Massive proliferation of retrotransposons contributes to genome size expansion in species of the *Pseudocercospora* genus

Sandra-Milena González Sáyer[1,2,3], Ibonne A Garcia [3], Cristian A Traslaviña[3], Alex Z Zaccaron[4], Ioannis Stergiopoulos[4], Fabio A Aristizabal[3], Ursula Oggenfuss[1,5,*,#], Daniel Croll[1,*,#]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Neuchâtel, Switzerland
[2] Centro de Biotecnología y Genómica de Plantas (UPM-INIA/CSIC), Madrid, Spain
[3] Universidad Nacional de Colombia, Instituto de Biotecnología, Laboratorio de Caracterización Molecular, Bogotá D.C., Colombia
[4] Department of Plant Pathology, University of California Davis, One Shields Avenue, Davis, California, United States of America
[5] Department of Microbiology and Immunology, University of Minnesota Medical School, Minneapolis, Minnesota, United States of America

[*]Corresponding authors: ursula.oggenfuss@gmail.com, daniel.croll@unine.ch

[#]Jointly supervised the work

ORCID: SMGS: 0000-0002-6052-2500, IAG: 0000-0002-8834-0899, AZZ: 0000-0001-5554-2059, IS: 0000-0002-2368-6119, FAA: 0000-0002-6405-70790 UO: 0000-0001-9291-9185, DC: 0000-0002-2072-380X

ABSTRACT

Genome size expansions are common among eukaryotic lineages. Enlarged genomes can be bioenergetically demanding, and active mobile elements can trigger chromosomal rearrangements and loss of gene function. What triggers genome size expansions remains largely unexplored in many biological clades, particularly within the fungal kingdom. Activation of large transposable elements (TEs), such as long-terminal repeats (LTRs), is a common contributor. Yet the mechanisms of LTR activation remain poorly understood. Here, we focus on the fungal genus *Pseudocercospora* and closely related species with known variation in genome size. In using an assembly-free approach, we found that TE content is highly variable among species, with species-specific retrotransposon families being the main drivers of independent genome expansions. We further focussed on the two species with the most expanded genomes and reference-quality genomes, *P. fijiensis* and *P. ulei*. We found that the *P. ulei* genome is compartmentalized, with highly variable TE densities among chromosomal regions, and a striking reduction in pathogenicity-associated genes. Overall, our study indicates that species of *Pseudocercospora* originally had reduced genome sizes, and genome expansions are species-specific, driven by heterogeneous sets of TE families. Furthermore, we found that in the species with the most expanded genome, TE activity might not have ceased yet as indicated by resequencing data analysis of six strains from diverse locations in Colombia. We discuss what might have caused TE activation and subsequent proliferation in the genus, including stress conditions and host adaptation. Surveys of clades with highly dynamic genome sizes are crucial for the investigation of causal factors driving long-term TE dynamics.

INTRODUCTION

Eukaryotic species have very diverse genome sizes, covering >10,000-fold changes for haploid genome sizes (1,2). Genome size expansions occurred widely among eukaryotic lineages and may be linked to speciation (3,4). Genome size expansions mostly arise from duplication of existing sequences and are thought to have minimal impact on functional complexity, including the number of gene families or regulatory sequences (5–7). Duplication of regions range in size from whole genome duplications to aneuploidy, to large structural variation and the duplication of genes or transposable elements (TEs) (1,8–12). TEs are mobile genetic units with the ability to translocate or create copies of themselves and subsequently insert into different genomic regions (13,14). TEs are not monophyletic in their origin, and either transpose via RNA intermediates, coding their own reverse transcriptase and creating a copy, or via excision and insertion cycles (15,16). TEs are generally autonomous, and contain all genes needed for their own transposition. However, non-autonomous TEs can transpose by parasitizing autonomous TEs (17). TEs are relatively small, ranging from a few dozen base pairs (bp) in certain non-autonomous elements to several tens of kilobases (kb) for typical retrotransposons, to 700 kb for more complex and clustered TE constructs like Starships (15,18). Individual TE insertions do not increase genome size substantially. However, their ongoing proliferation led, for example, the ~300 bp *Alu* element to cover almost half of the human genome, small miniature inverted TEs (MITE) proliferation to be responsible for remarkable genome size differences among rice genomes, and retrotransposons to make up around 60 % of the genome of the fungus *Cenococcum geophilum* (19–21). Different stress conditions can be a factor to induce bursts of TE activity, leading to increased copy numbers over short evolutionary timescales. Furthermore, even silenced or non-functional TEs can impact

genome size evolution, by inducing large-scale chromosomal rearrangements, duplications or deletions via ectopic recombination (22–24).

At the species and population level, TEs are a mutational force, and the impact of most new TE insertions is deleterious. Consequently, new TE insertions are selected against, and diverse mechanisms exist to suppress TE activity. Eukaryotes use epigenic silencing to prevent TEs from becoming active (25). Epigenetic silencing includes DNA methylation and histone modifications (26,27). Silencing can be reversible, especially under stress conditions (28). In addition, many ascomycete fungi have a defense mechanism against repeats called repeat-induced point mutation (RIP), which induces increased mutation rates in repeats (29). However, individual TE insertion can also have beneficial impacts. For instance, in the wheat pathogen *Zymoseptoria tritici*, a TE insertion upstream of the *Zmr1* promoter regulated the diversity of melanin accumulation (30). TE-derived chromosomal rearrangements in industrial and sea floor adapted strains of *Penicillium chrysogenum* likely increased penicillin production (31,32). Over longer evolutionary time frames, TE derived genes can also become co-opted by integrating into host gene functions (33,34).

The fungal genus *Pseudocercospora* contains over 300 species (35) predominantly consisting of host-specific fungal pathogens that pose a threat to agricultural and natural ecosystems (36). *Pseudocercospora* spp. are globally distributed, with a focus on tropical and subtropical environments (37). Key pathogens include *P. fijiensis*, *P. musae*, and *P. eumusae*, which collectively contribute to the black Sigatoka complex affecting banana crops, and *P. ulei*, responsible for South American Leaf Blight in natural rubber plants *Hevea brasiliensis* (38–40). Despite their significance in agriculture and the environment, only a few *Pseudocercospora*

92 species have been sequenced. The existing genome assemblies strongly indicate that the genus has
93 undergone significant genome size changes (40–45). *Pseudocercospora fijiensis* and *P. ulei*
94 exhibited genomes with sizes of 74 Mb and 93.8 Mb, respectively, which are among the largest
95 genomes in (45) the Capnodiales (21). Both genomes were reported to harbor a significant amount
96 of repeats including TEs (40,44,46). A TE hAT element captured H3 proteins and created 784
97 copies ultimately targeted by RIP throughout the repetitive regions in *P. fijiensis* (47). Compared
98 to *P. fijiensis*, in *P. ulei* lacks detailed analyses of the TE families responsible of the genome
99 expansion.
100
101 Here, we analyzed genome size expansion in the *Pseudocercospora* genus and closely related
102 sister clades. We found *P. musae*, *P. fijiensis* and *P. ulei* to have the most expanded genome sizes,
103 independent of the number of annotated genes. To detect genes potentially involved in the
104 interaction with the host, we made extensive analyses on pathogenicity-associated genes and found
105 a strong reduction in *P. ulei*. We then compared TE content among species and found substantial
106 variation in TE content, indicating ongoing TE activity after speciation. Specific retrotransposons
107 were the primary driver of genome size expansion with distinct contributions to independent
108 genome size expansions. Finally, we found that at least some TE activity persists within *P. ulei* by
109 analyzing genome sequencing data of six strains collected across Colombia.
110
111 **RESULTS**
112 ***Pseudocercospora* includes the largest known genomes in the Mycosphaerellaceae**
113 To provide a comprehensive characterization of *Pseudocercospora* spp. genome evolution, we
114 performed a comparative analysis including twenty fungal phytopathogen species grouped within

the Mycosphaerellaceae family for which genome assembly and gene annotation data were available in public databases (Supplementary Table S1). Of these, ten species belong to *Pseudocercospora*, i.e. *P. fijiensis*, *P. musae*, and *P. eumusae*, which are part of the Sigatoka disease complex affecting banana crops (41), *P. ulei*, a major threat to natural rubber crops (44), *P. macadamia*, the causal agent of husk spot in macadamia crops (42), *P. cruenta*, responsible for Cercospora leaf spot in cowpea crops, *P. pini-densiflorae*, a cosmopolitan pathogen affecting various pine species, *P. fuligena*, a tomato pathogen causing black leaf mold (43), *P. vitis*, the causal agent of isariopsis leaf spot in *Vitis* spp. crops (48), and *P. crystallina*, a fungal pathogen of eucalyptus crops (35,49). Another eight species from closely related sister clades were included, i.e. *Cercospora beticola*, *C. berteroae*, *C. zeina*, *Sphaerulina musiva*, *Dothistroma septosporum*, *Zasmidium cellare*, *Ramularia collo-cygni* and *Zymoseptoria tritici.* Genome assemblies and annotations of two additional species closely related to *Pseudocercospora*, *Paracercospora egenula* and *Rhachisphaerella mozambica*, were performed in this study.

BUSCO scores for *Pseudocercospora* and sister clade genomes indicated high completeness ranging from 93.8% in *P. musae* to 99.0% in *P. cruenta*, *C. berteroae* and *S. musiva* (Supplementary Figure S1). Genome assembly contiguity (*i.e.*, N50) ranged between 42.9 kb in *P. eumusae* to 5.9 Mb in *P. fijiensis* (Figure 1; Supplementary Table S1). The largest genomes were predominantly represented by the *Pseudocercospora* genus (Figure 1, highlighted in mauve). Genome sizes of all analyzed species ranged from 29.3 Mb in *S. musiva* to 93.7 Mb in *P. ulei*. To compare the number of predicted coding regions per genome, we performed gene annotations for the following eight species to fill gaps in publicly available data, i.e. *P. crystallina*, *P. cruenta*, *P. fuligena*, *P. vitis*, *P. pini-densiflorae*, and the outgroups *Pa. egenula* and *Rh. mozambica*

(annotations available on Zenodo: https://zenodo.org/records/15862053). The number of annotated protein-coding genes in Mycosphaerellaceae species ranged from 7,342 to 16,015 (Figure 1). *Za. cellare* and *P. macadamiae* exhibited the highest counts of annotated genes, with 16,015 and 15,430 genes, respectively. Conversely, *P. vitis* and *P. crystallina* displayed the lowest counts of annotated genes, with 7,342 and 8,716 predicted genes, respectively (Figure 1). The limited number of gene candidates may be attributed to the lack of specific training in the gene prediction algorithm. *Pseudocercospora ulei* has undergone a substantial genome expansion compared to all species in this study. The genome of *P. ulei* was three times larger than its closest relative, *P. vitis*, and 1.2 times larger than *P. fijiensis*, the second-largest genome in this study. The three largest genomes, *P. ulei*, *P. fijiensis* and *P. musae* each showed genome size increase, yet no increase in gene numbers.

