Title

The extrachromosomal circular DNAs of the rice blast pathogen *Magnaporthe oryzae* contain a wide variety of LTR retrotransposons, genes, and effectors

Pierre M. Joubert and Ksenia V. Krasileva

Abstract

One of the ways genomes respond to stress is by shedding 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. We present a whole circularome sequencing study of the rice blast pathogen *Magnaporthe oryzae*. We find that *M. oryzae* has a highly diverse circularome containing many genes and showing 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, which indicates that the presence of these genes on eccDNAs is selected against.

Introduction

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 (DMs), small polydisperse circular DNAs (spcDNAs), and microDNAs. They form through several mechanisms including through non-allelic homologous recombination (HR), double strand break repair, replication slippage, replication fork stalling, R-loop formation during transcription[1], and as a byproduct of LTR retrotransposon activity[2–4] (Figure 1A). EccDNAs can accumulate in cells through autonomous replication[5,6], high rates of formation[7], or through retention in ageing cells[8]. EccDNAs can contain genes, and amplification of gene-containing eccDNAs has been linked to adaptation to copper[7] and nitrogen[5] stress in yeast, herbicide resistance[6] in weeds, and drug resistance in cancer cells[9,10]. EccDNA formation is thought to cause genomic deletions[5,11,12] and reinsertion of eccDNAs after their formation has also been thought to generate structural variation[13,14]. Some evidence also indicates that eccDNAs could facilitate horizontal gene transfer[14]. 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.

One of the greatest threats to food security is the devastation of crops by fungal plant pathogens. These pathogens secrete molecules known as effectors which modify host functions and cause disease[15]. The most promising and intensely researched solution to these diseases is the genetic modification of crops to introduce new resistance genes, often by allowing the crops to detect effectors and trigger immune

responses[16]. Unfortunately, the deployment of these crops has often had only short-term impacts as some fungal pathogens have adapted to these defenses in very short time spans[17]. Similarly, fungicides are often used to mitigate the devastation caused by these pathogens but these fungi often evolve resistance to these drugs[18]. 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[19], accessory chromosomes[20,21], and horizontal gene transfer[22]. Additionally, the compartmentalized genome architecture of these pathogens, commonly referred to as the "two-speed" genome, is thought to facilitate adaptation to stress by harboring disease associated and stress response genes in rapidly evolving regions of their genomes that contain few genes and many repetitive elements[23]. However, fungal plant pathogen genome evolution remains an active area of research. 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 these organisms.

Magnaporthe oryzae (syn. *Pyricularia oryzae*) has been described as one of the most important fungal pathogens threatening agriculture[24] and is the causative agent of the rice blast disease[25]. Its ease of culture as well as the importance of this pathogen for global food security have propelled it to being one of the most studied plant pathogens resulting in over three hundred sequenced genomes as well as transcriptomic, and epigenetic datasets in addition to robust genetic tools including CRISPR/Cas9 mediated genome editing[26]. These tools and datasets make *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 sequence 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 suggests that many genes are never found on eccDNAs, and 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[27]. Briefly, after total DNA extraction, linear DNA was degraded using an exonuclease and the remaining circular DNA was amplified using rolling circle amplification (RCA). Depletion of linear DNA was verified using qPCR and primers to the *M. oryzae* actin gene which has previously been used as a marker for linear DNA since increased copies of the ACT1 gene are thought to be deleterious in yeast[28,29] (Supplemental Figure 1). Isolated eccDNAs were then sequenced using both paired-end Illumina sequencing and PacBio circular consensus sequencing (CCS).

To identify eccDNAs from our Illumina sequencing data, we developed a pipeline inspired by previously published methods[11]. In circularome sequencing data, split mapping reads often originate from sequencing of circularization junctions of eccDNAs and opposite facing read pairs represent sequencing

from paired-end sequencing fragments that span these circularization junctions. Our pipeline uses split reads in combination with opposite facing read pairs to find evidence of eccDNA formation. This allowed us to identify, with high confidence, genomic sequences belonging to eccDNAs, which we will hereafter refer to simply as "eccDNA forming regions." We will refer to split reads associated with these eccDNA forming regions simply as "junction split reads." Across all 8 sequenced samples, our pipeline identified 1,719,878 eccDNA forming regions using our Illumina paired-end sequencing data (Supplemental Data File 1).

We also inferred eccDNA forming regions using our PacBio sequencing data. This was done using a similar pipeline with less stringent criteria which was necessary given the much lower coverage of this sequencing data. This sequencing data allowed us to identify 147,335 eccDNA forming regions across all samples (Supplemental Data File 2). We compared these eccDNA forming regions to those called using Illumina data 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 (Supplemental Figure 2). We were able to attribute much of this discrepancy to our stringent criteria for calling eccDNA forming regions since simply looking for split reads in our Illumina data increased this rate to 90.36% (Supplemental Figure 2). 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. Aside from this validation, we chose not to include the PacBio data in our final analyses due to the lower read coverage.

Next, we quantified the potential false positive rate of our pipeline by running it on previously published whole genome sequencing data from *M. oryzae*[26,30,31]. Based off the number of eccDNA forming regions called from this data, we estimated this false positive rate at less than 3 junction split reads per million sequencing reads (Supplemental Table 1). For comparison, we found 38,585 junction split reads per million reads in our samples on average which indicates that the false positive rate of our pipeline is very low. Additionally, we could not rule out the presence of eccDNAs in the whole genome sequencing data we analyzed.

Finally, we compared previously published eccDNA forming regions in human tissue[11] to the output from our pipeline using the same data. This comparison showed that, on average, 74.62% of eccDNA forming regions called by our pipeline were also found in the published dataset (Supplemental Figure 3A). 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 (see Methods) and were likely due to differences in sequence data processing and different criteria for selecting split reads[11] (Supplemental Figure 3).

However, the two lists significantly differed in the number of eccDNA forming regions identified, with our pipeline identifying substantially less (Supplemental Figure 3B). 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 is in contrast to other methods of calling eccDNA forming regions which sometimes relies purely on peaks in sequencing coverage[11,28]. This meant that our pipeline could not detect eccDNAs formed from HR between identical repeats which do not result in split mapping 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. Regardless,

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 we 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[11,27,28,32] and reanalyzed them using our pipeline (Supplemental Data Files 3-18). This analysis revealed a very large number of eccDNA forming regions in *M. oryzae* compared to other previously sequenced organisms (Figure 1B). This result was still striking after normalizing for genome size and sequencing library size (Supplemental Figure 4A). While this difference could be attributed to differences in the methods used for eccDNA purification, our own eccDNA samples extracted from *Oryza sativa* had similar levels of diversity to previously sequenced samples (Supplemental Table 2). Additionally, *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* (Figure 1B, Supplemental Figure 4B).

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

LTR/Gypsy and LTR/Copia retrotransposons frequently generate eccDNAs through several mechanisms[2–4], 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 seemed to be composed of more than 90% LTR retrotransposon sequence indicating that these elements made up a large portion of the pathogen's circularome. 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 (Figure 1B, Supplemental Figure 4C and 4D). Additionally, this analysis also showed that the epigenetic regulation mutant *A. thaliana epi12* had an increased portion of its circularome made up by LTR/Copia sequences as previously described[27] which supports our re-analysis approach (Supplemental Figure 4D).

All six LTR retrotransposons identified in *M. oryzae* Guy11 formed eccDNAs (Figure 2A). However, the elements MAGGY, GYMAG1, and Copia1 made up the majority of eccDNA sequencing data (Figure 2B). 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 (Figure 2C, Supplemental Figure 5).

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

LTR retrotransposons can form eccDNAs through a variety of mechanisms[2–4]. EccDNA formation commonly occurs after transcription and reverse transcription of the transposon which results in a linear fragment of extrachromosomal DNA[33] (Figure 1A). The most common of these mechanisms are through nonhomologous end joining (NHEJ) of the two LTR ends to form eccDNAs containing 2 LTRs (scenario 1, Figure 3A), autointegration of the retrotransposon forming single LTR eccDNAs of various lengths, depending on where in the internal sequence of the transposon the autointegrations happens (scenario 2, Figure 3B), HR between the two LTRs to form single LTR eccDNAs (scenario 3, Figure 3C). Finally, LTR retrotransposons sequences can also become part of eccDNAs by other eccDNA formation mechanisms that do not rely on transposon activity, such as intrachromosomal HR between solo-LTRs or multiple

copies of the same transposon[4,5,28]. Given this diversity of mechanisms, we wanted to identify whether there were differences across LTR retrotransposons in *M. oryzae* in their tendency to form through one mechanism or another. To this end, 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 (Figure 3A-C). Furthermore, we also measured the prevalence of scenarios 1 and 2 by identifying specific split read variants. 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 (Supplemental Figures 6 and 8A). 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 (Supplemental Figure 7, 8B). 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.

Combining and interpreting this data gave us some insights into the mechanisms through which each LTR retrotransposon forms eccDNAs. For example, the shape of the profile plot for the observed read coverage for MAGGY indicated that it forms eccDNAs primarily through autointegration (Figure 3D). This was supported by a high correlation between the number of sequencing reads and LTR-internal split reads for MAGGY (Supplemental Figure 6) and a low correlation between sequencing reads and LTR-LTR split reads (Supplemental Figure 7). The data also pointed to MGRL3 and GYMAG1 forming eccDNAs primarily through autointegration (Figure 3E, 3G, Supplemental Figures 6 and 7). Copia1, on the other hand showed a clear pattern of read coverage corresponding to eccDNA formation through HR (Figure 3F), 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 (Supplemental Figure 7). In the case of GYMAG2, its sequencing read coverage resembled a pattern expected for LTR-eccDNAs formed through NHEJ (Figure 3H). The large amount of LTR-LTR split reads found corresponding to GYMAG2 elements compared to other retrotransposons supported this inference (Supplemental Figure 8). Finally, PYRET's distinct sequencing read coverage profile likely indicated that it mostly formed eccDNAs passively (Figure 3I). 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 (Supplemental Figures 6 and 7).

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[12,34]. 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 across all organisms though this enrichment was not found in in *M. oryzae* (Supplemental Figure 9). 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 in many of the other sequenced organisms (Supplemental Figure 9 and 10). 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 (Supplemental Figures 10 and 11). Additionally, eccDNAs are often associated with active transcription[1,7], and we found a slight but significant correlation between expression and junction split reads for large eccDNAs but not for microDNAs (Supplemental Figure 12).

In yeast, eccDNA amplification is thought to often occur with the help of autonomously replicating sequences (ARSs) which contain ARS consensus sequences (ACSs)[5,28,35]. 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 (Supplemental Figure 13).

Finally, microDNAs have been found to be associated with active chromatin marks and increased GC content[12,34]. However, we did not find any of these enrichments in microDNAs or large eccDNAs in *M. oryzae* (Supplemental Figure 14).

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 in an eccDNA forming region in at least one sample. These genes included TRF1 and PTP2 which have been shown to be involved in fungicide resistance in *M. oryzae*[36,37]. However, not all genes were observed in eccDNA forming regions at the same frequency. To further understand what types of genes are enriched in eccDNA forming regions, we selected a set of eccDNA-associated genes. This set was identified by first scoring each gene by the number of junction split reads that fully encompassed that gene. 986 genes that were in the top 10% of genes for this score were selected (Figure 4A, Supplemental Data File 19). We referred to this score as the number of “encompassing split reads” for each gene. We then performed gene ontology (GO) enrichment analysis and found that eccDNA-associated genes were enriched for GO terms related to vesicle trafficking among other terms (Figure 5A, Supplemental Figure 15, Supplemental Data Files 20-22).

EccDNA-associated genes are found in gene sparse and repeat dense regions of the genome

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. 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 (Figure 6). We observed a significant difference (permutation test for difference of medians, $p < 0.0001$, n = 10,000 replicates) between the median distance to the nearest gene of eccDNA-associated genes (593 base pairs) and other genes (478 base pairs). We also observed a significant difference (permutation test for difference of medians, $p = 0.0007$, n = 10,000 replicates) between the median distance to the nearest genomic repeat of eccDNA-associated genes (649 base pairs) and other genes (776 base pairs, Supplemental Figure 16). This difference in proximity was not observed for transposable elements specifically indicating that transposable elements alone were not responsible for this effect (Supplemental Figure 17).

EccDNA-associated genes are prone to presence-absence variation

There is evidence of eccDNAs generating structural variation in other organisms[13,14]. We therefore tested whether eccDNA formation is associated with genes prone to presence-absence variation in 162 rice-infecting *M. oryzae* isolates (Supplemental Data File 25). As expected from previous studies[38,39], our analysis indicated that predicted effectors were more likely to experience presence-absence variation (Supplemental Figure 18; 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 (Figure 4B and Supplemental Data File 19; X-squared = 28.511, df = 2, p-value = 6.44e-07). 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 (Supplemental Data File 24). However, none of these deletions matched our eccDNA forming regions and only 8 of them came within 50 bp. This result was unlikely to be conclusive though given that rarefaction analysis for both large eccDNAs and microDNAs revealed that there is likely to be a much greater diversity of these molecules than what we were able to capture at the sequencing depth of this study (Supplemental Figure 19).

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 pipeline that identified one previously described eccDNA-mediated translocation in wine yeast[14] (Supplemental Figure 20), we were unable to identify any such examples in any of the *M. oryzae* genomes we analyzed despite including isolates infecting a variety of hosts in this analysis (306 genomes in total, Supplemental Data File 25).

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 (Supplemental Figure 21).

Many genes that are never found on eccDNAs are cytoskeleton-related

Since most *M. oryzae* genes appeared in eccDNA forming regions in at least one sample, we were particularly interested in the 2,249 eccDNA-absent genes (Supplemental Data File 19). We first verified that eccDNA-absent genes were not due to insufficient sequencing coverage using rarefaction analysis. Given the high coverage of our data, we only expected to find approximately 468 genes in this category by chance (Figure 4C).

We next explored whether gene expression or other genomic markers could explain the genes in this category. However, we found no difference between eccDNA-associated genes and eccDNA-absent genes in gene expression, GC content, or histone marks aside from a slight difference in H3K36me3 (Supplemental Figure 22 and 23).

Finally, we performed GO enrichment analysis on these genes and found, amongst many other enriched terms, that terms related to cytoskeletal proteins were enriched within eccDNA-absent genes (Figure 6B, Supplemental Figure 24, Supplemental Data Files 26-28). This result is of particular interest given that cytoskeletal genes, specifically the actin gene, have been used in a previous study[28] as well as this

one, as a marker for linear DNA due to its negative fitness effect at high copy numbers in yeast[29]. This suggested 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

Next, 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* (Figure 7A, Supplemental Data File 19). Additionally, we found that many predicted effectors were found in eccDNA forming regions (Figure 7B, Supplemental Data File 19). We also found that many of these putative effectors were associated with large numbers of encompassing split reads and found this difference to be statistically significant (Figure 7C; 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, 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 effect is not simply due to their size.

