2018), and increase their copy number. These elements can also interfere with transcription activities of neighboring genes (Bourque et al. 2018). TEs are classified into two classes according to their replication mechanisms: Class I, called retrotransposons for TEs moving through an RNA intermediate (also called "copy and paste" mechanism) and Class II, or transposons for TEs moving through a DNA intermediate (also called "cut and paste" mechanism) (Wicker et al. 2007). Due to the numerous and potentially harmful consequences of their mobility, genomes developed several ways of fighting against the TE proliferation, for example, through epigenetic silencing (Slotkin and Martienssen 2007). TEs persistence can be seriously challenged, and there are few mechanisms allowing their survival.

One of these mechanisms is the escape from a genome followed by the invasion of a new one through horizontal transfer (HT) (Blumenstiel 2019). HTs are defined as the nonsexual transmission of nuclear and plastid genetic material between species. Although this phenomenon was considered uncommon (Panaud 2016), numerous HTs have been demonstrated in the last years in prokaryotes (Diao et al. 2006) and eukaryotes (Aubin et al. 2021). These transfers could contribute to the evolution and adaptation of species due to the diversity of the new acquired genes (Roulin et al. 2009; Acuña et al. 2012; Rice et al. 2013). HTs can involve nuclear and plastid genes (Aubin et al. 2021), while several studies have also reported the transfer of TEs (or HTT) (Roulin et al. 2008), which sometimes leads to deep genotypic and phenotypic consequences (Gilbert and Feschotte 2018).

In plant genomes, long terminal repeat retrotransposons (LTR-RTs) are the most frequent elements (Grandbastien 2015). These can represent up to 80% of the genome size, such as in wheat or barley. LTR-RTs are classified into *Copia* or *Gypsy* superfamilies according to the internal organization of the coding domains (Gao et al. 2012). Each *Copia* and *Gypsy* superfamily is subclassified into lineages (Wicker et al. 2007) based on specific coding regions and overall structure similarities (Llorens et al. 2009). Although LTR-RTs are frequently compared with retroviruses, the former are endogenous genome

elements generally lacking the genetic material to leave the cell and infect other organisms. In plant genomes, *Copia* and *Gypsy* are subclassified into 16 and 14 lineages, respectively, and some of these are closely related or others restricted to very few species known to date (Neumann et al. 2019). Now, bioinformatic tools allow the precise and rapid classification of LTR-RTs based on domain similarities (Orozco-Arias et al. 2018). HTs can be assessed through different methodologies that search for phylogenetic incongruence of the tree topology between TEs (i.e., if a phylogeny of TEs does not match the species phylogeny) and the host genome, the uneven ("patchy") phylogenetic distribution of conservation (i.e., if a TE shows a random distribution) among organisms, and indeed, the high nucleotide similarity of these elements between distantly related species (Wallau et al. 2012). Taken independently, each method does not indicate HT events, but other mechanisms such as stochastic losses or degradation of sequences (Wallau et al. 2012). More than 2800 HTTs have been documented in eukaryotes, mainly among insect genomes (2248 HTTs) (Peccoud et al. 2017) as well as between mammals and tetrapods (Pace II et al. 2008); fungi and plants (Novikova et al. 2010; Wang et al. 2020); arthropods and conifers (Lin et al. 2016); bivalves and other aquatic species (Metzger et al. 2018); and birds and nematodes (Suh et al. 2016). A significant number of LTR-RT HTs have been detected in 40 plant genomes (El Baidouri et al. 2014). However, no information has been reported about the classification down to the level of the precise lineage of LTR-RTs involved in these HTs.

To understand the evolution of LTR-RT lineages in plants and to explore the potential of some lineages to be involved in HT, we conducted an in silico reassessment of LTR-RTs shared with high sequence similarity among 69 genomes of green plants, including angiosperm and non-angiosperm species. The analysis of 88 450 LTR-RTs indicates that three *Copia* lineages share high sequence similarity between distantly related species. This finding raises several questions such as, are there more HTs than estimated so far in the plant genomes? What could be the relationship between the success of HT of LTR-RTs in plant genomes and these specific lineages? And what is their capacity to persist in a genome and invade a new one?

Materials and methods

Genomic data

We downloaded the genomes of 69 species from 34 plant families (Table S1), representing a total of 68.4 GB of data. The relationships between the species used in this analysis are illustrated in CoGePEDIA (https://genomevolution.org/wiki/index.php/Sequenced_plant_genomes, accessed 29 August 2019) and Timetree (http://www.timetree.org), and summarized in Fig. S1.

LTR-RT identification and classification

The genome sequences were first processed with LTR_STRUC (McCarthy and McDonald 2003) for de novo prediction of LTR-RTs. The 88 450 predicted LTR-RTs were then classified using Inpactor ((Orozco-Arias et al. 2018), https://github.com/simonorozcoarias/Inpactor) into superfamilies (i.e., *Copia* or *Gypsy*) and lineages (Table S2). Inpactor classifies the complete copies only, which resulted in a total of 46 898 LTR-RTs classified at lineage-level. The lineage names were rearranged according to the GyDB ((Lloréns et al. 2008), http://gydb.org) and REXDB classifications (Neumann et al. 2019) to reconcile them where possible. In some cases, synonymy of GyDB and REXDB was used as follows: *Del*/*Tekay*, *Ivana*/*Oryco*, *Ale*/*Retrofit*, and closely related lineages defined in REXDB were grouped to correspond to the GyDB classification as follows: *Tork*/*Tar*/*Ikeros*.

Comparison of predicted LTR-RTs

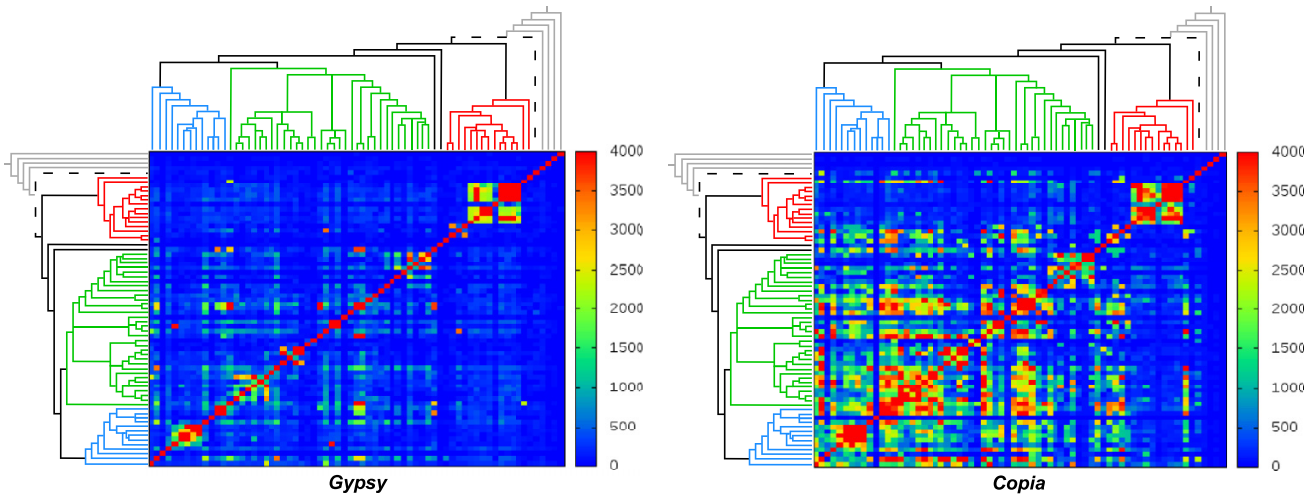
Pairwise comparisons of the predicted LTR-RTs were performed using the BLASTn algorithm (with evalue 1e-4 and the other parameters by default). The LTR-RT superfamilies (i.e., *Copia* and *Gypsy*) and lineages from one species were compared with those from the other species. The highest BLASTn "bit-score" for each pairwise comparison was kept and displayed on a graphical heatmap matrix using gnuplot (http://gnuplot.sourceforge.net). This score is directly associated with each pair of residues according to a nucleotide similarity matrix (match/mismatch scores: 2–3; gap cost: open 5, extension 2). More specifically, a higher score indicates a higher identity in a higher sequence proportion between the aligned residues. We tested different upper limits of similarity on the graphical heatmap, with BLASTn bit-score values of 3500, 3700, 4000, and 5000 (Fig. S2). We selected the score value of 4000 for the heat map representation of Fig. 1, because it allows to appreciate the distribution of the score among species with a large color panel.

