Catching up to fungal plant pathogens: A characterization of extrachromosomal circular DNAs and gene presence absence variation in *Magnaporthe oryzae*

By

Pierre M. Joubert

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor in Philosophy

in

Microbiology

and the Designated Emphasis

in

Computational and Genomic Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in Charge:

Professor Ksenia V. Krasileva, Chair
Professor Rachel B. Brem
Professor Benjamin K. Blackman

Spring 2023

Abstract

Catching up to fungal plant pathogens: A characterization of extrachromosomal circular DNAs and gene presence absence variation in *Magnaporthe oryzae*

By

Pierre M. Joubert

Doctor of Philosophy in Microbiology

Designated Emphasis in Computational and Genomic Biology

University of California, Berkeley

Professor Ksenia V. Krasileva, Chair

Fungal plant pathogens have major impacts on agriculture and global food security and are likely to have an even greater impact in the future. The current tools that we have available to combat them are insufficient, in part because these fungi can quickly adapt to these tools. Understanding fungal plant pathogen evolution is therefore essential to curbing the threat these pathogens pose. In Chapter 1, I describe the motivations for my dissertation work and the state of the field of fungal plant pathogen evolution. I also introduce the model organism I used in my research, *Magnaporthe oryzae*, which causes the blast disease. Chapter 2 describes my characterization of the extrachromosomal circular DNAs (eccDNAs) of *M. oryzae*. EccDNAs are a diverse class of molecules that can contribute to phenotypic and genotypic plasticity in eukaryotes, and I hypothesized that these may be involved in fungal plant pathogen evolution. I show that *M. oryzae* has a more diverse set of eccDNAs than other organisms and that these are enriched in LTR retrotransposons. I also show that many genes are found on eccDNAs in *M. oryzae*, and that effectors are enriched on eccDNAs. Finally, I show that eccDNAs are associated with gene presence-absence variation (PAV). Next, in Chapter 3, I discuss in greater detail the results presented in Chapter 2, as well as their implications and potential future directions. I also further discuss evidence in Chapter 2 that led me to believe that eccDNAs do not play a major role in fungal plant pathogen evolution and led me to focus directly on gene PAV in *M. oryzae* in the remainder of my dissertation. Subsequently, in Chapter 4, I describe my characterization of these events in *M. oryzae*. I find that genes experiencing PAV between lineages of *M. oryzae* are enriched in disease-causing and non-self-recognition genes. I describe how gene PAV events in the rice and wheat pathotypes show clear differences in their count and genomic location. Through comparing PAV genes to conserved genes, I show that these had distinct distances to TEs, distances to other genes, lengths, GC content, expression, and epigenetic marks. I also describe how a machine learning model can be trained to take advantage of these features to predict genes prone to PAV in the *M. oryzae* genome. Finally, in Chapter 5, I further discuss the implications of the results I describe in this dissertation, as well as future directions for implementing machine learning models and general knowledge of fungal plant pathogen evolution to help guide rational disease resistance engineering in crops.

Table of contents

Acknowledgements

First, I have to thank my advisor Ksenia Krasileva. She has been an incredible mentor to me and a constant source of support when I've needed it the most. Beyond research and communication skills, I am lucky to have learned so much about mentorship, management and leadership from her and these skills will be invaluable in my career. I would also like to thank my parents, Philippe and Cecile Joubert, for always believing in my success, for fostering my curiosity and problem-solving skills, and for providing me with the opportunities that drove me to graduate school and allowed me to be successful in my PhD. Of course, I have to thank my sisters, Rose and Louise, for their friendship and for teasing me the right amount when I'm being dramatic. I also thank Sharon Doty, my undergraduate mentor, for helping me build a strong intuition for biology and project design and for trusting me with so many opportunities that helped bring me to graduate school. Additionally, I'd like to give thanks to Anne Nakamoto for being a fantastic teammate and for so many interesting conversations about our research. Mentoring her has been a major highlight of my graduate school experience. I also thank Shruti Parikh for her friendship and emotional support throughout my PhD, for teaching me so much about mentorship, and for her help navigating the job market. I also have to thank all of the members of the Krasileva lab for their insightful feedback and support on all of my projects as well as many fascinating scientific discussions. Being part of a lab with so many creative and bright people that can conduct impactful research in so many different areas has been an incredible experience.

Finally, I thank Reena Debray for providing me so much support and happiness, especially in the most difficult times and for being my biggest cheerleader. I could not have made it through graduate school without her. Our conversations about mentorship, academia, research, science, and so many other things have taught me so much about the world and have really helped shape what I want to do with the rest of my life.

Chapter 1

Introduction to fungal plant pathogen genome evolution, *Magnaporthe oryzae*, and motivations for research

Fungal plant pathogens pose a major threat to all agricultural crops. These fungi can cause many different types of disease and cause massive losses in yield. Unfortunately, climate change will likely increase pathogen pressure on agriculture [1,2]. The recent emergence and rapid spread of devastating diseases like wheat blast and Panama disease of bananas also indicate that fungal plant pathogens will be a growing problem in the future [3,4].

Current methods to combat these fungi are unfortunately not enough to curb their threat. Fungicides have devastating environmental consequences and can often only slow the inevitable spread of these fungi [5,6]. Selective breeding and genetic engineering are more likely to be fruitful solutions. However, these tactics are often very slow to implement, and fungal plant pathogens can quickly adapt to overcome disease resistance [7]. Understanding how these pathogens can adapt to their host will help us design crops with more robust and long-lasting disease resistance in the future, and it is a central motivating question for my dissertation work.

Fungal plant pathogens secrete proteins called effectors to modify host functions and cause disease [8]. Their hosts often use receptors called NLRs to detect these effectors and trigger an immune response [9]. To avoid this immune response, pathogens secrete large suites of effectors with redundant functions. This means that they can simply stop secreting effectors that their host detects to escape recognition. The loss of effector production can be triggered through a variety of mechanisms including gene deletion, mutation, and transcriptional silencing [10]. Alternatively, these fungi can secrete mutant variants of effectors that cannot be detected by the plant or secrete effectors that suppress the plant's immune system. The result of these dynamics is that, as plant scientists produce new resistant crop varieties, often by genetically engineering or breeding NLRs with new binding specificities into them, fungal plant pathogens continue to evolve and eventually escape recognition. This has led to resistant crops sometimes losing their resistance after being in use for only a few years [10].

In addition to fast generation times, these pathogens take advantage of many features of their genomes to adapt quickly to their hosts. For example, many fungal plant pathogens have a "two-speed" genome architecture [7,11]. This genomic organization is characterized by the tendency for house-keeping genes to be present in gene-rich, repeat-sparse regions of the genome, and the tendency for effector genes to be present in gene-sparse, repeat-rich regions of the genome. This architecture is thought to enable effector genes to evolve rapidly while important house-keeping genes evolve more slowly. Transposable elements (TEs), accessory chromosomes, rapid evolution in sub-telomeric regions of the genome, and horizontal gene transfer events have also been implicated in the success of fungal plant pathogens [7,12,13]. Finally, extensive gene presence-absence variation (PAV) has been observed in many fungal plant pathogens, implying that they are able to quickly lose genes to escape recognition by their hosts [10,14]. However, fungal plant pathogen evolution remains an active area of research, and how these fungi shape their genomes to better take advantage of these features remains unknown. Additionally, whether we can construct predictive models that can predict effector evolution and help guide disease-resistance engineering has remained an open question in the field.

Magnaporthe oryzae (syn. *Pyricularia oryzae*) causes the blast disease and is one of the most important and well-studied fungal plant pathogens [15,16]. It causes massive losses in rice crops each year, equivalent to feeding 60 million people [17], and the emerging devastating wheat-infecting pathotype is likely to be an even bigger threat to global wheat production [4,16]. Because of its importance, it is amongst the fungal plant pathogens with the most available bioinformatics datasets. This includes hundreds of available genomes as well as transcriptomic datasets, and epigenetic datasets which makes it an ideal model for studying fungal plant pathogen evolution. TEs, accessory chromosomes, horizontal gene transfer, and extensive gene PAV have all been implicated in its evolution [18–27]. Additionally, *M. oryzae* reproduces clonally most of the time, which raises major questions about how it can produce enough genetic diversity to evolve in response to its hosts [26,28]. In this dissertation, I describe the research I carried out to improve our understanding of *M. oryzae* evolution and discuss how my findings might apply to the evolution of other fungal plant pathogens.

Genomes often respond to stress by shedding extrachromosomal circular DNAs (eccDNAs). These molecules can be generated through many different processes including DNA repair, transcription, and TE activity [29–32]. EccDNAs can accumulate in cells, massively amplifying the copy numbers of genes they contain [33–38]. EccDNA-mediated gene amplification can cause changes in phenotype and can result in adaptation to stress [33,34,37,39,40]. EccDNAs can also generate genomic structural variation [33,41–44]. Given the impressive evolutionary potential of eccDNAs, and the associations between eccDNAs and repetitive sequences, I hypothesized that eccDNAs might be involved in fungal plant pathogen evolution. If this were the case, eccDNAs could be an important target for crop disease prevention, as they are in cancer [29], and could allow us to observe fungal plant pathogen genome evolution under various stressors more easily in the lab.

As described in Chapter 2 of this dissertation, I sequenced the eccDNAs of *M. oryzae* and characterized them. I compared eccDNAs across multiple organisms and found that *M. oryzae* eccDNAs were particularly numerous and diverse and were more likely to contain LTR retrotransposon sequences. When I looked at eccDNAs generated from LTR retrotransposons in *M. oryzae*, I found that each of them used unique mechanisms for eccDNA formation. I also found that most genes in *M. oryzae* were present on eccDNAs in my data. However, a small subset was not, and this subset was enriched for genes related to cytoskeleton formation. I also found a set of genes that were particularly prone to forming eccDNAs. These genes were more likely to experience PAV than other genes and were more likely to be in the gene-sparse, repeat-dense compartment of the *M. oryzae* genome. Finally, I found that effectors were more likely to be found on eccDNAs than other genes. These results posed many interesting questions about eccDNA biology. However, I did not find evidence of eccDNA-mediated structural variation in *M. oryzae*, indicating that eccDNAs are unlikely to play a major role in fungal plant pathogen evolution. This led me to focus on other types of structural variation instead, and especially PAV.

As previously mentioned in this chapter, one of the ways fungal plant pathogens can escape recognition by their hosts is simply by removing the genes responsible for producing the detected effector from their genomes. Extensive effector PAV has been reported in the past in the rice pathotype of *M. oryzae* [20,25–27]. This PAV is thought to contribute to *M. oryzae*'s adaptation, and differences in gene content have been observed in isolated populations of the fungus [26]. However, other groups that have looked at PAV in *M. oryzae* in the past were mostly focused on effectors. They also did not dive deeper into where these PAV events happen in the genome and what features they might be associated with. Finally, PAV events had not been compared across different pathotypes of *M. oryzae*. I hypothesized that if we could gain a better understanding of these features, we could generate models that might help predict gene losses in the future.

In Chapter 4 of this dissertation, I describe my characterization of PAV in *M. oryzae*. First, I found extensive PAV of effectors and non-self-recognition genes in the fungus. When I looked at what features were associated with gene PAV, I found that these events were associated with high TE densities and low gene densities. I also found that genes prone to PAV had clear differences in their length, GC content, expression, and epigenetic marks when compared to conserved genes. When I compared orthogroups experiencing PAV in rice-infecting strains of *M. oryzae* to those prone to PAV in wheat-infecting strains, I found that the rice pathotype had fewer of these orthogroups and that they were more likely to be found in well-defined clusters in the *M. oryzae* genome. I also found clear differences in the genomic features associated with PAV genes between rice-infecting and wheat-infecting *M. oryzae*. Using the genomic features of PAV genes that I identified, I was able to train machine learning models to predict which genes are prone to PAV in the *M. oryzae* genome. The work presented in Chapter 4 of this dissertation, has important implications for the future of plant disease resistance engineering and supports the hypothesis that we will be able to use complex models to predict fungal plant pathogen evolution in the future and use that information to improve the engineering of robust disease resistance in crops.

In summary, this dissertation highlights the need for further investigation of eccDNAs, especially in fungal plant pathogens, and demonstrates unique features of the *M. oryzae* genome that could play an important role in its evolution. It also highlights differences in the evolution of the different pathotypes of *M. oryzae* that could have been shaped by differences in their history. Finally, I show evidence in this dissertation that aspects of genome evolution in *M. oryzae* can be linked to specific genomic features and that these can be used to construct predictive machine learning models, which supports the idea that predicting fungal plant pathogen evolution to guide disease engineering in crops could be possible in the future.

Chapter 2

Characterization of the extrachromosomal circular DNAs of *Magnaporthe oryzae*

The contents of this chapter are based on the following publication:

Joubert, P.M., Krasileva, K.V. The extrachromosomal circular DNAs of the rice blast pathogen Magnaporthe oryzae contain a wide variety of LTR retrotransposons, genes, and effectors. *BMC Biology* 20, 260 (2022). https://doi.org/10.1186/s12915-022-01457-2

Abstract

Background:

One of the ways genomes respond to stress is by producing extrachromosomal circular DNAs (eccDNAs). EccDNAs can contain genes and dramatically increase their copy number. They can also reinsert into the genome, generating structural variation. They have been shown to provide a source of phenotypic and genotypic plasticity in several species. However, whole circularome studies have so far been limited to a few model organisms. Fungal plant pathogens are a serious threat to global food security in part because of their rapid adaptation to disease prevention strategies. Understanding the mechanisms fungal pathogens use to escape disease control is paramount to curbing their threat.

Results:

We present a whole circularome sequencing study of the rice blast pathogen, *Magnaporthe oryzae*. We find that *M. oryzae* has a highly diverse circularome that contains many genes and shows evidence of large LTR retrotransposon activity. We find that genes enriched on eccDNAs in *M. oryzae* occur in genomic regions prone to presence-absence variation, and that disease associated genes are frequently on eccDNAs. Finally, we find that a subset of genes is never present on eccDNAs in our data, which indicates that the presence of these genes on eccDNAs is selected against.

Conclusions:

Our study paves the way to understanding how eccDNAs contribute to adaptation in *M. oryzae*. Our analysis also reveals how *M. oryzae* eccDNAs differ from those of other species, and highlights the need for further comparative characterization of eccDNAs across species to gain a better understanding of these molecules.

Background

Extrachromosomal circular DNAs (eccDNAs) are a broad and poorly understood category of molecules defined simply by the fact that they are circular and originate from chromosomal DNA. This group of molecules has been referred to by many names and includes many smaller categories of molecules such as episomes, double minutes, small polydisperse circular DNAs, and microDNAs. They form through several mechanisms including non-allelic homologous recombination (NAHR), double strand break repair, replication slippage, replication fork stalling, R-loop formation during transcription [29], and as a byproduct of LTR retrotransposon activity [30–32] (Fig. 1A). EccDNAs can accumulate in cells through autonomous replication [33–36], high rates of formation [37], or through retention in ageing cells [38]. EccDNAs can contain genes, and amplification of gene-containing eccDNAs has been linked to adaptation to copper

[37] and nitrogen [33] stress in yeast, herbicide resistance in weeds [34], and drug resistance in cancer cells [39,40]. EccDNA formation is thought to sometimes cause genomic deletions [33,41,42] and reinsertion of eccDNAs after their formation has also been thought to generate structural variation [43,44]. Some evidence also indicates that eccDNAs could facilitate horizontal gene transfer [44]. Despite their potential as important facilitators of genetic and phenotypic plasticity and presence in all eukaryotes, research efforts, and especially whole circularome sequencing experiments, have been limited to model organisms and human cancer. Therefore, how these molecules behave across the tree of life and how different species could take advantage of these molecules to rapidly adapt to their environments have remained largely unknown.

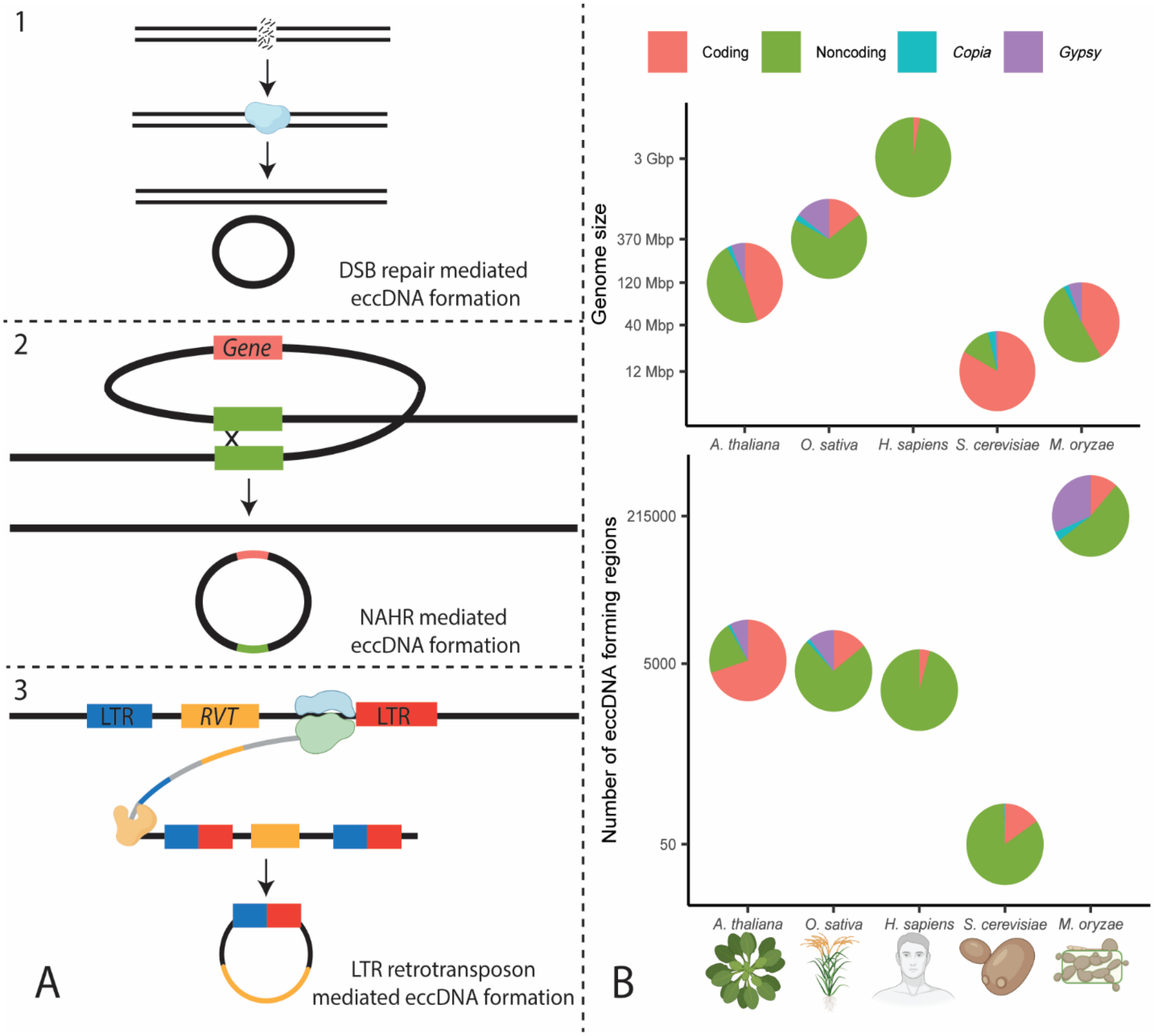


Fig. 1. Comparison of eccDNA formation in *M. oryzae* and other organisms. **A.** Examples of mechanisms of extrachromosomal circular DNA (eccDNA) formation. **1.** eccDNA formation as a result of double strand break repair. The blue enzyme represents several different types of DNA

repair mechanisms **2.** eccDNA formation as a result of non-allelic homologous recombination (NAHR). The green boxes represent homologous sequences. **3.** eccDNA formation as a result of LTR retrotransposon activity. The blue and green enzyme represents RNA polymerase, and the orange enzyme represents a reverse transcriptase (RVT). Rectangles that are partly blue and partly red represent hybrid LTRs formed from 5' and 3' LTRs during retrotransposition. DNA is drawn in black and RNA in gray. **B.** Comparison of genome size and number of eccDNA forming regions for *Arabidopsis thaliana* [45], *Oryza sativa* [46], *Homo sapiens* [41], *Saccharomyces cerevisiae* [47], and *Magnaporthe oryzae.* The number of eccDNA forming regions are shown as called by our pipeline in an average sample. Circularome data for *A. thaliana* and *O. sativa* leaf tissue, *H. sapiens* muscle tissue, and *S. cerevisiae* deletion collection samples are shown. The organism and protein icons were created with BioRender.com.

One of the greatest threats to food security is the devastation of crops by fungal plant pathogens. These pathogens secrete molecules known as effectors to modify host functions and cause disease [8]. The most promising solution to these diseases is the genetic modification of crops by introducing new disease resistance genes, often by allowing the crops to detect effectors and trigger immune responses [48]. Unfortunately, the deployment of disease resistant crops has often had only short-term success as some fungal pathogens have adapted to these defenses in very short time spans [10]. Similarly, fungicides are often used to mitigate the devastation caused by pathogens but fungi often evolve drug resistance [49]. A better understanding of how these pathogens adapt and overcome disease prevention efforts so quickly is vital to implementing future strategies. Sequencing and characterization of the genomes of fungal plant pathogens have implicated transposable elements [50], accessory chromosomes [13,51], and horizontal gene transfer [12]. Additionally, the compartmentalized genome architectures of some of these pathogens, commonly referred to as the "two-speed" genome, is thought to facilitate adaptation to stress by harboring stress response genes and disease associated genes, including effectors, in rapidly evolving regions of their genomes that contain few genes and many repetitive elements [11]. Given the potential for eccDNAs to be a source of phenotypic and genotypic plasticity, we sought to characterize the circularome of one of these pathogens to identify if eccDNAs could play a role in the rapid adaptation of the fungal plant pathogen, *Magnaporthe oryzae* (syn. *Pyricularia oryzae*).

M. oryzae, the causative agent of the rice blast disease [52], has been described as one of the most important fungal pathogens threatening agriculture [15] and is responsible for losses in rice crops equivalent to feeding 60 million people each year [53]. Its ease of culture as well as the importance of this pathogen for global food security have propelled *M. oryzae* to being one of the most studied plant pathogens; resulting in over three hundred sequenced genomes, transcriptomic and epigenetic datasets, as well as genetic tools including CRISPR/Cas9 mediated genome editing [54]. The availability of these extensive genomic datasets makes *M. oryzae* a prime candidate for understanding the role eccDNAs may play in adaptation to stress in a fungal plant pathogen.

We present here our analysis of circularome sequencing data for *M. oryzae* and identify eccDNA forming regions in its genome. We describe the high diversity of eccDNA forming regions that we found in the rice blast pathogen and compare it to previously sequenced circularomes. We find that most of the *M. oryzae* circularome is made up of LTR retrotransposon sequences and that genes on eccDNAs tend to originate from regions of the genome prone to presence-absence variation. Additionally, our characterization of the genes found on eccDNAs shows that many genes are never found on eccDNAs under the conditions we tested and suggests that selection may shape which genes are found on these molecules. Finally, our analysis reveals that many disease-causing effectors are found on eccDNAs in the pathogen.

Results

Identification of eccDNA forming regions in *Magnaporthe oryzae*

To characterize the circularome of *M. oryzae*, eccDNAs were purified and sequenced from pure cultures of *M. oryzae* Guy11 using a protocol adapted from previously published methods [46]. Briefly, after total DNA extraction of 3 biological replicates, linear DNA was degraded from 3 technical replicates for each biological replicate using an exonuclease, and the remaining circular DNA was amplified using rolling circle amplification (RCA). Depletion of linear DNA was verified with qPCR using markers to the *M. oryzae* actin gene (MGG_03982, Additional File 1: Fig. S1). This gene was used as a marker for linear DNA since increased copies of the ACT1 gene are thought to be deleterious in yeast [47,55]. Isolated eccDNAs were then sequenced using both paired-end Illumina sequencing and PacBio circular consensus sequencing (CCS). In total, we sequenced 8 samples as one technical replicate failed quality checks during library preparation. On average, Illumina sequencing yields were 6.5 Gbp per sample, and PacBio sequencing yields were 8 Gbp (subreads) and 500 Mbp (CCS) per sample.

To identify specific breakpoints indicating eccDNA formation in our Illumina sequencing data, we developed a pipeline inspired by previously published methods [41]. In circularome sequencing data, split mapping reads originate from sequencing circularization junctions of eccDNAs. Additionally, read pairs in the data that map in the opposite direction represent sequencing from paired-end sequencing fragments that span these circularization junctions. Our pipeline used split reads in combination with opposite facing read pairs to find evidence of eccDNA formation (Fig. 2). This allowed us to identify, with high confidence, genomic sequences belonging to eccDNAs, which we will hereafter refer to as “eccDNA forming regions.” We will refer to split reads associated with these eccDNA forming regions simply as “junction split reads.” Our analysis was limited to these eccDNA forming regions, rather than the fully resolved structure of each eccDNA molecule because of the complexity of eccDNAs and the techniques used to sequence them in this study. For example, eccDNAs can sometimes contain multiple copies of the same sequence [56] and our use of RCA, which generates long DNA fragments containing hundreds of tandem repeats of each circular molecule [57], prevents determination of whether a sequence is repeated many times on an eccDNA molecule or is just present once. Additionally, eccDNAs have also been shown to assemble with others, forming complex structures [58]. While our long-read PacBio sequencing may have been able to address this issue, our attempts at reference-free assembly of complete eccDNAs were unsuccessful, likely due to insufficient coverage of each molecule. While only eccDNA forming regions could be described in this study, these regions still enable a detailed description of the *M. oryzae* circularome. Across all 8 sequenced samples, our pipeline identified 1,719,878 eccDNA forming regions using Illumina paired-end sequencing data (Additional File 2). We validated 8 of these eccDNA forming regions using outward PCR and Sanger sequencing (Fig. 2 and Additional File 1: Fig. S2). These regions were chosen for validation as they fully contained genes of interest to the rest of the study, including well-known effectors.

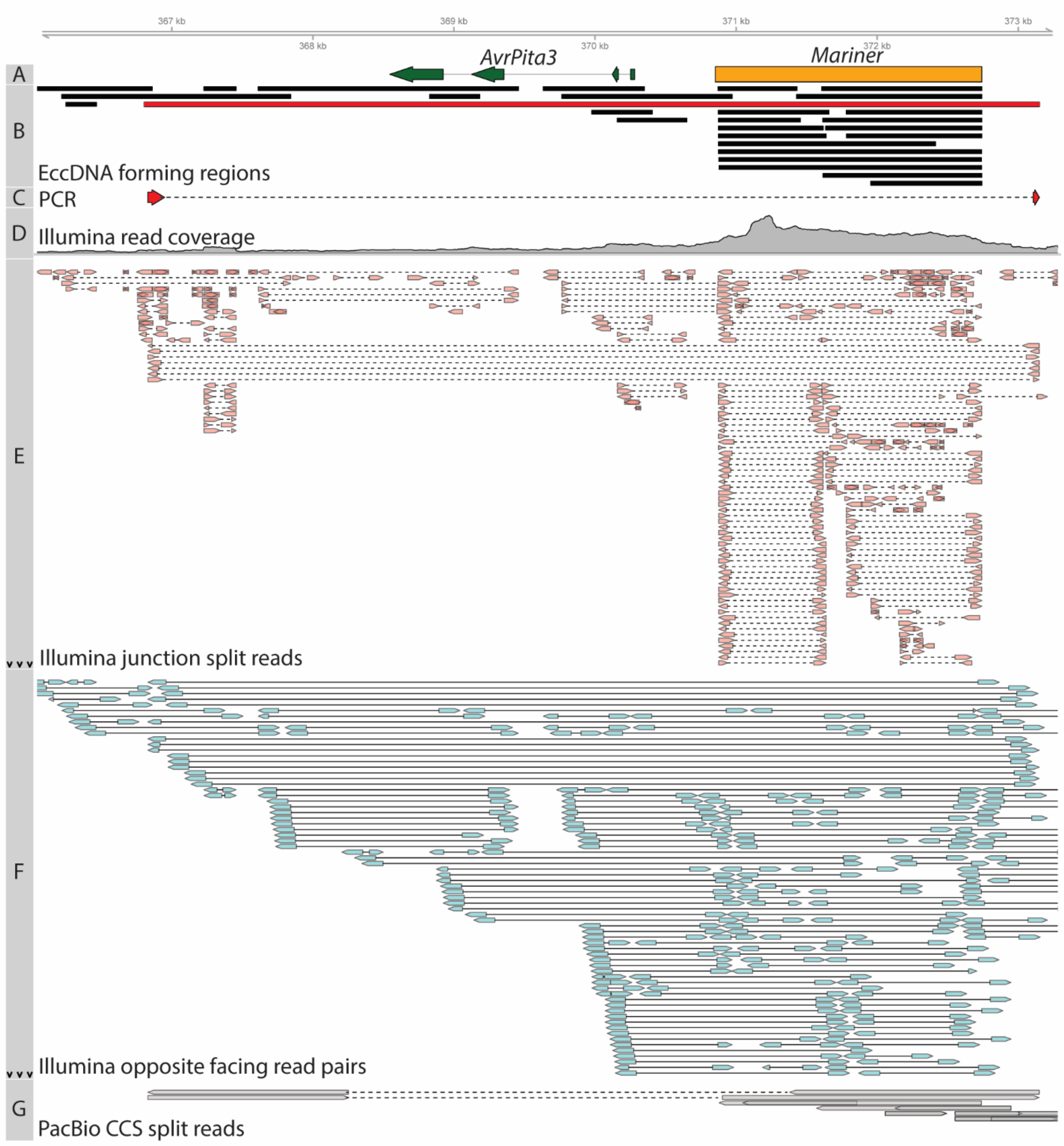


Fig. 2. Summary of evidence supporting an eccDNA forming region of interest in the *M. oryzae* genome. **A.** Location of effector *AvrPita3* and *Mariner* transposon. **B.** Location of eccDNA forming regions. The eccDNA forming region in red was chosen for validation using outward PCR. This eccDNA forming region was considered to fully encompass *AvrPita3*. **C.** Sanger sequencing read generated from outward PCR (Additional File 1: Fig. S2) that supports eccDNA

forming region highlighted in red in track B. **D.** Overall Illumina sequencing read coverage. **E.** Junction split reads obtained from Illumina data. Split reads are joined by a dashed line. Black arrows indicate not all reads were shown in areas with high counts. **F.** Opposite facing read pairs obtained from Illumina data. Read pairs are joined by a solid line. Black arrows indicate that not all reads were shown in areas with high counts. **G.** Split reads obtained from PacBio CCS data. Overlapping arrows indicate single reads mapped to the same location more than once. Split reads are joined by a dashed line. All data was obtained from a single sequenced sample (biological replicate 1, technical replicate A).

To determine how similar our technical and biological replicates were to each other, we compared the coordinates of eccDNA forming regions found in each sample. Overall, we found little overlap in eccDNA forming regions between technical replicates (14.16%, 10.09%, and 23.77%, for biological replicates 1, 2 and 3, respectively) and between biological replicates (9.41%) when comparing the exact start and end coordinates of these regions (Additional File 1: Fig. S3). Rarefaction analysis showed that these differences could be at least partially attributed to under sequencing, though this data could also be evidence of many low copy number eccDNAs being produced by the *M. oryzae* genome (Additional File 1: Fig. S4). However, principal component analysis using the coverage of junction split reads throughout the genome showed that technical replicates were more likely to be similar to other technical replicates within the same biological replicate than across biological replicates in the content of their eccDNA forming regions (Additional File 1: Fig. S5). Additionally, while exact coordinates of eccDNA forming regions did not have much overlap between samples, considering eccDNA forming regions whose start and end coordinates were within 100 bp of each other in two different samples to be the same increased this overlap greatly between technical replicates (48.46%, 45.55%, and 58.29% for biological replicates 1, 2 and 3, respectively) and between biological replicates (42.89%) (Additional File 1: Fig. S6). We performed a permutation analysis to simulate random formation of eccDNAs throughout the genome to verify that this result was meaningful and observed little overlap between replicates in this simulated scenario when increasing our overlap tolerance up to 100bp (Additional File 1: Fig. S6). All together, these results, as well as others presented throughout this study suggested that while the exact breakpoints of eccDNA forming regions were not identical across samples, the genomic loci, or hotspots, of eccDNA formation were highly similar.

Likely due to the great number of different eccDNAs in *M. oryzae*, the coverage of our PacBio sequencing data was too low to enable *de novo* assembly of eccDNA molecules. Therefore, we used our long read data to infer eccDNA forming regions by mapping them to the *M. oryzae* Guy11 genome and comparing these regions to those called using our short read data. This was done using a similar pipeline to the Illumina data with less stringent criteria which was better adapted to the lower read depth of the long read data. Our long read data allowed us to identify 147,335 eccDNA forming regions across all samples (Additional File 3). We compared these eccDNA forming regions to those called using Illumina data, allowing for up to a 10 bp difference between breakpoints to account for mapping ambiguity, and found that, on average, 81.42% of eccDNA forming regions called using PacBio data for one sample were also found in our eccDNA forming regions called using Illumina reads in the same sample (Additional File 1: Fig. S7). We were able to attribute much of this discrepancy to our stringent criteria for calling eccDNA forming regions since simply searching for split reads in our Illumina data increased this rate to 90.36% (Additional File 1: Fig. S7). The remaining differences are likely due to Illumina reads not being long enough to properly be mapped as split reads in certain regions of the genome. Such strong overlap between eccDNA forming regions called by long reads and short reads demonstrates the robustness of our short read data analysis. Aside from this validation, we chose not to include the PacBio data in our final analyses due to the low read depth.

Next, we quantified the potential false positive rate of our pipeline that could have originated from any undigested genomic DNA in our samples by running the pipeline on previously published whole genome sequencing data from *M. oryzae* Guy11 [19,54,59]. Based off the number of eccDNA forming regions called from this data, we estimated this false positive rate to be approximately 3 junction split reads per million sequencing reads (Additional File 4: Table S1). In comparison, we found 41,873 junction split reads per million reads in our eccDNA enriched samples, on average, indicating a very low false positive rate from our pipeline. Additionally, we could not completely rule out the presence of eccDNAs in the whole genome sequencing samples we analyzed. This validation showed that any remaining linear DNA in our samples after linear DNA degradation were unlikely to be called as eccDNA forming regions by our pipeline.

Finally, we benchmarked our pipeline on previously published eccDNA data in human tissue [41] (Additional Files 5 and 6). We found that, on average, 74.62% of eccDNA forming regions called by our pipeline were also described in the published dataset (Additional File 1: Fig. S8A). This number was even higher for eccDNA forming regions associated with 10 or more junction split reads (85.63%). The small fraction of eccDNA forming regions called by our pipeline that did not appear in the published list could not be attributed to how our pipeline handled multi-mapping reads (Additional File 1: Fig. S8A, see Methods) and were likely due to differences in sequence data processing and different criteria for selecting split reads between the two studies [41]. However, the two lists significantly differed in the number of eccDNA forming regions identified, with our pipeline identifying substantially less (Additional File 1: Fig. S8B). This difference can be attributed to our stricter evidence to call eccDNA forming regions. In our method, eccDNA forming regions were only called if split reads mapped to the region. This was in contrast to other methods of calling eccDNA forming regions which rely at least partly on peaks in sequencing coverage [41,47,60]. This meant that our pipeline could not detect eccDNAs formed from homologous recombination (HR) between identical repeats which do not result in split reads. We chose this method for *M. oryzae* because it showed circularome sequencing coverage throughout the entire genome in our samples and very few clear coverage peaks, which indicates that many low copy number eccDNAs were present in our samples. The high degree of overlap between our called eccDNA forming regions and those described by Møller *et al.* makes us confident that the eccDNA forming regions called using our pipeline are robust.

The *M. oryzae* circularome is more diverse and contains more noncoding sequences than the circularomes of other organisms

We were first interested in comparing the circularome of *M. oryzae* to those of other previously characterized organisms. To compare these datasets across different organisms, we gathered sequencing data from several previous studies [41,45–47] and reanalyzed them using our pipeline (Additional Files 5-20). Our analysis revealed a very large number of eccDNA forming regions in *M. oryzae* compared to other previously sequenced organisms (Fig. 1B). We also looked at the percentage of the genome that was found in eccDNA forming regions and found that while most organisms had 1-10% of their genome in eccDNA forming regions, our samples showed an average of 74.48% of the *M. oryzae* genome in eccDNA forming regions (Additional File 1: Fig. S9A). The difference in the number of eccDNA forming regions between organisms was still striking after normalizing for genome size and sequencing library size (Additional File 1: Fig. S9B). These results supported the idea that the low amount of overlap in eccDNA forming regions between our samples could be explained partly by the great number of eccDNAs produced by the *M. oryzae* genome. While the difference in the number of called eccDNA forming regions could be attributed to differences in the methods used for eccDNA purification (Additional File 4: Table S2), we extracted and sequenced eccDNAs from *Oryza sativa* and found similar levels of diversity to previously published samples (Additional File 1: Fig. S9B). We also found that *M. oryzae* had more eccDNA forming regions made up of noncoding sequences relative to the percentage of noncoding sequence in its genome than other organisms aside from *S. cerevisiae* (Fig. 1B, Additional File 1: Fig. S9C).

LTR retrotransposon sequences make up most of the *M. oryzae* circularome

Gypsy and *Copia* LTR retrotransposons frequently generate eccDNAs through several mechanisms [30–32], so we looked for the presence of these sequences in the *M. oryzae* circularome. Our analysis revealed that 54.12% of the eccDNA forming regions we identified were composed of more than 90% LTR retrotransposon sequence, indicating that these elements made up a large portion of the pathogen's circularome, despite only making up a small fraction of its genome (Fig. 1B, Additional File 1: Fig. S10). Further comparative analysis revealed that a much higher proportion of the *M. oryzae* circularome was made up of these LTR retrotransposon sequences than in other organisms (Fig. 1B, Additional File 1: Fig. S9D and S9E).

All six LTR retrotransposons identified in *M. oryzae* Guy11 formed eccDNAs (Fig. 3A). However, the elements *MAGGY*, *GYMAG1*, and *Copia1* made up the majority of the eccDNA sequencing data (Fig. 3B). When this data was normalized to the proportion of the genome made up by each transposon, *GYMAG1* stood out as making up a much greater percentage of the sequencing data than expected (Fig. 3C, Additional File 1: Fig. S11).

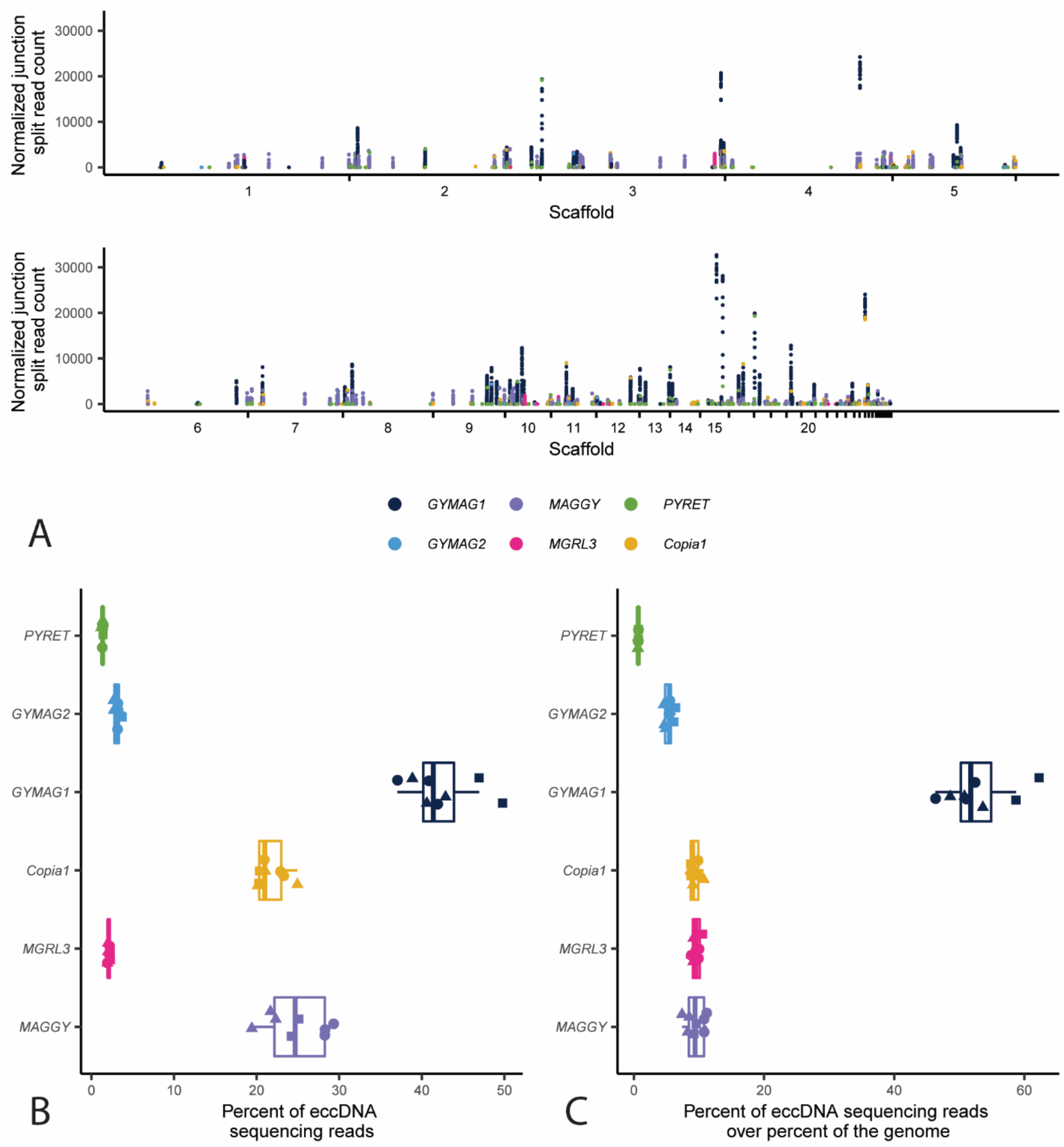


Fig. 3. The majority of eccDNAs in *M. oryzae* are made up of LTR retrotransposons. **A.** Manhattan plot showing the number of junction split reads per million averaged across biological replicates for all 100 bp bins that overlap an LTR retrotransposon in the *M. oryzae* Guy11 genome. Each point represents one of these bins. **B.** Boxplot showing the percentage of sequencing reads that map to LTR retrotransposons. Each point represents one sample, and the shape of the points represent the biological replicate that sample was taken from. **C.** Boxplot showing the ratio of the percentage of sequencing reads that map to LTR retrotransposons to the percentage of the *M. oryzae* Guy11 genome that is made up by that retrotransposon. Each

point represents one sample, and the shape of the points represent the biological replicate that sample was taken from.

LTR retrotransposons in *M. oryzae* form eccDNAs through a variety of mechanisms

LTR retrotransposons can form eccDNAs through a variety of mechanisms [30–32]. EccDNA formation commonly occurs after transcription and reverse transcription of the transposon which results in a linear fragment of extrachromosomal DNA [61] (Fig. 1A). Then, the most common circularization mechanisms are nonhomologous end joining (NHEJ) of the two LTR ends to form eccDNAs containing two LTRs (scenario 1, Fig. 4A), autointegration of the retrotransposon forming single LTR eccDNAs of various lengths, depending on where in the internal sequence of the transposon the autointegration event happens (scenario 2, Fig. 4B), and HR between the two LTRs to forming single LTR eccDNAs (scenario 3, Fig. 4C). Finally, LTR retrotransposon sequences can also become part of eccDNAs by other eccDNA formation mechanisms that do not rely on retrotransposition activity, such as intrachromosomal HR between solo-LTRs or between multiple copies of the same transposon [32,33,47]. Given this diversity of mechanisms, we wanted to evaluate which of them contributed to eccDNA formation in *M. oryzae.* To do this, we first simulated the expected read coverage for each of the three active LTR eccDNA formation mechanisms under ideal conditions where only one mechanism of formation was occurring (Fig. 4A-C). Then, we measured the prevalence of scenarios 1 and 2 by identifying specific split read variants in our data. LTR eccDNAs formed through NHEJ result in split reads that map to one end of an LTR and the other which we will refer to as LTR-LTR split reads (Additional File 1: Fig. S12 and S13A). Autointegration results in split reads that map to one LTR and to the internal region of the transposon which we will refer to as LTR-internal split reads (Additional File 1: Fig. S13B and S14). HR between two identical LTRs (scenario 3) would not result in a split read so we could not find this type of evidence in our data.

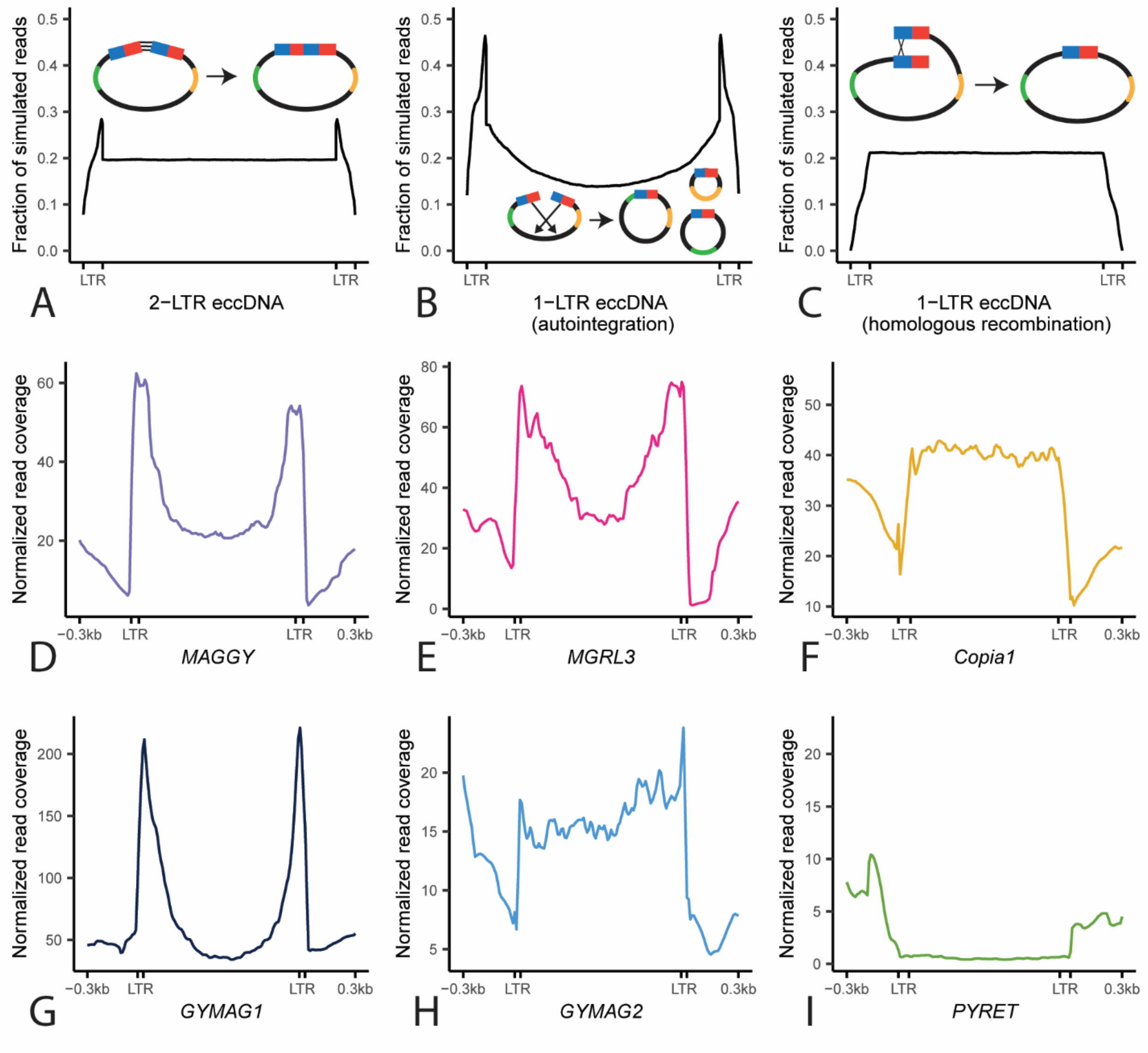


Fig. 4. LTR retrotransposons in *M. oryzae* form eccDNAs through a variety of mechanisms. **A-C.** Profile plots showing expected sequencing read coverage for each LTR retrotransposon eccDNA formation scenario as well as graphical representations of the scenario. In the graphics, blue and red rectangles represent hybrid LTRs formed from 5' and 3' LTRs during retrotransposition and green and orange lines represent areas of the internal region of the retrotransposon with distinct sequences. **D-I.** Profile plots showing observed sequencing read coverage for each LTR retrotransposon found in the *M. oryzae* Guy11 genome.

Comparisons between simulated and observed read coverage plots revealed contributions of several eccDNA formation mechanisms that varied by transposable element. For *MAGGY*, our analysis indicated that it forms eccDNAs primarily through autointegration (Fig. 4D). This was supported by a high correlation between the number of sequencing reads and LTR-internal split reads (Additional File 1: Fig. S13A) and a low correlation between sequencing reads and LTR-LTR split reads (Additional File 1: Fig. S12A). The data also pointed to *MGRL3* and *GYMAG1* forming eccDNAs primarily through autointegration (Fig. 4E and 4G, Additional File 1: Fig. S12BD and

S13BD). *Copia1*, on the other hand showed a clear pattern of read coverage corresponding to eccDNA formation through HR (Fig. 4F), though the high correlation between sequencing reads and LTR-internal split reads mapping to this element hinted that a small, but proportional, fraction of *Copia1* elements formed eccDNAs through autointegration (Additional File 1: Fig. S13C). In the case of *GYMAG2*, its sequencing read coverage resembled a pattern expected for LTR-eccDNAs formed through NHEJ (Fig. 4H). The large amount of LTR-LTR split reads per million mapped reads found corresponding to *GYMAG2* elements compared to other retrotransposons supported this inference (Additional File 1: Fig. S14A). *PYRET*'s distinct sequencing read coverage profile likely indicated that it mostly formed eccDNAs by other eccDNA formation mechanisms that do not rely on retrotransposition activity such as intrachromosomal HR (Fig. 4I). A low correlation between sequencing read coverage and both LTR-LTR split reads and LTR-internal split reads, as well as the fragmented nature of *PYRET* elements, which is a sign of low recent retrotransposon activity, supported this inference (Additional File 1: Fig. S12F and S13F). Finally, to determine whether the results we obtained were caused by bias in the length and completeness of the retrotransposon sequences in the *M. oryzae* genome, we generated profile plots for each retrotransposon using previously generated whole genome sequencing data [19,54,59]. The results from this analysis ruled out this possibility (Additional File 1: Fig. S15). In conclusion, it is clear that a variety of eccDNA formation mechanisms contributed to eccDNAs containing LTR retrotransposon sequences, and that these mechanisms varied by element.

