agronomy

MDPI

Article

Composition and Diversity of LTR Retrotransposons in the Coffee Leaf Rust Genome (*Hemileia vastatrix*)

Simon Orozco-Arias [1,2,*], Mariana S. Candamil [1,†], Paula A. Jaimes [1,†], Marco Cristancho [3], Reinel Tabares-Soto [4] and Romain Guyot [4,5,*]

1 Department of Computer Science, Universidad Autónoma de Manizales, Manizales 170001, Colombia; mariana.candamilc@autonoma.edu.co (M.S.C.); paula.jaimesb@autonoma.edu.co (P.A.J.)
2 Department of Systems and Informatics, Universidad de Caldas, Manizales 170004, Colombia
3 Vice-Presidency of Research and Creation, Universidad de los Andes, Bogotá 111711, Colombia; ma.cristancho29@uniandes.edu.co
4 Department of Electronics and Automation, Universidad Autónoma de Manizales, Manizales 170001, Colombia; rtabares@autonoma.edu.co
5 Institut de Recherche pour le Développement (IRD), CIRAD, Université de Montpellier, 34394 Montpellier, France
* Correspondence: simon.orozco.arias@gmail.com (S.O.-A.); romain.guyot@ird.fr (R.G.)
† These authors contributed equally to this work.

Abstract: Coffee leaf rust is the most damaging disease for coffee cultivation around the world. It is caused by a fungal pathogen, *Hemileia vastatrix* (*Hva*), belonging to the phylum Basidiomycota. Coffee leaf rust causes significant yield losses and increases costs related to its control, with evaluated losses of USD 1–2 billion annually. It attacks both the cultivated coffee species *Coffea canephora* (Robusta coffee) and *Coffea arabica* (Arabica coffee). New races, or pathotypes, are constantly emerging with increased virulence, suggesting a rapid evolution of the pathogen. Previous genetic and genomic studies have indicated a limited nucleotide diversity of *Hva* despite a high genetic diversity and large genome size estimated to be ~800 Mb, with a high content of repeated sequences (>74%). Despite several genomic resources and the release of a recent partial genome sequence, the diversity of these repeated sequences and how they may impact the evolution of the leaf rust genome have not been investigated in detail. In an attempt to characterize the transposable elements within the *Hva* genomes, we report here new lineages of long terminal repeat (LTR) retrotransposons, called CO-HUI, Soroa, and Baco, which are classified into Gypsy, and and Labe and Mapi, which are classified as Copia. The CO-HUI and Soroa elements represent the main part of all *Hva* transposable elements, as well as approximately 37% of the available genome assemblies. Mapi and CO-HUI are the main expressed families in RNA-seq data. Although Soroa is the lineage showing more insertions into exons and genes, Mapi seems to be more frequently involved in co-expression with genes. All these new families are also present in the Pucciniales, suggesting that they dynamically participate in their genome evolution.

Keywords: *Hva*; transposable elements; new lineages; genome size; Pucciniales; PucciDB

Citation: Orozco-Arias, S.; Candamil, M.S.; Jaimes, P.A.; Cristancho, M.; Tabares-Soto, R.; Guyot, R. Composition and Diversity of LTR Retrotransposons in the Coffee Leaf Rust Genome (*Hemileia vastatrix*). *Agronomy* **2022**, *12*, 1665. https://doi.org/10.3390/agronomy12071665

Academic Editor: Maria Céu Lavado da Silva

Received: 15 March 2022
Accepted: 10 May 2022
Published: 13 July 2022

1. Introduction

Coffee is one of the most traded commodities worldwide and has an important economic impact on growing regions globally. Approximately 70% of its production is destined to international trade (USDA, https://apps.fas.usda.gov/ accessed at 12 December 2021). Nearly 70 countries are coffee producers, with Brazil as the largest producer and exporter and Colombia as the third-largest producer (USDA, https://apps.fas.usda.gov/ accessed at 12 December 2021). In Colombia, the coffee sector has had an essential role in the economy because of its impact on more than 560,000 coffee growers and their families [1]. The most common pathogen affecting coffee crops worldwide is leaf rust, which is caused by

Hemileia vastatrix (*Hva*) [2]. In the absence of control methods, the pathogen has been reported to cause 35% of the losses in the susceptible varieties of *Coffea arabica* during epidemics [3–5]. It was responsible for the collapse of coffee production in India and Ceylon by the mid-19th century [6]. Over 55 *Hva* races have been reported throughout the world [4,7], and they have been shown to have a great genetic diversity using RAPD and AFLP [8,9] and sequences of internal transcribed sequences (ITS) [10].

Genomic studies of plant pathogens have provided insights into their evolution, the mechanisms that generate genetic variability, and the repertoire of genes involved in pathogenesis. These studies have also shown that rust fungi exhibit large genome size variations, for example, *Melampsora lini* [11] has a genome size of 189 Mb fungal compared to other fungi, such as *Ustilago maydis,* with a genome size of ~20 Mb [12]. Other non-rust Pucciniomycotina fungi such as *Mixia osmundae* [13], or others of the order Pucciniales, show huge genomes with an average size of 305.5 Mb [14].

Rust (Pucciniales) genomes are particularly difficult to sequence and assemble due to their repetitive content and large genome size. Here, the transposable elements represent up to 50% of the genome [15]. As a consequence, although there are studies at the RAD sequencing level [16,17], to date, only 19 genome assemblies (at the scaffold level (https://www.ncbi.nlm.nih.gov/datasets/genomes/?taxon=5258&utm_source=data-hub, accessed at 19 January 2022)) from these plant pathogens have been reported. These genome assemblies include: *Melampsora larici-populina* (101 Mb), *Puccinia graminis* (89 Mb) [18], *Puccinia triticina* (~106 Mb) [19], *Puccinia striiformis* (~88 Mb) [20], and *Austropuccinia psidii* (1.2 Gb), the largest assembled fungal genome to date [21]. For coffee leaf rust, there are currently two draft genome assemblies, one with a length of 201 Mb, released in 2014 [22], and the other with an assembly length of 543 Mb, as published in 2019 [23].

The first *Hva* genome was sequenced using pyrosequencing (454 Roche) and showed that the genome was larger than 700 Mb [24,25]. Unfortunately, this genome sequence was very fragmented and unsuitable for the development of molecular markers with the objective to study the diversity of the fungus. Some of the reasons for not obtaining the highest quality genome were the short-read technologies available at the time, the cost of these technologies, and the fact that the genome became much larger than expected for this species. More recently, using short-reads from Illumina and long-reads PacBio, a new *Hva* genome assembly was released, and it showed that approximately 80% of the genome is composed of repetitive sequences, of which 43.6% are transposable elements (TEs) [23]. However, the precise annotation and classification of TEs at the lineage level and their dynamism were not studied.

TEs are the dynamic part of a genome. They can move within and between genomes, increase their copy number, and, consequently, increase the size of their host's genome. They can also induce mutations and changes in gene expression [26]. TEs are classified according to various parameters, one of the most important of which is the mechanism of transposition [27,28], where they are classified into Class I or "retrotransposons" and Class II or "transposons". Retrotransposons transpose by a "copy and paste" process, making use of an RNA intermediate that is then transcribed to cDNA (complementary DNA) to be introduced into the genome [29]. Transposons, on the other hand, move by a cut-and-paste mechanism, where their intermediary is DNA [30]. Notably, Class I TEs contribute significantly to genome size, while Class II TEs contribute more to allele diversity [31]. TEs are found in abundance in eukaryotic species, being up to 80% in plant species, including wheat, barley, and maize [32–34], and also ~42% of the coffee Robusta genome [35]. TEs are able to generate a significant impact on function and structure from genetic disruption and genomic restructuring, which might cause the translocation, duplication, or deletion of some fractions of the genome [36], and adaptation, evolution [26], and centromere composition in plants [37].

In fungi, genome plasticity is evidenced by whole genome duplication, hybridization [38], and, sometimes, chromosomal rearrangements and mutations caused by TEs. These mutations can, in turn, generate new genes through the modification of exons and

the recruitment, or 'domestication', of TE genes, which can produce specific functions in the host [39]. On the other hand, the insertions of these elements in the existing coding genes can alter their function, and, therefore, the effector proteins [40] secreted by pathogenic fungi to stimulate plant infection. In the same way, it has been observed that the TEs are inserted close to the pathogen factors in genomes [39,41], as is the case of *M. oryzae* where the TEs are located at a distance of 1 kb from these coding genes [42]. The mutations generated by the insertions of the TEs can produce new pathogenic variants, which can invade host plants that previously would have acquired resistance, and therefore expand their range of hosts [41].

TEs can be associated with the adaptive evolution of fungi via recombination processes, leading to genomic diversification [41]. Since TEs are present in large quantities, they could be used as a bridge between host–pathogen relationships. The host genome could use TE-encoded genes by molecular recruitment or domestication processes, leading to the creation of new TE-derived genes, and thus new functions [43]. The coevolution generated between the pathogen and the host is influenced by the adaptation produced by the TEs, which guarantees the survival of the phytopathogen. However, this compartment drives the adaptive evolution of effector genes linked to pathogenicity [44]. Therefore, it is estimated that the influence of TEs on the adaptation of fungi to the host and the environment implies that a more dynamic pathogen is produced that is resistant to external changes related to the host and chemical controls such as fungicides [41].

In an attempt to characterize transposable elements within the *Hva* genomes, we report here new lineages of long terminal repeat (LTR) retrotransposons, called CO-HUI, Soroa, Baco, which are classified into Gypsy, and Labe and Mapi, which are classified as Copia. The CO-HUI and Soroa elements represent the main part of all *Hva* transposable elements, as well as a significant part of the available genome assemblies (up to 37%). While Mapi and CO-HUI are the main families expressed in the RNA-seq data, Soroa is the family with the most insertions in genes and exons.