The results were organized according to the order shown in the representation of the phylogeny of the plant species (Fig. S1). Moreover, a distribution plot representing the distribution of the sequence alignment size (minimum bit-score: 4000) for all BLASTn is available in Fig. S3. Similarity data are available in an open repository (doi: 10.5281/zenodo.7036190).

Characterization of highly similar LTR-RT elements

The shared LTR-RTs between two species were carefully analyzed and characterized to confirm the high similarity observed using the BLASTn bit-score matrix. The LTR-RTs were annotated using Artemis (Rutherford et al. 2000) and pairwise compared using dotter (Sonnhammer and Durbin 1996) and Stretcher from EMBOSS (Rice et al. 2000). Full-length LTR-RT sequences were aligned using nucmer with the following parameters: -l 10 -c 10—nosimplify –maxmatch, and displayed with mummerplot from the MUMer package (Kurtz et al. 2004). For seven plant species, namely, poplar, coffee, cannabis, orchid, banana, soybean, and kiwifruit, nuclear coding sequences (CDS) were retrieved from public annotation (released to GenBank). CDS were pairwise compared using BLASTn (setting the parameter -qcov_hsp_perc equal to 60) and the percentage of nucleotide identity was extracted. Synonymous substitution analyses (Ks) between

Fig. 1. Representation of the similarity of *Gypsy* and *Copia* elements per genome across 69 plant species. Each *Gypsy* or *Copia* set of elements per genome is pairwise compared with the others using BLASTn. The level of similarity is represented by a heatmap of the BLAST scores (0 in blue: no similarity, to a maximum score of 4000 in red). Species are organized in the matrix as shown in Fig. S1. Only the best score from all possible pairwise comparison scores is plotted for each comparison in the matrix. A symbolic tree of the species was placed laterally and horizontally. Blue: eudicots asterids, green: eudicots rosids, black: basal dicot species, red: monocots, black dashed: *Amborella*, and grey: non-angiosperm. The diagonal represents the similarity of *Gypsy* and *Copia* elements of a species against itself.



pair of plant coding regions were performed on CoGe using CodeML (https://genomevolution.org/). Synonymous substitution analyses (Ks) between pair of LTR-RT coding regions were carried out both on R using the "seqinr" package and with NGSEP (https://github.com/NGSEP/NGSEPcore) (Tello et al. 2022).

Nuclear phylogeny of plant species

The nuclear phylogeny of plant species used in this study was established using near-universal single-copy orthologous genes recovered with BUSCO v5.2.2 for the 69 plant genomes. The phylogeny was carried out using the BUSCO_phylogenomics pipeline (https://github.com/jamiemcg/BUSCO_phylogenomics; (McGowan et al. 2020)) and the supertree approach. This approach is based on the generation of a phylogeny from a set of input trees, which can be generated from different sets of genes, which may be fully or partially overlapping (Fitzpatrick et al. 2006).

Phylogeny of *Tork* LTR-RTs

The phylogeny of *Tork*/*Tar*/*Ikeros* LTR-RTs was carried out using reverse transcriptase domains (RT). RT domains were extracted from the predicted and classified LTR-RTs similarly to Ming et al. (2015). RT classified as *Tork* (1235 sequences) were aligned with MAFFT V. 7.471, (Katoh and Standley 2013) and FastTree V.2.1.10 (Price et al. 2010) was used to perform the phylogenetic analysis with default settings. FastTree was used to infer approximately-maximum-likelihood phylogenetic tree from RT alignments without taking gaps into account. Local support values were computed with the Shimodaira–Hasegawa (SH) test.

Results

Screening of LTR-RT conservations across plant genomes

LTR-RTs were mined from 69 available plant genomes (angiosperm and non-angiosperm species, Viridiplantae; Table S1) using LTR_STRUC (McCarthy and McDonald 2003). We preferred LTR_STRUC over other software since the latter still gives a few false detections of elements in plant genomes (Guyot R.; unpublished). Each set of predicted LTR-RTs for a given species was processed with Inpactor (Orozco-Arias et al. 2018) to classify elements into superfamilies (*Gypsy* or *Copia*) and subclassify them into lineages according to the similarities of five amino acid reference domains: capsid (GAG), aspartic protease (AP), reverse transcriptase (RT), RNAse H (RNAseH), and integrase (INT). Once classified into superfamilies and lineages, the LTR-RTs of a genome are aligned against the elements of other genomes by pairs using BLASTn. The BLASTn bit-scores representing the number of matching and mismatching residues (according to a given substitution matrix) between species (so called "all against all") were displayed by a heatmap. Only the best matching pairs in all possibilities are displayed. The heatmap is organized according to the general phylogeny of the species used in this study (displayed in Fig. S1) to appreciate both similarity and distribution of the elements analyzed (Fig. 1).

For *Gypsy*, high pairwise BLASTn bit-scores were mainly restricted to monocots and to Brassicaceae, legumes, and Solanaceae to a lesser extent. Some punctual high scores were also observed such as between cannabis and kiwi, castor bean and eggplant, or sacred lotus and clementine. For*Copia*, significant scores were observed among all groups of species,

except for *Amborella trichopoda*, *Utricularia gibba* (asterids), and the non-angiosperm species (*Chlamydomonas reinhardtii*, *Physcomitrella patens*, *Selaginella moellendorffii*, *Picea abies*) for which no similarity could be detected. (Fig. 1; Supplementary material S1). Interestingly, the distribution of the observed scores can be divided into three groups dicotyledonous, sharing most of the high similarity across the group; monocotyledonous, with a high similarity also across them (mainly cereals); and orchids, *Amborella* and the non-angiosperm species, for which no significant pairwise scores were observed.

In addition to superfamilies, LTR-RTs were subclassified into lineages or groups of closely related lineages. Briefly, *Copia* lineages (i.e., *Retrofit|Ale*, *Angela*, *Bianca*, *Oryco|Ivana*, *Tork|Tar|Ikeros*, and *SIRE*) and *Gypsy* lineages (i.e., *Athila*, *CRM*, *Del|Tekay*, *Galadriel*, *Reina*, and *TAT*) were identified according to the similarities in their internal amino acid domains (GAG, AP, INT, RT, RNAseH), as found in GyDB ((Lloréns et al. 2008), http://gydb.org) and REXDB (Neumann et al. 2019) (Table S2). The results (BLASTn score with an upper limit of 4000) were plotted as previously, using a heatmap representation organized according to the phylogenetic order of the species.