Discussion

EccDNAs have been shown to be a potential source of significant phenotypic[5–7,9,10] and genotypic[13,40] plasticity that can help organisms adapt to stress. While eccDNAs have been extensively studied in human cancer[1], 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, through this analysis, 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. We also found that eccDNA forming regions in *M. oryzae* were more commonly made up LTR retrotransposons than other organisms. These differences highlight the need to further characterize the circularome of other organisms as the current datasets are clearly not representative of all 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*. 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 *M. oryzae*. Additionally, 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 including inducing eccDNA formation in the lab 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.

We found some evidence that suggests that eccDNA formation could be a potential mechanism for phenotypic plasticity in *M. oryzae*. For example, we identified two genes associated with fungicide resistance in our eccDNA forming regions. 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[23]. The fact that eccDNA-associated genes were closer to repeats than other genes but not transposons specifically indicates that this effect was not simply caused by eccDNA formation by LTR retrotransposons. We also specifically found that predicted effectors were more commonly found fully encompassed by eccDNA forming regions than expected by chance. This result could be simply driven by the genomic context in which eccDNAs are commonly formed but could also point to eccDNAs playing a role in the evolution of these important genes.

Additionally, we identified a set of eccDNA-absent genes that were not explained simply by incomplete sequencing. The existence of this group of genes hints at selective pressure acting on eccDNA formation and preventing genes that are deleterious at increased copy numbers, such as cytoskeletal proteins, from accumulating high copy numbers through eccDNA formation. On the other hand, selective pressure caused by growth under stress conditions 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[5–7,9,10]. This could be caused either by extrachromosomal replication of eccDNAs, which the enrichment of ACSs in our eccDNA forming regions hinted at, or by high rates of formation either through random asymmetric segregation of eccDNAs at mitosis, as has been shown in cancer[10], or through retention of eccDNAs in ageing cells as has been shown in yeast[8]. 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. We attempted to perform such experiments by sequencing *O. sativa* tissue infected by *M. oryzae* but found that rice eccDNAs crowded out the circularome sequencing signal and prevented meaningful analysis.

In conclusion, this study begins 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 present in all other rice infecting isolates that we analyzed and that were in the eccDNA-absent group. These genes could be high potential targets for drug design or engineered resistance. Nevertheless, eccDNAs remain largely underexplored, especially in fungal plant pathogens, and, given their potential, further characterization of this circularome, especially under stress, is likely to prove valuable. Furthermore, this study presents a characterization of the circularome of a single fungal plant pathogen, and it will be important to perform a broader survey of these organisms to fully understand the role of eccDNAs in their adaptation to stress.

Methods

M. oryzae growth 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 15 ml complete medium[41] (CM) in petri dishes. 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 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 each of three biological replicates, the first true leaf was cut from one rice plant, its tip was removed, and it was 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 then placed in an airtight container with wet paper towels and 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 extraction was then performed using the Qiagen Plant DNeasy mini kit.

Circular DNA enrichment

Total DNA obtained from biological replicates was then split into three technical replicates before circular DNA enrichment. This enrichment was performed according to a protocol from Lanciano et al. with a few modifications[27]. To purify the samples and begin removing large linear DNA fragments, the samples were treated using a Zymo Research DNA Clean and Concentrator kit and 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 while the incubation continued. 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

ul 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 ul 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. Second, 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 5'-GTATGTGCAAGGCCGGTTTC-3' and 5'- GCACATCTGTCGACAAACCG-3' were used to amplify the *M. oryzae* actin gene (MGG_03982) along with Lightcycler 480 Sybr Green I master mix. 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. 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 60-minute sequencing strategy.

Inferring eccDNA forming regions from short read sequencing data

Illumina sequencing signal was analyzed using a custom pipeline inspired by previously published methods[11]. Illumina reads were first trimmed of Illumina TruSeq adapters using CutAdapt[42] version 2.4

with the nextseq-trim=20 option. Trimmed reads were then mapped to the *M. oryzae* Guy11 genome[43] and the 70-15 mitochondrial sequence[44] obtained from the Broad Institute (https://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/magnaporthe-comparative-genomics-proj) using BWA-MEM[45] 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, mapped to only two places in the genome, and where the start of 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 where one side 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 (Supplemental Figure 25A) was then used to probabilistically infer candidate eccDNA forming regions from multi-mapping reads (Supplemental Figure 25B). 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 their 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 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

Circular consensus sequences (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/). CCSs were then mapped to the *M. oryzae* Guy11 genome using minimap2[46] version 2.18-r1015. Only uniquely mapped reads were kept for analysis. We then identified eccDNA forming regions by looking for split reads that either mapped to the same orientation to the same exact region multiple times or pairs of split alignments that were less than 50 kb apart, mapped in the same orientation and oriented properly so that they were indicative of a circular junction rather than a deletion.

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. 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 were less than 50kb apart were considered eccDNA forming regions.

Comparing the eccDNA forming regions called using our pipeline to a previously published pipeline

EccDNA forming regions called using our pipeline were compared to eccDNA forming regions previously published for *H. sapiens*[11]. 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.

Gene annotation

The *M. oryzae* Guy11 genome along with 162 other rice-infecting *M. oryzae* genomes (Supplemental Data File 23) were annotated using the FunGAP[47] 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 (accessions GCA_000002495.2, GCA_000292605.1, and GCA_004346965.1, respectively) taken from GenBank. 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[48] fngrep version 25.10 with a *de novo* repeat library generated by RepeatModeler[49] version 2.0.1 run on the *M. oryzae* Guy11 genome with the LTRStuct argument 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. To make this assignment, *M. oryzae* Guy11 proteins were aligned to the *M. oryzae* 70-15 proteome using BLASTP[50] version 2.7.1+ and hits with greater than 80% sequence identity and that spanned more than 80% of the length of both the *M. oryzae* Guy11 protein and the *M. oryzae* 70-15 protein were kept.

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

High quality, full length, consensus sequences for known LTR/Gypsy elements in *M. oryzae* (MAGGY, GYMAG1, GYMAG2, PYRET, MGRL3) and one LTR/Copia element (Copia1) were generated using the WICKERsoft[51] suite of tools. Reference sequences from other genomes for each element were obtained from the RepBase[48] fngrep version 25.10 library. The *M. oryzae* Guy11 genome was then screened for the presence of these sequences using BLASTN[50] 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[52] 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[50] webserver in August 2020 to identify LTRs. These consensus sequences are available in Supplemental Data File 29. Finally, the locations of these elements in *M. oryzae* Guy11 genome were annotated with RepeatMasker[53] 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 filtered 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 (Supplemental Data File 30). 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.

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.

To simulate expected read coverage for different types of LTR eccDNAs, the Copia1 consensus sequence was taken as a reference, though other consensus sequences 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, Figure 3A) was simply made up of two LTR sequences and the internal sequence, and the expected 1-LTR circle sequence generated by HR (scenario 3, Figure 3C) 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, Figure 3B), 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[54] 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[45] 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[55] suite of tools version 3.5.1 using full length, high confidence LTR retrotransposon sequences.

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.

Analysis of eccDNA formation from specific genomic regions

Genome, gene annotation, and transposable element annotation files for each organism used for this analysis were as previously described (Supplemental Data File 30). 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[56] 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 fold enrichment was calculated by taking the log 2 of the observed average percentage over the expected average percentage.

Expression and junction split read counts per gene

Previously published RNAseq data from *M. oryzae* Guy11 grown in liquid culture in rich medium was obtained[57] (Supplemental Data File 31). The data was mapped to the *M. oryzae* Guy11 genome using STAR[58] 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 technical replicate 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.

Genes and junction split read counts per 100kbp

The *M. oryzae* genome was divided into 100kbp bins and the number of genes per bin was then 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 then averaged across technical replicates for each biological replicate, and then averaged across biological replicates.

ACS enrichment analysis

The published ACS sequence profile[35] was used to identify ACSs in eccDNA forming regions using the FIMO[59] 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 was obtained[57]. 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[45] version 0.7.17-r1188 with default settings. The bamCompare command from the DeepTools[55] suite of tools version 3.5.1 with the scaleFactorsMethod readCount option was then 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 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*[57]. 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.

Encompassing split read counts per gene

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[60] 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, then averaged across biological replicates to obtain the final encompassing split read count per gene.

GO enrichment analysis

GO terms were first assigned to annotated *M. oryzae* Guy11 genes using the PANNZER2[61] webserver on August 17th, 2020. The topGO[62] R package version 2.36.0 was then 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 then 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 then 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[63] 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[50] version 2.7.1+.

Effector annotation

Effectors were predicted among *M. oryzae* Guy11 genes by first selecting genes with signal peptides which were predicted using SignalP[64] version 5.0b Darwin x86_64. Genes with predicted transmembrane domains from TMHMM[65] version 2.0c were then excluded. Finally, EffectorP[66] version 2.0 was used to predict effectors from this secreted gene set. Previously well-characterized effectors were identified using previously published protein sequences[38] and DIAMOND[67] version 2.0.9.147.

Identification of small, genic deletions

Small, genic deletions were identified using orthologs identified by OrthoFinder[63] 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[68] suite of tools version 4.0.0rc1 to verify that a DNA deletion truly existed, and that this deletion overlapped the gene of interest.

693 **Identification of eccDNA-mediated translocations**

694 Identification of translocations with a potential eccDNA intermediate was done by first aligning two
695 genomes using the nucmer command of the MUMmer[68] suite of tools version 4.0.0rc1 with the
696 maxmatch option. The nucmer output was then parsed to look for portions of the reference genome
697 that had an upstream region that aligned to one query scaffold, followed by two separate adjacent
698 alignments to another query scaffold, followed by a downstream region that aligned to the original
699 query scaffold. We also required that the two adjacent alignments in the center of the region were to
700 adjacent regions in the query scaffold but their order was reversed compared to the reference.
701 Candidate eccDNA-mediated translocations were verified manually by inspecting alignment plots
702 generated using the mummerplot command. The *S. cerevisiae* EC1118 (GCA_000218975.1) and M22
703 genomes (GCA_000182075.2) obtained from GenBank were used to verify the ability of our pipeline to
704 detect these translocation events. The *M. oryzae* Guy11 genome was then compared to 306 *M. oryzae*
705 genomes (Supplemental Data File 25) to look for these events in the *M. oryzae* species. Before
706 alignment, transposable elements were masked from these *M. oryzae* genomes using RepeatMasker[53]
707 version 4.1.1 with the -cutoff 250, -nolow, -no_is, and -norna options, as well as a transposable
708 elements library generated by combining the RepBase[48] fngrep version 25.10 with the de novo repeat
709 library generated by RepeatModeler[49] version 2.0.1 run on the *M. oryzae* Guy11 genome with default
710 settings aside from the -LTRStuct argument.

711 **Minichromosome genes and eccDNAs**

712 Scaffolds corresponding to minichromosomes in the *M. oryzae* FR13 (GCA_900474655.3), CD156
713 (GCA_900474475.3), and US71 (GCA_900474175.3) genomes were extracted according to previously
714 published information[69]. Exonerate[70] version 2.4.0 was then used with the protein2genome model to
715 identify genes in the *M. oryzae* Guy11 genome that were found on minichromosomes in these other
716 isolates. Hits with greater than 70% sequence identity to any minichromosome scaffold were identified
717 as genes found on minichromosomes. Encompassing split reads were then counted for all genes. This
718 count was normalized to total number of junction split reads in each sample, then averaged across
719 technical replicates for each biological replicate, then averaged across biological replicates. Finally,
720 normalized encompassing split read counts for genes found on minichromosomes were compared to
721 genes not found on minichromosomes.

722 **Rarefaction analysis**

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

732 **Code and Code availability**

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[71] 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[60] version 2.28.0 and bedGraphToBigWig version 4 (https://www.encodeproject.org/software/bedgraphtobigwig/). SAM and BAM format files were processed with SAMtools[72] 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[73] version 0.25.1 and numpy[74] version 1.17.2 modules. The scipy[75] 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[76] version 1.13.6, tidyr[77] version 1.1.3, reshape2[78] version 1.4.4, and dplyr[79] version 1.0.4 packages. Plotting was performed using the ggplot2[80] version 3.3.5 package, with help from RColorBrewer[81] version 1.1.2, scales[82] version 1.1.1, cowplot[83] version 1.1.1, ggprepel[84] version 0.9.1 and ggpubr[85] version 0.4.0 packages. The Gviz[86] version 1.28.3 was used for BAM file visualization. Supplemental tables were made using gt[87] version 0.3.1.

Code for the pipeline used to call eccDNA forming regions for Illumina sequencing data is available in this 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 this GitHub repository: https://github.com/pierrj/moryzae_eccdnas_manuscript_analysis_and_plots

Data availability

All data in this study has been submitted to NCBI's Sequence Read Archive and Zenodo. 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.

Genomes and annotation files used for comparative circularome analysis are available here:

https://zenodo.org/record/5544950#.YVtyHe3MKUk

Annotated genes and predicted proteins for rice-infecting *M. oryzae* isolates are available here:

https://zenodo.org/record/5542597#.YVtxvO3MKUk

Output from OrthoFinder2 run on rice-infecting *M. oryzae* proteomes is available here:

https://zenodo.org/record/5544260#.YVtyWe3MKUk

All files used for statistical analysis and plotting are available here:

https://zenodo.org/record/5565321#.YWXknGLMKUk

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 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. Pierre M. Joubert has been supported by the Grace Kase graduate fellowship and Ksenia V. Krasileva has been supported by funding from Innovative Genomics Institute as well as the Gordon and Betty Moore Foundation (8802).