2. Materials and Methods

2.1. Genomic Data Used

In this analysis, two *Hva* assemblies available at NCBI were used. The first one, named by authors as HvCat (*Hva* of *Coffea arabica* var. Caturra), as published in 2014: GenBank assembly accession: GCA_003057935.1 [25], has a total length of 201 Mb, with a scaffold N50 = 3621 Kb, and was obtained by the 454-pyrosequencing technology (SRA accession number of raw reads: SRR833265). The second assembly, as reported in 2019: GenBank assembly accession: GCA_004125335.1 [23]), was named by authors as Hv33, and has a total length of 543 Mb, with a scaffold N50 = 9937 kb, and was obtained by the PacBio and Illumina HiSeq sequencing technologies. RNA-seq data (Illumina HiSeq 2000, San Diego, CA, USA) published by Cristancho and coworkers [25] (SRA: SRR1124793) was used to analyze the expression of LTR Retrotransposons (LTR-RTs). The RNA-seq data was obtained from infected coffee leaves with no evidence of the presence of the hyperparasitic fungus *Lecanicillium lecanii*.

2.2. Identification and Annotation of TEs

For the TE library creation, different software following distinct approaches, such as de novo, homology-based, and structure-based, were used. For detection, the Hv33 assembly was analyzed using REPET [45], RepeatModeler version 2 (http://www.repeatmasker.org/RepeatModeler/, accessed at 30 January 2022), LTR_STRUC [46], EDTA [47], and LTR_finder [48]. Then, the detected LTR-RTs were classified using Inpactor version 1 [49] using the domain sequences from the Gypsy database [50] as references. PASTEC [51] was used to classify the other types of TEs. Next, a clustering was made using CD-HIT [52], with a minimum of 95% of nucleotide identity and a maximum length of 50,000 bp (parameters -c 0.95 -M 50000) to remove redundancy.

Because of the high number of unclassified elements, a phylogenetic analysis was carried out using Inpactor's step 4 (which performed a multiple alignment and created the phylogenetic tree through MAFFT software [53]) with RT domains (with more than 200 amino acids). The phylogenetic analysis was confirmed using all domains found in the genome, and those trees were edited using Figtree (http://tree.bio.ed.ac.uk/software/figtree/, accessed at 24 February 2022). Then, a lineage-level classification of LTR-RTs was proposed based on the phylogenetic tree and each unknown lineage was named. One representative sequence for each lineage or family was extracted, which contained the maximum number of domains, and an alignment was made against the NCBI non-redundant database to search if these elements had been reported previously. Finally, TEs were annotated using RepeatMasker (http://repeatmasker.org, accessed at 30 January 2022) in both assemblies, and the raw 454 reads of the HvCat assembly were mapped with the TE library using Magicblast [54] with the parameters -s 500 –pi 80, with each assembly used against the same library to compare the results obtained by the raw reads. Then, the fragment copies identified for the RT domains were counted to calculate the contribution of each LTR-RT lineage to the genome size.

2.3. Structural Characterization of the LTR-RT Lineages

After annotating the TEs in the genome and searching those elements in the NCBI databases (https://www.ncbi.nlm.nih.gov, accessed at 9 October 2021), a characterization of the complete copies of the LTR-RTs was carried out. Thus, a BLASTn was made with the TE library against the genome and hits which satisfied the 80-80-80 rule (80% of similarity along 80% of the sequence with a minimum of 80 bp of the reference aligned against the target sequence) were extracted [27]. These sequences were classified with the same lineage or family of the TE in the library with its match. Using the complete copies obtained above, a search for the structural features of each LTR-RT lineage was performed. First, the presence or absence of domains needed for retrotransposition was estimated. Then, the distribution of sequence sizes of each lineage was found in order to provide key elements for the understanding of these new families. Finally, the size of the LTRs of each element was obtained and the distribution was calculated for each lineage.

2.4. Expression Analysis and Relationship with Coding Sequences

Due to the important portion of the genome represented by the LTR-RTs, we were interested in searching for the interaction of those elements with the coding regions, and so ~60 million HvCat RNA-seq reads (SRA: SRR1124793; [22]) were mapped against the TE library using Bowtie2 [55]. Then, total counts were normalized using the RPKM (reads per kilobase per million) method, taking into account the average length of each element's lineage. Next, we searched for CDS (CoDing Sequence) from the Pucciniales species found in the NCBI nucleotide database, and they were used in the co-expression analysis. Mapping was carried out between the RNA-seq reads and the TE library using Bowtie2 to extract the discordant reads. Later, an alignment was made between the unmapped reads and CDS using the BLASTn. Then, the reads with a similarity percentage of lower than 95% were filtered, and, finally, the count of hits of the LTR-RT was carried out.

Lastly, the LTR retrotransposon insertions were analyzed in the Hv33 genome with the aim of finding if those elements were inside genes or exons. Thus, the Hv33 annotation was used to first extract gene sequences using seqret from EMBOSS [56]. These sequences were then aligned by the BLASTn with the TE library to filter out the genes with more than 90% sequence length and more than 90% sequence similarity. Next, bedtools intersect [57] was run using the Hv33 annotation (without the filtered genes) and the LTR-RT annotation done by RepeatMasker. The final step was to count how many intersections were found between each lineage/family of the LTR-RTs and the genomic features marked as genes or exons.

2.5. LTR-RT Analysis in Other Pucciniales Genomes

A search for complete elements in 21 available Pucciniales genomes was performed (Supplemental Materials File S1). In order to understand the evolution dynamics of the TEs presented in these genomes, a phylogenetic tree was constructed using the most conserved genes. First, BUSCO v.4.0.5 [58] was executed and the basidiomycota_odb10 database included all species and 25,343 complete genes (single copies and duplicates). Then, all gene sequences by each species were concatenated into a single sequence (obtaining one sequence per genome) and a global alignment using MAFFT v.7.313 [53] was performed (with the parameter "–maxiterate 100"), and then the tree was edited using Figtree (http://tree.bio.ed.ac.uk/software/figtree/, accessed at 24 February 2022).

Using the TE library described above, an annotation of the repetitive sequences and LTR-RTs through RepeatMasker (http://www.repeatmasker.org, accessed at 30 January 2022) was performed in order to evaluate if the TEs found in the *Hva* were species-specific or conserved with other genomes in the Pucciniales order. Using the assemblies, the LTR-RT elements were detected and classified using Inpactor [49] (with input_type=fasta). Inpactor uses LTR_finder [48] to discover elements following a structure-based method. Then, Inpactor classified the LTR-RTs detected using a database of domains. At first, the Gypsy Database was used, but due to the low-quality results obtained, the strategy implemented in Inpactor's step 3 was implemented to extract the following domains from complete-size copies of the LTR-RTs found in *Hva*: INT, RT, GAG, AP, and RNAseH [49].

2.6. Setting Up a Reference Library of Domains and Complete LTR-RT Sequences of the Pucciniales Order (PucciDB)

First, the six main domains of the LTR-RTs were extracted from the *Hva* genome, i.e., the integrase (INT), reverse transcriptase (RT), aspartic protease (AP), RNAseH, envelope (ENV), and group-specific antigen (GAG) domains. The extraction was done by a BLASTx between the Hv33 assembly and the domains presented in GyDB (with the parameters -evalue 1×10^{-4} -max_hsps 1 -max_target_seqs 1). Then, the sequences with hits were extracted by the genewise command from WISE2 and implemented in a bash script. Next, those domains in the lineages and families found in the *Hva* assembly were classified using the proposed names. Subsequently, the extraction of the domains was carried out in the other 21 species of Pucciniales, but through utilizing the domains extracted in the previous step from *Hva*. Additionally, a full-length element version of PucciDB was created. The BLASTn between the intact LTR-RTs detected in *Hva* and the 21 Pucciniales genomes was used, and the LTR-RTs in the different genomes were extracted following the 80-80-80 rule (Wicker et al., 2007). A total of 34,449 domains were extracted, in which INT, RT, AP, RNAseH, and GAG contributed, respectively, for 15,714, 14,310, 1597, 4619, and 1209. In order to be able to handle the database in a simpler way, they were assigned a specific format for the domain IDs following the structure >Domain_Specie-name_NumericIndicative#Superfamily+Lineage [27]. Additionally, for the 37,431 full-length LTR-RTs found in the complete element's version of PucciDB, the sequence IDs were formatted in order to be compatible with RepeatMasker. Then, the following structure was proposed: >Domain_Specie-name_NumericIndicative#LTR/Lineage/family.

3. Results

3.1. Identification and Annotation of Transposable Elements in the Hva Genome Assemblies

The detection of TEs was performed using the Hv33 assembly and different bioinformatic approaches to represent all the diversity of the elements. Five softwares were finally used and detected a total of 2006 elements (consensuses and individual sequences) of all classes (Table S1). From the 2006 TEs detected, 1676 were categorized as LTR-RTs (autonomous and non-autonomous putative elements). After clustering to remove redundancy, the LTR-RT library contained 1315 distinct elements. The other types of TEs (DNA transposons and LINEs) were classified using PASTEC. Other types of retrotransposons (such as SINEs, DIRS, and PLEs) were not found in the assembly. In total, the TE library

contained 1264 elements, as follows: 91.7% (1159) were LTR-RTs, 7.3% (92) were DNA transposons, and LINEs represented ~1% (13).

The LTR-RTs were subclassified into superfamilies and lineages using Inpactor, but only the V_Clade lineage was identified, with 51 elements using the Gypsy Database as the domain references [50]. Therefore, to classify the LTR-RTs more efficiently, a phylogenetic analysis of all the LTR-RTs presented in the TE library was performed using their RT domains as a reference (data not shown). No clade seemed to be close to the RT reference domains of the Gypsy database, except for the V_Clade lineage. Then, these found clades were renamed as Soroa, Baco, CO-HUI, Mapi, and Labe. Soroa, Baco, CO-HUI, and V_Clade belong to the Gypsy superfamily, while Mapi and Labe belong to the Copia superfamily. To investigate these new groups, we extracted all the RT domains present in the Hv33 assembly (with a length greater than 200 amino acids) to construct an exhaustive phylogenetic tree.