MicroDNAs are distinct from other eccDNAs

MicroDNAs have previously been studied as a distinct set of molecules within the eccDNA category. Besides being small (less than 400bp), microDNAs are found to be enriched in genic regions, exons, 5'UTRs and CpG islands [42,62]. We examined if microDNAs in *M. oryzae* showed these characteristics by analyzing eccDNA forming regions less than 400 bp in length with less than 10% LTR retrotransposon sequence across different organisms. Enrichment of microDNAs in CpG islands was the most consistent result across all organisms we analyzed, though this enrichment was not found in *M. oryzae* (Additional File 1: Fig. S16). Similarly, we found no enrichment of microDNAs in 5'UTRs in *M. oryzae*. We did however find a small enrichment of microDNAs in genic regions in *M. oryzae* as seen in many of the other sequenced organisms (Additional File 1: Fig. S16 and S17). In general, our analysis suggested that the previously described characteristics of microDNAs are not common across all organisms and sample types.

MicroDNAs also displayed distinct features from the remaining subset of non-LTR eccDNAs which we called large eccDNAs. Among other differences, we found that, unlike microDNAs, large eccDNAs tended to be enriched in intergenic regions (Additional File 1: Fig. S17 and S18). Additionally, eccDNAs are often associated with active transcription [29,37], and we found a slight but significant correlation between expression and junction split reads for large eccDNAs but not for microDNAs (Additional File 1: Fig. S19).

In yeast, eccDNA amplification is thought to often occur with the help of autonomously replicating sequences (ARSs) which contain ARS consensus sequences (ACSs) [33,47,63]. In *M. oryzae*, we found that ACSs were enriched in large eccDNAs (permutation test, mean of expected: 5320.14 regions, observed: 6950 regions, $p < 0.01$, n = 100 replicates) but depleted in microDNAs (permutation test, mean of expected: 818.09 regions, observed: 714 regions, $p < 0.01$, $n = 100$ replicates). However, for both large eccDNAs and microDNAs, presence of an ACS in the eccDNA forming region did not result in an increased number of junction split reads (Additional File 1: Fig. S20). Finally, microDNAs have been found to be associated with chromatin marks and increased GC content [42,62]. However, we did not find any of these enrichments in microDNAs or large eccDNAs in *M. oryzae* (Additional File 1: Fig. S21).

Many genes are found encompassed by eccDNA forming regions

Many eccDNAs contain genes, and these eccDNAs can provide genotypic and phenotypic plasticity in other organisms. In *M. oryzae* we found that, out of the 12,115 genes in Guy11, 9,866 were fully contained by an eccDNA forming region in at least one sample (Fig. 2B and 5A). These genes included *TRF1* (MGG_04843) and *PTP2* (MGG_00912) which have been shown to be involved in fungicide resistance in *M. oryzae* [64,65]. EccDNA forming regions containing these two genes were validated using outward PCR (Additional File 1: Fig. S2). However, not all genes were observed in eccDNA forming regions at the same frequency, and their presence on eccDNAs was heterogenous across samples. To further understand what types of genes are enriched in eccDNA forming regions, we focused on a robust set of eccDNA-associated genes. To identify these genes, we first counted the number of times each gene was found fully contained by a junction split read in each sample. We referred to this count as the number of “encompassing split reads” for each gene. We then normalized this count to the number of junction split reads in each sample and averaged it across technical replicates for each biological replicate. Finally, we sorted the genes by their prevalence in each biological replicate and chose genes that were found in the top third of genes for this count in all three biological replicates. In total, using these metrics, we identified 558 eccDNA-associated genes shared across all biological replicates (Fig. 5A, Additional File 1: Fig. S22 and Additional File 21).

To identify biological processes enriched in eccDNA-associated genes, we performed gene ontology (GO) enrichment analysis. We found that eccDNA-associated genes were enriched for GO terms related to vesicle transport, mitosis, and the cytoskeleton among other terms (Fig. 6A, Additional File 1: Fig. S23 and Additional Files 22-24). We also explored whether eccDNA-associated genes showed differences in gene expression or other genomic features from other genes. However, we found no difference between eccDNA-associated genes and other genes in gene expression, GC content, or histone marks, aside from a significant difference in H3K36me3 (Additional File 1: Fig. S24 and S25).

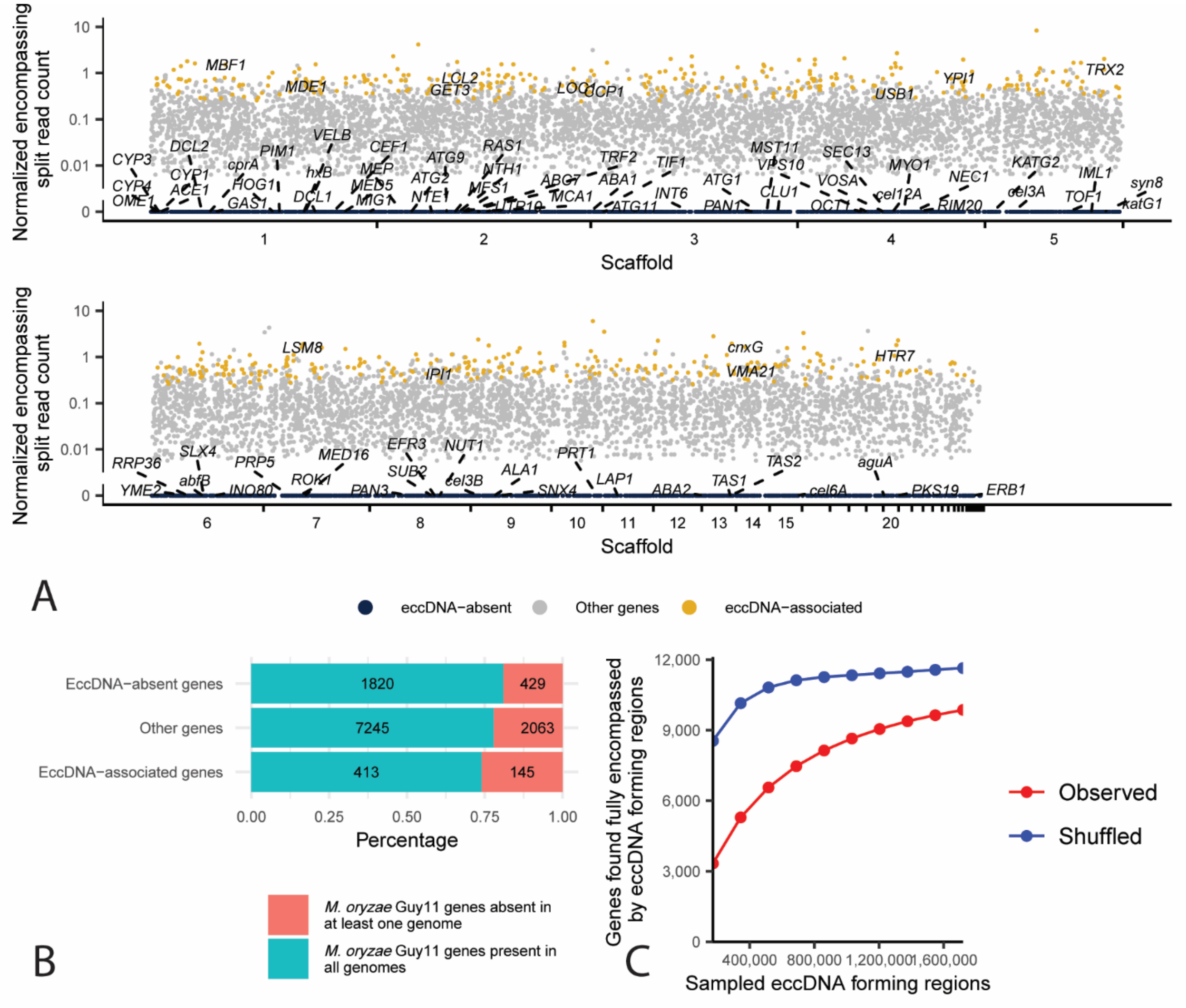


Fig. 5. EccDNA forming regions contain most *M. oryzae* genes, but not all, and many are associated with presence-absence variation. **A.** Manhattan plot showing the number of encompassing split reads per million junction split reads averaged across biological replicates for each gene in the *M. oryzae* Guy11 genome. Each dot represents one gene. EccDNA-associated genes with known gene names are labeled according to their normalized encompassing split read count and position in the genome. EccDNA-absent genes with known gene names are labeled with lines pointing to their location in the genome. **B.** Stacked bar plot showing the percentage of eccDNA-absent genes, other genes, and eccDNA-associated genes in the *M. oryzae* Guy11 genome that had an ortholog in all other 162 *M. oryzae* genomes analyzed or not. Numbers indicate the number of genes in each category. **C.** Rarefaction analysis of the observed number of genes found fully encompassed by eccDNA forming regions at different subsamples of all found eccDNA forming regions, compared to the same number of randomly selected genomic regions.

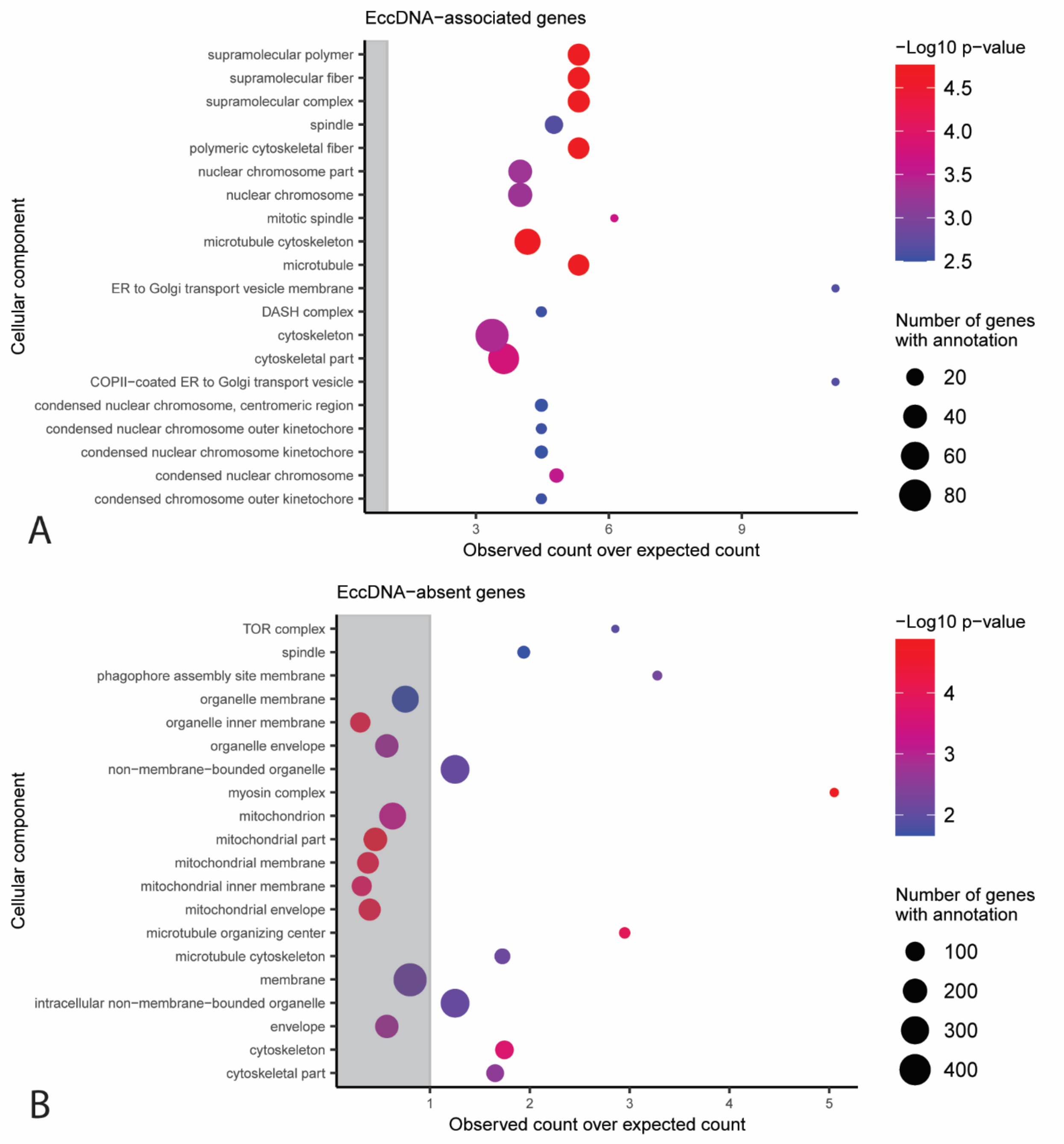


Fig. 6. Gene Ontology (GO) terms associated with eccDNA-associated and eccDNA-absent genes in *M. oryzae*. Functional categories in the cellular component GO with an observed number of **A.** eccDNA-associated genes or **B.** eccDNA-absent genes that is significantly different from the expected number with correction for gene length bias. The y-axis shows the different functional categories, and the x-axis represents the observed number of genes divided by the expected number of genes in this group. Dots outside of the grey rectangle represent functional categories that are observed more often than expected. The size of dots indicates the total number of genes in the *M. oryzae* genome that belong to each functional category. Only the 20 categories with the largest -log10 p-values according to a Chi-square test are shown.

EccDNA-associated genes are closer to gene sparse and repeat dense regions of the genome than other genes

Some plant pathogens are described as having “two-speed” genomes with housekeeping genes found close together in repeat-poor regions and environmentally responsive and disease-associated genes found in repeat-dense and gene-poor regions [11]. To determine if eccDNA-associated genes were enriched in either of these genomic contexts, we analyzed if eccDNA-associated genes were more distant from other genes than expected by chance (Fig. 7). We observed a significant difference (permutation test for difference of medians, p = 0.0117, n = 10,000 replicates) between the median distance to the nearest gene of eccDNA-associated genes (543 base pairs) and other genes (485 base pairs). We also observed a significant difference (permutation test for difference of medians, p = 0.0282, n = 10,000 replicates) between the median distance to the nearest genomic repeat of eccDNA-associated genes (663 base pairs) and other genes (769 base pairs, Additional File 1: Fig. S26). This difference in proximity was not observed for transposable elements, indicating that transposable elements alone were not responsible for this effect (Additional File 1: Fig. S27). The heterogeneity of eccDNAs and the mechanisms of their formation might be influencing this comparison. However, our data points to a link between genome architecture and eccDNA formation.

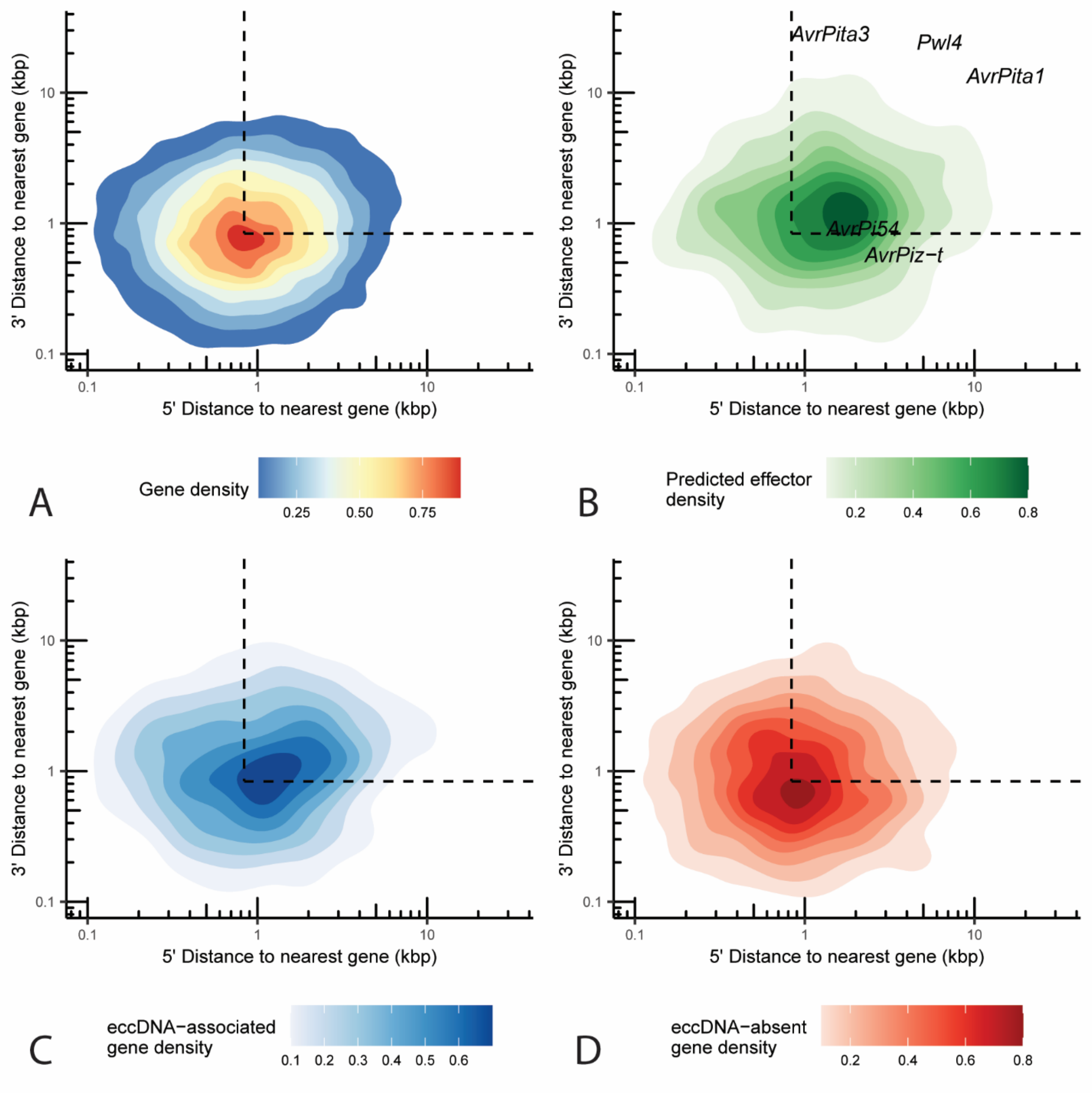


Fig. 7. EccDNA-associtated genes are often found in gene sparse regions of the *M. oryzae* genome. Two-dimensional density plot representing the 5' and 3' distance to the nearest gene in the *M. oryzae* Guy11 genome in kilobase pairs for each **A.** gene, **B.** predicted effector, **C.** eccDNA-associated genes, and **D.** eccDNA-absent genes. Known effectors are shown as text in **B.** Dashed lines represent median 5' and 3' distance to nearest gene.

EccDNA-associated genes are more prone to presence-absence variation than other genes

There is evidence of eccDNAs generating structural variation in other organisms [43,44]. We therefore tested whether eccDNA formation is associated with genes prone to presence-absence variation in 162 rice-infecting *M. oryzae* isolates (Additional File 25). As expected from previous studies [25,27], our analysis indicated that predicted effectors were more likely to experience presence-absence variation (Additional File 1: Fig. S28; X-squared = 146.33, df = 1, p-value < 2.2e-16). We also found that eccDNA-associated genes were more likely to be prone to presence-absence variation (Fig. 5B; X-squared = 16.262, df = 2, p-value = 2.95e-04). This result suggested that eccDNA formation and structural variation occur in similar regions of the genome but did not show whether they are directly linked. To see if a more direct link existed, we surveyed the genomes of the *M. oryzae* isolates for small deletions that completely or partially overlapped genes but did not disrupt neighboring genes. We were able to identify 257 such events (Additional File 26). However, none of these deletions matched our eccDNA forming regions and only 8 of them came within 50 bp. Our rarefaction analyses revealed that there is likely to be a much greater diversity of eccDNAs than what we were able to capture at the sequencing depth of this study, whether we considered samples individually or as a whole (Additional File 1: Fig. S4 and S29). Therefore, eccDNA formation that could have contributed to structural variation might have been missed due to either under sequencing or absence in the conditions tested in this study.

Similarly, we were interested in identifying any potential DNA translocations that may have occurred through an eccDNA intermediate. While we were able to successfully construct a bioinformatics pipeline that identified one previously described eccDNA-mediated translocation in wine yeast [44] (Additional File 1: Fig. S30), we were unable to identify any such examples in the *M. oryzae* genomes analyzed despite including isolates infecting a variety of hosts in this analysis (306 genomes in total, Additional File 27).

Finally, since mini-chromosomes have been hypothesized as playing important roles in fungal plant pathogen evolution, we also sought to determine whether genes that were previously found on *M. oryzae* mini-chromosomes were associated with eccDNA formation but found no such effect (Additional File 1: Fig. S31).

Many eccDNA-absent genes are myosin-complex related

Since most *M. oryzae* genes appeared in eccDNA forming regions in at least one sample, we were particularly interested in the 2,249 genes that never appeared fully encompassed by an eccDNA forming region in any of our technical or biological replicates, which we called eccDNA-absent (Fig. 5A, Additional File 21). We first verified that eccDNA-absent genes were not caused by insufficient sequencing coverage using rarefaction analysis. This analysis differed significantly from our previous ones (Additional File 1: Fig. S4 and S29). Here, we counted the number of genes found in eccDNA forming regions at various subsamples of eccDNA forming regions. This analysis revealed that our observations of eccDNA-absent genes were unlikely to be caused by the under sequencing we described previously as the number of genes found fully encompassed by eccDNA forming regions appeared to plateau at larger subsamples of eccDNA forming regions (Fig. 5C). Additionally, a permutation analysis showed that, given the high coverage of our data, we only expected to find 468 genes in this category by chance, which is far fewer than the 2,249 genes we observed (Fig. 5C).

We next explored whether gene expression or other genomic features could explain the observed eccDNA-absent genes. However, we found no strong differences between eccDNA-absent genes and other genes in gene expression, GC content, or histone marks (Additional File

1: Fig. S24 and S25). EccDNA-absent genes also did not differ from other genes in terms of their distance to the nearest gene, repeat or transposable element (Fig. 7, Additional File 1: Fig. S26 and S27).

Finally, we performed GO enrichment analysis on these genes and found, amongst many other enriched terms, that genes related to cytoskeletal proteins, and especially the myosin complex, were enriched within eccDNA-absent genes (Fig. 6B, Additional File 1: Fig. S32, and Additional Files 28-30). While genes related to the cytoskeleton were also enriched among eccDNA-associated genes, these were related to mitosis and microtubule polymerization, rather than the myosin complex (Fig. 6A, Additional File 1: Fig S23). This result is of particular interest given that the actin gene has also been used in a previous study [47] as a marker for linear DNA due to its negative fitness effect at high copy numbers in yeast [55]. As expected, the *M. oryzae* actin gene (MGG_03982) was one of the eccDNA-absent genes, meaning it was never found in an eccDNA forming region in its entirety in any of our samples. Furthermore, in agreement with our GO enrichment results, *MYO1* was another eccDNA-absent gene. To validate our bioinformatics analysis, we tested whether we could amplify the full sequences of these genes from our eccDNA samples using PCR. In agreement with our findings, we were only able to amplify these sequences from our genomic DNA sample (Additional File 1: Fig. S33). These results suggested that eccDNA formation is not random in *M. oryzae* and that certain groups of genes may be protected from eccDNA formation or maintenance of these eccDNAs in the cell.

Effectors are enriched in eccDNA forming regions compared to other genes

Finally, we wanted to identify whether eccDNA forming regions contained disease-causing effectors. We found that many known *M. oryzae* effectors were encompassed by eccDNA forming regions in at least one sample. This included *AvrPita3*, *AvrPita1*, *AvrPi9*, *AvrPi54*, *AvrPiz-t*, and *Pwl4* (Fig. 2,8, and Additional File 21). We validated eccDNA forming regions containing these effectors using outward PCR (Additional File 1: Fig. S2). Additionally, we found that many predicted effectors were found in eccDNA forming regions (Fig. 8 and Additional File 21). We also found that many of these putative effectors were associated with larger numbers of encompassing split reads and found this difference to be statistically significant (Additional File 1: Fig. S34; permutation test for difference in medians, $p < 0.0001$, n = 10,000 replicates). Effectors are often small genes and given the often-small size of eccDNA forming regions in our data ,which may have been caused by the bias of RCA towards small molecules [29,66] (Additional File 1: Fig. S35), we felt that our analysis could be affected by this bias. To address this issue, we repeated our permutation test, comparing predicted effectors to a set of non-effectors of similar lengths, and again found a significant difference in number of encompassing split reads (permutation test for difference in medians with correction for gene length distribution, $p = 0.0206$, n = 10,000 replicates). This result suggests that effectors are more likely to be found on eccDNAs than other genes in *M. oryzae,* and that this effect is not simply due to their size. Additionally, a small proportion of effectors are found among our eccDNA-absent genes (Fig. 8). These candidates might be more evolutionarily stable and therefore useful as targets for disease resistance.

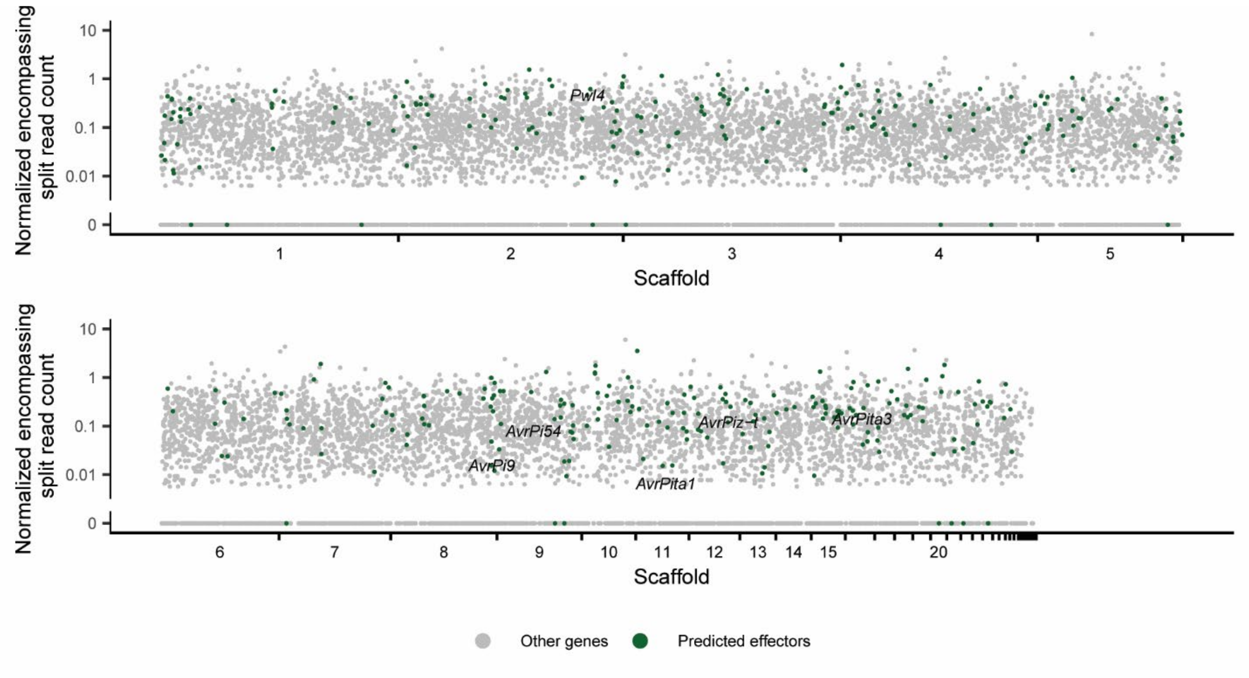


Fig. 8. Effectors are enriched in eccDNAs in *M. oryzae*. Manhattan plot showing the number of encompassing split reads per million reads averaged across biological replicates for each gene in the *M. oryzae* Guy11 genome. Each dot represents one gene. Predicted effectors are shown in green and known effectors are shown as text.

Discussion

EccDNAs have been shown to be a source of significant phenotypic [33,34,37,39,40] and genotypic [43,67] plasticity that can help organisms adapt to stress. While eccDNAs have been extensively studied in human cancer [29], very few studies have attempted to study the circularome of other organisms, and even fewer have generated high quality whole circularome sequencing data. To expand our understanding of eccDNAs across the tree of life, we studied the circularome of the fungal plant pathogen *M. oryzae* and, developed many tools to analyze whole circularome sequencing data, which can often be difficult to interpret. These include a new pipeline to identify eccDNA forming regions and frameworks for comparing this data across organisms, identifying mechanisms of eccDNA formation of LTR retrotransposons, identifying gene sets enriched or depleted in eccDNAs, and identifying structural variants that may have been caused by eccDNAs. Our analysis also revealed that the circularome of *M. oryzae* contains a wide diversity of eccDNA forming regions that appeared to exceed those of other previously characterized organisms. This wide diversity likely contributed to the under sequencing of our samples and a small overlap in exact eccDNA forming regions across samples. However, our analysis throughout this study showed that our samples clustered tightly together with regards to various features of the circularome, indicating that while exact eccDNA forming breakpoints were mostly not shared across samples, eccDNA formation hotspots were. We also found that eccDNA forming regions in *M. oryzae* were more commonly made up of LTR retrotransposons than other organisms. Though the results of our comparative analysis need to be verified using standardized protocols, these differences highlight the need to further characterize the circularome of other eukaryotes to obtain a better understanding of how they differ. Additionally, it is important to note that the data analyzed in this study only represent snapshots of the circularomes of the organisms described and could vary greatly across developmental stages and environmental stresses that were not included in these analyses. Further studies of eccDNAs across these different conditions are necessary to definitively describe and compare these molecules across organisms.

We analyzed the types of genes that were found on eccDNAs in *M. oryzae* and found that eccDNA-associated genes were often prone to presence-absence variation, hinting at a link between eccDNAs and genomic plasticity. However, we could not find direct evidence of gene deletions occurring through an eccDNA intermediate in *M. oryzae*. Similarly, we could not find any evidence of eccDNA-mediated translocations. These results could be due to our sequencing coverage and our bioinformatics pipelines not showing the full diversity of eccDNAs in *M. oryzae*. For example, our pipeline was unable to detect eccDNAs formed from HR between perfect repeats. Additionally, our scripts were able to identify an eccDNA-mediated translocation in wine yeasts but were limited to non-repetitive regions of the genome and may have missed some of these events in those regions in *M. oryzae*. Finally, it is possible that eccDNA-mediated translocations occur on a larger time scale than what we were able to sample within the *M. oryzae* species. However, it is likely that experimental approaches, such as inducing the formation of specific eccDNAs, are necessary to determine whether these events lead to chromosomal deletions or rearrangements. On a genome-wide scale, single cell sequencing of the circularome as well as genomic DNA could also lead to a more precise view of eccDNA formation and structural variation as they occur in the cell during vegetative growth. These techniques will likely also need to be paired with amplification-free eccDNA sequencing protocols as well as high coverage, long read sequencing to fully resolve the structure of eccDNA molecules. Additionally, we found that eccDNA-associated genes presented characteristics associated with the gene-sparse, repeat-rich, and “fast” part of the plant pathogen genome where rapid adaptation to stress occurs [11]. The fact that eccDNA-associated genes were closer to repeats than other genes, but not transposons specifically, indicated that this effect was not simply caused by eccDNA formation by LTR retrotransposons. We also found that predicted effectors were enriched in eccDNA forming regions. These results show that eccDNA formation occurs in the same genomic contexts as rapid genome evolution in *M. oryzae* and could also point to eccDNAs directly playing a role in the plasticity of important genes like effectors.

We also identified a set of eccDNA-absent genes, which were never found fully encompassed by eccDNA forming regions under our experimental conditions. This observation was not explained by incomplete sequencing. Histone marks, expression and proximity to repetitive DNA did not appear to set these genes apart either. Though it is possible that other factors contribute to this phenomenon and directly prevent eccDNA formation in these regions, our data indicates that eccDNA formation in *M. oryzae* is not a random process and hints at selective pressure acting against cells that accumulate high copy numbers of these genes through eccDNA formation. This idea is supported by the absence of genes related to the myosin complex, which are deleterious at high copy numbers in other organisms.

Selective pressure during growth under stress could favor *M. oryzae* cells containing higher copy numbers of genes important for survival under these conditions as has been extensively shown in other organisms [33,34,37,39,40]. For example, we identified two genes associated with fungicide resistance in our eccDNA forming regions which, if amplified, could lead to drug resistance, as previously observed [34,39,40]. Further experimentation and characterization of the *M. oryzae* circularome under stress is necessary to investigate if this eccDNA-mediated phenotypic plasticity is present in the plant pathogen. These experiments could also be used to assess how LTR retrotransposon activity changes in response to stress in *M. oryzae* and how the mechanisms of eccDNA formation that we described might be affected. We attempted to perform such experiments by sequencing *O. sativa* tissue infected by *M. oryzae* but found that *O. sativa* eccDNAs crowded out the circularome sequencing signal and prevented meaningful analysis, highlighting the need for a dedicated enrichment or single cell sequencing protocol. Additionally, analyzing the biological significance of the amplification of specific genes on eccDNAs, especially across treatments, may prove challenging and will require further tool development. For example, the same genes may be on eccDNAs of varying sizes and composition across samples. Multiple genes could also be on each eccDNA, further complicating the analysis. The complexity of eccDNAs combined with the limitations of current eccDNA sequencing techniques severely limits the analysis of circularome sequencing data, which is why we chose to focus our analysis on hotspots of eccDNA formation and groups of genes, rather than individual genes. In the future, high coverage, long read sequencing of eccDNAs collected without amplification will likely be necessary to perform more thorough analyses of eccDNAs; and this type of study is likely to become the gold standard for the field once cost is no longer prohibitive.

Conclusions

This study commences the characterization of the *M. oryzae* circularome and highlights its potential for generating phenotypic and genotypic plasticity. If eccDNAs were to facilitate these phenomena, they could become potential drug targets to prevent the rapid adaptation of the blast pathogen to environmental stress, fungicides, and resistant crop varieties. Furthermore, regions and genes prone to forming eccDNAs could be excluded as drug targets or as targets for engineered resistance in crops. On the other hand, we found 1,820 genes including several predicted effectors in the *M. oryzae* genome that were in the eccDNA-absent group and were conserved in all other rice infecting isolates that we analyzed. These genes could be high potential targets for fungicide design or engineered resistance. Our study also describes the great diversity of eccDNAs and the enrichment of LTR retrotransposons in the *M. oryzae* circularome. These observations, in addition to the potential consequences of eccDNA formation, highlights the need to study these molecules in more organisms, including other fungal plant pathogens.

Methods

M. oryzae cultures and DNA extraction

M. oryzae Guy11 was grown on Difco oatmeal agar plates for 21 days under constant light in a Percival Scientific Incubator Model CU-36L4 equipped with half fluorescent lights and half black lights. 1 cm^2 of mycelium was scraped from the colony edge and used to start 3 liquid cultures (biological replicates) in petri dishes with 15 ml complete medium [68] . Liquid cultures were incubated without shaking for 3 days in the same growth chamber.

Total DNA extraction was performed according to a protocol from the Prof. Natalia Requena group at the Karlsruhe Institute of Technology. Briefly, mycelium grown in liquid culture was washed 3 times with water and then ground in liquid nitrogen. Ground mycelium was incubated in extraction buffer (0.1M Tris-HCl pH 7.5, 0.05 M EDTA, 1% SDS, 0.5 M NaCl) at 65°C for 30 minutes. 5M potassium acetate was then added to the samples which were then incubated on ice for 30 minutes. The supernatant was then washed with isopropanol and ethanol. Finally, the DNA pellet was resuspended in water and treated with RNase A (Thermo Scientific).

O. sativa growth and DNA extraction

O. sativa samples were originally intended to serve as control samples to be compared to tissue infected by *M. oryzae* and therefore the methods below reflect this original intent. However, circularome sequencing data obtained from infected tissue was not included in this study as it included very little sequencing data that mapped to the *M. oryzae* Guy11 genome.

O. sativa cv. Nipponbare seeds were surface sterilized in 70% ethanol for 1 minute and 10% bleach for 10 minutes with thorough rinsing in sterile deionized water after each before being placed on wet filter paper in a petri dish. The petri dish was wrapped in foil and placed at 4°C for 2 days to germinate. Germinated seedlings were planted in potting mix made up of 50% Turface and 50% Super Soil. Seedlings were grown for three weeks in a greenhouse under standard conditions. For three samples, the first true leaf was cut from one rice plant, its tip removed, and then cut into two equal segments, approximately 10mm in length. This pair of segments was then placed on their abaxial surface on wet filter paper in a petri dish. Five hole-punches of filter paper soaked in 0.25% gelatin and 0.05% Tween-20 were then placed on each segment. The petri dishes were placed in an airtight container with wet paper towels and then placed on a windowsill for 7 days. Hole-punches were removed and non-chlorotic tissue in contact with hole-punches was ground in liquid nitrogen. DNA was extracted using the Qiagen Plant DNeasy mini kit.

Circular DNA enrichment

Total DNA obtained from DNA extractions (biological replicates) were then split into three samples (technical replicates) before circular DNA enrichment. This enrichment was performed according to a protocol from Lanciano *et al.* with a few modifications [46]. 5 µg of extracted DNA was used as input for circular DNA enrichment in *M. oryzae,* and 750 ng of extracted DNA were used for *O. sativa*. To purify the samples and begin removing large linear DNA fragments, the samples were treated using a Zymo Research DNA Clean and Concentrator kit with standard protocols. Linear DNA digestion was then performed using Epicentre PlasmidSafe DNase and incubated at 37°C for 24 hours. DNase, ATP, and reaction buffer were then added to the samples every 24 hours throughout the duration of the incubation. In total, the reaction was allowed to proceed for 96 hours. Remaining DNA was then precipitated overnight at 4°C by adding 0.1 volume 3M sodium acetate, 2.5 volumes ethanol and 1 µl glycogen (20 mg/ml). Rolling circle amplification was then performed using the Illustra TempliPhi 100 Amplification Kit (GE Healthcare). Precipitated DNA was resuspended directly in 20 µl of the Illustra TempliPhi sample buffer and the amplification reaction was allowed to proceed for 24 hours at 30°C.

Verification of circular DNA enrichment

In a separate experiment, 5 samples of *M. oryzae* mycelium were grown up in liquid culture and total DNA was extracted. Circular DNA enrichment was performed as before with some exceptions and without technical replicates. First, linear DNA digestion was only performed for 72 hours for 3 samples. Next, aliquots of the incubating samples were taken at 0 hours, 24 hours, 48 hours and 72 hours for these 3 samples, and 0 hours, 48 hours, 72 hours and 96 hours for the last 2 samples. qPCR was then used to verify linear DNA depletion in each sample using an Applied Biosystems QuantStudio 5 instrument and the QuantStudio Design and Analysis desktop software. Primers were used to amplify a portion of the *M. oryzae* actin gene (MGG_03982) along with Lightcycler 480 Sybr Green I master mix (Additional File 4: Table S3). Data from four qPCR technical replicates was obtained. Remaining linear DNA fraction in each sample at each timepoint was then calculated using the $2^{-\Delta\Delta Ct}$ method.

Illumina library preparation and sequencing

Library preparation was performed by the QB3-Berkeley Functional Genomics Laboratory at UC Berkeley. DNA was fragmented with an S220 Focused-Ultrasonicator (Covaris), and libraries prepared using the KAPA Hyper Prep kit for DNA (Roche KK8504). Truncated universal stub adapters were ligated to DNA fragments, which were then extended via PCR using unique dual indexing primers into full length Illumina adapters. Library quality was checked on an Agilent Fragment Analyzer. Libraries were then transferred to the QB3-Berkeley Vincent J. Coates Genomics Sequencing Laboratory, also at UC Berkeley. Library molarity was measured via quantitative PCR with the KAPA Library Quantification Kit (Roche KK4824) on a BioRad CFX Connect thermal cycler. Libraries were then pooled by molarity and sequenced on an Illumina NovaSeq 6000 S4 flowcell for 2 x 150 cycles, targeting at least 10Gb per sample. FastQ files were generated and demultiplexed using Illumina bcl2fastq2 version 2.20 and default settings, on a server running CentOS Linux 7. One technical replicate did not pass quality control before library preparation and was omitted.

PacBio library preparation and sequencing

Using a Covaris S220 Focused-Ultrasonicator, 2 ug of each DNA sample was sheared to an approximate fragment size of 5000 bp and purified using AMPure XP beads (Beckman Coulter). Library preparation was performed using the NEBNext Ultra DNA Library Prep Kit (kit number E7370L, New England Biolabs) and 8 cycles of PCR. Barcode sequences and barcodes assigned to each sample are described in Additional files 31 and 32. Libraries were then quality controlled using a Bioanalyzer high sensitivity DNA chip and the Agilent 2100 Bioanalyzer system. One technical replicate did not pass quality control before library preparation and was omitted. The samples were then submitted to Novogene (Tianjin, China) for PacBio sequencing which was performed on the PacBio Sequel platform using a 600-minute sequencing strategy and three SMRT cells.

Inferring eccDNA forming regions from short read sequencing data

Illumina sequencing signal was analyzed using a custom pipeline inspired by previously published methods [41]. Illumina reads were first trimmed of Illumina TruSeq adapters using CutAdapt [69] version 2.4 with the nextseq-trim=20 option. Trimmed reads were then mapped to the *M. oryzae* Guy11 genome [19] and the 70-15 mitochondrial sequence [70] obtained from the Broad Institute (https://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/magnaporthe-comparative-genomics-proj) using BWA-MEM [71] version 0.7.17-r1188 and the q and a options. Reads mapping to mitochondrial sequences were excluded. Uniquely mapped reads were then mined for split reads that mapped in the same orientation, had at least 20 bp of alignment on either side of the split, and mapped to only two places in the genome. We also only selected split reads where the start of the read mapped downstream from the end. This last filter sets these split reads apart from split reads that would indicate a deletion in the genome. Split reads for which one side of the split read mapped more than 50kbp away from the other, or to a different scaffold than the other, were excluded. Opposite facing read pairs were also obtained from uniquely mapped reads. Candidate eccDNA forming regions were then inferred by combining these two structural read variants. A split read that contained an opposite facing read pair that mapped no more than a combined 500 bp from the borders of the region contained within the two halves of the split read was considered a candidate eccDNA, and a junction split read. The length distribution of these candidate eccDNA forming regions (Additional File 1: Fig. S35A) was then used to probabilistically infer candidate eccDNA forming regions from multi-mapping reads (Additional File 1: Fig. S35B). For each multi-mapping split read, a list of potential combinations of alignments that satisfied the previously described criteria for split reads was generated, and one of these combinations was chosen at random, weighted by its length according to the generated length distribution. The chosen combinations were then used to infer additional candidate eccDNA forming regions by combining these with opposite facing read pairs as before, except this time obtained from unique and multi-mapping reads.

Each candidate eccDNA forming region was then validated by verifying that the region had over 95% read coverage and at least two junction split reads with the exact same coordinates. Candidate eccDNA forming regions that did not pass these criteria were considered low quality and were not included in the analysis.

Inferring eccDNA forming regions from long read sequencing data

CCS were first called from PacBio data using ccs version 3.4.1 (https://ccs.how/). Demultiplexing was then performed using lima version 1.9.0 (https://lima.how/) and sequences of barcodes used for library preparation (Additional Files 31 and 32). CCSs were then mapped to the *M. oryzae* Guy11 genome using minimap2 [72] version 2.18-r1015. Only uniquely mapped reads were kept for analysis. EccDNA forming regions were then identified by looking for split reads that either: 1) mapped in the same orientation to the same exact region multiple times or 2) mapped less than 50 kb apart, in the same orientation and oriented properly so that they were indicative of a circular junction rather than a deletion.

Outward PCR validation of eccDNA forming regions and PCR validation of eccDNA-absent genes

Outward facing primers were designed to 8 eccDNA forming regions of interest to validate their presence in our eccDNA sequencing samples. Primers were designed to amplify the junction of each eccDNA but not result in a product of the same size when used on genomic DNA (Additional File 4: Table S3). Primer3 [73] was used for primer design and the oligonucleotides were synthesized by Integrated DNA Technologies. PCR was performed using New England Biolab's Phusion High-Fidelity DNA polymerase on *M. oryzae* Guy11 genomic DNA and rolling circle amplification products for the sample each eccDNA forming region was found in. 5ng DNA of each sample was used per 50 µl PCR reaction as well as 5X Phusion HF buffer, 10 mM dNTPs, 10 µM forward primer, 10 µM reverse primer, and 1 unit of Phusion DNA polymerase. PCR conditions were as follows: initial denaturation at 98℃ for 30 seconds, 35 cycles of denaturation at 98℃ for 10 seconds, annealing at 64℃ or 65℃ for 30 seconds, extension at 72℃ for 10 seconds, and a final extension at 72℃ for 5 minutes. PCR products were run on a 2% agarose gel to check for amplification. Bands of the expected size were extracted from electrophoresis gels using Zymo Research's Zymoclean Gel DNA Recovery Kit. Sanger sequencing was performed by the UC Berkeley DNA Sequencing Facility, and Sanger sequences were examined for matches to corresponding eccDNA forming regions using BLASTN [74] version 2.2.9 and manual inspection.

PCR validation of eccDNA-absent genes was performed using similar methods. Primers were designed to amplify the entire annotated gene region of *MYO1* and the actin gene (MGG_03982) and a small segment of the *MAGGY* LTR retrotransposon from genomic DNA. 2ng DNA of each sample was used per 20 µl PCR reaction as well as 5X Phusion HF buffer, 10 mM dNTPs, 10 µM forward primer, 10 µM reverse primer, and 0.4 units of Phusion DNA polymerase. PCR conditions were as follows: initial denaturation at 98℃ for 30 seconds, 25 cycles of denaturation at 98℃ for 10 seconds, annealing at 64℃ or 65℃ for 30 seconds, extension at 72℃ for 5, 60 or 120 seconds, and a final extension at 72℃ for 5 minutes. PCR products were run on a 1% agarose gel to check for amplification.

Comparing eccDNA forming regions inferred from Illumina data and eccDNA forming regions inferred from PacBio data

EccDNA forming regions called using Illumina data and PacBio data were found to be identical if their start and end coordinates were within 10 bp of each other to account for mapping errors. EccDNA forming regions were then called with less stringent requirements to verify if any of the missing eccDNA forming regions were being filtered out somewhere in the pipeline. In this test, all uniquely mapped split reads that had 10 or more bp overlap on either side were properly oriented, and those less than 50kb apart were considered eccDNA forming regions.

Benchmarking eccDNA forming regions called using our pipeline on previously published data

EccDNA forming regions called using our pipeline were compared to eccDNA forming regions previously published for *H. sapiens* [41]. EccDNA forming regions were found to be identical if their start and end coordinates were within 10 bp of each other. EccDNA forming regions described as low quality by the authors were excluded from the published dataset before comparison. High coverage eccDNA forming regions were chosen for comparison if they had more than 10 associated junction split reads. Finally, multi-mapping reads were excluded from the pipeline to identify eccDNA forming regions called using only uniquely mapped reads.

Comparing eccDNA sequencing samples to each other

Overlaps in eccDNA forming regions between samples were first calculated based off the exact coordinates of the eccDNA forming regions and Venn diagrams based off these overlaps were generated using the ggVennDiagram R package [75] version 1.2.0. EccDNA forming regions found in all technical replicates taken from each biological replicate were first combined before looking for overlaps between biological replicates. Overlaps were then calculated with various levels of tolerance for the start and end coordinates of the eccDNA forming regions so that regions in one sample that were within 10, 100, or 1000 bp from the start and end coordinates of a region in another sample were considered to be found in both samples. Rarefaction analysis for eccDNA forming regions in all samples was performed by sampling mapped eccDNA sequencing reads at random in increasing 10% intervals. For each subsample, eccDNA forming regions were called as previously described and counted. Principal component analysis of read coverage was performed by first calculating junction split read coverage for all 10kbp windows in the genome for each sample. These values were then normalized to the total number of junction split reads in each sample. The matrix of normalized junction split read coverage for all samples was then processed using the prcomp function in R version 3.6.1 with the scale = TRUE option, and the first 6 principal components were plotted using the ggbiplot R package [76] version 0.55.

Gene and effector annotation

The *M. oryzae* Guy11 genome along with 162 other rice-infecting *M. oryzae* genomes (Additional File 25) were annotated using the FunGAP [77] version 1.1.0 annotation pipeline. For all genomes, RNAseq data (SRR8842990) obtained from GEO accession GSE129291 was used along with the proteomes of *M. oryzae* 70-15, P131, and MZ5-1-6 taken from GenBank (accessions GCA_000002495.2, GCA_000292605.1, and GCA_004346965.1, respectively). The 'sordariomycetes_odb10' option was used for the busco_dataset option and the 'magnaporthe_grisea' option was used for the augustus_species option. For repeat masking, a transposable element library generated by combining the RepBase [78] fngrep version 25.10 with a *de novo* repeat library, generated by RepeatModeler [79] version 2.0.1 run on the *M. oryzae* Guy11 genome with the LTRStruct option, was used for all genomes. Genes in *M. oryzae* Guy11 were assigned names according to the gene names listed on UniProtKB for *M. oryzae* 70-15 accessed in October 2021. To make this assignment, *M. oryzae* Guy11 proteins were aligned to the *M. oryzae* 70-15 proteome using BLASTP [74] version 2.7.1+. Hits with greater than 80% sequence identity that spanned more than 80% of the length of both the *M. oryzae* Guy11 protein and the *M. oryzae* 70-15 protein were assigned names.

Effectors were predicted among *M. oryzae* Guy11 genes by first selecting genes with signal peptides which were predicted using SignalP [80] version 5.0b Darwin x86_64. Genes with predicted transmembrane domains from TMHMM [81] version 2.0c were then excluded. Finally, EffectorP [82] version 2.0 was used to predict effectors from this secreted gene set. Previously well-characterized effectors were identified using previously published protein sequences [27] and DIAMOND [83] version 2.0.9.147.