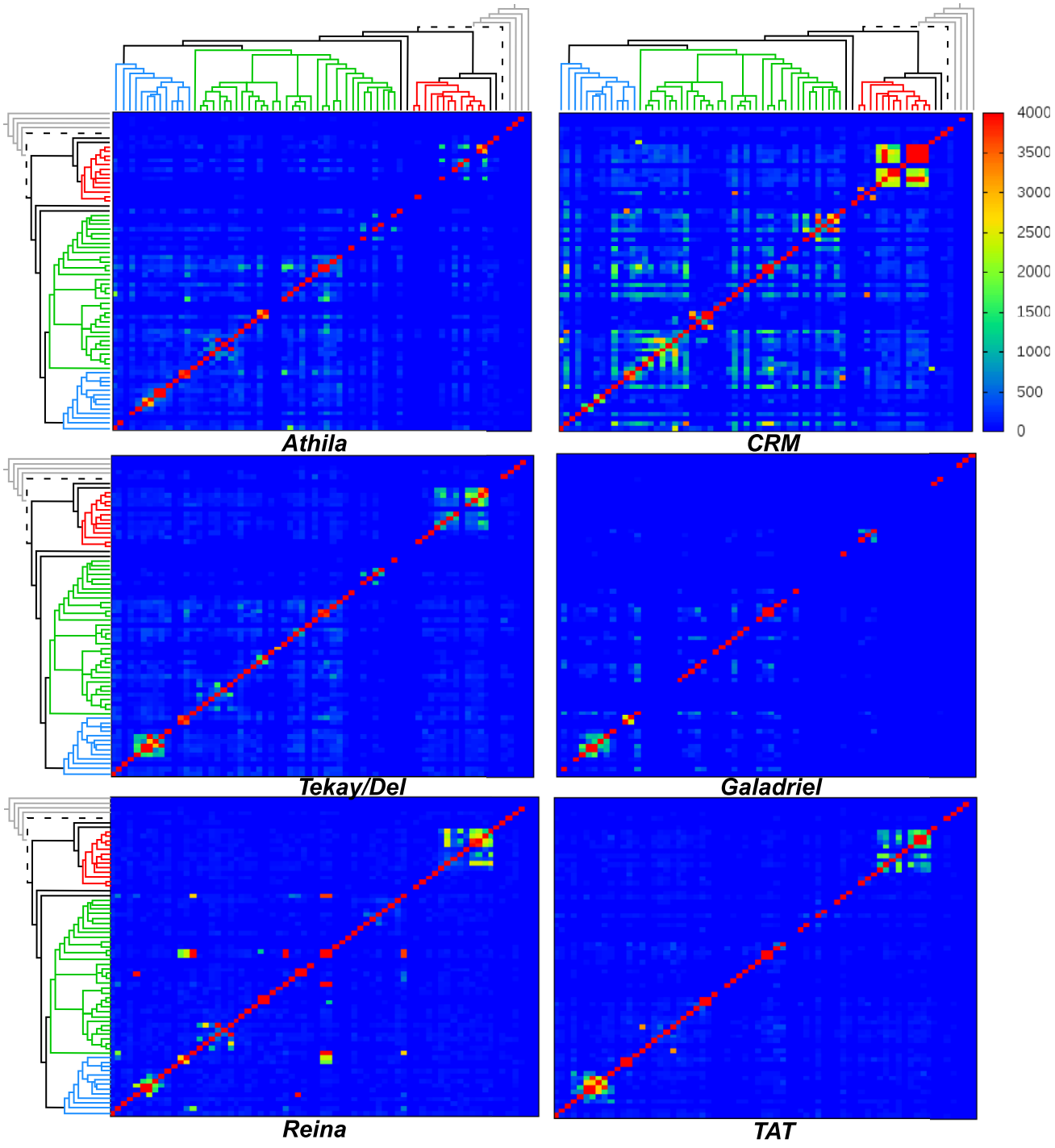
For *Gypsy*, a clear patchy distribution of similarity and difference in the similarity pattern across the six lineages was observed (Fig. 2). *CRM* appears as the *Gypsy* lineage with the highest scores. Breaks in the red diagonal (BLASTn analysis of the LTR-RTs set against itself) suggested that the lineage was not detected by LTR_STRUC and Inpactor. It is interesting to note that the cereal group (corresponding to the red lines in the phylogenetic tree representation) consistently showed a significant level of high scores across species and LTR-RT lineages. The *Chlamyvir*, *Phygy*, and *Selgy* LTR-RT lineages were not found similar between any species (data not shown). For *Copia*, the similarity pattern differs considerably from that of *Gypsy* (Fig. 3), although a patchy distribution of similarity is also evident. Similar to the superfamily analyses, high scores are observed for the lineages. The highest scores are unambiguously for the *Tork|Tar|Ikeros*, *Retrofit|Ale*, and *Oryco|Ivana* lineages. As with *Gypsy* lines, *Copia* lines of cereals have always shown strong pairwise similarities among cereal species. This indicates that the highest level of similarity is found among cereals species rather than with other angiosperms and non-angiosperm plant species. Finally, the *Bryco*, *Lyco*, *Gymco*, and *Osser* lineages were not found similar between species (data not shown).

Among the 69 plant species analyzed here, orange (*C. sinensis*), eucalyptus (*E. grandis*), and prunus (*P. avium*) showed the highest number of highly similar elements with other species (Fig. S4A). Most of the similarity was observed within the dicotyledonous family and little dicot/monocot similarity was noted at this stage of the analysis (BLASTn, score cutoff 4000). Most of these similarities involved *Copia* lineages (*Oryco|Ivana*, *Retrofit|Ale*, and *Tork|Tar|Ikeros*). This finding supports our previous observations using "heatmap" representations (Fig. S4B and Table S2). The *Oryco|Ivana*, *Retrofit|Ale*, and *Tork|Tar|Ikeros* lineages are also the most similar elements between the different plant families (Fig. S4C). In total, we noted 143 cases of high similarity, among them, 98 belong to three *Copia* lineages (*Oryco|Ivana*, *Retrofit|Ale*, and *Tork|Tar|Ikeros*), representing 69% of the highly similar elements. Interestingly, we found that a previously reported detailed case of HTT between coffee and banana (Dias et al. 2015) was clearly displayed on the heatmap of similarity (Fig. 3, white circle). Overall, these results indicate that there is a strong similarity between LTR-RTs involving mainly three lineages of *Copia*, detected in distant angiosperm species. Such similarity of LTR-RTs between distant species raises the question of whether some of them could be the result of HTs.