The tree obtained from the alignment of the 6476 RT domains confirmed the presence of the V_Clade lineage and the five additional clades (Soroa, Baco, CO-HUI, Mapi, and Labe) (Figure 1A). An analysis of tree topology and branch length also suggested a large difference in the number of RT domains in each clade and a different evolution. In particular, two large clades (CO-HUI and Soroa) showed the presence of subgroups that need to be analyzed in detail. Phylogenetic trees were constructed for the RT domains from CO-HUI (using 2858 RT domains) and Soroa (using 1226 RT domains), respectively (Figure 1B,C). Based on the tree topologies, CO-HUI can be subclassified into four families named Andes, Betania, Amani, and Peregrino (Figure 1B), and Soroa can be sub-grouped into three families named JCO, Inspector, and Retaso (Figure 1C). One complete representative element of each new family identified from the phylogenetic trees was extracted and annotated in detail (Supplemental Material File S2). The length of the different identified elements ranged from 4931 bp (Mapi) to 12,825 bp (JCO and Soroa) (Figure 2). The length of the LTR domains ranged from 177 bp (Mapi) to 1483 bp (Betania and CO-HUI). For most of the elements representing the families, the GAG domain could not be identified. However, a chromodomain (chromatin organization modifier domain) was identified for Andes, Betania, Amani, and Peregrino (CO-HUI) (Figure 2). Among these representative elements, three are classified as non-autonomous, or inactive, with the presence of several frameshifts in coding regions (Labe, Baco, and Retaso).

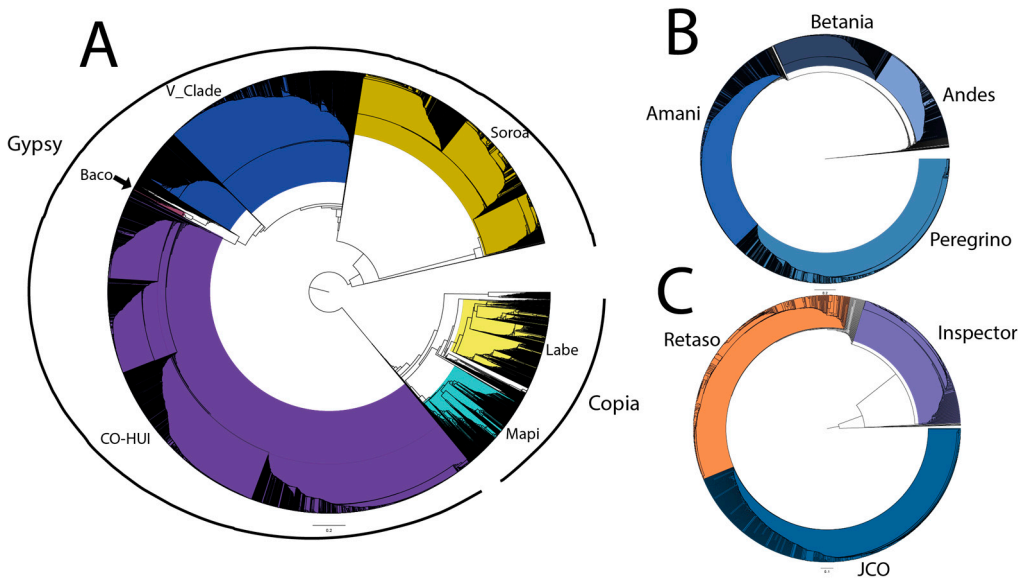

Figure 1. Phylogenetic trees of the 6476 identified RT domains (>200 amino acids) present in the Hv33 assembly. (**A**) the super families and their lineages. (**B**) The sub-families belonging to the CO-HUI lineage. (**C**) The sub-families that are present in the Soroa family.

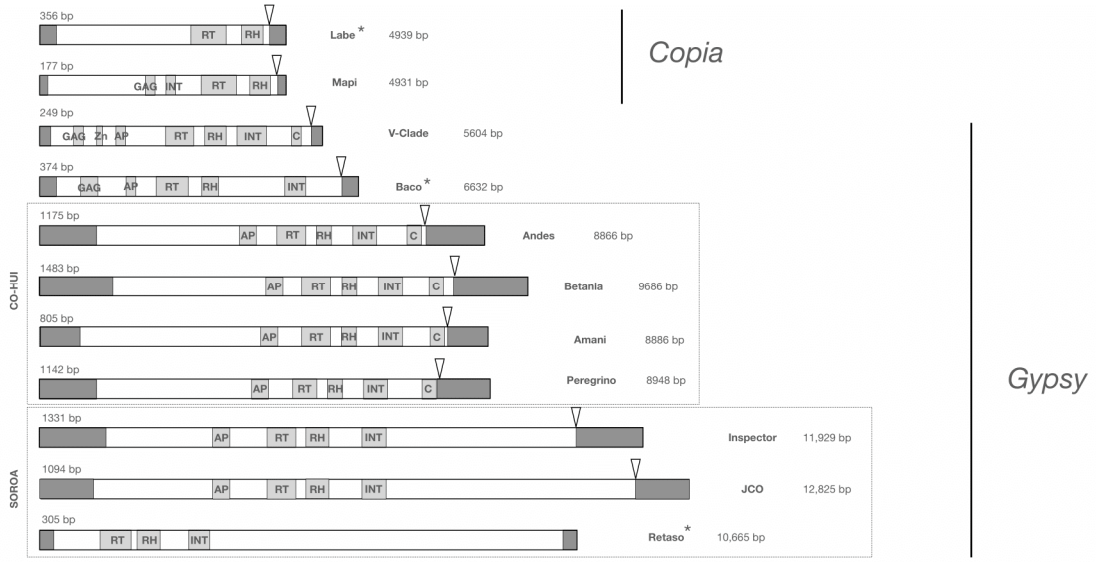

Figure 2. Structure of identified families of the LTR-retrotransposons in *Hva*. The following domains are shown: LTR (long terminal repeat), RT (reverse transcriptase), GAG (capsid), AP (aspartic protease), RH (RNAse H), INT (integrase) and C (chromodomain). The representative elements with * are considered as non-autonomous, or inactive, due to the presence of frameshifts in the coding regions.

To study the structure of the LTR-RTs in the Hv33 genome, and to compare it with the representative elements used in the Figure 2, 30,470 full copies of the LTR-RTs presented in the library were extracted using the 80-80-80 rule [27]. The elements were then used to investigate the full-size distribution of the LTR-RTs and LTR domains of the members of each lineage/family (Figure 3). The CO-HUI (with an average length of 9121 bp) families (Andes, Betania, Peregrino, and Amani) show an average size of between 8600 and 9500 bp, with Betania being the largest. Mapi, Baco, and Labe are the shortest elements, with averages of 5232 bp, 5985 bp, and 7335 bp, respectively. Soroa, at the opposite, is the largest, with an average length of 12,526 bp, and with its JCO elements being the longest. V_Clade demonstrated a very high variability in the lengths of its members. This result, which is the same as the composition shown in Figure 1A, may suggest that V_Clade is composed of more than one family. Nevertheless, the lack of contiguity of the assemblies (i.e., the low N50 values) did not all permit going deeper into the analysis [27].

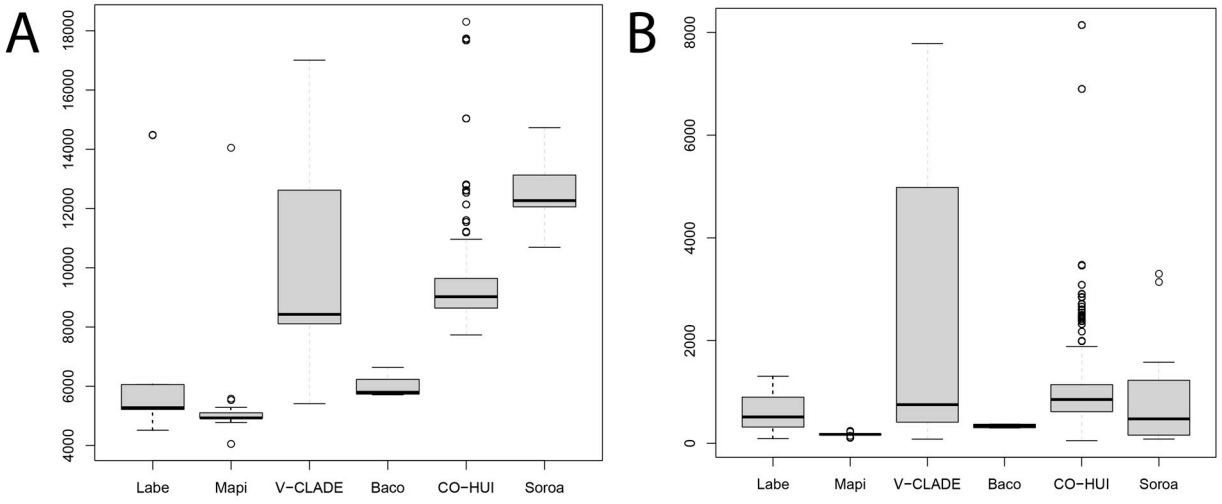

Figure 3. Length distribution of 54,125 full copies of the LTR-RTs detected following the 80-80-80 rule in the Hv33 assembly. (**A**) A full-length element sequence. (**B**) An LTR domain.

Figure 3B shows similar patterns to the elements in Figure 3A. Mapi and Baco had the shortest LTR sequences, with 166 bp and 338 bp, respectively. On the other hand, CO-HUI and Soroa have the largest LTR domains, with 939 bp and 709 bp, respectively, on average. Once again, V_Clade demonstrated a high variance. Interestingly, CO-HUI showed higher number of outliers for both full-length elements and LTR domains.

To better understand the LTR-RT structure of each lineage, the presence or absence of six domains was confirmed among the 54,125 full copies predicted, and included: integrase (INT), reverse transcriptase (RT), aspartic protease (AP), RNAseH, envelope (ENV), and group specific antigen (GAG) (Figures S1 and S2). Mapi and Labe, both from Copia, are lineages found more frequently within the GAG, INT, RNAseH, and RT domains, while the others were not identified in the GAG domains. This observation suggests that the GAG domains were not identified for the Gypsy superfamily, and so these domains might be different from the reference used in this study.