Figures

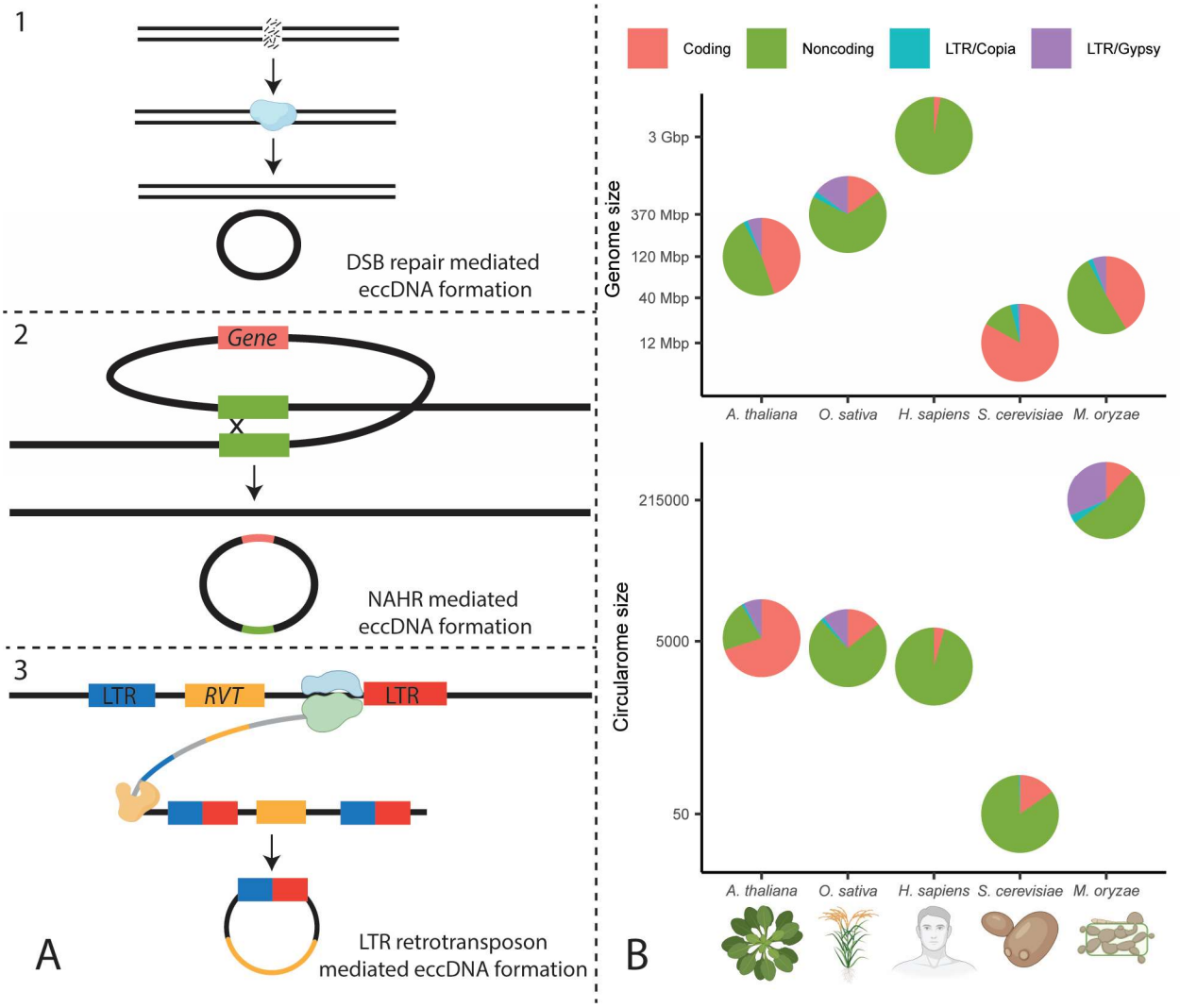


Figure 1. A. Examples of mechanisms of extrachromosomal circular DNA (eccDNA) formation. 1. eccDNA formation as a result of double strand break repair. The blue enzyme represents several different types of DNA repair mechanisms 2. eccDNA formation as a result of nonallelic homologous recombination (NAHR). The green boxes represent homologous sequences. 3. eccDNA formation as a result of LTR retrotransposon activity. The blue and green enzyme represents the ribosome, 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 circularome size for *Arabidopsis thaliana*, *Oryza sativa*, *Homo sapiens*, *Saccharomyces cerevisiae* and *Magnaporthe oryzae.* Circularome size is shown as the number of eccDNA forming regions called by our pipeline in an average sample. Only circularome data for *A. thaliana* and *O. sativa* leaf tissue, *H. sapiens* muscle tissue, and *S. cerevisiae* deletion collection samples are shown. Additionally, only *A. thaliana* circularome data from Wang et al. 2021 and *O. sativa* data from Lanciano et al. 2017 are shown. The organism icons were created with BioRender.com.

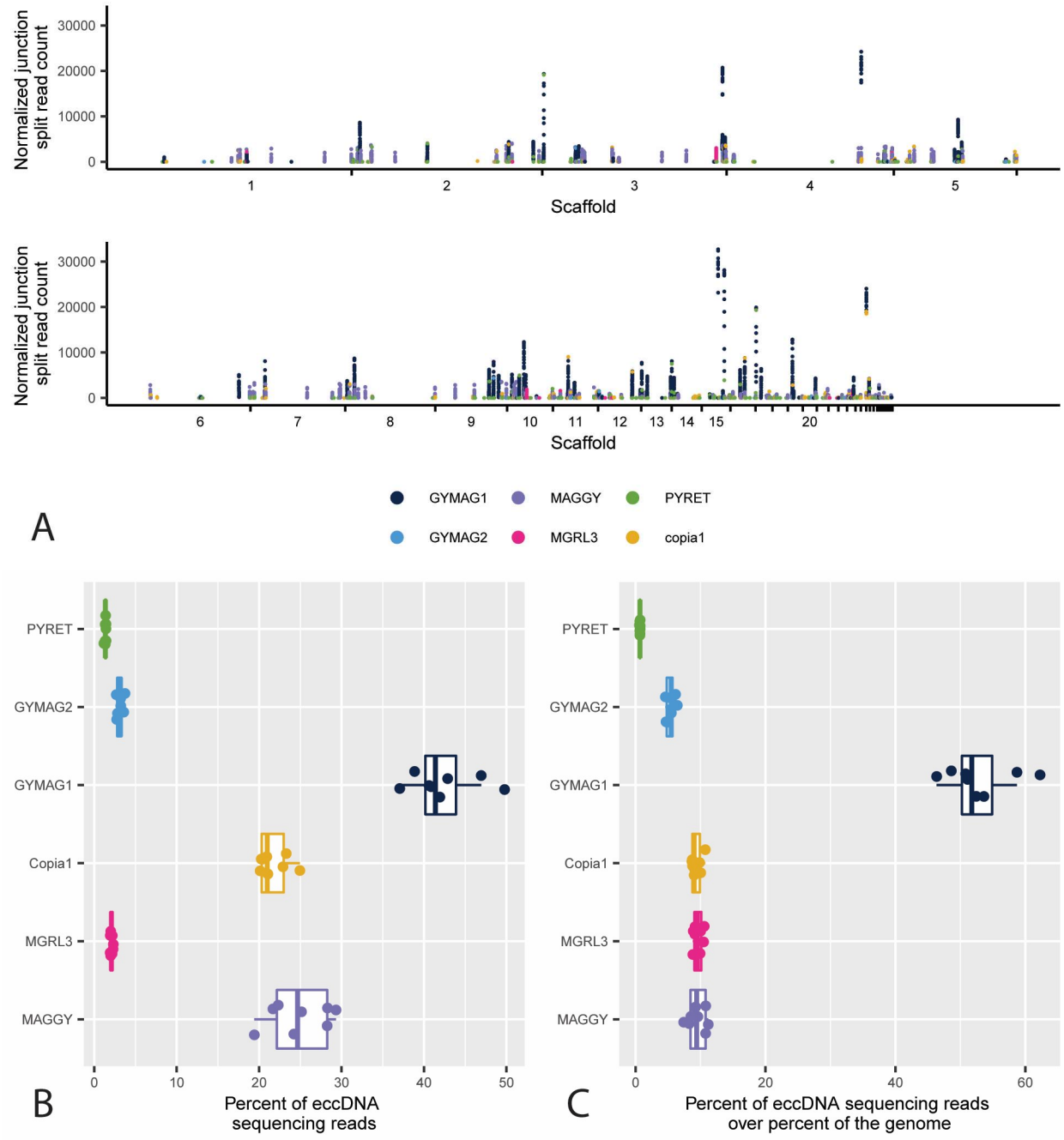


Figure 2. A. Manhattan plot showing the number of junction split reads per million averaged across all sequenced samples for all 100 bp bins that overlap an LTR retrotransposon. Each dot represents one of these bins. B. Boxplot showing the percentage of sequencing reads that map to LTR retrotransposons. Each dot represents one sequenced sample. 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 dot represents one sequenced sample.

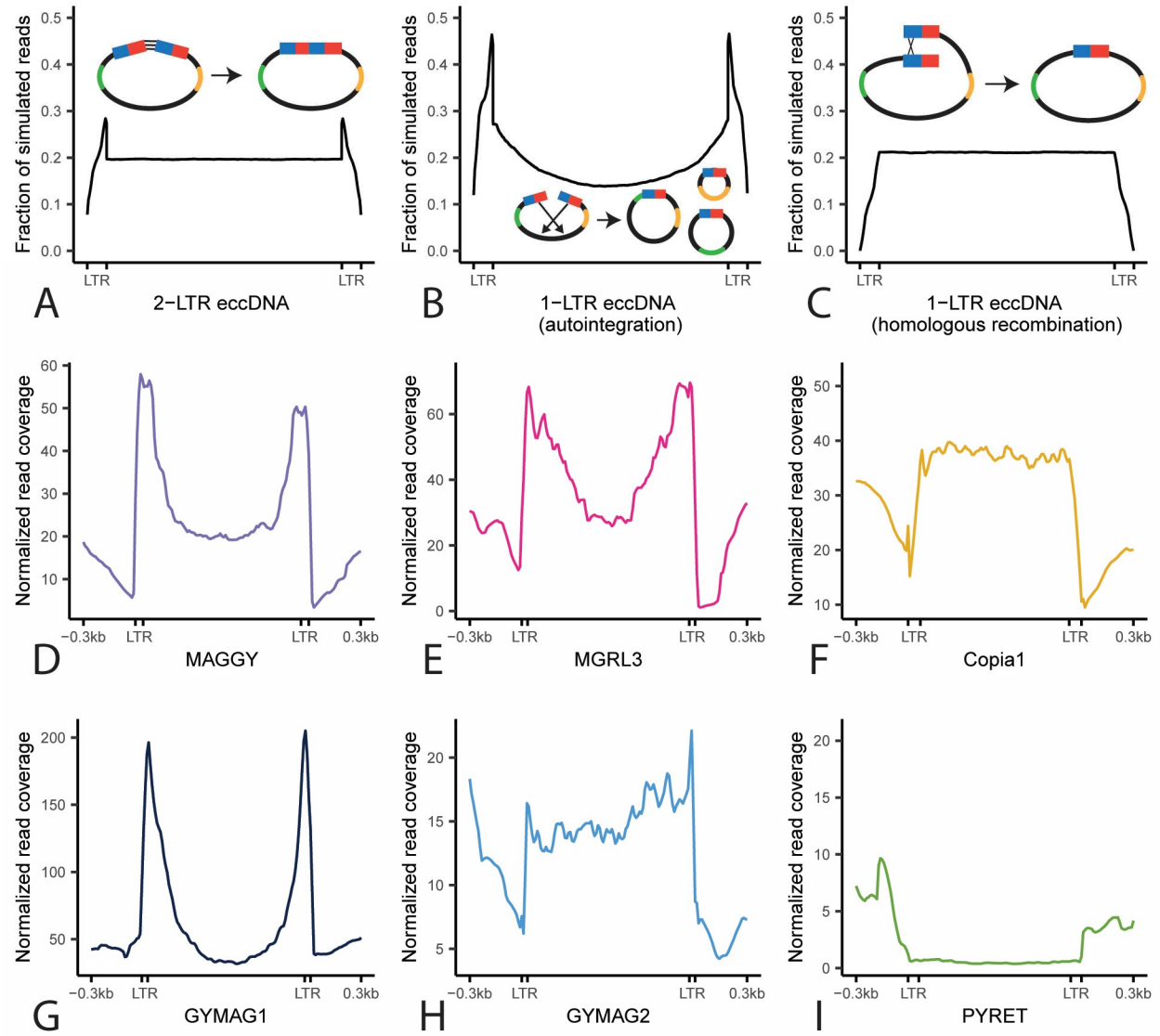


Figure 3. A-C. Profile plots showing expected sequencing read coverage for each LTR retrotransposon eccDNA formation scenario as well as graphical representations of the scenario. In the graphics, blue and red rectangles represent hybrid LTRs formed from 5' and 3' LTRs during retrotransposition and green and orange 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.

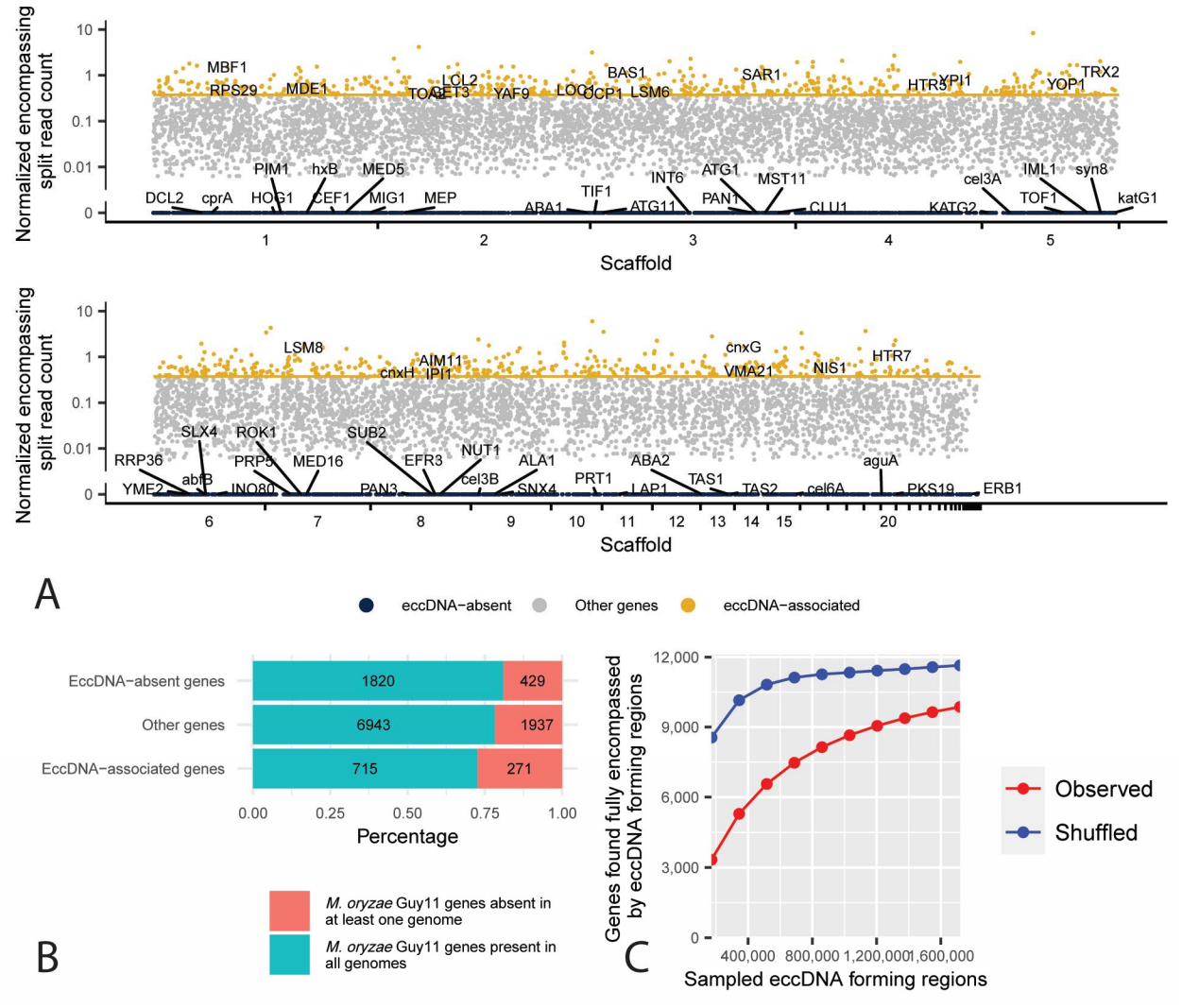


Figure 4. A. Manhattan plot showing the number of encompassing split reads per million junction split reads averaged across all sequenced samples for each gene in the *M. oryzae* Guy11 genome. Each dot represents one gene. The orange line represents the cutoff between eccDNA-associated genes and other genes. 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.