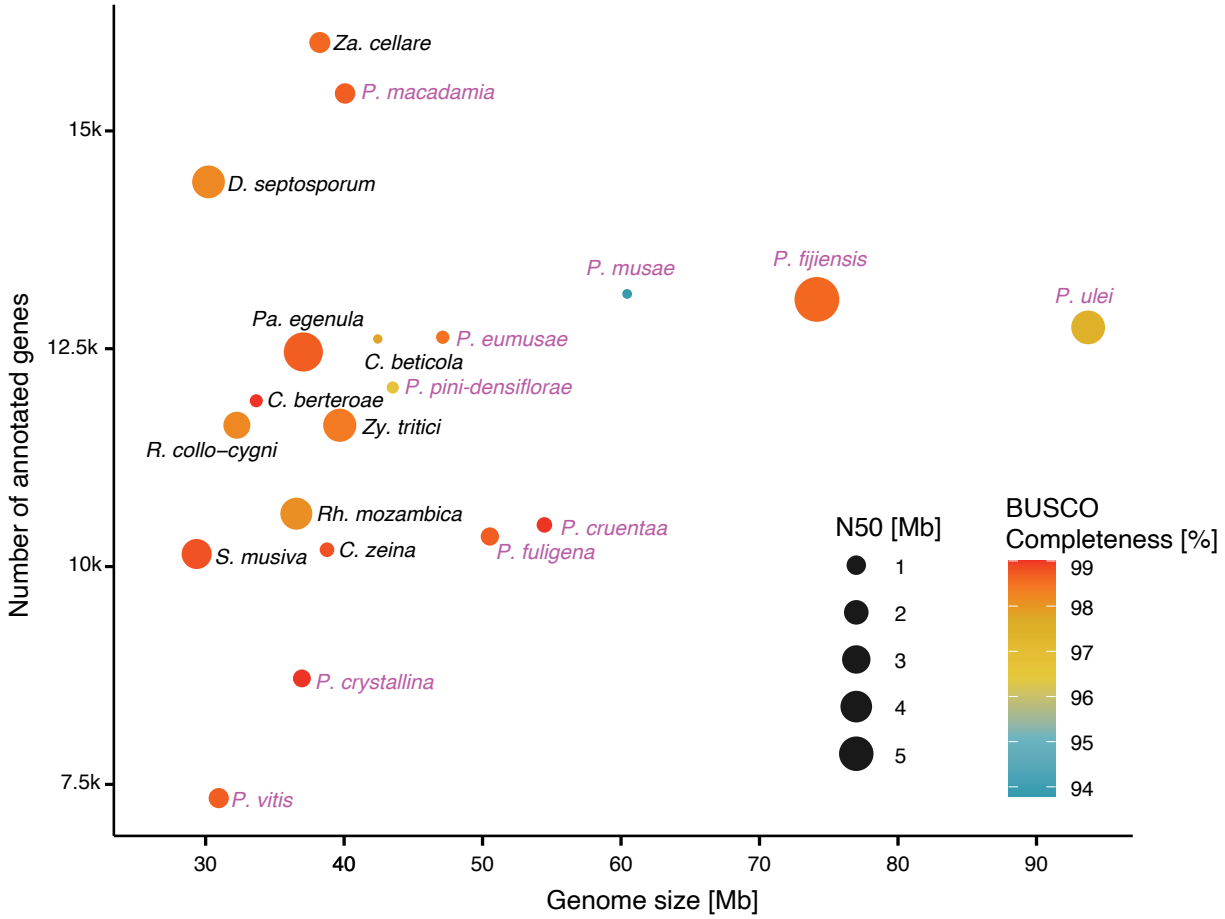
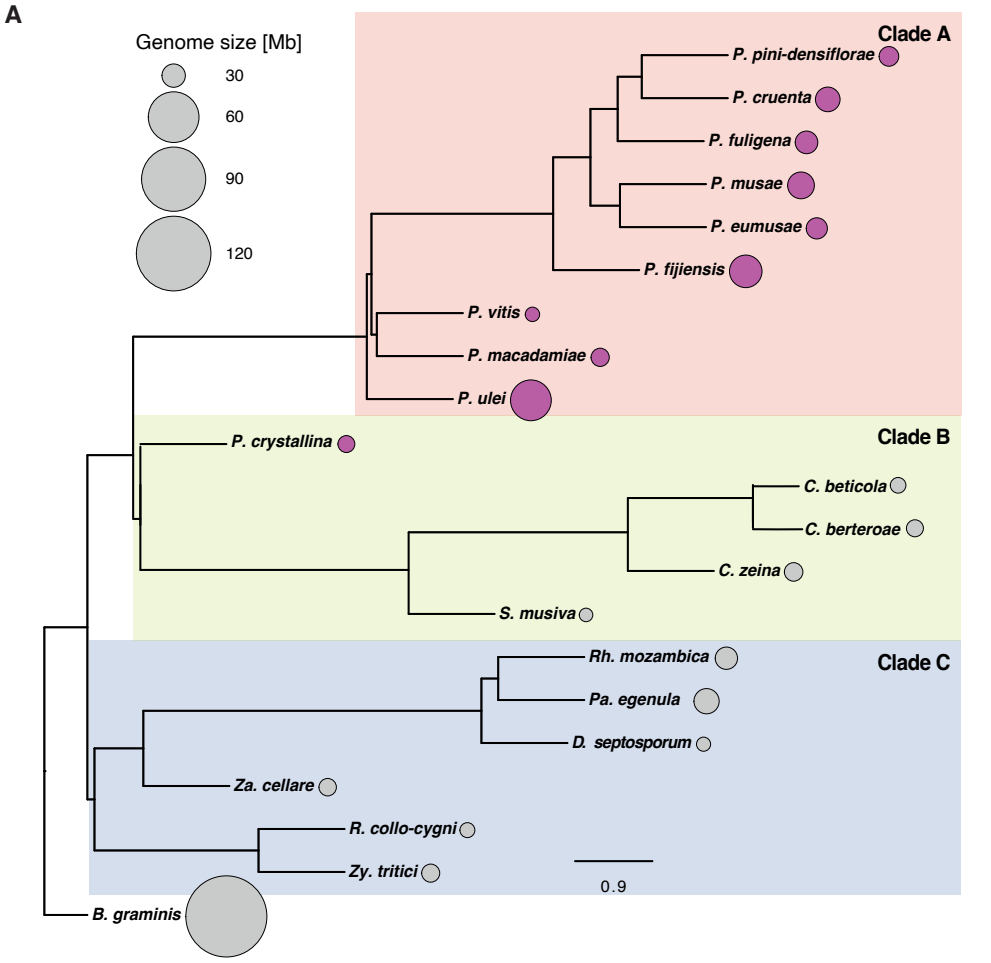


Figure 1. Genome assembly quality metrics of *Pseudocercospora* and closely related species including assembled genome size, gene content and N50 values. The dot color represents to genome assembly completeness score, which was assessed with the number of complete single copy BUSCO (Benchmarking Universal Single-Copy Orthologs) genes. Dot size indicates N50. Genomes from the *Pseudocercospora* genus are labeled in mauve.

Phylogenomic analyses reveal independent genome size expansions

The phylogenetic relationship of the *Pseudocercospora* genus and closely related species within the Mycosphaerellaceae was assessed using *Blumeria graminis* from the Erysiphaceae family as an outgroup to root the tree. The species grouped into three distinct clades as expected (Figure 2A). Clade A contained most of the *Pseudocercospora* species, except for *P. crystallina*. The two newly assembled species *Rh. mozambica* and *Pa. egenula* clustered with clade C. The species closest to massively expanded *P. fijiensis* and *P. ulei* genomes each showed small genome sizes.



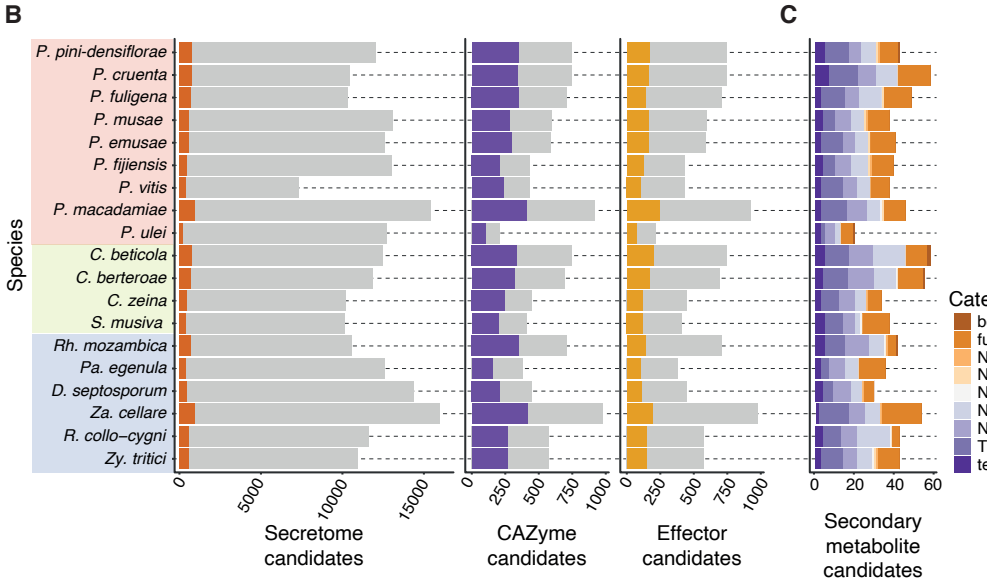


Figure 2. Phylogenetic relationship and pathogenicity-associated genes in Mycosphaerellaceae family genomes. A) Phylogenetic tree of species within the Mycosphaerellaceae family. Dot plots represent the genome size. Genomes belonging to the *Pseudocercospora* genus are filled in mauve. *Blumeria graminis* was used to root the tree as an outgroup. B) Secreted protein profiles of species within the Mycosphaerellaceae family. Left: The gray background represents the total proteome, and dark orange indicates the predicted secretome. Middle and right: The gray background represents the predicted secretome, while carbohydrate-active enzymes (CAZyme) are shown in purple and effector candidates are shown in yellow. C) Secondary metabolite gene clusters in species of the Mycosphaerellaceae family. The colors indicate the different categories of secondary metabolite gene clusters.

Reduction in pathogenicity-associated genes in *P. ulei*

To assess whether genomes experienced an expansion of pathogenicity-associated genes, we estimated the number of candidates for the secretome, with a focus on carbohydrate-active enzymes (CAZymes) and effectors (Figure 2B). As expected, the secretome candidates made up only a small share of the entire proteome. We identified differences in the *Pseudocercospora* genus with the largest secretome in *P. macadamiae* (n = 921 proteins) and the smallest in the closely related *P. ulei* (n = 212). *Pseudocercospora vitis* presented a reduced proteome with a similar number of secretome candidates compared to the genus. *Pseudocercospora ulei* showed a reduced number of CAZymes (n = 102) and effector candidates (n = 73) as well. However, CAZymes and effectors made up a larger proportion of the secretome compared to the other *Pseudocercospora* and closely related species. Secondary metabolite gene clusters showed similar numbers and proportions in categories among the genus (Figure 2C). *Pseudocercospora ulei* also showed a reduced number of secondary metabolite gene clusters, with a proportionally strongest reduction in T1PKS and NRPS categories. The number of pathogenicity-associated genes may be correlated with lifestyle in fungi. *Pseudocercospora ulei* was described as a biotroph, while the other species are likely either necrotrophic or hemibiotrophic. Consistent with total gene content, pathogenicity-associated genes and gene clusters were not correlated with genome size expansions in *P. fijiensis* and *P. ulei*.

194

195 **Genome size increases associated with TE expansions**

196 To objectively compare the repeat content among *Pseudocercospora* and closely related species,
197 we used an assembly-free approach based on short read sequencing. Such an approach likely
198 underrepresents repeat content but removes the bias stemming from unequal genome assembly
199 qualities. We used the tool dnaPipeTE for assembly-free repeat detection assessing the following
200 repeat types: low complexity, rRNA repeats, simple repeats, and TEs. TEs were further classified
201 into LTR retrotransposons, LINEs and DNA transposons. The repeat content varied strongly
202 within the *Pseudocercospora* genus and closely related species, ranging from 1.63% in *P.*
203 *macadamiae* to 71.02% in the closest relative *P. ulei* (Figure 3, Supplementary Figure S2A).
204 *Pseudocercospora macadamiae*, *P. pini-densiflorae* and *Rh. mozambica* showed very low repeat
205 contents of less than 5%, all of which have small genomes. In *P. cruenta* and *P. eumusae* around
206 a quarter of the genome was covered by repeats, and in *P. musae, P. fijiensis* and *Pa. egenula*,
207 around half of the genome was covered by repeats. Generally, closely related species showed
208 drastically different repeat contents. We observed a significant correlation (Pearson's, $r = 0.8$, $p =$
209 0.01) between genome size and the proportion of repetitive sequences across *Pseudocercospora*
210 species, indicating that genome size expansion is likely driven largely by the proliferation of
211 repetitive elements. Additionally, we found a strong negative correlation ($r = -0.84$, $p = 0.004$)
212 between the repetitive content and the proportion of pathogenicity-associated genes in
213 *Pseudocercospora* genomes. Most repeats in genomes with moderate to high repeat content were
214 unclassified TEs. Failure for classification by dnaPipeTE likely stems from fragmentation or low
215 coverage. We found variation in TE lengths among genomes with *e.g.*, most TEs being below 500
216 bp in *P. ulei* (Supplementary Figure S2B). Given that LTR retrotransposons can vary between a

few kb to tens of kb, the small repeats lengths in *P. ulei* indicates nested TE insertions resulting in fragmentation. LTR retrotransposons remained at low proportions, except for *P. ulei*, where these accounted for ~30% of the genome. Other TE types were only detected at low proportions. LINEs were only detected in *P. eumusae* and *P. ulei*, and DNA transposons were only detected in *P. musae*. Simple repeats expanded slightly in *P. cruenta* and *P. ulei*. Repeats of low complexity and rRNA remained at low proportions throughout the genus. Neither the repeat content nor the types of repeat content correlated with the phylogenetic position of the species, indicating likely independent bursts of repeat activation.