3.2. Copy Number Estimation and the Contribution of TEs to the Hva Genome Size

The copy number of the LTR-RT lineages and families in the available *Hva* genome assemblies was investigated. We first attempted to estimate the copy number of the retrotransposons in the Hv33 genome, taking into account only the presence of the RT domains for each lineage due to its high structure conservation [59]. The most numerous RTs belonged to the CO-HUI lineage in the Hv33 assembly (3504 domains). If we consider that each RT belongs to a full-length CO-HUI element, this lineage might contribute significantly to the size of the *Hva* genome (Figure S3).

The copy number of the transposable elements was also estimated using the 80-80-80 classification rule proposed by [27]. Here again, the CO-HUI lineage shows the highest copy numbers (12,380), followed by V_Clade (1803) (Figure S4A). At the CO-HUI sub-lineage level, Amani, Betania, and Peregrino show the highest copy numbers, with 4399, 2707, and 2593 copies, respectively (Figure S4B). Considering the average size of the complete elements of the CO-HUI lineage (about 9 kb), it can be estimated that, in the past, about 111 Mb of this lineage has been inserted into the *Hva* genome.

The TE library obtained with the analysis of the Hv33 assembly was used to annotate both the HvCat and Hv33 assemblies using RepeatMasker. The results indicate that the LTR-RTs represent the main part of the assemblies, with 32.64% and 46.92% for HvCat and Hv33, respectively. The LTR-RTs represent the vast majority of all the TEs identified (79.3% and 85.6% respectively for HvCat and Hv33) (Figure S5). At the TE family level, the most redundant elements are IS3EU and PIF-Harbinger for transposons, L1 for non-LTR retrotransposons, and CO-HUI (12.31% in HvCat and 23.26% in Hv33) and Soroa (~9% in HvCat and ~13.7% in Hv33) for LTR retrotransposons (Figure S6). The HvCat and Hv33 assemblies have a size of 201 Mb and 543 Mb, respectively, and only represent a fraction of the estimation of the genome size measured by flow cytometry (733.5 Mb [24], 796.80 Mb (Fungal Genome Size Database, http://www.zbi.ee/fungal-genomesize/, accessed at 19 March 2021), and 796.8 Mb [14]. This suggests that a significant part of the genome remains unassembled, potentially corresponding to repeated regions and transposable elements.

To tentatively estimate the proportion of LTR-RT retrotransposons in the *Hva* genome and not the proportion of the assemblies, raw reads were used (454 reads SRA: SRR833265, ~710,000 reads). LTR-RTs contribute to 53.3% of all 454 raw reads, which is significantly higher than the estimation of that in the assemblies. At the family level, CO-HUI and Soroa showed large variations for the presence of LTR retrotransposons according to the mapped 454 reads and the Hv33 and HvCat assemblies (Figure 4). Overall, these results strongly suggest that the contribution of TEs to genome size must be larger than estimated with current assemblies. In the future, a more complete sequenced and assembled *Hva* genome should allow a better estimation of the contribution of TEs to the genome composition. In the meantime, since the Hv33 genome seems to be more representative for the analysis of TEs, it will be used for further analysis.

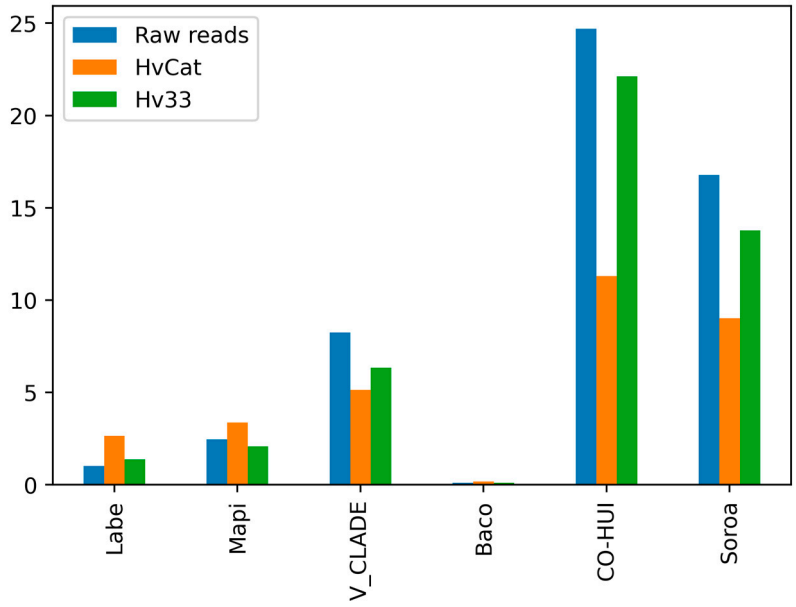

Figure 4. Comparison of the contribution of each LTR-RT lineage to the mapped 454 raw reads (blue), Hv33 assembly (green), and HvCat assembly (orange). The *y*-axis represents the percentage of genomic information belonging to each lineage.

3.3. Transcriptional Analysis of the LTR-RT Lineages and Families Using RNA-Seq, and Their Relationships with Coding Genes

More than 3025 insertions into exons and 1843 into annotated genes were identified for the Soroa LTR-RT lineage. Mapi, CO-HUI, and V_Clade showed less insertions in exons and annotated genes (Figure 5A). These results give the opportunity to test if the insertion of LTR-RTs in the vicinity of genes can generate a co-transcription of the host gene and the inserted element.

To test this hypothesis, 21,174 coding sequences (CDS) from Pucciniales were downloaded in order to perform a discordant reads mapping analysis. The paired RNA-seq reads were mapped against the *Hva* TE library. Pair-end (PE) reads with only one read mapped on the *Hva* TE library were kept. Then, the unmapped reads of these PE reads were aligned against the CDS using the BLASTn (with a nucleotide identity of >95%). The number of PE reads with both similarity to LTR-RTs and coding sequences is indicated in Figure 5B. Mapi is the LTR-RT family with more co-expressed transcripts, with 232,278 hits (75.78% of all hits), and Soroa is second, with 56,460 hits (18.42%). Finally, Baco contributed only 58 hits (0.019%) (Figure 5B).

Available RNA-seq for the *Hva* genome [25] was used to study the expression of LTR retrotransposons lineages and families. Sequences were mapped against the LTR-RT library and the hit counts were normalized using RPKM (reads per kilobase million), with the average length for each lineage/family. The normalized counts indicate a higher expression for the Mapi, CO-HUI, and V_Clade lineages (Figure 5C), while no expression was detected for Baco. Amani and Andes were the most-expressed sub-families (Figure S7).

3.4. Characterization of the Conserved LTR-RT Lineages in the Pucciniales Genomes

The lineages and families of the LTR-retrotransposons seem to contribute to different manners to the *Hva* genome structure and evolution. We studied the contribution of these elements to the evolution of the Pucciniales species. Then, 21 available Pucciniales genome assemblies (Supplemental Material File S1) were downloaded and analyzed. The TEs were identified and annotated following the same methodology used for the *Hva* genome. The percentage of LTR-RTs varies drastically, from 47% for *Hva* to 2.8% for *P. horiana*. Full-length LTR-RTs were not found in two species (*Melampsora allii-populina* and *Uromyces transversalis)*. The average percentage of LTR-RTs in these genomes was about 8%,

suggesting that the number of TEs is globally low in the Pucciniales genomes (Figure S8A). At the LTR retrotransposon level, all the lineages present in *Hva* are also identified in the Pucciniales genomes. However, large variations are observed, with a maximum genome contribution observed for both CO-HUI and Soroa in *Hva* (Figure S8B).

