Evolutionary dynamics of transposable elements in *Magnaporthe oryzae* reveal evidence of genomic transfer and key differences between rice and wheat blast pathotypes

Anne A. Nakamoto*, Pierre M. Joubert*, and Ksenia V. Krasileva

*These authors contributed equally

Abstract

Fungal plant pathogens pose a significant threat to biodiversity and food security worldwide. This threat is aggravated by their rapidly evolving genomes that adapt to evade host plant defenses, reducing the efficacy of deployed resistant crops. *Magnaporthe oryzae* infects rice, wheat, and many other grasses, resulting in significant crop losses each year. Transposable elements (TEs) are hypothesized to be involved in the evolution and rapid adaptation of *M. oryzae*. However, there is still much to understand about how these elements behave in different *M. oryzae* host-specific lineages. In this study, we completed the annotation and phylogenetic classification of TEs in five lineages of *M. oryzae*. We identified differences in TE content between these lineages, and showed that recent lineage-specific expansions of certain TEs have contributed to greater TE content in rice-infecting and *Setaria*-infecting lineages. We reconstructed the histories of LTR-retrotransposon expansions and found them to have experienced complex dynamics, where some were caused by the proliferation of one element, while others consisted of multiple elements from an older population of TEs that proliferated in parallel. Additionally, we found evidence suggesting the recent transfer of a DNA transposon between rice and wheat *M. oryzae* lineages, and a region showing evidence of recombination between those lineages, which could have facilitated such a transfer. These results point towards key differences in TE dynamics, evolutionary history, and adaptive potential between the rice-and-*Setaria*-infecting and the wheat-*Lolium*-and-*Eleusine*-infecting lineage groups.

Introduction

Magnaporthe oryzae, a causal agent of the blast disease, results in significant crop losses worldwide. Each year, 30% of rice crops are lost to this disease (1), making it one of the most important plant pathogens (2). *M. oryzae* has a wide host range, infecting many grasses including other important crops such as wheat and millet. It has been shown that *M. oryzae* is composed of many lineages that are specific to grass hosts including the *Oryza*, *Setaria*, *Triticum*, *Lolium*, and *Eleusine* genera (3). These lineages are recently diverged, with the *Oryza*-infecting *M. oryzae* (MoO) and *Setaria*-infecting (MoS) lineages having diverged from their common ancestor approximately 9,800 years ago (4), around the time of rice domestication (5). Wheat blast has recently become a major threat given its rapid emergence, spread, and particularly devastating disease symptoms (6). It was first discovered in Brazil in 1985, and has since spread to Bangladesh in 2016 and Zambia in 2018 (6). Previously, the *Triticum*-infecting lineage (MoT) was thought to have arisen via a host shift of *Lolium*-infecting *M. oryzae* (MoL) to wheat (7). However, a recent study has found that the recombination of standing variation in a multi-hybrid swarm of host-specific isolates, including an *Eleusine*-infecting (MoE) isolate and a relative

of MoO and MoS, likely gave rise to MoT and MoL within the past 60 years (8). Clearly, *M. oryzae* is a very successful pathogen that can infect many hosts and is actively evolving to infect new ones. However, there is still much to understand about the mechanisms that enable this success.

In addition to infecting many hosts, *M. oryzae* can quickly adapt to its host plant and overcome its defenses. For example, it often takes the pathogen only a few years to overcome newly introduced resistance genes in the field (9). Achieving durable resistance may, therefore, require fundamental understanding of the diversity generating mechanisms that allow the pathogen's rapid adaptation. Transposable elements (TEs) are hypothesized to generate genomic diversity due to their mobile and repetitive nature. Previous studies on TEs in *M. oryzae* have shown that they can have major effects on its genome and host specificity. For example, the *POT2* DNA transposon has been shown to insert into the *AVR-Pib* effector gene of rice blast isolates, allowing them to overcome resistance conferred by the *Pib* gene in rice (10). Additionally, the host jump from MoL to MoT was hypothesized to have occurred by selection for the loss of the *PWT3* effector, which is recognized by the *Rwt3* resistance gene in wheat (11). Insertions of the *MGR583* retrotransposon and *POT3* DNA transposon were found to cause the functional loss of *PWT3*, though recombination may have played a role in this loss as well (8). A notable way in which TEs can indirectly generate diversity is through repeat induced point mutation (RIP), a mutagenic mechanism that targets TEs and causes GC to TA mutations (12). RIP is only active during sexual reproduction (13), and previous studies have reported that it is minimally active in *M. oryzae* given its largely clonal life cycle (12–14). However, there is evidence that RIP occurred during the swarm event that formed MoT and MoL (8), and could have contributed to the progenies' success. Finally, recent studies have shown that *M. oryzae* experiences frequent gene gains and losses (15), and has multiple non-canonical DNA repair pathways (16) that could contribute to its rapid adaptation. *M. oryzae* has also been found to produce many extrachromosomal circular DNAs (17), which have been shown to have great adaptive potential (18). TEs are thought to be involved either directly or indirectly in all of these processes. Therefore, understanding TE dynamics and activity in *M. oryzae* is likely crucial to understanding how it is able to quickly adapt to its hosts.

To this aim, we assembled an unbiased library of TEs to produce robust annotations in each lineage. This allowed us to identify differences in TE content between the lineages and characterize TE dynamics. We observed that recent lineage-specific expansions of LTR-retrotransposons have contributed to the greater TE content in MoO and MoS. The histories and dynamics of these expansions were complex, with some having been caused by the proliferation of one element, while others consisted of multiple elements from an older population of TEs that proliferated in parallel. Additionally, we found evidence suggesting the recent transfer of a DNA transposon between rice and wheat *M. oryzae* lineages, and a potential region of recombination between those lineages that could have facilitated such a transfer. These results point towards key differences between the MoO-MoS and MoT-MoL-MoE lineage groups, suggesting unique genomic features in their adaptation.

Results

TE content in *M. oryzae* varies greatly across lineages and isolates

To analyze the TE content of all *M. oryzae* lineages, we first constructed a TE element library representative of all lineages. Since highly contiguous genome assemblies provide the most complete view of TE content (19), a set of *M. oryzae* genomes with the lowest number of scaffolds was gathered from NCBI GenBank for each lineage (Table S1). We then constructed a pipeline based on previous methods (20) to annotate TEs. The pipeline (Figure S1) utilized one representative genome of the highest quality from each lineage to perform *de novo* repeat annotation. We then added the RepBase (21) library of known TEs in fungi, and filtered the resulting library against a list of TE-associated domains (20). This ensured that only complete elements with the potential to be active were kept. Elements in the library were further refined by manual classification using protein domain-based phylogenies (Figure S1). Elements that formed a subclade with a known RepBase TE were classified as being the same element. This resulted in the classification of many *de novo* elements as known elements (*Ty3_MAG1, Ty3_MAG2, Grasshopper, MAG_Ty3, MGRL3, PYRET, MGR583, MoTeR1,* and *POT2*), or as new elements part of a known TE family (*Copia_elem* and *TcMar_elem*) (Figure 1, Table S2). TEs that did not group in a subclade containing a known TE were classified as 'unknown'. We then used the classified library to annotate TEs in the larger set of *M. oryzae* genomes (Table S1), and verified that each hit contained a domain associated with TE activity. This approach provided us with high-quality and unbiased copy number and positional information of near full-length TEs in each genome.

Using our TE annotations,we observed striking differences in TE content between genomes of different lineages, and these differences seemed to follow the evolutionary relationships between lineages (Figure 1, Figure S2). Most notably, MoO and MoS contained much higher TE content than MoT, MoL, and MoE (Figure 1A). In MoO and MoS, an average of 11.14% (5.1 Mbp) of the genome consisted of annotated TEs, while the average was 5.44% (2.4 Mbp) for the other three lineages (Table S1, Figure 1A). The *Magnaporthe grisea* isolate that was used as an outgroup had very similar TE content to the MoT, MoL, and MoE isolates, which suggested that the MoO-MoS clade may have acquired its higher TE content after its split from the other lineages (Figure 1A). The differences were not caused by genome duplication, since genome size across all isolates is relatively similar (Table S1). The results were also not due to assembly quality or completeness, as genomes with varying BUSCO scores all followed the trend of higher TE content in MoO and MoS (Table S1). Finally, we observed differences in the relative proportions of certain annotated elements, such as *MAG_Ty3* and *Copia_elem*, across the lineages (Figure 1B). This result hinted at complex lineage-specific TE dynamics, rather than genome-wide contraction or expansion of TE content.

While there was an overall greater number of TEs in MoO and MoS, some elements were more prevalent in the other lineages or in individual genomes. The *Grasshopper* LTR-retrotransposon made up a large portion of TE content in the MoE MZ5-1-6 genome specifically, but less in the other MoE, MoT, and MoL genomes, and was absent in MoO and MoS. Additionally, the MoTeR element had greater copy number in MoT's B71 and BR32, and MoE's CD156, but less in the other MoT, MoL, and MoE genomes, and was also absent from MoO and MoS (Figure 1B). This indicated that although MoT, MoL, and MoE have lower TE content, they may be more prone to isolate specific TE dynamics, in contrast to the greater TE content in MoO and MoS which is more uniform.

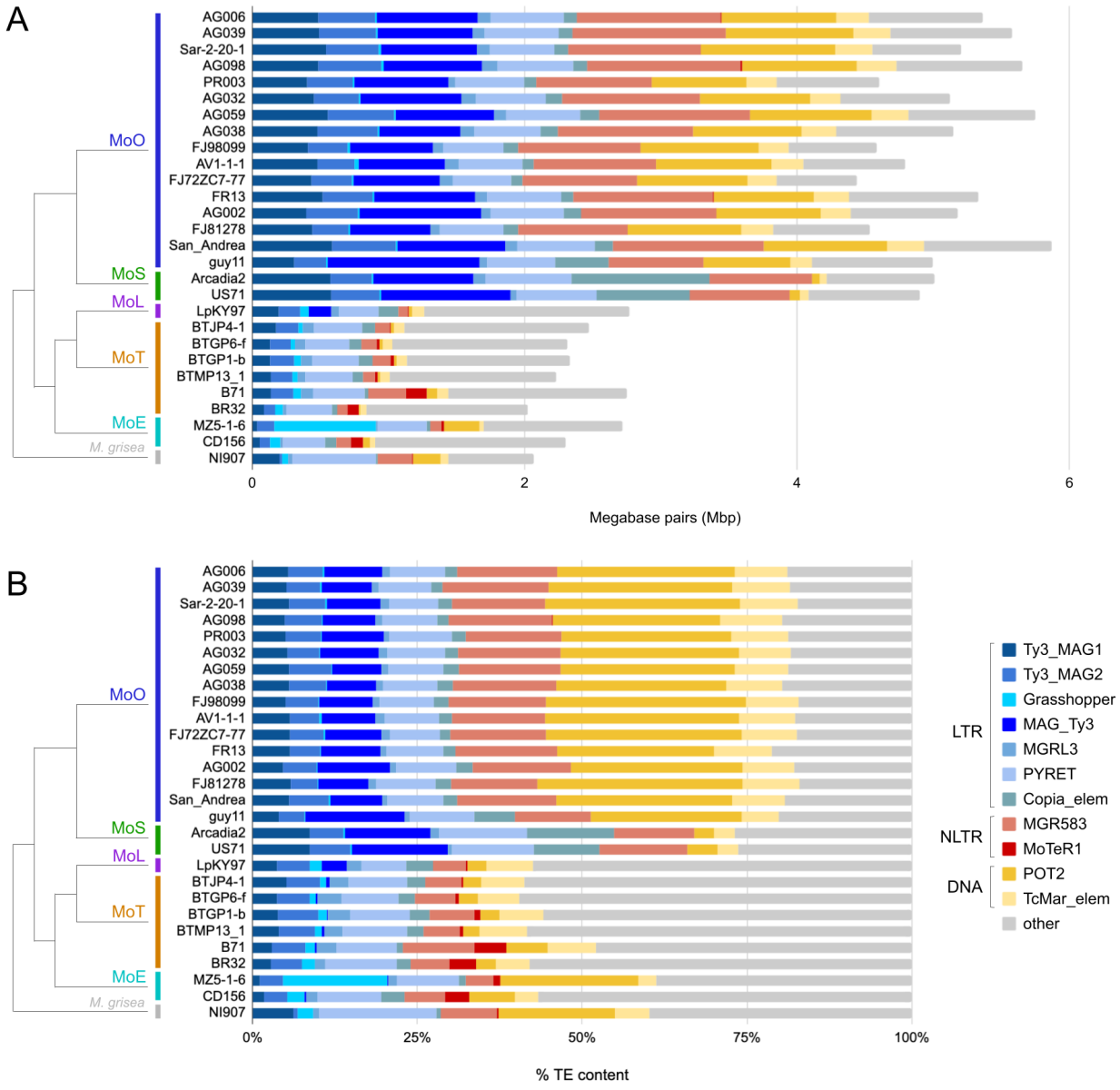


Figure 1: Differences in TE content between *M. oryzae* genomes of different lineages. **A,** Stacked bar plot showing the number of base pairs (bp) each TE occupies in each genome. **B,** Stacked bar plot showing the percentage that each individual TE makes up out of all TEs in each genome. At the left of both plots is shown what lineage each genome belongs to, and the evolutionary relationships between lineages (3) (branch lengths not to scale). Names of the individual TEs and their classification are shown in the key. LTR = long terminal repeat retrotransposon, NLTR = non-LTR retrotransposon, DNA = DNA transposon.

Recent lineage-specific expansions of LTR-retrotransposons led to differences in TE content between clades of *M. oryzae*

We next tested whether genome-wide or TE specific contraction or expansion dynamics led to the differences in TE content across lineages. Given that the bulk of the difference between MoO-MoS and

other lineages seemed to be explained by LTR retrotransposons (Figure 1B), we focused on these elements. We constructed domain-based maximum-likelihood (ML) phylogenies (Figure 2B-E, Figure S3) for each of the seven LTR-retrotransposons using the highest quality representative genome from each lineage. The trees were compared to the genome tree (Figure 2A), which was generated based on the alignment of single copy orthologous genes (SCOs) in order to identify recent lineage-specific expansions of these elements. Based on our analysis, three LTR-retrotransposons stood out as having experienced such lineage specific expansions. *MAG_Ty3* showed a large expansion in MoO and a smaller expansion in MoS (Figure 2B). *Copia_elem* expanded in both MoO and MoS (Figure 2C), and *Grasshopper* expanded only in the MoE MZ5-1-6 genome (Figure 2D). A helpful point of comparison was the *MGRL3* LTR retrotransposon, which was present at low copy number in all of the lineages. Elements in the *MGRL3* phylogeny didn't strictly group by lineage and were more interleaved (Figure 2E), which suggested that it experienced an older expansion before the *M. oryzae* lineages diverged and has not expanded recently. The other LTR-retrotransposons phylogenies of *Ty3_MAG1*, *Ty3_MAG2*, and *PYRET* (Figure S3) also resembled older proliferations of these elements. This shows that a genome-wide deregulation of TEs was likely not responsible for the higher TE content in MoO and MoS, but rather resulted from element-specific dynamics.