Detailed analysis of three selected cases of LTR-RT high similarity

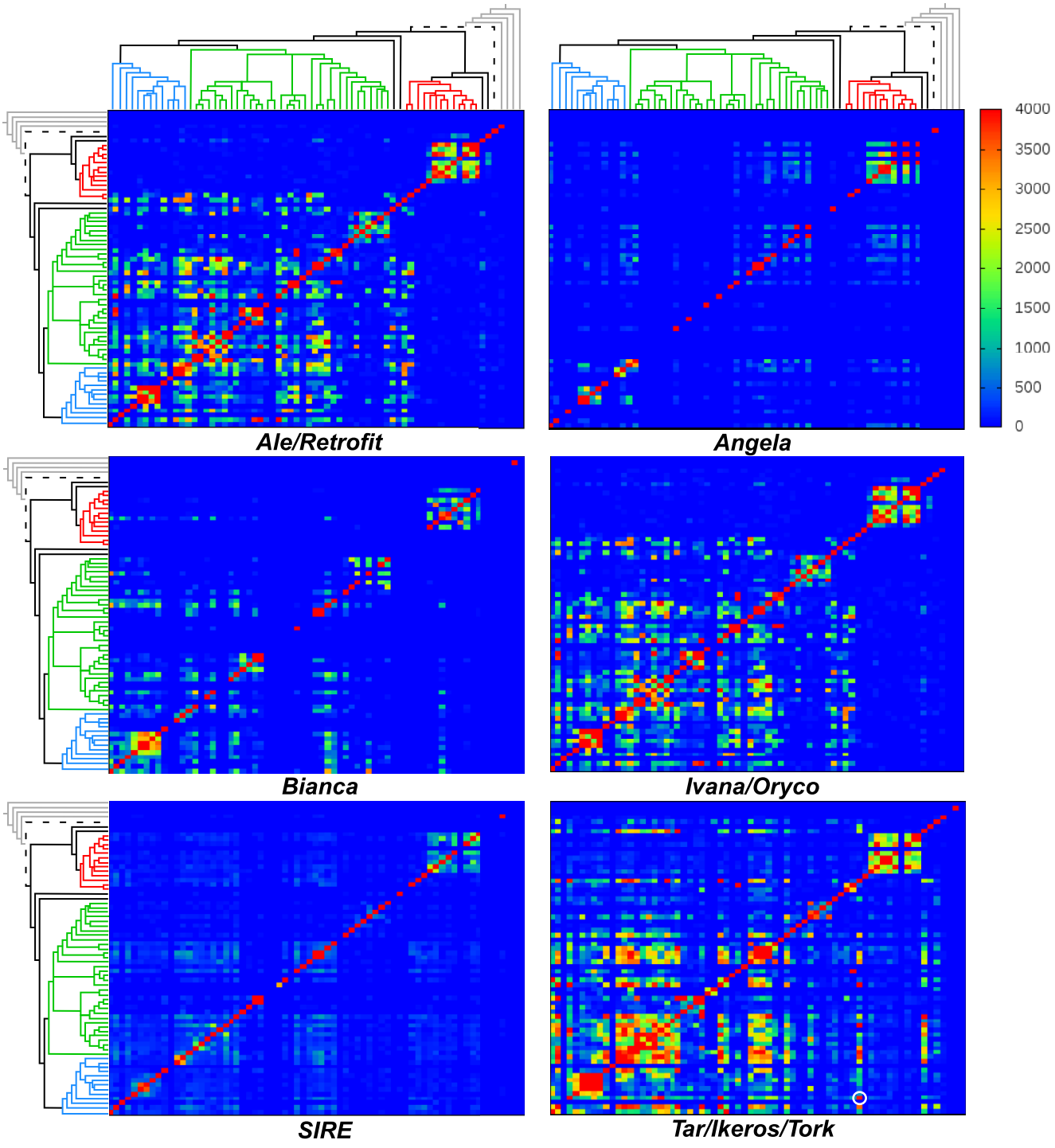
The high sequence similarities identified, and the patchy distribution above, are two of the criteria for identifying HT between species. To know whether this high similarity can be associated with potential mechanisms of HTs of LTR-RTs, other methods were applied on three different cases of high similarity identified between different plant families. The three cases were sampled according to different criteria: evolutionary distances between pairs of plant species (i.e., between monocots and dicots and between asterids and rosids), high conservation and LTR-RTs belonging to the *Tork* lineage (since this lineage is overrepresented), and relative conservation of LTR regions (Fig. S5). This small number of cases allowed us to perform a rapid detailed analysis. We first analyzed the percentage of nucleotide identity of paired elements (El Baidouri et al. 2014). We analyzed the high similarity found between orchid (*Phalaenopsis equestris*, monocots) and cannabis (*Cannabis sativa*, rosids; BLASTn score 4276), between banana (*Musa acuminata*, monocots) and soybean (*Glycine max*, rosids; BLASTn score 4414), and between poplar (*Populus trichocarpa*, rosids) and coffee (*Coffea canephora*, asteris; BLASTn score 4597). The elements were first graphically aligned to estimate the degree of similarity across the full-length elements (coding and non-coding regions), and then they were pairwise aligned using nucmer (Kurtz et al. 2004) (Fig. 4). All comparisons exhibited a high level of nucleotide identity between LTR-RTs, even in the non-coding regions (i.e., long terminal repeats). This suggests that the high scores observed in the previous analysis indeed corresponded to a very high nucleotide sequence similarity between the LTR-RTs. At the nucleotide level, orchid and cannabis elements show 75.2% of identity, while banana and soybean, and poplar and coffee exhibited 73.8% and 72.3% of shared residues, respectively. The LTRs (non-coding regions of LTR-RTs) show 50%–67% of nucleotide identity (Table S3). The similarity of the polyprotein genes (percentage of nucleotide identity) was also compared with the genome-wide sequence identity across all annotated genes for each genome pair analyzed (Fig. 4). The peak values of the identity distribution between pairs of coding genes were lower than that of the conserved LTR-RTs. In addition, synonymous distances (Ks) were calculated for all orthologous genes and for LTR-RTs between the three pairs of species (Fig. 4). The comparison of synonymous distances between genes and LTR-RTs suggests that the high similarity of LTR-RTs might be incompatible with a vertical transmission. The phylogenetic distribution of these three elements was also studied. We extracted the RT domains of all elements recovered from the LTR_STRUC

Fig. 2. Representation of the levels of similarity of the *Gypsy* lineage elements per genome across 69 plant species. Each lineage (*Athila*, *CRM*, *Tekay/Del*, *Galadriel*, *Reina*, and *TAT*) set of elements per genome is compared with the others using BLASTn. The level of similarity is represented by a heatmap of the BLASTn score (0 in blue: no similarity, to a max BLAST score of 4000 in red). Species are organized in the matrix as shown in Supplementary Figs. Only the best score from all possible pairwise comparison scores is plotted for each comparison in the matrix. The *Chlamyvir*, *Phygy*, and *Selgy* lineages are not displayed.



data set. RT classified as *Tork/Tar/Ikeros* (1235 from 58 species) were aligned and used for phylogenetic analysis (Fig. 5). Maximum likelihood tree showed phylogenetic incongruences for the three elements studied. *Tork/Tar/Ikeros* elements from orchid and cannabis (Fig. 5, A), banana and soybean (B), and poplar and coffee (C) clustered together showing a patchy distribution of these elements. Together with the phylogenetic tree incongruence between LTR-RT and species, we concluded that HT might be one of the most probable mechanisms explaining the high BLASTn scores and high nucleotide similarity of the three *Tork/Tar/Ikeros* elements studied here.

Fig. 3. Representation of the similarity levels of *Copia* lineages per genome across 69 plant species. Each lineage (*Bianca*, *Ivana/Oryco*, *Ale/Retrofit*, *SIRE*, and *Tork/Tar/Ikeros*) set of elements per genome is compared with the others using BLASTn. The level of similarity is represented by a heatmap of the BLASTn score (0 in blue: no similarity, to a max blast score of 4000). Species are organized in the matrix as shown in Fig. S1. Only the best score from all possible pairwise comparison scores is plotted for each comparison in the matrix. The *Bryco*, *Lyco*, *Gymco*, and *Osser* lineages were not displayed. The white circle and arrow on the *Tork/Tar/Ikeros* panel indicates similarities between coffee and banana (Dias et al. 2015).



Reclassification of horizontally transferred LTR-RTs

In 2014, using 40 different plant species, El Baidouri et al. (2014) found 32 putative HTTs of LTR retrotransposons, yet the horizontally transferred HT elements were not classified at the lineage level. Thus, a classification was carried out using the same methodology as implemented in this study. It was found that only three HTT events included *Gypsy* elements (*CRM*: 2, *Reina*: 1) and the remaining 29 included *Copia* elements (*Ale/Retrofit*: 14, *Bianca*: 1, *Ivana/Oryco*: 7, *Sire*:

Fig. 4. Detailed analysis of three potential cases of horizontal transfer identified in 69 sequenced plant genomes. The nuclear phylogenetic tree of the 69 plant genomes was computed using shared BUSCO genes. The cases of horizontal transfer analyzed are represented by colored lines connecting species and include the analysis of the nucleotide identity (%) of plant genes and elements, a dot-plot of conserved elements, and a graphical representation of the percentage of the identity (%) between shared elements. Comparison of synonymous distances (shown in black, log10 ks values) of genes and LTR-RTs (shown in red) between each pair of species. (A) Analysis of HT between cannabis and orchid. (B) Analysis of HT between banana and soybean. (C) Analysis of HT between poplar and coffee. (D) Relationships established in Dias et al. (2015).