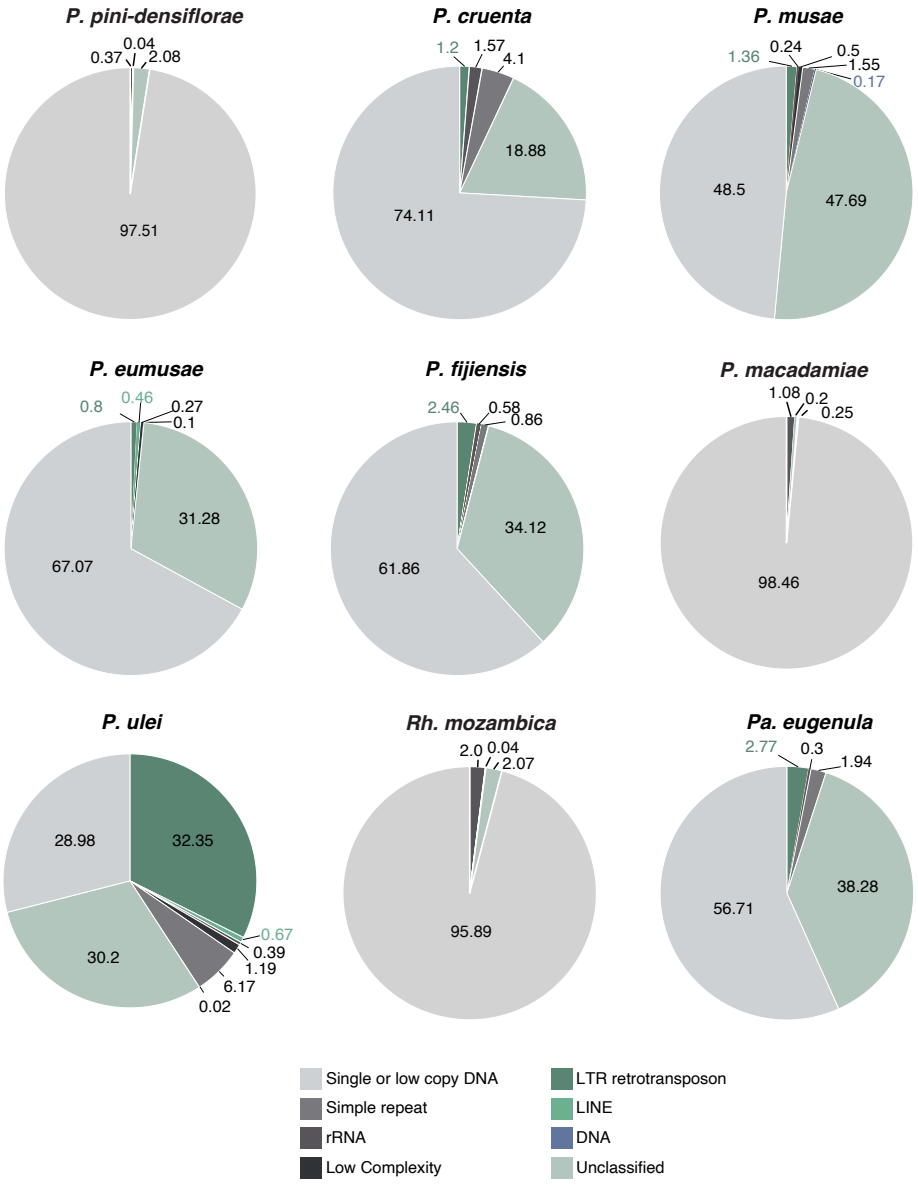


Figure 3. Repeat element distribution among *Pseudocercospora* genus genomes and two closely related species. Light gray indicates the estimated non-repetitive portion of the genomes. Dark grays indicate repeats of low complexity, rRNA and simple repeats. Green and blue colors indicate TEs. The order of the plots follows largely the phylogenetic grouping (Figure 2A).

LTR retrotransposons underpin genome size expansions in *Pseudocercospora*

The assembly-free TE detection with dnaPipeTE indicates strong and phylogeny-independent TE expansions. To further clarify the genome size expansion dynamics, we used the Earl Grey pipeline to produce high-quality TE annotations and classifications for the two high-quality expanded

genomes of *P. fijiensis* and *P. ulei*. TE coverage in the genome assessed by Earl Grey was similar to the estimations by dnaPipeTE (Figure 4A; Earl Grey TE family consensus sequences and annotations are available on Zenodo: https://zenodo.org/records/15862053; TE family names are species specific and do not indicate shared families). However, classification of the individual TE families was improved by access to the full genome sequence. Length estimates of TE fragments were also more robust. Most TE copies belong to LTR retrotransposons consistent with the assembly-free TE detection approach. In *P. fijiensis*, LTR retrotransposons covered 29.4% (21.8 Mb) of the genome, followed by unclassified TEs (13.3%, 9.8 Mb), DNA transposons (3.9%), Rolling-circle/Helitrons (2.7%), and LINE (1.7%). In *P. ulei*, LTR retrotransposons made up the largest part of the genome (74%, 69.5 Mb), followed by a small fraction of unclassified TEs (3.1%, 2.9 Mb), and satellite sequences (0.5%). The bulk of the detected TE fragments overall, which includes full-length elements and fragmented ones due to nested insertions were either LTR retrotransposons or remained unclassified (Figure 4B). The diversity of TE superfamilies is larger in *P. fijiensis* (n = 22), while the *P. ulei* genome includes only 8 TE superfamilies. Notably, most LTR retrotransposons belong to the RLG superfamily (formerly known as *Gypsy*, and to be renamed, see (50)), with 4,787 TE fragments in *P. fijiensis* and 26,795 TE fragments in *P. ulei*. Other LTR retrotransposons were only found at low copy numbers. Given the evidence for high degrees of TE fragmentation, we compared the lengths of each LTR retrotransposon fragments between the two species and *Z. tritici*, which was subject to an extensive manual TE curation (51,52). Consistent with the assembly-free approach, the mean length of LTR retrotransposons was lower in *P. ulei* compared to the other species, however, most TE fragments were found to be >500 bp (Figure 4C). Mean LTR retrotransposon length was 3,996 bp in *P. fijiensis*, 2,569 bp in *P. ulei*, and 3,790 in *Z. tritici*.

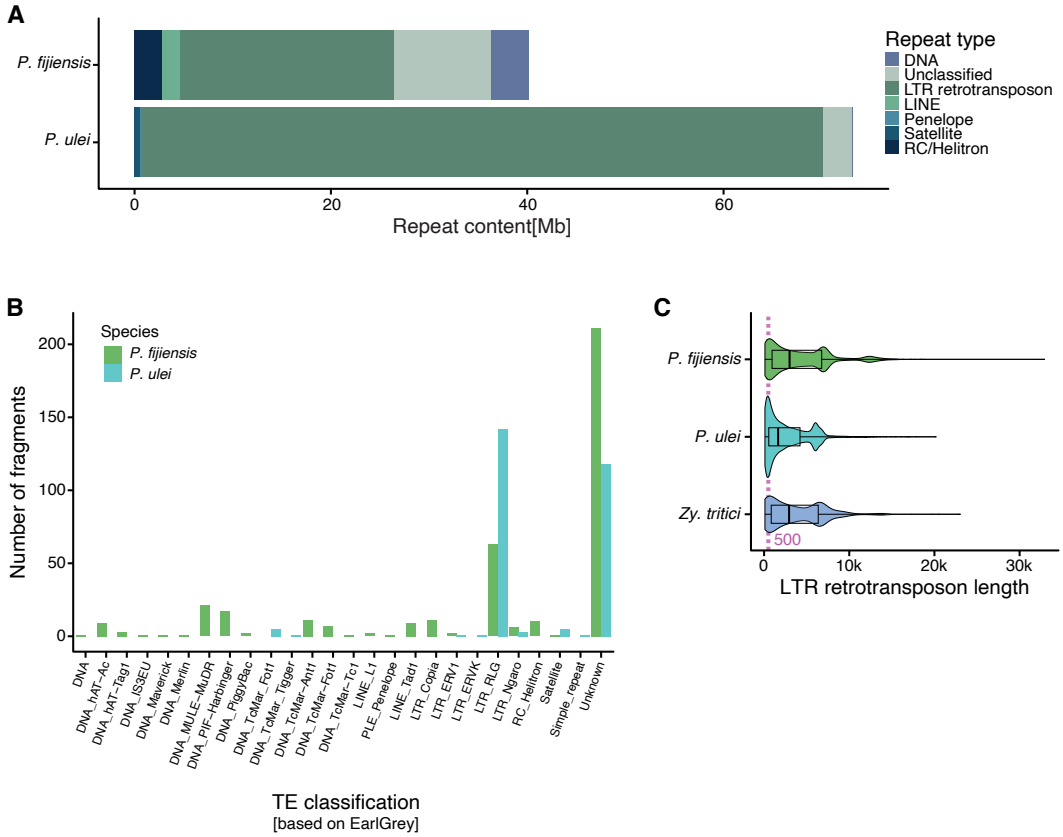


Figure 4. TE coverage in the two enlarged genomes of *P. fijiensis* and *P. ulei* based on high-quality detection and classification with Earl Grey. A) Total length of TEs per genome. The colors indicate the TE category. B) Number of TE copies assessed by Earl Grey for the different TE superfamilies. Copies were counted as both full-length TEs or fragments. Colors indicate the species. C) Distribution of LTR retrotransposon fragment lengths detected in *P. fijiensis* and *P. ulei* compared to the manually curated TE content of the *Z. tritici* outgroup. TEs include both full-length and fragmented TEs. The mauve line indicates 500 bp (as in Supplementary Figure S2B that shows length distribution of any repeat discovered with dnaPipeTE).

The genome-wide analysis of TEs revealed 391 species-specific families for *P. fijiensis*, and 277 for *P. ulei*, with no shared TE families. To identify candidate TE families responsible for recent TE activity bursts and subsequent genome size expansion, we filtered for families with copy numbers above 200. The *P. fijiensis* genome showed only two TE families with copy numbers above 200, but we found 29 such TE families in *P. ulei*, one of which had >800 copies (Figure

5A). The high-copy TE families in *P. fijiensis* belong to RLG and Helitrons, and RLG in *P. ulei*, with two TE families remaining unclassified. Taken together, this suggests that the repeat expansion in *P. fijiensis* likely stems from a higher diversity of low copy TEs, while the high-copy numbers of a few TE families in *P. ulei* suggests a more recent burst of fewer TE families.

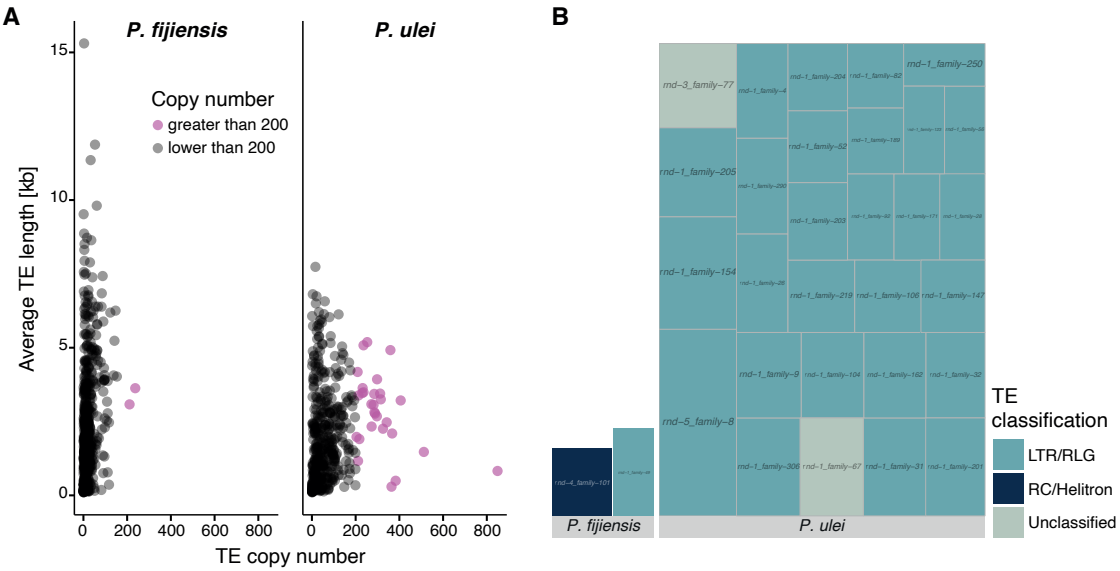
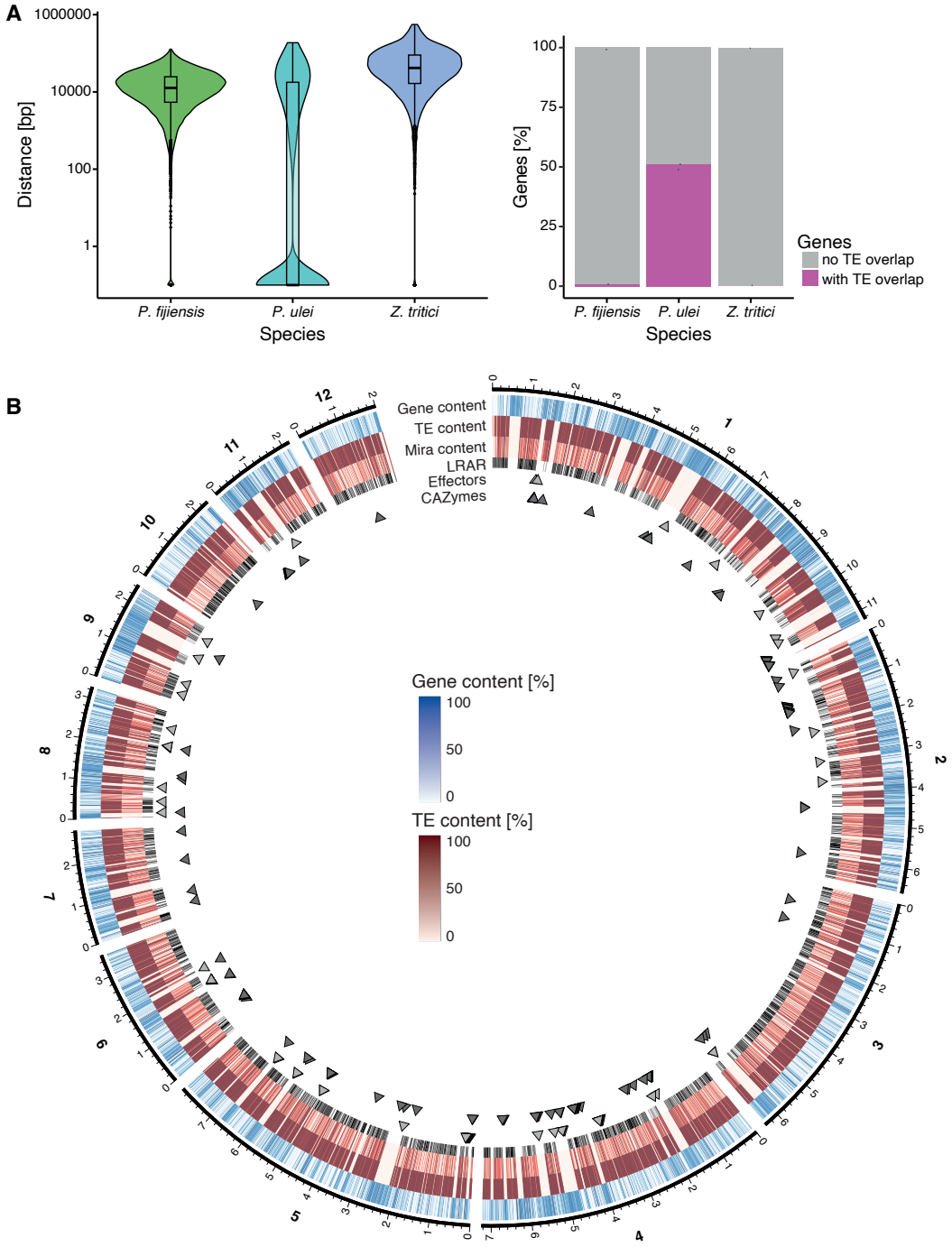