High quality LTR-retrotransposon annotations in *M. oryzae*

High quality, full length, consensus sequences for known *Gypsy* elements in *M. oryzae* (*MAGGY*, *GYMAG1*, *GYMAG2*, *PYRET*, *MGRL3*) and one *Copia* element (*Copia1*) were generated using the WICKERsoft [84] suite of tools. Reference sequences from other genomes for each element were obtained from the RepBase [78] fngrep version 25.10 library. The *M. oryzae* Guy11 genome was then scanned for the presence of these sequences using BLASTN [74] version 2.2.9 and then filtered to hits with 90% sequence identity and that contained 90% of the sequence length. Hits for each reference sequence were then extended to include 500 base pairs of genomic sequence upstream and downstream of the hit. A multiple sequence alignment of hits for each reference sequence was then generated using ClustalW [85] version 1.83 and boundaries were visually inspected and trimmed. Consensus sequences for each element were then generated from these multiple sequence alignments. These consensus sequences were split into LTR and internal regions by self-alignment using the BLASTN [74] webserver in August 2020 to identify LTRs. These consensus sequences are available in Additional File 33. Finally, the locations of these elements in *M. oryzae* Guy11 genome were annotated with RepeatMasker [86] version 4.1.1 with the -cutoff 250, -nolow, -no_is, and -norna options to identify their locations in the *M. oryzae* Guy11 genome. For read coverage plots as well as histone and GC content plots, full length LTR retrotransposon copies were required. These were identified by using the original full length consensus sequences with RepeatMasker as before and then filtering to hits greater than 3000 bp in length and greater than 90% sequence identity.

Comparative analysis of eccDNA forming regions

Analysis of eccDNA forming regions in organisms other than *M. oryzae* were performed as described above for Illumina sequencing data using previously published genome, gene annotation, and transposable element annotation files (Additional File 34). However, unlike the other data used in this study, the sequencing data in the *S. cerevisiae* dataset was single-end and therefore opposite facing read pairs could not be used to infer eccDNA forming regions. Instead, only eccDNA forming regions with three overlapping junction split reads were used for analysis. For all organisms, reads mapping to unplaced scaffolds and organellar genomes were removed after mapping as described above for the *M. oryzae* mitochondrial genome. These scaffolds were also removed from genome size, number of coding base pairs, and number of LTR retrotransposon base pairs calculations for comparative analysis. To calculate the percent of the genome that was covered in each sample, the genomecov command of the BEDtools [87] suite versions 2.28.0 was used with the -d option along with the coordinates of eccDNA forming regions for each sample. Any base pair with a coverage value greater than zero was counted as being a portion of the genome in an eccDNA forming region.

Characterization of eccDNA formation by LTR retrotransposons

To generate the Manhattan plot, junction split reads were filtered by selecting regions that were made up of 90% LTR retrotransposon sequences. Junction split read coverage was then calculated for each 100 bp window in the genome. Coverage values were then normalized to the total number of LTR eccDNA junction split reads per sample. These coverage values were then averaged across technical replicates for each biological replicate, and then averaged across biological replicates. Finally, only 100 bp bins that overlapped at least 50 bp with an LTR retrotransposon were plotted in Fig 3A. For Additional File 1: Fig. S10, only bins with coverage greater than 0 were plotted.

To simulate expected read coverage for different types of LTR eccDNAs, the *Copia1* consensus sequence was taken as a reference, though the *MAGGY* consensus sequence yielded identical results. Simulated DNA sequences were then generated for each type of LTR eccDNA. The expected 2-LTR circular sequence generated by NHEJ (scenario 1, Fig. 4A) was made up of two LTR sequences and the internal sequence, and the expected 1-LTR circle sequence generated by HR (scenario 3, Fig. 4C) was made up of one LTR sequence and the internal sequence. These sequences were shuffled 1000 times to generate 1000 sequences starting at various points of the expected circularized sequence. For the 1-LTR circle sequence generated by autointegration (scenario 2, Fig. 4B), the random autointegration events were simulated by choosing a random length segment of the internal sequence starting with its start or end, adding the LTR sequence to this sequence, and randomly shuffling the sequence to simulate a circular sequence. This process was repeated 1000 times to generate 1000 sequences. Finally, for each scenario, Illumina reads were simulated to reach 2000x coverage for each of the simulated sequences using ART Illumina [88] version 4.5.8 and the following parameters: 150 bp read length, 450 bp mean insert size, 50 bp insert size standard deviation, HiSeqX TruSeq. Reads were mapped to the simulated sequences using BWA-MEM [71] version 0.7.17-r1188 with default settings and coverage for each base pair was calculated.

To generate observed coverage for each element, sequencing read coverage across the genome was calculated for all 10 base pair windows in the *M. oryzae* Guy11 genome for each sample. Coverage values were then normalized to the total number of mapped sequencing reads in each sample. These coverage values were then averaged across technical replicates for each biological replicate, and then averaged across biological replicates. Finally, profile plot data was generated for full length, high confidence sequences for each LTR retrotransposon using computeMatrix scale-regions and plotProfile of the DeepTools [89] suite of tools version 3.5.1 using full length, high confidence LTR retrotransposon sequences. Profile plots were also generated using previously published whole genome sequencing data by averaging sequencing coverage across all three samples [19,54,59].

Identification of split reads associated with eccDNA formation from LTR retrotransposons

Split reads were first identified as any read that mapped to only two places in the genome with at least 20 base pairs of alignment on either side. LTR-LTR split reads were then selected from these split reads for each LTR retrotransposon if both sides of the split read had any overlap with any copy of that retrotransposon's LTR in the genome. LTR-internal split reads were selected if one side of the split read had any overlap with any copy of the retrotransposon's LTR in the genome and the other side had any overlap with any copy of the retrotransposon's internal region in the genome. Read coverage, LTR-LTR split read coverage, and LTR-internal coverage was then calculated for each annotation of each LTR retrotransposon. Coverage values were then normalized to the total number of mapped sequencing reads in each sample. These coverage values were then averaged across technical replicates for each biological replicate, and then averaged across biological replicates.

Comparison of microDNAs and large eccDNAs across organisms

Genome, gene annotation, and transposable element annotation files for each organism used for this analysis were as previously described (Additional File 34). Again, organellar genomes as well as unplaced contigs were filtered out of these files before analysis. Introns and UTRs were added to gene annotation files that were missing these elements using the 'agat_convert_sp_gff2gtf.pl' and 'agat_sp_add_introns.pl' commands from the AGAT toolkit version 0.6.2 (https://github.com/NBISweden/AGAT). Cpgplot of EMBOSS [90] version 6.6.0.0 was used to annotate CpG islands in each genome. Upstream and downstream regions were defined as being 2000 base pairs upstream from the transcription start site and downstream from the transcription end site, respectively. Genic regions were defined as being made up of all sequences between transcription start and end sites, and intergenic regions were the opposite. Junction split reads were counted as being from a specific region if they overlapped to any extent within that region.

The observed percentage of junction split reads overlapping with each region type was calculated for each sample for each organism and an average of these percentages was calculated. The junction split reads of each sample were then shuffled across the genome 10 times, excluding LTR retrotransposon locations, and an expected percentage for each region was calculated, averaged across all permutations, then averaged across all samples for each organism. Finally, the $\log_2$ of the fold enrichment was calculated by taking the $\log_2$ of the observed average percentage over the expected average percentage.

Correlation of expression and eccDNA formation

Previously published RNAseq data from *M. oryzae* Guy11 grown in liquid culture in rich medium was obtained [91] (Additional File 35). The data was mapped to the *M. oryzae* Guy11 genome using STAR [92] version 2.7.1a with the quantMode GeneCounts option. Read counts per gene were then divided by library size and multiplied by the length of each gene in order to obtain reads per kilobase million (RPKMs). RPKMs per gene were then averaged across all samples.

Junction split read counts per gene used to analyze the correlation of expression and eccDNA formation were generated for each gene by counting the number of junction split reads that intersect the gene to any extent. Counts per gene were first assessed for each sample and normalized to the number of junction split reads in that sample. Normalized counts were then averaged across technical replicates for each biological replicate. Average counts per biological replicate were then averaged to obtain the final result.

To compare gene content and eccDNA formation, the *M. oryzae* genome was divided into 100kbp bins and the number of genes per bin was calculated. Junction split reads per bin were calculated for each sample using the same method. Junction split read per bin values were then normalized to the total number of junction split reads in each sample. These values were averaged across technical replicates for each biological replicate, and then averaged across biological replicates.

ACS enrichment analysis

The published ACS sequence profile [63] was used to identify ACSs in eccDNA forming regions using the FIMO [93] software version 4.12.0. Only hits scoring greater than 17 were kept. In order to test for enrichment of these sequences, an expected distribution of ACS sequences was generated by randomly shuffling eccDNA forming regions across the *M. oryzae* Guy11 genome, excluding regions containing LTR retrotransposons. The observed number of ACS sequences in eccDNA forming regions was then compared to the expected distribution to generate a p-value.

Histone mark and GC content profile plots

Previously published ChIPSeq data for H3K27me3, H3K27ac, H3K36me3, and loading controls were obtained [91]. Sequencing reads for each technical replicate were combined before reads for each treatment for each biological replicate were mapped to the *M. oryzae* Guy11 genome using BWA-MEM [71] version 0.7.17-r1188 with default settings. The bamCompare command from the DeepTools [89] suite of tools version 3.5.1 with the scaleFactorsMethod readCount option was used to compare the signal from each treatment to the loading control for each biological replicate. computeMatrix scale-regions was then used in conjunction with the plotProfile command to generate processed data for profile plots. After verifying that all biological replicates resulted in similar profile plots, only the first biological replicate was chosen for presentation.

To generate tracks used for profile plots, a few different strategies were used. GC content profile plots were generated by calculating GC percentage for 50 base pair windows throughout the genome. Profile plot data was then generated using computeMatrix scale-regions and plotProfile commands as before. Methylated and acetylated genes were determined using the methylation and acetylation peaks published by Zhang *et al.* [91]. Marked genes were called when at least 50% of the gene overlapped with a peak. Large eccDNAs, microDNAs, and LTR-eccDNAs from all *M. oryzae* Guy11 samples were combined into a single list which was filtered for duplicates and used for the corresponding tracks in the profile plots. The genome baseline track was generated by combining all of these eccDNA forming regions and shuffling them randomly across the genome. Finally, the full length, high quality LTR-retrotransposon annotations described above were used for LTR retrotransposon tracks. The same approach was used for generating profile plots to compare histone marks and GC content for eccDNA-associated and eccDNA-absent genes.

Identification of eccDNA-associated and eccDNA-absent genes

Encompassing split read counts per gene for determining eccDNA-associated and eccDNA-absent genes were generated for each gene by counting the junction split reads that fully encompass the gene using the intersect command of the BEDTools [87] suite version 2.28.0 with the -f 1 option. This count was normalized to the total number of junction split reads in each sample, then averaged across technical replicates for each biological replicate. Genes with a count of zero were removed from each biological replicate before being sorted by this count.

Genes in the top third for this count were compared between biological reps using the ggVennDiagram R package [75] version 1.2.0. This count was averaged across biological replicates to obtain the encompassing split read count per gene for visualizations in Fig. 5A and Fig. 8 and for comparison between predicted effectors and other genes (Additional File 1: Fig. S34).

GO enrichment analysis

GO terms were first assigned to annotated *M. oryzae* Guy11 genes using the PANNZER2 [94] webserver on August 17th, 2020. Annotated GO terms were then filtered to annotations with a positive predictive value greater than 0.6. The topGO [95] R package version 2.36.0 was used to parse assigned GO terms and reduce the gene list to a list of feasible genes for analysis. Either eccDNA-associated or eccDNA-absent were assigned as significant genes, and the number of these genes belonging to each GO term was used as the observed value for the enrichment analysis. A kernel density function was then generated using the gene lengths of the significant gene set. The same number of genes as the significant gene set were sampled at random from the feasible gene set using weighted random selection with weights obtained from the kernel density function. This random sampling was repeated 100 times and the average of the number of genes belonging to each GO term was used as the expected value for the enrichment analysis. Finally, the Chi-square statistic was computed comparing observed and expected values to test for enrichment or depletion of each GO term.

Gene presence absence variation

In order to identify genes prone to presence absence variation in the *M. oryzae* Guy11 genome, OrthoFinder [96] version 2.5.1 with default settings was used on all of the *M. oryzae* proteomes and the *Neurospora crassa* proteome obtained from GenBank (accession GCA_000182925.2). Then, for each *M. oryzae* genome, we queried whether each gene annotated in the *M. oryzae* Guy11 genome had an ortholog identified by OrthoFinder in that genome. Finally, the absence of genes without orthologues was confirmed using BLASTN [74] version 2.7.1+.

Small, genic deletions were identified using orthologs identified by OrthoFinder [96] version 2.5.1 as before. For each genome, we looked for genes in the *M. oryzae* Guy11 genome that had no ortholog in that genome, but that were flanked by two genes with orthologs in that genome. One-to-many, many-to-many, and many-to-one orthologs were excluded from this analysis. Candidate gene deletions were validated using alignments performed using the nucmer and mummerplot commands of the MUMmer [97] suite of tools version 4.0.0rc1 to verify that a DNA deletion truly existed, and that this deletion overlapped the gene of interest.

Identification of eccDNA-mediated translocations

Identification of translocations with a potential eccDNA intermediate was done by first aligning two genomes using the nucmer command of the MUMmer [97] suite of tools version 4.0.0rc1 with the maxmatch option. The nucmer output was then parsed to look for portions of the reference genome that had an upstream region that aligned to one query scaffold, followed by two separate adjacent alignments to another query scaffold, followed by a downstream region that aligned to the original query scaffold. We also required that the two adjacent alignments in the center of the region were to adjacent regions in the query scaffold, but their order was reversed compared to the reference. Candidate eccDNA-mediated translocations were verified manually by inspecting alignment plots generated using the mummerplot command. The *S. cerevisiae* EC1118 (GCA_000218975.1) and M22 genomes (GCA_000182075.2) obtained from GenBank were used to verify the ability of our pipeline to detect these translocation events. The *M. oryzae* Guy11 genome was then compared to 306 *M. oryzae* genomes (Additional File 27) to look for these events in the *M. oryzae* species. Before alignment, transposable elements were masked from these *M. oryzae* genomes using RepeatMasker [86] version 4.1.1 with the -cutoff 250, -nolow, -no_is, and -norna options, as well as a transposable elements library generated by combining the RepBase [78] fngrep version 25.10 with the *de novo* repeat library generated by RepeatModeler [79] version 2.0.1 run on the *M. oryzae* Guy11 genome with default settings aside from the LTRStruct argument.

Minichromosome genes and eccDNAs

Scaffolds corresponding to minichromosomes in the *M. oryzae* FR13 (GCA_900474655.3), CD156 (GCA_900474475.3), and US71 (GCA_900474175.3) genomes were extracted according to previously published data [23]. Exonerate [98] version 2.4.0 was then used with the protein2genome model to identify genes in the *M. oryzae* Guy11 genome that were found on minichromosomes in these other isolates. Hits with greater than 70% sequence identity to any minichromosome scaffold were identified as genes found on minichromosomes. Encompassing split reads were then counted for all genes. This count was normalized to total number of junction split reads in each sample, then averaged across technical replicates for each biological replicate, then averaged across biological replicates. Finally, normalized encompassing split read counts for genes found on minichromosomes were compared to genes not found on minichromosomes.

Rarefaction analysis for eccDNA-absent genes and unique eccDNA forming regions

Rarefaction analysis for genes found fully encompassed by eccDNA forming regions were performed by first sampling eccDNA forming regions from all samples at random in increasing 10% intervals. For each subsample, the number of genes found fully encompassed by eccDNA forming regions was determined as before. Next, eccDNA forming regions were shuffled across the genome and sampled at random in increasing 10% intervals. Again, the number of genes found fully encompassed by eccDNA forming regions was determined for each sample. This analysis was performed 100 times with similar results as those represented in Fig. 5C. A similar approach was used for rarefaction analysis of eccDNA forming regions but the number of unique microDNAs, large eccDNAs and LTR-eccDNAs were counted at each subsample instead.

Data processing and analysis

Data processing was performed in a RedHat Enterprise Linux environment with GNU bash version 4.2.46(20)-release. GNU coreutils version 8.22, GNU grep version 2.20, GNU sed version 4.2.2, gzip version 1.5, and GNU awk version 4.0.2 were all used for file processing and handling. Conda version 4.8.2 (https://docs.conda.io/en/latest/) was used to facilitate installation of software and packages. Code parallelization was performed with GNU parallel [99] version

20180322. Previously published data was downloaded using curl version 7.65.3 (https://curl.se/) and sra-tools version 2.10.4 (https://github.com/ncbi/sra-tools). Image file processing was performed with the help of ghostscript version 9.25 (https://ghostscript.com/) and imagemagick version 7.0.4-7 (https://imagemagick.org/index.php). BED format files were processed using bedtools [87] version 2.28.0 and bedGraphToBigWig version 4 (https://www.encodeproject.org/software/bedgraphtobigwig/). SAM and BAM format files were processed with SAMtools [100] version 1.8 and Picard version 2.9.0 (https://broadinstitute.github.io/picard/).

Data processing was also facilitated by custom Python scripts written in Python version 3.7.4 with the help of the pandas [101] version 0.25.1 and numpy [102] version 1.17.2 modules. The scipy [103] version 1.4.1 and more-intertools version 7.2.0 (https://more-itertools.readthedocs.io/) modules were also used.

Data analysis and statistical analyses were performed in R version 3.6.1. Data handling was processed using data.table [104] version 1.13.6, tidyr [105] version 1.1.3, reshape2 [106] version 1.4.4, and dplyr [107] version 1.0.4 packages. Plotting was performed using the ggplot2 [108] version 3.3.5 package, with help from RColorBrewer [109] version 1.1.2, scales [110] version 1.1.1, cowplot [111] version 1.1.1, ggprepel [112] version 0.9.1 and ggpubr [113] version 0.4.0 packages. The Gviz [114] version 1.28.3 was used for BAM file visualization. Tables were made using gt [115] version 0.3.1.

Declarations

Availability of data and materials

The datasets supporting the conclusions of this article are available in the NCBI's Sequence Read Archive repository. Illumina circularome sequencing data for *M. oryzae* was submitted under BioProject accession PRJNA768097. PacBio circularome sequencing data for *M. oryzae* was submitted under BioProject accession PRJNA556909. Illumina circularome sequencing data for *O. sativa* was submitted under BioProject accession PRJNA768410. Additional datasets supporting the conclusions of this article are available on Zenodo. Genomes and annotation files used for comparative circularome are available under the DOI 10.5281/zenodo.5544950. Annotated genes and predicted proteins for rice-infecting *M. oryzae* isolates are also available under the DOI 10.5281/zenodo.5542597. Outputs from OrthoFinder2 run on rice-infecting *M. oryzae* proteomes are also available under the DOI 10.5281/zenodo.5544260. Finally, all files used for statistical analysis and plotting are available under the DOI 10.5281/zenodo. 7114261.

Code for the pipeline used to call eccDNA forming regions for Illumina sequencing data is available in a maintained GitHub repository (https://github.com/pierrj/ecc_caller). All other code used for raw data processing, data analysis, and figure generation is available in a GitHub repository (https://github.com/pierrj/moryzae_eccdnas_manuscript_code_final).

Authors' contributions

PMJ and KVK conceptualized and designed the study. PMJ collected and analyzed the data. PMJ wrote the original draft manuscript. PMJ and KVK reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Snighda Poddar for providing the *M. oryzae* Guy11 isolate and for advice for culturing the pathogen. We thank Ursula Oggenfuss for advice on using WICKERsoft for generating LTR retrotransposon consensus sequences. We thank Shruti Parikh for proofreading the manuscript. We also thank the Krasileva lab for feedback on manuscript preparation. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer). We also thank Novogene (Tianjin, China) for technical support.

Supplementary information

Additional File 1 (Supplementary Figures) and Additional File 4 (Supplementary Tables) have been included below in their entirety. Only the descriptions of all other additional files have been included below. All additional files are available as part of the original publication that this chapter was based on, as described at the beginning of the chapter.

Additional File 1: Supplementary Figures.

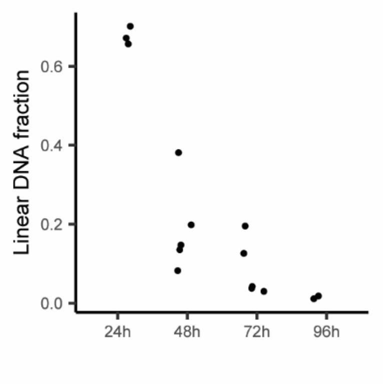


Fig. S1. Degradation of linear DNA using exonuclease treatment. Scatter plot showing the effect of exonuclease treatment on linear DNA fraction of total extracted DNA from *M. oryzae* tissue samples. Each dot represents one biological replicate averaged across four qPCR replicates.

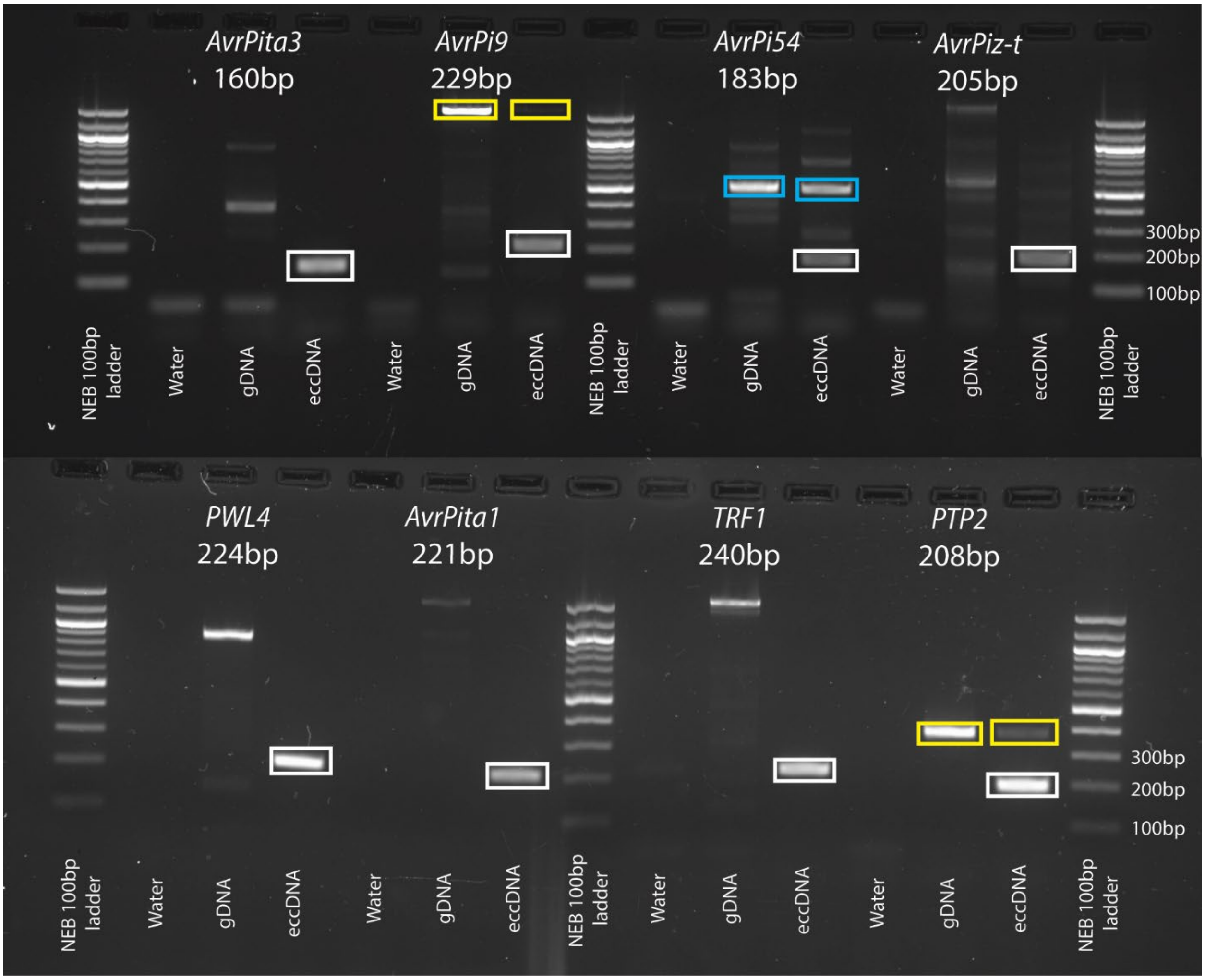


Fig. S2. Outward PCR validation of eccDNA forming regions. Genes of interest found in eccDNA forming regions are listed for each group of three samples. One primer set was used per group and the expected product size is written below the gene name. All samples for each product were from the same PCR reaction. All boxes indicate PCR products that were Sanger sequenced. White boxes indicate PCR products that matched the expected eccDNA junctions. Yellow boxes indicate PCR products that originated from continuous sequences of DNA present in both the genomic DNA and on a high confidence eccDNA forming region found in the eccDNA sample. Blue boxes indicate PCR products with different sequences. PCR for *AvrPita3*, *AvrPi9*, *AvrPi54*, *AvrPiz-t*, and *TRF1* junctions were all performed using biological replicate 1, technical replicate A. PCR for *AvrPita1*, and *PTP2* junctions were performed using biological replicate 1, technical replicate C. PCR for the *Pwl4* junction was performed using biological replicate 2, technical replicate A.

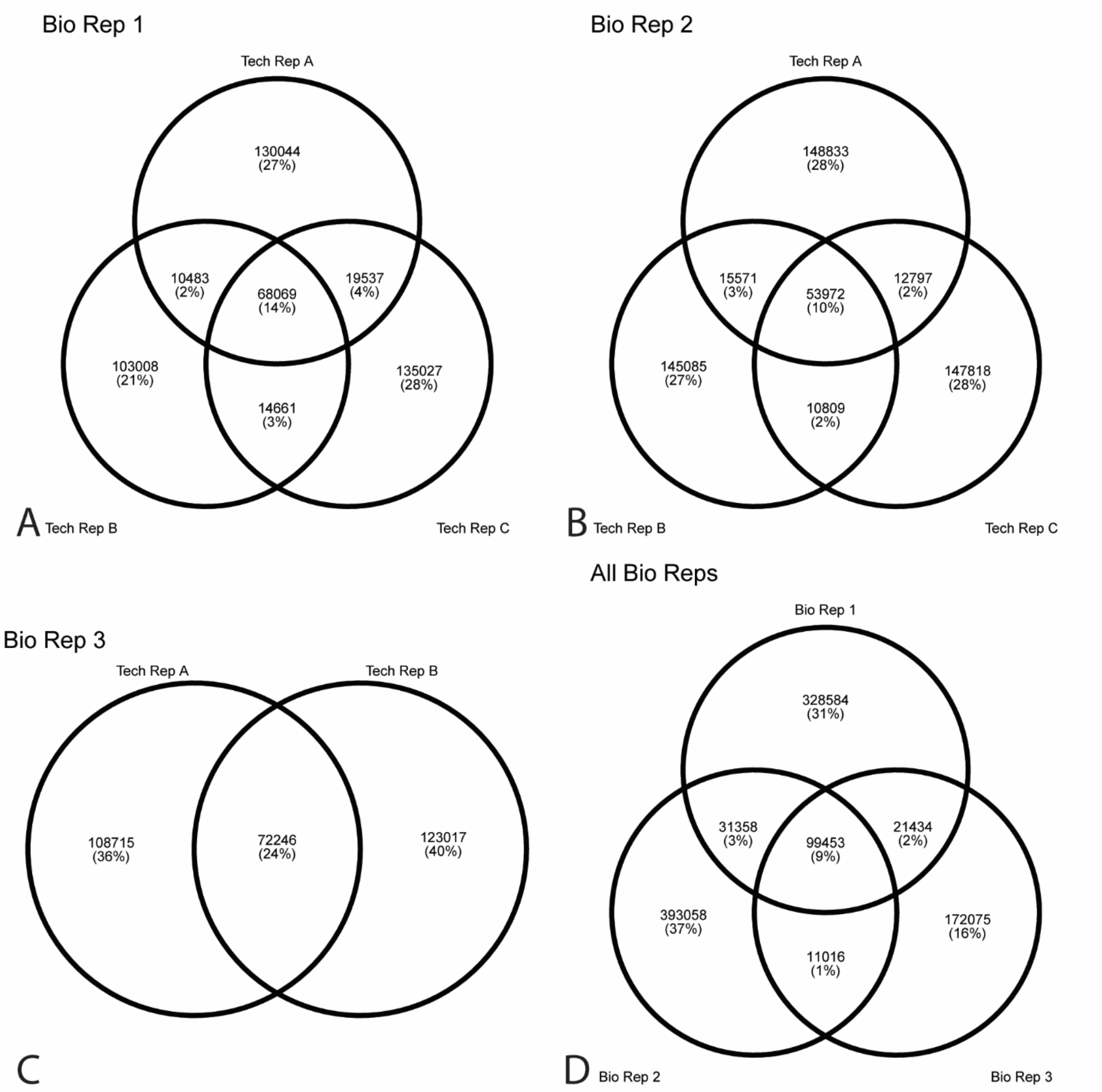


Fig. S3. Overlap in exact break points of eccDNA forming regions across samples. Venn diagrams showing the number of eccDNA forming regions sharing exact coordinates across technical replicates (**A-C**) and all biological replicates (**D**). EccDNA forming regions from all technical replicates for each biological replicate were merged before they were compared between biological replicates.

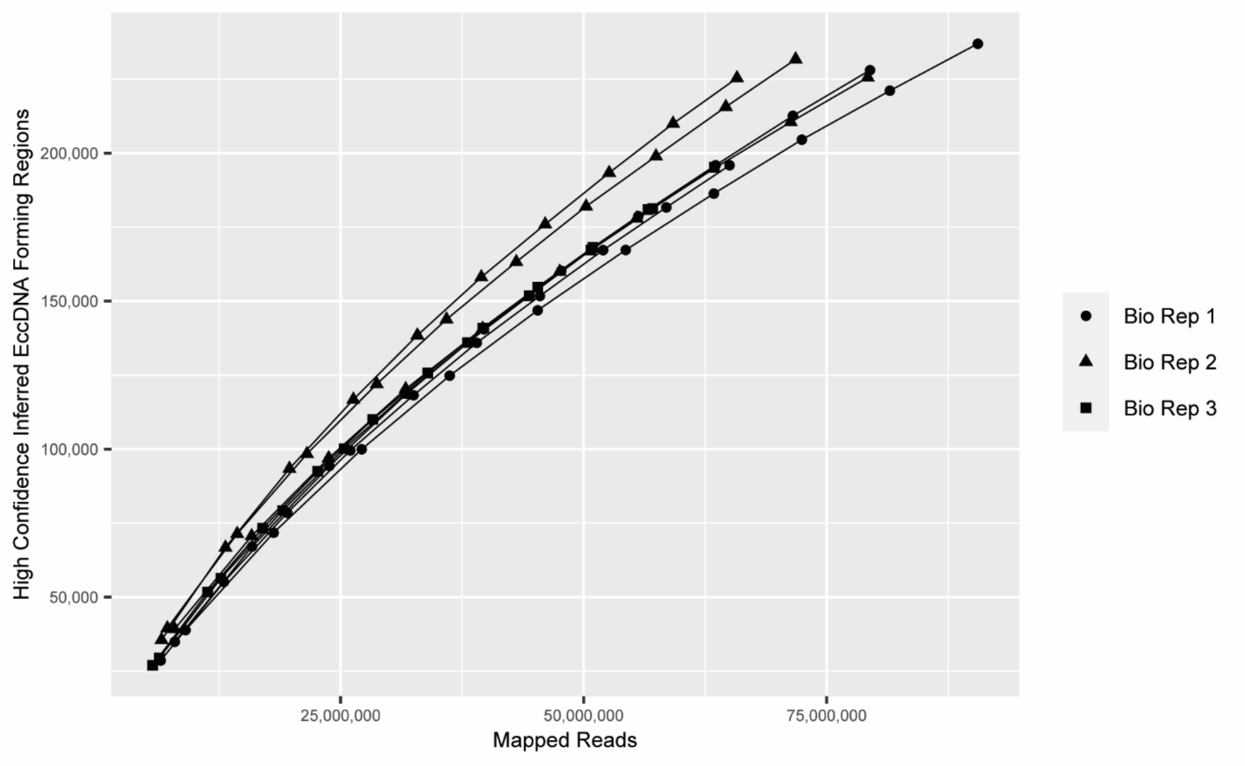


Fig. S4. Rarefaction analysis of sequencing coverage and eccDNA forming regions across all samples. Rarefaction curves showing the number of eccDNA forming regions called at each subset of total mapped reads for each sample. Each dot represents one subsample of mapped sequencing reads for one sequenced sample.

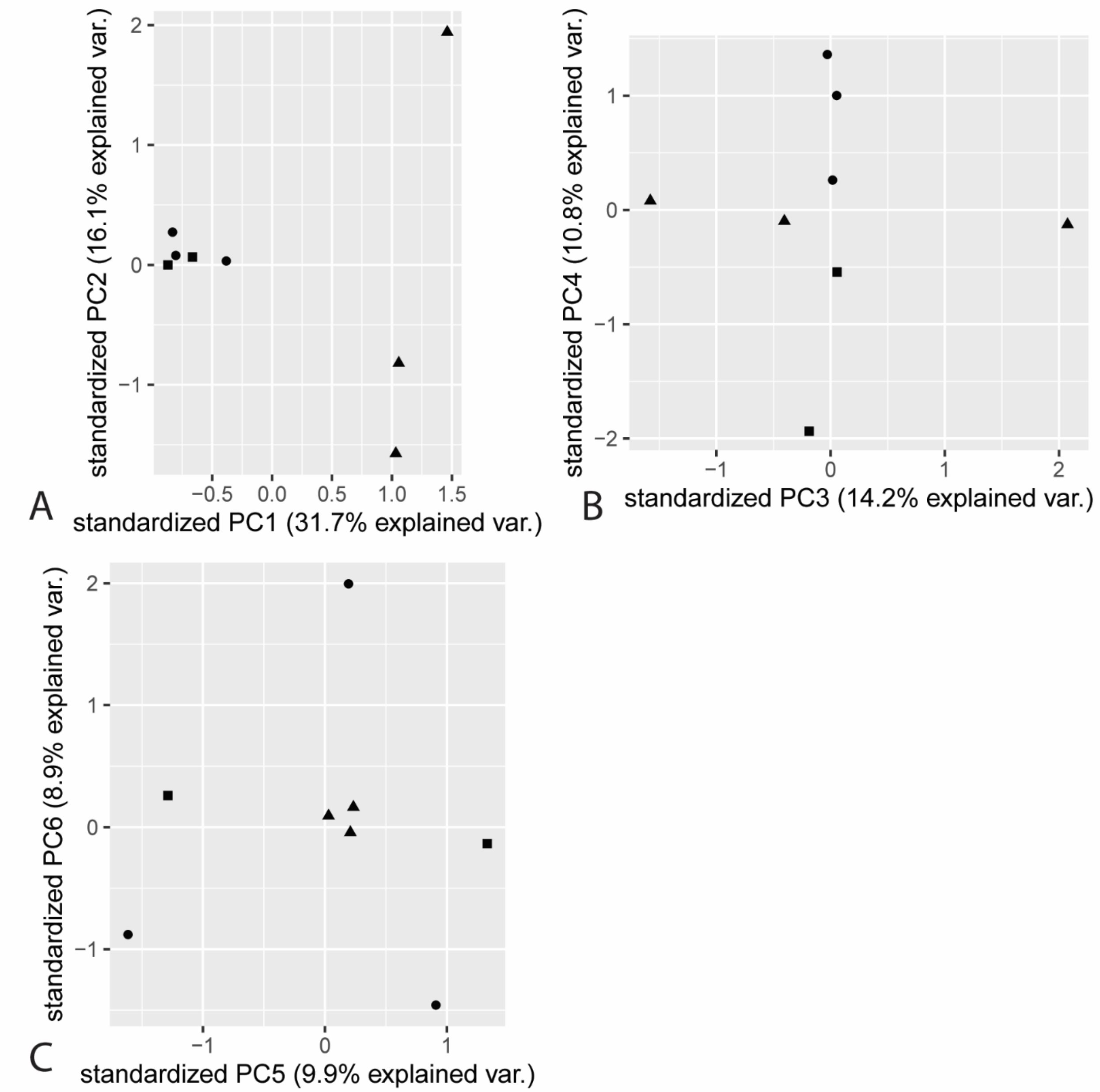


Fig. S5. Principal components analysis of sequencing coverage between samples. Biplots showing values of each sample for the first six principal components (PCs) generated from a principal components analysis performed using sequencing read coverage of all 10kbp bins across the *M. oryzae* Guy11 genome. Each dot represents one sample, and the shape of the dots represent the biological replicate each sample was taken from.

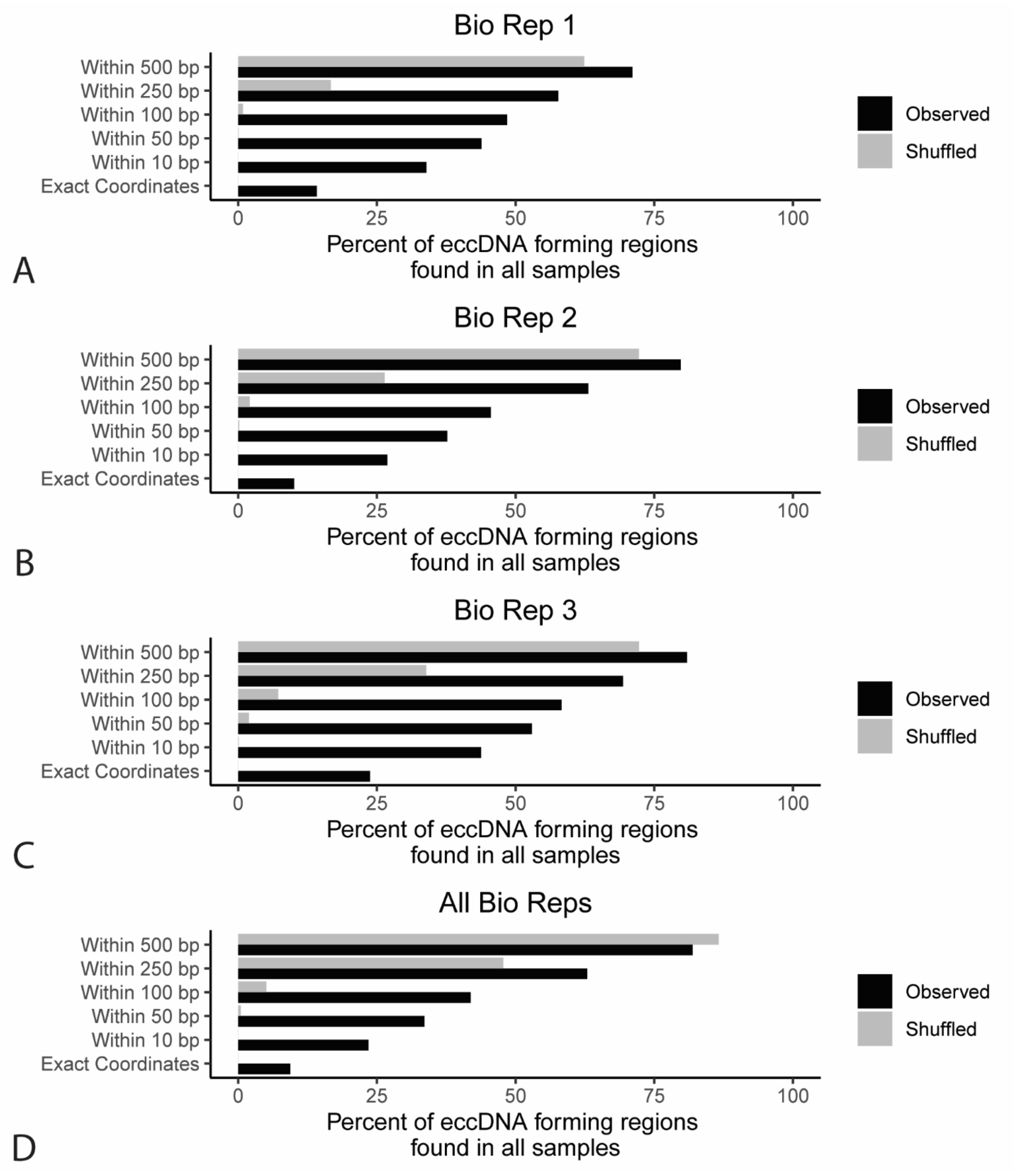


Fig. S6. Overlap in eccDNA forming regions across samples, with increasing tolerance for start and end coordinates. Histogram showing percentage of eccDNA forming regions found in all technical replicates for each biological replicate (**A-C**) as well as percentage of eccDNA forming regions found in all biological replicates (**D**). Percentages are shown for comparison of eccDNA forming regions based off exact coordinates as well as increasing levels of tolerance when comparing the start and end coordinates of the eccDNA forming regions. EccDNA forming

regions from all technical replicates for each biological replicate were merged before they were compared between biological replicates. This observed data was compared to data obtained by randomly placing eccDNA forming regions throughout the genome for each sample. Percentages shown for these shuffled data points are the mean of 100 randomized trials. Standard deviations were too small to visualize meaningfully in the figure.

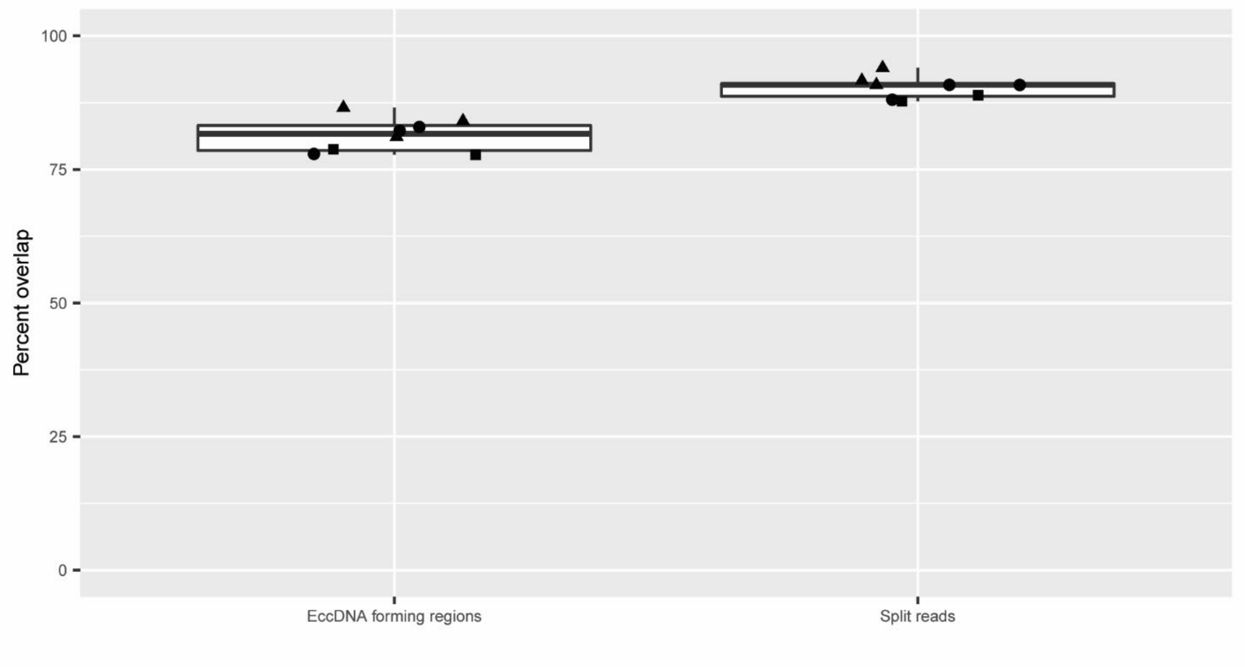


Fig. S7. Overlap between eccDNA forming regions called using PacBio sequencing data and Illumina sequencing data. Boxplot showing the percentage of eccDNA forming regions that were found in each sample using our PacBio sequencing data that were also represented in either eccDNA forming regions called using our Illumina sequencing data or split reads found in this data. Each point represents one sample, and the shape of the points represent the biological replicate that sample was taken from.

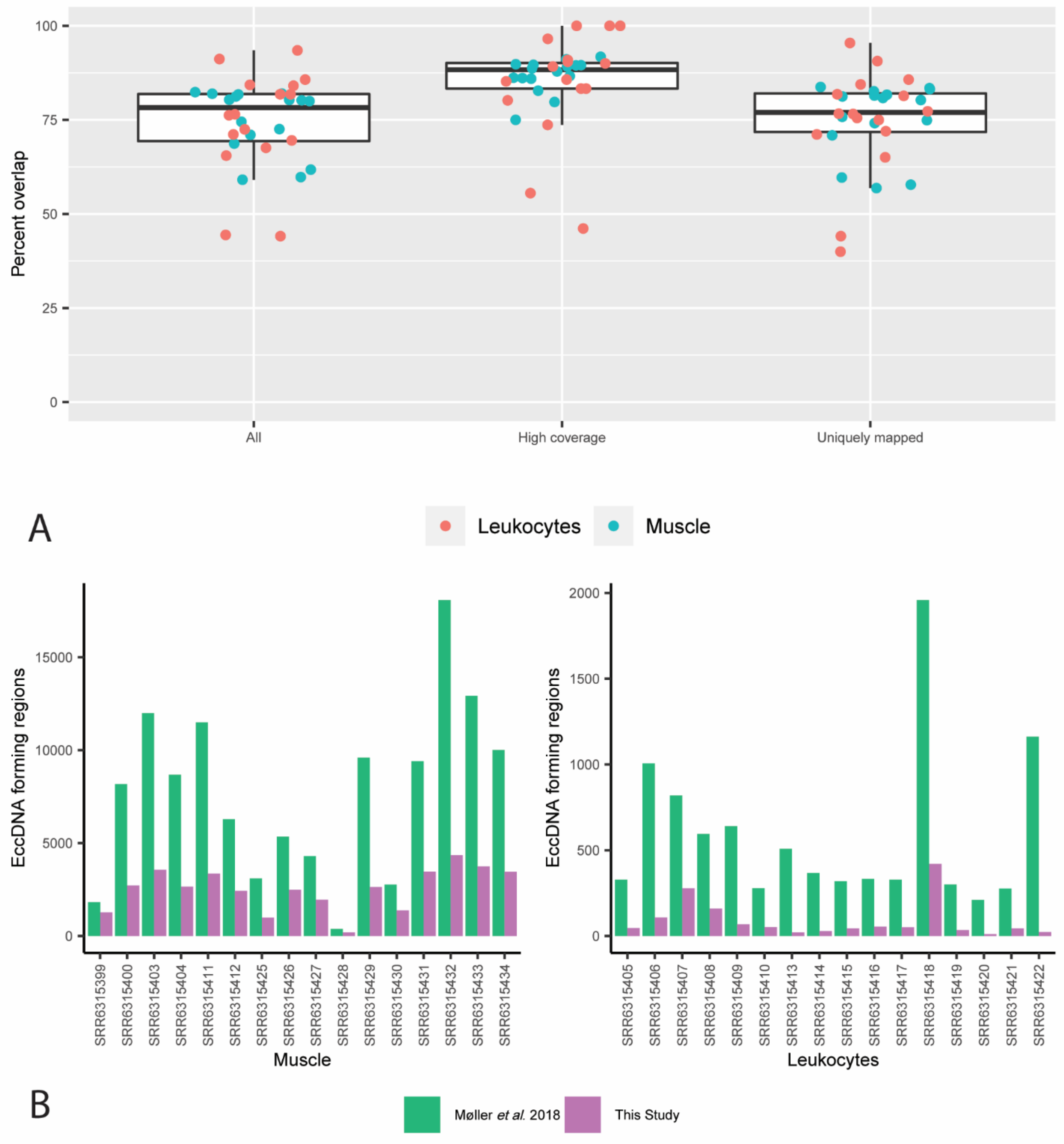


Fig. S8. Comparison between eccDNA forming regions in human samples called in this manuscript and in the original publication. **A.** Boxplot showing the percentage of eccDNA forming regions that were found using our pipeline that were also found in the published eccDNA forming regions for human samples. Each dot represents one sample. **B.** Bar plot showing counts of eccDNA forming regions generated from our pipeline compared to counts in published data. Sample IDs we taken from the Sequence Read Archive (SRA).

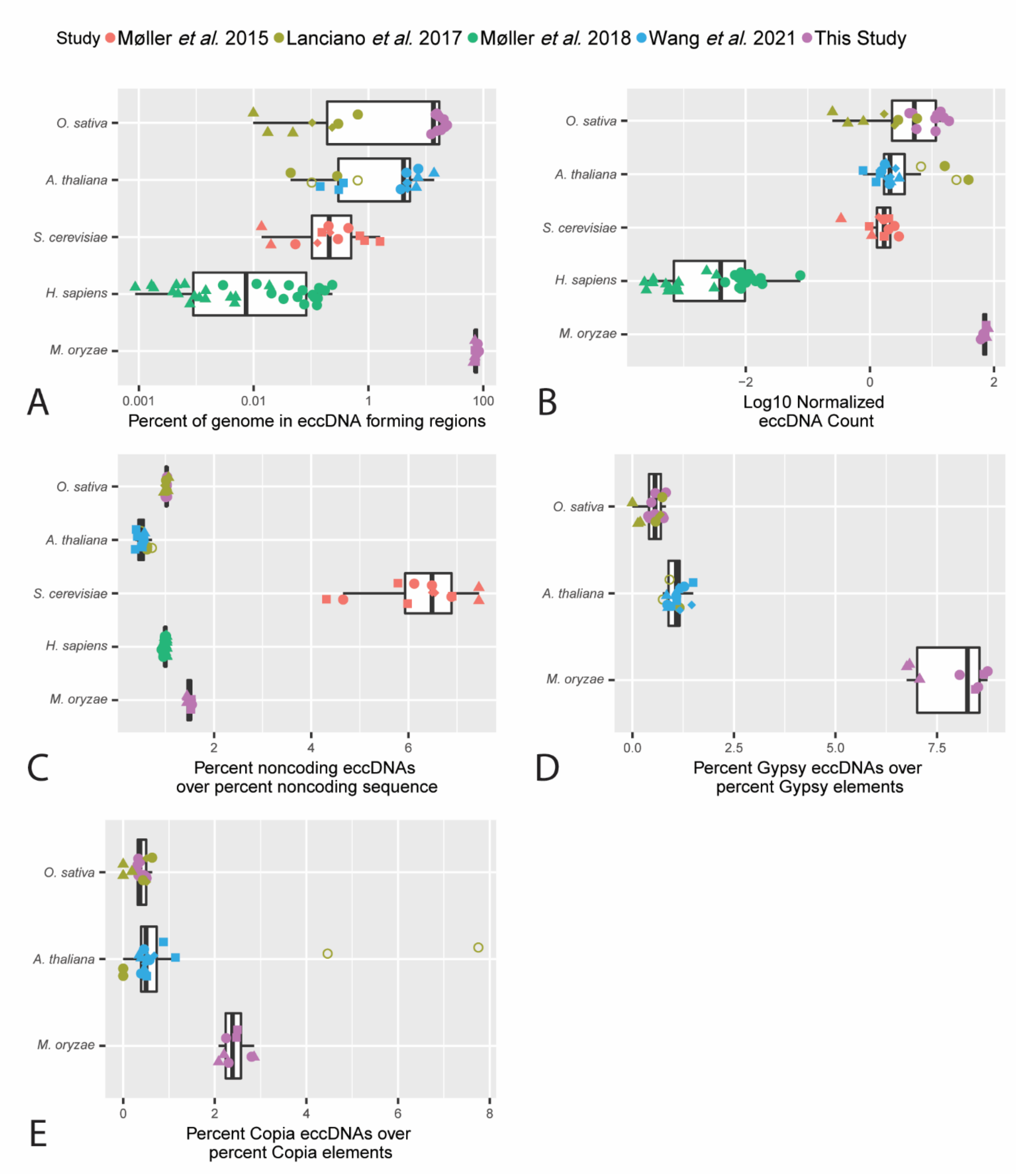


Fig. S9. Comparison of eccDNA forming regions between *M. oryzae* and other previously studied organisms. Box plots comparing **A.** the percentage of the genome found in eccDNA forming regions, **B.** log 10 count of eccDNA forming regions normalized to genome size and sequencing library size, **C.** percent of eccDNA forming regions that contain more than 50% noncoding sequences divided by percent of the genome made up of noncoding seqeuence, **D.** percent of eccDNA forming regions that contain more than 90% LTR/Gypsy retrotransposon

sequence divide by percent of the genome made up of LTR/Gypsy retrotransposon sequence, **E.** percent of eccDNA forming regions that contain more than 90% LTR/Copia retrotransposon sequence divided by percent of the genome made up of LTR/Gypsy retrotransposon sequence across multiple organisms and studies. Each dot represents one sequenced sample. Shapes represent variations in sample type within the same organism. For *M. oryzae*, shapes correspond to which biological replicate each sample was taken from. For *Oryza sativa*, circles represent leaf samples, triangles represent callus samples and diamonds represent seed samples. For *Homo sapiens*, circles represent muscle samples and triangles represent leukocyte samples. For *Arabidopsis thaliana*, circles represent wild type flower samples, empty circles represent *epi12* mutant flower samples, squares represent root samples, diamonds represent leaf samples and triangles represent stem samples. For *Saccharomyces cerevisiae*, circles represent samples from the yeast deletion collection, squares represent samples from the yeast deletion collection treated with zeocin, triangles represent samples from *GAP1* circle carrying yeast, diamonds represent samples from clonal isogenic haploid S228C yeast. For the retrotransposon boxplots, *H. sapiens* samples were excluded due to a lack of active LTR/Gypsy and LTR/Copia retrotransposons in their genome [116] and *S. cerevisiae* samples were excluded due to a small number of eccDNA forming regions containing retrotransposon sequences.