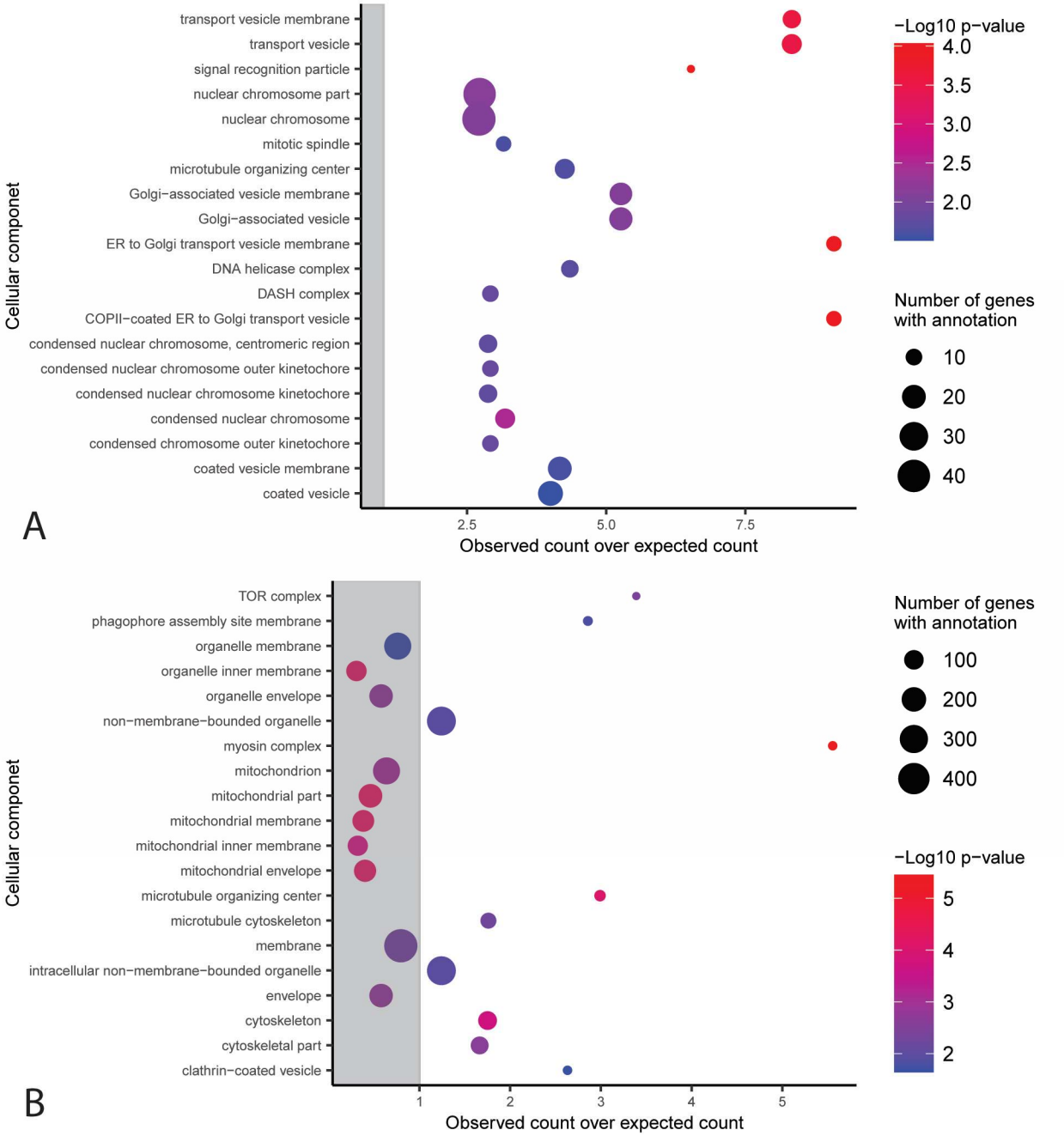
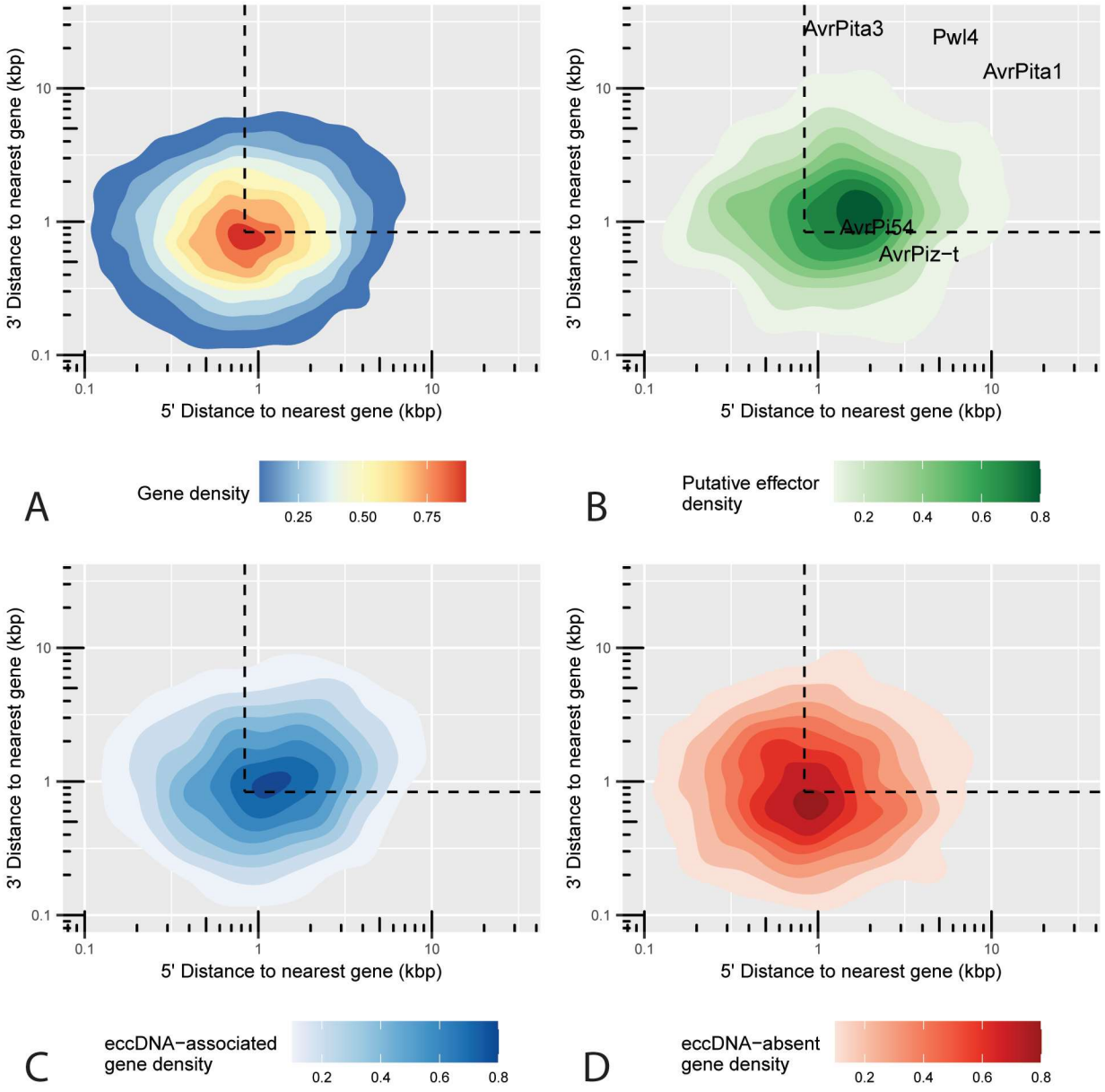


Figure 5. Functional categories in the cellular component Gene Ontology with an observed number of A. eccDNA-associated genes or B. eccDNA-absent genes that is significantly different from the expected number with correction for gene length bias (Chi-square test, $p < 0.05$). The y-axis shows the different 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.

834

835



836

837 **Figure 6.** Two-dimensional density plot representing the 5’ and 3’ distance to the nearest gene in the *M.*
838 *oryzae* Guy11 genome in kilobase pairs for each A. gene, B. predicted effector, C. eccDNA-associated
839 genes, and D. eccDNA-absent genes. Known effectors are shown as text in B. Dashed lines represent
840 median 5’ and 3’ distance to nearest gene.

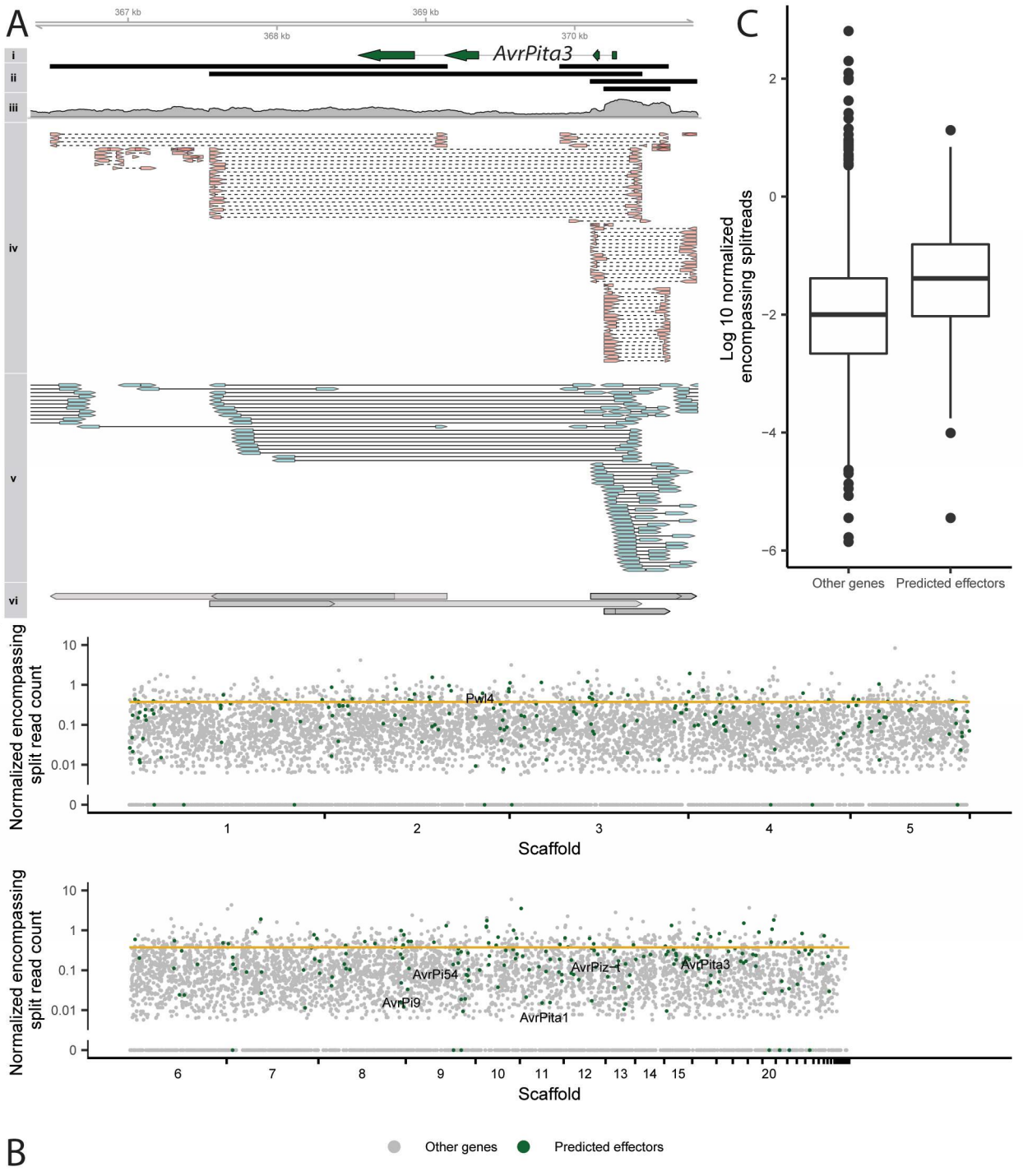
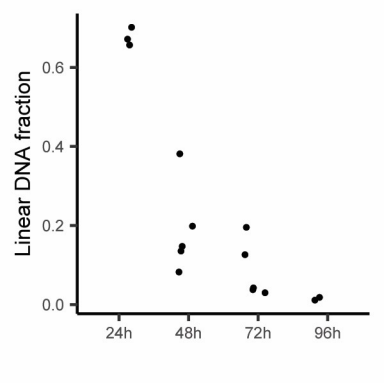


Figure 7. A. Example of an effector found on an eccDNA. i. location of effector *AvrPita3*. ii. Location of eccDNA forming regions. iii. Overall sequencing read coverage. iv. Junction split reads. Split reads are joined by a dashed line. v. Opposite facing read pairs. Read pairs are joined by a solid line. vi. PacBio split reads. Overlaps are from single reads mapped to the same location more than once. All data was obtained from a single sequenced sample. B. Manhattan plot showing the number of encompassing split reads per million averaged across all sequenced samples for each gene in the *M. oryzae* Guy11 genome.

848 Each dot represents one gene. The orange line represents the cutoff for eccDNA-associated genes.
849 Known effectors are shown as text. C. Box plot showing the log 10 of the number of fully encompassing
850 junction split reads per million junction split reads averaged across all sequenced samples for predicted
851 effectors compared to all other genes.