Figure 5. TE families with high copy numbers in *P. fijiensis* and *P. ulei* based on high-quality detection and classification with Earl Grey. A) Correlation of copy number and average length of species-specific LTR retrotransposons. TEs include both full-length TEs and fragments. Mauve dots indicate TE families with more than 200 copies. B) Copies of high-copy number TE families annotated in *P. fijiensis* and *P. ulei*. Colors indicate the superfamily. Each TE family is indicated by a box, and box size indicates the number of copies.

To improve the TE classification and to reduce fragments of TEs erroneously classified as full-length elements, we conducted a manual curation of TE families in *P. ulei*, created consensus sequences and renamed the remaining TE families according to the three-letter code from Wicker et al (2007). Many RLG or unclassified families detected by Earl Grey were mostly fragments of the newly named RLG_Mira family, followed by RLG_Ginan. Given the estimated length for the TE consensus sequences, we confirmed that most TEs in *P. ulei* were fragments of less than 80%

294 full-length (Supplementary Figure S3A). Among high-copy retrotransposons, lengths remained
295 highly variable, especially for the two high-copy TE families RLG_Mira and RLG_Ginan
296 (Supplementary Figure S3B). We created a phylogenetic tree for RLG_Mira coding regions and
297 identified two well differentiated clusters (Supplementary Figure S3C). GC content of RLG_Mira
298 coding regions showed almost exclusively a moderate to high GC content, indicating they the
299 element was not affected by RIP despite being recently active.
300
301 **Genomic landscape of the reference-quality *Pseudocercospora* genomes**
302 To identify TEs located close to genes, we calculated the distances between annotated genes and
303 the closest TE (Figure 6A). TEs were generally closer to genes in *P. ulei* (mean = 14,922 bp)
304 compared to *P. fijiensis* (mean = 17,785 bp) and *Z. tritici* (mean = 68,130 bp). Only a small number
305 of direct overlaps were detected in *P. fijiensis* (n = 98, 0.8% of all genes) and *Z. tritici* (n = 31,
306 0.3% of all genes), however, significant overlaps were found in *P. ulei* (n = 6,441, 51.1% of all
307 genes). Furthermore, we analyzed gene, TE in general and RLG_Mira contents in windows of 10
308 kb for the largest 12 scaffolds in *P. ulei* (Figure 6D). We found a strong compartmentalization
309 between TE-rich regions with a reduced number of genes and gene-rich, TE-depleted regions.
310 RLG_Mira elements were present in most TE-rich regions, but differed in the amount of overlap.
311 Next, we overlayed large RIP affected regions, which showed a similar distribution as the TE-rich
312 regions. Effector and CAZyme candidates were detected in each of the compartment types. The
313 strong compartmentalization of the *P. ulei* genome indicates strong purifying selection acting
314 against new TE insertions in gene-rich regions, and relaxed selection in TE-rich regions. The
315 proximity of TEs and some genes could also stem from some misannotated genes being in rather
316 TEs.

317



318

319 **Figure 6**. TE landscape in *P. fijiensis*, *P. ulei* and *Z. tritici*. A) Variation in distances between genes and
320 the closest TE. Genes with more than one TE insertion were treated as a single insertion. B) Circos plot
321 visualizing the genomic landscape for *P. ulei*. Genes, TEs and RLG_Mira content were calculated for
322 window sizes of 10 kb. Large RIP affected regions (LRAR) indicate regions of at least 4 kb with a high
323 RIP composite index.

TE content analyses of *P. ulei* strains

To determine whether some TE activity persist in *P. ulei,* we sampled and whole-genome sequenced strains from natural rubber tree infections in three different locations across Colombia (Figure 7A). We assessed the genetic structure of the *P. ulei* strain collection using 1,802,029 genome-wide SNPs. Strains clustered into three distinct groups according to geography (Figure 7A). We then mapped short-read data against the manually curated *P. ulei* consensus library to assess the coverage and estimate the number of copies for each TE family. TE copy numbers were largely stable and similar to the direct assessment in the reference genome (Figure 7B). Two strains had a slightly higher number of estimated TE copy number than the reference strain. Estimated TE copy numbers varied slightly, but independent of geographic origin. Like the reference genome, retrotransposons and the RLG superfamily were overrepresented in the additional strains as well. At the TE family level, RLG_Mira and RLG_Ginan families were the predominant components of the repetitive content of all the *P. ulei* strains assessed, as seen in the reference genome (Figure 7C). We observed small differences in TE family coverage among the six strains and the *P. ulei* reference genome data, however these differences may reflect limitations in TE detection with short reads rather than biological variation. High TE content is, hence, a broadly shared pattern within the species, and variability in content indicates that TE activity might be ongoing.

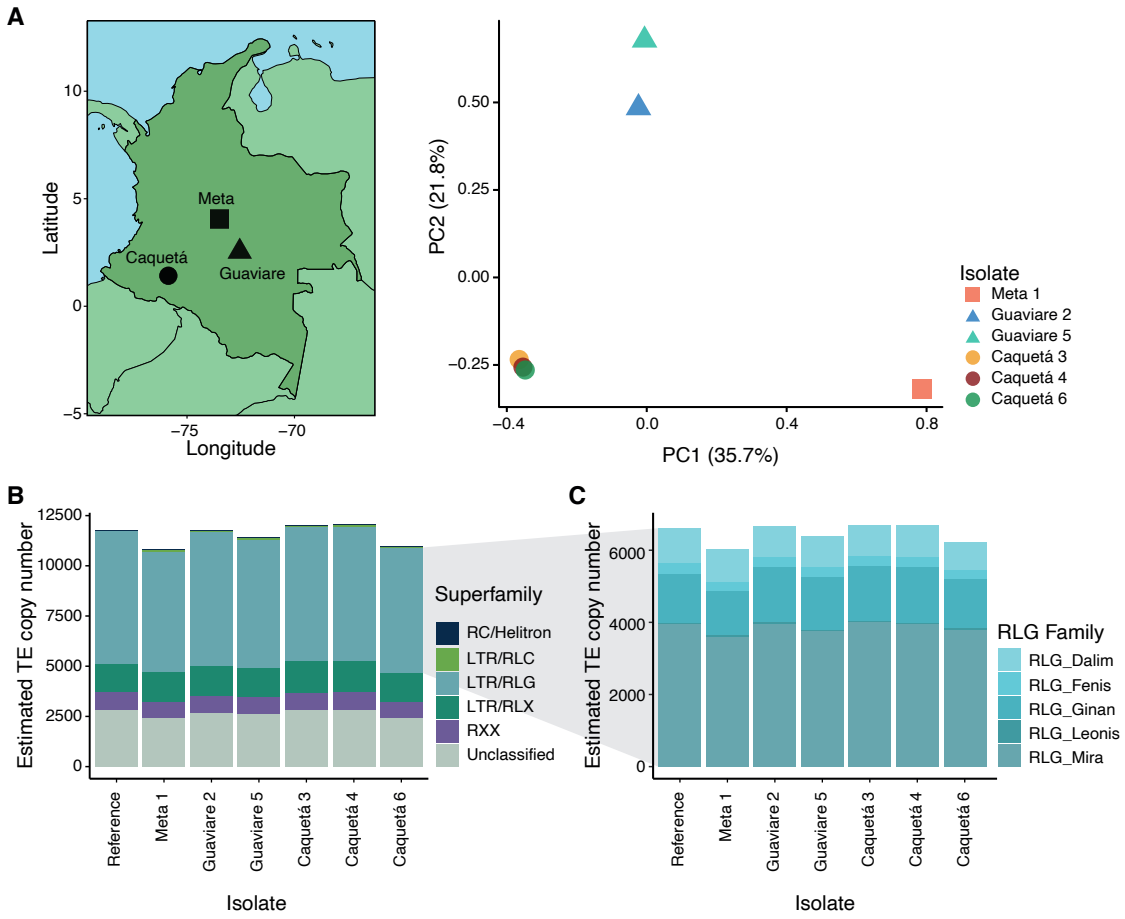


Figure 7. Whole-genome sequencing and TE content analyses of six *P. ulei* strains collected in Colombia. A) PCA analysis of *P. ulei* strains based on linkage-pruned genetic variants and their geographical locations. B) Estimated coverage of *P. ulei* strains using the McClintock coverage analysis based on a curated TEs library for the genus. The color represents the superfamily. C) Estimated coverage of species-specific RLG TE families.

DISCUSSION

Tracking genome size evolution among fungi remains limited to few groups including the Pucciniales, Erysiphaceae or Glomeraceae (53–57). Our study aimed to explore genome size dynamics within the species-rich *Pseudocercospora* genus consisting predominantly of host-specific plant pathogens (35). We found highly variable genome sizes even among closely related species. Our assembly-free approach based on low coverage short reads allowed us to create draft repeat coverage estimations that could be confirmed with high-quality genome assemblies.

Genome enlargement did not correlate with a higher number of genes or pathogenicity-associated genes. The largest genome, *P. ulei,* even showed a slight reduction in coding sequence content, and a dramatically increased number of TE insertions into predicted genes. Genome expansions were largely caused by TE expansions, but the expansion characteristics varied in terms of number and diversity of TE families involved. RLG retrotransposons were consistently involved in genome size expansions, yet no TE family was involved in more than one observed burst. Genome size expansions are likely phylogeny-independent, and might be caused by the activation of specific TE families during stress conditions, or by horizontal transfer of TEs with subsequent bursts in the new host.

To compare genomes of various quality and annotation, we used an assembly-free approach with short reads. dnaPipeTE detects TEs even in low coverage genomes. The repeat contents of the genomes remain a rough estimation, and most of the potential TE families were not classified. The use of short reads made it harder to estimate the coverage of the genome by TEs and other repeats, as repeats are too short to cover most TEs, and the presence of a TE fragment carries no information about the size of the full-length TE. However, when using the assembly based TE detection with Earl Grey in the two best assembled genomes, i.e. those of *P. fijiensis* and *P. ulei*, we found similar repeat contents. Earl Grey estimated larger numbers of classifiable TE families. However, the short read approach was not useful to fine-tune TE classifications or clearly assess locus specific TE insertions. In contrast, the tool-provided coverage analyses of the genomes provided a less skewed variation in TE content.

379 Our findings indicate that genome size expansions among *Pseudocercospora* species were largely
380 caused by differences in TE content. We found a striking difference in genome size and repeat
381 content between each *P. ulei* and its closest relative *P. macadamiae*, and between *P. fijiensis* and
382 its closest relatives *P. musae* and *P. eumusae*. In addition, different, species-specific TE families
383 were responsible for these individual expansions. *Pseudocercospora ulei* had an almost exclusive
384 expansion by LTR retrotransposons, namely by the two families RLG_Mira and RLG_Ginan.
385 *Pseudocercospora fijiensis* also had an expansion of RLG elements, although DNA transposons
386 and LINEs were part of the expansion as well, the number of TE families was higher, yet the copy
387 per family was a bit lower. *Pseudocercospora fijiensis* TEs are known to be severely impacted by
388 the defense mechanism RIP (46). Even though *P. ulei* shows many large RIP affected regions, the
389 expanded RLG_Mira coding sequences show no indications of RIP effects. This raised the
390 question whether RIP is still functional in *P. ulei*. Recent losses of the RIP machinery were
391 detected repeatedly among ascomycetes (58).
392
393 The TE induced genome size expansions were most likely initiated by independent triggers,
394 resulting from activated TE families not shared by a direct ancestor. TE activity is often caused by
395 stress conditions, in which previously silenced TEs are de-repressed and create new copies (28,59).
396 For fungal plant pathogens, stress includes the response of the host, fungicide application or
397 climatic factors. Such stressors might have induced TE activity, initiating stepwise genome size
398 increases. Expansions would have been countered if TEs were inserted frequently into conserved
399 regions, creating strong negative effects. Fungi might experience elevated stress when adapting to
400 a new host species. *H. brasiliensis*, the host of *P. ulei*, contains a poisonous latex with antifungal

properties (60). Depending on which host the last shared ancestor of *P. ulei* and *P. macadamiae* occupied, one or both species might have encountered similar new stress conditions.