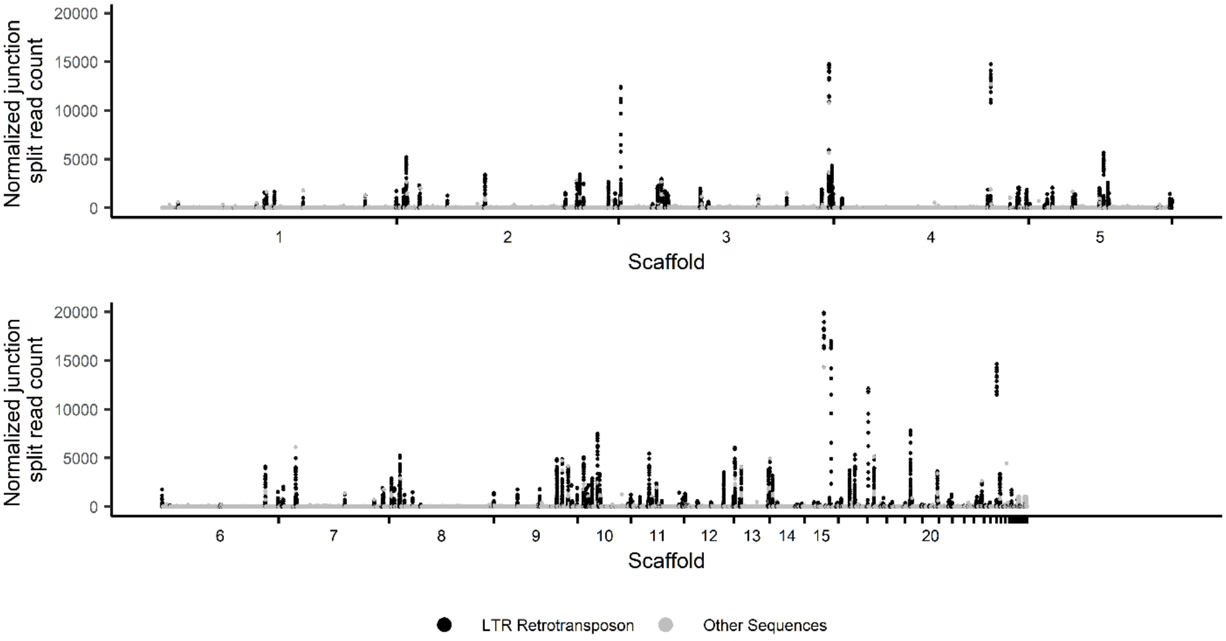


Fig. S10. EccDNA forming regions composed of more than 90% LTR retrotransposon sequence in *M. oryzae*. Manhattan plot showing the number of junction split reads per million averaged across biological replicates for all 100 bp bins with junction split read coverage greater than zero in the *M. oryzae* Guy11 genome. Each dot represents one of these bins. Bins made up of more than 90% LTR retrotransposon sequence are colored in black.

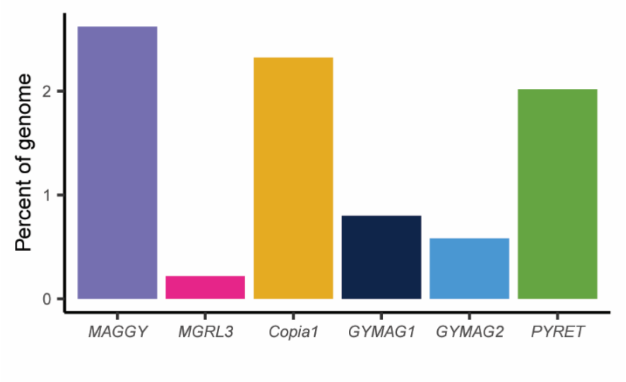


Fig. S11. Percentage of the *M. oryzae* Guy11 genome made up of each LTR retrotransposon.

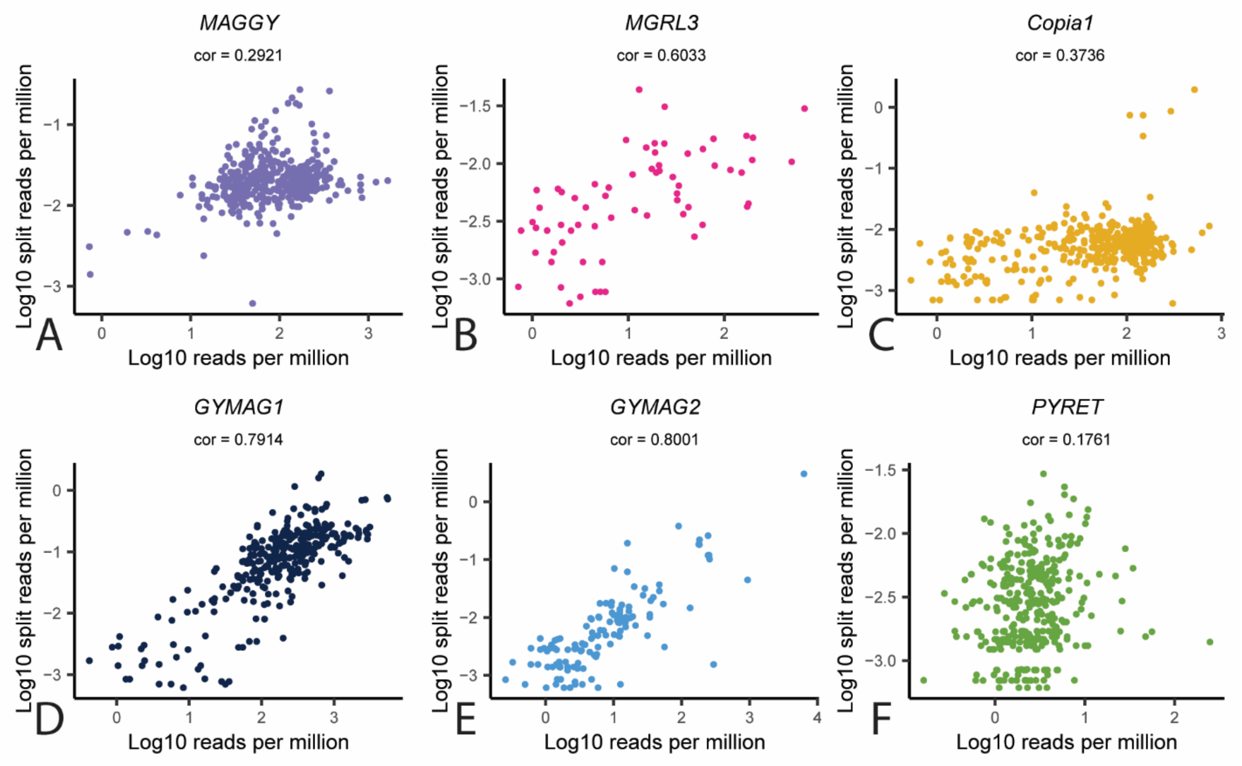


Fig. S12. Correlation between number of LTR-LTR split reads and sequencing reads in eccDNA sequencing samples for each LTR retrotransposon in *M. oryzae*. **A-F**. Scatter plots showing Pearson's correlation coefficient between log 10 sequencing reads per million reads and log 10 LTR-LTR split reads per million sequencing reads, averaged across biological replicates. Each dot represents one annotated portion of the LTR region of an LTR retrotransposon.

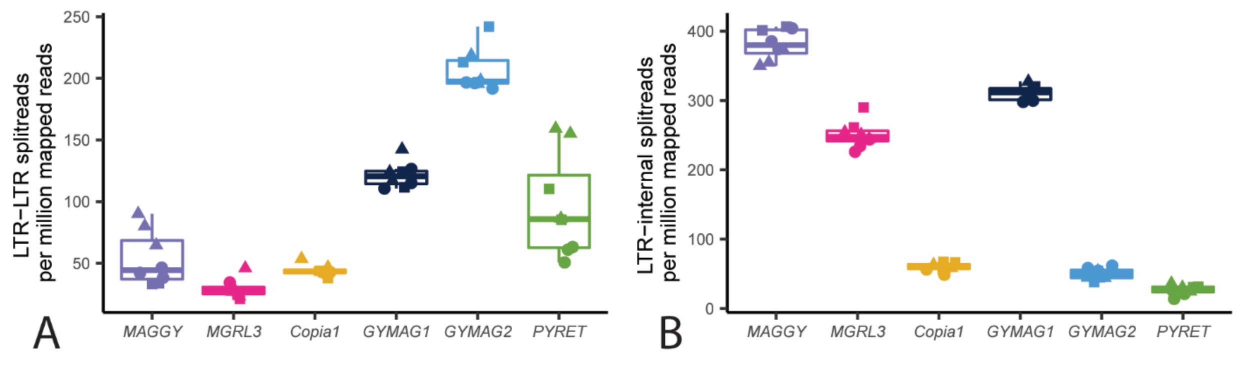


Fig. S13. Number of LTR-LTR split reads and LTR-internal split reads in eccDNA sequencing samples for each LTR retrotransposon in *M. oryzae.* **A.** Box plot showing identified LTR-LTR split reads per million reads mapped to each element for each LTR retrotransposon in the *M. oryzae* Guy11 genome. Each point represents one sample. **B.** Box plot showing identified LTR-internal split reads per million reads mapped to each element, for each LTR retrotransposon in the *M. oryzae* Guy11 genome. Each point represents one sample, and the shape of the points represent the biological replicate that sample was taken from.

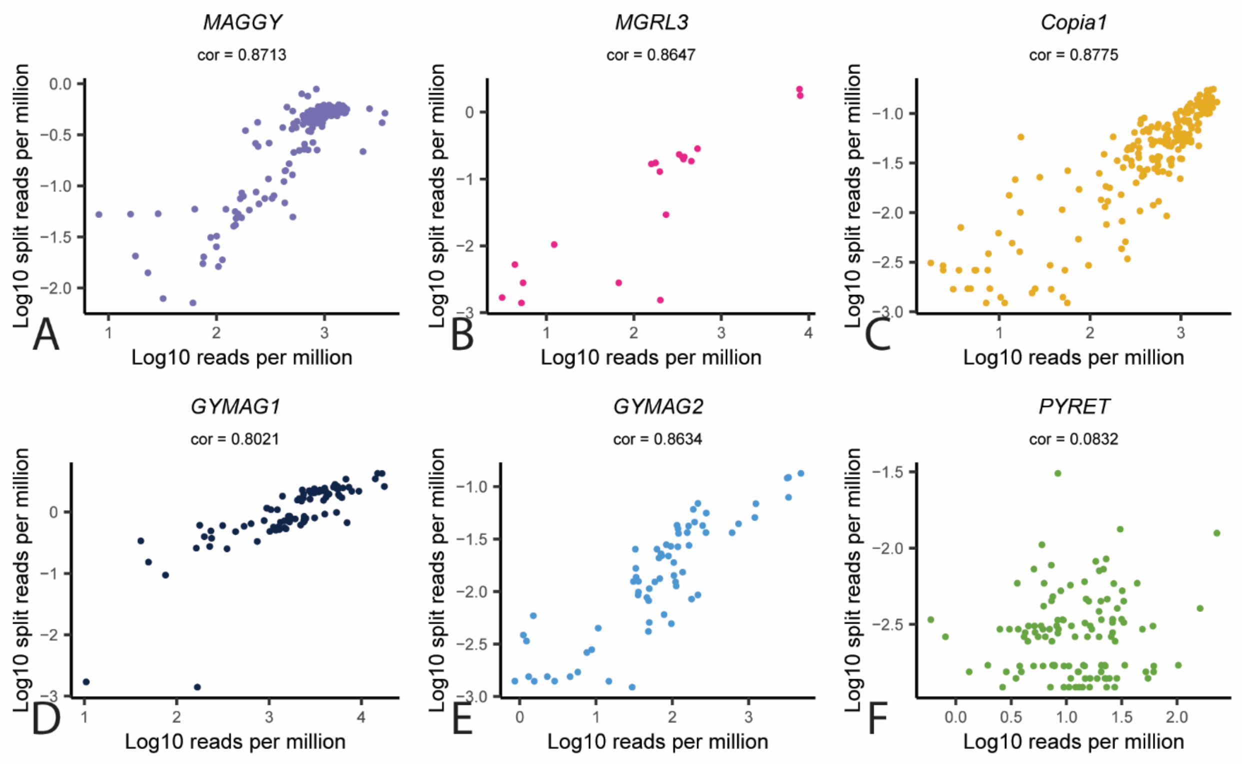


Fig. S14. Correlation between number of LTR-internal split reads and sequencing reads in eccDNA sequencing samples for each LTR-retrotransposon in *M. oryzae*. **A-F**. Scatter plots showing Pearson's correlation between log 10 sequencing reads per million reads and log 10

LTR-internal split reads per million sequencing reads, averaged across biological replicates. Each dot represents one annotated portion of the internal region of an LTR retrotransposon.

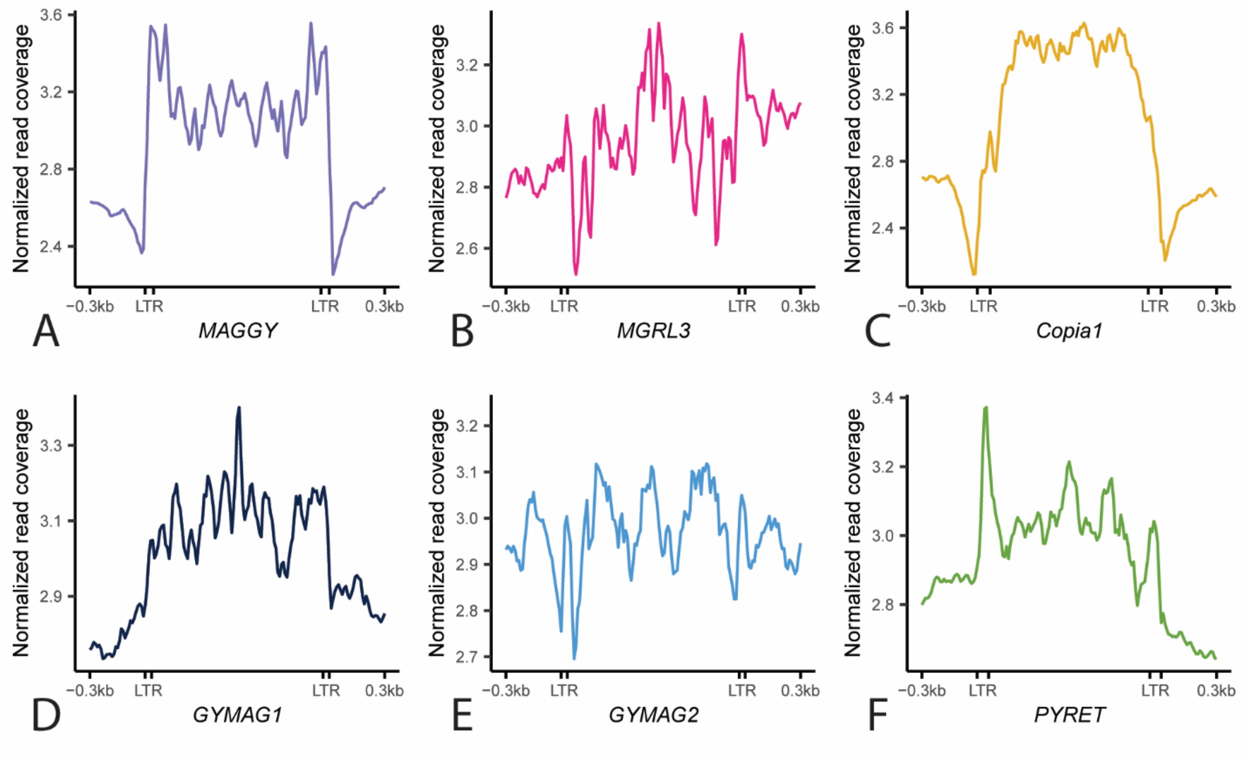


Fig. S15. Expected read coverage for LTR retrotransposons in *M. oryzae*. **A-F**. Profile plots showing observed whole genome sequencing read coverage for each LTR retrotransposon found in the *M. oryzae* Guy11 genome.

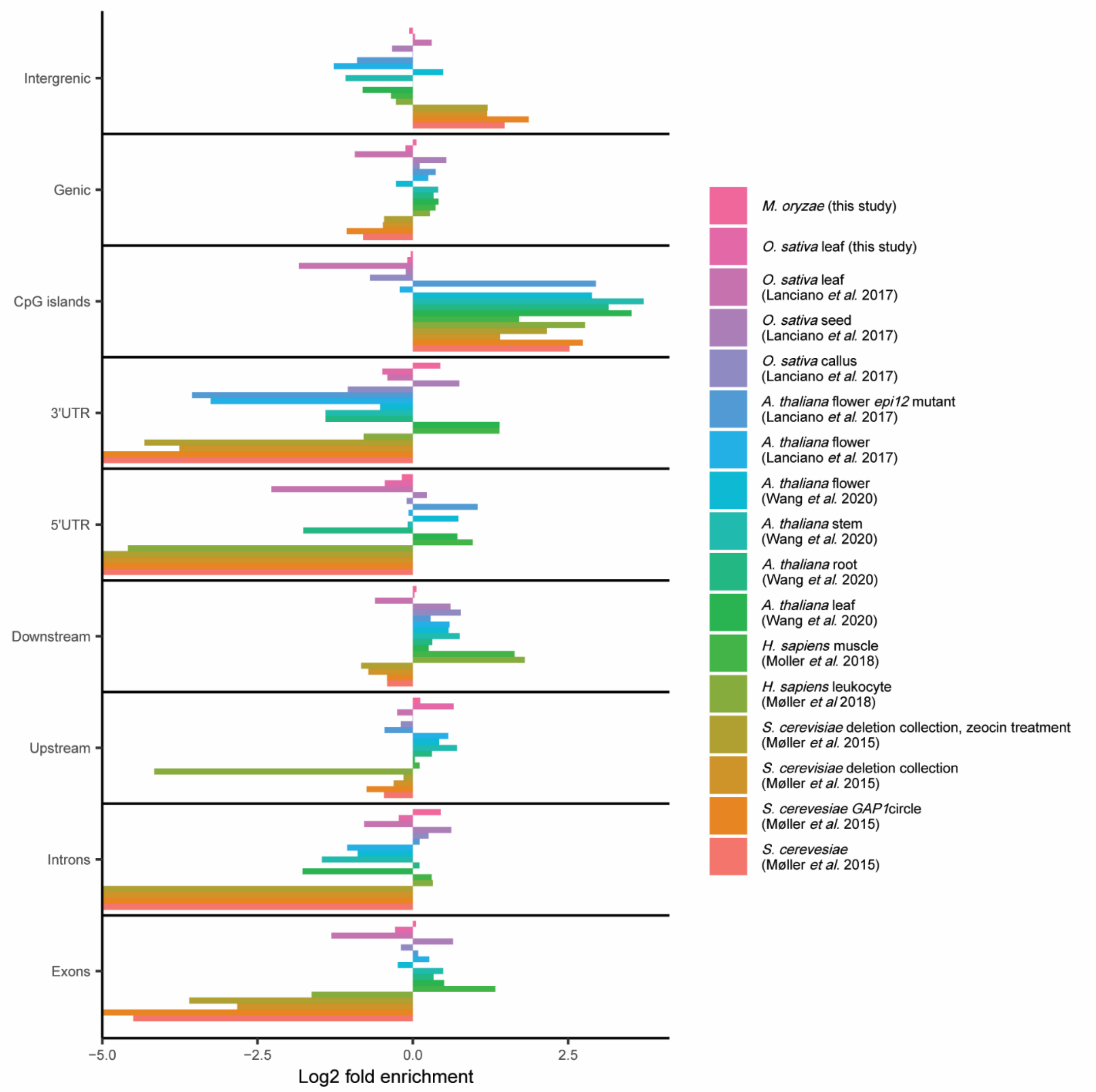


Fig. S16. MicroDNA enrichment and depletion in the genomes of various organisms. Bar plot showing observed enrichment of microDNAs across various regions of the genome across different previously sequenced organisms and sample types. Log2 fold enrichment of -5 represents samples where no microDNAs were found in that region. The presented data is an average of all sequenced samples of each type.

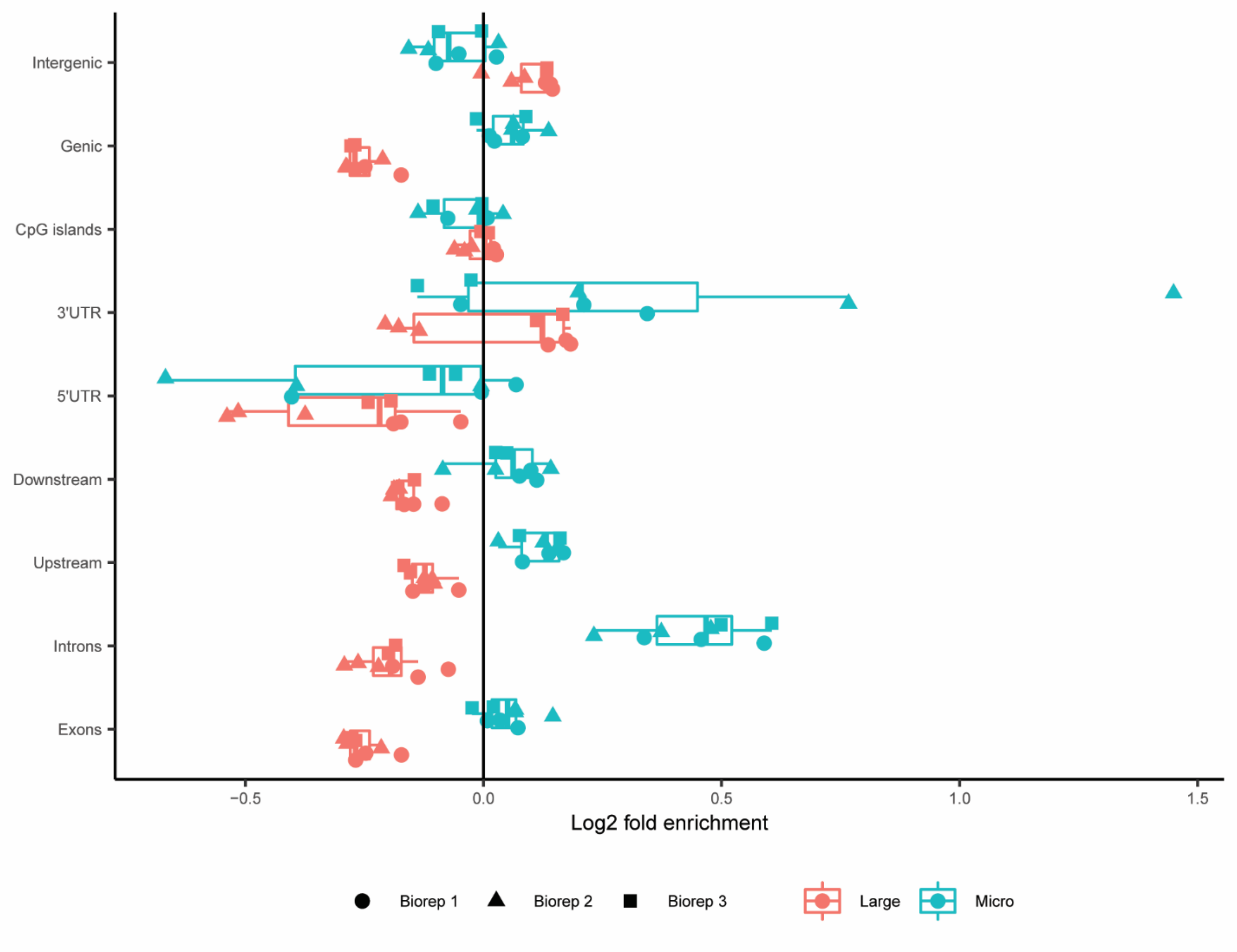


Fig. S17. Enrichment and depletion of microDNAs and large eccDNAs across various genomic regions in *M. oryzae*. Box plot showing observed enrichment of microDNAs and large eccDNAs across various regions of the genome. Each point represents one sample, and the shape of the points represent the biological replicate that sample was taken from.

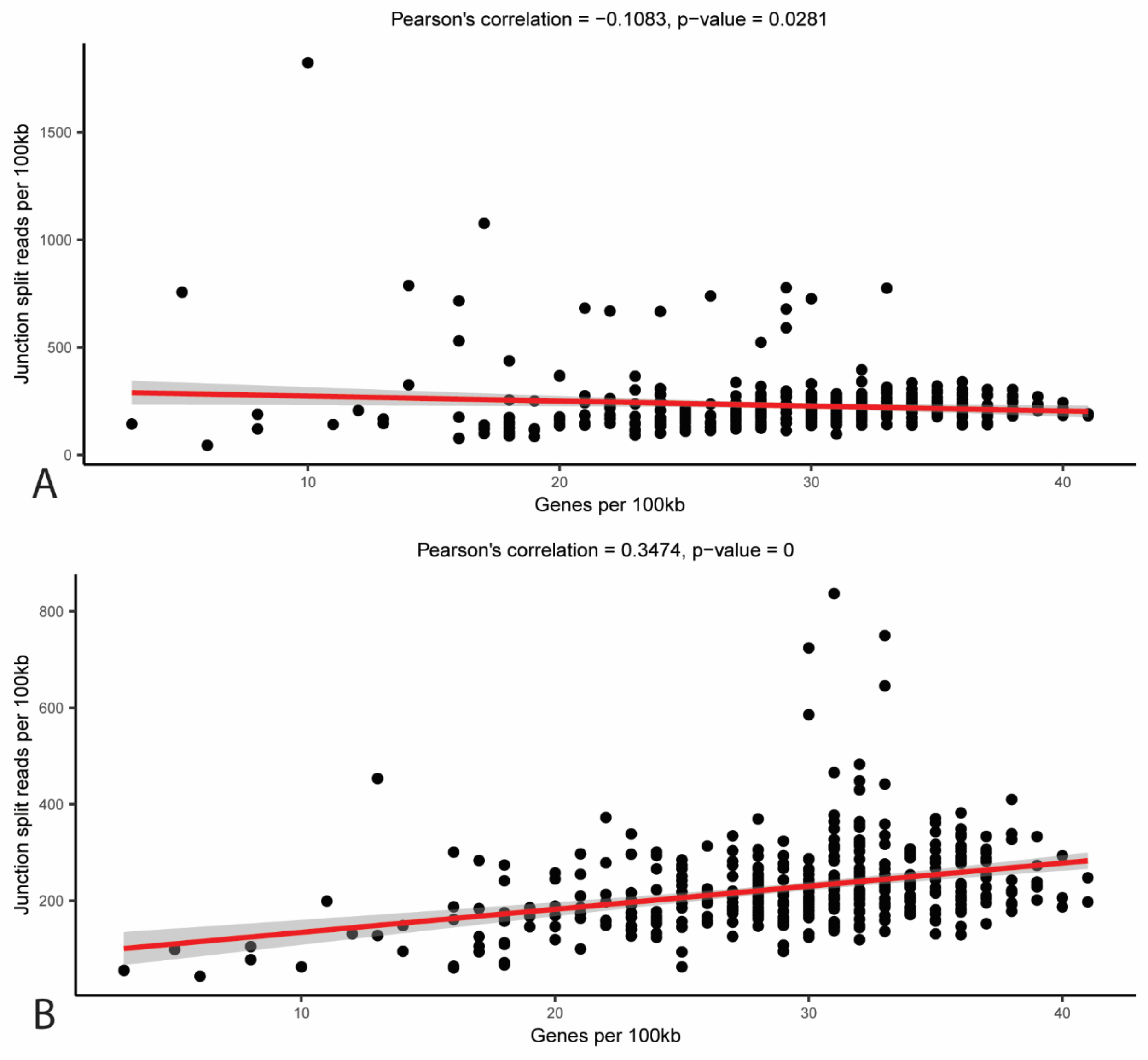


Fig. S18. Correlation between gene count and junction split read count across the *M. oryzae* genome. Scatter plot showing the number of genes and log 10 of the number of junction split reads per million per 100 kilobase pair bin in the *M. oryzae* Guy11 genome for **A.** large eccDNAs or **B.** microDNAs, averaged across biological replicates. The red line represents a linear regression line and the grey shadow represents 95% confidence intervals.

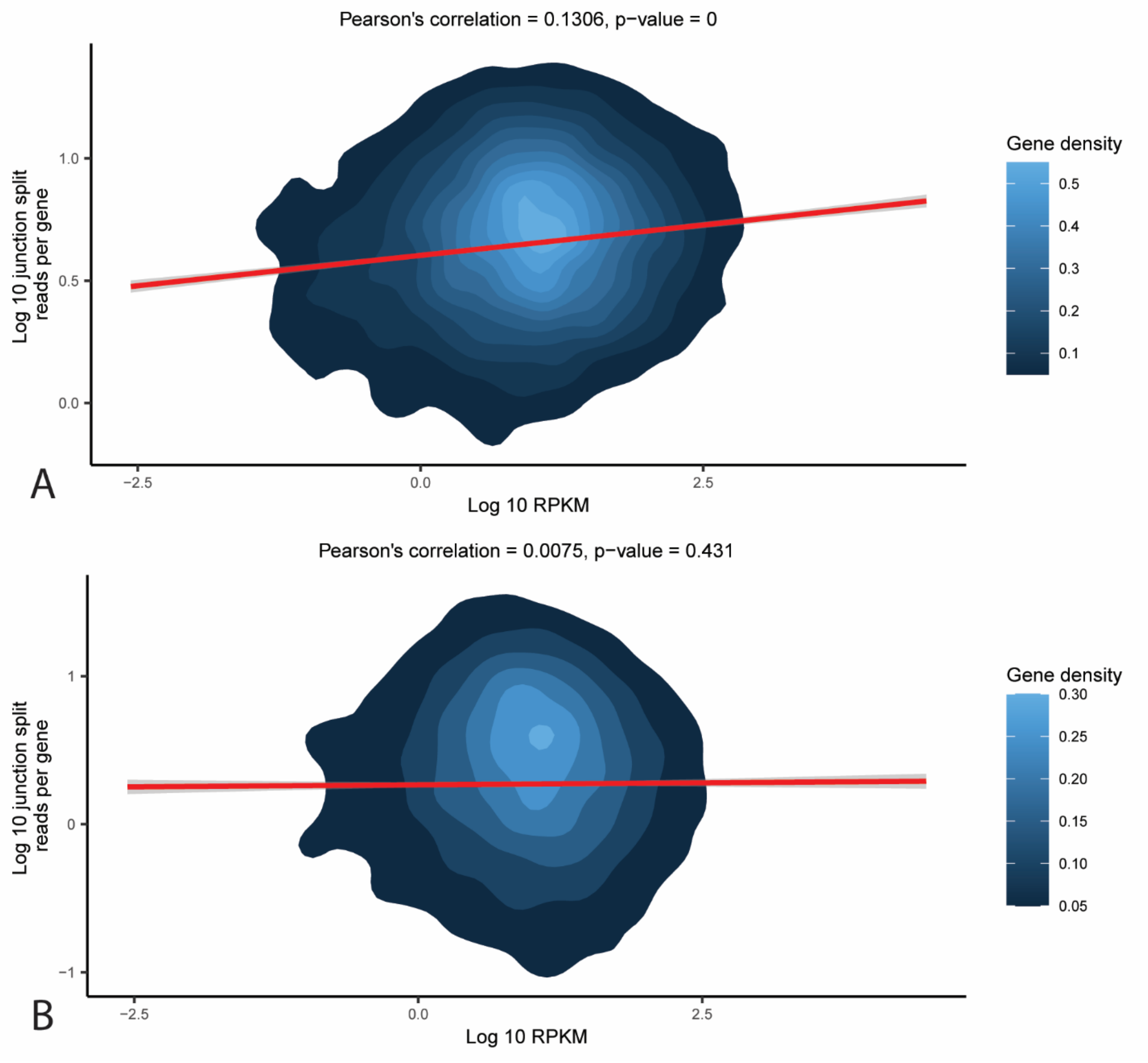


Fig. S19. Correlation between junction split read count and expression for *M. oryzae* genes. Two-dimensional density plot showing the log 10 of the reads per kilobase million averaged across multiple RNAseq samples and log 10 of the number of overlapping junction split reads per million for each gene for **A.** large eccDNAs and **B.** microDNAs, averaged across biological replicates. The red line represents a linear regression line and the grey shadow represents 95% confidence intervals.

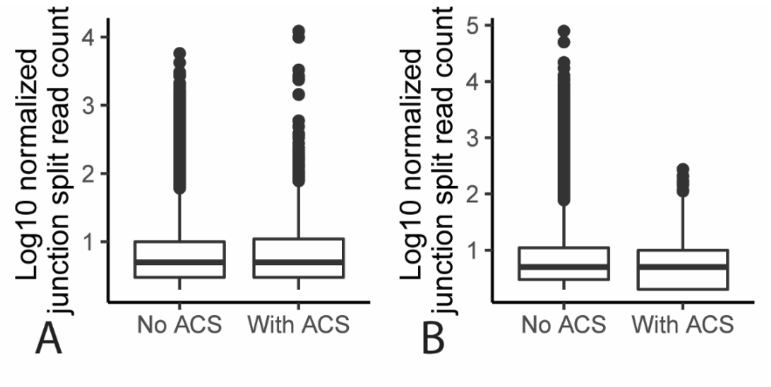


Fig. S20. Comparison of junction split read counts between eccDNA forming regions with and without an ACS. Box plot showing the log 10 of the number of junction split reads per million reads averaged across biological replicates for eccDNA forming regions that do and do not contain ACSs for **A.** large eccDNAs and **B.** microDNAs.

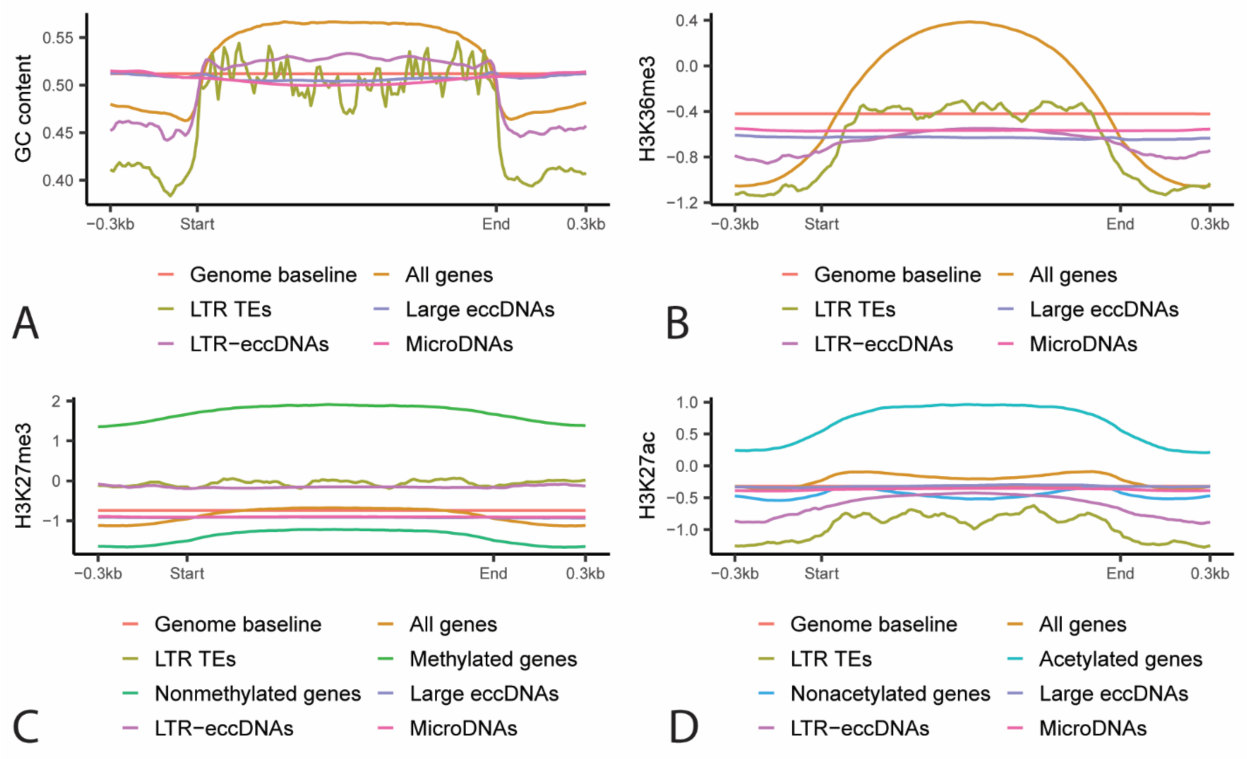


Fig. S21. GC content and chromatin marks of eccDNA forming regions in *M. oryzae*. Profile plots showing the average **A.** percent GC content, **B.** log2 ratio of read coverage for H3K36me3 chromatin immunoprecipitation and input control, **C.** log2 ratio of read coverage for H3K27me3 chromatin immunoprecipitation and input control and **D.** log2 ratio between read coverage for H3K27ac chromatin immunoprecipitation and input control for all *M. oryzae* genes, randomly selected regions of the genome, LTR retrotransposons, large eccDNAs, LTR-eccDNAs and

microDNAs. Methylated and nonmethylated genes and acetylated and nonacetylated genes, as defined by Zhang *et al.*, are also represented in **C.** and **D.**, respectively.

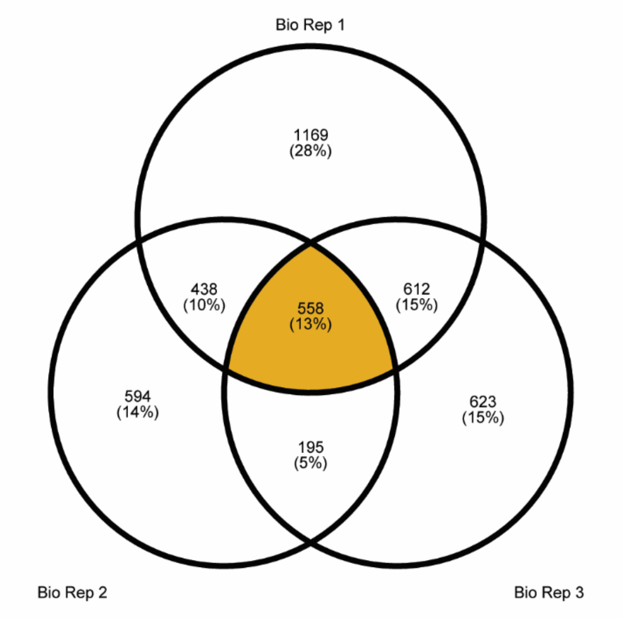


Fig. S22. Overlap between genes enriched on eccDNAs in biological replicates. Venn diagram showing overlap between genes in the top 33% for how often they were found fully encompassed by eccDNA forming regions in each biological replicate. Technical replicates for each biological replicates were normalized to the number of junction split reads in each sample then averaged. 558 genes found in the top 33% for all biological replicates were designated eccDNA-associated (colored in orange).

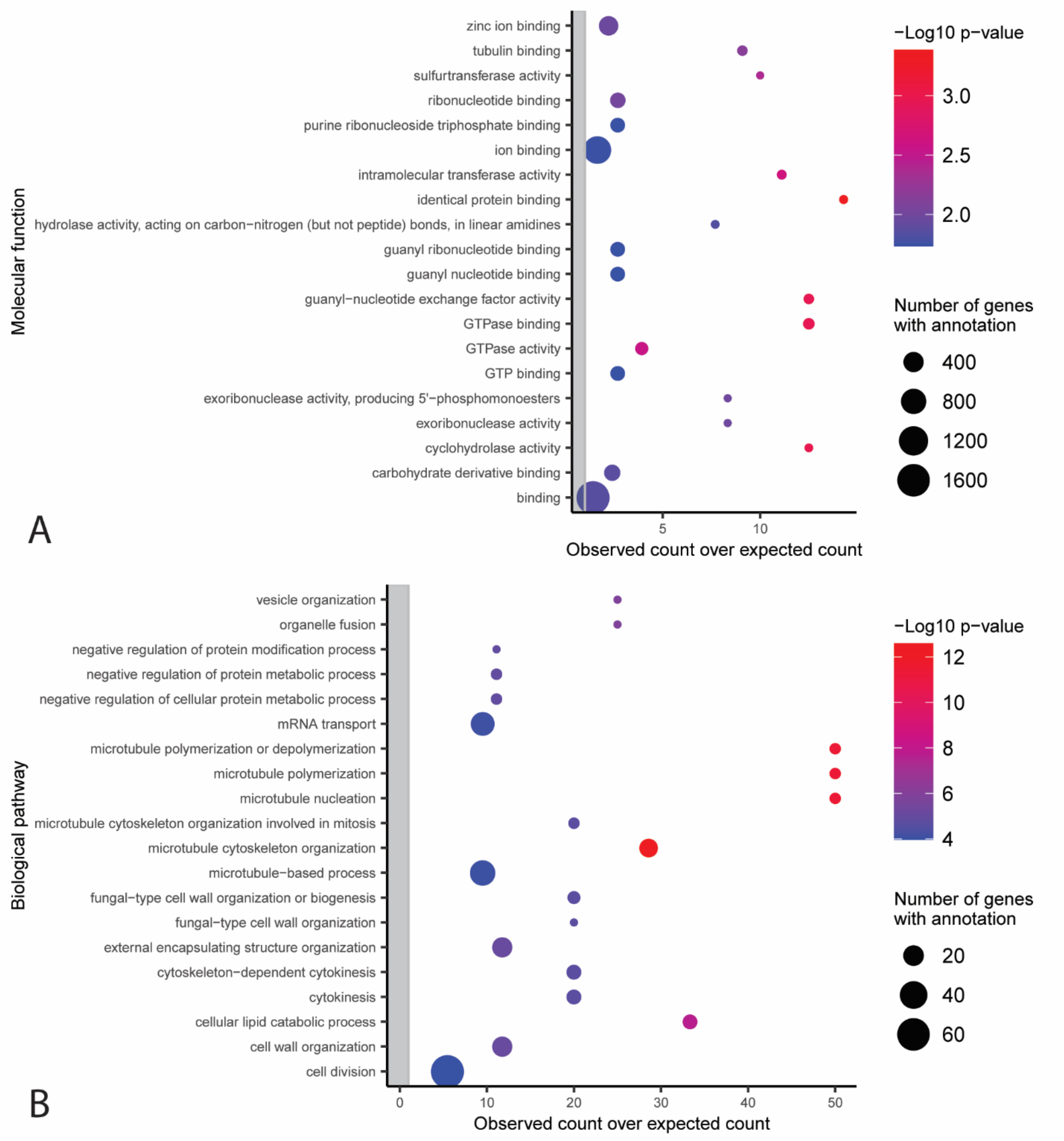


Fig. S23. GO terms associated with eccDNA-associated genes. Functional categories in the **A.** molecular function and **B.** biological pathway Gene Ontology with an observed number of eccDNA-associated genes that is significantly different from the expected number with correction for gene length bias (Chi-square test, p < 0.05). The y-axis shows the different functional categories, and the x-axis represents the observed number of genes divided by the expected number of genes in this group. Dots outside of the grey rectangle represent functional categories that are observed more often than expected. The size of dots indicates the total number of genes in the *M. oryzae* genome that belong to each functional category. Only the 20 categories with the largest -log10 p-values are shown.

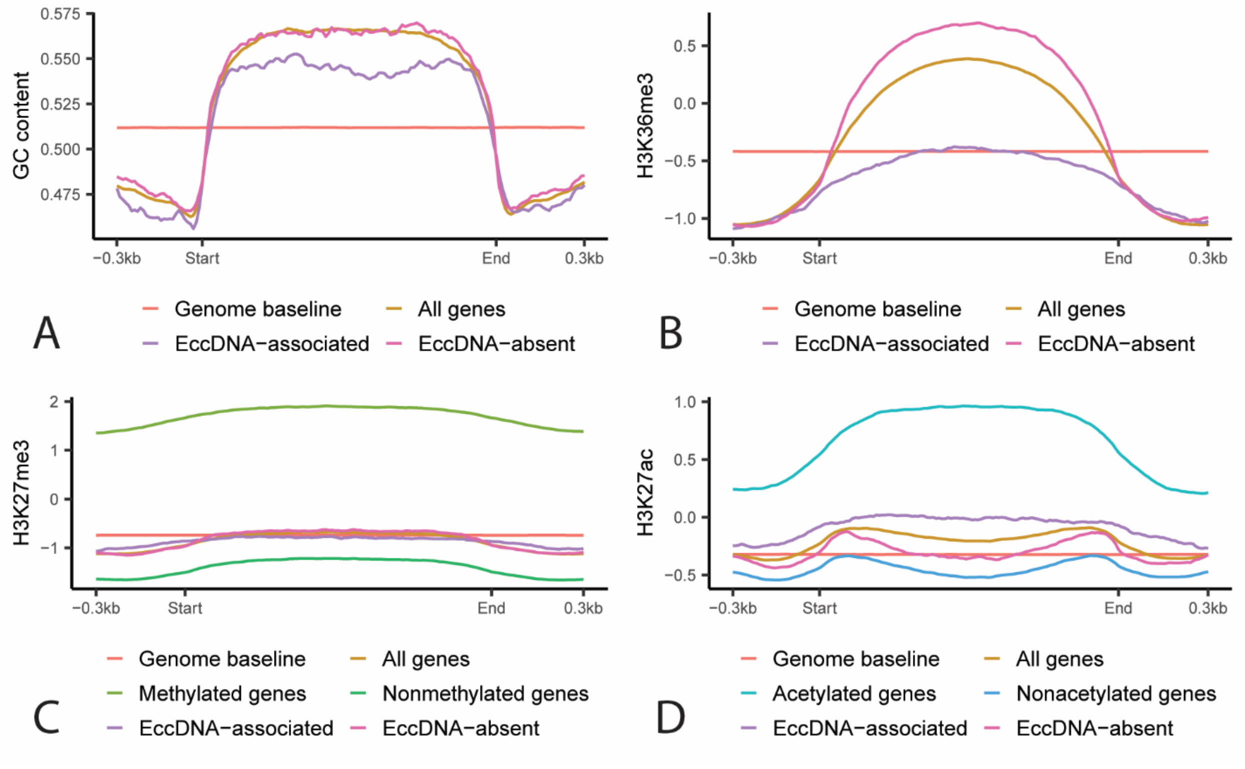


Fig. S24. GC content and chromatin marks of eccDNA-associated and eccDNA-absent genes in *M. oryzae*. Profile plots showing the average **A.** percent GC content, **B.** log2 ratio between read coverage for H3K36me3 chromatin immunoprecipitation and input control, **C.** log2 ratio between read coverage for H3K27me3 chromatin immunoprecipitation and input control and **D.** log2 ratio between read coverage for H3K27ac chromatin immunoprecipitation and input control for all *M. oryzae* genes, randomly selected regions of the genome, eccDNA-associated genes, and eccDNA-absent genes. Methylated and nonmethylated genes and acetylated and nonacetylated genes, as defined by Zhang *et al*., are also represented in **C.** and **D.**, respectively.

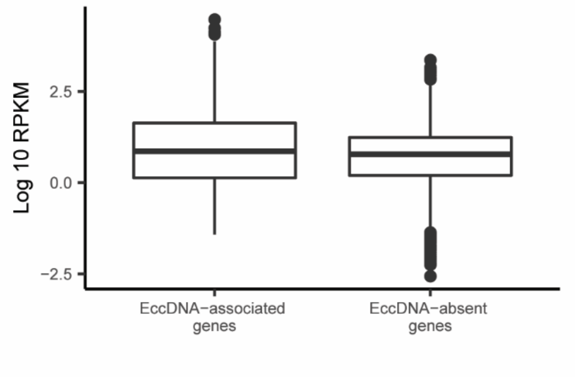


Fig. S25. Comparison of expression data between eccDNA-associated genes and eccDNA-absent genes in *M. oryzae*. Box plot showing the log 10 reads per kilobase million (RPKM) averaged across 12 previously published RNAseq samples for eccDNA-associated genes and eccDNA-absent genes.