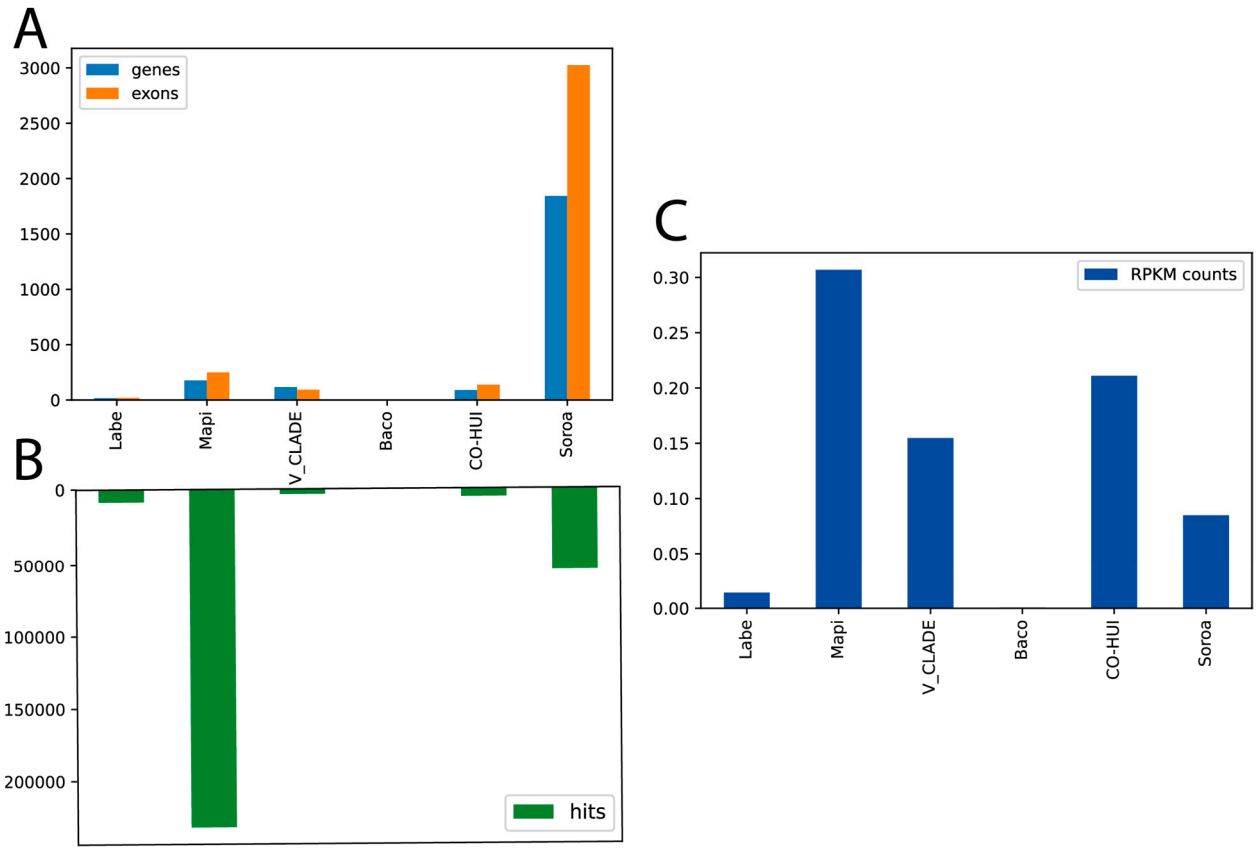

Figure 5. Number of LTR-RTs inserted into to annotated genes and exons. (**A**) The intersection between the LTR-RT elements and genes and exons in the Hv33 genome. (**B**) The co-expression analysis of Pucciniales between the annotated genes and LTR-RTs. The co-expression is represented as numbers of hits for each lineage. (**C**) RNAseq expression of the LTR-RTs. The RNAseq hits were normalized using RPKM (reads per kilobase million), which used the average length of each element's lineage.

In order to understand the relationships between the variations of LTR-RTs and the Pucciniales genome size evolution, the genome assembly length and the percentage of the LTR-RT lineages for 22 Pucciniales species were plotted in a phylogenetic context.

The tree shown in Figure 6 was constructed using concatenated BUSCO genes. Detailed information about the proportions of each LTR-RT lineage and family, as well as the non-autonomous LTR-RT elements (i.e., TRIM [60], LARD [61], and TR_GAG [62]) of the total number of elements found in the Pucciniales assemblies, can be found in Figure S9. At the phylogenetic level, several patterns of LTR-RT composition are present. The Cronartiaceae group contains significant LARD elements, while the Melampsoraceae group contains an increase in V_Clade lineage elements. The Pucciniaceae group appears to be heterogeneous in the LTR-RT compositions. Interestingly, the large genomes of *Hva* and *Apsidii* contain the largest numbers of CO-HUI lineages.

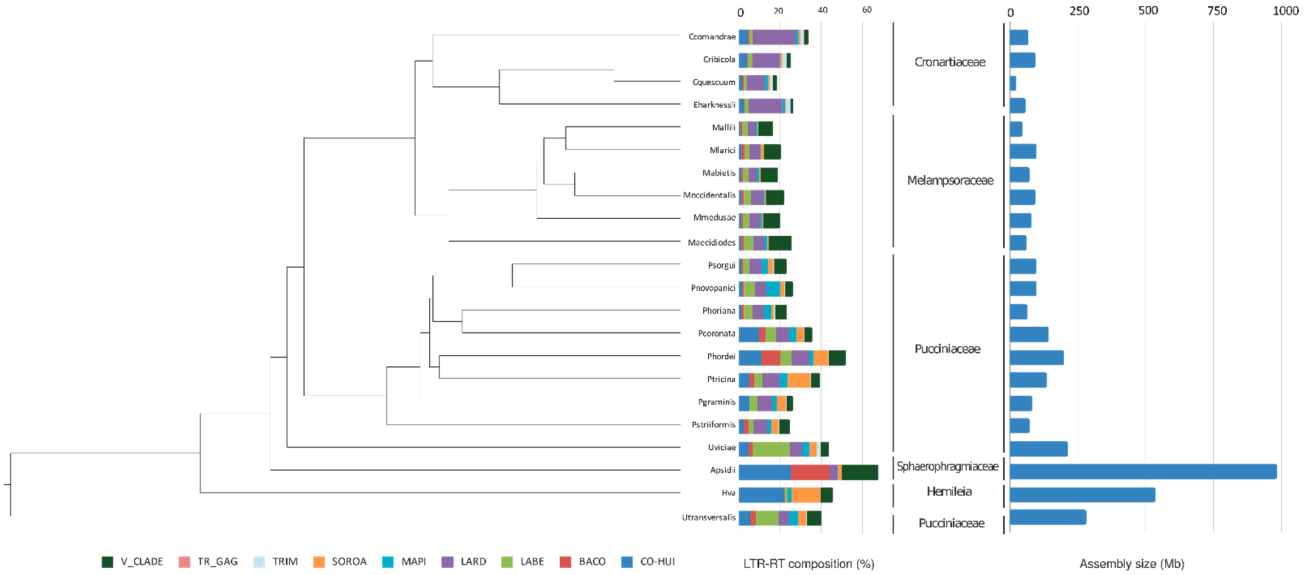

Figure 6. Contribution of LTR-RTs to the genome size of the Pucciniales species. The species were ordered according to their phylogenetic relationships using a tree constructed with BUSCO genes. The length of the bars is proportional to the percentage of LTR-RT composition (indicated as percentages and the genome assembly size) and the size for each assembly is represented in Mb.

4. Discussion

Rust species have been studied at the genomic, transcriptomic, and proteomic levels. However, due to the complexity of these obligate plant pathogen genomes [15], there are only 24 genomic assemblies reported so far in the NCBI database. These genomes are part of Pucciniales order, and include *Melampsora larici-populina* (101 Mb), *Puccinia graminis* (89 Mb) [18], *Puccinia triticina* (~106 Mb) [19], *Puccinia striiformis* (~88 Mb) [20], and *Austropuccinia psidii* (1.2 Gb) [21], among others. For coffee leaf rust, there are two draft genome assemblies, one with a length of 201 Mb, released in 2014 [22], and the other with an assembly length of 543 Mb, as published in 2019 [23].

4.1. Identification and Annotation of the Transposable Elements in Available Hva Genomes: A Challenge

The aim of this study was to understand the contribution of the transposable elements (TE) to the genome size of *Hva* using the available genomic data. The quality and contiguity of the sequenced genomes are the most important criteria to ensure the identification of a large number of full-length copies during the TE annotation process. Genome assemblies with small contigs length (called fragmented assembly) do not generally allow for the correct identification and annotation of complete TE structures. TEs such as LTR-RTs have a size ranging from 4 to 20 Kb [50]. The two assemblies used in this study show an N50 (a statistic similar to a mean or median length) of 3.6 Kb for HvCat [25] and 9.9 Kb for Hv33 [23], which represents a great limitation in the discovery of complete TEs and for the constitution of a quality reference database. Moreover, the available genomes represent only 25% and 67%, respectively, of the estimated genome size of *Hva*, clearly indicating that a significant part of the genome is missing in these available assemblies.

Nevertheless, while waiting for a better quality of genome assemblies, we have used two complementary de novo approaches, structure-based and homology-based methodologies (reviewed in [26]), to construct a first reference library of the transposable elements for *Hva*. Structure-based approaches are very sensible to the fragmentation of the genome assembly and allow the recovery of individual elements. At the opposite, the de novo homology-based approaches, as implemented by REPET [45], allow the reconstruction of the TE consensuses, and are less sensitive to the fragmentation level of the assembly. We

recovered 2006 elements in total, but of course, this reference cannot be complete because some elements may not be found in the assembled part of the available genomes, as was the case for the Pineapple genome [63], where an LTR-RT (Pusofa) was found mainly in raw reads. Here, we found that the *Hva* TEs are mainly composed of LTR-RTs (85.6%). This predominance of LTR-RTs is related to the high proportion of TEs in the *Hva* genomes. Sequencing 135 small- to medium-sized fungal genomes revealed that the proportion of LTR-RTs correlates well with the proportion of TEs in these genomes [64]. However, the high proportion of LTR_RTs does not necessarily correlate with the size of these genomes, as reported for basidiomycete [65]. The LTR-RTs were classified into superfamilies (i.e., Copia and Gypsy) and into lineages and families. Two new Copia lineages (Labe and Mapi) and three new Gypsy lineages (Soroa, Baco, and CO-HUI) were identified based on the phylogenetic analysis of their RT domains. The CO-HUI lineage, which was 8 to 11 kb size, carries a chromodomain (chromatin organization modifier domain) in the polyprotein (pol) open reading frame. Such structures have been previously detected as the dominant group of TEs in a systematic analysis of the LTR-RTs in 59 fungal genomes [66]. The presence of a chromodomain that classifies this group of elements as Chromoviridae [67] may suggest a preference for insertion into heterochromatin. The Soroa lineage contains the longest elements, which are 11 to 14 kb in size, and it has an LTR domain varying between 305 bp and 1331 bp. The elements of Mapi, Baco, and Retaso (a family of the Soroa lineage) do not seem to be complete or functional, since they present a number of mutations that create frameshifts in the polyprotein ORF. It is likely that complete copies are present in the genome, but that the fragmentation of the assemblies did not enable their identification. If our annotation identified five new LTR-RT lineages, an improved *Hva* genome with higher contiguity could allow for the identification of more new structures.

4.2. The Hva Genome Is Mainly Composed of Three LTR-RT Lineages

Here, we found that transposable elements comprised ~55% of the *Hva* draft genome (*Hv33*), which is slightly higher than the initial analysis [23]. Indeed, the LTR retrotransposons cover a large part of the assembly and represent 46.9% of the genome in total. Unfortunately, the raw (PacBio) reads of the Hv33 genome have not been made public by their authors, and the proportion of the LTR-RT elements could not be quantified using a mapping strategy of the unassembled reads. Similarly, the presence of other LTR-RT lineages in the unassembled raw reads, which were not present in the assembly, could not be tested. However, we used the 454 raw reads made available by [25] for the HvCat assembly. The 454 raw reads contained 53.3% of the LTR-RT elements compared to 31.6% for the HvCat assembly, suggesting that numerous copies of the identified LTR-RTs were unassembled, and so the proportion of TEs might be generally underestimated when based on assemblies. A conclusion was obtained for the *Amanita* fungi genomes [68]. Finally, for the Hv33 genomes, a similar dramatic variation between the assembly and the unassembled reads could be expected.