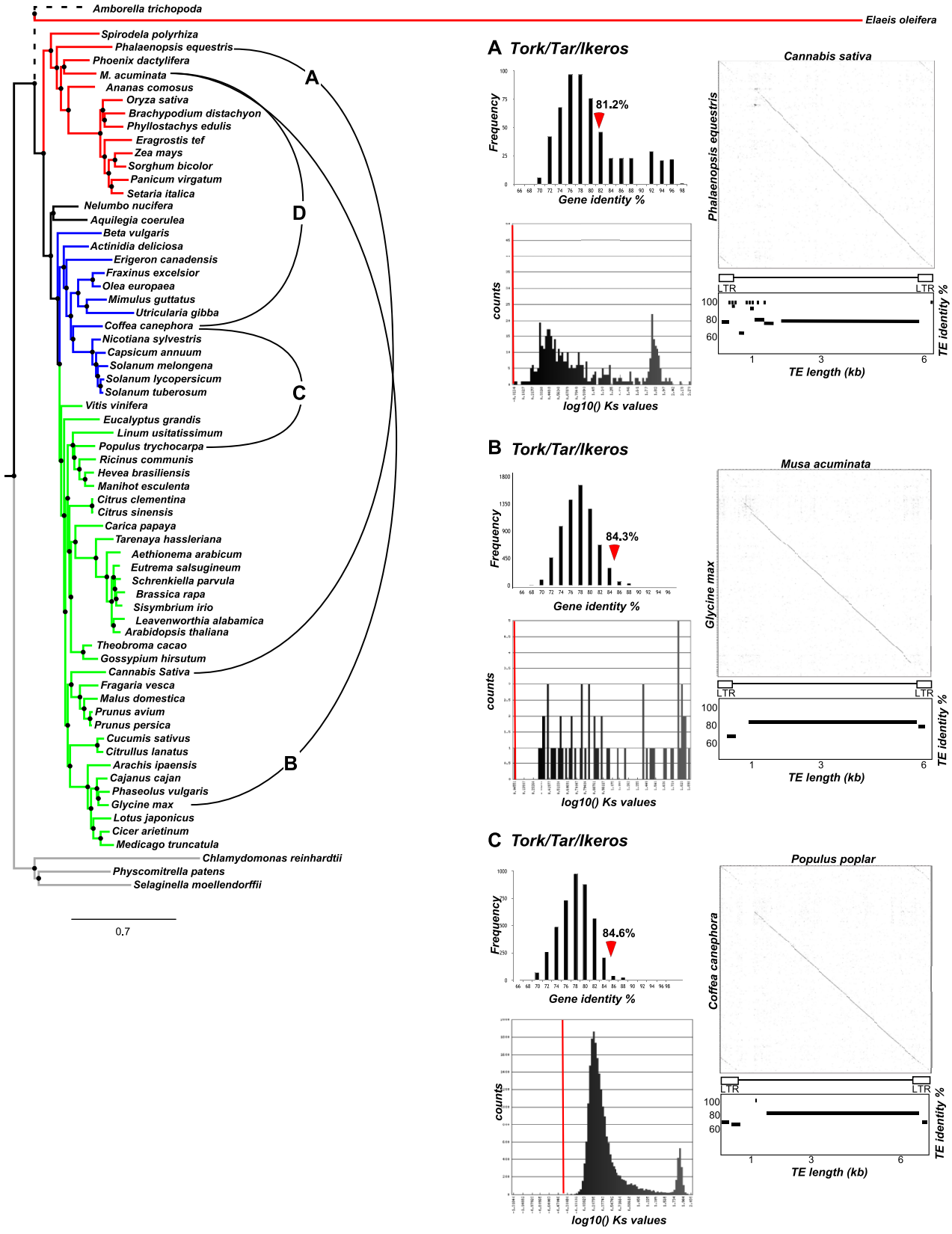
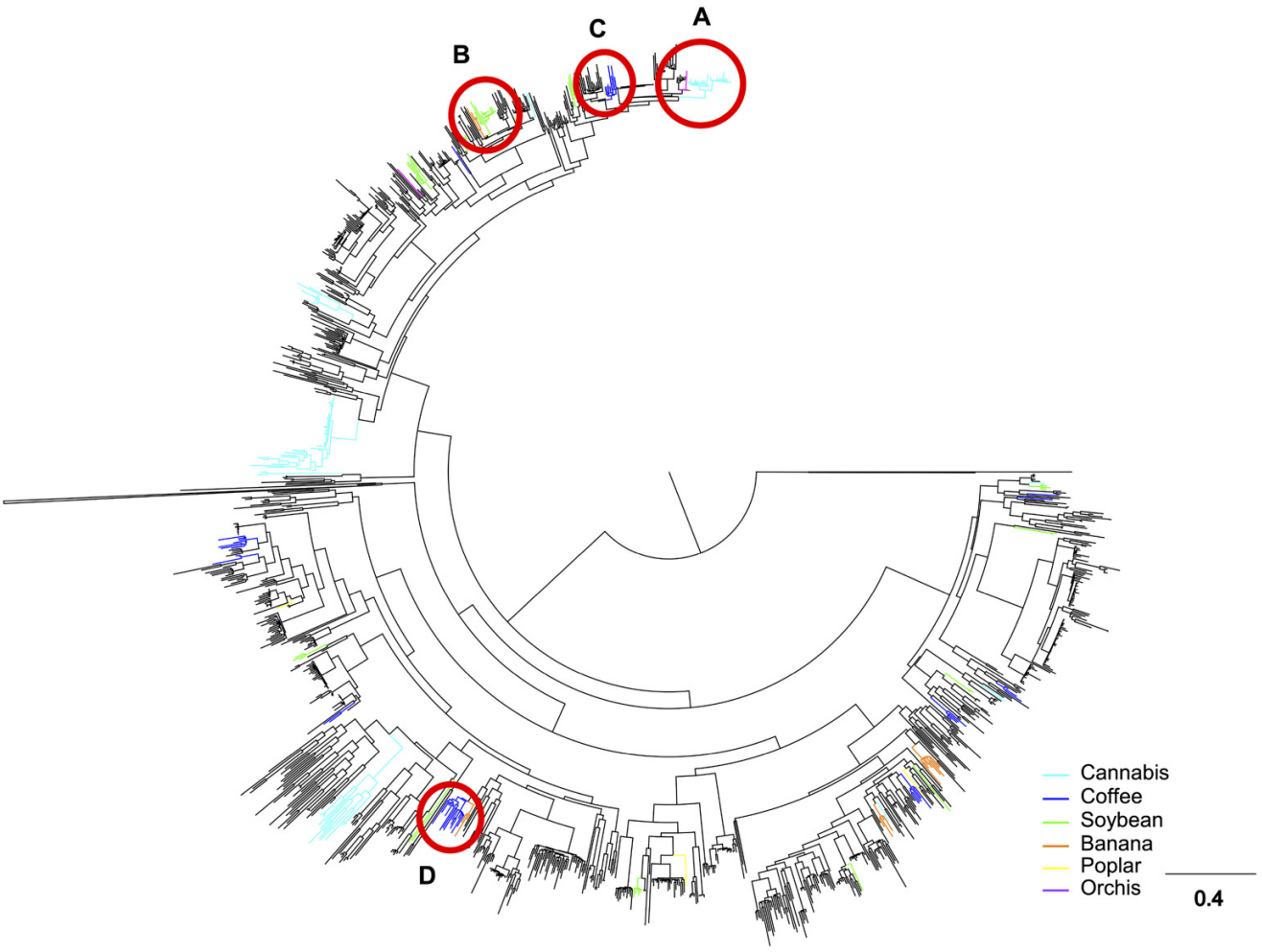


Fig. 5. Phylogenetic tree of *Tork* RT domain extracted from 69 species LTR-RT prediction. Branches corresponding to cannabis, coffee, soybean, banana, poplar, and orchid are colored. Red circles indicate studied cases in Fig. 5 (i.e., (A) cannabis and orchid, (B) soybean and banana, (C) coffee and poplar), and (D) illustrates a potential case of HT between banana and coffee (Dias et al. 2015).