852

853 **Supplemental Figures**

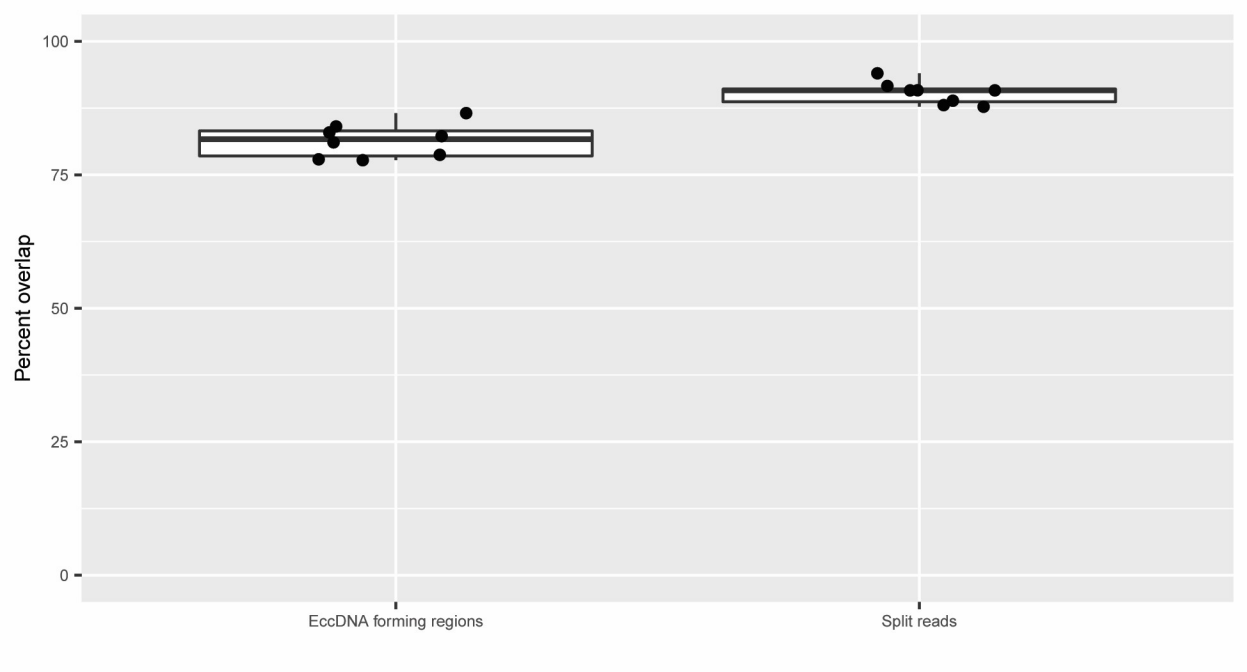


854

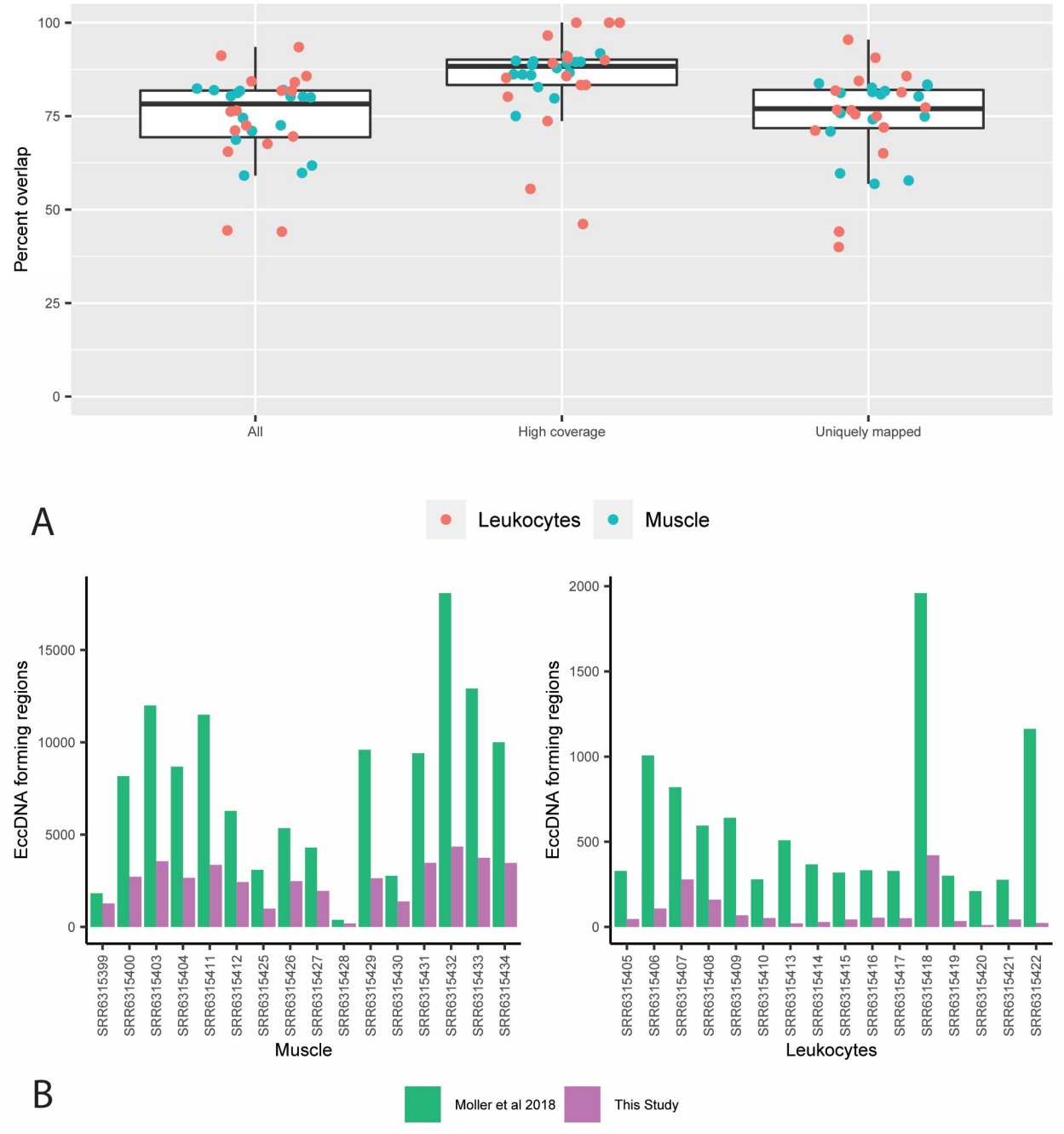
855 **Supplemental Figure 1.** Scatter plot showing the effect of exonuclease treatment on linear DNA fraction
856 of total extracted DNA from *M. oryzae* tissue samples. Each dot represents one biological replicate
857 averaged across four technical replicates.

858

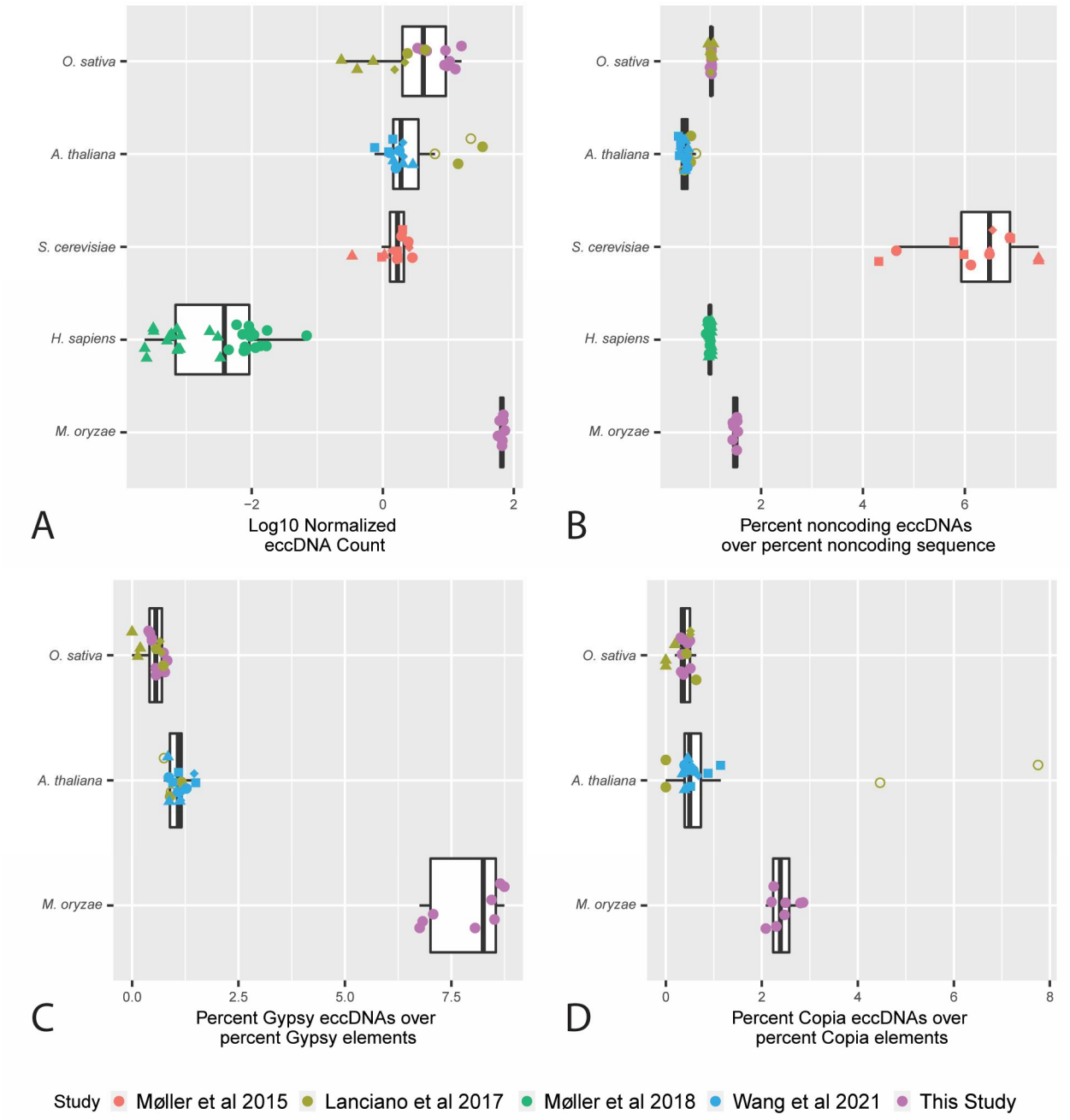
859



Supplemental Figure 2. 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.



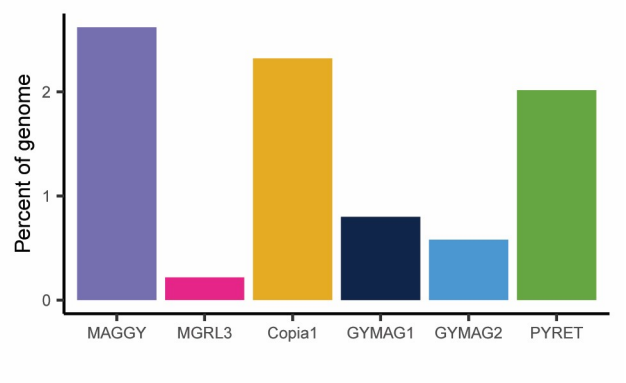
Supplemental Figure 3. A. Boxplot showing the percentage of eccDNA forming regions that were found using our pipeline that were also found in the published eccDNA forming regions for human samples. Each dot represents one sample. B. Bar plot showing counts of eccDNA forming regions generated from our pipeline compared to counts in published data. Sample IDs are from the Sequence Read Archive (SRA).



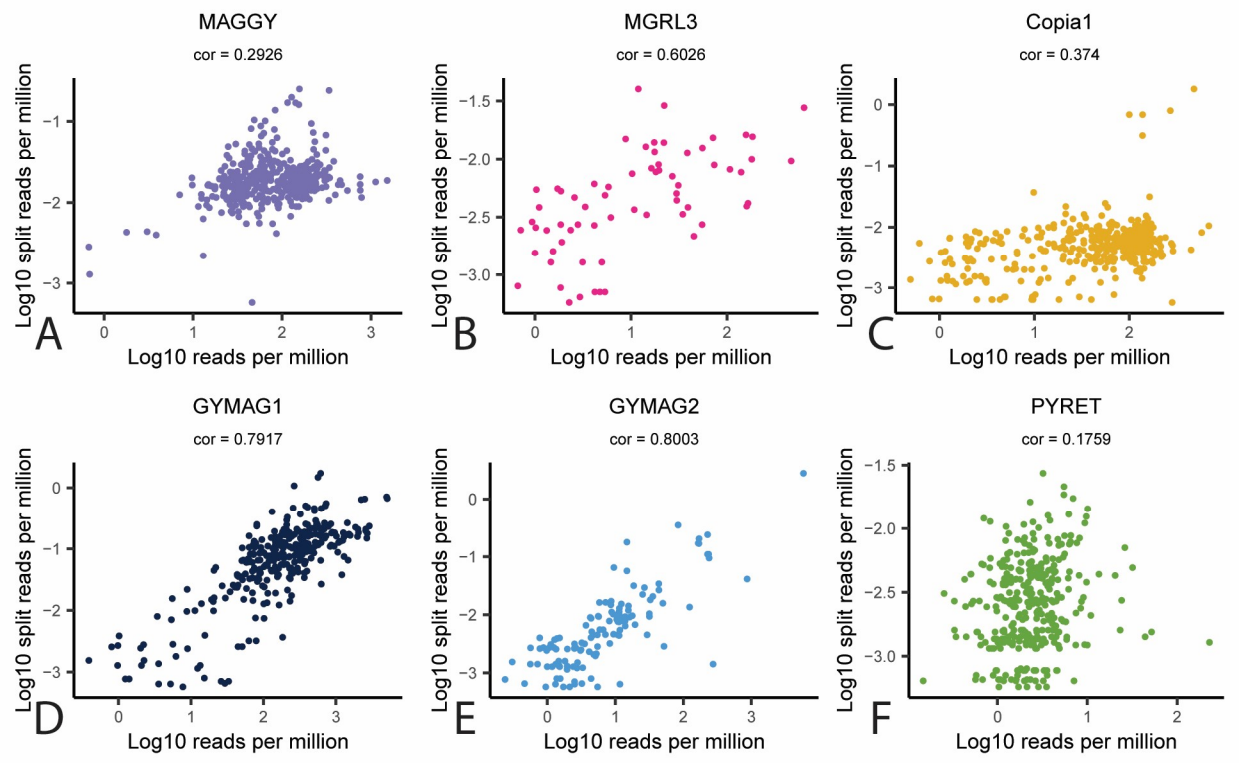
872

873 **Supplemental Figure 4.** Box plots comparing A. log 10 count of eccDNA forming regions normalized to
874 genome size and sequencing library size, B. percent of eccDNA forming regions that contain more than
875 50% noncoding sequences divided by percent of the genome made up of noncoding seqeuence, C.
876 percent of eccDNA forming regions that contain more than 90% LTR/Gypsy retrotransposon sequence
877 divide by percent of the genome made up of LTR/Gypsy retrotransposon sequence, D. percent of
878 eccDNA forming regions that contain more than 90% LTR/Copia retrotransposon sequence divided by
879 percent of the genome made up of LTR/Gypsy retrotransposon sequence across multiple organisms and
880 studies. Each dot represents one sequenced sample. Shapes represent variations in sample type within
881 the same organism. For *Oryza sativa*, circles represent leaf samples, triangles represent callus samples