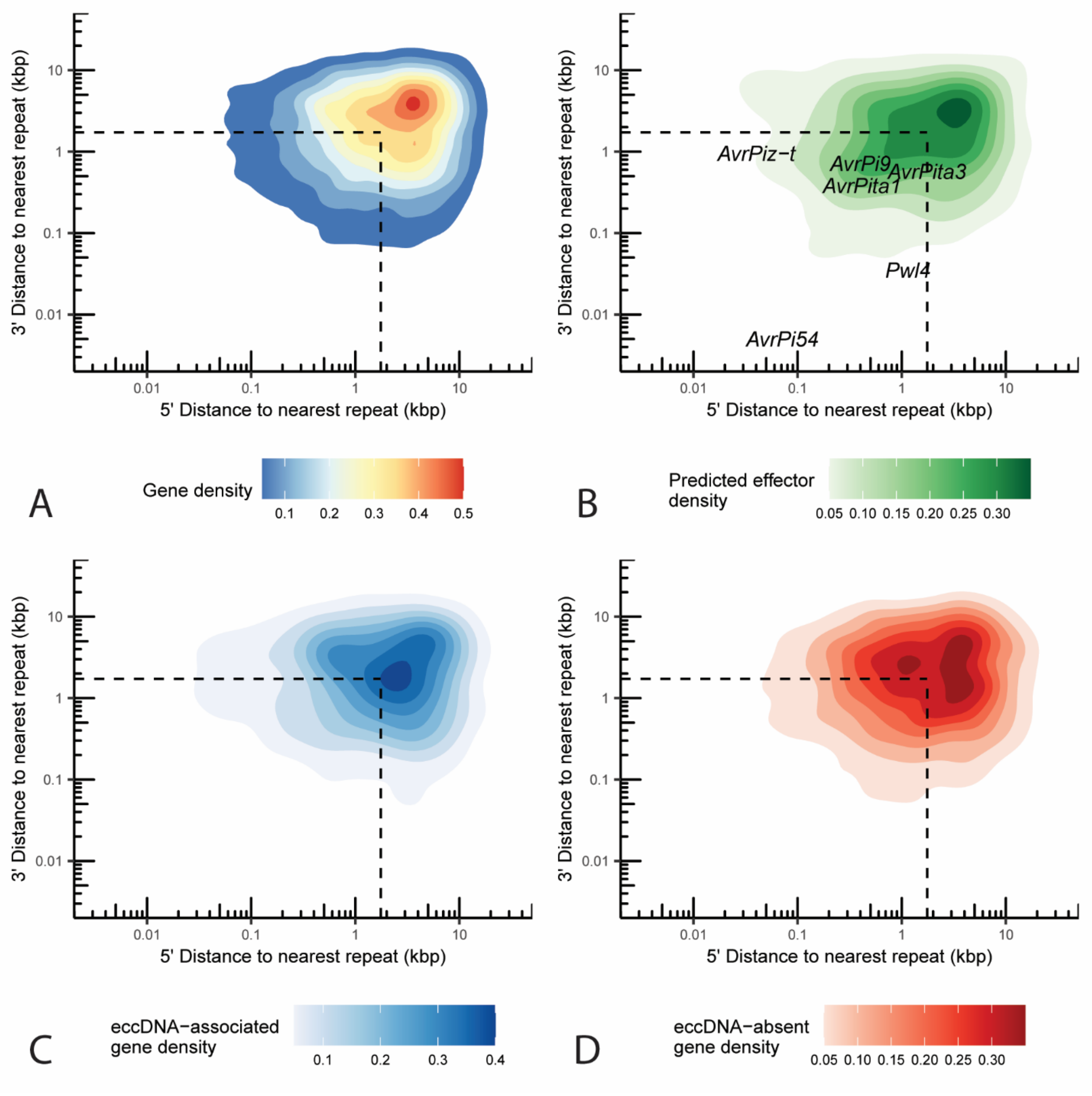


Fig. S26. Proximity of *M. oryzae* genes to repeats. Two-dimensional density plot representing the 5’ and 3’ distance to the nearest repeat in the *M. oryzae* Guy11 genome in kilobase pairs for each **A.** gene, **B.** predicted effector, **C.** eccDNA-associated genes, and **D.** eccDNA-absent genes. Known effectors are shown as text in **B.** Dashed lines represent median 5’ and 3’ distance to nearest gene.

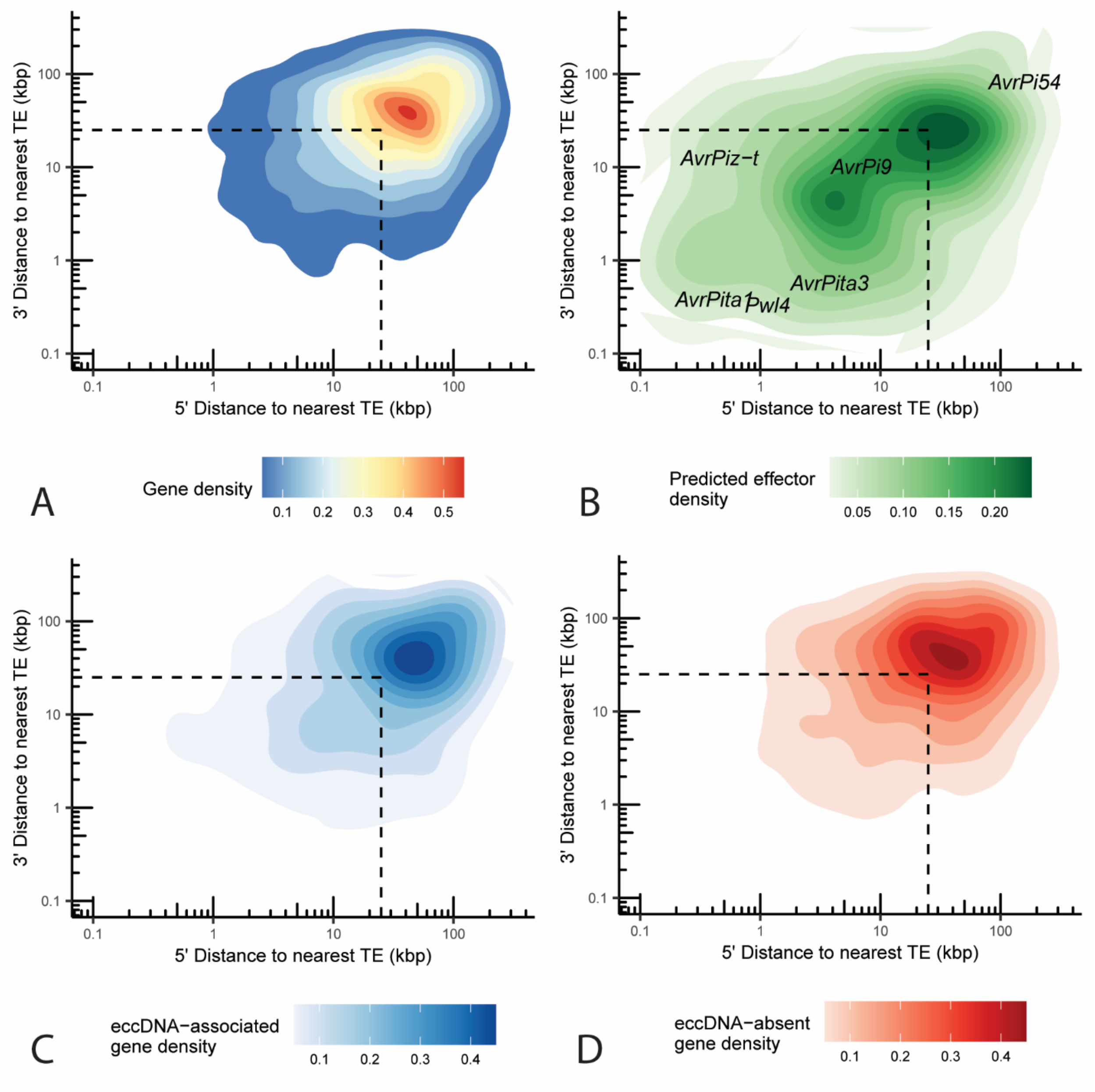


Fig. S27. Proximity of *M. oryzae* genes to TEs. Two-dimensional density plot representing the 5’ and 3’ distance to the nearest transposable element in the *M. oryzae* Guy11 genome in kilobase pairs for each **A.** gene, **B.** predicted effector, **C.** eccDNA-associated genes, and **D.** eccDNA-absent genes. Known effectors are shown as text in **B.** Dashed lines represent median 5’ and 3’ distance to nearest gene.

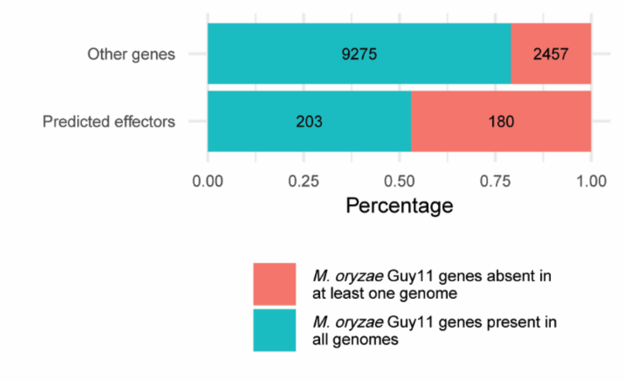


Fig. S28. Predicted effectors are prone to presence-absence variation in *M. oryzae*. Stacked bar plot showing the percentage of predicted effectors and all other genes in the *M. oryzae* Guy11 genome that had an ortholog in all other 162 *M. oryzae* genomes analyzed or not. Numbers indicate the number of genes in each category.

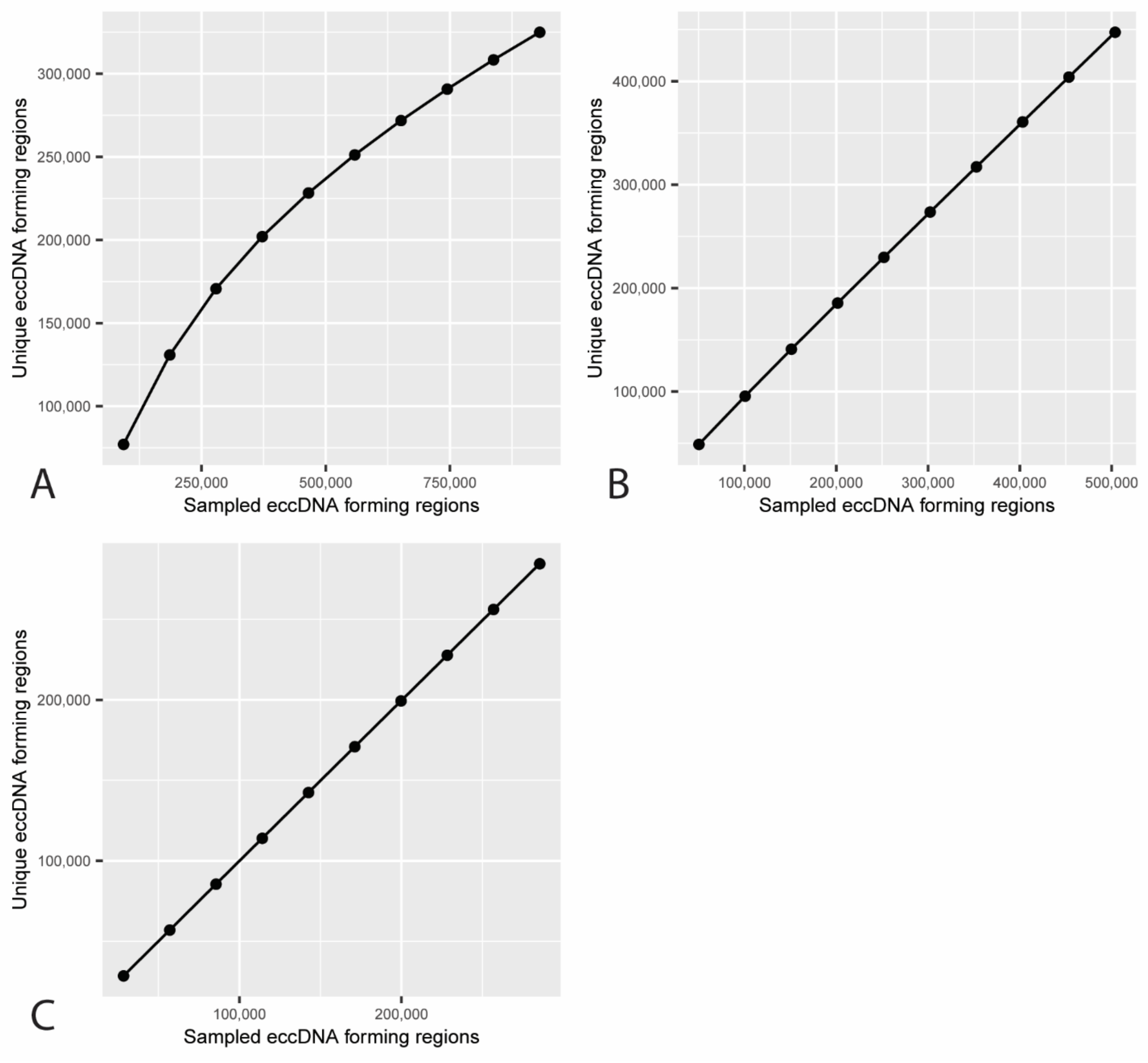


Fig. S29. Rarefaction curves for eccDNA forming regions in *M. oryzae*. Rarefaction analysis of the number of unique eccDNA forming regions at different subsamples of eccDNA forming regions across all samples for **A.** LTR-eccDNAs, **B.** large eccDNAs and **C.** microDNAs.

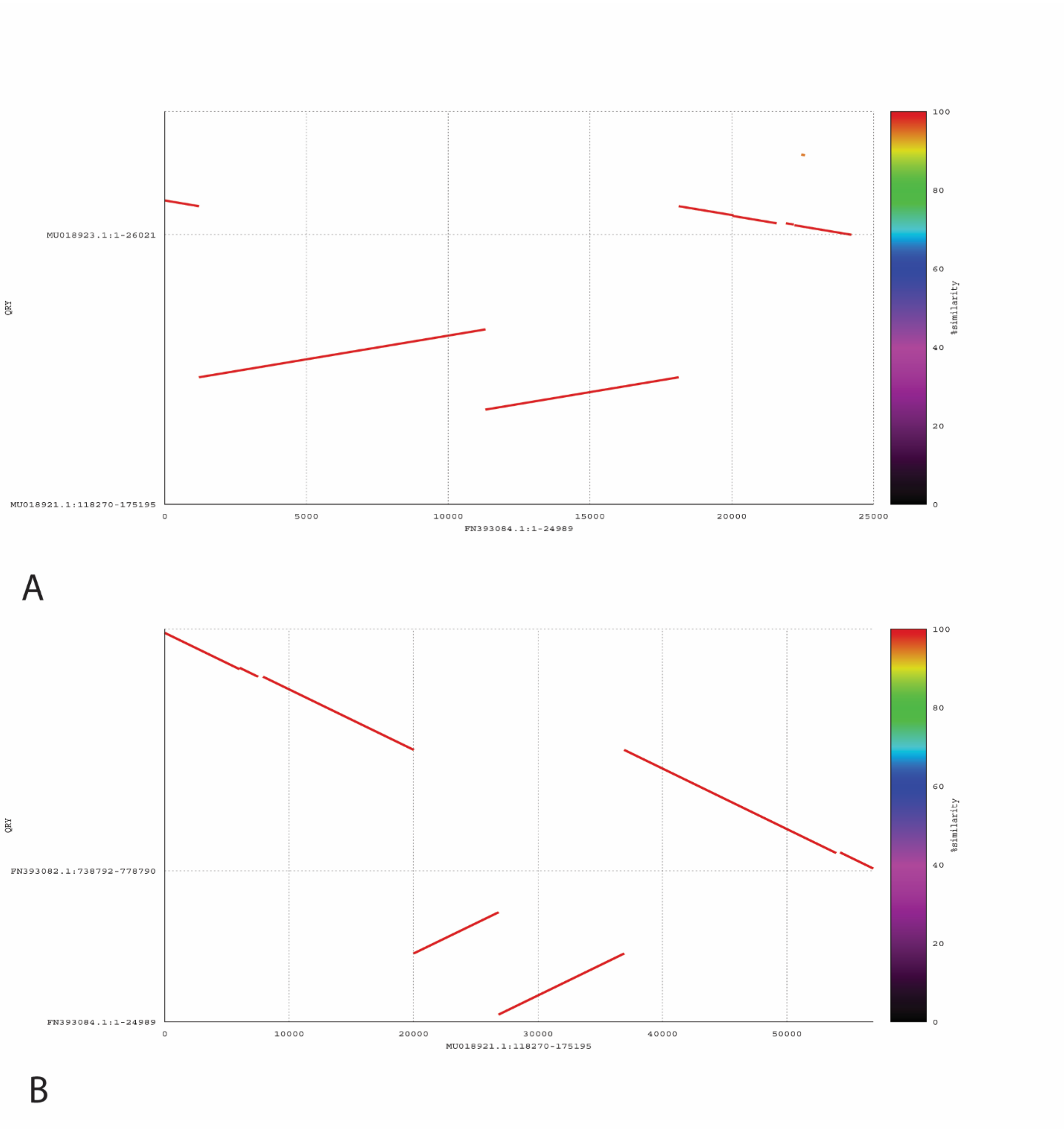


Fig. S30. Example of an eccDNA-mediated translocation in wine yeasts. Dot plot alignments between *S. cerevisiae* M22 and *S. cerevisiae* EC1118 genomes showing a DNA translocation likely caused by an eccDNA intermediate in yeast. **A.** A scaffold of the EC1118 genome aligns to two different scaffolds of the M22 genome. **B.** A scaffold of the M22 genome aligns to two different scaffolds of the EC1118 genome.

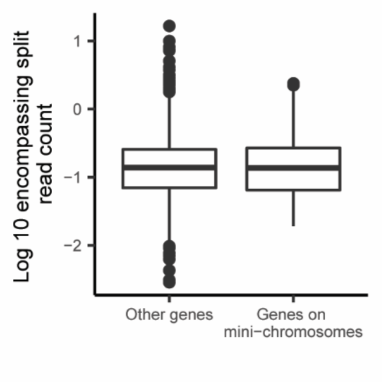


Fig. S31. Comparison of encompassing split read counts between genes found on mini-chromosomes in *M. oryzae* and other genes. Box plot showing the log 10 of the number of junction split reads per million reads averaged across biological replicates that fully encompass genes previously found on mini-chromosomes in other strains of *M. oryzae* and other genes.

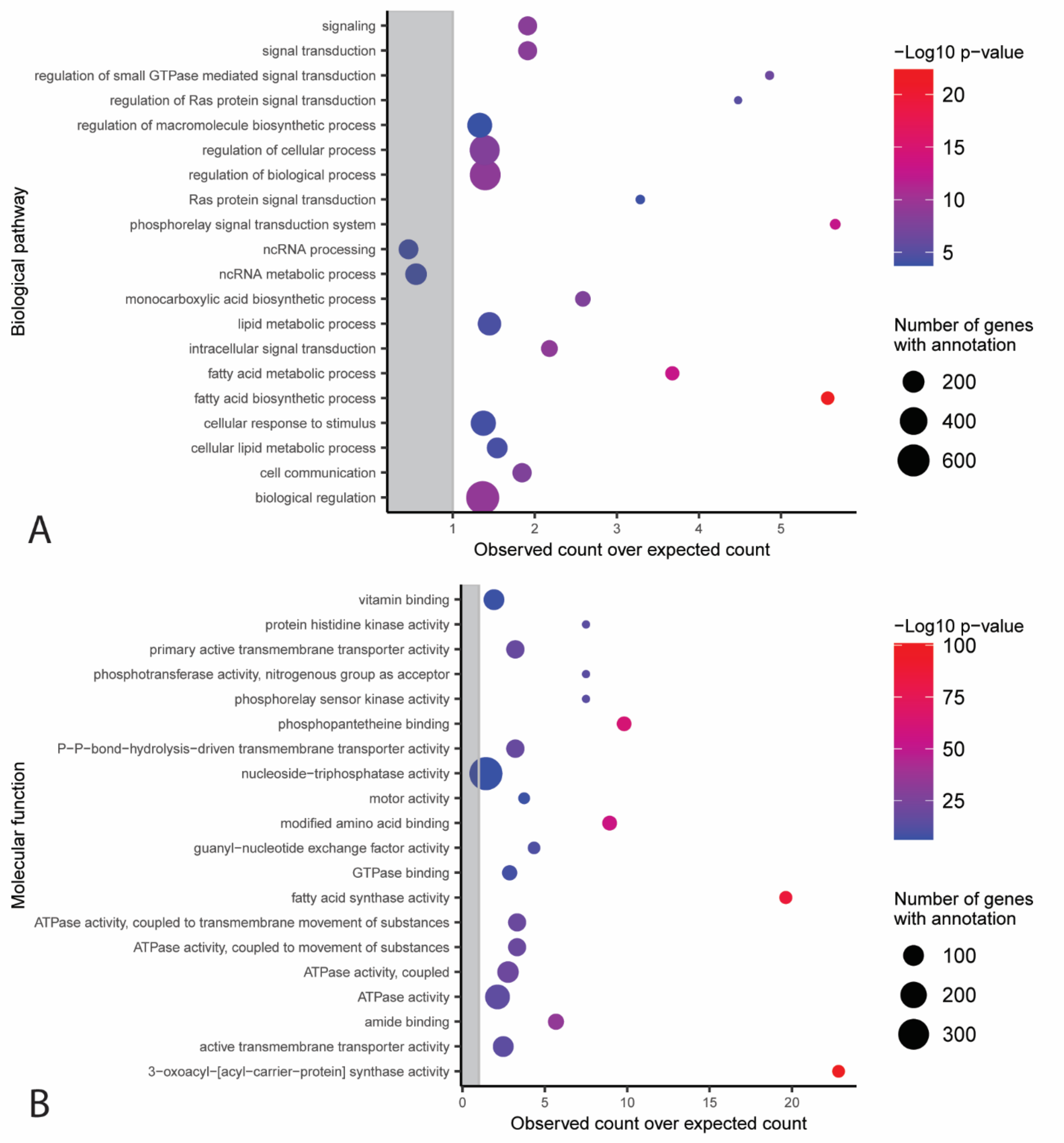


Fig. S32. GO terms associated with eccDNA-absent genes. Functional categories in the **A.** molecular function and **B.** biological pathway Gene Ontology with an observed number of eccDNA-absent genes that is significantly different from the expected number with correction for gene length bias (Chi-square test, $p < 0.05$). The y-axis shows the different functional categories, and the x-axis represents the observed number of genes divided by the expected number of genes in this group. Dots outside of the grey rectangle represent functional categories that are observed more often than expected. The size of dots indicates the total number of genes in the *M. oryzae* genome that belong to each functional category. Only the 20 categories with the largest -log10 p-values are shown.

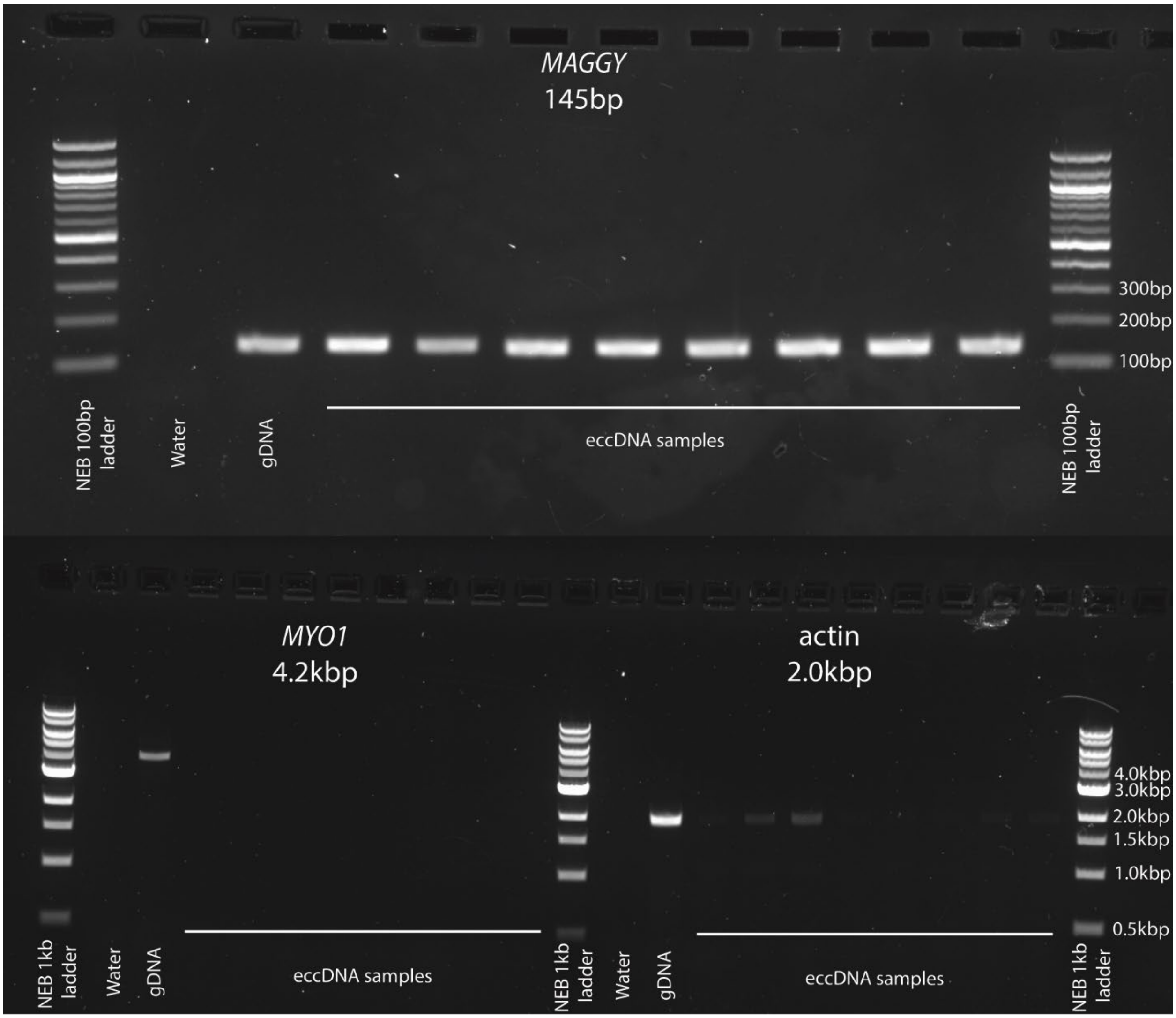


Fig. S33. PCR validation of eccDNA-absent genes. Features of interest are listed at the top of each group. One primer set was used per group and the expected product size is written below the feature name. A portion of the *MAGGY* LTR retrotransposon was used as a positive control for amplification. EccDNA samples were grouped by biological replicate and ordered within groups by technical replicate. All samples for each product were from the same PCR reaction.

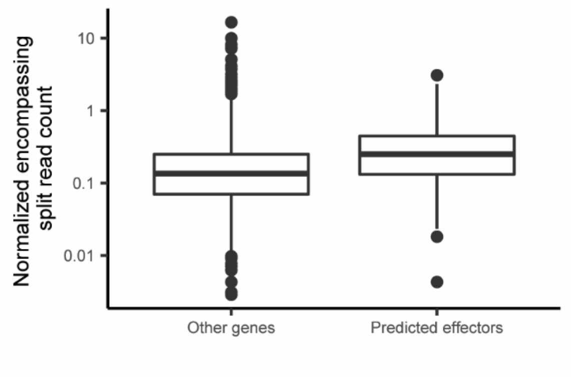


Fig. S34. Effectors are enriched in eccDNAs in M. oryzae. Box plot showing the number of fully encompassing junction split reads per million junction split reads averaged across biological replicates for predicted effectors compared to all other genes.

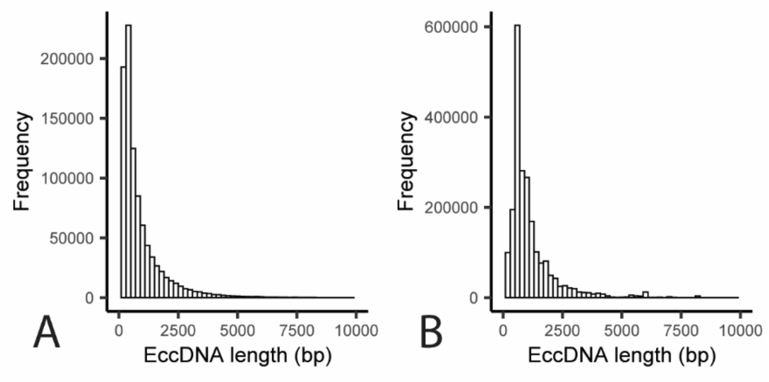


Fig. S35. Lengths of eccDNA forming regions in *M. oryzae*. Histograms showing the distribution of candidate eccDNA forming regions in *M. oryzae* for one sequenced sample. **A.** Length distribution of candidate eccDNA forming regions inferred from uniquely mapped reads. **B.** Length distribution of candidate eccDNA forming regions inferred from multi-mapping reads.

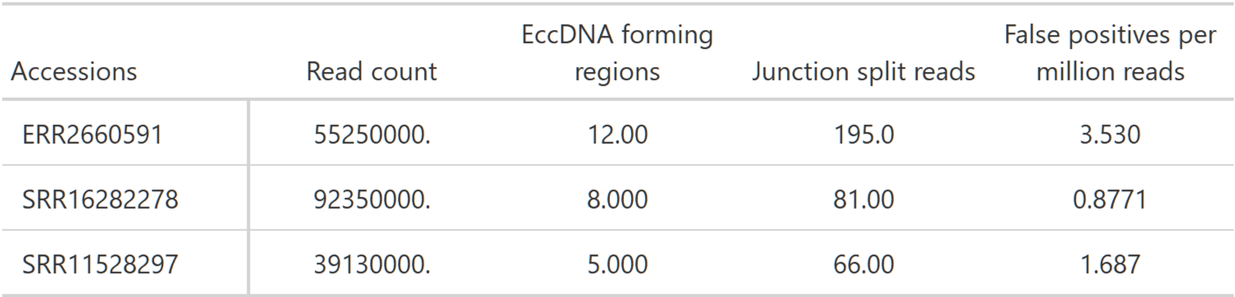
Additional File 2: List of eccDNA forming regions called using Illumina circularome sequencing data for *M. oryzae* in this study.

The first column describes the sample the eccDNA forming region was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 3: List of eccDNA forming regions called using PacBio circularome sequencing data for *M. oryzae* in this study.

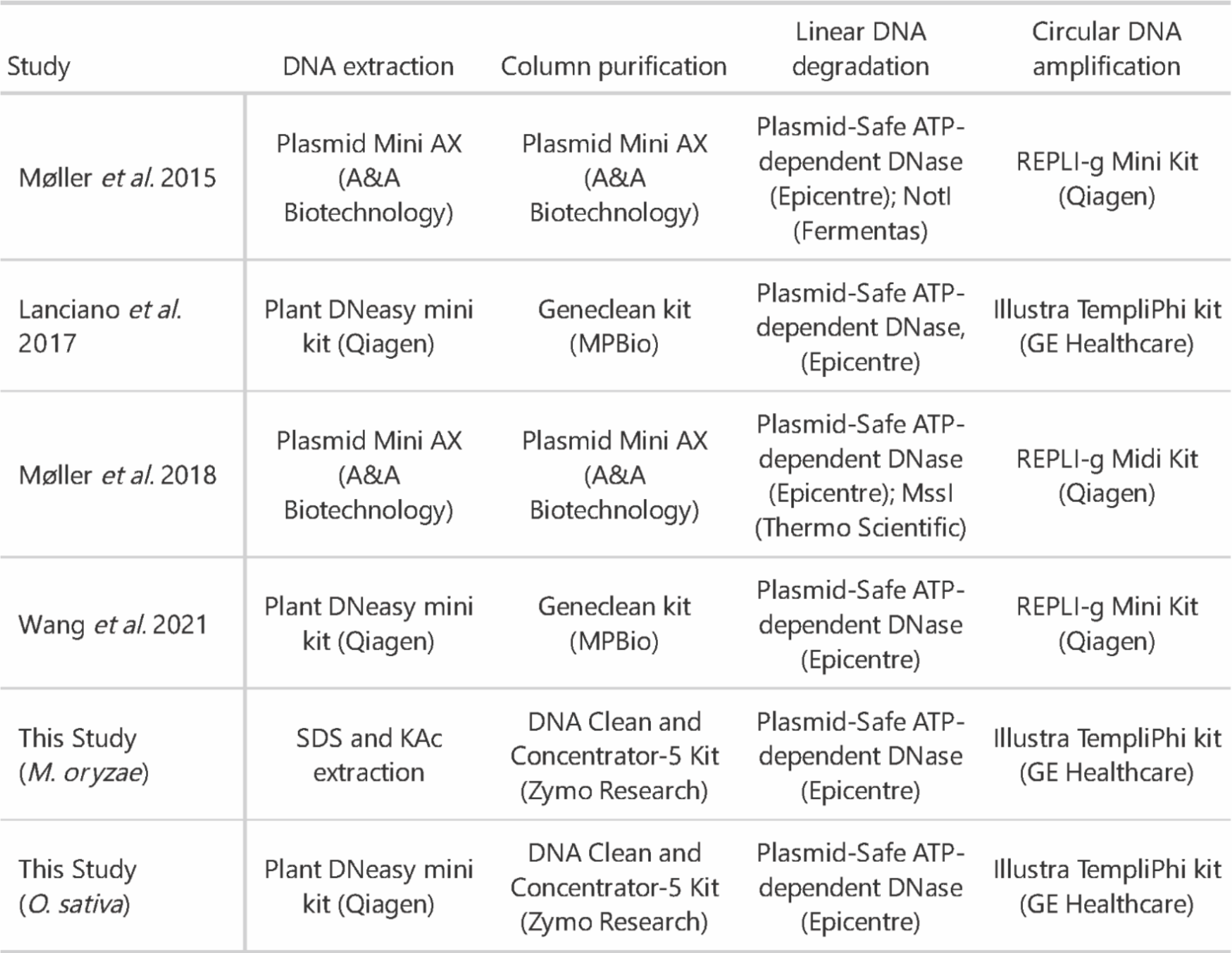
The first column describes the sample the eccDNA forming region was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 4: Supplementary Tables.



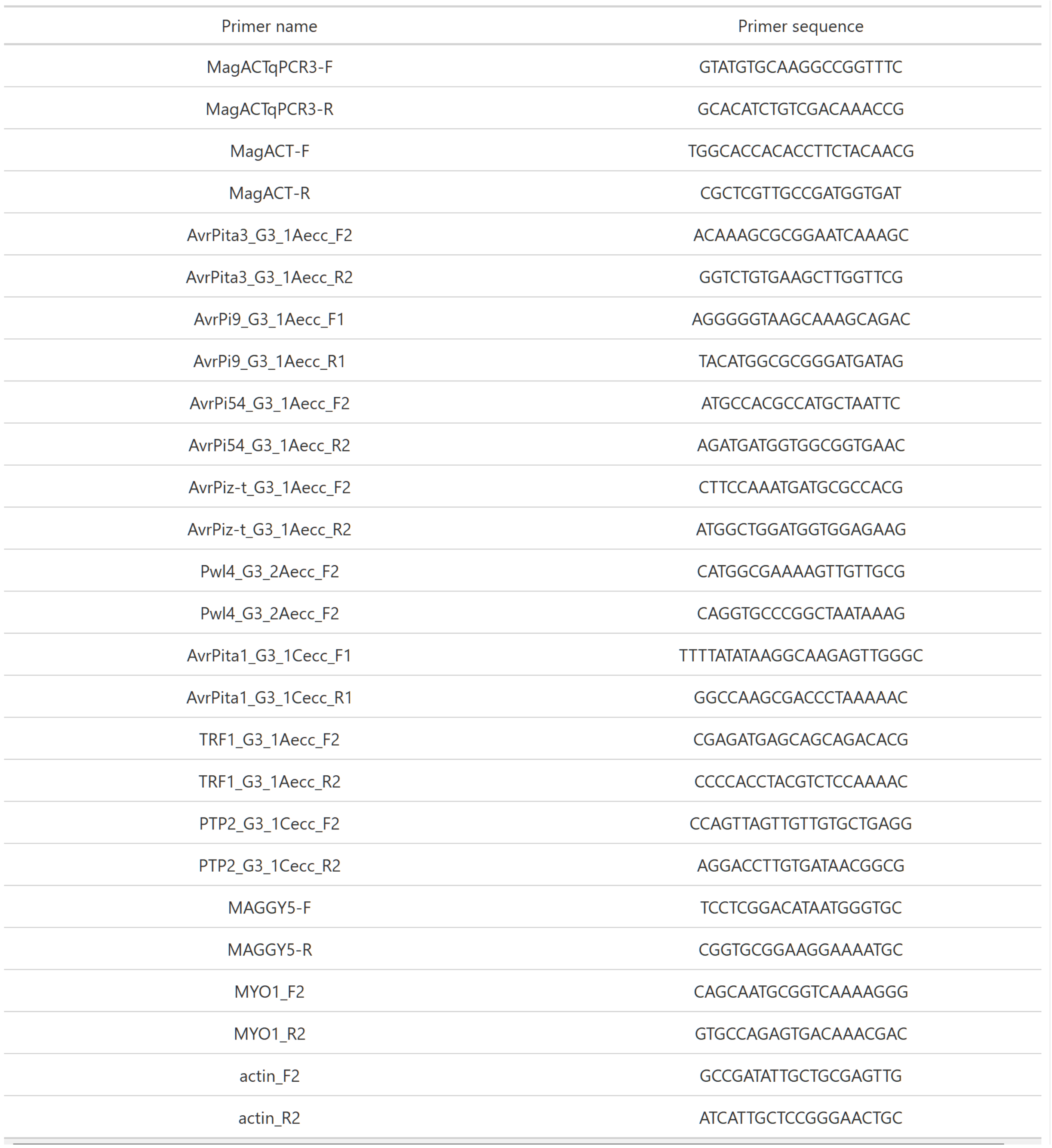
Accessions	Read count	EccDNA forming regions	Junction split reads	False positives per million reads
ERR2660591	55250000.	12.00	195.0	3.530
SRR16282278	92350000.	8.000	81.00	0.8771
SRR11528297	39130000.	5.000	66.00	1.687

Table S1. Number of eccDNA forming regions called using whole genome sequencing data. Read count, eccDNA forming regions inferred, and number of junction split reads found using our pipeline on three previously published whole genome sequencing datasets for *M. oryzae*.



Study	DNA extraction	Column purification	Linear DNA degradation	Circular DNA amplification
Møller *et al.* 2015	Plasmid Mini AX (A&A Biotechnology)	Plasmid Mini AX (A&A Biotechnology)	Plasmid-Safe ATP-dependent DNase (Epicentre); NotI (Fermentas)	REPLI-g Mini Kit (Qiagen)
Lanciano *et al.* 2017	Plant DNeasy mini kit (Qiagen)	Geneclean kit (MPBio)	Plasmid-Safe ATP-dependent DNase, (Epicentre)	Illustra TempliPhi kit (GE Healthcare)
Møller *et al.* 2018	Plasmid Mini AX (A&A Biotechnology)	Plasmid Mini AX (A&A Biotechnology)	Plasmid-Safe ATP-dependent DNase (Epicentre); MssI (Thermo Scientific)	REPLI-g Midi Kit (Qiagen)
Wang *et al.* 2021	Plant DNeasy mini kit (Qiagen)	Geneclean kit (MPBio)	Plasmid-Safe ATP-dependent DNase (Epicentre)	REPLI-g Mini Kit (Qiagen)
This Study (*M. oryzae*)	SDS and KAc extraction	DNA Clean and Concentrator-5 Kit (Zymo Research)	Plasmid-Safe ATP-dependent DNase (Epicentre)	Illustra TempliPhi kit (GE Healthcare)
This Study (*O. sativa*)	Plant DNeasy mini kit (Qiagen)	DNA Clean and Concentrator-5 Kit (Zymo Research)	Plasmid-Safe ATP-dependent DNase (Epicentre)	Illustra TempliPhi kit (GE Healthcare)

Table S2. Summary of protocols used to extract eccDNAs in studies analyzed in this manuscript. DNA extraction kit, column purification kit, linear DNA degradation enzymes and circular DNA amplification enzymes used for all studies whose data was used to compare the circularomes of the organisms discussed in this study.



Primer name	Primer sequence
MagACTqPCR3-F	GTATGTGCAAGGCCGGTTTC
MagACTqPCR3-R	GCACATCTGTCGACAAACCG
MagACT-F	TGGCACCACACCTTCTACAACG
MagACT-R	CGCTCGTTGCCGATGGTGAT
AvrPita3_G3_1Aecc_F2	ACAAAGCGCGGAATCAAAGC
AvrPita3_G3_1Aecc_R2	GGTCTGTGAAGCTTGGTTCG
AvrPi9_G3_1Aecc_F1	AGGGGGTAAGCAAAGCAGAC
AvrPi9_G3_1Aecc_R1	TACATGGCGCGGGATGATAG
AvrPi54_G3_1Aecc_F2	ATGCCACGCCATGCTAATTC
AvrPi54_G3_1Aecc_R2	AGATGATGGTGGCGGTGAAC
AvrPiz-t_G3_1Aecc_F2	CTTCCAAATGATGCGCCACG
AvrPiz-t_G3_1Aecc_R2	ATGGCTGGATGGTGGAGAAG
Pwl4_G3_2Aecc_F2	CATGGCGAAAAGTTGTTGCG
Pwl4_G3_2Aecc_F2	CAGGTGCCCGGCTAATAAAG
AvrPita1_G3_1Cecc_F1	TTTTATATAAGGCAAGAGTTGGGC
AvrPita1_G3_1Cecc_R1	GGCCAAGCGACCCTAAAAAC
TRF1_G3_1Aecc_F2	CGAGATGAGCAGCAGACACG
TRF1_G3_1Aecc_R2	CCCCACCTACGTCTCCAAAAC
PTP2_G3_1Cecc_F2	CCAGTTAGTTGTTGTGCTGAGG
PTP2_G3_1Cecc_R2	AGGACCTTGTGATAACGGCG
MAGGY5-F	TCCTCGGACATAATGGGTGC
MAGGY5-R	CGGTGCGGAAGGAAAATGC
MYO1_F2	CAGCAATGCGGTCAAAAGGG
MYO1_R2	GTGCCAGAGTGACAAACGAC
actin_F2	GCCGATATTGCTGCGAGTTG
actin_R2	ATCATTGCTCCGGGAACTGC

Table S3. Primers used for qPCR validation of linear DNA degradation and outward PCR validation of eccDNA forming regions.

Additional File 5: List of eccDNA forming regions called using Illumina circularome sequencing data for *H. sapiens* muscle tissue published by Møller *et al.* [41]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 6: List of eccDNA forming regions called using Illumina circularome sequencing data for *H. sapiens* leukocytes published by Møller *et al.* [41]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 7: List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* in this study.

The first column describes the sample the eccDNA forming region was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 8: List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* leaf tissue published by Lanciano *et al.* [46]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 9: List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* seed tissue published by Lanciano *et al.* [46]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 10: List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* callus tissue published by Lanciano *et al.* [46]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 11: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* wild type tissue published by Lanciano *et al.* [46]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 12: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana epi12* mutant tissue published by Lanciano *et al.* [46]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 13: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* leaf tissue published by Wang *et al.* [45].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 14: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* root tissue published by Wang *et al.* [45].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 15: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* stem tissue published by Wang *et al.* [45].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 16: List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* flower tissue published by Wang *et al.* [45].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 17: List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* wild type cells published by Møller *et al.* [41].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 18: List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* GAP1^{circle} cells published by Møller *et al.* [47].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 19: List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* cells from the deletion collection published by Møller *et al.* [47].

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 20: List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* cells from the deletion collection treated with zeocin published by Møller *et al.* [47]**.**

The first column describes the sample the eccDNA was called with, the next three columns represent the genomic coordinates of the eccDNA forming region, and the last column represents the number of junction split reads used to call the eccDNA forming region.

Additional File 21: List of genes annotated in the *M. oryzae* Guy11 genome along with other information discussed in this study for each gene.

The first three columns describe the genomic coordinates of the gene, the fourth column is the gene's ID, the fifth column describes whether the gene was predicted to be an effector, the sixth column lists its name if it is a known effector, the seventh column lists the name of the protein in the *M. oryzae* 70-15 proteome, the eighth column describes whether it is an eccDNA-associated or eccDNA-absent gene, and the last column describes whether this gene was kept in all rice-infecting *M. oryzae* genomes analyzed.

Additional File 22: Enriched GO terms in the cellular components ontology for eccDNA-associated genes.

The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-associated category, the fourth column list the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 23: Enriched GO terms in the molecular function ontology for eccDNA-associated genes.

The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-associated category, the fourth column lists the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 24: Enriched GO terms in the biological pathway ontology for eccDNA-associated genes.

The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-associated category, the fourth column list the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 25: List of GenBank accessions for the genomes of rice-infecting *M. oryzae* isolates used in this study for gene annotation.

Additional File 26: List of small, genic deletions identified in the *M. oryzae* Guy11 genome.

The first three columns describe genomic coordinates of the deletion, the fourth column is the missing gene's ID, and the last column is the name of the genome where the deletion is present.

Additional File 27: List of GenBank accessions for the genomes of *M. oryzae* used in this study to search for eccDNA-mediated translocations.

Additional File 28: Enriched GO terms in the cellular components ontology for eccDNA-absent genes.

The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-absent category, the fourth column list the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 29: Enriched GO terms in the molecular function ontology for eccDNA-absent genes.

The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-absent category, the fourth column list the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 30: Enriched GO terms in the biological pathway ontology for eccDNA-absent genes. The first column lists the GO term, the second column lists the number of genes annotated with each term, the third column lists the number of genes observed in the eccDNA-absent category, the fourth column list the number of genes expected in that category, the fifth column shows is a description of the go term, the sixth column lists the Chi-square value for that GO term, and the final column lists the ratio of the observed number of genes in the eccDNA-associated category divided by the expected number of genes in that category.

Additional File 31: List showing names of barcodes used for each PacBio sequencing sample.

Additional File 32: Sequences of barcodes used for library preparation of PacBio sequencing samples in FASTA format.

Additional File 33: Consensus sequences of LTR retrotransposons in the *M. oryzae* Guy11 genome in FASTA format.

Additional File 34: Genome, gene annotation, and transposable element annotation files used for comparative circularome analysis.

Additional File 35: List of Sequence Read Archive accessions for RNAseq data used in this study.

Chapter 3

Extended Discussion and Conclusions of Chapter 2 and Transition to Chapter 4

In Chapter 2, I sequenced and characterized the eccDNAs of *M. oryzae*. Through a comparative analysis to previously published datasets, I found that *M. oryzae* eccDNAs are more diverse than those of other organisms and more likely to contain repetitive elements. While this difference was stark no matter what normalization or re-analysis approach I used, it is unclear why *M. oryzae* produces so many different eccDNAs. A few experiments could be done to help elucidate this question. In general, more organisms need to have their eccDNAs sequenced to perform comparative analyses and to start to understand what factors generate diversity in eccDNA profiles. This data is especially lacking in fungi as the only other fungus that has had their eccDNAs sequenced is *S. cerevisiae*. It would also be interesting to sequence the eccDNAs of many fungal plant pathogens to see if they all show similar levels of diversity as *M. oryzae*. This could help support the hypothesis that eccDNAs play a role in fungal plant pathogen evolution.

The *M. oryzae* species contains many host-specific populations, with distinct evolutionary histories [117,118]. Therefore, it could also be interesting to see how these evolutionary histories may affect their eccDNA profiles, and whether rice-infecting *M. oryzae* produce more eccDNAs than other pathotypes. Observing differences in eccDNA profiles between *M. oryzae* pathotypes could help narrow in on what factors define eccDNA content as well. To better understand the mechanisms that generate such a great diversity of eccDNAs, *M. oryzae* DNA repair mutant could also have their eccDNA profiles characterized. Interesting work is already being done on *M. oryzae* DNA repair mutants [119] and it would be quite interesting to see how *M. oryzae*'s unique DNA repair biology is related to its unique eccDNA profile. Previously published studies on the effect of mutations in DNA repair genes supports the idea that these two factors may be related [29].

Through my analysis, I also found that LTR retrotransposons are particularly overrepresented in *M. oryzae* eccDNAs. As a follow up to this observation, I described which ones were most likely to contribute to eccDNA formation and found that different LTR retrotransposons formed eccDNAs through different mechanisms. Since it is well known that TE activity varies between conditions [21,120], I hypothesized that the number and type of eccDNAs formed due to TE activity would vary in these conditions. I was also interested to see if eccDNA sequencing could be a better indicator of LTR retrotransposon activity than RNA sequencing given that RNA sequencing measures the expression of all LTR retrotransposons, even those that have mutated to the point that they are no longer able to be reverse-transcribed. Additionally, it is important to emphasize that the eccDNA sequencing data that I analyzed in Chapter 2 was only done in one specific condition. It is possible that the eccDNA profile and the number of eccDNAs produced by *M. oryzae* could change under different growth conditions, such as within the plant, under different stressors, or at different growth stages.

With these ideas in mind, I attempted to sequence eccDNAs under various conditions: during infection, after hydrogen peroxide treatment, and after methyl viologen treatment. Unfortunately, my analysis of these datasets was not successful due to technical reasons. Substantial troubleshooting was necessary to address the issues, but I did not end up performing this troubleshooting and therefore never got meaningful data from these experiments. However, the central question of Chapter 2 of my thesis was whether eccDNAs might directly contribute to *M. oryzae*'s adaptation, which is why I chose instead to focus my efforts on analyses that centered on the genic content of *M. oryzae* eccDNAs.

First, I identified sets of eccDNA-absent and eccDNA-associated genes. The existence of eccDNA-absent genes is an important finding as it shows that eccDNA formation in *M. oryzae* is not completely random and is likely shaped by some kind of biological selection. This selection could be acting at the level of individual cell fitness. For example, it is possible that a cell that generates eccDNAs containing copies of the actin gene expresses too much of this gene, resulting in an unstable cytoskeleton and an unhealthy cell. Of course, this hypothesis assumes that genes are expressed directly from eccDNAs. While this has been observed in other organisms, the only evidence I gathered during my research relevant to this hypothesis was that there is a weak correlation between eccDNA formation and previously published RNA sequencing data. This result needs to be validated with side-by-side RNA and eccDNA sequencing. Additionally, directly demonstrating expression from eccDNAs is a challenging task and showing a phenotypic effect associated with the amplification of a gene on an eccDNA under stress conditions is likely much easier. An alternative hypothesis for the existence of eccDNA-absent genes in *M. oryzae* is that there is some sort of mechanism that directly prevents the formation of these eccDNAs. H3K36me3 marks were the only feature of eccDNA-absent genes that set them apart from other genes in my analysis. Sequencing eccDNAs in histone methylation mutants could also help reveal whether a direct mechanism for controlling eccDNA-formation exists in *M. oryzae*, which could be particularly interesting to look at if evidence of eccDNAs affecting the phenotype of some *M. oryzae* cells under stress was uncovered.