1, *Tork/Tar/Ikeros*: 6). Furthermore, the three most common lineages found in the previously proposed HTT events corresponded to those found by this study (*Ale/retrofit*, *Ivana/Oryco*, and *Tork/Tar/Ikeros*) with the 84.3% of cases (i.e., 27 out of 32). Because the LTR-RT discovery methodology used by El Baidouri et al. (2014) was different from the one used in this study, we found only three of the 32 TEs involved in HTT events described by El Baidouri et al. (2014) (Table S4A). However, the main objective of this analysis was to observe whether the HTTs found by El Baidouri et al. (2014) were mostly concentrated in the same lineages described in this study or not. This shows that regardless of the strategy employed to discover LTR-RTs, the *Ale/retrofit*, *Ivana/Oryco*, and *Tork/Tar/Ikeros* lineages are much more frequently involved in interspecies conservation.

Discussion

The objective of this study was to report and understand why some superfamilies and lineages of LTR-RTs displayed high similarities, even between distantly related plant species. Of course, such exceptional similarity suggests potential cases of HTs since it is known that such transfers can happen between plants (El Baidouri et al. 2014). The strong similarity of two copies of an element between distant species is one of the required criteria to consider cases of HTs, together with the uneven ("patchy") distribution of the element across the phylogenetic tree of the species, and the phylogenetic tree incongruence between species and elements (Aubin et al. 2021).

Here, using 69 species, we show that there are many instances of very high similarity of *Copia* and *Gypsy* LTR-RTs between distantly related species, with a clear and patchy phylogenetic distribution, raising the question of whether these similarities could be considered as cases of HTs. Among the 143 high similarities observed, we analyzed in detail three of them, showing at the same time a high similarity, a patchy distribution, and a phylogenetic incongruence, which corresponds to the three required criteria to consider HT events. These three cases involved plant species belonging to different clades (eudicots/monocots and rosids/asterids). Of course, the detailed analysis of three randomly selected cases

cannot be generalized to the 143 cases of high similarity identified, and a large-scale exhaustive analysis should be undertaken. However, taking into account the re-analysis of the results of the El Baidouri study, we can hypothesize that a significant part of the conservations could be potential cases of HTT.

Alternative mechanisms to HTs must also be considered, like unequal substitution rates in TE sequences in different species (Silva and Kidwell 2000). For example, if paralogous copies of a same TE family in different lineages were vertically transmitted through speciation events, these copies could then be sampled in different species, leading to a TE phylogeny not matching the species tree (Loreto et al. 2008). This scenario being possible mostly for related species, it can be excluded for some of the potential HTs detected in this study, involving very divergent plant species. Another mechanism could be the extinction of some inactive families in several ancestors of the analyzed species (stochastic loss), explaining a patchy distribution of elements (Du et al. 2010; Wallau et al. 2012). However, this scenario can also be excluded because it does not explain the phylogenetic incongruities.

Although we have to be careful with the hypothesis of HT events for closely related plant species, the three potential HT cases studied in more detail involve highly divergent species (monocot versus dicots, and asterids versus rosids), and the high similarity between the copies of each *Copia* element, including in their non-coding parts is better explained by HT events than the latter hypotheses. Although we cannot rule out scenarios other than HT events for the set of elements found with high similarity in this study, these results provide basic evidence that may suggest that LTR-RT HTs may well be more frequent than expected in plants.

One of the key questions about HTs in plants is the identification of mechanisms or vectors able to carry TEs between different organisms. Insects, mite parasites, endosymbiotic bacteria, and viruses have been found to participate in the transfer of TEs (reviewed in Gilbert et al. 2014; Wallau et al. 2018; Aubin et al. 2021). In plants, strong suspicions are placed on viruses since they are able to infect a large range of different species, thus providing a strong argument in favor of their involvement in the transfer of host genetic material. Moreover, the ability of viruses to co-encapsulate host RNA, including LTR-RTs has been recently demonstrated (Shrestha et al. 2018), suggesting the formation of active infectious viral particles. Other carriers, such as pests like aphid or parasitic plants have been suggested, but to our knowledge, no experimental study brought evidence for specific vector-mediated HTs (Fortune et al. 2008).

Based on our analysis, *Sire* (*Copia*) and *Athila* (*Gypsy*) show a very low level of interspecies similarity, although these lineages carry an envelope-like gene (*env,* allowing a retrovirus particle to leave the host cell) (Havecker et al. 2004) and are thus considered potential endogenous retroviruses. This finding strongly suggests that the presence of an *env*-like gene is not the crucial factor involved in HTs of LTR-RTs across plant species. This is clearly different from *Drosophila*, where *Gypsy* elements are frequently found involved in HTs (Bartolomé et al. 2009), higher than *Copia* elements (Schaack et al. 2010). In particular, *Gypsy* in *Drosophila* possess an *env*-like open reading frame and have been proposed as retrovirus-like particles able to infect other cells and organisms. In plants, different mechanisms allowing HTs might operate, and the presence of *env*-like genes in some LTR-RT lineages could represent an obstacle to successful HT events, contrary to what is observed in animal genomes.

Our analysis indicates that among *Copia* and *Gypsy* lineages classified so far in plants, only three related *Copia* lineages or groups of lineages, *Ale|retrofit*, *Ivana|Oryco*, and *Tork|Tar|Ikeros* are frequently found with high similarities and patchy distribution, and so potentially involved in HTs. This observation is supported by the reclassification of the 32 LTR-RTs horizontally transferred in plants (El Baidouri et al. 2014) in which most of them (27) are *Copia* from *Ivana|Oryco*, *Ale|Retrofit* and *Tork|Tar|Ikeros*. A recent analysis in the *Vitis* genus has confirmed the transfer of 34 LTR-RTs from other species of which 30 belong to *Copia* (Park et al. 2021*b*). Moreover, a re-analysis of the recent bibliography on plant HTs indicates that *Copia* elements are clearly more frequently transferred than *Gypsy* (Roulin et al. 2007, 2009; Cheng et al. 2009; Dias et al. 2015; Huang et al. 2017; Hou et al. 2018; Aubin et al. 2021; Park et al. 2021*a*, 2021*b*). Although classification data are not homogeneous, among the 175 cases of HT found in the literature, 65% of the LTR-RTs involved in HT are classified as *Copia*. When classification at the lineage level is available, the most transferred lineages are *Tork* and *Ale|Retrofit* (Table S5B). This observation is quite unexpected and raises important questions: what are the mechanisms allowing a high similarity between elements belonging to these three lineages of *Copia*? What would be the mechanisms of HT that would specifically select/favor these lineages?

Different tracks/leads should be pursued in the future to understand why some lineages are more conserved than others or more subjected to HTs, such as their transcriptional and insertional activities and their copy number in the genomes.