We then wanted to validate our hypothesis that *Grasshopper*, *MAG_Ty3*, and *Copia_elem* had experienced lineage-specific expansions that occurred after all *M. oryzae* lineages had diverged, as opposed to expansions in all lineages and subsequent losses in some. We used the presence of solo LTRs to address this possibility, which are often left behind when LTR retrotransposons are excised from the genome (22). So, a large number of solo LTRs and few full elements would suggest recent contraction of an LTR-retrotransposon population, while few solo LTRs and many full elements suggest a recent expansion. We observed that lineage-specific expansions are largely responsible for LTR-retrotransposon copy number variation, rather than removal of these elements from some lineages (FIgure 2F). *MAG_Ty3* had less than 13 solo LTRs present in each of the MoL, MoT, and MoE lineages, whose removal could not account for the 227 and 188 full-length *MAG_Ty3* in MoO and MoS, respectively (Figure 2F). Thus, *MAG_Ty3's* higher copy number in MoO and MoS was likely due to expansion in those lineages only. The Copia element had a lot more solo LTRs present in all genomes (>150), however there were still many more in MoO and MoS (>300) (Figure 2F). This suggests that older expansions of the Copia element may have occurred before the divergence of the lineages and were partially removed, however expansions unique to MoO and MoS were likely responsible for higher copy numbers of full-length elements in those lineages. *Grasshopper* had less than 50 solo LTRs in all genomes besides MoE MZ5-1-6, which had 114 full-length elements (Figure 2F), indicating expansion in that genome only. In contrast, *MGRL3* had a relatively high copy number of solo LTRs (>96) and a low copy number of full elements (<26) in all lineages (Figure 2F). This supports the idea that *MGRL3* was expanded before the divergence of the lineages, then was largely removed from all of them over time. Thus, although both expansion and contraction play roles in determining LTR-retrotransposon copy number in *M. oryzae*, large expansions were the main cause of lineage-specific copy number variation for the LTR-retrotransposons of interest.

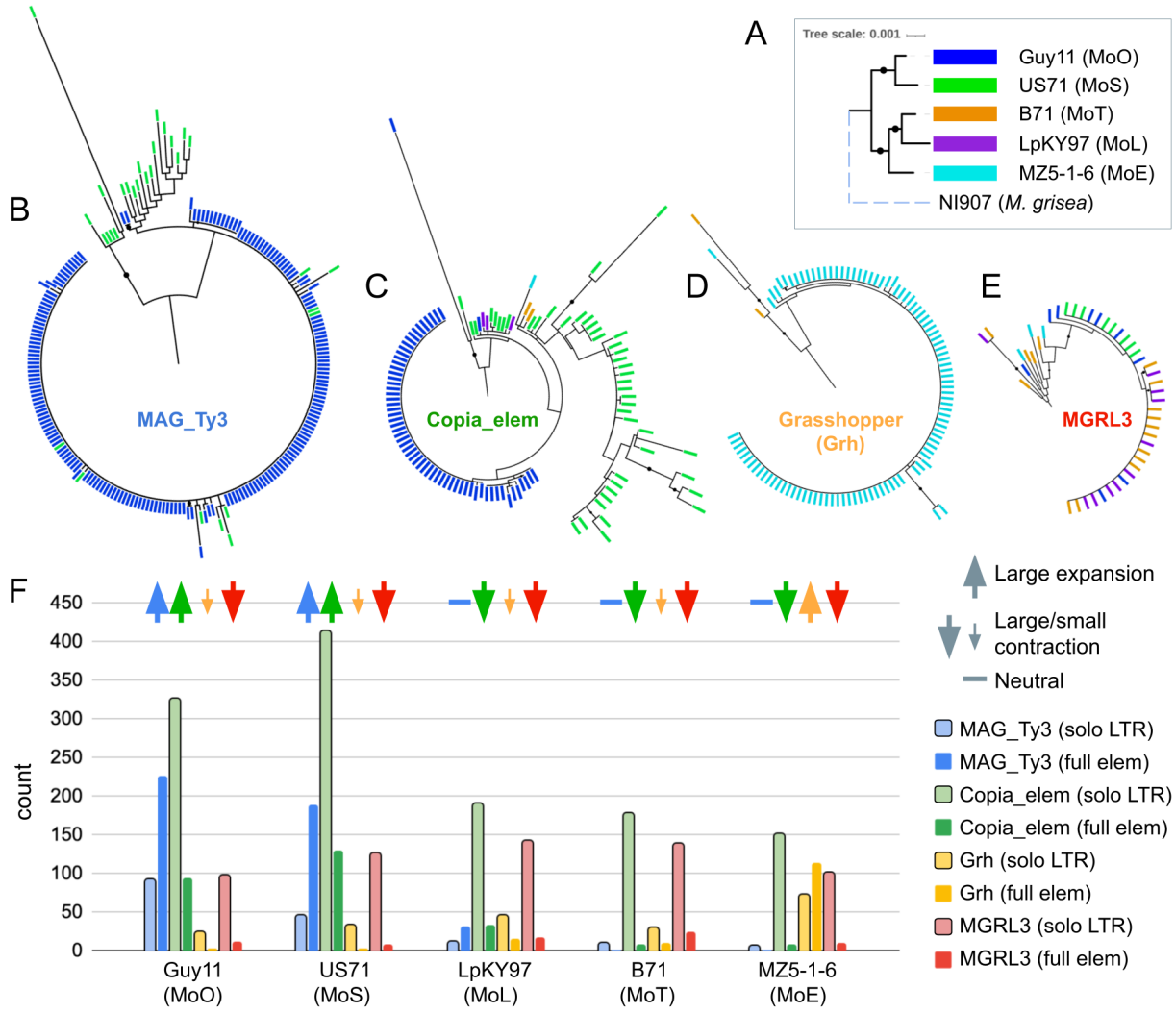


Figure 2: Certain LTR retrotransposons have experienced lineage-specific expansions. **A,** Maximum-likelihood (ML) phylogeny of representative genomes of each lineage, based on the alignment of 8,655 SCOs. Branch lengths are to scale, except for the dashed line of the *M. grisea* outgroup. Domain-based ML phylogenies of TEs **B,** *MAG_Ty3*, **C,** *Copia_elem*, **D,** *Grasshopper*, and **E,** *MGRL3* are shown. Colored rectangle tips correspond to the genome each element is from, as shown in A. **F,** In this barchart, solo LTR copy number is compared to the number of full length TEs to represent the expansion and contraction dynamics of the element in each genome. Each bar is colored according to the specific element, corresponding to the color of the label within each TE phylogeny (B-E). Lighter bars outlined in black represent solo LTRs. Arrows above each set of bars indicate our interpretation of the degree to which a particular LTR-retrotransposon in a certain genome has experienced expansion or contraction, based on the ratio of full element to solo LTRs.

Complex LTR-retrotransposon proliferation history and dynamics explain lineage-specific expansions in *M. oryzae*

Next, we sought to better understand the timing and history of the LTR-retrotransposon expansions we observed. To do so, we used nucleotide sequence comparison and sequence divergence tests, which make assumptions about sequence divergence rate; however, such tests could be violated by the presence RIP in the sequences. To measure how prevalent RIP was in our sequences we calculated the GC content of all TE copies in each representative genome and compared them to the genome-wide average GC content (12). We found that, although each TE had a different mean GC content, the distributions we observed were clustered closely about that value for most elements (Figure S4A). While we observed some trailing copies with low GC content, these were likely older elements that had been affected by RIP in the past. The *Pyret* LTR-retrotransposon had a GC content distribution with a strong skew that suggested it may have experienced RIP. However, our data indicated that Pyret had not been recently active, since it had a similar number of copies in each lineage (Figure 1A) and elements in its phylogeny didn't group by lineage (Figure S3), similar to *MGRL3*. Thus, RIP affecting Pyret had likely not occurred recently, and this element served as a good contrast to the recent LTR-retrotransposon expansions we focused on. As a final test, we compared the GC content of TEs in an MoO isolate originating from a recombining lineage (Guy11) to a clonal MoO isolate (FJ98099) (23), since RIP is only active during sexual reproduction (13). There were no differences in the GC content between these two isolates, indicating very little RIP activity in even the sexually reproducing rice blast isolates (Figure S4B). These analyses strongly suggest that although RIP may have been active in the past, it has had little effect on recently expanded TEs, and so divergence tests could be performed on our recently-expanded TEs.

To investigate the timing of LTR-retrotransposon expansions, we first sought to date individual TE insertions. A method for determining the age of LTR-retrotransposon insertions is to calculate the divergence between the flanking LTR sequences of an element (22). LTR-retrotransposons consist of an internal region, and two flanking LTRs that are identical upon insertion of the element (24). Flanking LTRs of older elements would be more divergent, since they have had more time to accumulate mutations, while newer elements would have highly similar LTRs (22). We determined LTR sequence divergence for *MAG_Ty3*, *Copia*, *Grasshopper*, and *MGRL3* retrotransposons (Figure 3A-D). Our results indicated that the expanded LTRs (*MAG_Ty3*, *Copia*, and *Grasshopper*) were inserted very recently, as many LTR pairs had zero sequence differences between them. The fact that LTR sequences are quite short (250 to 500 bp) combined with a reported mutation rate of 1.98e-8 substitutions/site/year in *M. oryzae* (4) likely contributed to this result. Nevertheless, given this mutation rate, a 500bp sequence would be expected to mutate once every 100,000 years, indicating that these expansions have the potential to have occurred after the divergence of the MoO and MoS lineages, which happened 9,800 years ago (4). The *Copia_elem* in the MoS genome, on the other hand, showed a broader range of LTR divergence values, indicating proliferation events spread out over time (Figure 3B). *MGRL3* had slightly higher divergences between its flanking LTRs that were generally similar for all the lineages (Figure 3D), supporting the idea that it experienced an older expansion in a single event before the divergence of the lineages. Overall, these findings support our interpretation that the lineage specific LTR-retrotransposon expansions occurred recently.

Next, using the Jukes-Cantor distance metric, we estimated the sequence divergence of full-length TEs, following a previously published method (25). For each TE, a consensus sequence was generated by aligning all copies of the element across all lineages. Then, the divergence of each copy from the consensus was determined and corrected by the Jukes-Cantor formula (26). The same procedure was then repeated for each lineage. The former method (one consensus for all lineages) showed the distance of each TE from the supposed common ancestor of that TE in all lineages, and indicated whether elements in each genome might have proliferated from the same original element (Figure 3E-H). The latter method (separate consensus for each lineage) showed how diverged the copies within one lineage were, and provided information on the recency of each expansion, and the population structure of TEs that contributed to it (Figure 3I-L). This analysis provided additional information beyond the TE phylogenies (Figure 2A-D), which were based only on the reverse transcriptase domain that each LTR-retrotransposon contains. For *Grasshopper*, we found that the Jukes-Cantor distance metrics could have suggested two expansions of this element, one older and one more recent (Figure 3G). However, when taking into account our LTR divergence analysis (Figure 3C), it was more likely that the entire *Grasshopper* expansion occurred recently and consisted of multiple elements expanding in parallel (Figure 3O). We also looked at where the TE proliferations were localized in the genome and found that the *Grasshopper* expansion occurred globally regardless of the location of the original proliferating element, as elements from both Jukes-Cantor peaks were distributed throughout MZ5-1-6's seven chromosomes (Figure S5A). In contrast to *Grasshopper*, *MAG_Ty3* appeared to have expanded by a single element only (Figure 3E), but also very recently (Figure 3A). These analyses on the Copia LTR-retrotransposon revealed a more complex scenario. Firstly, Copia elements in MoO and MoS appeared to have proliferated from the same original element, since most were about the same distance from the consensus of both lineages (Figure 3J). MoS Copia elements were more similar to each other than elements in MoO (Figure 3F). Yet, most MoO Copia had zero LTR divergence while many MoS Copia had further diverged LTRs (Figure 3B). The most likely explanation for this was that the Copia expansion in MoO occurred very recently but consisted of multiple elements from an older population of TEs with sequence differences. Meanwhile, the expansion in MoS was older and caused by just one element proliferating (Figure 3N). Finally, *MGRL3* looked to have proliferated from the same original element in all lineages (Figure 3L), which was consistent with the data supporting an old expansion of this element (Figure 2D,E) before the lineages diverged. Overall, we have demonstrated that LTR-retrotransposons in *M. oryzae* have experienced complex proliferation dynamics, resulting in different histories of each lineage-specific expansion.

To place the TE expansion events in the context of the overall history of *M. oryzae*, we attempted to compare the TE Jukes-Cantor distances with distances between single copy orthologous genes (SCOs) in all lineages, which has been previously used to indicate when isolates diverged from one another in relation to TE expansion events (25). However, the SCO gene distances were an order of magnitude smaller than that of the TEs (Figure S6). One potential explanation was that all the TE expansions occurred before the lineages diverged, which was very unlikely given the rest of the evidence presented here. Instead, this was likely due to the *M. oryzae* lineages being very closely related, and thus not giving enough signal for meaningful analysis. The *M. grisea* outgroup, the closest species with an available genome to *M. oryzae*, is highly diverged as compared to the divergence between the lineages (Figure

S2). Regardless, these results did indicate that TEs in *M. oryzae* are diverging at a much faster rate than single copy genes.