The strong clustering of TEs into chromosomal compartments in *P. ulei*, and the large amount of fragmented TEs rather than retention of full-length copies indicates that existing TE-rich regions are preferred insertion sites for new TEs, by this contributing to fragmentation. *Pseudocercospora ulei* shows a reduced set of effectors, a characteristic shared with *Oidium heveae*, another pathogen of *H. brasiliensis* (61). Further examples do not follow this same pattern though with other fungal and oomycete pathogens of *H. brasiliensis* showing no reductions, or even increases in effector gene content. Hence, reduction of effector genes is unlikely to be a general adaptation to the host plant (62–66). Genus-wide analyses and beyond can identify broad patterns of factors influencing TE activity. Our study reinforces the observation that genome size expansions are initiated mostly in terminal branches making direct observation of causal factors very challenging. However, broad surveys of agriculturally relevant plant pathogen genera will build towards a more complete picture of the interplay of TEs, genome sizes and pathogen functions.

METHODS

Data acquisition and genome assembly

We obtained genome assemblies for 20 species of the Mycosphaerellaceae family from the National Center for Biotechnology Information (NCBI, Supplementary Table S1). Our comparative genome analysis focused on ten species from public databases from the *Pseudocercospora* genus: *P. cruenta*, *P. crystallina*, *P. eumusae*, *P. fuligena*, *P. fijiensis*, *P. musae*,

P. macadamiae, *P. ulei*, *P. pini-densiflorae*, and *P. vitis*. An additional eight related species genomes were accessed from public databases: *C. beticola*, *C. berteroae*, *C. zeina*, *S. musiva*, *D. septosporum*, *Za. cellare*, *R. collo-cygni* and *Z. tritici*. Because no assembly was available, we used short-read sequencing data to produce draft genome assemblies for two additional closely related species, *Pa. egenula* and *Rh. mozambica*. *Pa. egenula* and *Rh. mozambica* genome assemblies were constructed from Illumina short-reads. For that end, Illumina raw reads were initially assessed with FastQC v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were pre-processed with Trimmomatic v.0.39 using the following parameters: ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10:1:true TRAILING:2 SLIDINGWINDOW:4:15 MINLEN:90 (67). For phylogenetic analyses, we included *B. graminis* as an outlier group to root the tree (68)**.** Trimmed reads were assembled with SPAdes v.3.13.0 using the --careful parameter (69). Assemblies were evaluated with QUAST v.5.2.0 (70). Genome completeness was assessed based on orthologous gene composition obtained with BUSCO v.5.7.1 using the ascomycota_odb10 database (71).

Phylogenetic analyses

We performed phylogenomic inference using 1315 orthologous BUSCO gene sequences that were shared by all genomes. Amino acid sequences from single copy orthologous genes shared between the assessed species were aligned with MAFFT v.7.310 with the parameters --genafpair --maxiterate 1000 (72). A maximum likelihood tree was estimated with IQTree v.2.2.5 using 1000 bootstrap replicates (73). For this, we initially concatenated independent tree files previously inferred and then estimated the tree with Astral v.5.7.8 using the default parameters and *B. graminis* as outgroup to root the tree (74).

447

448 **Structural and functional annotation**

449 To compare functional annotations among genomes, we performed structural annotations for
450 genomes were this information was lacking: *P. crystallina*, *P. cruenta*, *P. fuligena*, *P. vitis*, *P. pini-*
451 *densiflorae*, *Pa. egenula* and *Rh. mozambica*. For this, we used Augustus v.3.5.0 (75), training the
452 gene predictors with gene models from the *P. ulei* reference genome (44). Predicted proteomes
453 were extracted from the gene candidate catalogs using gffread v0.12.1 (76). To predict proteins
454 interacting with the plant host, we used the Predector pipeline v.1.2.7 (77). Predector integrates a
455 range of fungal secretome and effector discovery tools, and ranks the effector candidates based on
456 machine learning methods. Predector also includes CAZyme identification which is predicted by
457 homology mapping the amino acid sequences against dbCAN v.10 database (78). Secondary
458 metabolite gene clusters were predicted from genome assemblies using the antiSMASH web server
459 v.7.0 (79).

460

461 **Transposable element annotations**

462 Given the variable quality of genome assemblies in the *Pseudocercospora* genus, we resorted to
463 analyzing TEs with dnaPipeTE, an assembly-free tool optimized for TE detection in low coverage
464 short read datasets (80). Briefly, dnaPipeTE analyzes repetitive elements by processing raw
465 genomic reads. It selects three subsamples of low coverage (<1x) reads to assemble TEs into
466 contigs. The assembled contigs are then annotated through homology comparisons using the
467 DFAM database (81). DnaPipeTE was used for TE annotation only in the eight *Pseudocercospora*
468 species and two closely related species *Pa. egenula* and *Rh. mozambica* for which short-read
469 sequencing data were available (Supplementary Table S1).

Sampling of *P. ulei* strains

We gathered a total of six *P. ulei* strains isolated in the main rubber producer regions of Colombia (Figure 7A). *Pseudocercospora ulei* strains were isolated from *H. brasiliensis* clones established in different regions of Colombia. Two strains were isolated from the leaves of the FX 3864 clone located in Vereda Santa Rosa in the Guaviare Department and provided by the Guaviare Rubber Producers and Marketers Association. Three strains were isolated from leaves of the IAN 873 clone located in the municipality of Belén de los Andaquíes on the Los Gomas farm and provided by the Universidad de la Amazonia. One strain was isolated from the leaves of the RRIM600 clone located in the clonal gardens of Villavicencio - La Libertad and provided by the Corporación Colombiana de Investigación Agropecuaria (Agrosavia) and . All the samples were obtained under the Addendum No. 20 of the Framework Contract for Access to Genetic Resources and their Derivatives (No. 121 of January 22, 2016) established between the Ministry of Environment and Sustainable Development and the National University of Colombia. The information about each geo-referenced sampling point is shown in Supplementary Table S2.

Propagules were isolated from single foliar, sporulating lesions from which conidia were collected and cultured on M3 solid medium at 25ºC in the dark for 45 days until visible stroma formation according to the protocol (82). Once the stroma reached a size of 5x5 mm these were macerated into 2 mL microcentrifuge tubes (Eppendorf®, Germany) and transferred to 125 mL flasks containing M4 sporulation solid medium. The M4 medium consists of potato broth, amino acids, and peptone (82). Sporulation was stimulated by exposing the cultures to white light for 90 minutes for six days (83).

493 **Genomic DNA extraction and sequencing of *P. ulei***

494 High molecular weight DNA from sporulated stromata was extracted following Stirling's protocol
495 (84), modified by adding phenolic extraction followed by three phases of chloroform extractions.
496 DNA concentration, integrity, and purity were assessed by fluorometry (Qubit®), agarose gel
497 electrophoresis (1%) with Tris-borate-EDTA (TBE) buffer stained with SYBR safe (0.5 mg/L),
498 and spectrophotometry (Nanodrop®), respectively. Short-read sequencing was performed using
499 the DNBSEQ Platform Sequencing on the DNBSEQ PE150 instrument by MGI Inc. (China) at
500 the BGI Hong Kong Tech Solution NGS Lab, utilizing the DNBseq DNA library construction kit.
501 Short read raw data was generated in paired-end mode (2x150 bp).
502
503 **Genome-wide SNP analyses**
504 To assess genetic variation among *P. ulei* strains, whole-genome resequencing data from six strains
505 collected from different departments in Colombia were analyzed. Raw paired-end reads were
506 quality-checked using FastQC v.0.11.9
507 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and adapters and low-quality
508 sequences were trimmed using fastp v.1.01 with a minimum quality threshold of Q20 (85).
509 Trimmed reads were aligned to the *P. ulei* reference genome using BWA-MEM v0.7.17 (86). SAM
510 files were converted to BAM and sorted with samtools v.1.20, and duplicates were marked using
511 Picard v.2.27.4 (https://github.com/broadinstitute/picard). Variant calling was performed using
512 GATK (v.4.3.0.0) HaplotypeCaller in GVCF mode for each sample, followed by joint genotyping
513 with GenotypeGVCFs (87). The resulting variant calls were filtered using GATK VariantFiltration
514 based on the following thresholds: QD < 20.0, QUAL < 10000.0, MQ < 30.0, ReadPosRankSum
515 < -2.0 or > 2.0, MQRankSum < -2.0 or > 2.0, and BaseQRankSum < -2.0 or > 2.0. To explore

516 population structure of the six *P. ulei* strains, we performed a principal component analysis (PCA)
517 based on filtered SNPs. The filtered VCF file was first converted to the PLINK binary format using
518 PLINK v.1.9 (88). Linkage disequilibrium pruning was applied with the option --indep-pairwise
519 50 10 0.2 in PLINK. A PCA was then conducted using the --pca option in PLINK, which generated
520 eigenvalues and eigenvectors. The first two principal components were plotted in R (89), and
521 sample metadata on geographic origin was integrated into the PCA plot to visualize clustering
522 patterns associated with geographic regions.
523
524 **Transposable element genome annotation**
525 To obtain high-quality TE libraries for the two largest and most contiguous genomes of species *P.*
526 *fijiensis* and *P. ulei*, we ran the Earl Grey pipeline v4.1 (90). The Earl Grey pipeline combines
527 identification of TEs based on preexisting libraries and *de novo* approaches for TE annotation.
528 Repetitive elements were first identified and masked by RepeatMasker v4.1.2
529 (http://www.repeatmasker.org), ignoring low-complexity repeats and small RNA genes. The
530 masked genome was subsequently used for *de novo* TE identification performed with
531 RepeatModeler v.2.0.2 (91) using RepBase v.23.08 and Dfam v.3.3 databases for the DNA and
532 amino acid sequence identification. TEs were classified based on the similarity between *de novo*
533 annotated and known TEs, creating a new combined library. Finally, full-length long terminal
534 repeat retrotransposons (LTRs) were identified with LTR_Finder v1.07 (92). To estimate TE
535 distribution and a potential impact on gene integrity and expression, we compared the annotations
536 of TEs and genes in *P. fijiensis* and *P. ulei* and *Z. tritici* separately. We used BEDtools *closest*
537 v.2.30.0 with the parameter -D a (93). Genes with more than one TE insertions were counted as
538 just a single occurrence.

539

540 **Manual TE consensus identification**

541 To obtain high-quality TE family consensus sequences, a manual curation as described in (52) was
542 conducted. In short, the RepeatModeler and Earl Grey consensus sequences were first curated with
543 WICKERsoft (94): similar sequences were searched genome-wide with blastn v.2.13.0 (95). 15-
544 25 sequences of a subset of hits with 300 bp added each up- and downstream were extracted, and
545 a multiple sequence alignment was(G Higgins & M Sharp, 1988)2.1 (96). Visual inspection, as
546 well as information on the sequences of target site duplications and expected start and end
547 sequences were used to define the actual boundaries of each TE family (15), and higher quality
548 consensus sequences were created. New TE families were classified depending on the homology
549 of encoded proteins and the presence and type of terminal repeats, and named after the three letter
550 classification system (15). To remove redundancy and predicted TE families created from TE
551 fragments, each new TE consensus sequence was compared against the already curated consensus
552 sequences with blastn. A large number of previously predicted families turned out to be redundant,
553 as they were fragments of full-length consensus sequences.
554 A second round TE curation was done to identify non-autonomous TE families that do not contain
555 some or all protein sequences. LArge Retrotransposon Derivates (LARD) and Terminal Repeat
556 retrotransposons In Miniature (TRIM) were detected with LTR-Finder and the filters -d 2001 -D
557 6000 -l 30 -L 5000 and -d 30 -D 2000 -l 30 -L 500 respectively. Miniature Inverted-repeat
558 Transposable Elements (MITE) were detected with MITE Tracker (97). Short Interspersed Nuclear
559 Elements (SINE) were detected with SINE-Finder in Sine-Scan (98,99). Predicted consensus
560 sequences were compared with WICKERsoft as described above, and removed if less than 5 copies
561 were detected in the whole genome or if a TE consensus sequence already existed. The *P. ulei*