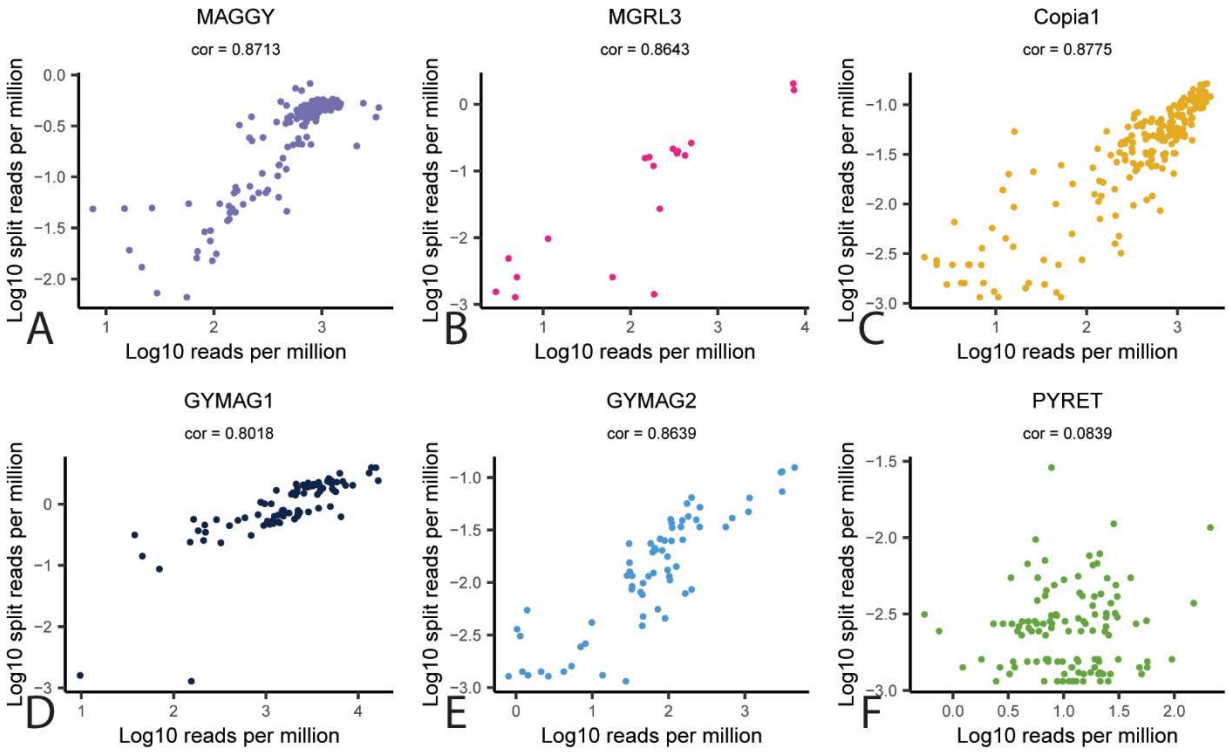
and diamonds represent seed samples. For *Homo sapiens*, circles represent muscle samples and triangles represent leukocyte samples. For *Arabidopsis thaliana*, circles represent wild type flower samples, empty circles represent *epi12* mutant flower samples, squares represent root samples, diamonds 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[88] and *S. cerevisiae* samples were excluded due to a small number of eccDNA forming regions containing retrotransposon sequences.



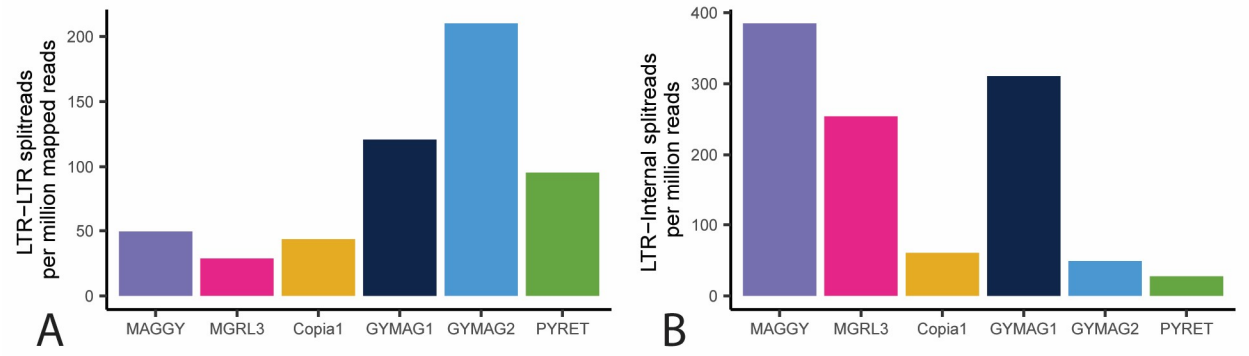
Supplemental Figure 5. Bar plot showing percentage of the *M. oryzae* Guy11 genome made up of each LTR retrotransposon.



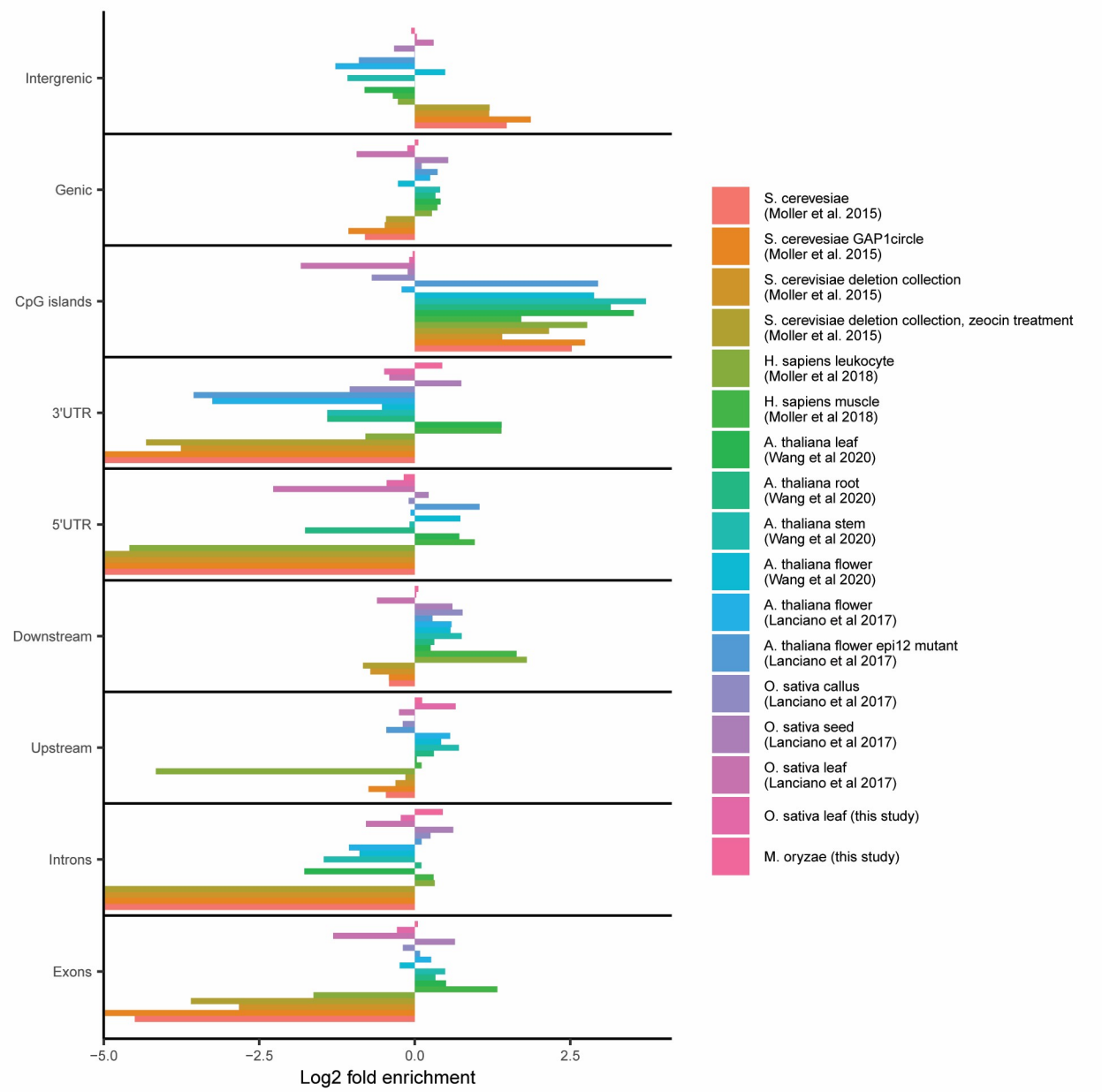
Supplemental Figure 6. A-F. Scatter plots showing Pearson's correlation coefficient between log 10 sequencing reads per million reads and log 10 LTR-LTR split reads per million sequencing reads, averaged across all sequenced samples. Each dot represents one annotated portion of the LTR region of an LTR retrotransposon.



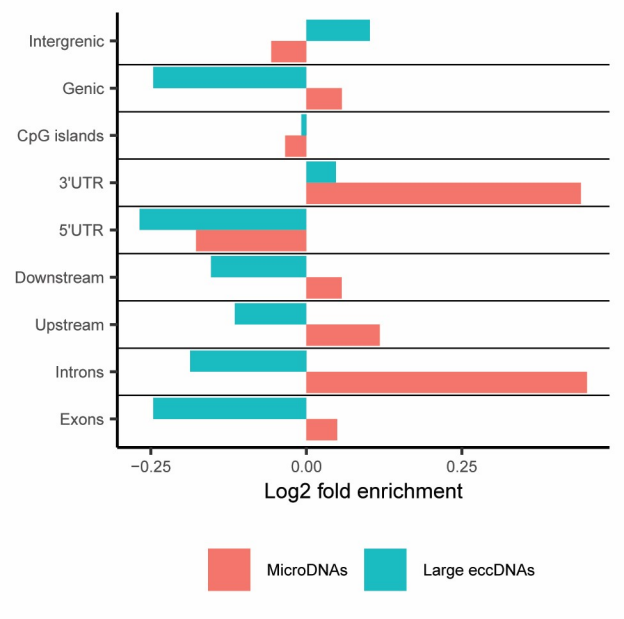
Supplemental figure 7. A-F. Scatter plots showing Pearson's correlation between log 10 sequencing reads per million reads and log 10 LTR-internal split reads per million sequencing reads, averaged across all sequenced samples. Each dot represents one annotated portion of the internal region of an LTR retrotransposon.



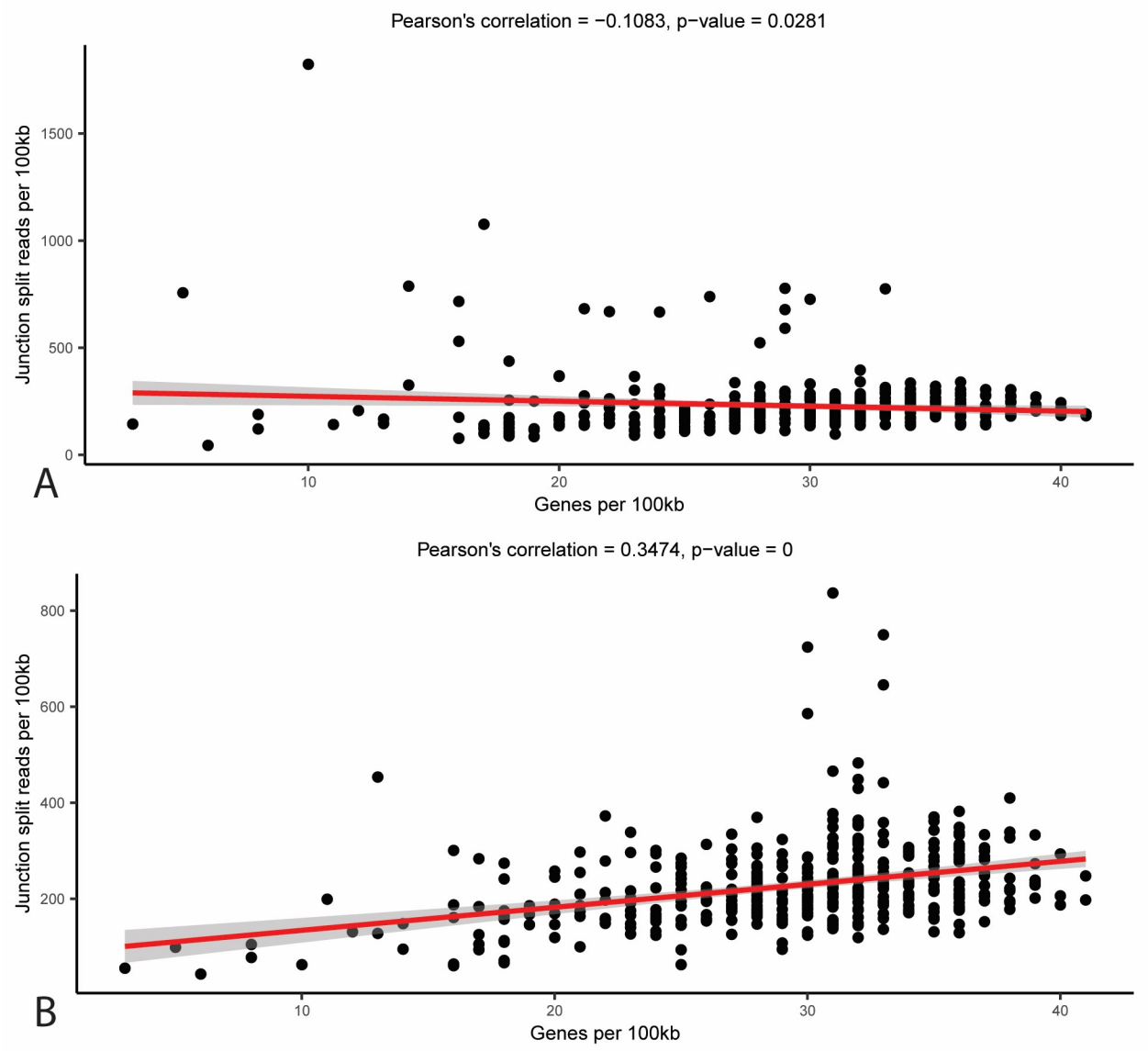
Supplemental Figure 8. A. Bar plot showing identified LTR-LTR split reads per million reads, averaged across all sequenced samples for each LTR retrotransposon in the *M. oryzae* Guy11 genome. B. Bar plot showing identified LTR-internal split reads per million reads, averaged across all sequenced samples for each LTR retrotransposon in the *M. oryzae* Guy11 genome.