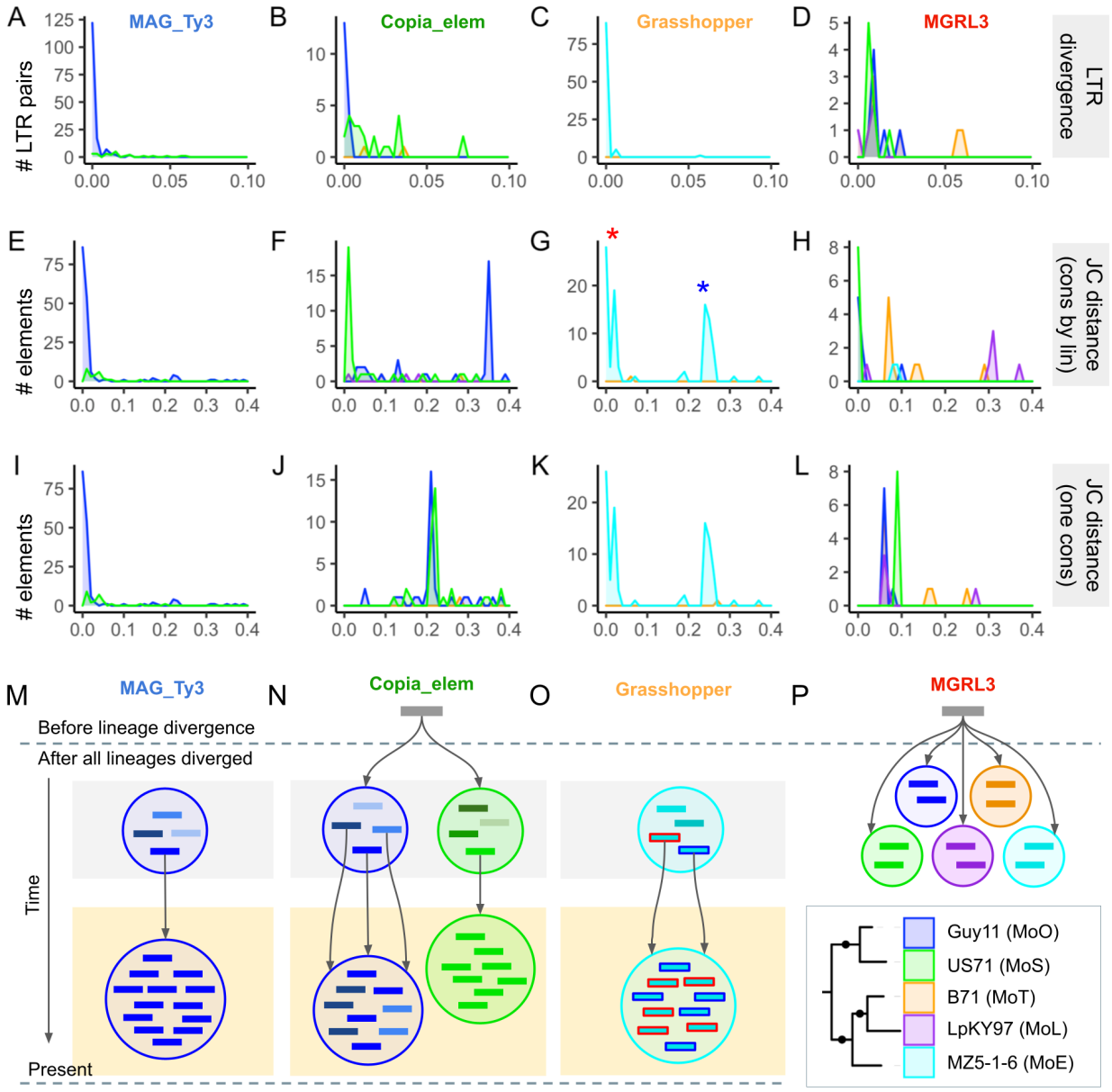


Figure 3: LTR divergence and Jukes-Cantor distance analyses suggest different methods of TE proliferation. Columns correspond to *MAG_Ty3*, *Copia_elem*, *Grasshopper*, and *MGRL3* (from left to right, for each row of the figure). **A-D,** Divergence between flanking LTR sequences of LTR retrotransposons. **E-H,** The Jukes-Cantor distance calculated using a separate consensus for each lineage. **I-L,** The Jukes-Cantor distance calculated using one consensus for all lineages. **M-P,** Schematic diagrams representing our hypothesis for the history of TE expansion events for each of the elements. The representation of the current population for each TE is highlighted in yellow, and these expansions occurred either by one or multiple elements from an older population of TEs with sequence differences

(highlighted in gray) proliferating recently. We indicate that TEs from different lineages might have proliferated from the same original element for Copia_elem and MGRL3. Blue and red outlined *Grasshopper* rectangles correspond to the two labeled peaks in G. *MGRL3* is an example of an old expansion. The tree in the bottom right corner serves as a key for color-coding of the lineages for all parts of the figure.

A DNA transposon, POT2, appears to have been transferred from the rice to the wheat *M. oryzae* pathotype

Although LTR-retrotransposons played a major role in the greater TE content of MoO and MoS, the *POT2* DNA transposon also stood out as being a large contributor. As shown by the ML phylogeny based on alignment of *POT2's* transposase domain, we found it to have greatly expanded in MoO and MoE, with a smaller expansion in MoT (Figure 4A). The phylogeny also suggested a potential transfer of *POT2* between MoO and MoT due to the unexpectedly high similarity of certain copies from MoT B71 and MoO Guy11. Previously published criteria for identifying potential transfers based on phylogenies are: (i) unexpectedly high similarity between TEs in lineages that aren't closely related, (ii) a patchy distribution of the element in one of those lineages, as well as absence from its sister lineage, and (iii) discordance between the TE tree and genome tree (27). When comparing the *POT2* phylogeny to the genome tree based on SCOs (Figure 2A), we observed clear discordance between the two. We expected MoT *POT2* to be more closely related to MoE *POT2*, since MoT was closer to MoE than to MoO; however, this was not the case. Additionally, *POT2* from another MoT genome (BR32) was not found in the clade containing closely related MoO and MoT *POT2*. BR32 *POT2* copies were also not found to be expanded, potentially indicating a patchy distribution of *POT2* in MoT. Finally, *POT2* was generally absent from MoL, the most closely related lineage to MoT. Although there were some older copies of *POT2* in all genomes, including MoL, no *POT2* from MoL were found in or near the clade containing potentially transferred MoT *POT2*. These results are in line with the criteria for a potential transfer of *POT2*.

We also observed additional lines of evidence pointing to a potential transfer of *POT2* between MoO and MoT. Our analysis of GC content revealed that the MoO Guy11 genome had two distinct sets of *POT2*, one with higher GC content and one with slightly lower GC content (Figure 4B). Many *POT2* in MoT had the same higher GC content as the former set, and most *POT2* in MoE had the same lower GC content as the latter set. This difference could have been caused by the ancestor of the MoO-MoE *POT2* being slightly RIPped, while the MoO-MoT *POT2* ancestor had not, then each element had its own evolutionary trajectory thereafter. The Jukes-Cantor analysis further supported the trend observed with GC content. When comparing *POT2* copies from all lineages to their consensus, we saw the same two MoO-MoT and MoO-MoE groupings (Figure 4D). This supported the idea that the two groups didn't come from the same original *POT2* element, rather they likely came from separate original elements with sequence differences. Although the MoO-MoE grouping of *POT2* by GC content and Jukes-Cantor distance might resemble a transfer between these lineages as well, it is not supported by the phylogeny (Figure 4A). So, the original MoO-MoE *POT2* was likely present in all lineages but only expanded in MoO and MoE, while the MoO-MoT *POT2* expanded only in MoO then transferred to MoT. Comparing each *POT2* copy to the consensus of its separate lineage (Figure 4C) showed that *POT2* in MoT and MoE were more closely related within their respective lineage than *POT2* within MoO. This suggested that either *POT2*

expansions in MoT and MoE occurred much more recently than in MoO, or that they consisted of one element expanding, while multiple elements from a population of TEs with sequence differences expanded in MoO. Either interpretation supported the idea that an individual *POT2* was recently transferred from MoO to MoT and subsequently expanded.

POT2 also experienced differential locality of its expansions in different lineages. Local proliferation was displayed by *POT2* in MoT, where most of its copies were located on the minichromosome sequences of the B71 genome (Figure S5B). In contrast, *POT2* in MoE was evenly distributed throughout the seven chromosomes (Figure S5C), similar to the LTR-retrotransposon expansions we characterized. Minichromosomes have been reported to harbor many repetitive sequences as well as virulence factors (28). Of the genomes used in this study, it is known that MZ5-1-6 (MoE), BR32 (MoT), and Guy11 (MoO) do not have minichromosomes, while B71 (MoT), LpKY97 (MoL), FR13 (MoO), US71 (MoS), and CD156 (MoE) do (28,29). Despite the existence of genomes both with and without minichromosomes in many of the lineages, their presence did not affect the lineage-specific patterns of TE content (Figure 1). Since the MoE MZ5-1-6 genome does not contain minichromosomes and experienced a global *POT2* expansion, while B71 does have minichromosomes and had a local *POT2* expansion there, perhaps the presence of minichromosomes could affect the locality of TE content. Regardless, this result further highlights the different expansion histories and dynamics of *POT2* in different lineages.

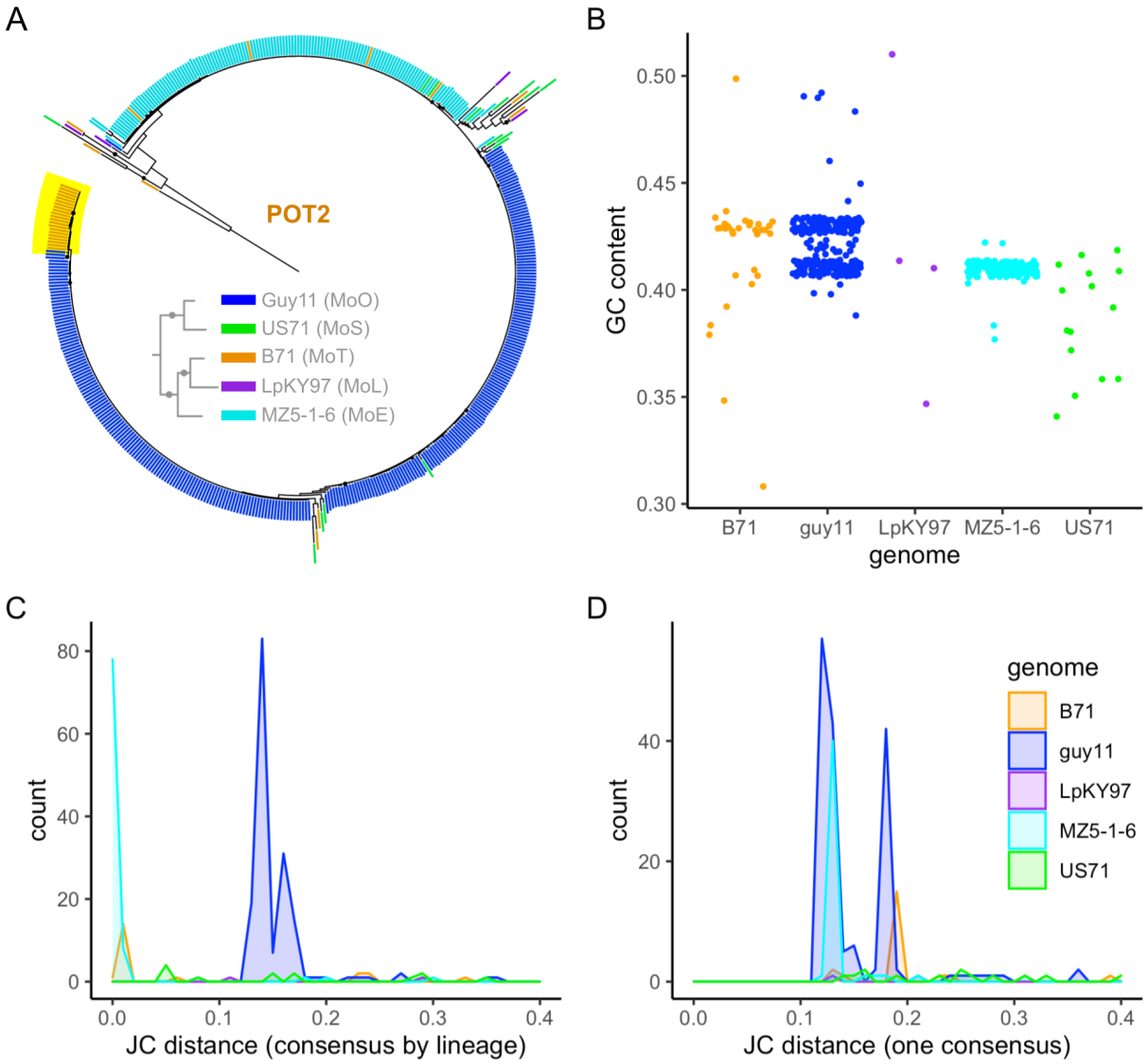


Figure 4: Evidence supporting a *POT2* transfer between MoO and MoT. **A,** Domain-based ML phylogeny of *POT2* in each representative genome (Guy11, US71, B71, LpKY97, MZ5-1-6) shows that it has experienced lineage-specific expansions. The yellow highlighted MoT elements are unexpectedly closely related to MoO elements. The smaller phylogeny within shows the expected relationships between the lineages (same as Figure 2A) and the color coding, which represents the lineage an element is from. **B,** Dots represent GC content in each *POT2* copy, showing two distinct groupings of copies from MoO. One group has similar GC content to most MoT *POT2* and the other is similar to most MoE *POT2*. **C,** Jukes-Cantor distance analysis of POT2 based on a separate consensus for each lineage, showing *POT2* in MoT and MoE have high intra-lineage similarity compared to *POT2* in MoO. **D,** Jukes-Cantor distance analysis of *POT2* based on one consensus for all lineages, also showing two distinct groups in MoO, one having similar distance from the consensus with MoT *POT2* and the other being similar to MoE *POT2*.

***POT2* was potentially transferred during a recombination event between wheat and rice blast**

To investigate how *POT2* may have been transferred between the MoO and MoT lineage, we first looked for any regions of the Guy11 and B71 genomes that may have been transferred in a larger transfer event. *POT2* elements and their flanking regions were compared using DNA alignments and synteny analysis; however, none of these segments containing *POT2* stood out as being potentially transferred regions. We then considered the possibility that *POT2* could have moved as part of a larger region but then transposed out of that region. We looked for evidence of genes that might have been transferred between Guy11 and B71 by filtering for gene trees that follow the same topology as the *POT2* phylogeny. One region in B71 on chromosome 7 stood out as having many of these genes, with 29 out of the 38 genes that matched the *POT2* phylogeny being located there (Figure S7). The other 9 genes were scattered among various chromosomes. We located the orthologs of the 29 B71 genes in the other genomes, and found them to be syntenic. In LpKY97 and MZ5-1-6, the other two chromosome level assemblies, the genes were in the same location on chromosome 7. We aligned the full-length nucleotide sequence from each genome and produced an ML phylogeny (Figure 5A), which showed that this entire region followed a *POT2*-like tree topology rather than the expected evolutionary relationships between the *M. oryzae* lineages (Figure 2A). Since B71 grouped with Guy11 in the MoO-MoS clade, it was likely that an MoO isolate was the donor of this region in B71. Additionally, since this region was syntenic in all genomes, the most likely explanation for this transfer event was homologous recombination.