Acknowledgements

The authors acknowledge the IRD itrop HPC (South Green Platform) at IRD Montpellier for providing HPC resources that contributed to the research results reported in this paper. The authors thank the Universidad Autónoma de Manizales, Manizales, Colombia for support under the project 752-115 and the LMI BIO-INCA for supporting RG. We thank Moaine El Baidouri for his valuable comments on the manuscript and Laura Gonzalez and Jorge Duitama for help in synonymous distance analyses.

Article information

History dates

Received: 25 February 2022
Accepted: 16 December 2022
Accepted manuscript online: 9 January 2023
Version of record online: 25 January 2023

Data availability
All data used in this study are public data.

Author information

Author ORCIDs
Simon Orozco-Arias https://orcid.org/0000-0001-5991-8770
David Gutiérrez-Duque https://orcid.org/0000-0001-8493-6452
Romain Guyot https://orcid.org/0000-0002-7016-7485

Author contributions
Conceptualization: SO-A, RG
Formal analysis: SO-A, MD, DG-D, RT-S, RG
Writing – original draft: SO-A, MD, RG
Writing – review and editing: SO-A, MD, RG

Competing interests
The authors declare there are no competing interests.

Funding information
SO-A is supported by a PhD grant from the Ministry of Science, Technology and Innovation (Minciencias) of Colombia, grant call 785/2017. This work was supported by Ecos-Nord N° C21MA01 and STICAMSUD 21-STIC-13.

Supplementary material

Supplementary data are available with the article at https://doi.org/10.1139/GEN-2022-0026.

References

Acuña, R., Padilla, B.E., Flórez-Ramos, C.P., Rubio, J.D., Herrera, J.C., Benavides, P., et al. 2012. Adaptive horizontal transfer of a bacterial gene to an invasive insect pest of coffee. Proc. Natl. Acad. Sci. U.S.A. **109**(11): 4197–4202. doi:10.1073/pnas.1121190109. PMID: 22371593.

Aubin, E., El Baidouri, M., and Panaud, O. 2021. Horizontal gene transfers in plants. Life, **11**(8): 857. doi:10.3390/life11080857. PMID: 34440601.

Bartolomé, C., Bello, X., and Maside, X. 2009. Widespread evidence for horizontal transfer of transposable elements across *Drosophila* genomes. Genome Biol. **10**(2): R22. doi:10.1186/gb-2009-10-2-r22. PMID: 19226459.

Blumenstiel, J.P. 2019. Birth, school, work, death, and resurrection: the life stages and dynamics of transposable element proliferation. Genes, 14.

Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., et al. 2018. Ten things you should know about transposable elements. Genome Biol. 12.

Cheng, X., Zhang, D., Cheng, Z., Keller, B., and Ling, H.-Q. 2009. A new Family of Ty1-*copia*-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. Genetics, **181**(4): 1183–1193. doi:10.1534/genetics.108.099150. PMID: 19153256.

Diao, X., Freeling, M., and Lisch, D. 2006. Horizontal transfer of a plant transposon. PLoS Biol. **4**(1): 0119–0127. doi:10.1371/journal.pbio.0040005.

Dias, E.S., Hatt, C., Hamon, S., Hamon, P., Rigoreau, M., Crouzillat, D., et al. 2015. Large distribution and high sequence identity of a *Copia*-type retrotransposon in angiosperm families. Plant Mol. Biol. **89**(1–2): 83–97. doi:10.1007/s11103-015-0352-8. PMID: 26245353.

Du, J., Tian, Z., Hans, C.S., Laten, H.M., Cannon, S.B., Jackson, S.A., et al. 2010. Evolutionary conservation, diversity and specificity of LTR-retrotransposons in flowering plants: insights from genome-wide analysis and multi-specific comparison. Plant J. **63**(4): 584–598. doi:10.1111/j.1365-313X.2010.04263.x. PMID: 20525006.

El Baidouri, M., Carpentier, M.-C., Cooke, R., Gao, D., Lasserre, E., Llauro, C., et al. 2014. Widespread and frequent horizontal transfers of transposable elements in plants. Genome Res. **24**(5): 831–838. doi:10.1101/gr.164400.113. PMID: 24518071.

Fitzpatrick, D.A., Logue, M.E., Stajich, J.E., and Butler, G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol. Biol. **6**(1): 99. doi:10.1186/1471-2148-6-99. PMID: 17121679.

Fortune, P.M., Roulin, A., and Panaud, O. 2008. Horizontal transfer of transposable elements in plants. Commun. Integr. Biol. **1**(1): 74–77. doi:10.4161/cib.1.1.6328.

Gao, D., Chen, J., Chen, M., Meyers, B.C., and Jackson, S. 2012. A highly conserved, small LTR retrotransposon that preferentially targets genes in grass genomes. PLoS ONE, **7**(2): e32010. doi:10.1371/journal.pone.0032010. PMID: 22359654.

Gilbert, C., and Feschotte, C. 2018. Horizontal acquisition of transposable elements and viral sequences: patterns and consequences. Curr. Opin. Genet. Dev. **49**: 15–24. doi:10.1016/j.gde.2018.02.007.

Gilbert, C., Chateigner, A., Ernenwein, L., Barbe, V., Bézier, A., Herniou, E.A., and Cordaux, R. 2014. Population genomics supports baculoviruses as vectors of horizontal transfer of insect transposons. Nat. Commun. **5**(1): 3348. doi:10.1038/ncomms4348. PMID: 24556639.

Grandbastien, M.-A. 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim. Biophys. Acta, **1849**(4): 403–416. doi:10.1016/j.bbagrm.2014.07.017.

Havecker, E.R., Gao, X., and Voytas, D.F. 2004. The diversity of LTR retrotransposons. Genome Biol. **5**: 225.1–225.6. doi:10.1186/gb-2004-5-6-225.

Hou, F., Ma, B., Xin, Y., Kuang, L., and He, N. 2018. Horizontal transfers of LTR retrotransposons in seven species of Rosales. Genome, **61**(8): 587–594. doi:10.1139/gen-2017-0208. PMID: 29958091.

Huang, J., Wang, Y., Liu, W., Shen, X., Fan, Q., Jian, S., and Tang, T. 2017. *EARE-1*, a transcriptionally active Ty1/*copia*-like retrotransposon has colonized the genome of *Excoecaria agallocha* through horizontal transfer. Front. Plant Sci. **8**. doi:10.3389/fpls.2017.00045.

Katoh, K., and Standley, D.M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 9.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. 2004. Versatile and open software for comparing large genomes. Genome Biol. 9.

Li, Z.-W., Hou, X.-H., Chen, J.-F., Xu, Y.-C., Wu, Q., and Gonzalez, J. 2018. Transposable elements contribute to the adaptation of *Arabidopsis thaliana*. Genome Biol. Evol. 11.

Lin, X., Faridi, N., and Casola, C. 2016. An ancient transkingdom horizontal transfer of *Penelope*-like retroelements from arthropods to conifers. Genome Biol. Evol. **8**(4): 1252–1266. doi:10.1093/gbe/evw076. PMID: 27190138.

Lloréns, C., Futami, R., Bezemer, D., and Moya, A. 2008. The Gypsy Database (GyDB) of mobile genetic elements. Nucleic Acids Res. **36**: 38–46. doi:10.1093/nar/gkm697.

Llorens, C., Muñoz-Pomer, A., Bernad, L., Botella, H., and Moya, A. 2009. Network dynamics of eukaryotic LTR retroelements beyond phylogenetic trees. Biol. Direct, **4**: 41. doi:10.1186/1745-6150-4-41. PMID: 19883502.

Loreto, E., Carareto, C., and Capy, P. 2008. Revisiting horizontal transfer of transposable elements in *Drosophila*. Heredity, **100**(6): 545–554. doi:10.1038/sj.hdy.6801094. PMID: 18431403.