The analysis of the proportion of the different LTR-RT lineages indicates that three lineages from the Gypsy superfamily, CO-HUI, Soroa, and V_Clade, contribute to 51.7% (281.2 Mb) of the Hv33 assembly size and correspond to 93.3% of all the LTR-RTs identified.

CO-HUI full-length elements contribute more than 215 Mb (23,651 copies, with an average length of 9121 bp), V_Clade elements contribute 64.3 Mb (6584 copies, with average length of 9762 bp), and Soroa, 1.9 Mb (152 copies and an average length of 12,526 bp) to the genome of *Hva*. For CO-HUI only, 3504 RT domains were recovered from the assembly, indicating a discrepancy with the number of estimated copies. Such a discrepancy suggests that most of the elements were fragmented, partial, or non-autonomous. This can be due to the fragmentation of the assembly or to processes of unequal and illegitimate intra- or inter-element recombination [69]. The insertion time of full-length LTR-RTs can be estimated by the divergence between the two LTR domains [70]. Unfortunately, too few full-length LTR-RTs have been identified in the assemblies to conduct this analysis and to study the dynamics of the insertion of elements. However, the phylogenetic tree of the RT domains

let us hypothesize the past activity of the LTR-RTs. For the Gypsy superfamily, the CO-HUI, Soroa, and, to a lesser extent, V_Clade lineages, numerous shorter branches indicate recent insertion activity. Together with the contribution of these three lineages to the assembly length, we suggest that recent activity of CO-HUI, Soroa, and V_Clade might be at the origin of the genome size expansion of *Hva*. The propensity of retrotransposons to suddenly increase the size of their host genome is now well established, and has been highlighted by the doubled genome size of Australian rice (*Oryza australiensis*) through the activity of three LTR-RT families [71]. For fungi, a burst expansion of the Gypsy LTR-RTs was identified in a strain of *Microbotryum* [72]. However, for *Hva*, the timing of the insertions remains to be determined in order to understand if the expansions of the CO-HUI, Soroa, and V_Clade lineages were simultaneous, or if the events were successive, sudden, or progressive.

Finally, we tried to understand the impact of the LTR-RTs on the coding sequences. The CO-HUI and V_Clade lineages were not very frequently found inserted in exons or annotated genes, which are supported by the presence of a chromodomain with a preference of insertion in the heterochromatin. On the other hand, Soroa showed a high number of insertions in the coding regions. Together with a significant expression found in RNAseq data, Soroa could have a significant impact on the functions and expressions of genes. Several mutations induced by TE insertions were reported in fungi, with negative or positive effects [65,73]. However, an analysis of 625 publicly available fungal genomes [66] identified numerous overlaps between the TE remnants and coding genes, but had no correlation with gene function. Nevertheless, the transposable elements could modulate the expression of adjacent genes with a repressive effect [39]. If we can consider most of the insertion of the TEs as neutral for gene function, the specific insertion pattern of Soroa into the coding regions and its consequences should be studied in detail.

4.3. Pattern of LTR-RT Fractions in the Pucciniales Genomes

The genome size of Pucciniales varies over tenfold from 70 to 893 Mb [14], suggesting that genome size expansion is one characteristic of rust fungi. Here, the main objective was to assess the abundance of the LTR-RT lineages and families in the context of the Pucciniales's phylogenetic relationships and genome size variations. The *Hva* TE library was initially used for the Pucciniales genomes annotation, but a low proportion was recovered. Only 3.9%, 2.3%, 5.1%, and 14.4% of the *P. graminis, M. larici-populina, P. triticina*, and *A. psidii*, genomes, respectively, were identified as TEs. It has been previously established that the repeat content of these genomes is 45% for *P. graminis* and *M. larici-populina*, ~40% for *P. triticina*, and ~27% for *A. psidii* [74]. A de novo re-annotation of 21 Pucciniales genomes allows us to construct a database of annotated elements (called PucciDB) with a similar process of annotation for comparison purposes. The abundance of LTR-RT generally varies with the genome size of the species. Most of the LTR-RTs identified as belong to the Gypsy superfamily were commonly found in other fungi [66]. However, *U. viciae* (86.7%), *P. sorgui* (60%), and *P. novopanici* (56.7%) show a high proportion of Copia elements, suggesting that this superfamily is also able to increase its copy number in Pucciniales.

Interestingly, correlations were observed between the patterns of the LTR-RT lineage abundances and rust families. Species of Cronartiaceae and Melampsoraceae each have a specific LTR-RT lineage abundancy pattern, suggesting that this pattern has been conserved and transmitted in the evolution of each family. In contrast, the Pucciniaceae family shows heterogeneous patterns that correlate with the genome size variations, and the largest genomes show an expansion in the CO-HUI and Soroa lineage abundances. This finding of the expansion of some LTR-RT lineages in large genome sizes needs to be expanded to more closely related species, with the advance of sequencing in rust families [14,22].

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/agronomy12071665/s1, Table S1: Number of TEs detected by each software in the Hv33 assembly. Figure S1: Domains present in the LTR-RTs used to annotate the assemblies. (**A**) Labe, (**B**) Soroa, (**C**) Baco, (**D**) Andes, (**E**) Betania, (**F**) Peregrino, (**G**) Amani, and (**H**) Mapi. Andes, Betania, Peregrino, and Amani correspond to the CP-HUI lineage. Figure S2:

Counting the presence of each domain in the lineages identified for *Hva*. Figure S3: Number of RT domain copies in the Hv33 assembly. Figure S4: Copy numbers of the TEs identified in the Hv33 assembly. (**A**) The LTR-RTs linages (in blue) and DNA TEs superfamilies (in orange), and (**B**) the CO-HUI families. 'CO-HUI elements' here means the elements which could not be classified in one of the four families. LINEs were not shown due to their few copies (less than 40). Figure S5: Composition of DNA transposons, LTR retrotransposons, and LINEs in the HvCat and Hv33 assemblies. Figure S6: Composition of the DNA transposons (**A**), LINEs (**B**), and LTR-RTs (**C**) families in the HvCat and Hv33 assemblies. Figure S7: RNA-seq RPKM normalized counts for the LTR retrotransposon CO-HUI lineage families. Figure S8: Identification and annotation of the LTR-retrotransposons in 22 Pucciniales genomes, including *Hva* (Hv33). (**A**) The percentage of TEs for each Pucciniales assembly. (**B**) The percentage of LTR-RT lineages for each Pucciniales assembly. Species were sorted based on the phylogenetic tree shown in Figure 6. Figure S9: Proportion of each type of full-length LTR-RT found in the 19 Pucciniales genomes. The LTR-RTs from Cquescuum (*Cronartium quercuum*) could not be classified into any superfamily or lineage. RLG: Gypsy superfamily, RLC: Copia superfamily. The species were sorted by their position in the phylogenetic tree shown in Figure 6.

Author Contributions: Conceptualization, S.O.-A. and R.G.; methodology, S.O.-A., M.S.C., P.A.J. and R.T.-S.; writing—original draft, S.O.-A., M.S.C., P.A.J., M.C., R.T.-S. and R.G.; writing—review and editing, S.O.-A., M.S.C., P.A.J., M.C., R.T.-S. and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: S.O.-A. was supported by a Ph.D. grant from the Ministry of Science, Technology, and Innovation (Minciencias) of Colombia, Grant Call 785/2017. S.O.-A., R.T.-S. and R.G. were supported by the Universidad Autónoma de Manizales, Manizales, Colombia, under project Colombia 752-115, and the research was funded by the Minciencias-Ecos Nord N° C21MA01 and 285-2021 grants. R.G. was supported by the LMI BIO-INCA and M.C. was funded by the UKRI-BBSRC 'Capacity building for bioinformatics in Latin America' (CABANA) grant on behalf of the Global Challenges Research Fund [BB/P027849/1]. The funding institutions had no role in the study design, the data collection and analysis, the decision to publish, or the preparation of the manuscript.

Data Availability Statement: PucciDB can be downloaded from: https://zenodo.org/record/5655349#.YjDOaHqZNPY (accessed on 9 May 2022).