We then looked at the TE insertions in the region we identified to determine the timing of the TE expansions we characterized in relation to the transfer of the region. There were no full length TEs contained in this region besides in the Guy11 genome, where *MAG_Ty3*, *Ty3_MAG1*, and *Ty3_MAG2* elements were likely inserted after the transfer event. There were a few solo LTRs located in the region, including a *MAG_Ty3* solo LTR that was present at the same location in all genomes. Located upstream of the transferred region there was a unique set of many TEs in each genome, indicating lineage-specific TE activity. There were no *POT2* within or nearby this region in B71, however there were many *POT2* copies upstream of the region in Guy11 (Figure 5B), supporting the possibility that one of these elements were included in the recombination event.

We next looked to see if any genes of importance were transferred along with this region. There were a few predicted effectors (Figure 5B), however they were not under presence-absence variation and did not include any AVRs or members of expanded *M. oryzae* ART and MAX effector families (30). We then characterized the genes in this region by obtaining their Gene Ontology (GO) terms (Additional File 1) and PFAM domain terms (Additional File 2). The most common terms included a putative ssRNA binding PFAM domain (RRM_1), iron ion binding molecular function (MF), zinc ion binding MF, proteolysis biological process (BP), glycolytic process BP, DNA repair BP, and mitochondrion cellular component (CC). However, the region was too small (182 genes out of 12,658 total) to perform meaningful enrichment analysis.

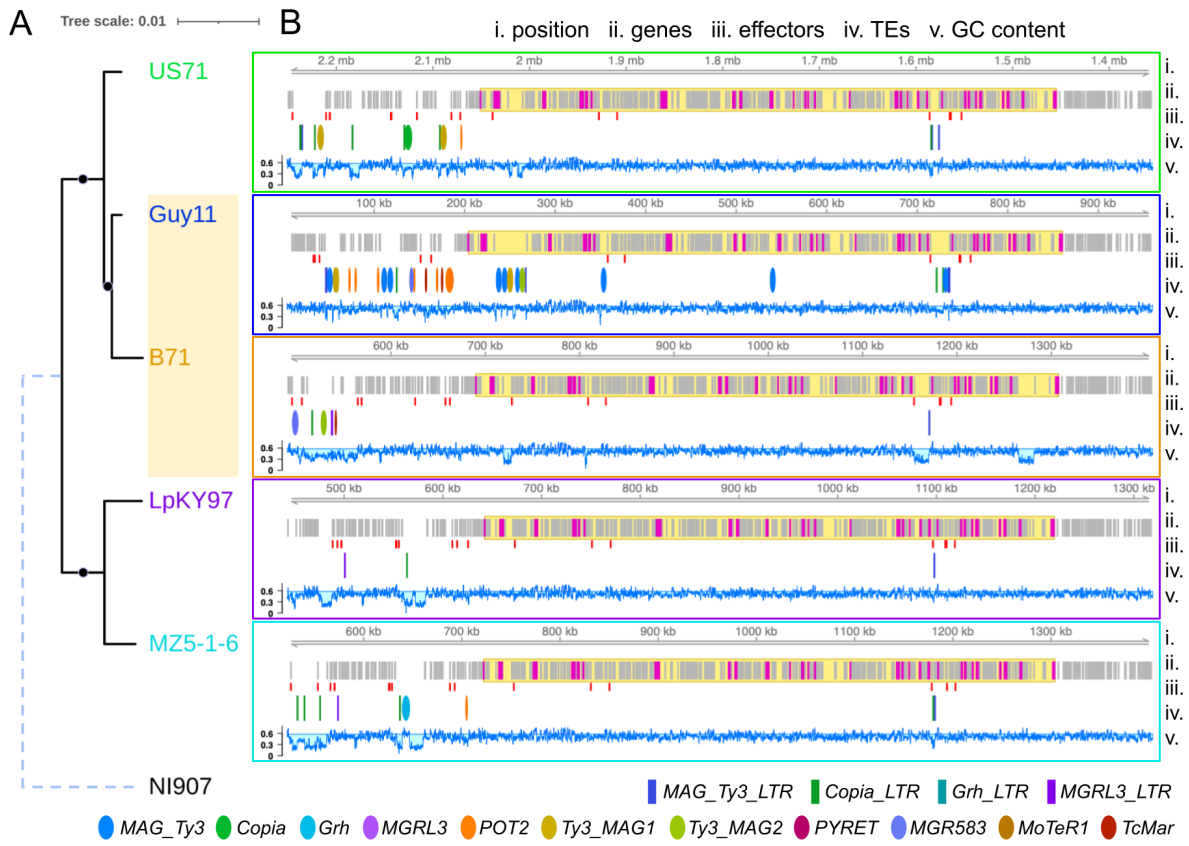


Figure 5: A potential region of recombination between MoO and MoT isolates. **A,** The ML phylogeny of an ~583 kb syntenic region on chromosome 7 that contains many genes following a *POT2*-like gene tree topology. Black circles indicate a bootstrap value of 100. Branch lengths are to scale, except for the dashed outgroup branch of NI907 (*M. grisea*). **B,** Genomic tracks show features of the potential region of recombination in each genome. Tracks from top to bottom: 1) Position along the scaffold or chromosome (B71: CM015706.1, Guy11: MQOP01000008.1, US71: UCNY03000007.1, LpKY97: CP050926.1, MZ5-1-6: CP034210.1, NI907: CM015044.1). 2) All genes, where magenta represents a *POT2*-like topology gene, and the yellow highlighted area indicates the region containing all of those genes. 3) Position of candidate effectors. 4) Position of TEs, where ellipses are full elements (blue=*MAG_Ty3*, green=*Copia_elem*, teal=*Grasshopper*, purple=*MGRL3*, orange=*POT2*, mustard=*Ty3_MAG1*, yellow-green=*Ty3_MAG2*, magenta=*PYRET*, skyblue=*MGR583*, light-brown=*MoTeR1*, tomato=*TcMar_elem*) and rectangles are solo LTRs (dark-blue=*MAG_Ty3_LTR*, dark-green=*Copia_LTR*, dark-teal=*Grasshopper_LTR*, dark-purple=*MGRL3_LTR*). 5) GC content, where the horizontal line is the genome-wide average GC content (0.577891).

Discussion

It is important to understand how fungal phytopathogens quickly adapt to their hosts in order to better manage diseases of crops and protect food security. TEs are hypothesized to be involved in the evolution and rapid adaptation of *M. oryzae*, however their role in shaping the evolutionary trajectories and host

specificity of its various lineages has not been explored in depth. To this end, we constructed a *de novo* TE library that represents the full diversity of TEs in all lineages of *M. oryzae*, and avoided biases in our analysis. We then annotated TEs in all of the lineages and found that MoO and MoS contain much greater TE content than the other lineages. While we focused in this study on near full-length TEs, further research is needed to study the full set of all repetitive DNA across the *M. oryzae* lineages. Most notably, protein domain-lacking elements such as MITEs are not included here, so our analyses likely underestimate overall TE content. Additionally, our analysis was restricted to highly contiguous genome assemblies which are few in number for *M. oryzae*, especially for the MoS, MoL and MoE lineages. While we are confident that our analyses are robust, we may have missed certain TE expansion events due to this limit in the scope of our analysis.

Despite these limitations, we found strong evidence that recent lineage-specific TE expansions contributed to the greater number of TEs in MoO and MoS, rather than removal of TEs in the other lineages. Our solo-LTR copy number analysis allowed us to verify that some LTR retrotransposons were expanded in certain lineages, rather than having been expanded in all lineages and subsequently removed in only some. By synthesizing the results of our LTR divergence and Jukes-Cantor distance analyses, we were able to construct a model showing differences in the method of TE expansion between various types of TEs and between the same TE in different lineages. Some expansions were caused by the proliferation of one element, while others consisted of multiple elements from an older population of TEs with sequence differences that proliferated in parallel. Most expansions occurred globally, with elements being distributed throughout the genome, however the *POT2* DNA transposon proliferated locally in the MoT isolate (B71) minichromosomes. Solo-LTR and LTR divergence analyses are not possible for DNA transposons, so it is difficult to determine expansion versus contraction dynamics for *POT2*, or how recently its copies proliferated. Nevertheless, our reconstruction of TE expansion histories points to the complexity of TE activity in *M. oryzae*.

Through our analyses, we found multiple lines of evidence suggesting the recent transfer of a DNA transposon between rice and wheat-infecting *M. oryzae* lineages. The phylogeny, Jukes-Cantor distances, and GC contents of *POT2* copies all showed that MoT *POT2* grouped unexpectedly with MoO *POT2* when considering the evolutionary relationships between the lineages. Notably, *POT2* has been found to insert into the *AVR-Pib* effector gene in MoO field isolates and modulate their virulence (10), so its transfer to other lineages has the potential to contribute to their adaptability. This observation could have been caused either by a horizontal transfer (HT) event or recombination between the lineages. Notably, *POT2* is a DDE-type DNA transposon of the Tc1/Mariner family, which are reported to be prone to HT (24). However, we could not identify direct evidence of such an HT event. It is possible that an individual *POT2* transferred by itself, which would be impossible to detect through a comparative genomics approach. Additionally, we did not find evidence of any non-syntenic, horizontally transferred regions that could have carried *POT2.* Since the potentially transferred *POT2* copies are localized on B71's minichromosomes, it is also possible that minichromosome dynamics allowed *POT2's* transfer or resulted in its locality. The HT of minichromosomes between isolates has been previously observed (28), as has the acquisition of core chromosomal regions by minichromosomes (29). An alternative explanation for the transfer of *POT2* is gene flow between the lineages through recombination during sexual

reproduction. A previous study has shown evidence of historical gene flow in *M. oryzae*, most of which was caused by events that occurred before the divergence of the lineages (3). However, the limited evidence of recent gene flow between MoO and MoT (3) allows for the possibility that the isolates may have interacted at some point. Additionally, the hypothesis that MoT and MoL arose recently via recombination in a multi-hybrid swarm of pre-existing host-specific isolates in South America (8) makes it possible that *POT2* was acquired by MoT in that event.

In searching for genes that might have accompanied *POT2* in a potential transfer event, we identified a region of recombination that could have facilitated the transfer. This region on chromosome 7 contained many genes whose phylogenies followed the topology of the *POT2* phylogeny, and were located syntenically in each lineage. The region we identified was also identified by Rahnama *et al.* through chromosome painting, where SNP analysis indicated that the region originates from a currently unsampled (cryptic) relative of MoO and MoS (8). They also showed that the recombination likely occurred during the swarm event, before the divergence of MoT and MoL from each other, since a few MoL isolates appear to contain the region and a few MoT isolates do not (8). This confirms that the region we identified had likely experienced recombination. The TE insertions in this region in the MoO isolate provide further evidence that the transferred region originates from a relative of MoO that participated in the swarm event, since it's unlikely that this region accumulated eight new LTR-retrotransposon insertions in the past 60 years (Figure 5B). Unfortunately, we did not find a copy of *POT2* in this region, and given the fact that DNA transposons leave almost undetectable scars (31), we could not find direct evidence that *POT2* was transferred through this mechanism.

While there is not sufficient evidence to claim that this particular recombination event was responsible for the transfer of *POT2* between the lineages, we still highlight this region as an example of a possible way that *POT2* could have been transferred. We describe a potential mechanism where *POT2* is transferred from the cryptic relative of MoO and MoS to an ancestor of MoT and MoL in the recombination event involving this region. Subsequently, *POT2* could have transposed out of the transferred region onto B71's minichromosomes, where it proliferated.

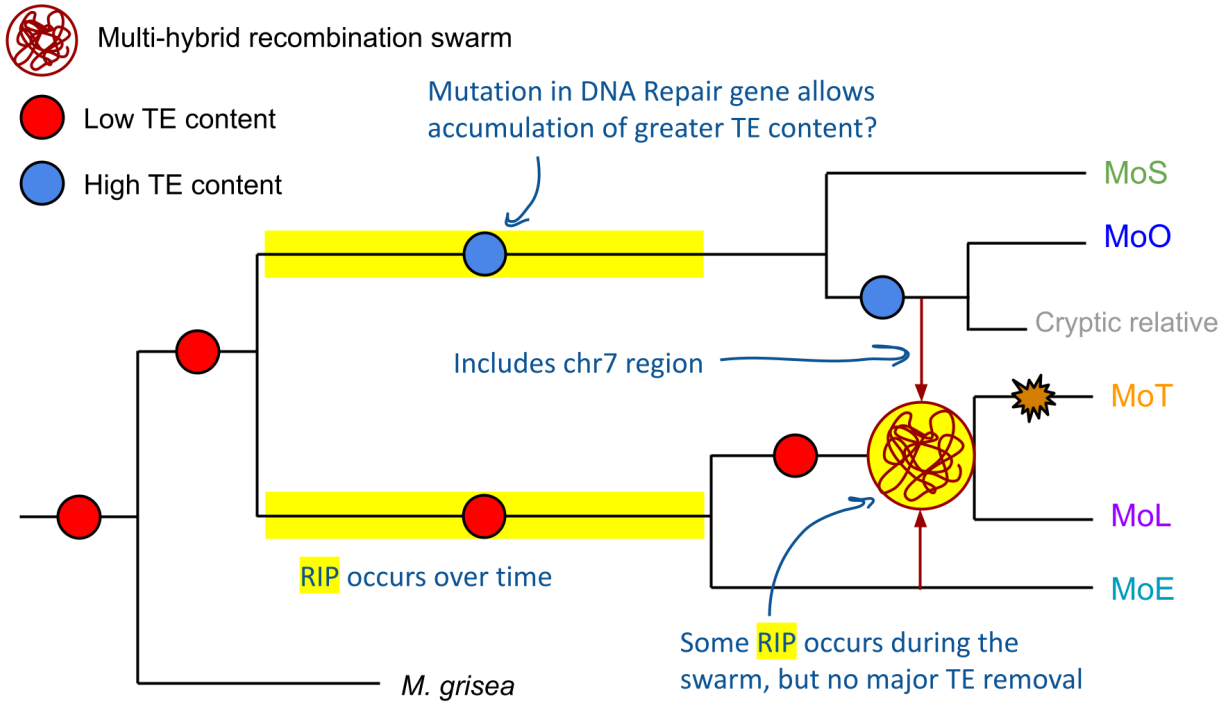


Figure 6: Proposed model explaining the increased TE content in MoO and MoS lineages. Phylogenetic analysis shows that a lower TE content was likely the ancestral state. A mutation in a DNA repair pathway may have contributed to the increased TE content in MoO and MoS. RIP (shown in yellow) may have occurred during the multi-hybrid swarm, but the swarm does not explain the higher TE content in MoO and MoS. During this swarm, a region on chr7 was transferred from a cryptic relative of the MoO lineage and may have contained a *POT2* element, which subsequently expanded in the MoT lineage (orange starburst).