Due to the co-localization of TEs and rapid evolution in the *M. oryzae* genome, the overwhelming presence of TEs on eccDNAs hinted at an association between eccDNAs and the evolution of *M.* oryzae. Identifying eccDNA-associated genes allowed me to further explore this correlation. I found that eccDNA-associated genes often experience PAV. I also found that eccDNA-associated genes are often in gene-poor regions and repeat-rich regions of the genome. Finally, I found that effectors are enriched on eccDNAs. This was encouraging evidence that supported the idea that eccDNAs play a role in the rapid evolution of *M. oryzae*. However, the main effect of eccDNA formation is gene amplification and it is not obvious how amplification of effector genes on eccDNAs alone could help *M. oryzae* adapt to its host. While increased expression of effectors could, in theory, help *M. oryzae* defeat host defenses more quickly, it is unlikely that a stochastic process like eccDNA formation would be very useful in accomplishing this. It is much easier to imagine stochastic gene copy number variation being useful in adaptation to environmental stressors, as has been discussed many times in the past [33,34,37,39,40]. This is why I was excited to find two genes known to be associated with fungicide-resistance captured by eccDNAs in my data. However, formation of eccDNAs containing these genes is not necessarily meaningful on its own and, again, fungicide resistance is not necessarily the result of increased expression of fungicide targets. To directly answer whether copy number variation caused by eccDNA formation could help *M. oryzae* adapt to its environment, eccDNA sequencing under different treatments, including fungicide treatments, would be required. However, as mentioned previously, while I attempted to perform these experiments, I was not able to draw meaningful conclusions from them.

EccDNAs can also contribute to genome evolution by causing genomic rearrangements. In this context, it is much easier to imagine how *M. oryzae* effectors could benefit from diversity generation through eccDNA formation. EccDNAs could shuttle effectors to different areas of the genome or to mini chromosomes, generating epigenetic diversity. If eccDNAs form without generating genomic deletions, they could reinsert resulting in effector gene duplications and enabling subsequent diversification. If an eccDNA forms while causing a genomic deletion, it could help *M. oryzae* escape detection by its host by removing an effector gene from the genome. EccDNAs could also mutate independently and then reinsert into the genome, generating diversity.

These eccDNA reinsertion events often leave characteristic scars that can be mined bioinformatically. Therefore, I focused much of my time during this project on finding these scars in the *M. oryzae* genome. However, despite using several exhaustive approaches, I could not find evidence of eccDNA translocation in *M. oryzae*. While it is possible that I did not sample quite enough genomes to observe this evidence and that there were some blind spots in my searches, this result heavily implies that eccDNA reinsertion is not a common phenomenon in *M. oryzae.* Routine DNA transformation protocols have been used in *M. oryzae* to integrate exogenous plasmids into the fungus' genome [121]. Therefore, the fact that I did not observe evidence of eccDNA reinsertion in *M. oryzae* came as a bit of a surprise. It is possible that some mechanism exists through which *M. oryzae* can quickly purge eccDNAs from its cells, perhaps similar to that employed by ageing yeast cells, that prevents the reintegration of eccDNAs into its genome [38]. It is also possible that *M. oryzae* does not produce enough eccDNAs for reinsertions to be observable in the time frame that I sampled.

Regardless, since many of the potential roles for eccDNAs in effector evolution involve reinsertion, this result indicates that it is unlikely that eccDNAs have a large impact on *M. oryzae*'s adaptation to crops. The one eccDNA-mediated genomic diversity generation process that does not involve reinsertion is gene deletion. While I did find that many eccDNA-associated genes experience PAV, I did not find any overlap between genomic deletions in *M. oryzae* and eccDNA formation. Again, to conclusively confirm this negative result, many more experiments are needed. For example, tandem eccDNA and DNA sequencing, especially at single-cell resolution, could directly confirm whether an eccDNA in one cell corresponds to a deletion in the same cell. Unfortunately, this technology simply does not exist today and would be extremely challenging to implement.

Ultimately, my work in this chapter resulted in many interesting discoveries about eccDNA biology in *M. oryzae* and it laid the groundwork for many follow up experiments that would help further explain these discoveries. However, many questions remain unanswered about eccDNA biology in general, and many tools still need to be developed to properly answer them. It is unlikely that *M. oryzae* is the best model to answer these questions and develop these tools. Models like *S. cerevisiae* and human cancer cells are, of course, far more appropriate. Furthermore, as described in this chapter and Chapter 2, several of the analyses I performed indicated that it is unlikely that eccDNAs play a major role in the evolution of the rice blast fungus. Given that my priority was to improve our understanding of genome evolution in *M. oryzae*, I chose to turn my focus away from eccDNAs in Chapter 4 of my dissertation.

Through presenting and publishing the work I described in Chapter 2, I realized that while extensive effector PAV had been reported many times in *M. oryzae* [25–27], the rules that shaped these events had not been clearly defined. As I realized during the GSA's 2022 Fungal Genetics Conference in Asilomar, CA, the *M. oryzae* genomics community was instead focused on population genetics, effector evolution and mini-chromosome biology. Given the fact that PAV appeared to play an important role in *M. oryzae*'s evolution, it seemed to me that understanding the patterns shaping these PAV events could help us tackle the threat *M. oryzae* poses to global food security. Specifically, I wanted to see if a deeper understanding of these patterns could help us predict PAV events in the future and therefore engineer better disease resistant crops. To study these patterns, I wanted to look at PAV in all genes, as I believed that focusing too much on effectors could lead me to miss some interesting biology. Additionally, I was interested in characterizing PAV in the emerging wheat-infecting strains of *M. oryzae* as this had not been done before. In a project not described in this dissertation, Anne Nakamoto and I characterized the TE content of many pathotypes of *M. oryzae* and found that rice-infecting strains harbored much greater TE content than wheat-infecting ones [122]. This observation made me particularly interested in whether the patterns of PAV might be different between the two pathotypes. All these ideas lead to the work I present in Chapter 4, where I characterized PAV events in rice and wheat-infecting *M. oryzae* and built machine learning models that could accurately predict these events.

Chapter 4

Characterization of gene presence-absence variation in *Magnaporthe oryzae*

The contents of this chapter are based on the following preprint:

Joubert, P.M., Krasileva, K.V. Distinct genomic contexts predict gene presence-absence variation in different pathotypes of a fungal plant pathogen. *bioRxiv* (2023). https://doi.org/10.1101/2023.02.17.529015

Abstract

Background

Fungi use the accessory segments of their pan-genomes to adapt to their environments. While gene presence-absence variation (PAV) contributes to shaping these accessory gene reservoirs, whether these events happen in specific genomic contexts remains unclear. Additionally, since pan-genome studies often group together all members of the same species, it is uncertain whether genomic or epigenomic features shaping pan-genome evolution are consistent across populations within the same species. Fungal plant pathogens are useful models for answering these questions because members of the same species often infect distinct hosts, and they frequently rely on gene PAV to adapt to these hosts.

Results

We analyzed gene PAV in the rice and wheat blast fungus, *Magnaporthe oryzae*, and found that PAV of disease-causing effectors, antibiotic production, and non-self-recognition genes may drive the adaptation of the fungus to its environment. We then analyzed genomic and epigenomic features and data from available datasets for patterns that might help explain these PAV events. We observed that proximity to transposable elements (TEs), gene GC content, gene length, expression level in the host, and histone H3K27me3 marks were different between PAV genes and conserved genes, among other features. We used these features to construct a random forest classifier that was able to predict whether a gene is likely to experience PAV with high precision (86.06%) and recall (92.88%) in rice-infecting *M. oryzae*. Finally, we found that PAV genes in the wheat- and rice-infecting pathotypes of *M. oryzae* differed in their number and their genomic context.

Conclusions

Our results suggest that genomic and epigenomic features of gene PAV can be used to better understand and even predict fungal pan-genome evolution. We also show that substantial intra-species variation can exist in these features.

Background

Microbial species have expansive pan-genomes that allow them to adapt to their environments. While bacteria typically gain and lose genes in the form of large horizontal gene transfer events [123], the accessory portion of fungal pan-genomes, which is defined in contrast to the conserved set of genes found in all members of a species, are typically shaped by small gene duplication and deletion events, which contribute to gene presence-absence variation (PAV) [124]. Previous fungal pan-genome studies have focused on the roles and functions of core and accessory genes [124–127], but our knowledge of which genomic and epigenomic features shape fungal pangenomes remains limited. Some studies have highlighted an association of accessory genes with subterminal chromosomal regions and transposable elements (TEs) [124,125], but it is uncertain whether these associations are strong enough to be predictive of gene PAV. While one study constructed models that could successfully predict meiotically derived structural variation generation events, and identified TEs, histone marks and GC content as particularly important predictors, it did not expand these findings to pan-genome evolution [128]. Finally, it is unclear whether any patterns in genomic or epigenomic features of PAV events would be generalizable to all populations of the same species, as pan-genomes are typically assembled for entire species without consideration of differential evolution between populations.

Fungal plant pathogens are ideal candidates to study pan-genome evolution, and especially gene PAV. They have dynamic pan-genomes that allow them to adapt to their hosts [125–127]. Specifically, fungal plant pathogens secrete a wide range of rapidly evolving effector proteins to cause disease. These effectors can become a disadvantage, however, when the immune receptors of their hosts acquire new binding specificities that detect these effectors and trigger an immune response [9]. Gene PAV is therefore particularly important in fungal plant pathogen evolution [10]. In these fungi, rapid genome evolution, especially of effectors, tends to occur in TE-dense and gene-poor regions of the genome while slower evolution and house-keeping genes occur in TE-poor and gene-dense regions of the genome [7,11]. This idea is often referred to as the “two-speed” genome concept. While effectors are particularly prone to PAV, it is unclear whether this concept extends to gene PAV and especially PAV of non-effectors. Many fungal plant pathogens of the same species also infect distinct hosts, which could facilitate the characterization and comparison of these PAV events in isolated populations.

Magnaporthe oryzae causes the blast disease of rice and wheat and is amongst the most important and well-studied pathogens with hundreds of available genomes and next-generation sequencing datasets [15,16]. The fungus has been reported to experience substantial gene PAV but these analyses have been largely restricted to effectors, and the genomic and epigenomic features associated with these PAV events remain largely unexplored [25–27]. *M. oryzae* reproduces mostly clonally, which makes the study of how its pan-genome can evolve without substantial recombination possible [26,28,118]. Finally, the blast fungus infects several different hosts, enabling the comparison of gene PAV between pathotypes within the same species [28]. While rice blast has been a long-standing threat, the rapid spread of wheat blast throughout the world as well as its particularly devastating effect on wheat crops has strongly encouraged research into this pathotype of *M. oryzae* and especially how it was able to jump hosts from rice to wheat and become such a devastating pathogen [16]. Altogether, *M. oryzae* offers a unique opportunity for studying gene PAV and the genomic and epigenomic features that shape these events as well as how these events vary within a species.

In this study, we sought to characterize and compare gene PAV in rice-infecting (MoO) and wheat-infecting *M. oryzae* (MoT). We first identified orthogroups experiencing PAV that distinguished isolated MoO lineages and found that they were enriched in effectors, as well as functions related to antibiotic production, and non-self-recognition. Next, we characterized the genomic contexts in which all gene PAV occurs in MoO and MoT and found that TEs were often found in proximity to these genes. Additionally, we found that gene length, GC content, expression, and histone H3k27me3 marks were distinct for PAV genes. We used these features to construct a random forest classifier and found that the differences we observed were strong enough to produce a model that predicted whether a gene is likely to experience PAV with high precision (86.06%) and recall (92.88%). Finally, we found significant differences in the number of PAV events and the features that predict PAV in MoO and MoT, which could reflect their differing evolutionary history and could be evidence of distinct mechanisms contributing to gene loss in the two recently diverged lineages.

Results

Genes associated with pathogenicity, non-self-recognition and antibiotic production, are enriched among orthogroups experiencing lineage-differentiating presence-absence variation in *M. oryzae*

Differences in gene PAV events between isolated lineages of *M. oryzae* could be evidence of local adaptation. MoO isolates can be grouped into four lineages, called lineages 1, 2, 3, and 4 [28]. Lineages 2, 3, and 4 are monophyletic within the MoO phylogeny and propagate clonally [28]. All lineages show evidence of local adaptation [26]. To generate a table of all gene PAV events in MoO, we analyzed 123 previously published genomes [26,129,130]. These genomes were re-annotated, and the proteomes were clustered into orthogroups. This enabled us to identify putative gene absences in all genomes. These were then validated by using TBLASTN [74] against the genome and comparing hits to the missing orthogroup using BLASTP[74]. This approach helped ensure that gene absences were not annotation errors. We also constructed a phylogeny from a multiple sequence alignment of all of our single copy orthologs (SCOs) and found each of the three clonal MoO lineages formed separate monophyletic groups in our data, as previously observed (Additional File 1: Fig. S1) [26,28].

To identify whether differences in gene PAV events existed between the three clonal lineages of MoO, we performed a principal components analysis (PCA) on our table of PAV events. We found that the top 2 principal components (PCs) of our PCA clearly separated the lineages demonstrating that different PAV events had occurred in each lineage since their separation (Fig. 1A). Next, we identified 587 orthogroups that represented 70.53% and 62.17% of the variance in PCs1 and 2, respectively, and labeled these orthogroups as experiencing lineage-differentiating PAV. We then identified, among all orthogroups, 594 putative effector orthogroups and found, as previously reported [25,26], that PAV of effector orthogroups alone were sufficient to separate the MoO lineages in a follow-up PCA (Fig. 1B). Given that we identified 4.30% of all orthogroups as putative effectors, the fact that 8.67% of lineage-differentiating PAV orthogroups were effectors represented a clear enrichment. However, non-effector orthogroups still represented 91.33% of lineage-differentiating PAV orthogroups, showing that many orthogroups besides effectors experience lineage-differentiating PAV (Fig. 1C).

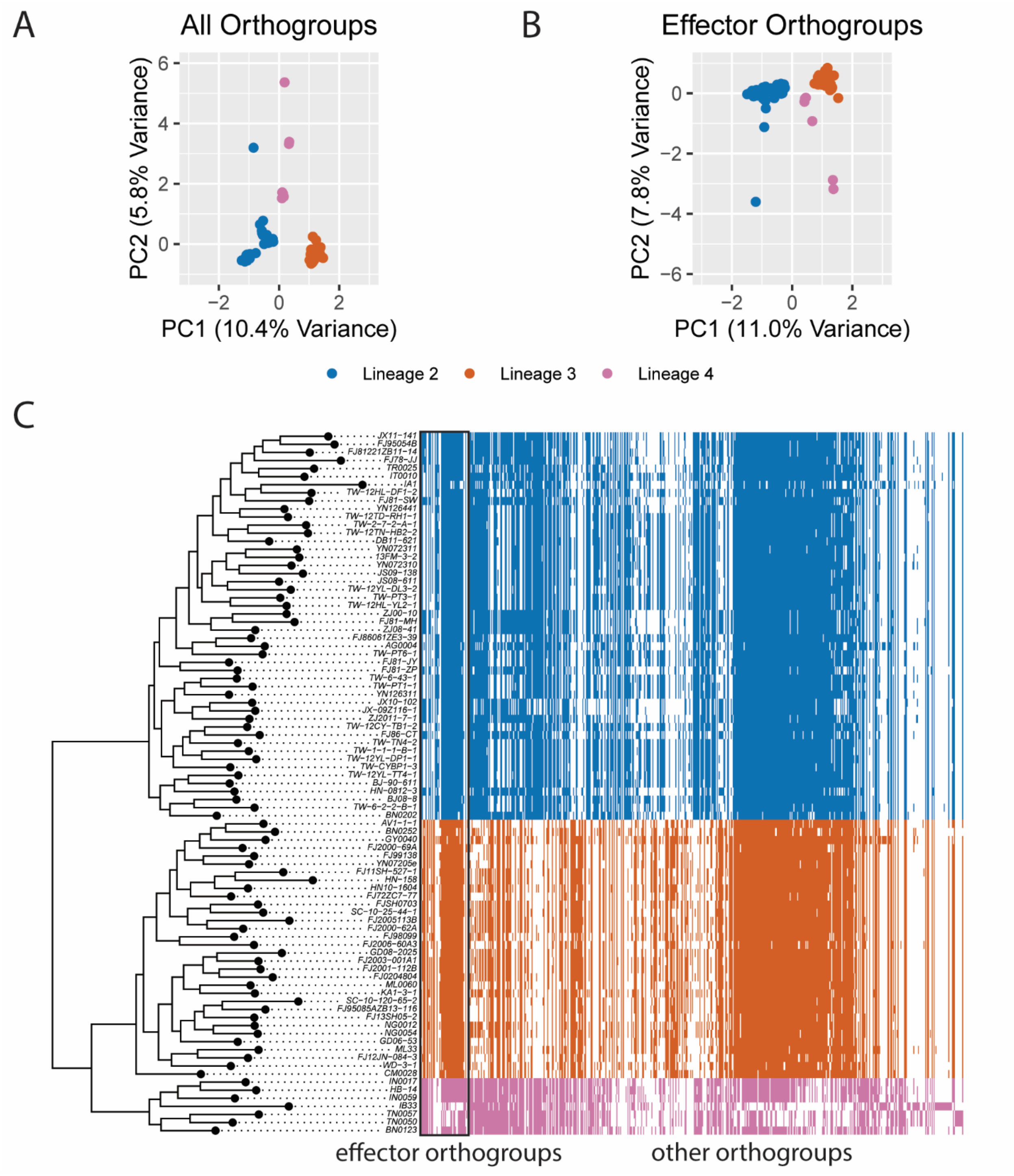


Fig. 1. PAV of effector and non-effector orthogroups differentiate the clonal lineages of rice-infecting *M. oryzae*. A. Scatter plot of values for principal components (PCs) 1 and 2 resulting from a PCA of orthogroup PAV. Each point represents one isolate. B. Scatter plot of values for PCs 1 and 2 resulting from a PCA of effector orthogroup PAV. Each point represents one isolate. C. Heat map representing which orthogroups are present (color) or absent (white) in each genome. Effector orthogroups are separated from other orthogroups by a black box. The

phylogeny was generated using a multiple-sequence alignment of SCOs and fasttree and is a subset of the full MoO phylogeny generated from our data (Additional File 1: Fig. S1). In all panels, colors represent the clonal lineages of MoO. Blue represents lineage 2, orange represents lineage 3 and pink represents lineage 4. Lineages were named as previously described [28].

To identify what other types of genes were enriched amongst lineage-differentiating PAV orthogroups, we performed gene ontology (GO) and protein family (PFAM) enrichment analysis. This analysis revealed that lineage-differentiating PAV orthogroups were enriched for GO terms related to secondary metabolite production and biosynthesis of membrane components, among other terms (Fig. 2A). Lineage-differentiating PAV orthogroups were enriched for PFAM domains related to antibiotic production, among other domains (Fig. 2B). Genes without PFAM domains were also strongly enriched in PAV orthogroups (6040 annotated, 407 observed, 256.55 expected, p-value < 0.001, Fisher's exact test). Notably, the HET domain, which is associated with heterokaryon incompatibility in fungi, was also enriched among these orthogroups (Fig. 2B). Finally, while NACHT and NB-ARC domains did not appear enriched on their own due to a small number of lineage-differentiating PAV orthogroups having these annotations, NOD-like receptors (NLRs) which may play a import role in fungal immunity and contain either a NACHT or an NB-ARC domain [131], were enriched amongst lineage-differentiating PAV orthogroups (23 annotated, 4 observed, 1.15 expected, p-value = 0.026, Fisher's exact test). These results indicated that antibiotic production and non-host recognition, in addition to effectors, may play an important role in driving adaptation in these three isolated lineages of rice-infecting *M. oryzae*.

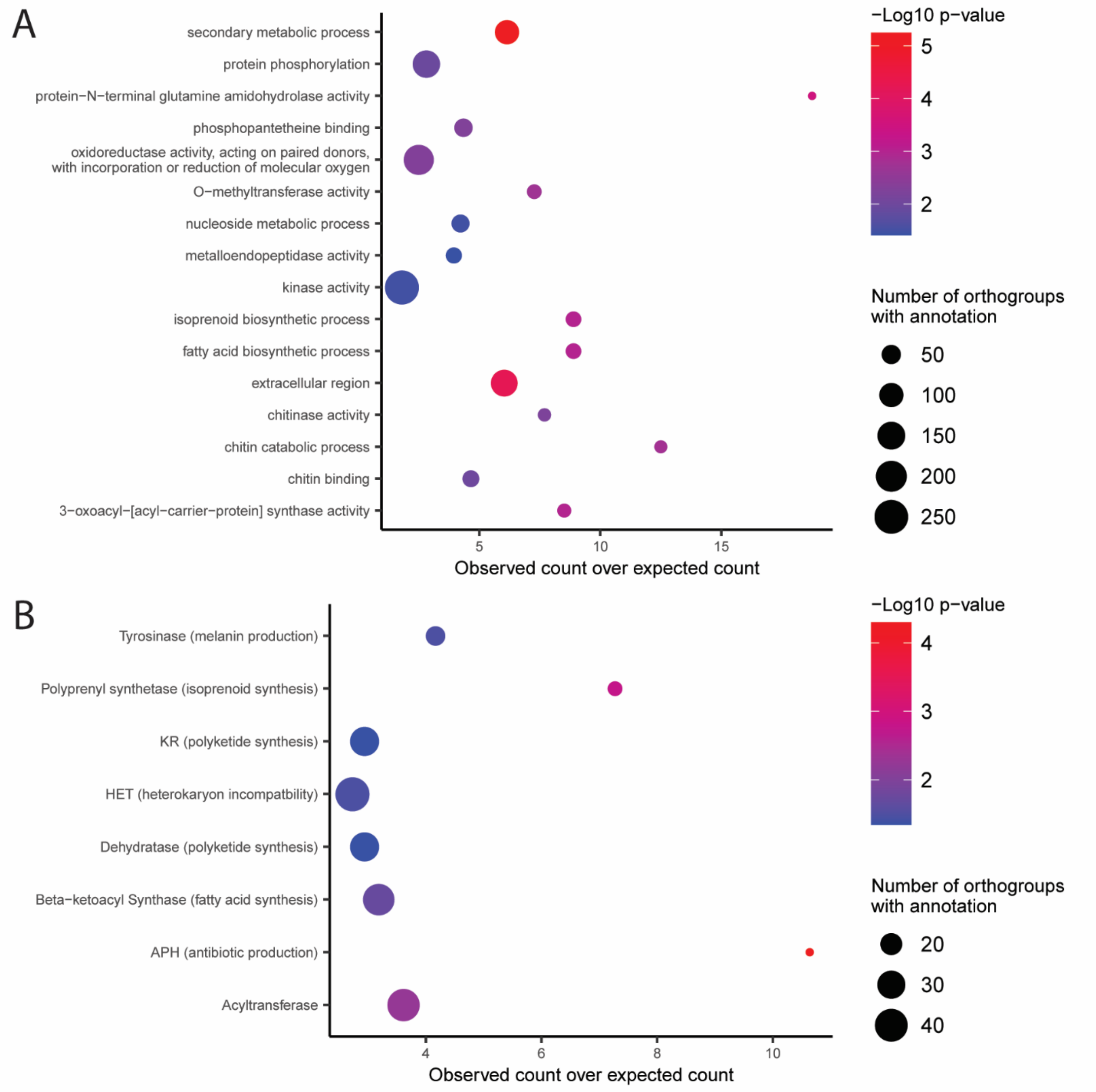


Fig. 2. Lineage-differentiating PAV orthogroups in rice-infecting *M. oryzae* contain many genes related to antibiotic production and non-self-recognition. A. Gene ontology (GO) enrichment analysis of lineage-differentiating PAV orthogroups. B. Protein family (PFAM) domain enrichment analysis of lineage-differentiating PAV orthogroups. P-values shown are the results of Fisher's exact tests.

Presence-absence variation genes are more common and more spread out throughout the genome in wheat-infecting *M. oryzae* than in rice-infecting *M. oryzae*

We next sought to identify whether there were specific patterns in the genomic contexts of PAV events in *M. oryzae*. To expand our analyses beyond lineage-differentiating PAV orthogroups and to compare PAV orthogroups to conserved orthogroups, we first needed a systematic way to label them. To avoid erroneously calling single gene gain or loss events as PAV, we chose to incorporate phylogenetic information in these definitions and therefore identified PAV and conserved orthogroups for each clonal, monophyletic lineage of *M. oryzae*. In our data, orthogroups were labeled as PAV if they were present in all isolates of at least two subclades within a lineage and absent in all isolates of at least two subclades within a lineage. This definition meant that at least two phylogenetically independent loss or gain events needed to be observed in our data for an orthogroup to be labeled PAV. All orthogroups that were present in all but two or fewer isolates in a lineage were labeled as conserved orthogroups. All orthogroups that did not fit this definition were labeled as “other”. Genes belonging to PAV orthogroups or conserved orthogroups were labeled PAV genes and conserved genes, respectively. This approach allowed us to label 1,269 and 1,029 PAV orthogroups in lineage 2 and 3 of MoO, respectively (Fig. 3A). We did not include lineage 4 in our analysis because of its small sample size and omitted lineage 1 because it is thought to be recombining and would therefore would have violated the assumptions of our definition of PAV and conserved orthogroups [28].

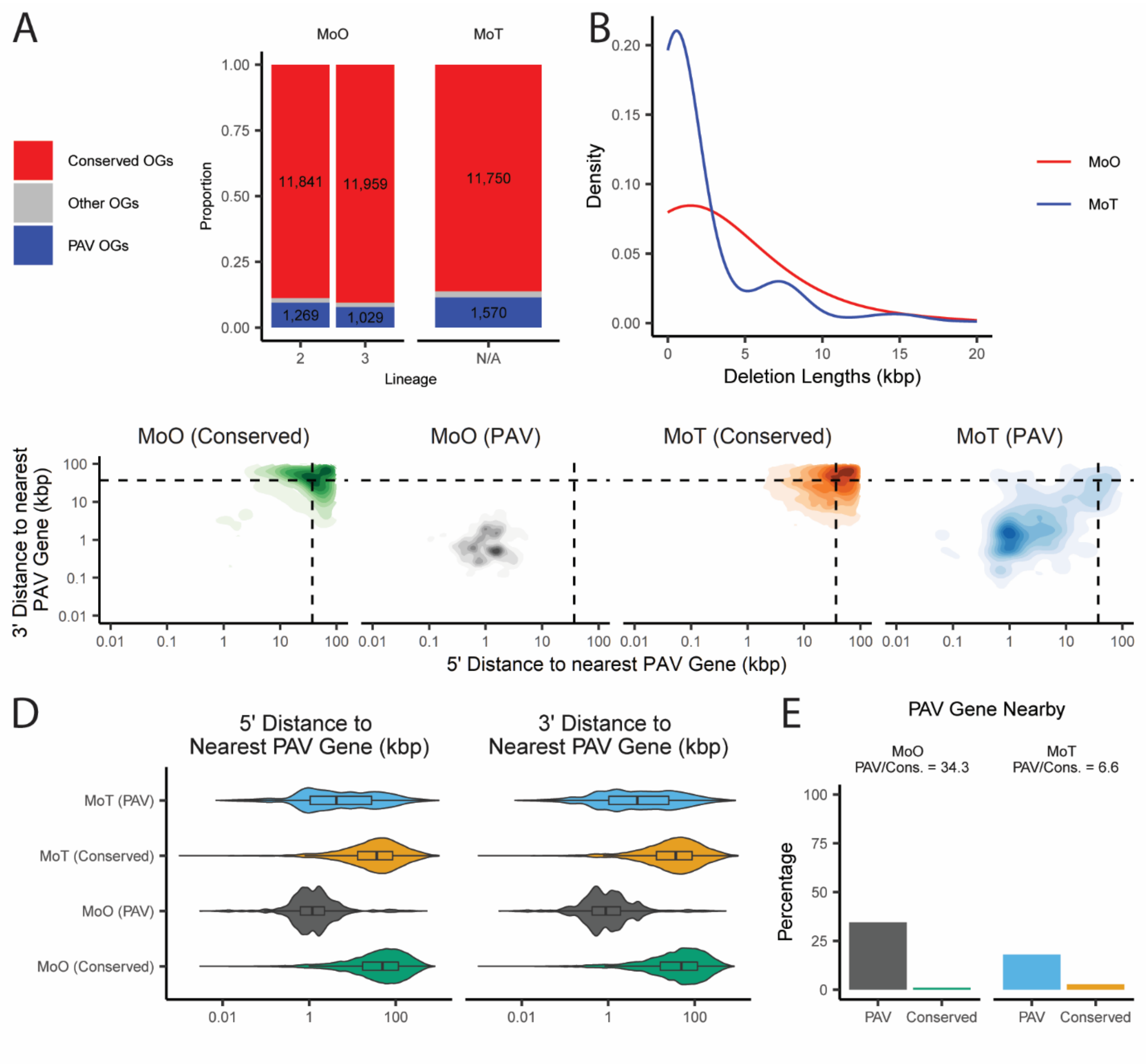


Fig. 3. PAV genes are more common and more spread out throughout the genome in wheat-infecting *M. oryzae* than in rice-infecting *M. oryzae*. A. Stacked barplot comparing the number of PAV orthogroups (OGs) and conserved orthogroups in MoO and MoT. "Other OGs" denote orthogroups that did not satisfy our definitions for either category. B. Distribution of the lengths of genomic deletions in MoO and MoT. C. Density plot showing the distribution of the distances to the nearest PAV gene for conserved and PAV genes in MoO and MoT. Dashed lines in density plots represent the median values for all genes in both pathotypes. D. Violin plot showing the distribution of the distances to the nearest PAV gene for conserved and PAV genes in MoO and MoT. E. Percentages and proportions of PAV and conserved genes that are within 1000bp of a PAV gene in MoO and MoT. Median values and statistical comparisons for data

shown in panels C through E are listed in Additional File 7, Additional File 8, and Additional File 9.

To compare PAV across rice and wheat-infecting *M. oryzae* isolates, we annotated, called orthogroups and validated missing orthologs for 36 previously published MoT genomes (Additional File 2). Unlike for MoO, MoT isolates have not been formally assigned into lineages in the past, though they are thought to have propagated primarily clonally since their recent appearance [118]. Given the small number of MoT genomes we used in our analysis, we chose not to separate them into different lineages (Additional File 1: Fig. S2). In these isolates, we identified substantially more PAV orthogroups than in the MoO lineages (Fig. 3A). To assess whether this contrast was reflected in genomic deletions, we used 117 MoO and 47 MoT Illumina whole-genome sequencing datasets to call genomic deletions based on a high-quality reference genome for each pathotype (Additional File 3 and Additional File 4). This approach allowed us to identify 1,870 deletions in MoO and 1,862 deletions in MoT despite using more than double the number of datasets for MoO than MoT (Additional File 5 and Additional File 6). We also found that genomic deletions were larger in MoO than in MoT, with a median of 1,818bp in MoO and 960bp in MoT (Fig. 3B). Correspondingly, when we compared the density of PAV genes in MoO and MoT, we found that genes belonging to PAV orthogroups were much more likely to be in proximity with other genes belonging to PAV orthogroups in MoO than in MoT (Fig. 3C-E, Additional File 7, Additional File 8, and Additional File 9). Taken together, these results indicated that genomic deletions, especially those involving genes, were more likely to involve multiple genes in MoO than in MoT. These results also hinted that gene PAV happens in defined regions of the genome in MoO while in MoT these events are more likely to be randomly spread out throughout the genome.

Genes prone to presence-absence variation are closer to transposable elements than other genes

The two-speed genome hypothesis defines two genomic compartments in fungal plant pathogens, one characterized by rapid evolution, few genes and many TEs, and the other characterized by slow evolution, many genes and few TEs [7,11]. We investigated whether orthogroups experiencing PAV followed this model in *M. oryzae*. We found that genes in PAV orthogroups were much closer to TEs than genes in conserved orthogroups in both MoO and MoT (Fig. 4, Additional File 7 and Additional File 8). While the differences in distance to the nearest gene between conserved and PAV orthogroups in MoO or MoT were typically quite small (median difference <100bp), we did find that genes in PAV orthogroups were less likely to be close to genes than conserved genes, though the effect was not as strong as for TEs (Additional File 1: Fig. S3, Additional File 7, and Additional File 8). We also observed differences in these patterns for MoO and MoT. Specifically, we found that PAV orthogroups in MoO were more likely to be close to TEs than those in MoT (Fig. 4C and Additional File 9). We also found that MoO PAV genes were more likely to be far away from genes that MoT PAV genes (Additional File 1: Fig. S3C and Additional File 9).

To understand if these observations also applied to genomic deletions in MoT and MoO, we measured TE and gene densities within the genomic deletions we previously identified and within their flanking regions. This analysis revealed that genomic deletions and their flanking regions were enriched in TEs and depleted in genes, though the effect was stronger for TE density than for gene density (Additional File 1: Fig. S4).

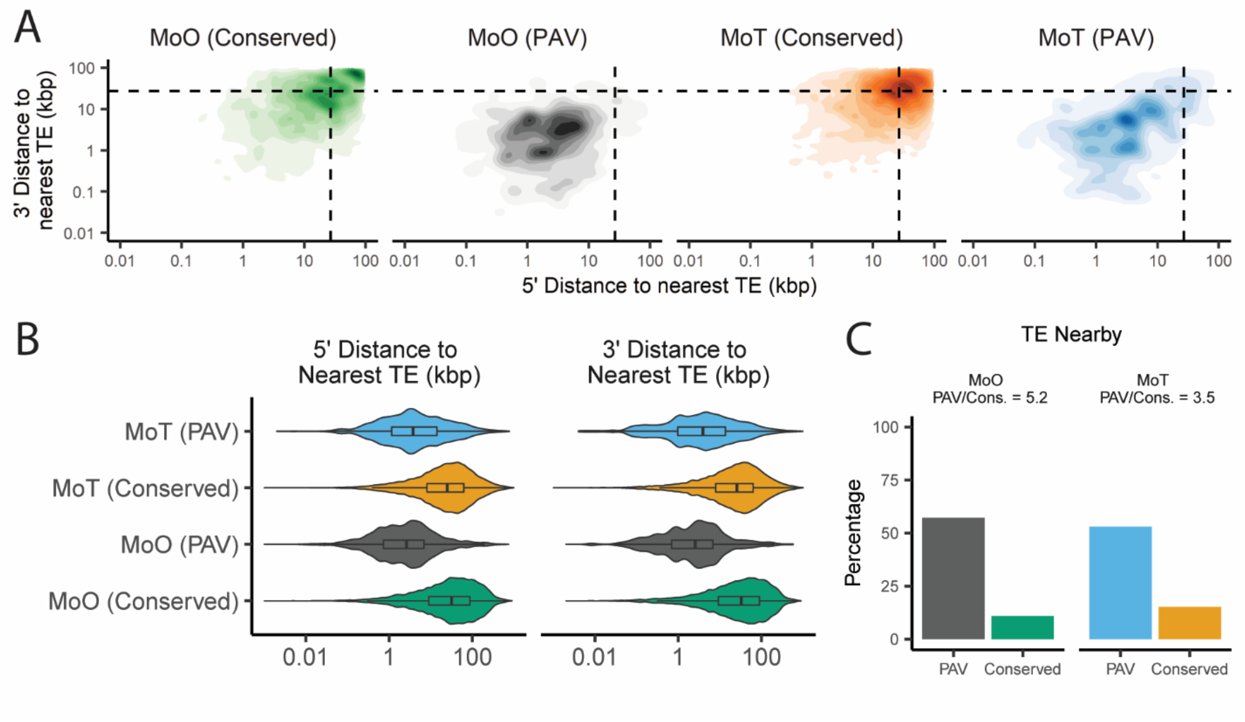


Fig. 4. PAV genes are more likely to be found near transposable elements (TEs) than conserved genes. A. Density plots showing the distribution of the distances to the nearest TE for conserved and PAV genes in MoO and MoT. B. Violin plot showing the distribution of the distances to the nearest TE for conserved and PAV genes in MoO and MoT. C. Percentages and proportions of PAV and conserved genes that are within 5000bp of a TE in MoO and MoT. Dashed lines in density plots represent the median values for all genes in both pathotypes. Median values and statistical comparisons for data shown are listed in Additional File 7, Additional File 8, and Additional File 9.

Genes prone to presence-absence variation show distinct genomic and epigenomic features and some differences in these features exist between rice and wheat-infecting *M. oryzae*

A previous report has shown that MoO has a much greater TE content than MoT [122]. Therefore, given this fact and the increased number of PAV orthogroups in MoT compared to MoO we observed (Fig. 3A), it is unlikely that TEs alone define whether a gene is prone to PAV or not. We therefore chose to investigate whether we could identify other differences in genomic features between PAV genes and conserved genes in *M. oryzae*. We first looked at the GC content of these genes and the regions that flank them. PAV genes were more likely to have lower GC content than conserved genes (Fig. 5A and Additional File 10), as did the regions that flank them, though the effect was more subtle for the flanking regions (Additional File 1: Fig. S5A and Additional File 10). We also found that PAV genes were shorter than conserved genes (Fig. 5B and Additional File 10). We next performed various functional annotations of PAV and conserved genes and found that PAV genes were more likely to be predicted effectors, in accordance with our previous results, and less likely to have GO or PFAM annotations than conserved genes (Additional File 1: Fig. S6 and Additional File 8).

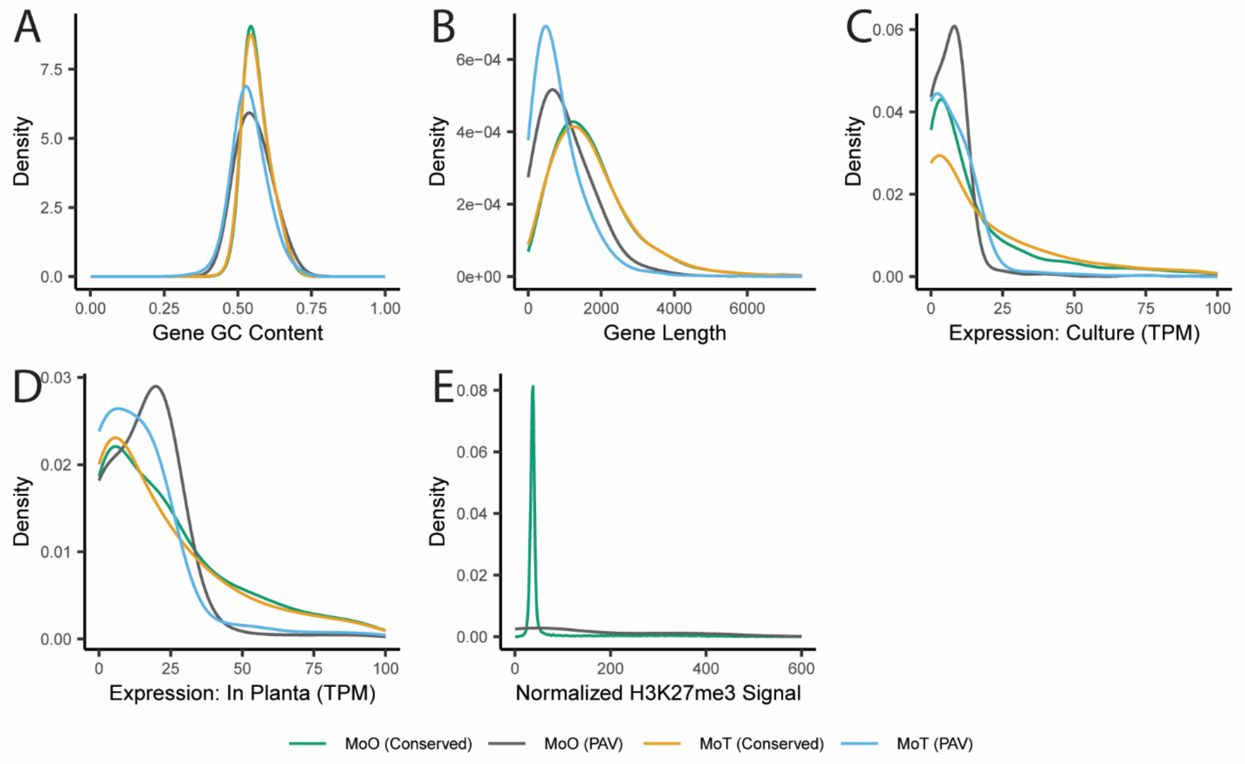


Fig. 5. PAV genes are distinct from conserved genes in many ways beyond their proximity to TEs. Density plots showing the distributions of A. gene GC content, B. gene lengths, C. expression in culture, D. expression *in planta*, and E. normalized H3K27me3 histone mark ChIP-Seq signal for PAV and conserved genes in MoO and MoT. In panel E, MoT genes were not included as this data is not available for MoT. Statistics describing the distributions shown and statistical comparisons between these statistics are listed in Additional File 8, Additional File 9, Additional File 10, and Additional File 11.

Next, we gathered histone mark, transcription, methylation, and extrachromosomal circular DNA sequencing (eccDNA) data from the literature for both MoO and MoT to further characterize PAV genes. Unfortunately, these datasets were only available for some strains of MoO or MoT but not for all. Therefore, we analyzed these datasets for one reference MoO strain and one reference MoT strain, and then generalized this signal to our orthogroups to impute the signal in other strains of *M. oryzae*. This allowed us to observe that average expression was higher both in culture and *in planta* for conserved genes than for PAV genes (Fig. 5C and D and Additional File 10). Additionally, PAV genes were more likely to show signal from chromatin immunoprecipitation sequencing of H3K27me3 and H3K36me3 histone marks and less likely to show signal from H3K27ac histone marks (Fig. 5E, Additional File 1: Fig. S5B and C, and Additional File 10). We also looked at bisulfite sequencing data and found that PAV genes were less methylated and showed a greater variation in methylation percentage than conserved genes (Additional File 1: Fig. S5D and Additional File 10). Finally, we found that PAV genes had a much tighter distribution of eccDNA sequencing signal than conserved genes (Additional File 1: Fig. S5E and Additional File 10). Overall, these results indicated clear differences in the genomic and epigenomic features of PAV genes compared to conserved genes.

Many differences between PAV genes in MoO and MoT were discovered through this process including in gene length, where PAV genes were smaller in MoT than MoO (Fig. 5B and Additional File 11), and in expression, where PAV genes in MoT showed less expression on average than MoO PAV genes both in culture and *in planta* (Fig. 5C and D and Additional File 11). Additionally, PAV genes were more likely to have GO and PFAM annotations in MoO than in MoT (Additional File 1: Fig. S6E and F and Additional File 9). These observations further supported the idea that PAV may be occurring in different genomic contexts in MoO and MoT.

Finally, we analyzed a similar set of features in the genomic deletions we identified in MoT and MoO. We found that while some of the trends we observed in PAV genes were similar in genomic deletions compared to a genomic baseline, such as H3K27me3 signal and GC content, other trends like the increase in expression of PAV genes did not translate to differences in RNAseq signal in deleted regions (Additional File 1: Fig. S7 and Additional File 12).

Genomic and epigenomic features of genes prone to presence-absence variation can be used to generate predictive models for rice and wheat-infecting *M. oryzae*

Our previous results demonstrated the differences in genomic contexts between PAV genes and conserved genes. We then wanted to determine whether these features in aggregate could provide enough signal to predict whether a gene was prone to PAV using a machine learning approach. To this end, we trained a random forest classifier on all features we described for MoO. We selected this algorithm because of its ease of implementation as well as its robustness to correlated features [132]. When we trained this model on data from all but 8 strains of MoO and tested the model on the remaining strains, we observed that the model performed very well and was able to predict PAV genes with 86.06% precision and 92.88% recall on average (F1 = 89.34%, Fig. 6A). Our model also allowed us to determine how important each feature was in predicting PAV genes by calculating the decrease in the F1 statistic when the variable in our testing data was permuted. This approach identified histone H3K27me3 as being the most predictive feature for PAV genes in MoO (Fig. 6B). Although the accuracy of predictions by the random forest classifier is robust to correlated features, the variable importances we observed were likely influenced by the fact that several variables in our model were correlated with each other and that many showed high dependences, which meant that the information encoded in these variables could also be described by other variables in the model (Additional File 1: Fig. S8 and Fig. S9). These importances should therefore be interpreted with caution. Next, we trained a model to predict PAV genes in MoT and found that the model performed even better with a precision of 94.81% and a recall of 96.43% (F1 = 95.61%, Additional File 1: Fig. S10A). In this reduced model, gene expression *in planta* stood out as being particularly predictive of MoT PAV genes (Fig. 6C). Finally, we trained another MoO model using a reduced set of features that matched the data that was available for MoT, and found that the MoO model still performed well with an 86.11% precision and 92.21% recall (F1 = 89.05%, Additional File 1: Fig. S10B). The similar performance of the two

MoO models could likely be explained by the high dependences of our variables (Additional File 1: Fig. S8 and Fig. S9). When comparing the reduced MoO model to the MoT model, we noticed some differences between the importances of the features in each model (Fig. 6C and D). For example, in culture expression and the presence of functional annotations was more important in the reduced MoO model than in the MoT model. These differences in importances may have been influenced by the previously described differences in the features of PAV genes in MoO and MoT.

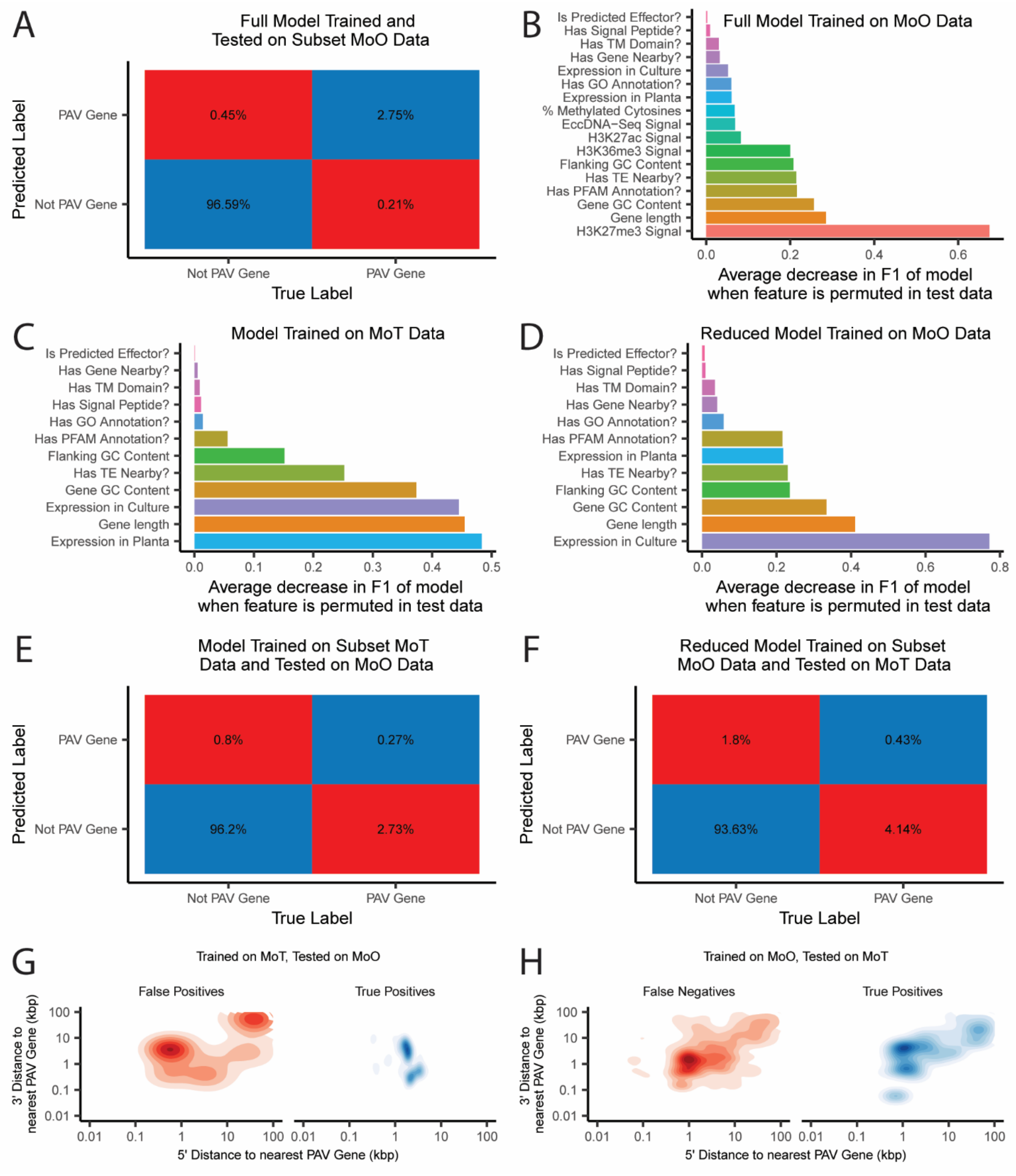


Fig. 6. Random forest classifiers accurately identify PAV genes in rice and wheat-infecting *M. oryzae*, but the models perform poorly on genes from the host they were not trained on. A. Confusion matrix showing predictions of the MoO random forest classifier when tested on MoO genes that it was not trained on. B. Decrease in the F1 statistic of the MoO random forest classifier when each feature is permuted in the testing data. Features described as questions are binary, all other features are continuous. C. Decrease in the F1 statistic of the MoT random

forest classifier when each feature is permuted in the testing data. D. Decrease in the F1 statistic of the MoO random forest classifier trained on a subset of features when each variable is permuted in the testing data. F. Confusion matrix showing predictions of the MoT random forest classifier when tested on MoO genes. E. Confusion matrix showing predictions of the MoO random forest classifier trained on a reduced subset of features when tested on MoT genes. G. Density plots showing the distribution of the distances to the nearest PAV gene for false positive and true positive predictions by the MoT random forest classifier when tested on MoO genes. H. Density plots showing the distribution of the distances to the nearest PAV gene for false negative and true positive predictions by the MoO random forest classifier trained on a subset of features when tested on MoT genes.