562 reference genome was then annotated with the curated consensus sequences using RepeatMasker
563 with a cut-off value of 250, and simple repeats and low complexity region hits were filtered out.
564
565 **Phylogenetic reconstruction of RLG_Mira coding regions**
566 To test if the high-copy TE family RLG_Mira underwent a recent burst, we performed multiple
567 sequence alignment and phylogenetic analyses of its coding regions, following an approach
568 established by Oggenfuss et al. 2023. All full-length sequences and fragments of RLG_Mira copies
569 detected with RepeatMasker in *P. ulei* and a copy from *P. macadamiae* as an outlier were extracted
570 with samtools faidx from the reference genome. Sequences on the negative strand were reverse-
571 complemented. The coding sequence of RLG_Mira was extracted with a blastx search against the
572 PTREP18 TE protein database (https://trep-db.uzh.ch/), and the best hit was retained. A multiple
573 sequence alignment was created containing all sequences from *P. ulei*, the copy from *P.*
574 *macadamiae* and the coding sequence using MAFFT and the parameters --reorder --local-pair --
575 maxiterate 1000 -nomemsave--leavegappyregion. The multiple sequence alignment was then
576 trimmed at the start and end positions of the coding sequence using extractalign from EMBOSS.
577 Sequences and fragments that covered less than 50% of the coding region were removed with
578 trimAl v.1.4.rev15 (100). To prevent structural variants in a subset of RLG_Mira copies to from
579 distorting the phylogeny, conserved blocks were extracted with Gblocks v.0.91b, using the
580 parameters -t = d -b3 = 10 -b4 = 5 -b5 = a -b0 = 5 (101). The GC content of each sequence was
581 calculated with geecee in EMBOSS. Maximum likelihood trees were estimated with RAxML v.8.2
582 (102). First, 10 independent maximum likelihood tree searches were conducted using the
583 parameters with the parameters raxmlHPC-PTHREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -
584 # 10 --print-identical-sequences. The best maximum likelihood tree was retained. Second,

bootstrap analysis was performed to obtain branch support values with the parameters raxmlHPC-P- THREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -b 12345 -# 50 --print-identical-sequences. Finally, bipartitions were added to the best maximum likelihood tree with the parameters raxmlHPC-PTHREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -f b --print-identical- sequences. The best scoring maximum likelihood tree was then visualized in R, using read.tree from the package treeio v.1.10.0 to import, ape v.5.7.1 to root the tree based on the *P. macadamiae* copy, tibble v.3.0.1 to add the GC content information to the tree and ggtree(103–106). To detect if RLG_Mira entered *P. ulei* via horizontal transfer, we performed blastx and found that best hits are found in fungi including *Metarhizium anisopliae*.

Genomic environment of the high-quality reference genome of *P. ulei*

To characterize the genomic environment of *P. ulei*, the largest 12 scaffolds of the reference genome were split into non-overlapping 10 kb windows using EMBOSS splitter v.6.6.0 (107). The percentages coverage by annotated TEs, by the high-copy TE family RLG_Mira and genes per window were calculated using BEDtools intersect v.2.30.0 (93). To calculate a potential impact by RIP mutations, large RIP affected regions in the reference genome were detected using The RIPper (108). The visualization was made with circos (109).

TE copy number estimation for *P. ulei strains*

The reference genome is not always representative for the whole species, and might be an outlier, which could explain the high TE density. To determine if field strains from different regions contain similar numbers of TEs, we estimated the coverage for each manually curated TE family. Raw reads were first trimmed with Trimmomatic v.0.33 with the parameters:

608 ILLUMINACLIP:TruSeq3-PE-2.fa:2:30:10 LEADING:3 TRAILING:3
609 SLIDINGWINDOW:4:15 MINLEN:36. Copy numbers for each TE family were then estimated
610 based on normalized coverage and using the method coverage in the McClintock pipeline (110).
611 We attempted to track the positions of the annotated TEs; however, due to their high abundance,
612 it was not possible to identify homologous sites with matches spanning both TEs and non-
613 repetitive genomic regions, preventing their accurate localization within the *P. ulei* genome.
614
615

Declarations

Data availability: Sequence data are deposited at the NCBI Sequence Read Archive under the accession numbers SRR34278610 (*P. ulei*), SRR34278609 (*Pa. egenula*), SRR34278616 (*Rh mozambica*). The additional *P. ulei* isolates were deposited under SRR34278611-SRR34278614 and SRR34278617-SRR34278618. Genome assemblies for *Pa. egenula* and *Rh. mozambica*, gene annotations for the *Pseudocercospora* species and TE annotations for *P. fijiensis* and *P. ulei*, and the TE family consensus sequences are available on Zenodo: https://zenodo.org/records/15862053.

Acknowledgements: We are grateful to the individuals and institutions who provided natural rubber samples that enabled the isolation of *Pseudocercospora ulei* and the generation of Illumina sequencing data from several departments in Colombia. Specifically, we thank Lyda Constanza Galindo Rodríguez from the University of Amazonia for samples collected in the Caquetá department; the ASOPROCAUCHO Association of Rubber Producers and Traders of Guaviare for samples from the Guaviare department; and Olga María Castro from the Research Group on Conservation Agriculture for Lowland Tropical Soils at AGROSAVIA, La Libertad Research Center, for samples from the Meta department. We thank Tobias Baril from the University of Neuchâtel for help with the Earl Grey pipeline.

Funding: SMGS was supported by the Internship Excellence - Foreign Students FCS Postdoctoral Fellowship, granted by the Federal Commission for Scholarships of the Swiss Confederation. UO was supported by the Swiss National Science Foundation (P5R5PB_225522).

Competing interests: The authors declare that no competing interests exists.

638 **Author contributions**: GSSM, UO and DC designed the study. IBA, CAT and FAA provided
639 biological material and performed experiments. AZZ and IS contributed datasets. GSSM and UO
640 conducted analyses. GSSM, UO and DC wrote the manuscript. GSSM and UO acquired funding.
641 UO and DC supervised the work. All authors approved the final manuscript version.
642
643
644

Bibliography

1. Zaccaron AZ, Stergiopoulos I. The dynamics of fungal genome organization and its impact on host adaptation and antifungal resistance. Journal of Genetics and Genomics [Internet]. 2025 May 1;52(5):628–40. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1673852724002844
2. Elliott TA, Gregory TR. What's in a genome? The C-value enigma and the evolution of eukaryotic genome content. Philosophical Transactions of the Royal Society B: Biological Sciences [Internet]. 2015 Sep 26;370(1678):20140331. Available from: https://royalsocietypublishing.org/doi/10.1098/rstb.2014.0331
3. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol. 2012;10(6):417–30.
4. Puttick MN, Clark J, Donoghue PCJ. Size is not everything: rates of genome size evolution, not C -value, correlate with speciation in angiosperms. Proceedings of the Royal Society B: Biological Sciences [Internet]. 2015 Dec 7;282(1820):20152289. Available from: https://royalsocietypublishing.org/doi/10.1098/rspb.2015.2289
5. Lynch M, Conery JS. The Origins of Genome Complexity. Science (1979). 2003;302(5649):1401–4.
6. Lynch M. Complexity myths and the misappropriation of evolutionary theory. Proceedings of the National Academy of Sciences [Internet]. 2025 Jun 10;122(23):1–5. Available from: https://pnas.org/doi/10.1073/pnas.2425772122
7. Lower SS, McGurk MP, Clark AG, Barbash DA. Satellite DNA evolution : old ideas , new approaches. Curr Opin Genet Dev [Internet]. 2018;49:70–8. Available from: https://doi.org/10.1016/j.gde.2018.03.003
8. Plissonneau C, Stürchler A, Croll D. The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. mBio [Internet]. 2016 Nov 2;7(5):1–13. Available from: http://mbio.asm.org/lookup/doi/10.1128/mBio.01231-16
9. Scott AL, Richmond PA, Dowell RD, Selmecki AM. The Influence of Polyploidy on the Evolution of Yeast Grown in a Sub-Optimal Carbon Source. Mol Biol Evol. 2017;34(10):2690–703.
10. Todd RT, Forche A, Selmecki A. Ploidy Variation in Fungi: Polyploidy, Aneuploidy, and Genome Evolution. Heitman J, Stukenbrock EH, editors. Microbiol Spectr [Internet]. 2017 Aug 25;5(4):139–48. Available from: https://journals.asm.org/doi/10.1128/microbiolspec.FUNK-0051-2016
11. Sipos G, Prasanna AN, Walter MC, O'Connor E, Bálint B, Krizsán K, et al. Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nat Ecol Evol [Internet]. 2017;1(12):1931–41. Available from: http://dx.doi.org/10.1038/s41559-017-0347-8
12. Frantzeskakis L, Németh MZ, Barsoum M, Kusch S, Kiss L, Takamatsu S, et al. The Parauncinula polyspora Draft Genome Provides Insights into Patterns of Gene Erosion and Genome Expansion in Powdery Mildew Fungi. mBio. 2019;10(5):10:e01692-19.
13. McClintock B. Induction of Instability at Selected Loci in Maize. Genetics [Internet]. 1953;38(6):579–99. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17247459%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1209627

14. Fedoroff N, Wessler S, Shure M. Isolation of the transposable maize controlling elements Ac and Ds. Cell. 1983;35(1):235–42.
15. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8(12):973–82.
16. Wells JN, Feschotte C. A Field Guide to Transposable Elements. Annu Rev Genet. 2020;54:7–34.
17. Wessler S. LTR-retrotransposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev [Internet]. 1995 Dec;5(6):814–21. Available from: https://linkinghub.elsevier.com/retrieve/pii/0959437X9580016X
18. Hill R, Smith D, Canning G, Grey M, Hammond-Kosack KE, McMullan M. Starship giant transposable elements cluster by host taxonomy using k -mer-based phylogenetics. Rokas A, editor. G3: Genes, Genomes, Genetics [Internet]. 2025 Jun 4;15(6). Available from: https://academic.oup.com/g3journal/article/doi/10.1093/g3journal/jkaf082/8110972
19. Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet [Internet]. 2002 May;3(5):370–9. Available from: https://www.nature.com/articles/nrg798
20. Castanera R, Vendrell-Mir P, Bardil A, Carpentier M, Panaud O, Casacuberta JM. Amplification dynamics of miniature inverted-repeat transposable elements and their impact on rice trait variability. The Plant Journal [Internet]. 2021 Jul 31;107(1):118–35. Available from: https://onlinelibrary.wiley.com/doi/10.1111/tpj.15277
21. Peter M, Kohler A, Ohm RA, Kuo A, Krützmann J, Morin E, et al. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus Cenococcum geophilum. Nat Commun [Internet]. 2016 Nov 7;7(1):1–15. Available from: http://www.nature.com/articles/ncomms12662
22. Fouché S, Oggenfuss U, McDonald BA, Croll D. Recurrent chromosome destabilization through repeat-mediated rearrangements in a fungal pathogen [Internet]. bioRxiv. 2023. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.07.14.549097
23. Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 2002;12(7):1075–9.
24. Schrader L, Schmitz J. The impact of transposable elements in adaptive evolution. Mol Ecol [Internet]. 2019 Mar 4;28(6):1537–49. Available from: https://onlinelibrary.wiley.com/doi/10.1111/mec.14794
25. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet. 2007;8(4):272–85.
26. Bewick AJ, Hofmeister BT, Powers RA, Mondo SJ, Grigoriev I V, James TY, et al. Diversity of cytosine methylation across the fungal tree of life. Nat Ecol Evol [Internet]. 2019 Mar 18;3(3):479–90. Available from: http://dx.doi.org/10.1038/s41559-019-0810-9
27. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res [Internet]. 2011;21(3):381–95. Available from: http://dx.doi.org/10.1038/cr.2011.22
28. Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. Arkhipova I, editor. Mol Biol Evol [Internet]. 2020 Jan 1;37(1):221–39. Available from: https://academic.oup.com/mbe/article/37/1/221/5573762
29. Galagan JE, Selker EU. RIP: the evolutionary cost of genome defense. Trends in Genetics [Internet]. 2004;20(9):417–23. Available from: http://ac.els-

cdn.com/S0168952504001878/1-s2.0-S0168952504001878-main.pdf?_tid=38a51930-2f11-11e7-9d3b-00000aab0f02&acdnat=1493713847_37be5e43095b9599eb20f3958d0dc3ac

30. Krishnan P, Meile L, Plissonneau C, Ma X, Hartmann FE, Croll D, et al. Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. BMC Biol. 2018;16(1):1–18.