McCarthy, E.M., and McDonald, J.F. 2003. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics, **19**(3): 362–367. doi:10.1093/bioinformatics/btf878. PMID: 12584121.

McGowan, J., O'Hanlon, R., Owens, R.A., and Fitzpatrick, D.A. 2020. Comparative genomic and proteomic analyses of three widespread phytophthora species: *Phytophthora chlamydospora*, *Phytophthora gonapodyides* and *Phytophthora pseudosyringae*. Microorganisms, 31.

Metzger, M.J., Paynter, A.N., Siddall, M.E., and Goff, S.P. 2018. Horizontal transfer of retrotransposons between bivalves and other aquatic species of multiple phyla. Proc. Natl. Acad. Sci. U.S.A. **115**(18). doi:10.1073/pnas.1717227115.

Ming, R., VanBuren, R., Wai, C.M., Tang, H., Schatz, M.C., Bowers, J.E., et al. 2015. The pineapple genome and the evolution of CAM photosynthesis. Nat. Genet. **47**(12): 1435–1442. doi:10.1038/ng.3435. PMID: 26523774.

Neumann, P., Novák, P., Hoštáková, N., and Macas, J. 2019. Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polyprotein domains and provides a reference for element classification. Mob. DNA, **10**(1): 1. doi:10.1186/s13100-018-0144-1. PMID: 30622655.

Novikova, O., Smyshlyaev, G., and Blinov, A. 2010. Evolutionary genomics revealed interkingdom distribution of Tcn1-like chromodomain-containing Gypsy LTR retrotransposons among fungi and plants. BMC Genomics, **11**(1): 231. doi:10.1186/1471-2164-11-231. PMID: 20377908.

Orozco-Arias, S., Liu, J., Tabares-Soto, R., Ceballos, D., Silva Domingues, D., Garavito, A., et al. 2018. Inpactor, integrated and parallel analyzer and classifier of LTR retrotransposons and its application for pineapple LTR retrotransposons diversity and dynamics. Biology, **7**(2): 32. doi:10.3390/biology7020032. PMID: 29799487.

Pace, J.K., II, Gilbert, C., Clark, M.S., and Feschotte, C. 2008. Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. Proc. Natl. Acad. Sci. U.S.A. **105**(44): 17023–17028. doi:10.1073/pnas.0806548105. PMID: 18936483.

Panaud, O. 2016. Horizontal transfers of transposable elements in eukaryotes : the flying genes. C. R. Biol. 13–16. doi:10.1016/j.crvi.2016.04.013. PMID: 27939232.

Park, M., Christin, P.-A., and Bennetzen, J.L. 2021a. Sample sequence analysis uncovers recurrent horizontal transfers of transposable elements among grasses. Mol. Biol. Evol. **38**: 12.

Park, M., Sarkhosh, A., Tsolova, V., and El-Sharkawy, I. 2021b. Horizontal transfer of LTR retrotransposons contributes to the genome diversity of *Vitis*. Int. J. Mol. Sci. **22**(19): 10446. doi:10.3390/ijms221910446. PMID: 34638784.

Peccoud, J., Loiseau, V., Cordaux, R., and Gilbert, C. 2017. Massive horizontal transfer of transposable elements in insects. Proc. Natl. Acad. Sci. U.S.A. **114**(18): 4721–4726. doi:10.1073/pnas.1621178114. PMID: 28416702.

Price, M.N., Dehal, P.S., and Arkin, A.P. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE, **5**(3): e9490. doi:10.1371/journal.pone.0009490. PMID: 20224823.

Rice, D.W., Alverson, A.J., Richardson, A.O., Young, G.J., Sanchez-Puerta, M.V., Munzinger, J., et al. 2013. Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science, **342**(6165): 1468–1473. doi:10.1126/science.1246275. PMID: 24357311.

Rice, P., Ian, L., and Bleasby, A. 2000. EMBOSS: the European Molecular Biology open software suite. Trends Genet. **16**(6): 276–277. doi:10.1016/S0168-9525(00)02024-2.

Roulin, A., Piegu, B., Wing, R.A., and Panaud, O. 2007. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. Plant J. **53**(6): 950–959. doi:10.1111/j.1365-313X.2007.03388.x. PMID: 18088314.

Roulin, A., Piegu, B., Wing, R.A., and Panaud, O. 2008. Evidence of multiple horizontal transfers of the long terminal repeat retrotransposon *RIRE1* within the genus *Oryza*. Plant J. **53**: 950–959. doi:10.1111/j.1365-313X.2007.03388.x. PMID: 18088314.

Roulin, A., Piegu, B., Fortune, P.M., Sabot, F., D'Hont, A., Manicacci, D., and Panaud, O. 2009. Whole genome surveys of rice, maize and sorghum reveal multiple horizontal transfers of the LTR-retrotransposon *Route66* in *Poaceae*. BMC Evol. Biol. **9**(1): 58. doi:10.1186/1471-2148-9-58. PMID: 19291296.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.-A., and Barrell, B. 2000. Artemis : sequence visualization and annotation. Bioinformatics, **16**(10): 944–945. doi:10.1093/bioinformatics/16.10.944. PMID: 11120685.

Schaack, S., Gilbert, C., and Feschotte, C. 2010. Promiscuous DNA: horizontal transfer of transposable elements and why it matters for eukaryotic evolution. Trends Ecol. Evol. **25**(9): 537–546. doi:10.1016/j.tree.2010.06.001.

Shrestha, N., Weber, P.H., Burke, S.V., Wysocki, W.P., Duvall, M.R., and Bujarski, J.J. 2018. Next generation sequencing reveals packaging of host rnas by brome mosaic virus. Virus Res. **252**: 82–90. doi:10.1016/j.virusres.2018.05.011. PMID: 29753892.

Silva, J.C., and Kidwell, M.G. 2000. Horizontal transfer and selection in the evolution of *P* elements. Molecular Biol. Evol. **17**(10): 1542–1557. doi:10.1093/oxfordjournals.molbev.a026253. PMID: 11018160.

Slotkin, R.K., and Martienssen, R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat. Rev. Genet. **8**(4): 272–285. doi:10.1038/nrg2072. PMID: 17363976.

Sonnhammer, E.L.L., and Durbin, R. 1996. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene, **167**: 1–10.

Suh, A., Witt, C.C., Menger, J., Sadanandan, K.R., Podsiadlowski, L., Gerth, M., et al. 2016. Ancient horizontal transfers of retrotransposons between birds and ancestors of human pathogenic nematodes. Nat. Commun. 9.

Tello, D., Gonzalez-Garcia, L.N., Gomez, J., Zuluaga-Monares, J.C., Garcia, R., Angel, R., et al. 2022. NGSEP 4: efficient and accurate identification of orthogroups and whole-genome alignment. Bioinformatics. doi:10.1101/2022.01.27.478091. PMID: 36284273.

Wallau, G.L., Ortiz, M.F., and Loreto, E. 2012. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. Genome Biol. Evol. **4**(8): 801–811. doi:10.1093/gbe/evs055.

Wallau, G.L., Vieira, C., and Loreto, É.L.S. 2018. Genetic exchange in eukaryotes through horizontal transfer: connected by the mobilome. Mob. DNA, **9**(1): 6. doi:10.1186/s13100-018-0112-9. PMID: 29422954.

Wang, H., Sun, S., Ge, W., Zhao, L., Hou, B., Wang, K., et al. 2020. Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science, **368**(6493): eaba5435. doi:10.1126/science.aba5435. PMID: 32273397.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J.L., Capy, P., Chalhoub, B., et al. 2007. A unified classification system for eukaryotic transposable elements. Nat. Rev. Genet. **8**(12): 973–982. doi:10.1038/nrg2165. PMID: 17984973.