Although the TE expansions we characterized help to explain the lineage-specific copy number variation, it is still unclear how MoO and MoS accumulated approximately twice as much TE content as other lineages. Based on the multi-hybrid swarm hypothesis presented by Rahnama *et al.* (8), one might infer that this event could explain the low TE content of MoT and MoL, where much TE content may have been lost due to RIP and removal via recombination. However, our data supports the alternative hypothesis that all lineages originally had lower TE content, and MoO and MoS independently accumulated their greater TE content after divergence from the common ancestor of all lineages (Figure 6). MoE genomes have low TE content, so this is not unique to the MoT and MoL that originated from the swarm event. Likewise, the fact that the TE content of the *M. grisea* outgroup is most similar to MoT, MoL, and MoE supports the idea that the common ancestor of all lineages had very few TEs. Additionally, if many TEs were removed during the swarm event, we'd expect large numbers of solo-LTRs in MoT and MoL, since recombination between flanking LTRs is a common mechanism of removal. However, our results indicate no extensive removal in MoT and MoL to explain the large difference in TE content.

A lack of severe RIP in the TEs we analyzed also refutes that RIP removed TEs from the genomes of all lineages except MoO and MoS. Through our GC content analysis (Figure S4) we concluded that there had not been recently active RIP in any *M. oryzae* lineages. It is very likely that RIP was active during the swarm due to the many sexual recombination events, and Rahnama *et al.* found regions that had clearly been RIPped (8). However, this RIP activity was not substantial enough to explain the differences in TE content we observed. The strongest evidence showing that RIP had not severely affected TEs during the swarm is that the *MGRL3* LTR-retrotransposon, which we found to be an old element that proliferated before the divergence of the lineages, has not experienced recent RIP in any of the lineages (Figure S4). Furthermore, had RIP affected *MGRL3* during the swarm, we would expect to find less copies in MoT and MoL compared to the others; however, its copy number of full elements and solo-LTRs is very uniform throughout all lineages.

Thus, we propose that there is a biological difference between the MoO-MoS and MoT-MoL-MoE clades responsible for the drastic difference in TE content (Figure 6). It would be interesting to investigate whether TEs in MoO and MoS are more often deregulated, and if this could be caused by a loss of or mutation in a DNA repair or TE suppression gene. Genes involved in DNA repair are of particular interest due to the recent finding that multiple non-canonical and error-prone DNA repair pathways exist in *M. oryzae*, and their influence on genomic variation are not well understood (16). Additionally, analyzing epigenetic data could offer insight into potential reasons for the complex TE proliferation histories we observed. Finally, future work comparing the distribution of TEs throughout the genomes of isolates with and without minichromosomes would be interesting to see whether they affect TE locality; this may soon be possible as more chromosome-level *M. oryzae* assemblies known to contain or not contain minichromosomes become available. Regardless of the approach, studying the underlying reasons for lineage-specific differences in TE content could reveal valuable insight into diversity generating mechanisms and the adaptive potential of *M. oryzae*.

Methods

Genomic datasets used and quality assessment

All genome sequences were retrieved from NCBI GenBank in December 2020, along with information on the host they were isolated from, year they were collected, their GenBank accession, assembly quality, number of scaffolds, and genome size (Table S1). Isolates were chosen primarily based on having the lowest number of contiguous scaffolds, which is most ideal for TE annotation (19). We assessed the completeness of the genomes using BUSCO (32) version 5.2.2 software with 'sordariomycetes_odb10' as the busco_dataset option.

TE annotation, classification, and phylogenetic analysis

The highest quality representative genomes of each lineage (Guy11 for MoO, US71 for MoS, B71 for MoT, LpKY97 for MoL, and MZ5-1-6 for MoE) were used as input into Inverted Repeat Finder (33) version 3.07 and RepeatModeler (34) version 2.0.2 software to obtain *de novo* annotations of TEs representing all lineages. Inverted Repeat Finder was called with options '2 3 5 80 10 20 500000 10000 -a3 -t4 1000 -t5 5000 -h -d -ngs', and RepeatModeler was called with options '-engine ncbi -LTRStruct'. These libraries

were combined with the RepBase (21) fngrep version 25.10 library of known TEs in fungi, and clustering was performed using cd-hit-est from CD-HIT (35) version 4.7 with options '-c 1.0 -aS 0.99 -g 1 -d 0 -M 0'. The resulting comprehensive TE library was then scanned against a list of domains that are associated with TE activity (20), containing both CDD profiles and PFAM profiles. The TE library was scanned for CDD profiles using rpsblastn from BLAST (36) version 2.7.1+ with option '-evalue 0.001'. PFAM profiles were retrieved from Pfam-A.hmm of HMMER (37) version 3.1b2. PFAM hmms were scanned for using pfam_scan.pl (38) version 1.6 with options '-e_dom 0.01 -e_seq 0.01 -translate all'. Elements not containing such a domain were filtered out, resulting in a comprehensive TE library of elements that contain a domain associated with TE activity.

Next, a curated approach was used to classify *de novo* elements in this library. Domain-based ML phylogenies were constructed using the most common PFAM domains found in the TE library, which were RVT_1 (PF00078.29), DDE_1 (PF03184.21), rve (PF00665.28), Chromo (PF00385.26), RNase_H (PF00075.26), and RVT_2 (PF07727.16). Each domain was aligned to the TE library using HMMER (37) hmmalign with options '--trim --amino --informat fasta'. The alignments were processed using esl-reformat and esl-alimanip from Easel version 0.48, which is part of the HMMER (37) package. Columns containing all gaps were removed by calling esl-reformat with the '--mingap' option, so that the length of the alignment was the same as the hmm length. Then, sequences that didn't match at least 70% of the hmm were filtered out by calling esl-alimanip with the '--lmin' option specifying 70% of the hmm length. Finally, esl-reformat was used to convert the alignment to fasta format. RAxML (39) version 8.2.11 with options '-f a -x 12345 -p 12345 -# 100 -m PROTCATJTT' was then used to construct the domain-based ML phylogeny of TEs containing the domain. This process was repeated for each of the six domains. The phylogenies were visualized in the Interactive Tree of Life (iTOL) (40) online tool, and clades where *de novo* elements grouped with elements of known classification were copied using the "copy leaf labels" feature of iTOL. *De novo* elements in the clade were then classified as the known element (Figure S1), generating a TE library with many more elements having classifications.

This classified TE library as well as the full set of genomes were then used as input to RepeatMasker (41) version 4.1.1 to generate copy number and positional data for TEs in all the genomes. These hits were converted to fasta format using bedtools (42) getfasta version 2.28.0 with the '-s' option to force strandedness, then filtered once more for elements containing a domain associated with TE activity, in the same way as described previously. This produced TE annotations for each genome of elements that were predicted to be complete, and many of which were now classified.

Domain-based ML phylogenies of each individual TE were constructed in the same way as those used to classify *de novo* elements as known elements. The domains used for each TE were: RVT_1 reverse transcriptase for *MAG_Ty3*, *Grasshopper*, and *MGRL3*, RVT_2 reverse transcriptase for *Copia*, and DDE_1 transposase for *POT2*.

Phylogeny of *M. oryzae* genomes

The genome tree of the *M. oryzae* isolates was generated by first annotating genes in each genome using FunGAP (43) version 1.1.0 with arguments '--augustus_species magnaporthe_grisea --busco_dataset

sordariomycetes_odb10'. RNAseq data for genome annotation was retrieved from the NCBI SRA database in June 2021. RNAseq for Guy11 (accession SRX5630771) was used as input for genomes of MoO and MoS lineage, RNAseq for B71 (accession SRX5900622) was used for MoT and MoL genomes, and RNAseq for MZ5-1-6 (accession SRX5092987) was used for MoE genomes. This resulted in predicted genes for each genome, which were input to OrthoFinder (44) version 2.5.4 along with the *M. grisea* NI907 proteome as the outgroup (retrieved from NCBI GenBank, accession GCA_004355905.1). OrthoFinder was run with options '-M msa -S diamond_ultra_sens -A mafft -T fasttree', and the output identified 8,655 SCOs. These were aligned using MAFFT (45) version 7.312 with parameters '--maxiterate 1000 --globalpair', and the alignments were concatenated. The ML phylogeny was produced from the alignment using raxmlHPC-PTHREADS-SSE3 (39) with options '-m PROTGAMMAGTR -T 24 -f a -x 12345 -p 12345 -# 100', and was visualized in iTOL (40).

Divergence analysis

To address potential RIP in *M. oryzae*, GC content was calculated using geecee from EMBOSS (46) version 6.6.0.0 in TEs and in coding sequences of the representative genomes.

LTR divergence analysis was performed by first determining a consensus sequence for each flanking LTR. Elements from the refined *MAG_Ty3*, *Copia*, *Grasshopper*, and *MGRL3* domain-based phylogenies were extracted from each representative genome, plus 1,000 bp on either side, using bedtools slop (42). These sequences were then blasted against the clustered TE library from an intermediate step in the TE annotation pipeline, before LTRs were removed when filtering for domain containing elements. This helped to manually determine the element that best represented the LTR sequence of each TE, which was blasted back to the set of LTR retrotransposon sequences plus flanking regions to extract LTRs. These extracted LTRs were then aligned using MAFFT (45), and a consensus sequence was generated using EMBOSS cons (46). The resulting LTR consensus sequences were used as the input library to RepeatMasker (41), which produced positional information for all LTRs. This was used along with the original full sequence plus flanking regions to find which LTRs belonged to which full elements using bedtools intersect (42). Finally, EMBOSS (46) needle was used to find the divergence of flanking LTR pairs.

Jukes-Cantor distance analysis was performed on all full-length TEs of interest, where the distance of each element to the consensus of its lineage, and to the consensus of all copies of that TE from any lineage were calculated. Following previous methods (25), we first produced the two types of consensus sequences by aligning TEs using MAFFT (45), then using EMBOSS (46) cons to generate the consensus of the alignment. The divergence of a TE from the consensus was found using EMBOSS (46) needle, and this divergence was corrected by the Jukes-Cantor distance formula, $d = -3/4 * \ln(1 - 4/3 * p)$, where p is the divergence and d is the corrected distance (26). Using Copia as an example, a consensus for all lineages was generated by aligning all copies of Copia present in its domain-based ML phylogeny, then the distance of all Copia from that consensus was found and plotted separately for each lineage. Also, a consensus was generated separately for Copia from MoO, and the distance was computed as previously described, except using this consensus specific to the lineage. This was done for Copia elements of each lineage separately and plotted. This process for making both plots was done for each of *MAG_Ty3*,

Grasshopper, *POT2*, and *MGRL3* as well. To find the Jukes-Cantor distances between genomes, coding sequences were extracted using gffread (47) version 0.12.7 with the '-x' option, then the output was filtered to only keep SCOs. This was done for the representative genomes of each lineage, and an additional MoO genome (FJ98099) as the reference point. The distance of each representative genome SCO to the reference genome SCO was determined using EMBOSS (46) needle, corrected by the Jukes-Cantor formula (26), and plotted separately for each lineage.

Solo LTR analysis

Solo LTRs were identified by determining which LTRs (from the annotations previously generated for LTR divergence analysis) did not belong to an LTR-retrotransposon found by the TE annotation pipeline. Using the '-v' option for bedtools intersect (42) returned only the LTR sequences that had no overlap with an annotated TE, and thus were considered solo-LTRs. The number of solo LTRs compared to the number of their full-length LTR-retrotransposon counterparts was used to determine whether the retrotransposon experienced expansion or removal from the genome.

Analyses for investigating potential *POT2* HT

To investigate potential larger HT regions containing *POT2*, synteny analyses were performed between all *POT2* regions in Guy11 and B71. *POT2* sequences plus 50,000 bp on either side were extracted using bedtools (42) slop and getfasta. These regions were compared using nucmer and mummerplot from MUMmer (48) version 4.0.0. To align the sequences, nucmer was called with the '--maxmatch' option, and to visualize the alignment, mummerplot was called with options '--postscript --color'. This produced synteny plots that were visually screened through for long segments of synteny between Guy11 and B71 flanking the position of *POT2*.

In order to find any genes that may have been transferred along with *POT2*, gene trees produced by OrthoFinder based on amino acid sequence were screened through to select those that follow the same topology as the *POT2* phylogeny. The ete2 (49) python package was used to determine which gene trees were structured such that the gene from Guy11 (MoO) and the gene from B71 (MoT) had the smallest distance from each other than from any other gene. Out of all SCOs, 388 genes had trees following this topology, and these were further refined by aligning their nucleotide sequences and determining topology in the same way as before. The remaining 38 genes whose trees based on nucleotide sequence followed this topology were visualized in IGV (50) to determine any localization in the B71 genome.

Investigating the potential region of recombination

The full segments of chromosome 7 from each representative genome that contained genes following a *POT2* topology were extracted using bedtools (42) getfasta, and the nucleotide sequences were aligned using MAFFT (45). A phylogeny was produced using raxmlHPC-PTHREADS-SSE3 (39), with options '-m GTRGAMMA -T 20 -f a -x 12345 -p 12345 -# 100' based on the alignment.

To characterize the genes located in the potential region of recombination, we obtained their GO terms using the PANNZER (51) webserver and their PFAM terms using pfam_scan.pl (38) with options '-e_dom

0.01 -e_seq 0.01' against the Pfam-A.hmm library of HMMER (37). The output from PANNZER was then filtered for GO terms with PPV value > 0.6.