31. Liu X, Wang X, Zhou F, Xue Y, Liu C. Genomic insights into Penicillium chrysogenum adaptation to subseafloor sedimentary environments. BMC Genomics [Internet]. 2024 Jan 2;25(1):4. Available from: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-023-09921-1

32. Wong VL, Ellison CE, Eisen MB, Pachter L, Brem RB. Structural Variation among Wild and Industrial Strains of Penicillium chrysogenum. Moreno-Hagelsieb G, editor. PLoS One [Internet]. 2014 May 13;9(5):e96784. Available from: https://dx.plos.org/10.1371/journal.pone.0096784

33. Agrawal A, Eastman QM, Schatz DG. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature [Internet]. 1998 Aug;394(6695):744–51. Available from: http://link.springer.com/10.1007/BF02786485

34. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. Science (1979) [Internet]. 2021 Feb 19;371(6531):eabc6405. Available from: https://www.sciencemag.org/lookup/doi/10.1126/science.abc6405

35. Groenewald JZ, Chen YY, Zhang Y, Roux J, Shin H -D., Shivas RG, et al. Species diversity in Pseudocercospora. Fungal Syst Evol [Internet]. 2024 Jun 30;13(1):29–89. Available from: https://www.ingentaconnect.com/content/10.3114/fuse.2024.13.03

36. Liang C, Jayawardena RS, Zhang W, Wang X, Liu M, Liu L, et al. Identification and Characterization of Pseudocercospora Species Causing Grapevine Leaf Spot in China. Journal of Phytopathology [Internet]. 2016 Feb 1;164(2):75–85. Available from: https://onlinelibrary.wiley.com/doi/10.1111/jph.12427

37. Crous PW, Braun U, Hunter GC, Wingfield MJ, Verkley GJM, Shin HD, et al. Phylogenetic lineages in Pseudocercospora. Stud Mycol [Internet]. 2013 Jun;75:37–114. Available from: https://www.ingentaconnect.com/content/10.3114/sim0005

38. Friesen TL. Combating the Sigatoka Disease Complex on Banana. McDowell JM, editor. PLoS Genet [Internet]. 2016 Aug 11;12(8):e1006234. Available from: https://dx.plos.org/10.1371/journal.pgen.1006234

39. Guyot J, Le Guen V. A Review of a Century of Studies on South American Leaf Blight of the Rubber Tree. Plant Dis [Internet]. 2018 Jun 1;102(6):1052–65. Available from: https://apsjournals.apsnet.org/doi/10.1094/PDIS-04-17-0592-FE

40. Chang TC, Salvucci A, Crous PW, Stergiopoulos I. Comparative Genomics of the Sigatoka Disease Complex on Banana Suggests a Link between Parallel Evolutionary Changes in Pseudocercospora fijiensis and Pseudocercospora eumusae and Increased Virulence on the Banana Host. Hane JK, editor. PLoS Genet [Internet]. 2016 Aug 11;12(8):e1005904. Available from: https://dx.plos.org/10.1371/journal.pgen.1005904

41. Arango Isaza RE, Diaz-Trujillo C, Dhillon B, Aerts A, Carlier J, Crane CF, et al. Combating a Global Threat to a Clonal Crop: Banana Black Sigatoka Pathogen Pseudocercospora fijiensis (Synonym Mycosphaerella fijiensis) Genomes Reveal Clues

for Disease Control. McDowell JM, editor. PLoS Genet [Internet]. 2016 Aug 11;12(8):e1005876. Available from: https://dx.plos.org/10.1371/journal.pgen.1005876

42. Akinsanmi OA, Carvalhais LC. Draft Genome of the Macadamia Husk Spot Pathogen, Pseudocercospora macadamiae. Phytopathology [Internet]. 2020 Sep 1;110(9):1503–6. Available from: https://apsjournals.apsnet.org/doi/10.1094/PHYTO-12-19-0460-A

43. Zaccaron AZ, Stergiopoulos I. First Draft Genome Resource for the Tomato Black Leaf Mold Pathogen Pseudocercospora fuligena. Molecular Plant-Microbe Interactions® [Internet]. 2020 Dec 1;33(12):1441–5. Available from: https://apsjournals.apsnet.org/doi/10.1094/MPMI-06-20-0139-A

44. González Sáyer SM, Oggenfuss U, García I, Aristizabal F, Croll D, Riaño-Pachon DM. High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees. Genet Mol Biol [Internet]. 2022;45(1):1–5. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-47572022000100802&tlng=en

45. Sinha S, Navathe S, Anjali, Vishwakarma S, Prajapati P, Chand R, et al. Whole genome sequencing and annotation of Pseudocercospora abelmoschi, a causal agent of black leaf mould of okra. World J Microbiol Biotechnol [Internet]. 2025 May 15;41(5):174. Available from: https://link.springer.com/10.1007/s11274-025-04398-4

46. Santana MF, Silva JC, Batista AD, Ribeiro LE, da Silva GF, de Araújo EF, et al. Abundance, distribution and potential impact of transposable elements in the genome of Mycosphaerella fijiensis. BMC Genomics. 2012;13(1):1–11.

47. Dhillon B, Kema GH, Hamelin RC, Bluhm BH, Goodwin SB. Variable genome evolution in fungi after transposon-mediated amplification of a housekeeping gene. Mob DNA [Internet]. 2019 Dec 27;10(1):37. Available from: http://dx.doi.org/10.1101/550798

48. Chen Q, Bakhshi M, Balci Y, Broders KD, Cheewangkoon R, Chen SF, et al. Genera of phytopathogenic fungi: GOPHY 4. Stud Mycol [Internet]. 2022 Mar 1;101(1):417–564. Available from: https://www.ingentaconnect.com/content/10.3114/sim.2022.101.06

49. Crous PW, Wingfield MJ, Cheewangkoon R, Carnegie AJ, Burgess TI, Summerell BA, et al. Foliar pathogens of eucalypts. Stud Mycol [Internet]. 2019 Sep 1;94(1):125–298. Available from: https://www.ingentaconnect.com/content/10.1016/j.simyco.2019.08.001

50. Wei K, Aldaimalani R, Mai D, Zinshteyn D, PRV S, Blumenstiel JP, et al. Rethinking the "gypsy" retrotransposon: A roadmap for community-driven reconsideration of problematic gene names. OSFpreprints [Internet]. 2022;10.31219/o. Available from: https://osf.io/fma57/

51. Baril T, Croll D. A pangenome-guided manually curated library of transposable elements for Zymoseptoria tritici. BMC Res Notes [Internet]. 2023;16(1):23–6. Available from: https://doi.org/10.1186/s13104-023-06613-7

52. Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biol [Internet]. 2020 Dec 11;18(1):12. Available from: http://biorxiv.org/content/early/2019/10/13/803098.abstract

53. Miyauchi S, Kiss E, Kuo A, Drula E, Kohler A, Sánchez-García M, et al. Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits. Nat Commun. 2020;11(1):1–17.

54. Castanera R, Borgognone A, Pisabarro AG, Ramírez L. Biology, dynamics, and applications of transposable elements in basidiomycete fungi. Appl Microbiol Biotechnol

[Internet]. 2017 Feb 10;101(4):1337–50. Available from: http://dx.doi.org/10.1007/s00253-017-8097-8

55. Tavares S, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, et al. Genome size analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci [Internet]. 2014 Aug 26;5(AUG). Available from: http://journal.frontiersin.org/article/10.3389/fpls.2014.00422/abstract

56. Aime MC, McTaggart AR, Mondo SJ, Duplessis S. Phylogenetics and Phylogenomics of Rust Fungi. In: Advances in Genetics [Internet]. Academic Press Inc.; 2017. p. 267–307. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0065266017300391

57. Murat C, Payen T, Noel B, Kuo A, Morin E, Chen J, et al. Pezizomycetes genomes reveal the molecular basis of ectomycorrhizal truffle lifestyle. Nat Ecol Evol [Internet]. 2018 Nov 12;2(12):1956–65. Available from: https://www.nature.com/articles/s41559-018-0710-4

58. van Wyk S, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. Front Microbiol [Internet]. 2021 Feb 1;11(February). Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.622368/full

59. Miousse IR, Chalbot MCG, Lumen A, Ferguson A, Kavouras IG, Koturbash I. Response of transposable elements to environmental stressors. Mutat Res Rev Mutat Res [Internet]. 2015;765:19–39. Available from: http://dx.doi.org/10.1016/j.mrrev.2015.05.003

60. Van Parijs J, Broekaert WF, Goldstein IJ, Peumans WJ. Hevein: an antifungal protein from rubber-tree (Hevea brasiliensis) latex. Planta [Internet]. 1991 Jan;183(2):258–64. Available from: http://link.springer.com/10.1007/BF00197797

61. Mei S, Hou S, Cui H, Feng F, Rong W. Characterization of the interaction between Oidium heveae and Arabidopsis thaliana. J Dig Dis. 2016;17(9):1331–43.

62. Mazlan S, Md JAAFAR N, Wahab A, Sulaiman Z, Rajandas H, Zulperi D. Major Diseases of Rubber (Hevea brasiliensis) in Malaysia. 2019;5(2):10–21. Available from: http://www.pjsrr.upm.edu.my/

63. Evangelisti E, Gogleva A, Hainaux T, Doumane M, Tulin F, Quan C, et al. Time-resolved dual transcriptomics reveal early induced Nicotiana benthamiana root genes and conserved infection-promoting Phytophthora palmivora effectors. BMC Biol. 2017 May 11;15(1).

64. Hsieh DK, Chuang SC, Chen CY, Chao YT, Lu MYJ, Lee MH, et al. Comparative Genomics of Three Colletotrichum scovillei Strains and Genetic Analysis Revealed Genes Involved in Fungal Growth and Virulence on Chili Pepper. Front Microbiol [Internet]. 2022 Jan 27;13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.818291/full

65. Ali SS, Asman A, Shao J, Balidion JF, Strem MD, Puig AS, et al. Genome and transcriptome analysis of the latent pathogen Lasiodiplodia theobromae, an emerging threat to the cacao industry. Genome. 2020;63(1):37–52.

66. Longsaward R, Viboonjun U, Wen Z, Asiegbu FO. In silico analysis of secreted effectorome of the rubber tree pathogen Rigidoporus microporus highlights its potential virulence proteins. Front Microbiol. 2024;15.

67. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

68. Vaghefi N, Kusch S, Németh MZ, Seress D, Braun U, Takamatsu S, et al. Beyond Nuclear Ribosomal DNA Sequences: Evolution, Taxonomy, and Closest Known Saprobic Relatives of Powdery Mildew Fungi (Erysiphaceae) Inferred From Their First Comprehensive Genome-Scale Phylogenetic Analyses. Front Microbiol [Internet]. 2022 Jun 9;13. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2022.903024/full
69. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol [Internet]. 2012;19(5):455–77. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22506599%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC3342519
70. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics [Internet]. 2013 Apr 15;29(8):1072–5. Available from: https://academic.oup.com/bioinformatics/article/29/8/1072/228832
71. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.
72. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–80.
73. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530–4.
74. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics [Internet]. 2018 May 8;19(S6):153. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2129-y
75. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics [Internet]. 2008 Mar 1;24(5):637–44. Available from: https://academic.oup.com/bioinformatics/article/24/5/637/202844
76. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Res [Internet]. 2020 Apr 28;9:304. Available from: https://f1000research.com/articles/9-304/v1
77. Jones DAB, Rozano L, Debler JW, Mancera RL, Moolhuijzen PM, Hane JK. An automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant pathogens. Sci Rep [Internet]. 2021 Oct 5;11(1):19731. Available from: https://www.nature.com/articles/s41598-021-99363-0
78. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46(W1):W95–101.
79. Blin K, Shaw S, Kautsar SA, Medema MH, Weber T. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Res [Internet]. 2021 Jan 8;49(D1):D639–43. Available from: https://academic.oup.com/nar/article/49/D1/D639/5957162
80. Goubert C, Modolo L, Vieira C, ValienteMoro C, Mavingui P, Boulesteix M. De Novo Assembly and Annotation of the Asian Tiger Mosquito (Aedes albopictus) Repeatome with dnaPipeTE from Raw Genomic Reads and Comparative Analysis with the Yellow

Fever Mosquito (Aedes aegypti). Genome Biol Evol [Internet]. 2015 Apr 1;7(4):1192–205. Available from: https://academic.oup.com/gbe/article-lookup/doi/10.1093/gbe/evv050

81. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12(1):1–14.
82. Gasparotto L, Clério J, Pereira R, Técnicos E. Doenças da seringueira no Brasil 2ª edição revista e atualizada [Internet]. 2012. Available from: www.embrapa.br/liv
83. Lieberei R. South American Leaf Blight of the Rubber Tree (Hevea spp.): New Steps in Plant Domestication using Physiological Features and Molecular Markers. Ann Bot [Internet]. 2007 Sep 19;100(6):1125–42. Available from: https://academic.oup.com/aob/article-lookup/doi/10.1093/aob/mcm133
84. Stirling D. DNA extraction from fungi, Yeast and Bacteria. In: Stirling D, Barlet J, editors. PCR protocols Method in Molecular Biology. 2004. p. 53–4.
85. Chen S. Ultrafast one-pass FASTQ data preprocessing, quality control, and deduplication using fastp. iMeta [Internet]. 2023 May 8;2(2). Available from: https://onlinelibrary.wiley.com/doi/10.1002/imt2.107
86. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM [Internet]. 2013. Available from: http://github.com/lh3/bwa.
87. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.
88. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience [Internet]. 2015 Dec 1;4(1):1–16. Available from: https://academic.oup.com/gigascience/article/doi/10.1186/s13742-015-0047-8/2707533
89. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2022;
90. Baril T, Galbraith J, Hayward A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. Arkhipova I, editor. Mol Biol Evol [Internet]. 2024 Apr 2;41(4). Available from: https://academic.oup.com/mbe/article/doi/10.1093/molbev/msae068/7635926
91. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci U S A. 2020;117(17):9451–7.
92. Xu Z, Wang H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35(SUPPL.2):265–8.
93. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.
94. Breen J, Wicker T, Kong X, Zhang J, Ma W, Paux E, et al. A highly conserved gene island of three genes on chromosome 3B of hexaploid wheat: Diverse gene function and genomic structure maintained in a tightly linked block. BMC Plant Biol. 2010;10.
95. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–402.