A predictive model trained on wheat-infecting *M. oryzae* data does not accurately predict presence-absence variation in rice-infecting *M. oryzae* and vice versa, highlighting differences in the genomic contexts of presence-absence variation in the two pathotypes

Finally, we tested if the model trained on MoT data could predict whether genes are prone to PAV in MoO and vice versa. The MoT model performed very poorly on MoO data, with a precision of 25.40% and a recall of 9.06% (F1 = 13.35%,Fig. 6E). Similarly, the reduced MoO model performed very poorly on the MoT data with a precision of 19.30% and a recall of 9.41% (F1 = 12.65%, Fig. 6F). This result could be explained by a variety of factors including differences in genomic features between the two pathotypes, differences in the importances of each feature in the model, and overfitting. When we analyzed the conserved genes that the MoT model falsely labeled as PAV, we found that many of them were found in isolated regions far away from true PAV genes (Fig. 6G). Similarly, many of the PAV genes in MoT that were not detected by the MoO model were found in isolated regions (Fig. 6H). These results followed the patterns we observed in PAV clusters in MoO and MoT genomes (Fig. 3A). Our observations, combined with the differences we observed in the genomic and epigenomic features of PAV genes in MoO and MoT described previously, indicated that the patterns and genomic contexts of PAV between the two pathotypes are significantly different, despite being within the same species.

Discussion

Gene PAV plays an important role in fungal pan-genome evolution [124–127]. To improve our understanding of these events, we designed a robust pipeline to identify orthogroups experiencing PAV in *M. oryzae*. We found that PAV of these orthogroups differentiates isolated lineages of MoO, and found that these lineage-differentiating PAV orthogroups are enriched for effectors, as previously published [25,26]. We also found that genes related to antibiotic production and non-self-recognition were also enriched among them. This result could point to the local and rice-associated microbiome playing an important role in *M. oryzae*'s evolution. All three clonal lineages are geographically isolated and experience different climates [26]. They also tend to infect different rice varieties and cause disease of varying severity [26]. Geography and host genotype could have major influences on the microbiome the fungus encounters. Microbiome sampling of rice varieties used in these areas as well as the environment could therefore give better insight into the results we present here and how the microbiome might shape the fungus' fitness. Additionally, it is important to note that adaptation to the host microbiome and the environment in general are often forgotten when discussing fungal plant pathogen evolution. Our results point to the importance of considering these factors when studying the success of these pathogens. Unfortunately, we could not extend these analyses to MoT as lineages of MoT have not been formally defined in the past and neither has their geography or host phenotypes.

We then looked to find features of PAV orthogroups that might help us better understand where these events are occurring in the genome. We found that these events were associated with a high TE density and a low gene density, though the effect was stronger for TE density than gene density. We also found that PAV genes are shorter, have lower GC content, and are more likely to be effectors. Finally, PAV genes are less expressed and display stronger histone H3K27me3 signal than conserved genes. When we combined all of these features into a predictive model, we found that the model performed very well and predicted PAV genes with 86.06% precision and 92.88% recall, on average. We were also able to identify histone H3K27me3 as the most predictive feature, though gene length and GC content stood out as well. We could not clearly state whether genomic deletions showed similar features to PAV genes in our data. While we found that these genomic deletions occurred frequently in TE-dense and gene-sparse areas of the genome, and that GC content and H3K27me3 ChIP-Seq signal for these regions resembled that of PAV genes, other features were not similar between the two. These results may have been confounded by the need for reference-based identification of these deletions, an unclear baseline for comparison, and events like transposon insertion polymorphisms.

Many of the features that were particularly important in our classifier were related to the two-speed genome concept which supported the idea that gene PAV in *M. oryzae* is strongly associated with the rapidly evolving compartment of the genome [7,11]. Our findings support the idea that these features may play an important role in the evolution of the pathogen. However, the fact that the presence of TEs were important features in our random forest classifier but not amongst the most important, supports the idea that the correlation between TEs and rapid evolution is not always a causal one and that complex correlations are at play. In short, other variables may be shaping the PAV-prone compartment of the *M. oryzae* genome and driving both rapid evolution and TE activity. Our findings also reflect previous findings on the association between TEs and the evolution of the accessory portion of fungal pan-genomes [125]. While the gene space for the genomes we analyzed were well assembled, most of the genomes we performed our analysis on were not chromosome-level assemblies. Therefore, although we observed features associated with subterminal regions in our PAV genes, we could not confirm previous findings on the association between subterminal regions and accessory genes in other fungi [124]. This analysis should be repeated once more high-quality genomes become available for *M. oryzae* to fully determine whether these findings apply to the blast fungus as well. Similarly, though the features we identified in this study should be kept in mind when studying other pan-genomes, it is unclear whether the features of gene PAV we identified are applicable to other fungi, and therefore more in-depth studies of these genomic and epigenomic features are necessary to assess how broad these findings are as datasets become available for more fungi.

Though our random forest classifier performed well, many of the important features we used in our model had to be propagated from a single strain which likely led to many biologically inaccurate values in our data and potentially errors in how we ranked the importances of each variable. To truly validate our results, our approach would need to be repeated using more data for each isolate. However, even a model using a subset of our features performed well, indicating that RNAseq data for each isolate may be enough to further substantiate these results. Regardless, our model showed that PAV genes can be identified simply using features in the genome, therefore establishing a method to identify genes prone to PAV in *M. oryzae* without relying on phylogenetics. This could be useful for identifying genes prone to PAV in lineages of MoO with very few isolates, like lineage 4, or for studying PAV in groups of genes with complicated evolutionary relationships like sequence unrelated structurally similar (SUSS) effectors [133]. While our models performed well, they also identified many genes that had features of PAV genes but did not experience PAV. These false positives could help us better understand which genes are under strong selection to be kept in the *M. oryzae* genome or which genes' genomic contexts are changing to look more like conserved genes. Notably, our results support the exciting possibility of using genomics to predict targets for disease-prevention strategies that will remain in the genome, therefore making these strategies more robust.

Finally, we found distinct patterns in the genomic contexts of PAV genes in MoO and MoT. Specifically, we found that PAV in MoT appeared to occur more frequently and was more spread out throughout the genome than in MoO. Though we used a similar number of isolates from MoT for our analysis as we did for each lineage of MoO, the evolutionary distances between isolates in lineages of the two pathotypes were different, which may have contributed to the differences in the number of PAV orthogroups we observed (Additional File 1: Fig. S1 and Fig. S2). However, the stark difference in the number of PAV orthogroups, as well as supporting evidence from our analysis of genomic deletions, suggest that our observations are valuable despite this caveat. We also found that many of the genomic and epigenomic features of PAV that we identified in MoO were different in MoT. These differences may have explained why our MoO random forest classifier performed poorly on MoT data and vice versa, since the patterns in the false positives and false negatives of these tests reflected the observed differences in PAV between MoO and MoT. These results in aggregate indicated differences in the evolution of the rice and wheat pathotypes of *M. oryzae*.

The two *M. oryzae* pathotypes share some major differences in their TE content [122] and very different life histories, with MoO originating 9,800 thousand years ago [28] and propagating mostly clonally since then, while MoT is thought to have emerged approximately 60 years ago from a multi-hybrid swarm of many different *M. oryzae* pathotypes [16,118]. We propose that the differences in PAV across the two pathotypes may reflect these life histories, with MoO exhibiting more of a stable equilibrium and much slower paced evolution, where PAV events happen in specifically defined compartments of the genome, while MoT is rapidly losing and gaining genes, even in areas of the genome where most of the conserved genes in MoO are located. It is unclear at this point whether MoT is heading towards an equilibrium that will resemble MoO, or whether there are key differences between the two pathotypes that are shaping their genomes beyond their evolutionary histories. MoT, which appears to lose genes at a faster rate than MoO and evolve rapidly in general, will pose a significant challenge for disease prevention. A better understanding of these evolutionary dynamics and the differences between MoO and MoT could help us better comprehend why MoT is such a devastating emerging pathogen and help us curb its threat. Finally, these results highlight the need to study isolated populations of a species separately as well as in aggregate to understand whether observations made for the pan-genome applies to every population within a species, especially if they are adapted to different hosts or environments and if they have different evolutionary histories.

Conclusions

Our study demonstrates that gene PAV can be associated with specific genomic and epigenomic features in fungi and that these associations can be predictive. We also show that major variation can exist in these features between different populations of the same species. This study therefore highlights the need for more studies of fungal pan-genomes and the genomic and epigenomic features that define them to better understand how fungi adapt to their environments. These studies could also lead to a greater understanding of how fungal plant pathogens adapt to their hosts. Predicting these adaptations could help us develop more effective disease prevention strategies in the future. Finally, it is important that future pan-genome studies be done in a way that considers intra-species variation and evolutionary history of different populations to avoid generalizing based on a reference strain or pathotype.

Methods

Genome annotation, proteome orthogrouping, and phylogeny

The set of 123 MoO genomes were obtained from a previously published study [26,129,130], while 36 MoT genomes (Additional File 2) as well as a single *M. grisea* proteome (GCA004355905.1) were obtained from NCBI's GenBank. All genomes were verified to have more than 90% completeness using BUSCO version 5.2.2 and the "sordariomycetes_odb10" option [134]. Genomes were annotated using FunGAP [77] version 1.1.0 and RNAseq data obtained from Sequence Read Archive (SRA) accession ERR5875670. The "sordariomycetes_odb10" option was used for the busco_dataset argument and the "magnaporthe_grisea" option was used for the augustus_species argument. For repeat masking, a TE library generated by combining the RepBase [78] fngrep version 25.10 with a *de novo* repeat library, generated by RepeatModeler [79] version 2.0.1 run on the *M. oryzae* Guy11 genome (GCA002368485.1) with the LTRStruct option, was used for all genomes. Annotated proteomes were then used as input for OrthoFinder [96] version 2.5.4 to form two separate sets of orthogroups, one for MoO genomes and one for MoT genomes. The *M. grisea* proteome was included in both as an outgroup. Orthogrouping was performed using the "diamond_ultra_sens" parameter for sequence search, the "mafft" parameter for species tree generation and the "fasttree" parameter for gene tree generation. Single copy orthologs (SCOs) were then obtained from the OrthoFinder output, aligned using mafft [135] version 7.487 with the --maxiterate 1000 parameter and the --globalpair parameter, concatenated, and then trimmed using trimal [136] version 1.4.rev22 and a 0.8 gap threshold parameter. Finally, fasttree [137] version 2.1.10 with the gamma parameter was used to generate a phylogeny and ape [138] version 5.5 was used to root each tree on the *M. grisea* outgroup.

Gene absence validation

A preliminary set of missing orthogroups in each genome was obtained from the OrthoFinder outputs. Gene absences were validated by first using TBLASTN [74] version 2.7.1+ with the -max_intron_length 3000 parameter to align all protein sequences from an orthogroup to the genome that was missing that orthogroups. Any orthogroup that resulted in two or more hits above 55% sequence identity, 55% query coverage and an e-value smaller than 10^{-10} when aligned to the target genome were selected for further verification. These cutoffs were optimized so that less than 1% of orthogroups were misclassified as absent in a testing set of orthogroups that were known to be present in a target genome. Finally, TBLASTN hits were extracted as protein sequences using agat_sp_extract_sequences.pl version 0.9.1 from the AGAT toolkit (https://github.com/NBISweden/AGAT), and aligned against all protein sequences in all orthogroups using BLASTP [74] version 2.7.1+. The top 100 hits were collected, and majority vote was used to determine which orthogroup the TBLASTN hit would have belonged to had it been annotated by FunGAP. If no TBLASTN hits were found or if the BLASTP hits did not match the original missing orthogroup, the absence was counted as a validated absence, otherwise it was removed from the preliminary set of missing orthogroups.

Effector annotation

Effectors were predicted in all proteomes by first selecting genes with signal peptides which were predicted using SignalP [139] version 4.1 using the “euk” organism type and using 0.34 as a D-cutoff for both noTM and TM networks. Genes with predicted transmembrane domains from TMHMM [81] version 2.0c were then excluded. Finally, EffectorP [140] version 3.0 was used to predict effectors from this secreted gene set. Effector orthogroups were then called if at least half of the orthologs within the orthogroup were annotated as predicted effectors.

Principal component analysis and identification of lineage-differentiating PAV orthogroups

The matrix of missing effector orthogroups for each MoO isolate was used for PCA using the prcomp function in R version 3.6.1. PCA was performed a second time using the matrix of all missing orthogroups. Lineage-differentiating PAV orthogroups were then chosen by selecting the orthogroups that contributed more than 0.1% of the variance to PCs 1 and 2 using the get_pca_var function in R version 3.6.1.

Gene ontology and protein family enrichment analyses

All proteins were annotated for GO terms using the PANNZER2 [94] webserver and command line software SANSPANZ version 3 in October 2022. Only annotations with a positive predictive value greater than 0.6 and an ARGOT rank of 1 were kept. All GO terms assigned to genes within an orthogroup were then transferred to their orthogroup. GO term enrichment analysis was then performed using TopGO [95] version 2.36.0 and enrichment was calculated using the Fisher’s exact test and the “weight” algorithm. Only GO terms that were assigned to 3 or more lineage-differentiating PAV orthogroups and who’s enrichment was significant at a p-value of less than 0.05 were reported. PFAM enrichment analysis was performed by annotating PFAM domains using pfam_scan.pl [141] version 1.6-4 and the PFAM-A database. The output from pfam_scan.pl was parsed using K-parse_Pfam_domains_v3.1.pl (https://github.com/krasileva-group/plant_rgenes) [142] and an e-value cutoff of 0.001, and domain names were simplified by removing numbers and additional letters attached to domain names. Orthogroups were called as containing a domain if at least half of their orthologs had that domain annotation. Fisher's exact test for enrichment was performed using the scipy.stats Python module [103] version 1.9.0. Only domains which were observed in three or more lineage-differentiating PAV orthologs and with enrichment p-values less than 0.05 were reported.

Identification of genomic deletions

Illumina sequencing data was obtained from 117 datasets for MoO and 47 datasets for MoT from the SRA (Additional File 3 and Additional File 4). Reads were mapped to the *M. oryzae* Guy11 genome (GCA002368485.1) for MoO datasets and to the *M. oryzae* B71 genome (GCA004785725.2) for MoT datasets using BWA MEM [71] version 0.7.17-r1188. Read duplicates were marked using Picard (https://broadinstitute.github.io/picard/) version 2.9.0. Structural variants were then called using smoove (https://github.com/brentp/smoove) version 0.2.8, wham [143] version 1.7.0-311-g4e8c, Delly [144] version 0.9.1, and Manta [145] version 1.6.0 using default settings. The Delly output was processed using bcftools [146] version 1.6 to keep only called structural variants that passed Delly's quality control. Structural variants were then merged and filtered using SURVIVOR [147] version 1.0.7. Structural variants that were the same type, were on the same strand, and had breakpoints within 1000bp were merged. Only structural variants that were called by three or more callers and were larger than 50 bp were kept. Finally, the structural variants called for each dataset were all merged as before except breakpoints within 100bp were merged together. From this list of all structural variants, only genomic deletions were kept for further analysis.

Definition of PAV orthogroups and conserved groups

For each lineage, PAV orthogroups were defined by first taking the matrix of validated PAVs and filtering this matrix to orthogroups that were present in at least two isolates and absent in at least two isolates. The SCO phylogeny of the lineage was then analyzed for each candidate PAV orthogroup. If the orthogroup was only absent in strains that formed a monophyletic group, the orthogroup was not considered to be a PAV orthogroup. Additionally, if the orthogroup was only found in strains that formed a monophyletic group, the orthogroup was not considered to be a PAV orthogroup either. All orthogroups that were therefore present in two independent groups and absent in two independent groups were labeled PAV orthogroups. All orthogroups that were missing one or fewer strains were considered conserved orthogroups. All other orthogroups were considered "other".

Transposable element annotation

TE annotation was performed using RepeatMasker [86] version 4.1.1 and a reference TE library for all pathotypes of *M. oryzae* generated by Nakamoto *et al.* [122]. The parameters -cutoff 250, -nolow, -no_is, and -norna were used for the RepeatMasker command.

Next-generation sequencing data and GC content analysis

RNAseq data for MoO was obtained from SRA (Additional File 13) from a previous study [91] and mapped to the *M. oryzae* Guy11 genome (GCA002368485.1) for in culture data and the *M. oryzae* Guy11 genome combined with the *Oryza sativa* Nipponbare genome (GCA001433935.1) for the *in planta* data. RNAseq data for MoT was obtained from SRA accessions SRR9127598 through SRR9127602 from a previously published study [22] and mapped to the *M. oryzae* B71 genome (GCA004785725.2) for in culture data and the *M. oryzae* B71 genome combined with the wheat *Triticum aestivum* genome (GCA900519105.1) for the in planta data. Mapping was performed using STAR [92] version 2.7.1a and index files for mapping were made using the previously mentioned genomes and genome combinations along with corresponding gene annotation files obtained from FunGAP for the *M. oryzae* genomes, or from GenBank for the rice and wheat genomes. Read counts for each gene were calculated using the –quantMode GeneCounts parameter in STAR. These read counts were normalized to gene size as reads per kilobase values (RPK), then the total number of RPKs were summed for each sample and divided by one million. This sum was used to normalize read counts in each sample to obtain transcript per million (TPM) values for each sample. These TPM values were then averaged across treatments.

Published ChIPSeq data for H3K27me3, H3K27ac and H3K36me3 histone marks were obtained from a study published by Zhang *et al.* [91]. Published eccDNA sequencing data were obtained from a previous study by Joubert and Krasileva [148]. Reads were mapped to the *M. oryzae* Guy11 genome using BWA MEM [71] version 0.7.17-r1188. Read counts per gene were obtained using the coverage command from the BEDtools suite of tools [87] version 2.28.0. Read counts were normalized for gene and library size and averaged per treatment as for RNAseq data.

Methylation data from *M. oryzae* mycelium was obtained from a previous study published by Jeon *et al.* [149]. Reads were mapped to the *M. oryzae* genome and processed using the Bismark pipeline [150] version 0.24.0. Methylation percentage for all cytosines were extracted while ignoring the first 2 bases of all reads. The percentage of methylated cytosines was then calculated for a gene by averaging the methylation percentage of all cytosines in that gene.

To assign signal from next-generation sequencing datasets to orthogroups, signal for all orthologs in *M. oryzae* Guy11 and *M. oryzae* B71 within each orthogroup were averaged. Any orthogroups that did not have orthologs from B71 and Guy11 within them were given a value equal to the median value for all other orthogroups. *M. oryzae* Guy11 was not included in the original orthogrouping so a separate set of orthogroups were generated which included the *M. oryzae* Guy11 proteome annotated using FunGAP [77] as previously described in order to transfer next-generation sequencing signals.

Finally, GC content values for genes and flanking regions were calculated using the nuc command in BEDTools [87] version 2.28.0.

The same methods were used for calculating these values for genomic deletions.

Profile plots

10bp windows were first generated for each *M. oryzae* reference genome. The number of TEs and the number of genes in each window were then calculated using the coverage command in BEDTools [87] version 2.28.0 and stored as bedgraph files. Bigwig files were generated from bedgraph files using the bedGraphToBigWig tool (https://www.encodeproject.org/software/bedgraphtobigwig/) version 4. Finally, data for profile plots of genomic deletions were generated using the computeMatrix scale-regions and the plotProfile commands of the DeepTools suite of tools [89] version 3.5.1.

Random forest classification and feature importances calculation

Random forest classifiers were trained and performance statistics were calculated using the scikit-learn Python module [151] version 1.1.1. The hyperparameters used to train the model were as follows: 2000 estimators, a minimum of two samples to split a node, no minimum number of samples per leaf, no maximum tree depth, no maximum number of features per tree, and bootstrapping enabled. Classifiers were trained only on data for genes belonging to lineages 2 and 3 for MoO. Before training, all genes belonging to four genomes from each lineage were removed. From the remaining data, 50% of the genes not labeled as PAV were removed to improve the balance between PAV genes and non-PAV genes in the training data. The model was then trained and tested on the genes from the eight genomes that were removed before testing. The training and testing data split was repeated 100 times to generate average precision, recall, and F1 values as well as average number of true positives, false positives, true negatives, and false negatives for all models.

Feature importances were calculated according to methods described within the rfpimp Python module (https://github.com/parrt/random-forest-importances). Briefly, a random forest classifier was trained and tested as before to measure a baseline F1 statistic. Each variable in the testing data was then permuted in turn and a new F1 statistic for the model was generated on the permuted data. The difference between the baseline F1 and the new F1 were then calculated. This process was then repeated 100 times and the average decrease in the F1 statistic when each variable was permuted were reported.

Spearman and point biserial correlation coefficients between variables were calculated using the cor function in R version 3.6.1. Phi correlation coefficients were calculated using the psych package [152] version 2.2.9. To calculate dependence statistics for each variable in the complete MoO model, a random forest classifier or a random forest regressor was used to predict each variable originally used to train the PAV gene prediction model using all remaining variables. The same hyperparameters and train-test split were used to train and test each model as for the original PAV gene prediction model. Baseline F1 or R^2 values for each model were then calculated and the change in these values when each variable within the model was permuted were calculated as before. However, the results reported were only from a single run of this analysis.

Data processing and analysis

Data processing was performed in a RedHat Enterprise Linux environment with GNU bash version 4.2.46(20)-release. GNU coreutils version 8.22, GNU grep version 2.20, GNU sed version

4.2.2, gzip version 1.5, and GNU awk version 4.0.2 were all used for processing and handling. Conda (https://docs.conda.io/en/latest/) was used to facilitate installation of software and packages. Code parallelization was performed with GNU parallel [99] version 20180322. Previously published data was downloaded using curl version 7.65.3 (https://curl.se/) and sra-tools version 2.10.4 (https://github.com/ncbi/sra-tools). BED format files were processed using BEDtools [87] version 2.28.0. VCF format files were processed using bcftools [146] version 1.6. SAM and BAM format files were processed using SAMtools [146] version 1.8. FASTA format files were processed using seqtk (https://github.com/lh3/seqtk) version 1.2-r102-dirty.

Data processing and analysis were performed using custom Python scripts written in Python version 3.10.5 with the help of pandas [101] version 1.4.3 and numpy [102] version 1.23.1. GFF format files were parsed in Python using BCBio GFF version 0.6.9 (https://github.com/chapmanb/bcbb/tree/master/gff). FASTA format files were processed in python using SeqIO from Biopython [153] version 1.80.

Data processing and analysis were also performed using custom R scripts written in R version 3.6.1 with the help of data.table [104] version 1.13.6, tidyr [105] version 1.1.3, reshape2 [106] version 1.4.4, and dplyr [107] version 1.0.4. Plotting was performed using the ggplot2 package [108] version 3.3.5 and the ggnewscale package [154] version 0.4.8. Phylogenies were analyzed and plotted using the ape [138] package version 5.5 and the phytools package [155] version 0.7.90.

Declarations

Availability of data and materials

All code for data generation, analysis, and plotting is available on GitHub: https://github.com/pierrj/moryzae_pav_manuscript_code

All files used for analysis and plotting are available on Zenodo under the DOI 10.5281/zenodo.7444379.

Authors' contributions

PMJ conceptualized and designed the study with input from KVK. PMJ designed the analyses and analyzed the data. PMJ wrote the original draft manuscript. PMJ and KVK reviewed and edited the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Anne Nakamoto and other Krasileva lab members for thoughtful comments and feedback on the manuscript. This research used the Savio computational cluster resource provided by the Berkeley Research Computing program at the University of California, Berkeley (supported by the UC Berkeley Chancellor, Vice Chancellor for Research, and Chief Information Officer).

Supplementary Information

Additional File 1 (Supplementary Figures) has been included below in its entirety. Only the descriptions of all other additional files have been included below. All additional files are available as part of the original publication that this chapter was based on, as described at the beginning of the chapter.

Additional File 1: Supplementary Figures.

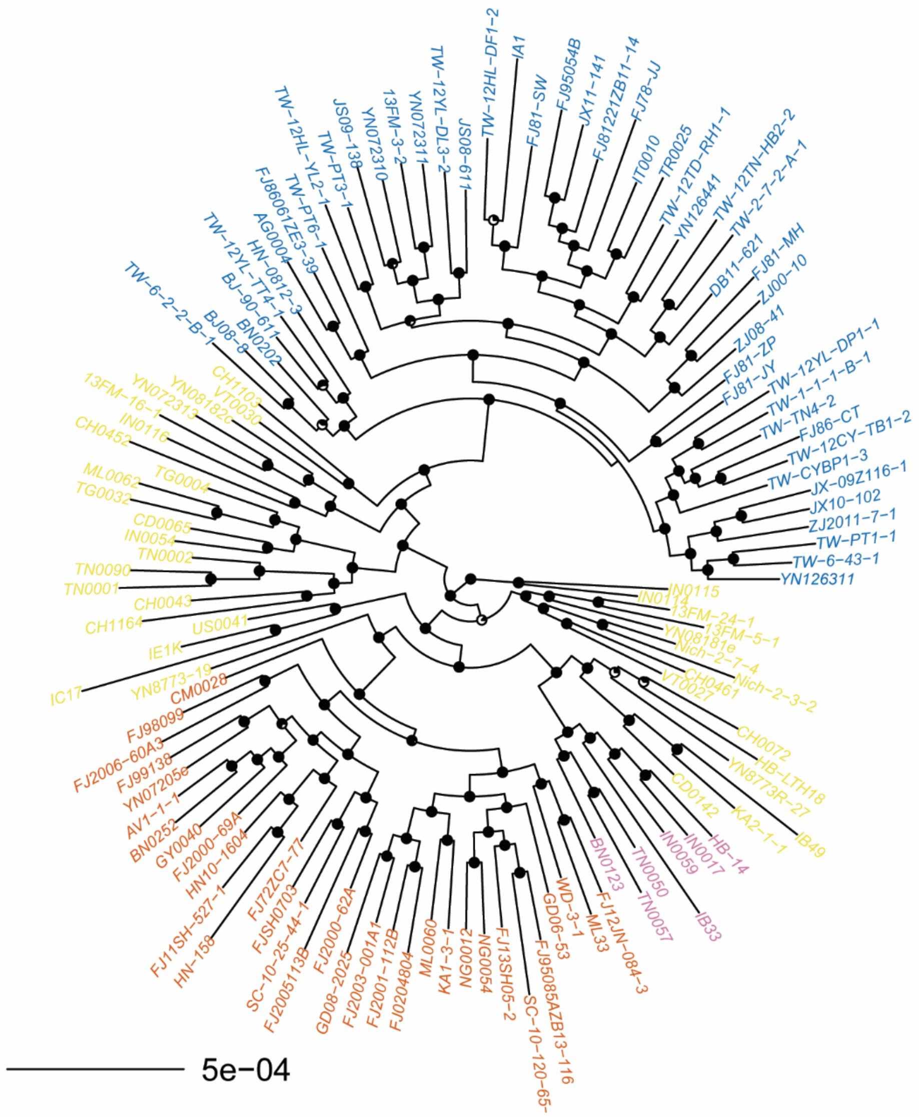


Fig. S1. Phylogeny of rice-infecting *M. oryzae* isolates used in this study. Phylogeny was generated using a multiple-sequence alignment of SCOs and fasttree [137]. Pie charts on nodes represent the fraction of bootstrap replicates that support the node. Isolates belonging to lineage 1 are colored yellow, isolates belonging to lineage 2 are colored orange, isolates

belonging to lineage 3 are colored blue, and isolates belonging to lineage 4 are colored pink. Lineages were named as previously described [28].

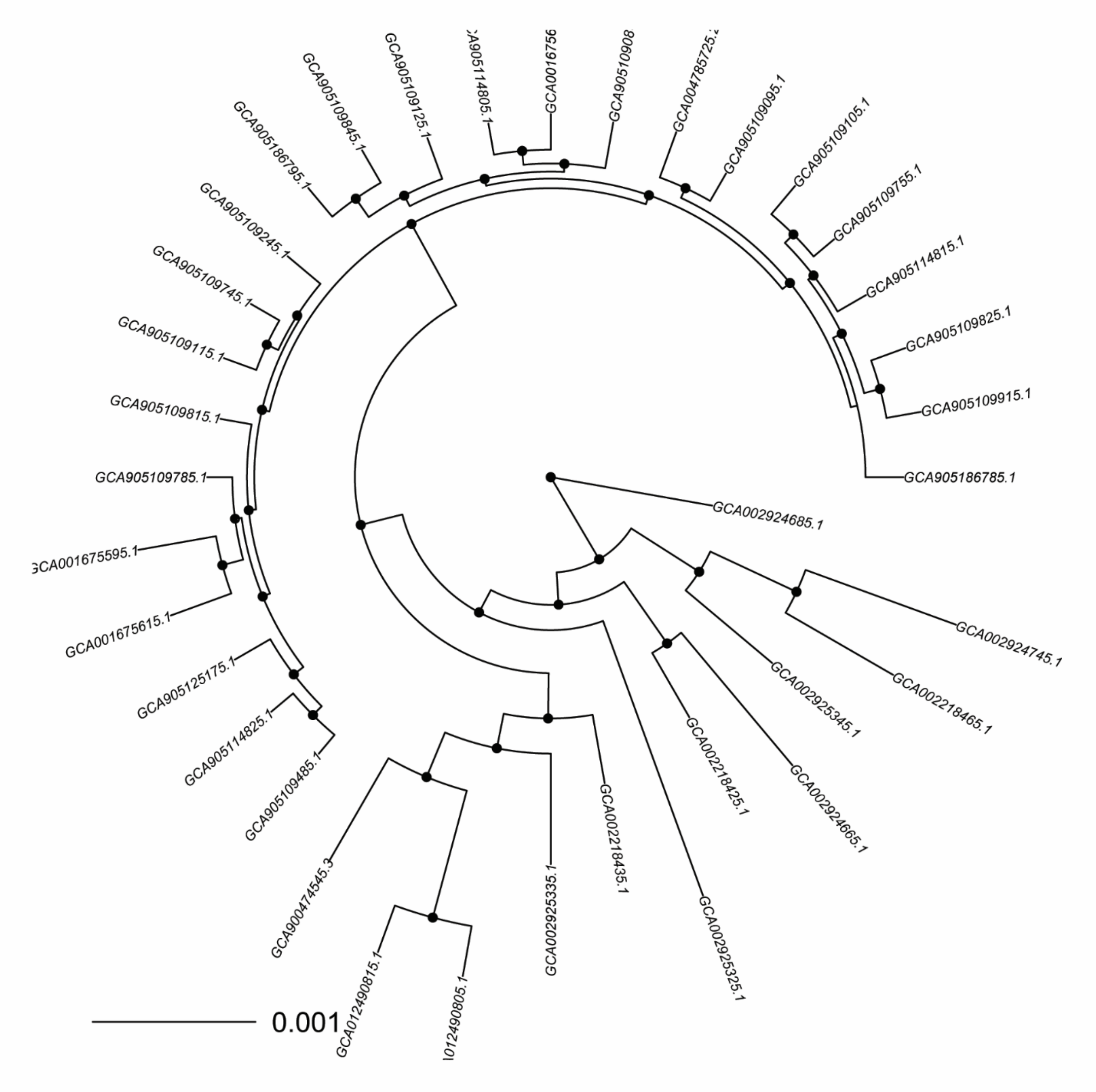


Fig. S2. Phylogeny of wheat-infecting *M. oryzae* isolates used in this study. Phylogeny was generated using a multiple-sequence alignment of SCOs and fasttree [137]. Pie charts on nodes represent the fraction of bootstrap replicates that support the node.

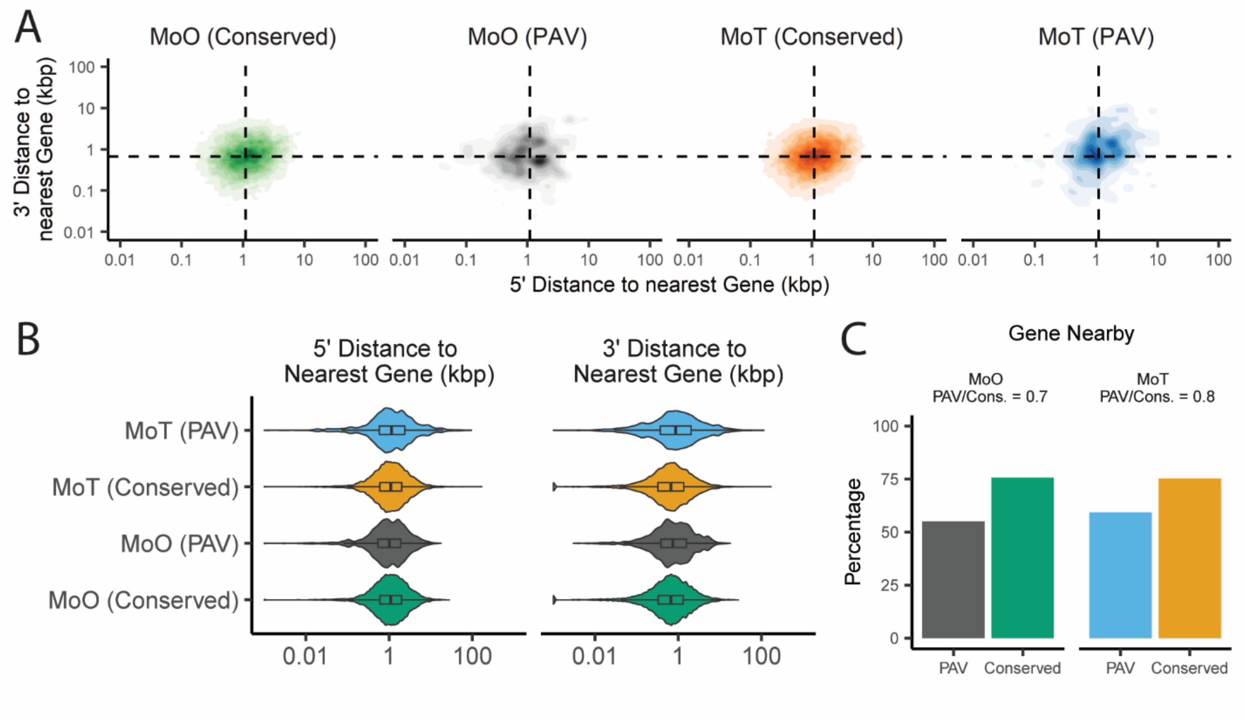


Fig. S3. Distances to the nearest gene for PAV and conserved genes in MoO and MoT. A. Density plots showing the distribution of the distances to the nearest gene for conserved and PAV genes in MoO and MoT. B. Violin plot showing the distribution of the distances to the nearest gene for conserved and PAV genes in MoO and MoT. C. Percentages and proportions of PAV and conserved genes that are within 1000bp of another gene in MoO and MoT. Dashed lines in density plots represent the median values for all genes in both pathotypes. Median values and

statistical comparisons for data shown are listed in Additional File 7, Additional File 8, and Additional File 9.

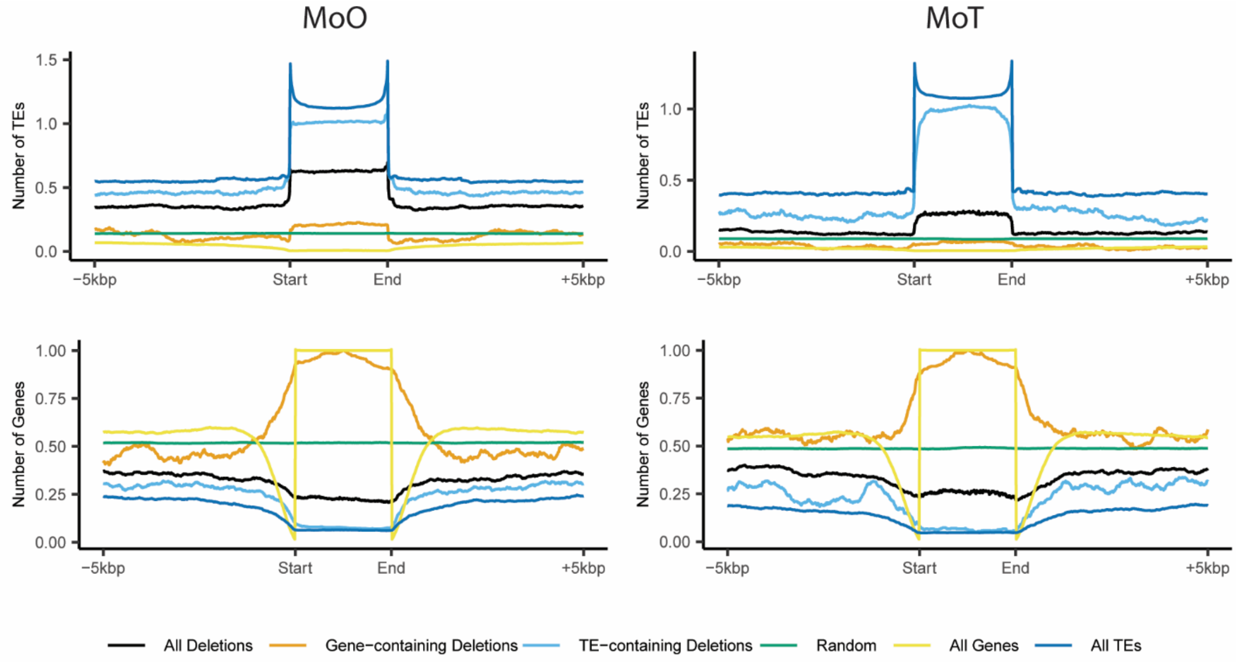


Fig. S4. Profile plots showing transposable element (TE) and gene density within genomic regions of the rice and wheat-infecting *M. oryzae* genomes. The flanking regions of these regions are also shown. Gene- and TE-containing regions represent the subet of all deletions that overlapped at least 50% with a gene or TE sequence, respectively. Genomic deletions were

shuffled throughout the genome 100 times to generate the data for random regions in the plots.

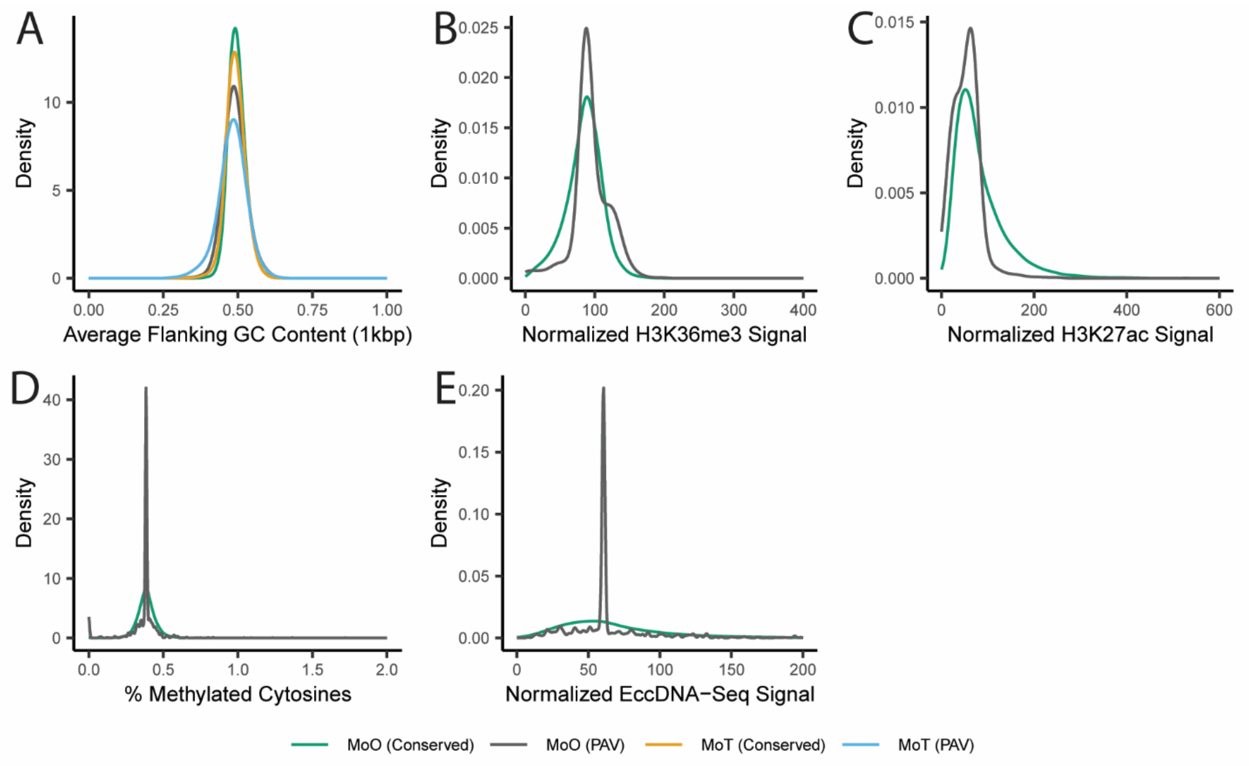


Fig. S5. Density plots of additional features of PAV and conserved genes. Density plots showing the distributions of A. average flanking GC content, B. normalized H3K36me3 histone mark ChIP-Seq signal, C. noramlized H3K27ac histone mark ChIP-Seq signal, D. average % methylation of cytosines, and E. normalized extrachromosomal DNA (eccDNA) sequencing signal for PAV and conserved genes in MoO and MoT. In panel A, the line representing the data for MoO PAV genes appears behind the line representing data for MoT PAV genes. In panels B, C, D, and E, MoT genes were not included as this data is not available for MoT. Statistics describing

distributions and statistical comparisons between these statistics are listed in Additional File 10 and Additional File 11.

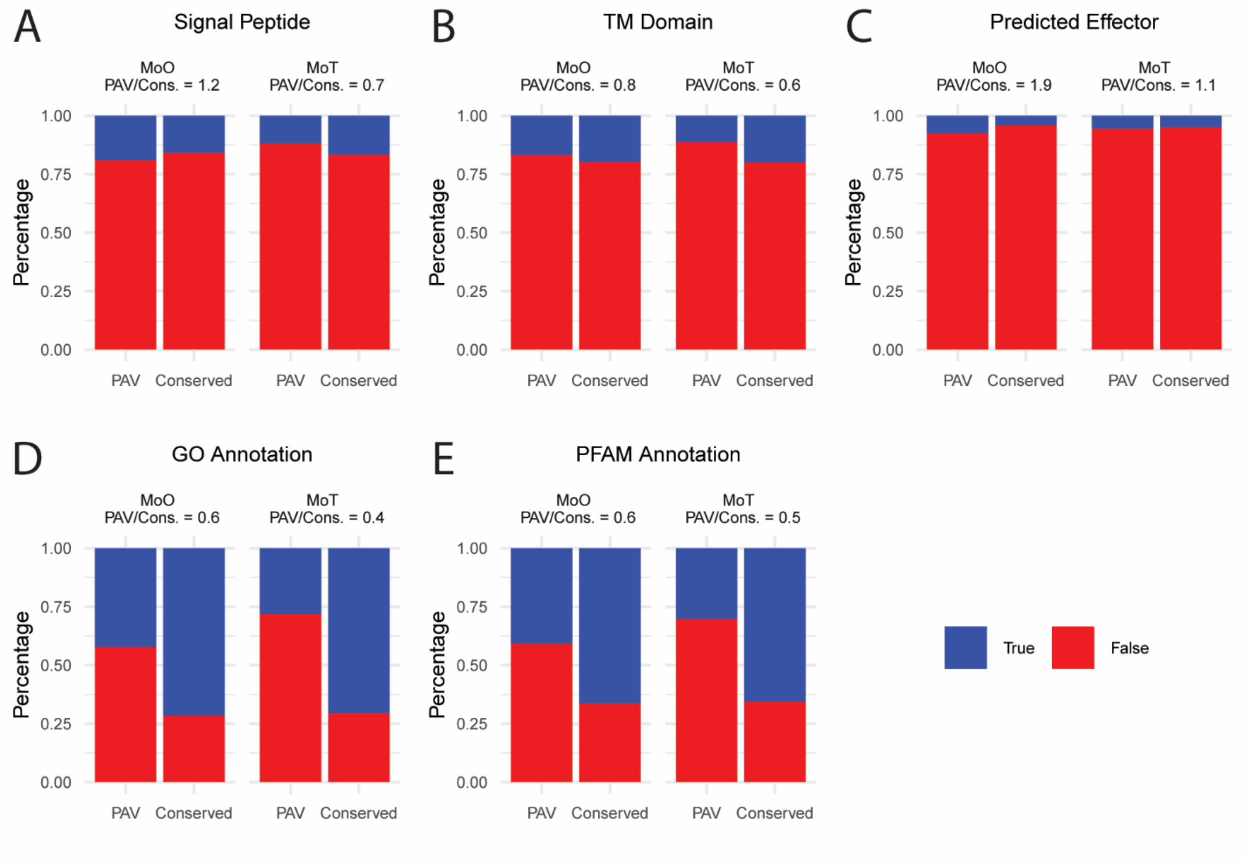


Fig. S6. Comparison of various functional annotations of PAV and conserved genes. Comparison of percentages and ratios of PAV and conserved genes annotated as A. having a signal peptide, B. having a transmembrane (TM) domain, C. being a predicted effector, D. having a GO annotation, and E. having a protein family (PFAM) domain annotation for MoO and MoT genes.

Counts for each category and stastical comparisons of these counts are listed in Additional File 8 and Additional File 9.

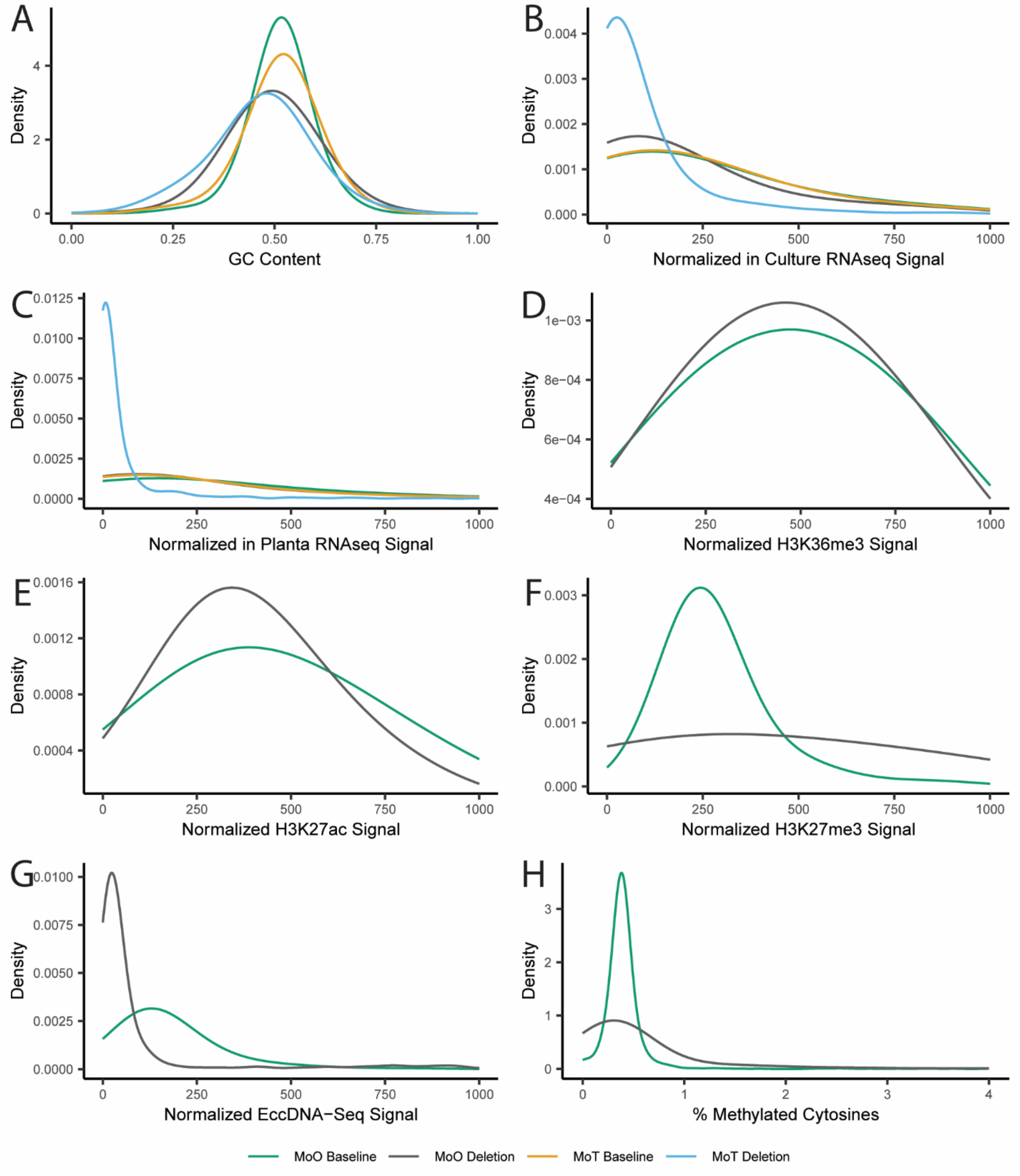


Fig. S7. Density plots showing the distributions of various features of MoO and MoT genomic deletions. Density plots showing the distributions of A. AT content, B. normalized in culture RNAseq signal, C. normalized in planta RNAseq signal, D. normalized H3K36me3 histone mark ChIP-Seq signal, E. normalized H3K27ac histone mark ChIP-Seq signal, F. normalized H3K27ac

histone mark ChIP-Seq signal, G. normalized eccDNA sequencing signal, and H. average % methylation of cytosines for genomic deletions in MoO and MoT, as compared to baseline. Genomic baseline values were generated by shuffling the deletions throughout the portions of the genome that were not deleted in any isolate. Statistics describing distributions and statistical comparisons between these statistics are listed in Additional File 12.