Effector annotation and analysis

Effectors were predicted by following a previously established pipeline (52). Proteomes from FunGAP (43) output were input into SignalP (53) version 5.0 to filter for proteins containing a signal peptide. The output of SignalP was then input to tmhmm (54) version 2.0, which filtered out proteins containing a transmembrane domain. Finally, the remaining proteins were input to EffectorP (55) version 3.0.

Data processing and analysis

Analyses were conducted in a Linux environment with GNU bash version 4.2.46, GNU coreutils version 8.22, GNU Awk version 4.0.2, GNU grep version 2.20, and gzip version 1.5. Conda version 4.10.1 was used to install software. Scripts for parsing data were written in Python version 3.7.4, using biopython version 1.79. R version 4.1.0 was used for plotting and permutation tests, with packages ggplot2, RColorBrewer, tidyverse, and scales.

Availability of Data and Code

Datasets and intermediate analyses files are provided as additional data files, or available on Zenodo at DOI 10.5281/zenodo.7366416. Code and scripts used for all analyses are located in a GitHub repository (https://github.com/annenakamoto/moryzae_tes).

Acknowledgements

We thank Dr. Anna Muszewska for advice on TE annotation, Dr. Pierre Gladieux for help with substitution rates in *M. oryzae*, and Dr. Michael Seidl for insight on using the Jukes-Cantor distance metric. We also thank Dr. Ursula Oggenfuss and Dr. Emile Gluck Thaler from the Daniel Croll Lab for feedback on results. We thank members of the Krasileva Lab for feedback on manuscript preparation. Finally, we thank the Berkeley Research Computing program at the University of California, Berkeley for use of the Savio computational cluster resource.

Funding

AAN has been supported by the Berkeley Summer Undergraduate Research Fellowship, and PMJ has been supported by the Grace Kase-Tsujimoto Graduate Fellowship. KVK's work on this project has been supported by the Innovative Genomics Institute and the National Institute of Health New Innovator Director's Award, which also provided funding to AAN and PMJ.

Competing Interests

The authors declare no competing interests.

Contributions

AAN, PMJ, and KVK developed the project. AAN designed methods, performed data collection and analyses, interpreted results, produced figures, and wrote the original manuscript draft. All authors reviewed, edited, and approved the final manuscript. PMJ and KVK supervised research.

Supplemental Material

Table S1: *M. oryzae* genomes used in this study, with the lineage they belong to, host they were isolated from, collection year, NCBI accession, assembly quality, number of scaffolds, quality of the genome indicated by the percentage of complete BUSCOs, genome size in base pairs, and the percentage of the genome containing TEs (as found by our TE annotation pipeline). All genomes were retrieved from NCBI GenBank December 2020.

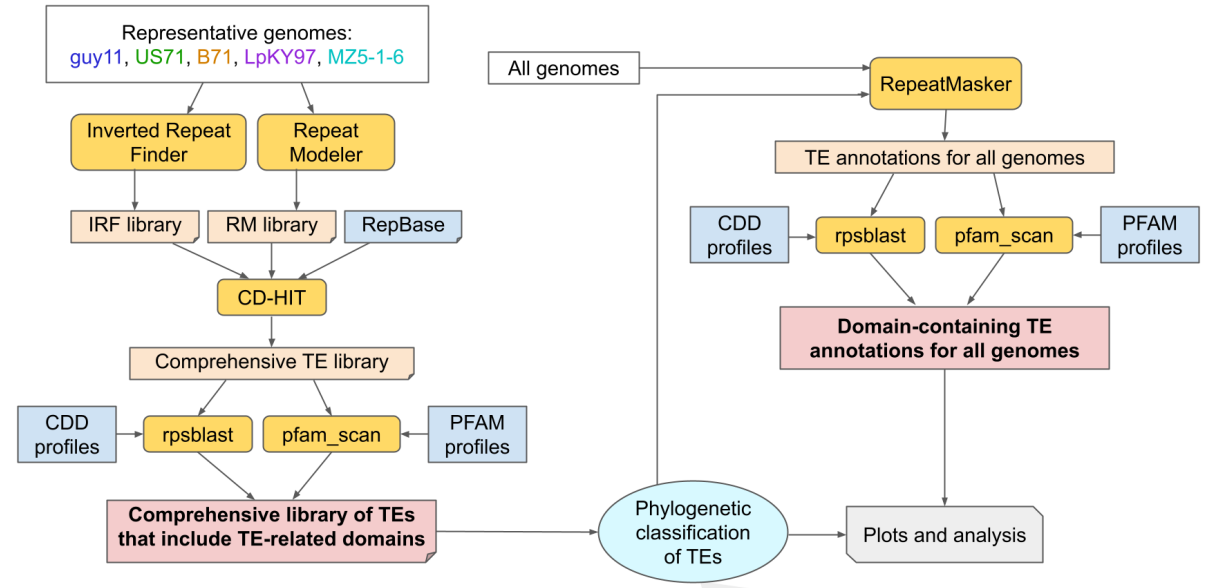
Isolate	Lineage	Host	Year	NCBI accession	Assembly Quality	# Scaffolds	N50	L50	% Complete BUSCOs	Genome size (bp)	% TEs in genome
AG006	*Oryza*	*Oryza sativa*	2011	GCA_905067025.2	Contig	24	6,618,557	3	95	47,005,811	11.41
AG039	*Oryza*	*Oryza sativa*	2011	GCA_905067035.2	Contig	26	6,249,570	3	95.1	47,495,958	11.75
Sar-2-20-1	*Oryza*	*Oryza sativa*	2013	GCA_011799915.1	Contig	16	6,161,260	4	98.2	46,284,791	11.25
AG098	*Oryza*	*Oryza sativa*	2011	GCA_905067015.2	Contig	33	6,094,221	4	96.9	47,810,826	11.83
PR003	*Oryza*	*Oryza sativa*	2003	GCA_905067075.2	Contig	16	6,063,740	4	95.7	44,615,198	10.32
AG032	*Oryza*	*Oryza sativa*	2011	GCA_905067055.2	Contig	24	5,961,411	4	96	45,916,910	11.17
AG059	*Oryza*	*Oryza sativa*	2011	GCA_905066965.2	Contig	37	5,900,770	4	97	47,743,121	12.05
AG038	*Oryza*	*Oryza sativa*	2011	GCA_905067005.2	Contig	19	5,673,907	4	95.7	46,291,169	11.13
FJ98099	*Oryza*	*Oryza sativa*	1998	GCA_011799925.1	Contig	10	5,637,639	4	98.3	44,637,475	10.27
AV1-1-1	*Oryza*	*Oryza sativa*	2015	GCA_011799965.1	Contig	13	5,503,597	4	98.2	44,970,614	10.66
FJ72ZC7-77	*Oryza*	*Oryza sativa*	1992	GCA_011799905.1	Contig	13	5,454,130	3	97.8	43,369,826	10.24
FR13	*Oryza*	*Oryza sativa*	1988	GCA_900474655.3	Contig	31	5,398,440	4	98.2	46,410,415	11.49
AG002	*Oryza*	*Oryza sativa*	2010	GCA_905067045.2	Contig	36	4,555,161	5	97	46,086,469	11.25
FJ81278	*Oryza*	*Oryza sativa*	1981	GCA_002368475.1	Contig	54	4,134,126	5	98.2	43,846,566	10.34
San_Andrea	*Oryza*	*Oryza sativa*	2001	GCA_905067085.2	Contig	35	4,122,582	5	95.7	48,509,714	12.1
guy11	*Oryza*	*Oryza*	1978	GCA_002368485.1	Scaffold	56	3,275,692	5	98.5	42,869,699	11.66

		sativa									
Arcadia2	*Setaria*	*Setaria italica*	1989	GCA_012654115.1	Scaffold	23	5,982,914	4	78.7	45,684,507	10.97
US71	*Setaria*	*Setaria italica*	1998	GCA_900474175.3	Contig	55	2,812,411	5	98.3	45,580,691	10.76
LpKY97	*Lolium*	*Lolium perenne*	1989	GCA_012272995.1	Complete genome	9	-	-	97.9	45,612,971	6.07
BTJP4-1	*Triticum*	*Triticum aestivum*	2016	GCA_900474225.2	Contig	59	4,344,896	4	69.3	44,506,711	5.55
BTGP6-f	*Triticum*	*Triticum aestivum*	2017	GCA_900474435.2	Contig	57	3,705,381	5	64.1	44,234,332	5.22
BTGP1-b	*Triticum*	*Triticum aestivum*	2017	GCA_900474635.2	Contig	74	2,814,025	6	61.1	44,406,101	5.25
BTMP13_1	*Triticum*	*Triticum aestivum*	2016	GCA_900474375.2	Contig	16	6,037,509	3	56.6	43,978,086	5.07
B71	*Triticum*	*Triticum aestivum*	2012	GCA_004785725.1	Chromosome	13	6,442,091	3	98.6	44,516,808	6.18
BR32	*Triticum*	*Triticum aestivum*	1990	GCA_900474545.3	Contig	17	5,096,353	3	98.3	41,805,140	4.84
MZ5-1-6	*Eleusine*	*Eleusine coracana*	1976	GCA_004346965.1	Complete genome	7	-	-	98.5	42,703,282	6.37
CD156	*Eleusine*	*Eleusine indica*	1989	GCA_900474475.3	Contig	27	5,531,649	4	98.1	43,939,965	5.24
NI907	*M. grisea*	*Digitaria sanguinalis*	1974	GCA_004355905.1	Chromosome	43	5,912,490	3	97.1	44,557,582	4.64

Table S2: Names and classifications of TEs discussed. The name we used throughout the paper for each element is shown, along with its original name in RepBase (21) fngrep version 25.10, the class, and the family each element belongs to. We adopted a naming convention for *Ty3* (formerly *Gypsy*) elements, where any "*GY*" in the RepBase name was replaced with "*Ty3*" in order to use a non-discriminatory and respectful naming scheme (56). *Grasshopper* is the original name of the *GYPSY1* RepBase element, so it is used instead (57). Elements that don't correspond to a specific element in RepBase are indicated by "N/A," and are named by their family (i.e. *Copia_elem*). LTR = long terminal repeat retrotransposon, NLTR = non-LTR retrotransposon, DNA = DNA transposon.

Name used	**RepBase name**	**Class**	**Family**
Ty3_MAG1	*GYMAG1*	LTR	*Ty3*
Ty3_MAG2	*GYMAG2*	LTR	*Ty3*
Grasshopper (Grh)	*GYPSY1*	LTR	*Ty3*
MAG_Ty3	*MAGGY*	LTR	*Ty3*
MGRL3	*MGRL3*	LTR	*Ty3*

PYRET	*PYRET*	LTR	*Ty3*
Copia_elem	N/A	LTR	*Ty1/Copia*
MGR583	*MGR583*	NLTR	*LINE/Tad1*
MoTeR1	*MoTeR1*	NLTR	*LINE/CRE*
POT2	*POT2*	DNA	*Tc1/Mariner*
TcMar_elem	N/A	DNA	*Tc1/Mariner*



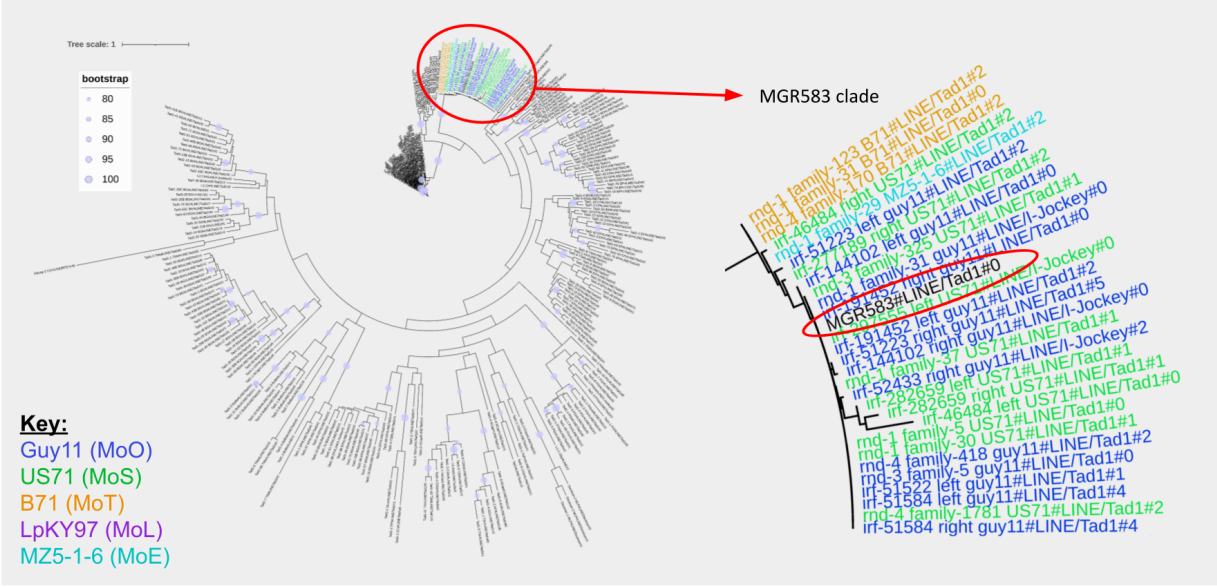


Figure S1: TE annotation pipeline diagram. The five representative genomes, Guy11 (MoO), US71 (MoS), B71 (MoT), LpKY97 (MoL), and MZ5-1-6 (MoE) were used for *de novo* TE annotation to produce an unbiased TE library that is representative of TE content in all lineages. The bottom gray box provides an example of how TEs were classified. Shown is a tree based on the Exo_endo_phos_2 domain with a phylogenetically defined TE subclade indicated by the red circle. Subclades of *de novo* elements (in color) that grouped with a known RepBase element (*MGR583* in this example, circled in red) were classified as that element. Colored text names of *de novo* elements represent the genome they were annotated in, as shown in the key.