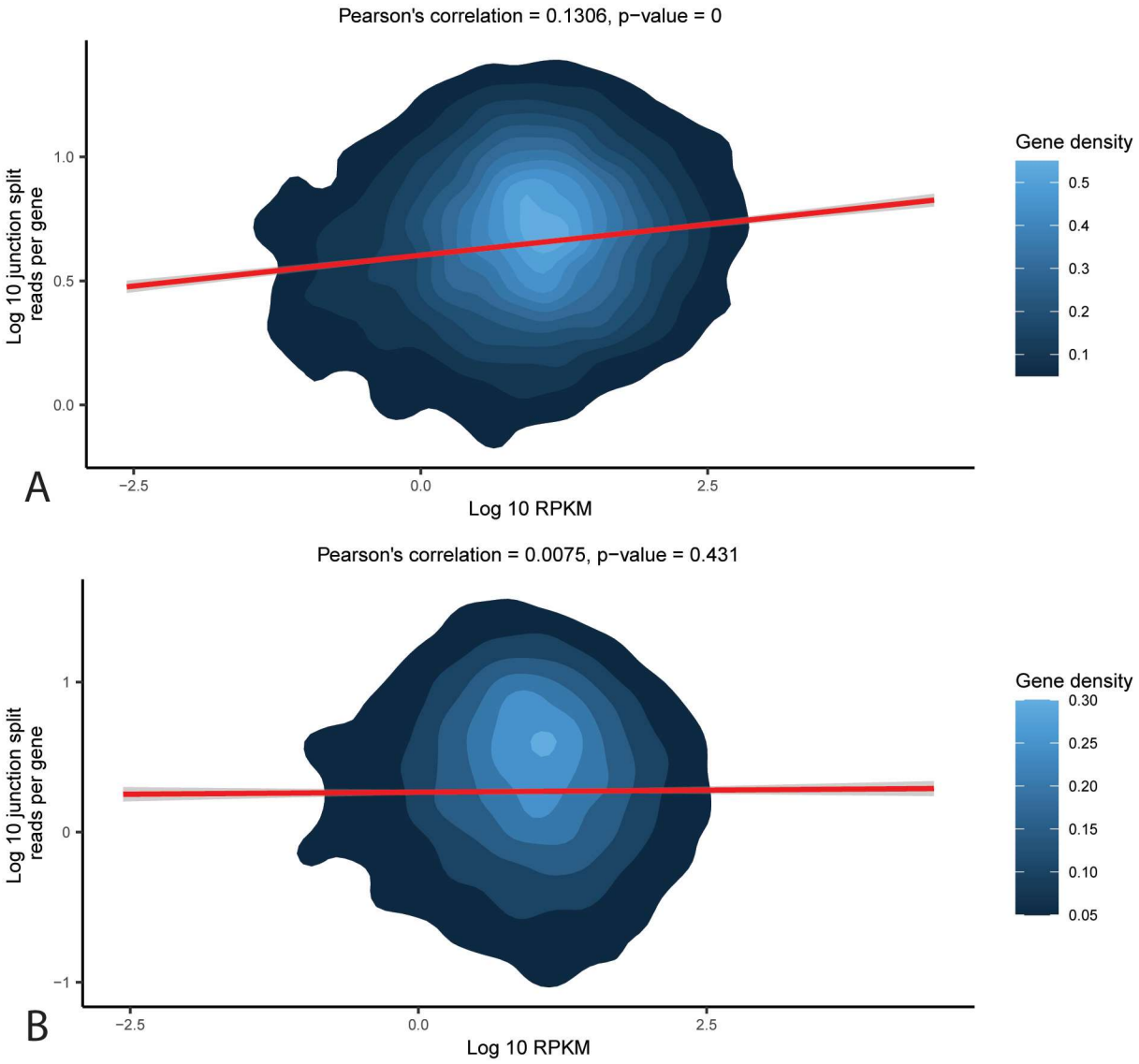
Supplemental Figure 9. Bar plot showing observed enrichment of microDNAs across various regions of the genome across different previously sequenced organisms and sample types. Log2 fold enrichment of -5 represents samples where no microDNAs were found in that region. The presented data is an average of all sequenced samples of each type.



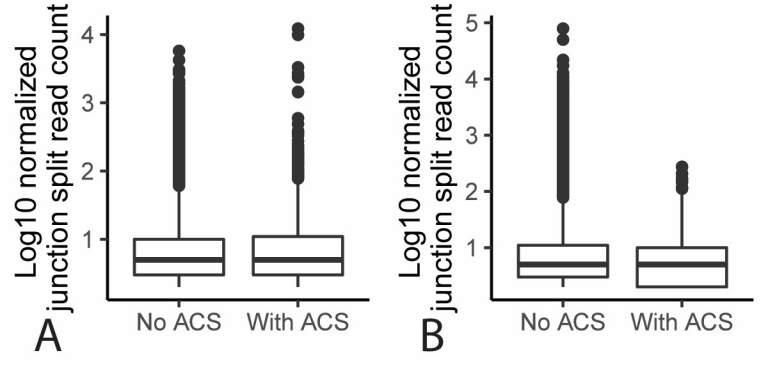
Supplemental Figure 10. Bar plot showing observed enrichment of microDNAs and large eccDNAs across various regions of the genome. The presented data is an average of all sequenced samples from *M. oryzae* Guy11 in this study.



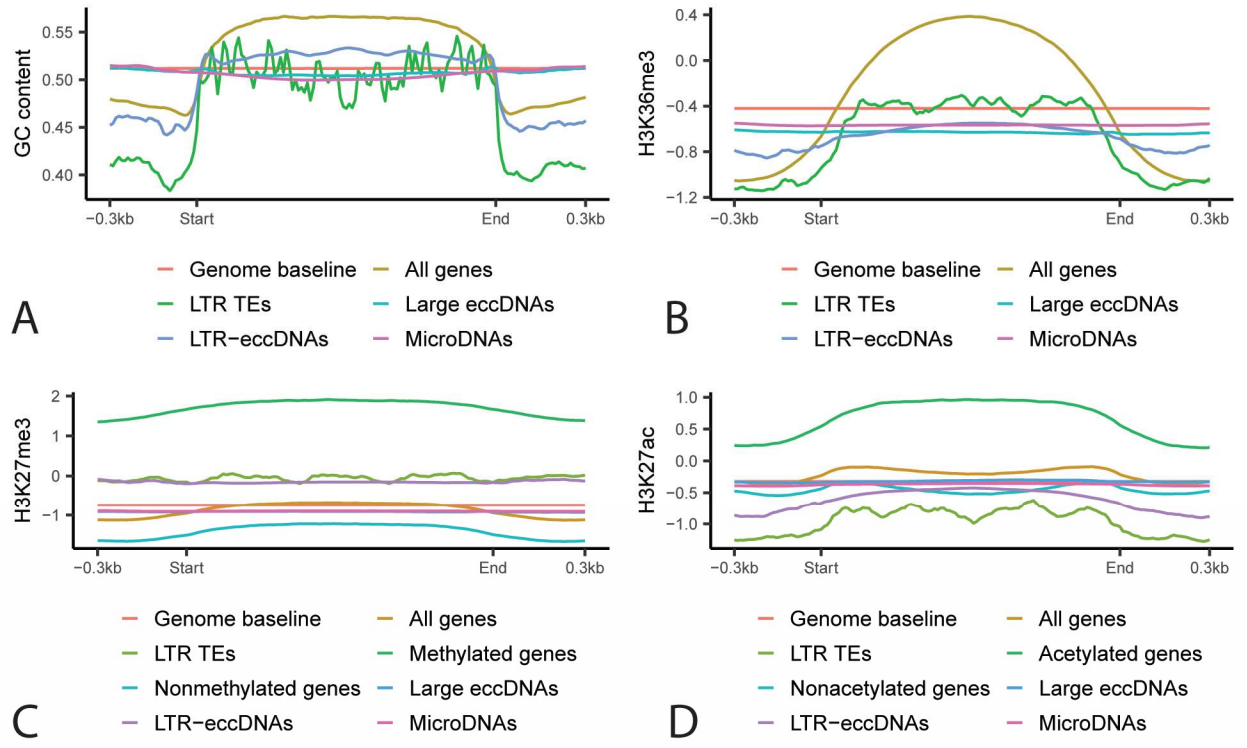
Supplemental Figure 11. Scatter plot showing the number of genes and log 10 of the number of junction split reads per 100 kilobase pair bin in the *M. oryzae* Guy11 genome for A. large eccDNAs or B. microDNAs. The red line represents a linear regression line and the grey shadow represents 95% confidence intervals.



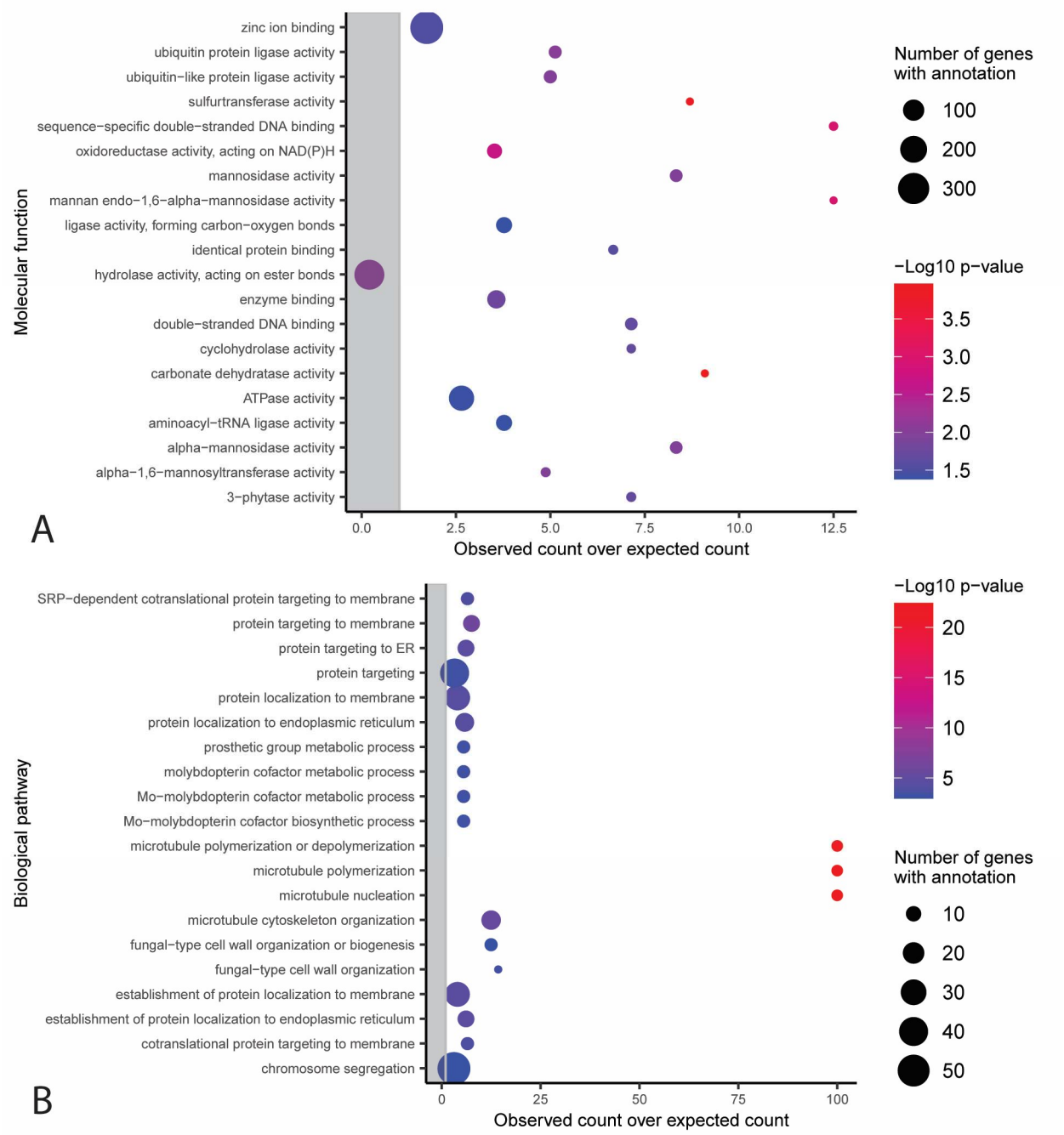
Supplemental Figure 12. Two-dimensional density plot showing the log 10 of the reads per kilobase million averaged across multiple RNAseq samples and log 10 of the number of overlapping junction split reads for each gene for A. large eccDNAs and B. microDNAs. The red line represents a linear regression line and the grey shadow represents 95% confidence intervals.



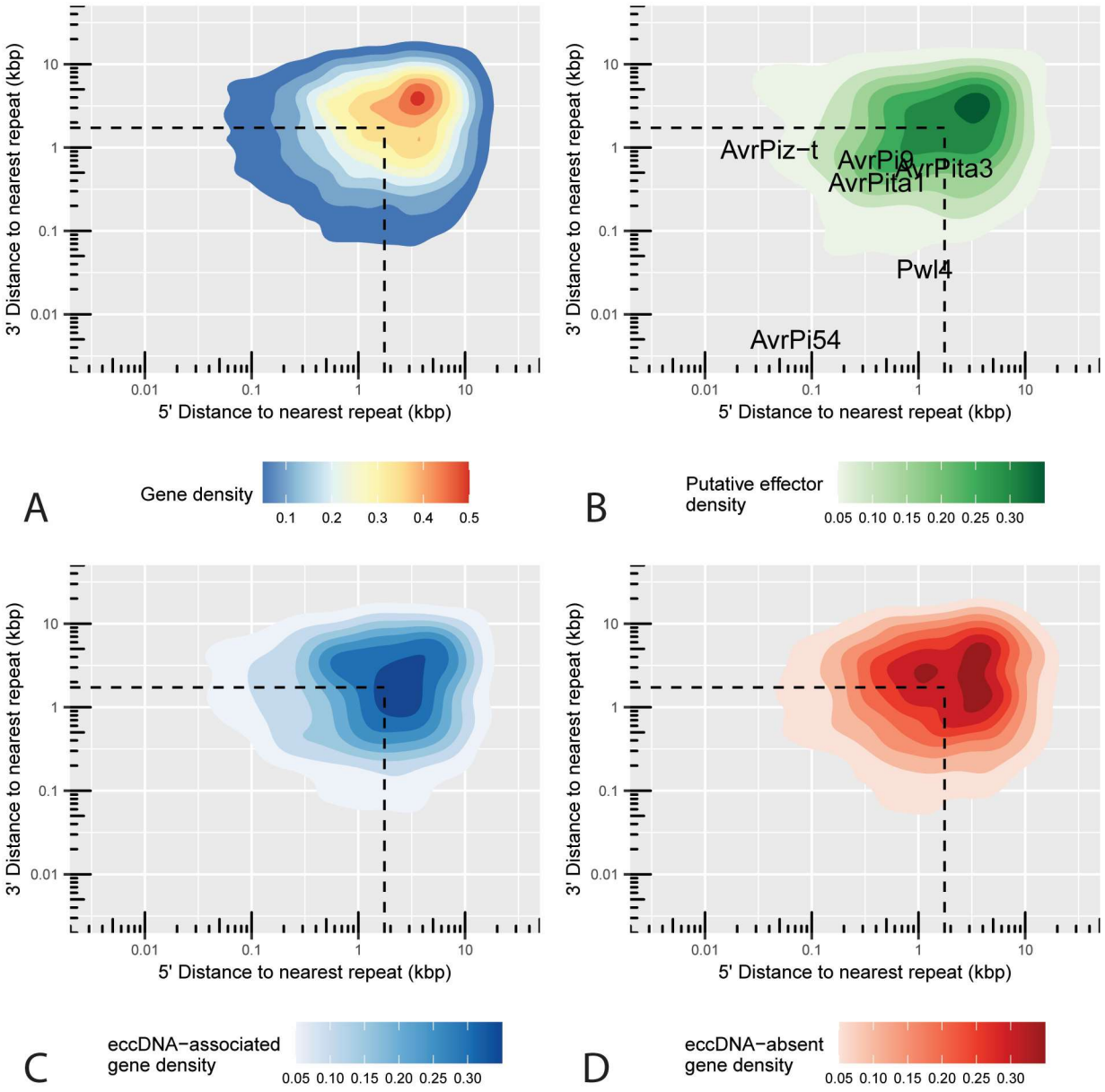
Supplemental Figure 13. Box plot showing the log 10 of the number of junction split reads per million reads averaged across all sequenced samples for eccDNA forming regions that do and do not contain ACSs for A. large eccDNAs and B. microDNAs.