Acknowledgments: The authors acknowledge the IFB Core Cluster that is part of the National Network of Computing Resources (NNCR) of the Institut Français de Bioinformatique (https://www.france-bioinformatique.fr, accessed on 9 May 2022), the Genotoul Bioinformatics platform (http://bioinfo.genotoul.fr/, accessed on 9 May 2022), and the IRD itrop (https://bioinfo.ird.fr/, accessed on 9 May 2022) at IRD Montpellier for providing HPC resources that have contributed to the research results reported in this paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lopez, O.L.O.; Alvarez-Herrera, L.M. Tendencia de la producción y el consumo del café en Colombia. *Apunt. Cenes* **2017**, *36*, 139–166. [CrossRef]
2. Cristancho-Ardila, M.A.; Escobar-Ochoa, C.; Dickson Ocampo-Muñoz, J. Evolución de razas de *Hemileia vastatrix* en Colombia. *Cenicafé* **2007**, *58*, 340–359.
3. Rhiney, K.; Guido, Z.; Knudson, C.; Avelino, J.; Bacon, C.M.; Leclerc, G.; Aime, M.C.; Bebber, D.P. Epidemics and the future of coffee production. *Proc. Natl. Acad. Sci. USA* **2021**, *118*. [CrossRef] [PubMed]
4. Talhinhas, P.; Batista, D.; Diniz, I.; Vieira, A.; Silva, D.N.; Loureiro, A.; Tavares, S.; Pereira, A.P.; Azinheira, H.G.; Guerra-Guimarães, L.; et al. The coffee leaf rust pathogen *Hemileia vastatrix*: One and a half centuries around the tropics. *Mol. Plant Pathol.* **2017**, *18*, 1039–1051. [CrossRef] [PubMed]
5. Monaco, L.C. Consequences of the introduction of coffee rust into Brazil. *Ann. N. Y. Acad. Sci.* **1977**, *287*, 57–71. [CrossRef]
6. McCook, S. Global rust belt: *Hemileia vastatrix* and the ecological integration of world coffee production since 1850. *J. Glob. Hist.* **2006**, *1*, 177–195. [CrossRef]
7. Silva, M.D.C.; Guerra-Guimarães, L.; Diniz, I.; Loureiro, A.; Azinheira, H.; Pereira, A.P.; Tavares, S.; Batista, D.; Várzea, V. An Overview of the Mechanisms Involved in Coffee-*Hemileia vastatrix* Interactions: Plant and Pathogen Perspectives. *Agronomy* **2022**, *12*, 326. [CrossRef]
8. Gouveia, M.M.C.; Ribeiro, A.; Várzea, V.M.; Rodrigues, C.J., Jr. Genetic diversity in Hemileia vastatrix based on RAPD markers. *Mycologia* **2017**, *97*, 396–404. [CrossRef]

9. Cabral, P.G.C.; Maciel-Zambolim, E.; Oliveira, S.A.S.; Caixeta, E.; Zambolim, L. Genetic diversity and structure of Hemileia vastatrix populations on Coffea spp. *Plant Pathol.* **2016**, *65*, 196–204. [CrossRef]
10. Santana, M.F.; Zambolim, E.M.; Caixeta, E.T.; Zambolim, L. Population genetic structure of the coffee pathogen Hemileia vastatrix in Minas Gerais, Brazil. *Trop. Plant Pathol.* **2018**, *43*, 473–476. [CrossRef]
11. Nemri, A.; Saunders, D.G.O.; Anderson, C.; Upadhyaya, N.M.; Win, J.; Lawrence, G.; Jones, D.; Kamoun, S.; Ellis, J.; Dodds, P. The genome sequence and effector complement of the flax rust pathogen Melampsora lini. *Front. Plant Sci.* **2014**, *5*, 98. [CrossRef] [PubMed]
12. Schirawski, J.; Mannhaupt, G.; Münch, K.; Brefort, T.; Schipper, K.; Doehlemann, G.; Di Stasio, M.; Rössel, N.; Mendoza-Mendoza, A.; Pester, D.; et al. Pathogenicity Determinants in Smut Fungi Revealed by Genome Comparison. *Science* **2010**, *330*, 1546–1548. [CrossRef] [PubMed]
13. Toome, M.; Ohm, R.A.; Riley, R.W.; James, T.Y.; Lazarus, K.L.; Henrissat, B.; Albu, S.; Boyd, A.; Chow, J.; Clum, A.; et al. Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. *New Phytol.* **2014**, *202*, 554–564. [CrossRef] [PubMed]
14. Tavares, S.; Ramos, A.P.; Pires, A.S.; Gil Azinheira, H.; Caldeirinha, P.; Link, T.; Abranches, R.; Silva, M.D.C.; Voegele, R.T.; Loureiro, J.; et al. Genome size analyses of Pucciniales reveal the largest fungal genomes. *Front. Plant Sci.* **2014**, *5*, 422. [CrossRef]
15. Aime, M.C.; McTaggart, A.R.; Mondo, S.J.; Duplessis, S. Phylogenetics and Phylogenomics of Rust Fungi. *Adv. Genet.* **2017**, *100*, 267–307. [CrossRef]
16. Silva, D.N.; Várzea, V.; Paulo, O.S.; Batista, D. Population genomic footprints of host adaptation, introgression and recombination in coffee leaf rust. *Mol. Plant Pathol.* **2018**, *19*, 1742–1753. [CrossRef]
17. Rodrigues, A.S.; Silva, D.N.; Várzea, V.; Paulo, O.S.; Batista, D. Worldwide population structure of the coffee rust fungus *Hemileia vastatrix* is strongly shaped by local adaptation and breeding history. *Phytopathology* **2022**. [CrossRef]
18. Duplessis, S.; Cuomo, C.A.; Lin, Y.-C.; Aerts, A.; Tisserant, E.; Veneault-Fourrey, C.; Joly, D.; Hacquard, S.; Amselem, J.; Cantarel, B.L.; et al. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 9166–9171. [CrossRef]
19. Kiran, K.; Rawal, H.C.; Dubey, H.; Jaswal, R.; Devanna, B.; Gupta, D.K.; Bhardwaj, S.C.; Prasad, P.; Pal, D.; Chhuneja, P.; et al. Draft Genome of the Wheat Rust Pathogen (*Puccinia triticina)* Unravels Genome-Wide Structural Variations during Evolution. *Genome Biol. Evol.* **2016**, *8*, 2702–2721. [CrossRef]
20. Cantu, D.; Segovia, V.; MacLean, D.; Bayles, R.; Chen, X.; Kamoun, S.; Dubcovsky, J.; Saunders, D.G.; Uauy, C. Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. triticireveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genom.* **2013**, *14*, 270. [CrossRef]
21. McTaggart, A.R.; A Duong, T.; Le-Quy, V.; Shuey, L.S.; Smidt, W.; Naidoo, S.; Wingfield, M.J.; Wingfield, B.D. Chromium sequencing: The doors open for genomics of obligate plant pathogens. *BioTechniques* **2018**, *65*, 253–257. [CrossRef] [PubMed]
22. Cristancho, M.A.; Botero-Rozo, D.O.; Giraldo, W.; Tabima, J.; RiañO-Pachón, D.M.; Escobar, C.; Rozo, Y.; Rivera, L.F.; Durán, A.; Restrepo, S.; et al. Annotation of a hybrid partial genome of the coffee rust (*Hemileia vastatrix*) contributes to the gene repertoire catalog of the Pucciniales. *Front. Plant Sci.* **2014**, *5*, 1–11. [CrossRef] [PubMed]
23. Porto, B.N.; Caixeta, E.T.; Mathioni, S.M.; Vidigal, P.M.P.; Zambolim, L.; Zambolim, E.M.; Donofrio, N.; Polson, S.W.; Maia, T.; Chen, C.; et al. Genome sequencing and transcript analysis of Hemileia vastatrix reveal expression dynamics of candidate effectors dependent on host compatibility. *PLoS ONE* **2019**, *14*, e0215598. [CrossRef] [PubMed]
24. Carvalho, G.M.A.; Carvalho, C.R.; Barreto, R.W.; Evans, H.C. Coffee rust genome measured using flow cytometry: Does size matter? *Plant Pathol.* **2013**, *63*, 1022–1026. [CrossRef]
25. Cristancho, M.; Giraldo, W.; Botero, D.; Tabima, J.; Ortiz, D.; Peralta, A.; Gaitán, À.; Restrepo, S.; Riaño, D. Application of Genome Studies of Coffee Rust. *Adv. Comput. Biol.* **2014**, 133–139. [CrossRef]
26. Orozco-Arias, S.; Isaza, G.; Guyot, R. Retrotransposons in Plant Genomes: Structure, Identification, and Classification through Bioinformatics and Machine Learning. *Int. J. Mol. Sci.* **2019**, *20*, 3837. [CrossRef]
27. Wicker, T.; Sabot, F.; Hua-Van, A.; Bennetzen, J.L.; Capy, P.; Chalhoub, B.; Flavell, A.; Leroy, P.; Morgante, M.; Panaud, O.; et al. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **2007**, *8*, 973–982. [CrossRef]
28. Bourque, G.; Burns, K.H.; Gehring, M.; Gorbunova, V.; Seluanov, A.; Hammell, M.; Imbeault, M.; Izsvák, Z.; Levin, H.L.; Macfarlan, T.S.; et al. Ten things you should know about transposable elements. *Genome Biol.* **2018**, *19*, 1–12. [CrossRef]
29. Boeke, J.D.; Garfinkel, D.J.; Styles, C.A.; Fink, G.R. Ty elements transpose through an RNA intermediate. *Cell* **1985**, *40*, 491–500. [CrossRef]
30. Feschotte, C.; Pritham, E.J. DNA Transposons and the Evolution of Eukaryotic Genomes. *Annu. Rev. Genet.* **2007**, *41*, 331–368. [CrossRef]
31. Han, Y.; Qin, S.; Wessler, S.R. Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. *BMC Genom.* **2013**, *14*, 71. [CrossRef] [PubMed]
32. Choulet, F.; Alberti, A.; Theil, S.; Glover, N.; Barbe, V.; Daron, J.; Pingault, L.; Sourdille, P.; Couloux, A.; Paux, E.; et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science* **2014**, *345*, 1249721. [CrossRef] [PubMed]
33. Orozco-Arias, S.; Isaza, G.; Guyot, R.; Tabares-Soto, R. A systematic review of the application of machine learning in the detection and classification of transposable elements. *PeerJ* **2019**, *7*, e8311. [CrossRef] [PubMed]