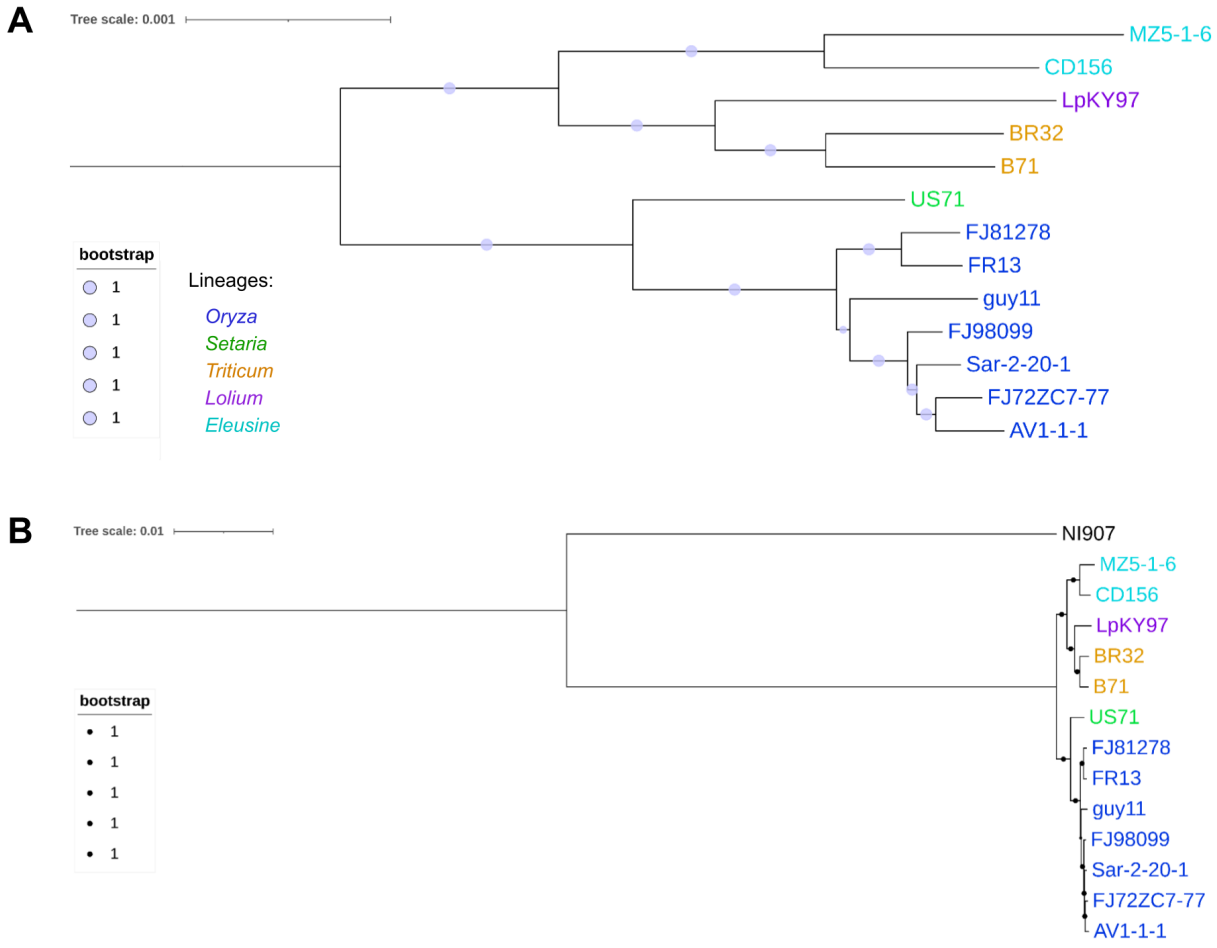


Figure S2: Maximum-likelihood (ML) phylogeny of *M. oryzae* genomes based on the alignment of 8,655 single copy orthologous genes (SCOs), **A,** Zoomed in without outgroup and **B,** including *Magnaporthe grisea* outgroup (NI907). Only genomes with BUSCO score greater than 97% were included.

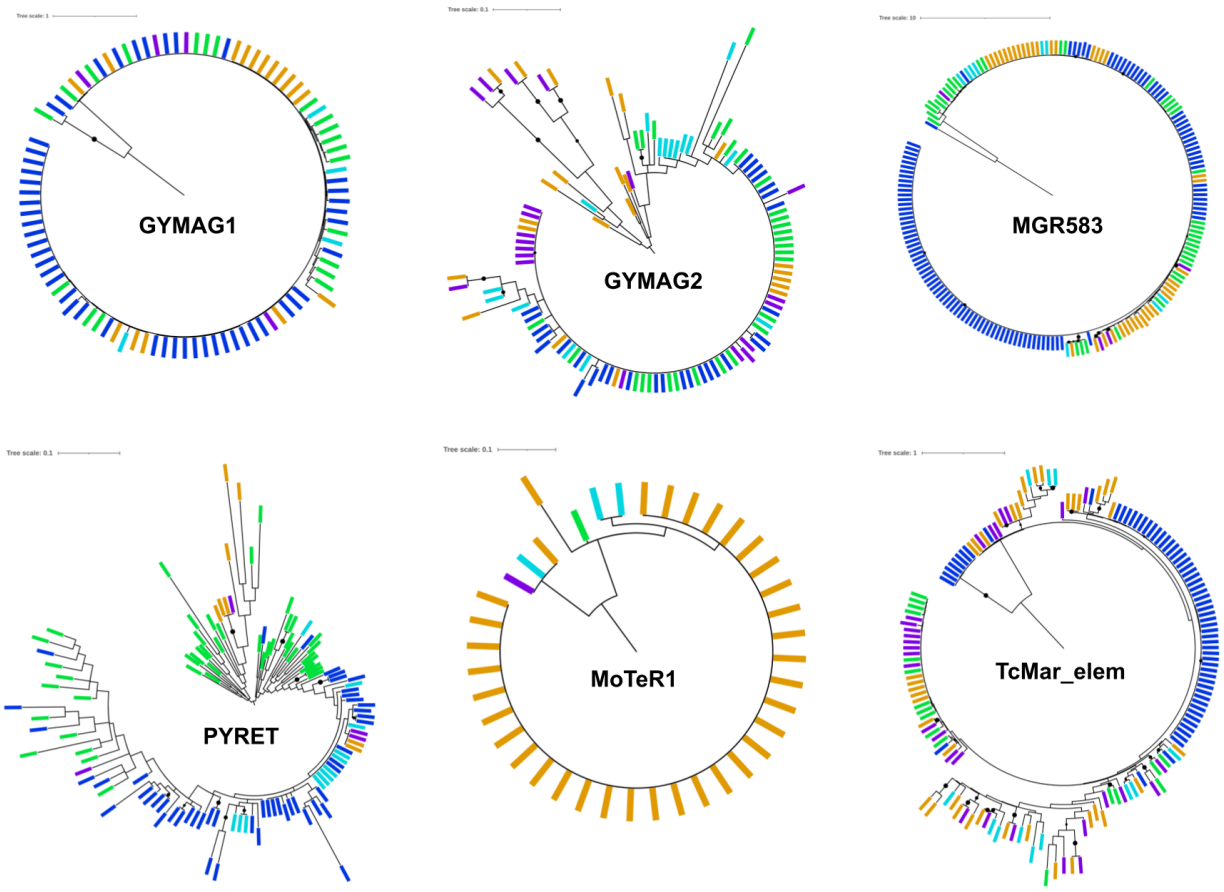


Figure S3: Domain-based maximum-likelihood (ML) phylogenies for *Ty3_MAG1*, *Ty3_MAG2*, *MGR583*, *PYRET*, *MoTeR1*, and *TcMar_elem*. Colored rectangle tips correspond to the genome each element is from: blue=Guy11 (MoO), green=US71 (MoS), orange=B71 (MoT), purple=LpKY97 (MoL), and cyan=MZ5-1-6 (MoE) .

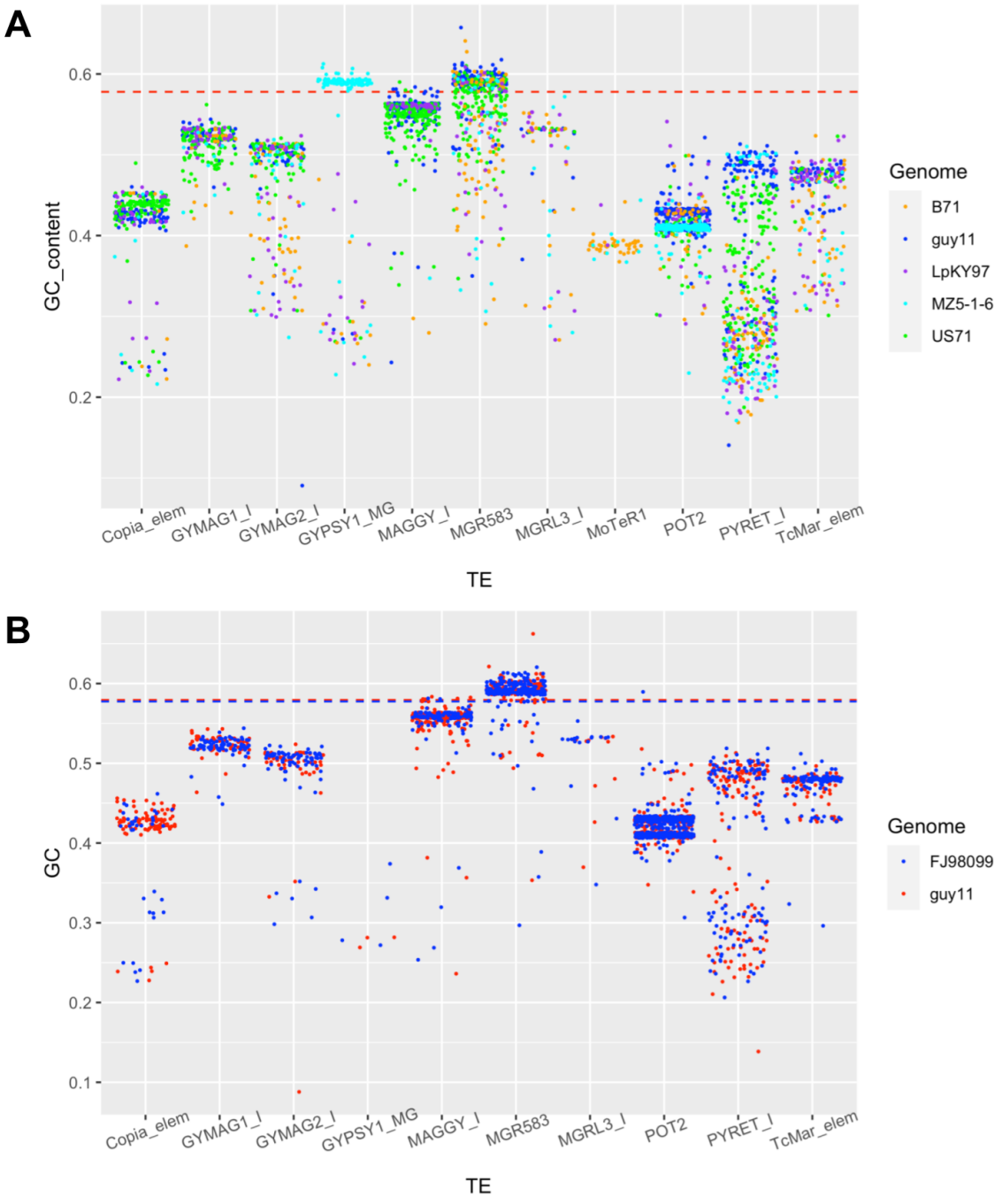


Figure S4: Analyzing the potential for repeat induced point mutations (RIP) using GC content of each TE. **A,** Jitter-plot showing GC content in each individual TE, color-coded by lineage. Each dot represents one TE copy, and the dashed red line represents the genome-wide average GC content of coding regions. **B,** Jitter-plot showing GC content in each individual TE, for the sexual guy11 genome (red) and the clonal FJ98099 genome (blue). Each dot represents one TE copy, and dashed lines represent the genome-wide average GC-content of coding regions for each genome.

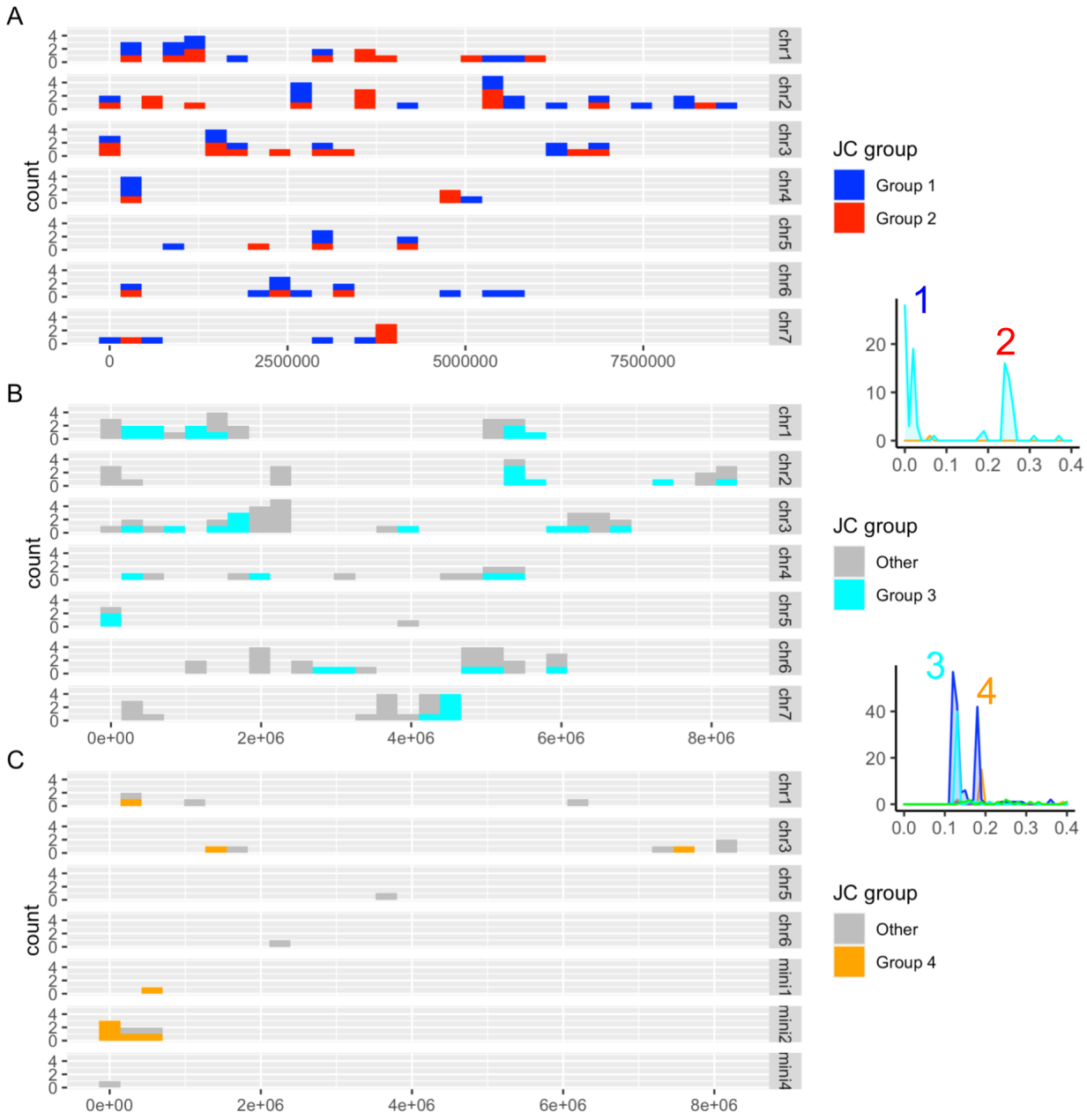


Figure S5: Global versus local proliferation of various expanded TEs. **A,** The location of *Grasshopper* elements of lower (Group 1 in red) and higher (Group 2 in blue) Jukes-Cantor distance throughout MZ5-1-6's seven chromosomes. The Jukes-Cantor plot for *Grasshopper* with groups 1 and 2 peaks labeled is shown for reference. **B,** The location of *POT2* elements in MZ5-1-6 that group with the lower Guy11 *POT2* Jukes-Cantor peak (Group 3 in cyan). **C,** The location of *POT2* elements in B71 that group with the higher Guy11 *POT2* Jukes-Cantor peak (Group 4 in orange). The Jukes-Cantor plot for *POT2* with groups 3 and 4 peaks labeled is shown for reference.