96. G Higgins D, M Sharp P. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988;73(1):237–44.
97. Crescente JM, Zavallo D, Helguera M, Vanzetti LS. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics. 2018;19(1):1–10.
98. Mao H, Wang H. SINE-scan: An efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics. 2017;33(5):743–5.
99. Wenke T, Dobel T, Sorensen TR, Junghans H, Weisshaar B, Schmidt T. Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes. the Plant Cell Online [Internet]. 2011;23(9):3117–28. Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.111.088682
100. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25(15):1972–3.
101. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17(4):540–52.
102. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312–3.
103. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. R package version 1.0.0. 2020.
104. Paradis E, Schliep K. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics. 2019;35(3):526–8.
105. Wang LG, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol Biol Evol. 2020;37(2):599–603.
106. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol Evol. 2017;8(1):28–36.
107. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics. 2000;16(6):276–7.
108. van Wyk S, Harrison CH, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. The RIPper, a web-based tool for genome-wide quantification of Repeat-Induced Point (RIP) mutations. PeerJ [Internet]. 2019;7:e7447. Available from: https://peerj.com/articles/7447
109. Krzywinski M, Schein J, Birol İ, Connors J, Gascoyne R, Horsman D, et al. Circos: An information aesthetic for comparative genomics. Genome Res [Internet]. 2009 Sep;19(9):1639–45. Available from: http://genome.cshlp.org/lookup/doi/10.1101/gr.092759.109
110. Nelson MG, Linheiro RS, Bergman CM. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3; Genes|Genomes|Genetics [Internet]. 2017 Aug;7(8):2763–78. Available from: http://g3journal.org/lookup/doi/10.1534/g3.117.043893

Supplementary Tables

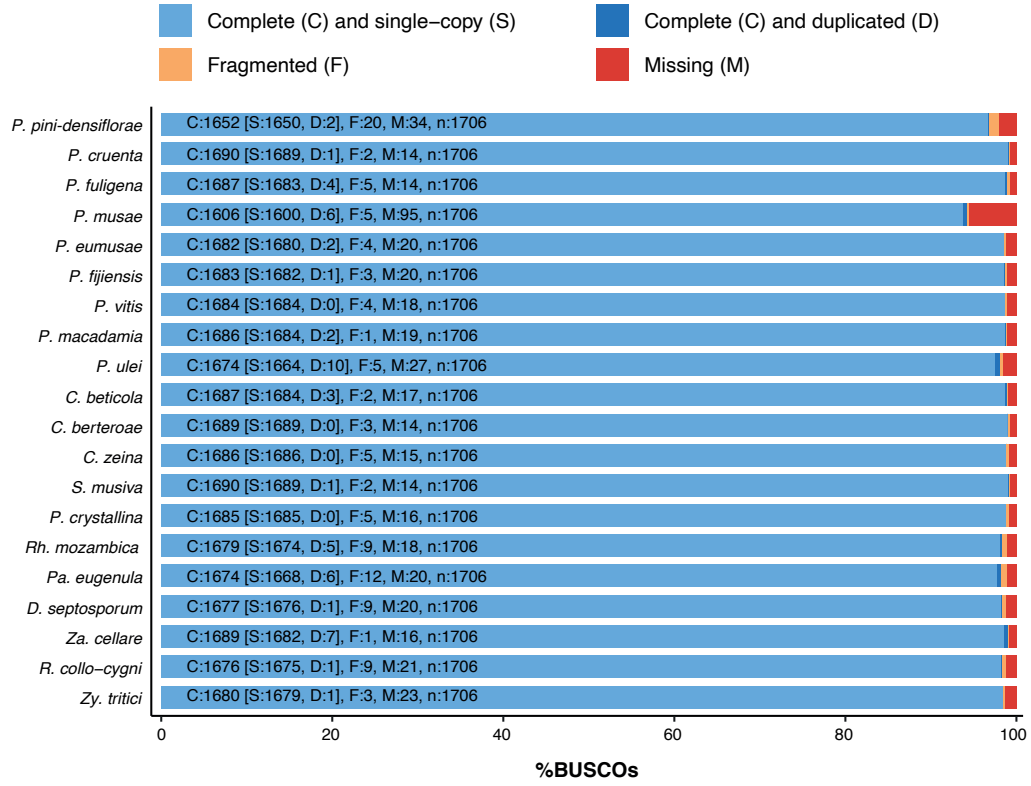
(see separate file)

Supplementary Table S1: Accession numbers and data resources for the genomes analyzed in this study.

Supplementary Table S2: *Pseudocercospora ulei* isolates analyzed in this study.

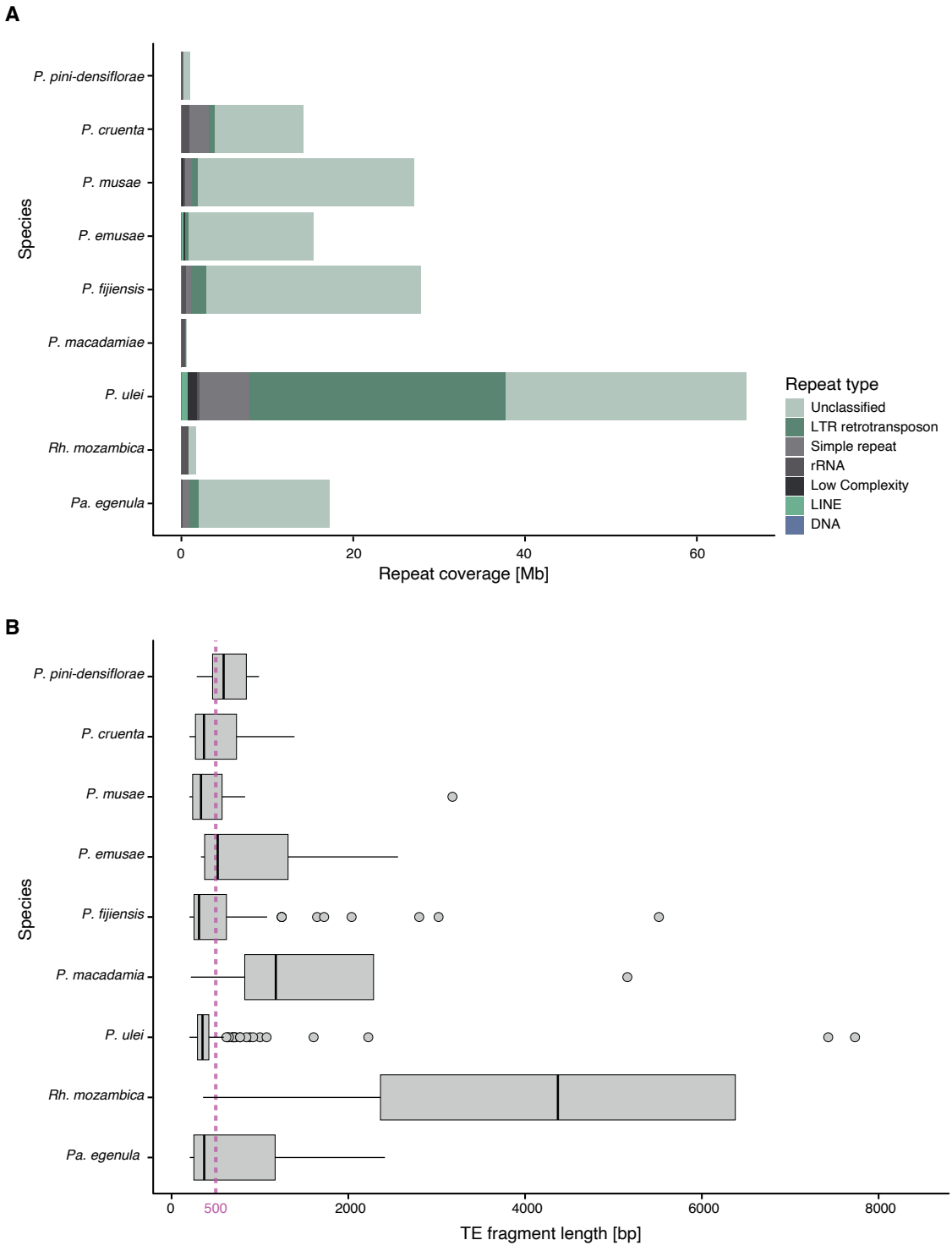
1016 **Supplementary Figures**

1017

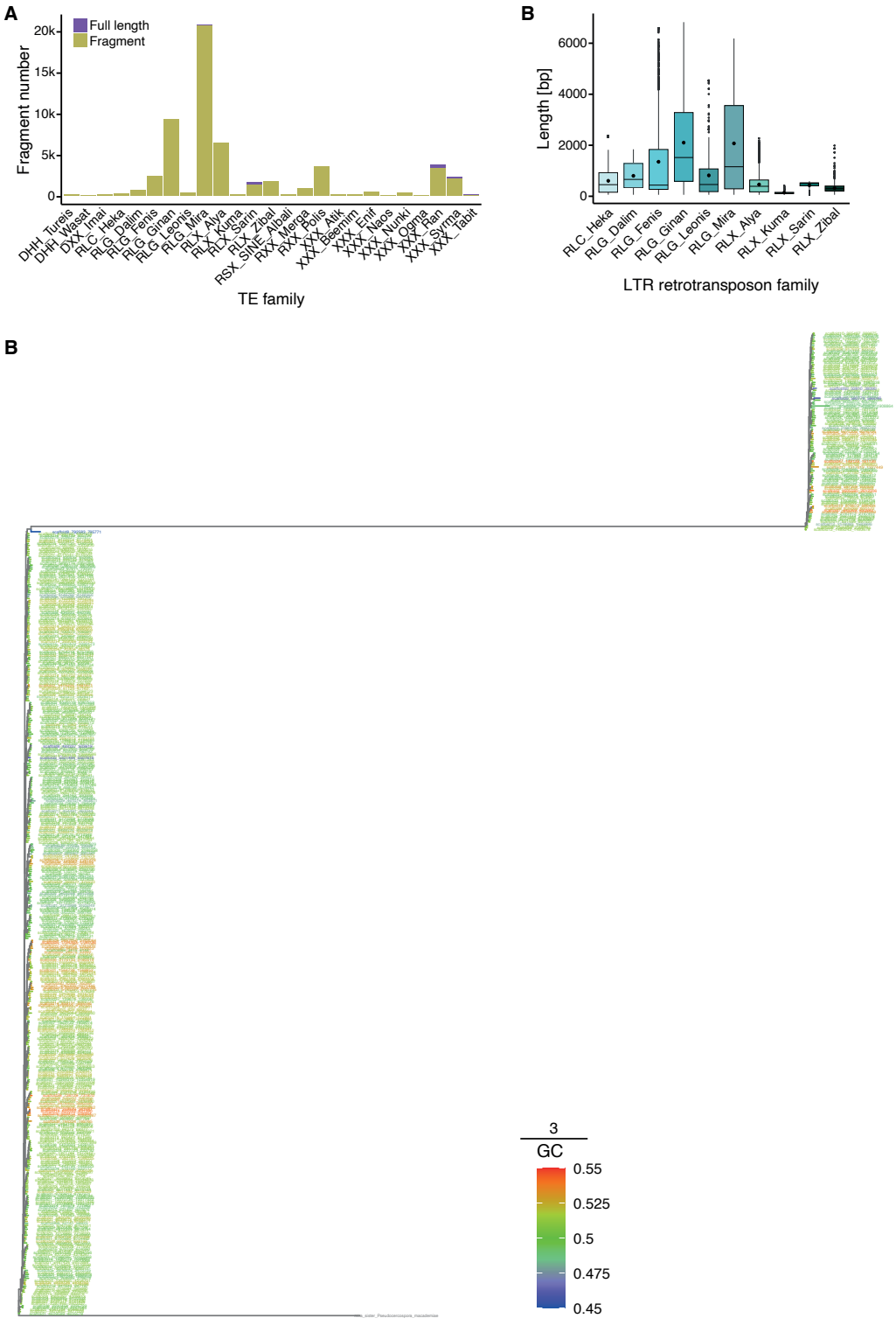


1018

1019 **Supplementary Figure S1**. Detailed BUSCO (Benchmarking Universal Single-Copy Orthologs) scores
1020 for Mycosphaerellaceae genome assemblies. 1706 BUSCO orthologs from the ascomycota_odb10 database
1021 were analyzed, and the complete (single copy or duplicated), fragmented and missing orthologs were listed.
1022



Supplementary Figure S2. Repeat element lengths in the seven strains with the most expanded genome size of the *Pseudocercospora* genus and two closely related species. A) The total length of repeats in each genome. Colors indicate the type of repeat, with green and blue colors indicating TEs. B) Length distribution of TE fragments per genome. The mauve line indicates 500 bp.



Supplementary Figure S3. Manually curated TE families in *P. ulei*. A) Copy numbers of TE families. The color indicates if the length of the TE copy was the same (or shorter by >20%) as the consensus. B) Length distribution of TEs for retrotransposons. C) Phylogenetic tree for the coding regions of the RLG_Mira family. The color indicates the GC content, with a low GC content (blue) indicating a potential impact of RIP.