34. Su, W.; Gu, X.; Peterson, T. TIR-Learner, a New Ensemble Method for TIR Transposable Element Annotation, Provides Evidence for Abundant New Transposable Elements in the Maize Genome. *Mol. Plant* **2019**, *12*, 447–460. [CrossRef] [PubMed]
35. Denoeud, F.; Carretero-Paulet, L.; Dereeper, A.; Droc, G.; Guyot, R.; Pietrella, M.; Zheng, C.; Alberti, A.; Anthony, F.; Aprea, G.; et al. The coffee genome provides insight into the convergent evolution of caffeine biosynthesis. *Science* **2014**, *345*, 1181–1184. [CrossRef]
36. Seidl, M.F.; Thomma, B.P. Transposable Elements Direct The Coevolution between Plants and Microbes. *Trends Genet.* **2017**, *33*, 842–851. [CrossRef]
37. Nunes, R.D.C.; Orozco-Arias, S.; Crouzillat, D.; Mueller, L.A.; Strickler, S.R.; Descombes, P.; Fournier, C.; Moine, D.; de Kochko, A.; Yuyama, P.M.; et al. Structure and Distribution of Centromeric Retrotransposons at Diploid and Allotetraploid Coffea Centromeric and Pericentromeric Regions. *Front. Plant Sci.* **2018**, *9*, 175. [CrossRef]
38. Ayala-Usma, D.A.; Cardenas, M.; Guyot, R.; Mares, M.C.D.; Bernal, A.; Munoz, A.R.; Restrepo, S. A whole genome duplication drives the genome evolution of *Phytophthora betacei*, a closely related species to *Phytophthora infestans*. *BMC Genom.* **2021**, *22*, 795. [CrossRef]
39. Castanera, R.; López-Varas, L.; Borgognone, A.; LaButti, K.; Lapidus, A.; Schmutz, J.; Grimwood, J.; Pérez, G.; Pisabarro, A.G.; Grigoriev, I.V.; et al. Transposable Elements versus the Fungal Genome: Impact on Whole-Genome Architecture and Transcriptional Profiles. *PLoS Genet.* **2016**, *12*, e1006108. [CrossRef]
40. Belyayev, A. Bursts of transposable elements as an evolutionary driving force. *J. Evol. Biol.* **2014**, *27*, 2573–2584. [CrossRef]
41. Mat Razali, N.; Cheah, B.H.; Nadarajah, K. Transposable Elements Adaptive Role in Genome Plasticity, Pathogenicity and Evolution in Fungal Phytopathogens. *Int. J. Mol. Sci.* **2019**, *20*, 3597. [CrossRef] [PubMed]
42. Bao, J.; Chen, M.; Zhong, Z.; Tang, W.; Lin, L.; Zhang, X.; Jiang, H.; Zhang, D.; Miao, C.; Tang, H.; et al. PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in *Magnaporthe oryzae*. *Mol. Plant* **2017**, *10*, 1465–1468. [CrossRef] [PubMed]
43. Hua-Van, A.; Le Rouzic, A.; Boutin, T.S.; Filée, J.; Capy, P. The struggle for life of the genome's selfish architects. *Biol. Direct* **2011**, *6*, 19. [CrossRef]
44. Plissonneau, C.; Benevenuto, J.; Mohd-Assaad, N.; Fouché, S.; Hartmann, F.E.; Croll, D. Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. *Front. Plant Sci.* **2017**, *8*, 119. [CrossRef]
45. Flutre, T.; Duprat, E.; Feuillet, C.; Quesneville, H. Considering Transposable Element Diversification in De Novo Annotation Approaches. *PLoS ONE* **2011**, *6*, e16526. [CrossRef] [PubMed]
46. McCarthy, E.M.; McDonald, J.F. LTR_STRUC: A novel search and identification program for LTR retrotransposons. *Bioinformatics* **2003**, *19*, 362–367. [CrossRef] [PubMed]
47. Ou, S.; Su, W.; Liao, Y.; Chougule, K.; Agda, J.R.A.; Hellinga, A.J.; Lugo, C.S.B.; Elliott, T.A.; Ware, D.; Peterson, T.; et al. Benchmarking transposable element annotation methods for creation of a streamlined, comprehensive pipeline. *Genome Biol.* **2019**, *20*, 275. [CrossRef]
48. Xu, Z.; Wang, H. LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **2007**, *35*, W265–W268. [CrossRef]
49. Orozco-Arias, S.; Liu, J.; Tabares-Soto, R.; Ceballos, D.; Domingues, D.S.; Garavito, A.; Ming, R.; Guyot, R. Inpactor, Integrated and Parallel Analyzer and Classifier of LTR Retrotransposons and Its Application for Pineapple LTR Retrotransposons Diversity and Dynamics. *Biology* **2018**, *7*, 32. [CrossRef]
50. Llorens, C.; Futami, R.; Covelli, L.; Dominguez-Escriba, L.; Viu, J.M.; Tamarit, D.; Aguilar-Rodríguez, J.; Vicente-Ripolles, M.; Fuster, G.; Bernet, G.P.; et al. The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Res.* **2011**, *39*, D70–D74. [CrossRef]
51. Hoede, C.; Arnoux, S.; Moissette, M.; Chaumier, T.; Inizan, O.; Jamilloux, V.; Quesneville, H. PASTEC: An Automatic Transposable Element Classification Tool. *PLoS ONE* **2014**, *9*, e91929. [CrossRef] [PubMed]
52. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* **2012**, *28*, 3150–3152. [CrossRef] [PubMed]
53. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
54. Boratyn, G.M.; Thierry-Mieg, J.; Thierry-Mieg, D.; Busby, B.; Madden, T.L. Magic-BLAST, an accurate RNA-seq aligner for long and short reads. *BMC Bioinform.* **2019**, *20*, 405. [CrossRef]
55. Langmead, B.; Trapnell, C.; Pop, M.; Salzberg, S.L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **2009**, *10*, R25. [CrossRef] [PubMed]
56. Rice, P.; Longden, I.; Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **2000**, *16*, 276–277. [CrossRef]
57. Quinlan, A.R.; Hall, I.M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **2010**, *26*, 841–842. [CrossRef]
58. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210–3212. [CrossRef]
59. Sundaravaradan, V.; Hahn, T.; Ahmad, N. Conservation of functional domains and limited heterogeneity of HIV-1 reverse transcriptase gene following vertical transmission. *Retrovirology* **2005**, *2*, 1–17. [CrossRef]

60. Witte, C.-P.; Le, Q.H.; Bureau, T.; Kumar, A. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 13778–13783. [CrossRef]
61. Kalendar, R.; Vicient, C.M.; Peleg, O.; Anamthawat-Jonsson, K.; Bolshoy, A.; Schulman, A.H. Large Retrotransposon Derivatives: Abundant, Conserved but Nonautonomous Retroelements of Barley and Related Genomes. *Genetics* **2004**, *166*, 1437–1450. [CrossRef] [PubMed]
62. Chaparro, C.; Gayraud, T.; De Souza, R.F.; Domingues, U.S.; Akaffou, S.; Vanzela, A.L.L.; De Kochko, A.; Rigoreau, M.; Crouzillat, M.; Hamon, S.; et al. Terminal-Repeat Retrotransposons with GAG Domain in Plant Genomes: A New Testimony on the Complex World of Transposable Elements. *Genome Biol. Evol.* **2015**, *7*, 493–504. [CrossRef] [PubMed]
63. Ming, R.; VanBuren, R.; Wai, C.M.; Tang, H.; Schatz, M.C.; E Bowers, J.; Lyons, E.; Wang, M.-L.; Chen, J.; Biggers, E.; et al. The pineapple genome and the evolution of CAM photosynthesis. *Nat. Genet.* **2015**, *47*, 1435–1442. [CrossRef] [PubMed]
64. Miyauchi, S.; Kiss, E.; Kuo, A.; Drula, E.; Kohler, A.; Sánchez-García, M.; Morin, E.; Andreopoulos, B.; Barry, K.W.; Bonito, G.; et al. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. *Nat. Commun.* **2020**, *11*, 5125. [CrossRef]
65. Castanera, R.; Borgognone, A.; Pisabarro, A.G.; Ramírez, L. Biology, dynamics, and applications of transposable elements in basidiomycete fungi. *Appl. Microbiol. Biotechnol.* **2017**, *101*, 1337–1350. [CrossRef]
66. Muszewska, A.; Hoffman-Sommer, M.; Grynberg, M. LTR Retrotransposons in Fungi. *PLoS ONE* **2011**, *6*, e29425. [CrossRef]
67. Marín, I.; Lloréns, C. Ty3/Gypsy Retrotransposons: Description of New Arabidopsis thaliana Elements and Evolutionary Perspectives Derived from Comparative Genomic Data. *Mol. Biol. Evol.* **2000**, *17*, 1040–1049. [CrossRef]
68. Hess, J.; Skrede, I.; Wolfe, B.E.; LaButti, K.; Ohm, R.A.; Grigoriev, I.V.; Pringle, A. Transposable Element Dynamics among Asymbiotic and Ectomycorrhizal Amanita Fungi. *Genome Biol. Evol.* **2014**, *6*, 1564–1578. [CrossRef]
69. Devos, K.M.; Brown, J.K.M.; Bennetzen, J.L. Genome Size Reduction through Illegitimate Recombination Counteracts Genome Expansion in Arabidopsis. *Genome Res.* **2002**, *12*, 1075–1079. [CrossRef]
70. SanMiguel, P.; Tikhonov, A.; Jin, Y.-K.; Motchoulskaia, N.; Zakharov, D.; Melake-Berhan, A.; Springer, P.S.; Edwards, K.J.; Lee, M.; Avramova, Z.; et al. Nested Retrotransposons in the Intergenic Regions of the Maize Genome. *Science* **1996**, *274*, 765–768. [CrossRef]
71. Piegu, B.; Guyot, R.; Picault, N.; Roulin, A.; Saniyal, A.; Kim, H.; Collura, K.; Brar, D.S.; Jackson, S.; Wing, R.A.; et al. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res.* **2006**, *16*, 1262–1269. [CrossRef] [PubMed]
72. Horns, F.; Petit, E.; Hood, M.E. Massive Expansion of Gypsy-like Retrotransposons in Microbotryum Fungi. *Genome Biol. Evol.* **2017**, *9*, 363–371. [CrossRef] [PubMed]
73. Fowler, T.J.; Mitton, M.F. Scooter, a New Active Transposon in Schizophyllum commune, Has Disrupted Two Genes Regulating Signal Transduction. *Genetics* **2000**, *156*, 1585–1594. [CrossRef] [PubMed]
74. Tan, M.-K.; Chen, Z.; Wilkins, M.R.; Collins, D.; Englezou, A. A brief overview of the size and composition of the myrtle rust genome and its taxonomic status. *Mycology* **2014**, *5*, 52–63. [CrossRef]