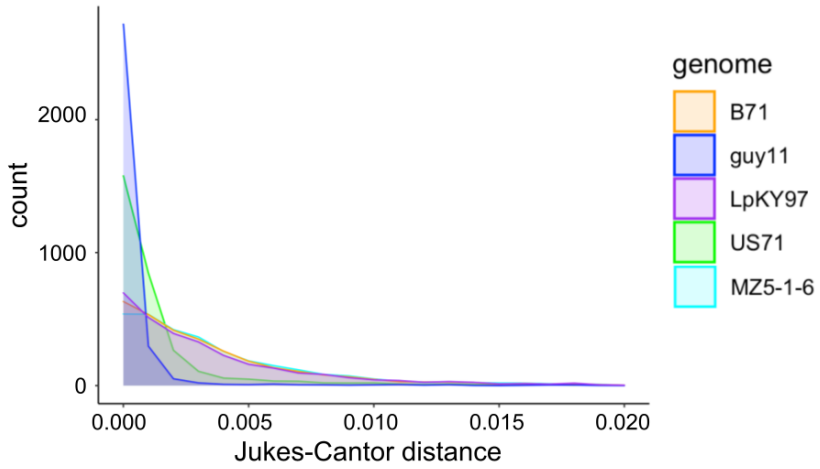


Figure S6: Jukes-Cantor distances between SCOs in each lineage are an order of magnitude smaller than distances between TEs and their consenses. The x-axis shows the Jukes-Cantor distance of an SCO from each genome compared to the orthologous gene in the reference MoO genome (Guy11). The y-axis shows the number of SCOs with a particular Jukes-Cantor distance from the reference.

Figure S7: Genes following *POT2* tree topology (in pink) are localized in a region on B71's chromosome 7. The first IGV track shows all seven chromosomes, and the second track shows just chromosome 7. The tree on the left side shows the phylogeny constructed from an alignment of the full-length region in each isolate. This doesn't follow the expected relationships between the lineages (tree on the right).

Additional Files

Additional File 1: GO terms output from PANNZER, filtered for >0.6 PPV value, for genes in the region on chromosome 7 following *POT2* topology.

Additional File 2: PFAM terms output from pfam_scan, filtered for E-value <0.01, for genes in the region on chromosome 7 following *POT2* topology.

References

1. Nalley L, Tsiboe F, Durand-Morat A, Shew A, Thoma G. Economic and Environmental Impact of Rice Blast Pathogen (*Magnaporthe oryzae*) Alleviation in the United States. Wang Z, editor. PLOS ONE. 2016 Dec 1;11(12):e0167295.
2. Dean R, Van Kan JAL, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, et al. The Top 10 fungal pathogens in molecular plant pathology: Top 10 fungal pathogens. Mol Plant Pathol. 2012 May;13(4):414–30.
3. Gladieux P, Condon B, Ravel S, Soanes D, Maciel JLN, Nhani A, et al. Gene Flow between Divergent Cereal- and Grass-Specific Lineages of the Rice Blast Fungus *Magnaporthe oryzae*. Taylor JW, editor. mBio. 2018 Mar 7;9(1):e01219-17.
4. Gladieux P, Ravel S, Rieux A, Cros-Arteil S, Adreit H, Milazzo J, et al. Coexistence of Multiple Endemic and Pandemic Lineages of the Rice Blast Pathogen. Guttman D, editor. mBio. 2018 May 2;9(2):e01806-17.
5. Zhong Z, Chen M, Lin L, Han Y, Bao J, Tang W, et al. Population genomic analysis of the rice blast fungus reveals specific events associated with expansion of three main clades. ISME J. 2018 Aug;12(8):1867–78.
6. Singh PK, Gahtyari NC, Roy C, Roy KK, He X, Tembo B, et al. Wheat Blast: A Disease Spreading by Intercontinental Jumps and Its Management Strategies. Front Plant Sci. 2021 Jul 23;12:710707.
7. Ceresini PC, Castroagudín VL, Rodrigues FÁ, Rios JA, Aucique-Pérez CE, Moreira SI, et al. Wheat blast: from its origins in South America to its emergence as a global threat. Mol Plant Pathol. 2019 Feb;20(2):155–72.
8. Rahnama M, Condon B, Ascari JP, Dupuis JR, Del Ponte E, Pedley KF, et al. Recombination of standing variation in a multi-hybrid swarm drove adaptive radiation in a fungal pathogen and gave rise to two pandemic plant diseases [Internet]. Evolutionary Biology; 2021 Nov [cited 2022 Oct 25]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.11.24.469688
9. Sánchez-Vallet A, Fouché S, Fudal I, Hartmann FE, Soyer JL, Tellier A, et al. The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens. Annu Rev Phytopathol. 2018 Aug 25;56(1):21–40.
10. Li J, Lu L, Li CY, Wang Q, Shi ZF. Insertion of transposable elements in AVR-Pib of *Magnaporthe oryzae* [Internet]. In Review; 2022 Mar [cited 2022 Oct 25]. Available from: https://www.researchsquare.com/article/rs-1440283/v1
11. Inoue Y, Vy TTP, Yoshida K, Asano H, Mitsuoka C, Asuke S, et al. Evolution of the wheat blast fungus through functional losses in a host specificity determinant. Science. 2017 Jul 7;357(6346):80–3.
12. Pereira D, Oggenfuss U, McDonald BA, Croll D. Population genomics of transposable element activation in the highly repressive genome of an agricultural pathogen. Microb Genomics [Internet]. 2021 Aug 23 [cited 2022 Sep 15];7(8). Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000540
13. Ikeda K ichi, Nakayashiki H, Kataoka T, Tamba H, Hashimoto Y, Tosa Y, et al. Repeat-induced point

mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context: Repeat-induced point mutation in *Magnaporthe grisea*. Mol Microbiol. 2002 Sep 2;45(5):1355–64.

14. 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. 2021 Feb 1;11:622368.
15. Thierry M, Charriat F, Milazzo J, Adreit H, Ravel S, Cros-Arteil S, et al. Maintenance of divergent lineages of the Rice Blast Fungus *Pyricularia oryzae* through niche separation, loss of sex and post-mating genetic incompatibilities. Stukenbrock EH, editor. PLOS Pathog. 2022 Jul 25;18(7):e1010687.
16. Huang J, Rowe D, Zhang W, Suelter T, Valent B, Cook DE. CRISPR-Cas12a induced DNA double-strand breaks are repaired by locus-dependent and error-prone pathways in a fungal pathogen [Internet]. Microbiology; 2021 Sep [cited 2022 Oct 25]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.09.08.459484
17. Joubert PM, Krasileva KV. The extrachromosomal circular DNAs of the rice blast pathogen *Magnaporthe oryzae* contain a wide variety of LTR retrotransposons, genes, and effectors [Internet]. Genomics; 2021 Oct [cited 2022 Sep 15]. Available from: http://biorxiv.org/lookup/doi/10.1101/2021.10.12.464130
18. Paulsen T, Kumar P, Koseoglu MM, Dutta A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet. 2018 Apr;34(4):270–8.
19. Rech GE, Radío S, Guirao-Rico S, Aguilera L, Horvath V, Green L, et al. Population-scale long-read sequencing uncovers transposable elements associated with gene expression variation and adaptive signatures in *Drosophila*. Nat Commun. 2022 Dec;13(1):1948.
20. Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. Transposable elements contribute to fungal genes and impact fungal lifestyle. Sci Rep. 2019 Dec;9(1):4307.
21. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015 Dec;6(1):11.
22. Jedlicka P, Lexa M, Kejnovsky E. What Can Long Terminal Repeats Tell Us About the Age of LTR Retrotransposons, Gene Conversion and Ectopic Recombination? Front Plant Sci. 2020 May 20;11:644.
23. Latorre SM, Reyes-Avila CS, Malmgren A, Win J, Kamoun S, Burbano HA. Differential loss of effector genes in three recently expanded pandemic clonal lineages of the rice blast fungus. BMC Biol. 2020 Dec;18(1):88.
24. Wells JN, Feschotte C. A Field Guide to Eukaryotic Transposable Elements. Annu Rev Genet. 2020 Nov 23;54(1):539–61.
25. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM, Wittenberg AHJ, et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res. 2016 Aug;26(8):1091–100.
26. Jukes TH, Cantor CR. Evolution of Protein Molecules. In: Mammalian Protein Metabolism [Internet]. Elsevier; 1969 [cited 2022 Sep 15]. p. 21–132. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9781483232119500097
27. Bergman CM. Horizontal transfer and proliferation of Tsu4 in *Saccharomyces paradoxus*. Mob DNA. 2018 Dec;9(1):18.
28. Langner T, Harant A, Gomez-Luciano LB, Shrestha RK, Malmgren A, Latorre SM, et al. Genomic rearrangements generate hypervariable mini-chromosomes in host-specific isolates of the blast fungus. Lin X, editor. PLOS Genet. 2021 Feb 16;17(2):e1009386.
29. Peng Z, Oliveira-Garcia E, Lin G, Hu Y, Dalby M, Migeon P, et al. Effector gene reshuffling involves dispensable mini-chromosomes in the wheat blast fungus. Lin X, editor. PLOS Genet. 2019 Sep 12;15(9):e1008272.
30. Seong K, Krasileva KV. Computational Structural Genomics Unravels Common Folds and Novel

Families in the Secretome of Fungal Phytopathogen *Magnaporthe oryzae*. Mol Plant-Microbe Interactions®. 2021 Nov;34(11):1267–80.

31. Luo G, Ivics Z, Izsvák Z, Bradley A. Chromosomal transposition of a *Tc1/mariner*-like element in mouse embryonic stem cells. Proc Natl Acad Sci. 1998 Sep;95(18):10769–73.
32. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Kelley J, editor. Mol Biol Evol. 2021 Sep 27;38(10):4647–54.
33. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G. Inverted Repeat Structure of the Human Genome: The X-Chromosome Contains a Preponderance of Large, Highly Homologous Inverted Repeats That Contain Testes Genes. Genome Res. 2004 Oct;14(10a):1861–9.
34. 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. 2020 Apr 28;117(17):9451–7.
35. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012 Dec;28(23):3150–2.
36. NCBI Resource Coordinators, Agarwala R, Barrett T, Beck J, Benson DA, Bollin C, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018 Jan 4;46(D1):D8–13.
37. Eddy SR. Accelerated Profile HMM Searches. Pearson WR, editor. PLoS Comput Biol. 2011 Oct 20;7(10):e1002195.
38. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 2016 Jan 4;44(D1):D279–85.
39. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014 May 1;30(9):1312–3.
40. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019 Jul 2;47(W1):W256–9.
41. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013. Available from: http://www.repeatmasker.org
42. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010 Mar 15;26(6):841–2.
43. Min B, Grigoriev IV, Choi IG. FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. Birol I, editor. Bioinformatics. 2017 Sep 15;33(18):2936–7.
44. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 2019 Dec;20(1):238.
45. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 2013 Apr 1;30(4):772–80.
46. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000 Jun;16(6):276–7.
47. Pertea G, Pertea M. GFF Utilities: GffRead and GffCompare. F1000Research. 2020 Sep 9;9:304.
48. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. Darling AE, editor. PLOS Comput Biol. 2018 Jan 26;14(1):e1005944.
49. Huerta-Cepas J, Dopazo J, Gabaldón T. ETE: a python Environment for Tree Exploration. BMC Bioinformatics. 2010 Dec;11(1):24.
50. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011 Jan;29(1):24–6.
51. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids

Res. 2018 Jul 2;46(W1):W84–8.

52. Singh PK, Mahato AK, Jain P, Rathour R, Sharma V, Sharma TR. Comparative Genomics Reveals the High Copy Number Variation of a Retro Transposon in Different *Magnaporthe* Isolates. Front Microbiol. 2019 May 6;10:966.
53. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, et al. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 2019 Apr;37(4):420–3.
54. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11. Edited by F. Cohen. J Mol Biol. 2001 Jan;305(3):567–80.
55. Sperschneider J, Dodds PN. EffectorP 3.0: Prediction of Apoplastic and Cytoplasmic Effectors in Fungi and Oomycetes. Mol Plant-Microbe Interactions®. 2022 Feb;35(2):146–56.
56. Wei KHC, Aldaimalani R, Mai D, Zinshteyn D, Satyaki P, Blumenstiel J, et al. Rethinking the “gypsy” retrotransposon: A roadmap for community-driven reconsideration of problematic gene names. Available from: https://osf.io/fma57/download
57. Dobinson KF. *Grasshopper,* a Long Terminal Repeat (LTR) Retroelement in the Phytopathogenic Fungus *Magnaporthe grisea*. Mol Plant Microbe Interact. 1993;6(1):114.