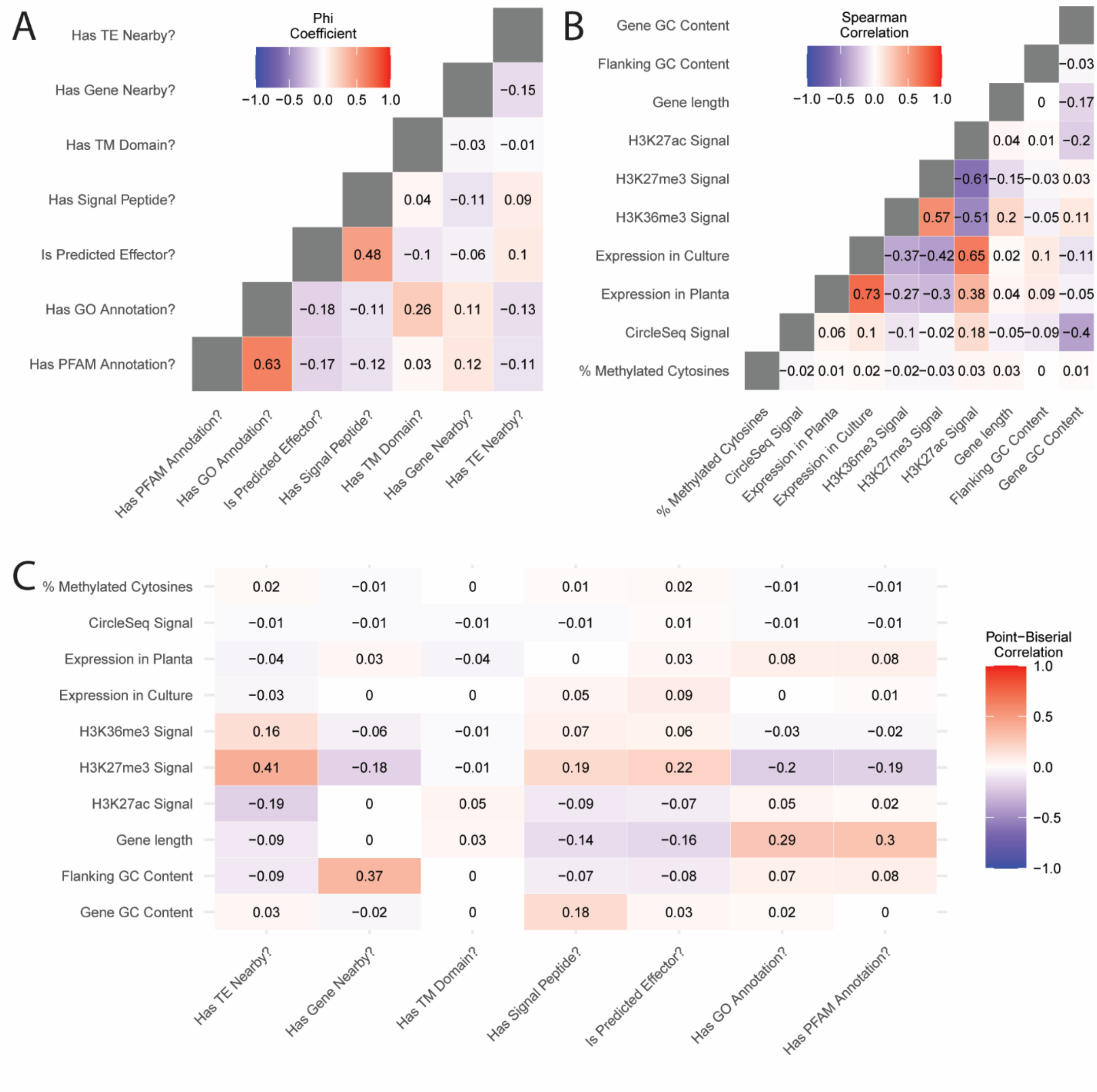


Fig. S8. Correlation coefficients for variables included in the MoO random forest classifier. Heat map representing A. Phi coefficient between binary variables, B. Spearman rank correlation

coefficient between continuous variables, and C. Point-Biserial correlation coefficient between continuous and binary variables.

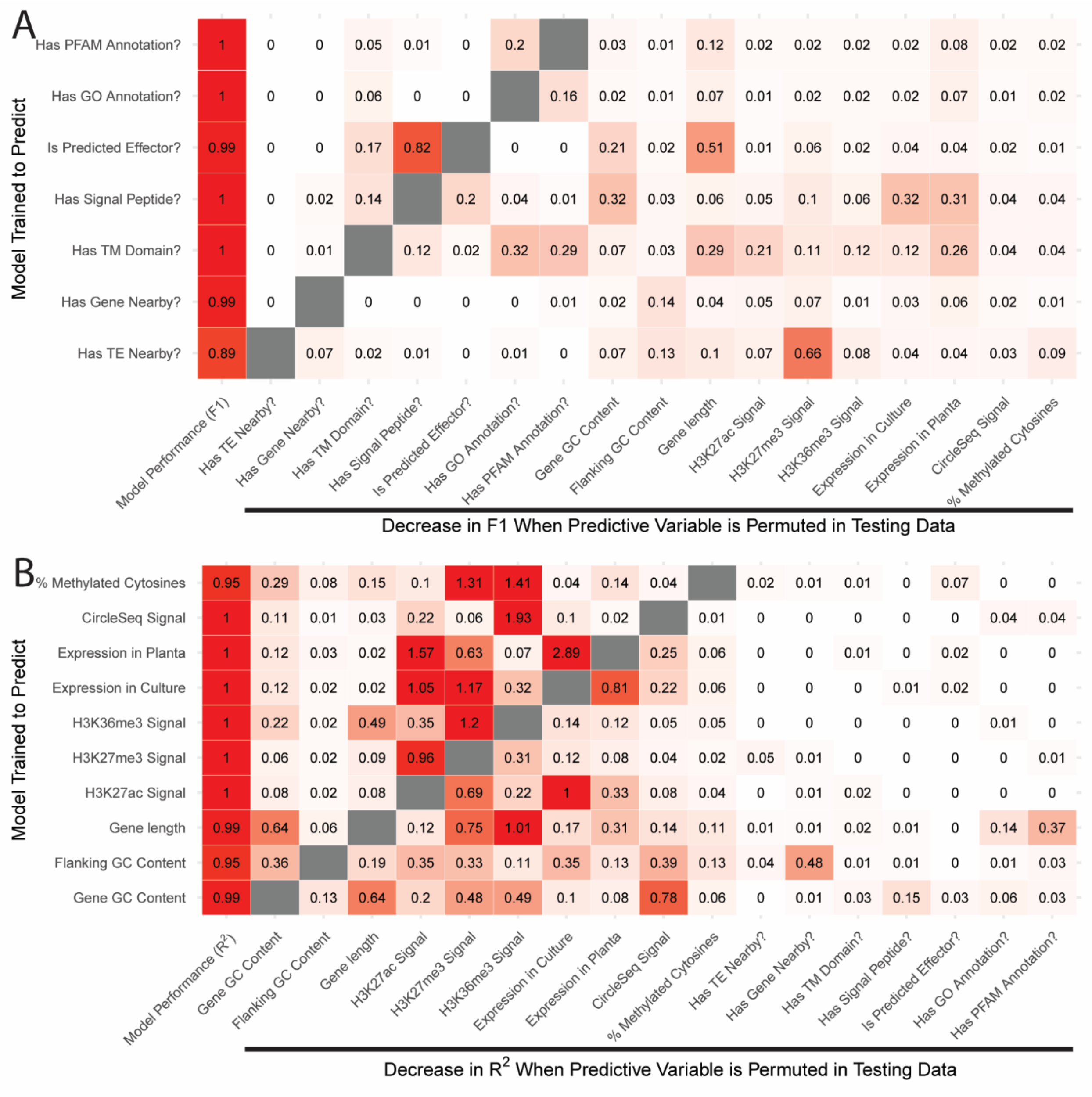


Fig. S9. Dependence matrix of variables included in the MoO random forest classifier. A model was trained to predict each variable used in our MoO random forest classifier using the remaining variables. A. Heatmap representing the F1 statistic of each model when trained to predict categorical variables and decrease in F1 when predictive variables were permuted in the testing data. B. Heatmap representing the R^2 statistic of each model when trained to predict

categorical variables and decrease in R^2 when predictive variables were permuted in the testing data.

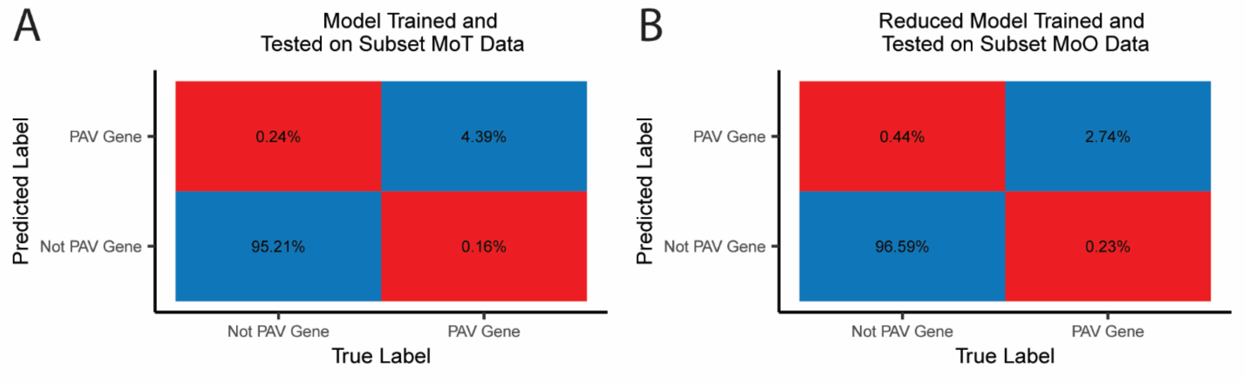


Fig. S10. Confusion matrices for the MoT random forest classifier and the MoO random forest classifier trained on a subset of features. A. Confusion matrix showing predictions of the MoT random forest classifier when tested on MoT genes that it was not trained on. B. Confusion matrix showing predictions of the MoO random forest classifier trained on a subset of features when tested on MoO genes that it was not trained on.

Additional File 2: List of accessions for MoT genomes.

Additional File 3: List of accessions for MoO Illumina data.

Additional File 4: List of accessions for MoT Illumina data.

Additional File 5: List of genomic deletions called using MoO Illumina sequencing data.

Additional File 6: List of genomic deletions called using MoT Illumina sequencing data.

Additional File 7: Table of median upstream and downstream distances to nearest PAV gene, TE, and gene for MoO and MoT. The p-values shown are two-tailed p-values resulting from permutation tests for the differences in medians between PAV and conserved genes for each pathotype with 1,000 permutations.

Additional File 8: Table showing the number of PAV and conserved genes that are near PAV genes, near TEs, near genes, have a TM domain, have a signal peptide, are predicted effectors, have a GO annotation, and have a PFAM domain annotation for MoO and MoT. The p-values shown were the results of Chi-squared tests used to test for indepedence between the PAV/conserved gene label and each feature for each pathotype.

Additional File 9: Table showing the number of PAV and conserved genes that are near PAV genes, near TEs, near genes, have a TM domain, have a signal peptide, are predicted effectors, have a GO annotation, and have a PFAM domain annotation for MoO and MoT. The p-values shown were the results of Chi-squared tests used to test for indepedence between the pathotype and each feature for each PAV/conserved gene label.

Additional File 10: Table showing the mean, median, standard deviation, 25th percentile and 75th percentile for the distributions of various continuous variables that describe PAV and

conserved genes in MoO and MoT. The p-values shown are two-tailed p-values resulting from permutation tests for the differences in each statistic between PAV and conserved genes for each pathotype with 1,000 permutations.

Additional File 11: Table showing the mean, median, standard deviation, 25th percentile and 75th percentile for the distributions of various continuous variables that describe PAV and conserved genes in MoO and MoT. The p-values shown are two-tailed p-values resulting from permutation tests for the differences in each statistic between pathotypes for PAV and conserved genes with 1,000 permutations.

Additional File 12: Table showing the mean, median, standard deviation, 25th percentile and 75th percentile for the distributions of various continuous variables that describe genomic deletions and baseline genomic regions in MoO and MoT. The p-values shown are two-tailed p-values resulting from permutation tests for the differences in each statistic between deletions and baseline for each pathotype with 1,000 permutations.

Additional File 13: List of accessions for MoO RNAseq data.

Chapter 5

Extended Discussion and Conclusions of Chapter 4 and Overall Future Outlooks

In Chapter 4, I analyzed gene PAV in the rice and wheat infecting pathotypes of *M. oryzae*. I began this work by designing robust tools to label PAV orthogroups in the fungus. When I started this project, other researchers that had looked at PAV in *M. oryzae* were simply using the outputs from the OrthoFinder software as a PAV matrix [25,26,96]. For my project, this method had a serious disadvantage: it did not account for annotation errors. As previously discussed in Chapters 1 and 3, I was interested in characterizing PAV in *M. oryzae* as a proxy for the generation of structural variation and genomic deletions. Therefore, while I used OrthoFinder to generate a list of candidate missing genes, I also used a TBLASTN and BLASTP-based approach to verify that the sequence of the gene was fully missing from the genome. This helped me verify that the genes that I was labeling missing were fully deleted from the genome; they were not missed by the annotation software, and the genomes did not have partial or mutated copies of the gene within them. This approach ignores gene silencing through processes like mutation but allowed me to take a focused look at gene PAV generated by structural variation.

Additionally, to perform statistical comparisons between PAV and conserved genes, I wanted to establish a stronger definition of what a PAV orthogroup was. Previously, researchers had simply reported the presence of extensive gene PAV in *M. oryzae* without clearly defining PAV orthogroups [25,26]. To this end, I focused on the clonal lineages of *M. oryzae* and came up with a strict definition for PAV orthogroups. In my project, these orthogroups were ones that showed evidence of at least two independent loss events. Independence of deletion events was defined using the genome phylogeny and relied on the assumption that the genomes were non-recombining. Therefore, when a gene was lost in two isolates that were not directly connected in the phylogeny, I assumed that this represented two independent genomic deletions. I believe that this robust definition of PAV orthogroups allowed me to conduct a more precise and thorough look at PAV in *M. oryzae* than what was previously possible and enabled the discoveries that I presented in Chapter 4.

As previously mentioned in Chapter 3, I am concerned that genomic studies in *M. oryzae* and other fungal plant pathogens are too focused on effectors and not mindful of other genes that are important for their evolution. Knowing that the existence of differential PAV has been previously shown between rice-infecting lineages of *M. oryzae* [25,26], I dug deeper into this result and found that, while these PAV orthogroups are enriched in disease-causing genes, only a small fraction of PAV orthogroups are disease-causing. When I looked at the rest of these orthogroups, I also found an enrichment in genes related to non-self-recognition and antibiotic production. To me this was a significant result that indicated that the evolution of geographically separated populations of *M. oryzae* involves adaptation to the local rice genotypes but also to the local microbiome. It would be interesting to sequence and characterize the rice and soil-associated microbiomes of each of the regions where clonal *M. oryzae* lineages are present and analyze how differences between them might relate to differences in antibiotic production and non-self-recognition genes between the isolates found there. This fascinating result goes back to my initial motivation for this avenue of research: focusing too much on effectors can make us miss important insights into fungal plant pathogen genomics and biology. It also further motivated my efforts to look at the genomic contexts that are associated with all gene PAV in *M. oryzae*, rather than just looking at PAV of effectors.

One of the most significant findings that I describe in Chapter 4 was the discovery of clear differences in PAV between the rice and wheat pathotypes of *M. oryzae*. I observed differences in the genomic features associated with PAV genes in each pathotype. I also observed differences in the distribution of PAV and conserved genes throughout the genomes of each pathotype. In the rice pathotype, PAV genes appear to be tightly clustered in specific, well-defined regions of the genome where no conserved genes are found. This is not the case in the wheat pathotype where many conserved genes are found next to PAV genes. Finally, I showed that wheat-infecting *M. oryzae* appears to experience gene PAV at an accelerated rate compared to the rice-infecting pathotype.

These findings could have important implications for the study of the wheat pathogen's emergence and how it was able to spread and become a devastating threat to agriculture so rapidly. It is possible that these differences in PAV are the result of changes in the way that its genome evolves, including potential changes to DNA repair mechanisms. This idea is supported by the decreased TE content that Anne Nakamoto and I found in the wheat-infecting lineage of *M. oryzae* [122]. EccDNA sequencing might help support this hypothesis as well, as modifications in DNA repair would likely influence MoT's eccDNA profile. On the other hand, accelerated evolution in MoT could also simply be the product of a substantial influx of genetic diversity which could have originated from a multi-hybrid swarm [118]. If this was the case, it is possible that the wheat pathotype would eventually reach some sort of equilibrium in its evolution that resembles the rice pathotype. Regardless, these novel observations could help explain why *M. oryzae* is such a successful pathogen and especially why the wheat pathotype has been so devastating. My thesis work makes it clear that there are significant differences in evolutionary history and genomics between the rice and wheat-infecting pathotypes of *M. oryzae* and it is possible that understanding these differences could lead to significant insights into the fungus' evolution.

A major goal of Chapter 4 was to understand the patterns that shape PAV in *M. oryzae*. Knowing that TEs had been intimately associated with fungal plant pathogen genome evolution in the past, I started by analyzing whether TEs were near genes prone to PAV in *M. oryzae* and found a clear association between the two. I then looked to see whether PAV genes were found in gene-poor regions of the genome, which is another hallmark of the two-speed genome hypothesis. While I did find a negative association between PAV-prone genes and gene density, this effect was much weaker than the association with TE density. These results showed that there were clear associations between genomic context and PAV and drove me to analyze every possible gene feature that I could to further understand these associations. Through comparing PAV and conserved genes, I found many additional features that set genes prone to PAV apart from conserved genes including GC content, length, expression, eccDNA production, and certain histone marks. The association between eccDNAs and PAV was no surprise given the results I published in Chapter 2, and further supported the association between PAV, eccDNAs, and *M. oryzae* genome evolution. Given the variety and strength of the evidence pointing to clear differences between genes prone to PAV and conserved genes, it was no surprise that I was able to train a machine learning model on these gene features that could predict whether a gene was prone to PAV in wheat and rice-infecting *M. oryzae*.

A significant caveat to these findings, however, is that the NGS datasets that I re-analyzed for Chapter 4 were only available for single isolates. This meant that I had to make many assumptions and imputations to fill in this data for other isolates, which likely resulted in many erroneous values in my dataset. However, many of the features I used do not rely on this imputation. I also show evidence in Chapter 4 that an RNA sequencing dataset for each isolate would likely be sufficient to generate an accurate predictive model. Despite this caveat, my results imply that, once a higher quality model is trained using one RNAseq dataset per isolate, we could quickly know which genes are prone to experience PAV in any emerging, particularly virulent strain of *M. oryzae* by simply sequencing its genome and its transcriptome, which is becoming more and more easy to do every year.

Of course, the success of this approach in *M. oryzae* raises an important question: could this approach be applied to other fungal plant pathogens? And if so, could we use it to predict which genes will be prone to being deleted in an emerging pathogen? I showed that a model trained on the wheat pathotype of *M. oryzae* could not be used to predict PAV in the rice pathotype and *vice versa* which appears to refute the idea that a general model that works well for all plant pathogens could be constructed. However, part of this result may be attributed to overfitting, and it is possible that more work could be done to generalize the model. Additionally, more complex models could be trained with data from other pathogens to see whether more general patterns that apply to all fungal pathogens could be used to construct such a model. Unfortunately, *Zymoseptoria tritici* is currently the only other fungal plant pathogen that has been sequenced to the same extent as *M. oryzae* [156]. Therefore, more genomes and transcriptomes of other pathogens need to be generated to see whether this approach is viable. Given the potential of this approach to predict gene PAV in an emerging fungal plant pathogen and help guide molecular biologists and geneticists in designing robust resistant crops that cannot be overwhelmed by gene deletion, constructing a general PAV prediction model should be attempted once the data becomes available.

While PAV gene prediction models show great potential in these fungi, there is much more to effector evolution than PAV. According to both my models and my phylogenetic data, most effectors in *M. oryzae* are conserved, which would imply that, if PAV was the only way that it adapted to crops, it would be easy to engineer resistant crops. Of course, this is not the case. There are several hypotheses that could help explain this observation and complement the over simplified model that genomic context is the only thing that shapes PAV and adaptation in *M. oryzae*. Firstly, while the loss of conserved genes is less likely than that of genes that my model labels “prone to PAV”, they can still be lost from the genome. It is also very likely that my model was too conservative in labeling genes prone to PAV and may also have been overfit. My model was optimized to only predict genes that had been previously observed to experience PAV and not all possible PAV genes, which is an important distinction. Additionally, it is possible that large effective population sizes allow the fungus to generate enough gene content diversity to overcome disease resistance. Though I excluded them from my analysis, this could make recombining lineages of *M. oryzae* particularly hard to combat. Another hypothesis is the frequent exchange of genomic information between pathotypes of *M. oryzae* [18,23]. According to this hypothesis, avirulent rice-infecting *M. oryzae* strains could obtain new effector alleles through sexual recombination, mini-chromosome exchange, or even asexual recombination through heterokaryon formation [157] with other pathotypes. The hypothesis describing a multi-hybrid host swarm that created the wheat-infecting pathotype could support these ideas [118].

Of course, mutations could play a more important role in *M. oryzae*'s evolution than PAV and transcriptional silencing, epigenetics, and TE insertions could be a big part of the story as well. It is very likely that machine learning models could be trained to predict which genes are prone to greater mutation rates. Proximity to TEs is also associated with an increased mutation rate [7,11], and my second chapter shows a clear association between characteristics associated with the two-speed genome hypothesis and other genomic and epigenomic features. Models trained to predict genes prone to increased mutation rates would therefore likely conclude that genes prone to PAV are also prone to mutations, which would result in the same problems as for the PAV model. Therefore, it is possible that the future of this type of modeling will need to involve much more complex models, that can not only identify which genes are prone to being deleted or mutated but can also predict which specific amino acid residues might be prone to change, allowing the pathogen to escape detection. Kyungyong Seong's work using computational structural predictions to model NLR-effector interactions and rationally engineer better NLRs could potentially lead to these types of models. Regardless, there is a lot of research that needs to be done to understand how each of these evolutionary mechanisms contribute to the adaptation of *M. oryzae* and other fungal plant pathogens to their hosts. This research will likely help guide predictive modeling efforts in the future.

Though it is still in early stages at the time of writing this dissertation, Kyungyong Seong and I have started a project in collaboration with Pierre Gladieux that seeks to develop our understanding of these evolutionary mechanisms. Pierre's group previously characterized the phenotypes of 45 rice varieties when infected by 70 *M. oryzae* isolates and found that different lineages of *M. oryzae* caused disease in different rice varieties [26]. Pursuing this goal of using computational biology to find targets for robust disease resistance engineering, Kyungyong and I decided to sequence and analyze the genomes of these 70 *M. oryzae* isolates. In this project we are looking to combine our expertise to find patterns in genomic features, PAV, structural variants, SNPs, or protein structures that might help further our understanding of disease resistance. We will be using methods such as genome-wide or k-mer association studies, structural effectoromics, protein docking, and machine learning to look for patterns in our dataset, with the goal of identifying new effector targets for disease resistance engineering as well as mechanisms the fungus uses to escape detection by its host. Many of the rice varieties that were tested against the 70 isolate panel also have genomic or NLR sequencing data available which could prove to be extremely valuable for these analyses. Of course, a goal of this research is to produce machine learning models that can predict disease phenotype from genomic sequences, which would be a significant breakthrough for the field. This could be used to identify which rice varieties should be planted in response to an outbreak of a particularly dangerous *M. oryzae* strain, for example. This type of project that combines various types of computational approaches will likely be a big part of the future of disease resistance engineering.

Fungal plant pathogens pose a serious threat to agriculture and the pervasiveness of intensive monoculture farming and climate change has made this threat even worse [1,2]. Additionally, several major pandemics are on the horizon. For example, the rapid spread of the devastating wheat blast fungus and the emergence of *Fusarium oxysporum* strains that infect the Cavendish banana could have massive economic impacts [3,4]. Current disease prevention strategies are unfortunately insufficient. While pesticides can keep fungal plant pathogens in check, they often only delay the spread of the fungi and result in significant environmental consequences [5,6]. Selective breeding and disease resistant engineering in crops is likely to be a far more successful solution. Unfortunately, the deployment of genetically engineered crops is currently extremely slow, in part due to technical challenges. Additionally, once resistant crops are deployed, their resistance can often be short-lived, as pathogens continue to adapt to these crops [10]. A better understanding of the plant immune system and its evolution will help accelerate the deployment of resistant crops and make these deployments more successful. Understanding fungal plant pathogen evolution will likely be equally important as this understanding will help determine which effectors make for good disease resistance engineering targets. It is clear that the partnership between these fields of studies will be essential to solving the plant disease crisis we face today and the key to sustainable agriculture in the future.

References

1. Fones HN, Bebber DP, Chaloner TM, Kay WT, Steinberg G, Gurr SJ. Threats to global food security from emerging fungal and oomycete crop pathogens. Nat Food. 2020;1: 332–342. doi:10.1038/s43016-020-0075-0

2. Raza MM, Bebber DP. Climate change and plant pathogens. Curr Opin Microbiol. 2022;70: 102233. doi:https://doi.org/10.1016/j.mib.2022.102233

3. Viljoen A, Ma L, Molina AB. CHAPTER 8: Fusarium Wilt (Panama Disease) and Monoculture in Banana Production: Resurgence of a Century-Old Disease. Emerging Plant Diseases and Global Food Security. The American Phytopathological Society; 2020. pp. 159–184. doi:https://doi.org/10.1094/9780890546383.008

4. Ceresini PC, Castroagudín VL, Rodrigues FÁ, Rios JA, Eduardo Aucique-Pérez C, Moreira SI, et al. Wheat Blast: Past, Present, and Future. Annu Rev Phytopathol. 2018;56: 427–456. doi:10.1146/annurev-phyto-080417-050036

5. Zubrod JP, Bundschuh M, Arts G, Brühl CA, Imfeld G, Knäbel A, et al. Fungicides: An Overlooked Pesticide Class? Environ Sci Technol. 2019;53: 3347–3365. doi:10.1021/acs.est.8b04392

6. Steinberg G, Gurr SJ. Fungi, fungicide discovery and global food security. Fungal Genet Biol. 2020;144: 103476. doi:https://doi.org/10.1016/j.fgb.2020.103476

7. Torres DE, Oggenfuss U, Croll D, Seidl MF. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biol Rev. 2020;34: 136–143. doi:https://doi.org/10.1016/j.fbr.2020.07.001

8. Selin C, de Kievit TR, Belmonte MF, Fernando WGD. Elucidating the role of effectors in plant-fungal interactions: Progress and challenges. Front Microbiol. 2016;7: 1–21. doi:10.3389/fmicb.2016.00600

9. Tamborski J, Krasileva K V. Evolution of Plant NLRs: From Natural History to Precise Modifications. Annu Rev Plant Biol. 2020;71: 355–378. doi:10.1146/annurev-arplant-081519-035901

10. 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;56: 21–40. doi:10.1146/annurev-phyto-080516-035303

11. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: Waltz with plants. Curr Opin Genet Dev. 2015;35: 57–65. doi:10.1016/j.gde.2015.09.001

12. Soanes D, Richards TA. Horizontal Gene Transfer in Eukaryotic Plant Pathogens. Annu Rev Phytopathol. 2014;52: 583–614. doi:10.1146/annurev-phyto-102313-050127

13. Bertazzoni S, Williams AH, Jones DA, Syme RA, Tan K-C, Hane JK. Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. Mol Plant-Microbe Interact. 2018;31: 779–788. doi:10.1094/mpmi-06-17-0135-fi

14. Fouché S, Plissonneau C, Croll D. The birth and death of effectors in rapidly evolving

filamentous pathogen genomes. Curr Opin Microbiol. 2018;46: 34–42. doi:10.1016/j.mib.2018.01.020

15. 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. Mol Plant Pathol. 2012;13: 414–430. doi:10.1111/j.1364-3703.2011.00783.x

16. 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;20: 155–172. doi:https://doi.org/10.1111/mpp.12747

17. 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. PLoS One. 2016;11: 1–15. doi:10.1371/journal.pone.0167295

18. Chuma I, Isobe C, Hotta Y, Ibaragi K, Futamata N, Kusaba M, et al. Multiple translocation of the avr-pita effector gene among chromosomes of the rice blast fungus magnaporthe oryzae and related species. PLoS Pathog. 2011;7. doi:10.1371/journal.ppat.1002147

19. Bao J, Chen M, Zhong Z, Tang W, Lin L, Zhang X, et al. PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae. Mol Plant. 2017;10: 1465–1468. doi:10.1016/j.molp.2017.08.008

20. Yoshida K, Saunders DGO, Mitsuoka C, Natsume S, Kosugi S, Saitoh H, et al. Host specialization of the blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. BMC Genomics. 2016;17: 1–18. doi:10.1186/s12864-016-2690-6

21. Chadha S, Sharma M. Transposable elements as stress adaptive capacitors induce genomic instability in fungal pathogen Magnaporthe oryzae. PLoS One. 2014;9. doi:10.1371/journal.pone.0094415

22. 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. PLoS Genet. 2019;15: 1–23. doi:10.1371/journal.pgen.1008272

23. Langner T, Harant A, Gomez-luciano LB, Shrestha RK, Win J. Genomic rearrangements generate hypervariable mini- chromosomes in host-specific lineages of the blast fungus. PLoS Genet. 2021;17(2). doi:10.1371/journal.pgen.1009386

24. Gomez Luciano LB, Tsai IJ, Chuma I, Tosa Y, Chen YH, Li JY, et al. Blast fungal genomes show frequent chromosomal changes, gene gains and losses, and effector gene turnover. Mol Biol Evol. 2019;36: 1148–1161. doi:10.1093/molbev/msz045

25. 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;18: 88. doi:10.1186/s12915-020-00818-z

26. 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. PLOS Pathog. 2022;18: 1–33. doi:10.1371/journal.ppat.1010687

27. Kim K, Ko J, Song H, Choi G, Kim H, Jeon J, et al. Evolution of the Genes Encoding Effector Candidates Within Multiple Pathotypes of Magnaporthe oryzae. Front Microbiol. 2019;10: 1–15. doi:10.3389/fmicb.2019.02575

28. 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. MBio. 2018;9. doi:10.1128/mBio.01806-17

29. Paulsen T, Kumar P, Koseoglu MM, Dutta A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. Trends Genet. 2018;34: 270–278. doi:10.1016/j.tig.2017.12.010

30. Kilzer JM, Stracker T, Beitzel B, Meek K, Weitzman M, Bushman FD. Roles of host cell factors in circularization of retroviral DNA. Virology. 2003;314: 460–467. doi:10.1016/S0042-6822(03)00455-0

31. Garfinkel DJ, Stefanisko KM, Nyswaner KM, Moore SP, Oh J, Hughes SH. Retrotransposon Suicide: Formation of Ty1 Circles and Autointegration via a Central DNA Flap. J Virol. 2006;80: 11920–11934. doi:10.1128/jvi.01483-06

32. Møller HD, Larsen CE, Parsons L, Hansen AJ, Regenberg B, Mourier T. Formation of extrachromosomal circular DNA from long terminal repeats of retrotransposons in Saccharomyces cerevisiae. G3 Genes, Genomes, Genet. 2016;6: 453–462. doi:10.1534/g3.115.025858

33. Gresham D, Usaite R, Germann SM, Lisby M, Botstein D, Regenberg B. Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. Proc Natl Acad Sci. 2010;107: 18551–18556. doi:10.1073/pnas.1014023107

34. Koo D-H, Molin WT, Saski CA, Jiang J, Putta K, Jugulam M, et al. Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed *Amaranthus palmeri*. Proc Natl Acad Sci. 2018;115: 3332–3337. doi:10.1073/pnas.1719354115

35. Molin WT, Yaguchi A, Blenner M, Saski CA. Autonomous replication sequences from the Amaranthus palmeri eccDNA replicon enable replication in yeast. BMC Res Notes. 2020;13: 330. doi:10.1186/s13104-020-05169-0

36. Molin WT, Yaguchi A, Blenner M, Saski CA. The EccDNA Replicon: A Heritable, Extranuclear Vehicle That Enables Gene Amplification and Glyphosate Resistance in Amaranthus palmeri[OPEN]. Plant Cell. 2020;32: 2132–2140. doi:10.1105/tpc.20.00099

37. Hull R, King M, Pizza G, Krueger F, Vergara X, Houseley J. Transcription-induced formation

of extrachromosomal DNA during yeast ageing. PLoS Biol. 2019;17. doi:10.1371/journal.pbio.3000471

38. Shcheprova Z, Baldi S, Frei SB, Gonnet G, Barral Y. A mechanism for asymmetric segregation of age during yeast budding. Nature. 2008;454: 728–734. doi:10.1038/nature07212

39. Nathanson DA, Gini B, Mottahedeh J, Visnyei K, Koga T, Gomez G, et al. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. Science (80-). 2014;343: 72–76. doi:10.1126/science.1241328

40. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, et al. Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature. 2017;543: 122–125. doi:10.1038/nature21356

41. Møller HD, Mohiyuddin M, Prada-Luengo I, Sailani MR, Halling JF, Plomgaard P, et al. Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. Nat Commun. 2018;9: 1–12. doi:10.1038/s41467-018-03369-8

42. Shibata Y, Kumar P, Layer R, Willcox S, Gagan JR, Griffith JD, et al. Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. Science (80-). 2012;336: 82–86. doi:10.1126/science.1213307

43. Durkin K, Coppieters W, Drögüller C, Ahariz N, Cambisano N, Druet T, et al. Serial translocation by means of circular intermediates underlies colour sidedness in cattle. Nature. 2012;482: 81–84. doi:10.1038/nature10757

44. Galeote V, Bigey F, Beyne E, Novo M, Legras JL, Casaregola S, et al. Amplification of a Zygosaccharomyces bailii DNA segment in wine yeast genomes by extrachromosomal circular DNA formation. PLoS One. 2011;6: 1–10. doi:10.1371/journal.pone.0017872

45. Wang K, Tian H, Wang L, Wang L, Tan Y, Zhang Z, et al. Deciphering extrachromosomal circular DNA in Arabidopsis. Comput Struct Biotechnol J. 2021;19: 1176–1183. doi:10.1016/j.csbj.2021.01.043

46. Lanciano S, Carpentier MC, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, et al. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLoS Genet. 2017;13: 1–20. doi:10.1371/journal.pgen.1006630

47. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci. 2015;112: E3114–E3122. doi:10.1073/pnas.1508825112

48. Dong OX, Ronald PC. Genetic engineering for disease resistance in plants: Recent progress and future perspectives. Plant Physiol. 2019;180: 26–38. doi:10.1104/pp.18.01224

49. Hollomon DW. Fungicide Resistance: 40 Years on and Still a Major Problem. Springer. 2015; 3–11. doi:10.1007/978-4-431-55642-8

50. Fouché S, Oggenfuss U, Chanclud E, Croll D. A devil's bargain with transposable elements in plant pathogens. Trends Genet. 2021; 1–9. doi:10.1016/j.tig.2021.08.005

51. Croll D, McDonald BA. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 2012;8: 8–10. doi:10.1371/journal.ppat.1002608

52. Fernandez J, Orth K. Rise of a Cereal Killer: The Biology of Magnaporthe oryzae Biotrophic Growth. Trends Microbiol. 2018;26: 582–597. doi:10.1016/j.tim.2017.12.007

53. 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. PLoS One. 2016;11: 1–15. doi:10.1371/journal.pone.0167295

54. Foster AJ, Martin-Urdiroz M, Yan X, Wright HS, Soanes DM, Talbot NJ. CRISPR-Cas9 ribonucleoprotein-mediated co-editing and counterselection in the rice blast fungus. Sci Rep. 2018;8: 1–12. doi:10.1038/s41598-018-32702-w

55. Magdolen V, Drubin DG, Mages G, Bandlow W. High levels of profilin suppress the lethality caused by overproduction of actin in yeast cells. FEBS Lett. 1993;316: 41–47. doi:10.1016/0014-5793(93)81733-G

56. Cohen S, Segal D. Extrachromosomal circular DNA in eukaryotes: Possible involvement in the plasticity of tandem repeats. Cytogenet Genome Res. 2009;124: 327–338. doi:10.1159/000218136

57. Ali MM, Li F, Zhang Z, Zhang K, Kang D-K, Ankrum JA, et al. Rolling circle amplification: a versatile tool for chemical biology, materials science and medicine. Chem Soc Rev. 2014;43: 3324–3341. doi:10.1039/C3CS60439J

58. Storlazzi CT, Lonoce A, Guastadisegni MC, Trombetta D, D'Addabbo P, Daniele G, et al. Gene amplification as double minutes or homogeneously staining regions in solid tumors: origin and structure. Genome Res. 2010;20: 1198–1206. doi:10.1101/gr.106252.110

59. Zhong Z, Chen M, Lin L, Chen R, Liu D, Norvienyeku J, et al. Genetic Variation Bias toward Noncoding Regions and Secreted Proteins in the Rice Blast Fungus Magnaporthe oryzae. mSystems. 2020;5. doi:10.1128/msystems.00346-20

60. Zhang P, Peng H, Llauro C, Bucher E, Mirouze M. ecc_finder: A Robust and Accurate Tool for Detecting Extrachromosomal Circular DNA From Sequencing Data. Front Plant Sci. 2021;12. doi:10.3389/fpls.2021.743742

61. Havecker ER, Gao X, Voytas DF. The diversity of LTR retrotransposons. Genome Biol. 2004;5. doi:10.1186/gb-2004-5-6-225

62. Dillon LW, Kumar P, Shibata Y, Wang YH, Willcox S, Griffith JD, et al. Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity. Cell Rep. 2015;11: 1749–1759. doi:10.1016/j.celrep.2015.05.020

63. Breier AM, Chatterji S, Cozzarelli NR. Prediction of Saccharomyces cerevisiae replication

origins. Genome Biol. 2004;5. doi:10.1186/gb-2004-5-4-r22

64. Wang Z-Q, Meng F-Z, Zhang M-M, Yin L-F, Yin W-X, Lin Y, et al. A Putative Zn2Cys6 Transcription Factor Is Associated With Isoprothiolane Resistance in Magnaporthe oryzae. Front Microbiol. 2018;9: 2608. doi:10.3389/fmicb.2018.02608

65. Bohnert S, Heck L, Gruber C, Neumann H, Distler U, Tenzer S, et al. Fungicide resistance toward fludioxonil conferred by overexpression of the phosphatase gene MoPTP2 in Magnaporthe oryzae. Mol Microbiol. 2019;111: 662–677. doi:https://doi.org/10.1111/mmi.14179

66. Norman A, Riber L, Luo W, Li LL, Hansen LH, Sørensen SJ. An Improved Method for Including Upper Size Range Plasmids in Metamobilomes. PLoS One. 2014;9: 1–12. doi:10.1371/journal.pone.0104405

67. Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, et al. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of saccharomyces cerevisiae. PLoS Genet. 2011;7. doi:10.1371/journal.pgen.1001287

68. Foster AJ, Jenkinson JM, Talbot NJ. Trehalose synthesis and metabolism are required at different stages of plant infection by Magnaporthe grisea. EMBO J. 2003;22: 225–235. doi:10.1093/emboj/cdg018

69. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17: 10–12. doi:10.14806/ej.17.1.200

70. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, et al. The genome sequence of the rice blast fungus Magnaporthe grisea. Nature. 2005;434: 980–986. doi:10.1038/nature03449

71. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv. 2013. doi:10.48550/arXiv.1303.3997

72. Li H. Minimap2: Pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34: 3094–3100. doi:10.1093/bioinformatics/bty191

73. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, et al. Primer3--new capabilities and interfaces. Nucleic Acids Res. 2012;40: e115. doi:10.1093/nar/gks596

74. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10: 421. doi:10.1186/1471-2105-10-421

75. Gao C-H. ggVennDiagram: A “ggplot2” Implement of Venn Diagram. 2021. Available: https://cran.r-project.org/package=ggVennDiagram

76. Vu VQ. ggbiplot: A ggplot2 based biplot. 2011. Available: http://github.com/vqv/ggbiplot

77. Min B, Grigoriev I V., Choi IG. FunGAP: Fungal Genome Annotation Pipeline using

evidence-based gene model evaluation. Bioinformatics. 2017;33: 2936–2937. doi:10.1093/bioinformatics/btx353

78. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6: 4–9. doi:10.1186/s13100-015-0041-9

79. 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;117: 9451–9457. doi:10.1073/pnas.1921046117

80. 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;37: 420–423. doi:10.1038/s41587-019-0036-z

81. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 2001;305: 567–580. doi:10.1006/jmbi.2000.4315

82. Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. Mol Plant Pathol. 2018;19: 2094–2110. doi:10.1111/mpp.12682

83. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18: 366–368. doi:10.1038/s41592-021-01101-x

84. 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: 98. doi:10.1186/1471-2229-10-98

85. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22: 4673–4680. doi:10.1093/nar/22.22.4673

86. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. Available: http://www.repeatmasker.org

87. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26: 841–842. doi:10.1093/bioinformatics/btq033

88. Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. Bioinformatics. 2012;28: 593–594. doi:10.1093/bioinformatics/btr708

89. Ramírez F, Ryan DP, Grüning B, Bhardwaj V, Kilpert F, Richter AS, et al. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 2016;44: W160–W165. doi:10.1093/nar/gkw257

90. Rice P, Longden L, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16: 276–277. doi:10.1016/S0168-9525(00)02024-2

91. Zhang W, Huang J, Cook DE. Histone modification dynamics at H3K27 are associated with altered transcription of in planta induced genes in Magnaporthe oryzae. PLoS Genet. 2021;17: 1–29. doi:10.1371/JOURNAL.PGEN.1009376

92. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29: 15–21. doi:10.1093/bioinformatics/bts635

93. Grant CE, Bailey TL, Noble WS. FIMO: scanning for occurrences of a given motif. Bioinformatics. 2011;27: 1017–1018. doi:10.1093/bioinformatics/btr064

94. Törönen P, Medlar A, Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res. 2018;46: W84–W88. doi:10.1093/nar/gky350

95. Alexa A, Rahnenfuhrer J. topGO: Enrichment Analysis for Gene Ontology. 2019.

96. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. Genome Biol. 2019;20: 1–14. doi:10.1186/s13059-019-1832-y

97. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLoS Comput Biol. 2018;14: 1–14. doi:10.1371/journal.pcbi.1005944

98. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics. 2005;6: 1–11. doi:10.1186/1471-2105-6-31

99. Tange O. GNU Parallel. 2018. doi:10.5281/zenodo.1146014

100. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25: 2078–2079. doi:10.1093/bioinformatics/btp352

101. pandas development team T. pandas-dev/pandas: Pandas. Zenodo; 2020. doi:10.5281/zenodo.3509134

102. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with {NumPy}. Nature. 2020;585: 357–362. doi:10.1038/s41586-020-2649-2

103. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17: 261–272. doi:10.1038/s41592-019-0686-2

104. Dowle M, Srinivasan A. data.table: Extension of `data.frame`. 2020. Available: https://cran.r-project.org/package=data.table

105. Wickham H. tidyr: Tidy Messy Data. 2021. Available: https://cran.r-project.org/package=tidyr

106. Wickham H. Reshaping Data with the {reshape} Package. J Stat Softw. 2007;21: 1–20. Available: http://www.jstatsoft.org/v21/i12/

107. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation. 2021. Available: https://cran.r-project.org/package=dplyr

108. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York; 2016. Available: https://ggplot2.tidyverse.org

109. Neuwirth E. RColorBrewer: ColorBrewer Palettes. 2014. Available: https://cran.r-project.org/package=RColorBrewer

110. Wickham H, Seidel D. scales: Scale Functions for Visualization. 2020. Available: https://cran.r-project.org/package=scales

111. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for “ggplot2.” 2020. Available: https://cran.r-project.org/package=cowplot

112. Slowikowski K. ggrepel: Automatically Position Non-Overlapping Text Labels with “ggplot2.” 2021. Available: https://cran.r-project.org/package=ggrepel

113. Kassambara A. ggpubr: “ggplot2” Based Publication Ready Plots. 2020. Available: https://cran.r-project.org/package=ggpubr

114. Hahne F, Ivanek R. Statistical Genomics: Methods and Protocols. In: Mathé E, Davis S, editors. New York, NY: Springer New York; 2016. pp. 335–351. doi:10.1007/978-1-4939-3578-9_16

115. Iannone R, Cheng J, Schloerke B. gt: Easily Create Presentation-Ready Display Tables. 2021. Available: https://cran.r-project.org/package=gt

116. Hancks DC, Kazazian HH. Roles for retrotransposon insertions in human disease. Mob DNA. 2016;7. doi:10.1186/s13100-016-0065-9

117. Gladieux P, Condon B, Ravel S, Soanes D, Maciel JLN, Nhani AJ, et al. Gene Flow between Divergent Cereal- and Grass-Specific Lineages of the Rice Blast Fungus Magnaporthe oryzae. MBio. 2018;9: 1–19. doi:10.1128/mBio.01219-17

118. 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. bioRxiv. 2021. doi:10.1101/2021.11.24.469688

119. Huang J, Rowe D, Subedi P, Zhang W, Suelter T, Valent B, et al. CRISPR-Cas12a induced DNA double-strand breaks are repaired by multiple pathways with different mutation profiles in Magnaporthe oryzae. Nat Commun. 2022;13: 7168. doi:10.1038/s41467-022-34736-1

120. Ikeda K, Nakayashiki H, Takagi M, Tosa Y, Mayama S. Heat shock, copper sulfate and oxidative stress activate the retrotransposon MAGGY resident in the plant pathogenic fungus Magnaporthe grisea. Mol Genet Genomics. 2001;266: 318–325. doi:10.1007/s004380100560

121. Parsons KA, Chumley FG, Valent B. Genetic transformation of the fungal pathogen responsible for rice blast disease. Proc Natl Acad Sci. 1987;84: 4161–4165. doi:10.1073/pnas.84.12.4161

122. Nakamoto AA, Joubert PM, Krasileva K V. Evolutionary dynamics of transposable elements in Magnaporthe oryzae reveal evidence of genomic transfer and key differences between rice and wheat blast pathotypes. bioRxiv. 2022. doi:10.1101/2022.11.27.518126

123. McInerney JO, McNally A, O’Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2: 17040. doi:10.1038/nmicrobiol.2017.40

124. McCarthy CGP, Fitzpatrick DA. Pan-genome analyses of model fungal species. Microb Genomics. 2019;5. doi:https://doi.org/10.1099/mgen.0.000243

125. 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. 2020;18: 12. doi:10.1186/s12915-020-0744-3

126. Moolhuijzen PM, See PT, Shi G, Powell HR, Cockram J, Jørgensen LN, et al. A global pangenome for the wheat fungal pathogen Pyrenophora tritici-repentis and prediction of effector protein structural homology. Microb Genomics. 2022;8. doi:https://doi.org/10.1099/mgen.0.000872

127. Kaushik A, Roberts DP, Ramaprasad A, Mfarrej S, Nair M, Lakshman DK, et al. Pangenome Analysis of the Soilborne Fungal Phytopathogen Rhizoctonia solani and Development of a Comprehensive Web Resource: RsolaniDB. Front Microbiol. 2022;13. doi:10.3389/fmicb.2022.839524

128. Badet T, Fouché S, Hartmann FE, Zala M, Croll D. Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. Nat Commun. 2021;12: 3551. doi:10.1038/s41467-021-23862-x

129. 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;12: 1867–1878. doi:10.1038/s41396-018-0100-6

130. Pordel A, Ravel S, Charriat F, Gladieux P, Cros-Arteil S, Milazzo J, et al. Tracing the Origin and Evolutionary History of Pyricularia oryzae Infecting Maize and Barnyard Grass. Phytopathology®. 2020;111: 128–136. doi:10.1094/PHYTO-09-20-0423-R

131. Dyrka W, Lamacchia M, Durrens P, Kobe B, Daskalov A, Paoletti M, et al. Diversity and Variability of NOD-Like Receptors in Fungi. Genome Biol Evol. 2014;6: 3137–3158. doi:10.1093/gbe/evu251

132. Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, et al. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography (Cop). 2013;36: 27–46. doi:https://doi.org/10.1111/j.1600-0587.2012.07348.x

133. Seong K, Krasileva K V. Computational Structural Genomics Unravels Common Folds and Novel Families in the Secretome of Fungal Phytopathogen Magnaporthe oryzae. Mol Plant-Microbe Interact. 2021;34: 1267–1280. doi:10.1094/MPMI-03-21-0071-R

134. 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: 3210–3212. doi:10.1093/bioinformatics/btv351

135. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30: 772–780. doi:10.1093/molbev/mst010

136. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25: 1972–1973. doi:10.1093/bioinformatics/btp348

137. Price MN, Dehal PS, Arkin AP. FastTree 2 - Approximately maximum-likelihood trees for large alignments. PLoS One. 2010;5. doi:10.1371/journal.pone.0009490

138. Paradis E, Schliep K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in {R}. Bioinformatics. 2019;35: 526–528.

139. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8: 785–786. doi:10.1038/nmeth.1701

140. Sperschneider J, Dodds PN. EffectorP 3.0: Prediction of Apoplastic and Cytoplasmic Effectors in Fungi and Oomycetes. Mol Plant-Microbe Interact. 2022;35: 146–156. doi:10.1094/MPMI-08-21-0201-R

141. Madeira F, Pearce M, Tivey ARN, Basutkar P, Lee J, Edbali O, et al. Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. 2022; gkac240. doi:10.1093/nar/gkac240

142. Sarris PF, Cevik V, Dagdas G, Jones JDG, Krasileva K V. Comparative analysis of plant immune receptor architectures uncovers host proteins likely targeted by pathogens. BMC Biol. 2016;14: 8. doi:10.1186/s12915-016-0228-7

143. Kronenberg ZN, Osborne EJ, Cone KR, Kennedy BJ, Domyan ET, Shapiro MD, et al. Wham: Identifying Structural Variants of Biological Consequence. PLOS Comput Biol. 2015;11: e1004572. Available: https://doi.org/10.1371/journal.pcbi.1004572

144. Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28: i333–i339. doi:10.1093/bioinformatics/bts378

145. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics. 2016;32: 1220–1222. doi:10.1093/bioinformatics/btv710

146. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. Gigascience. 2021;10: giab008. doi:10.1093/gigascience/giab008

147. Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, et al. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. Nat Commun. 2017;8: 14061. doi:10.1038/ncomms14061

148. Joubert PM, Krasileva K V. The extrachromosomal circular DNAs of the rice blast pathogen Magnaporthe oryzae contain a wide variety of LTR retrotransposons, genes, and effectors. BMC Biol. 2022;20: 260. doi:10.1186/s12915-022-01457-2

149. Jeon J, Choi J, Lee GW, Park SY, Huh A, Dean RA, et al. Genome-wide profiling of DNA methylation provides insights into epigenetic regulation of fungal development in a plant pathogenic fungus, Magnaporthe oryzae. Sci Rep. 2015;5: 1–11. doi:10.1038/srep08567

150. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27: 1571–1572. doi:10.1093/bioinformatics/btr167

151. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12: 2825–2830. Available: http://jmlr.org/papers/v12/pedregosa11a.html

152. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research. Evanston, Illinois; 2022. Available: https://cran.r-project.org/package=psych

153. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25: 1422–1423. doi:10.1093/bioinformatics/btp163

154. Campitelli E. ggnewscale: Multiple Fill and Colour Scales in “ggplot2.” 2022. Available: https://cran.r-project.org/package=ggnewscale

155. Revell LJ. phytools: An R package for phylogenetic comparative biology (and other things). Methods Ecol Evol. 2012;3: 217–223.

156. Feurtey A, Lorrain C, McDonald MC, Milgate A, Solomon PS, Warren R, et al. A thousand-genome panel retraces the global spread and adaptation of a major fungal crop pathogen. Nat Commun. 2023;14: 1059. doi:10.1038/s41467-023-36674-y

157. Strom NB, Bushley KE. Two genomes are better than one: history, genetics, and biotechnological applications of fungal heterokaryons. Fungal Biol Biotechnol. 2016;3: 1–14. doi:10.1186/s40694-016-0022-x