Supplemental Figure 14. 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 are also represented in C and D, respectively.



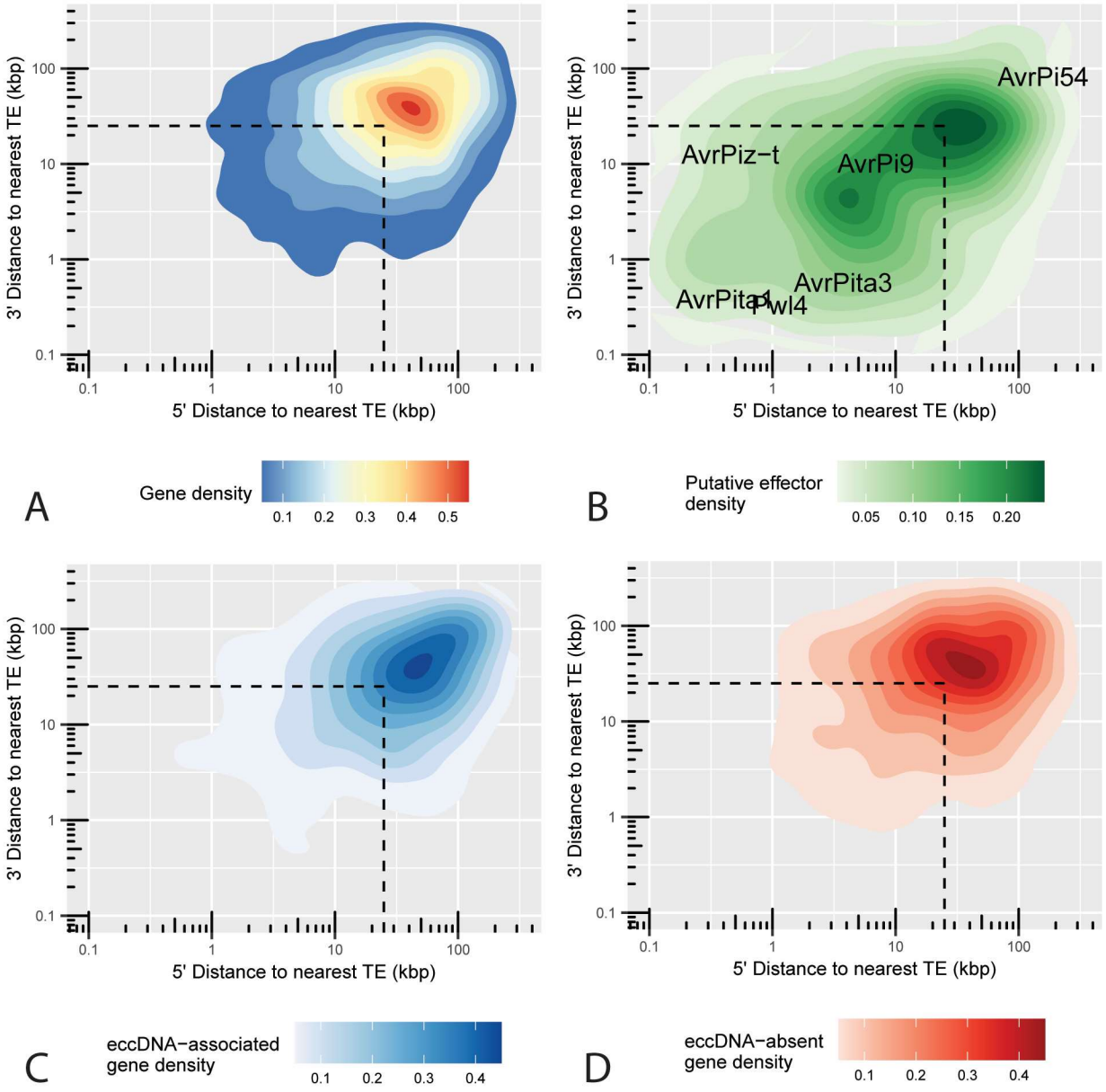
Supplemental Figure 15. Functional categories in the A. molecular function and B. biological pathway Gene Ontology with an observed number of eccDNA-associated genes that is significantly different from the expected number with correction for gene length bias (Chi-square test, $p < 0.05$). The y-axis 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.



954

955 **Supplemental Figure 16.** Two-dimensional density plot representing the 5' and 3' distance to the
956 nearest repeat in the *M. oryzae* Guy11 genome in kilobase pairs for each A. gene, B. predicted effector,
957 C. eccDNA-associated genes, and D. eccDNA-absent genes. Known effectors are shown as text in B.
958 Dashed lines represent median 5' and 3' distance to nearest gene.

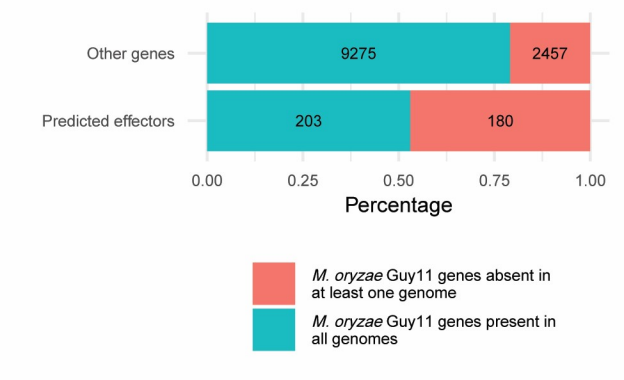
959



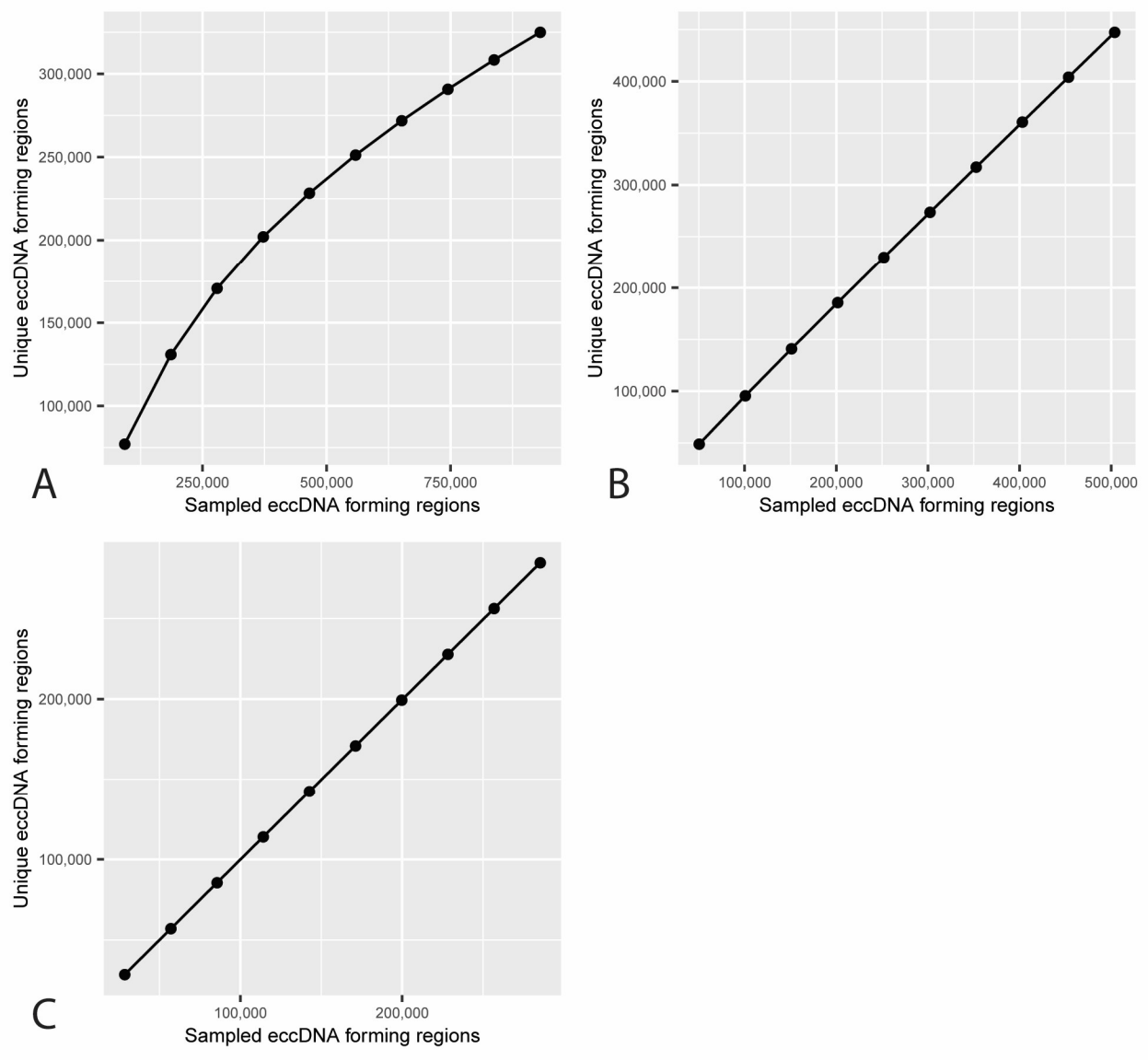
960

961 **Supplemental Figure 17.** Two-dimensional density plot representing the 5' and 3' distance to the
962 nearest transposable element in the *M. oryzae* Guy11 genome in kilobase pairs for each A. gene, B.
963 predicted effector, C. eccDNA-associated genes, and D. eccDNA-absent genes. Known effectors are
964 shown as text in B. Dashed lines represent median 5' and 3' distance to nearest gene.

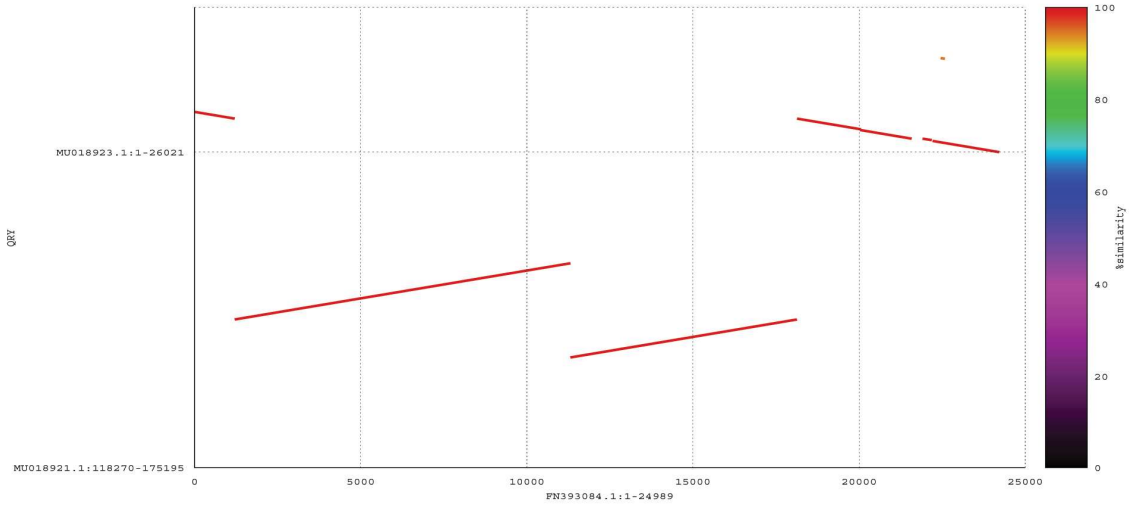
965



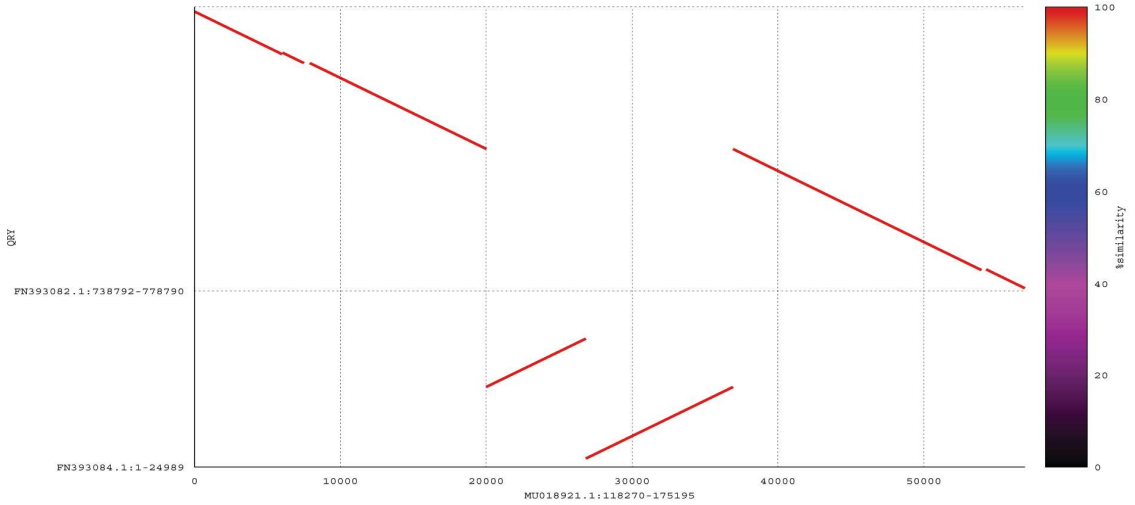
Supplemental Figure 18. Stacked bar plot showing the percentage of predicted effectors and all other genes in the *M. oryzae* Guy11 genome that had an ortholog in all other 162 *M. oryzae* genomes analyzed or not. Numbers indicate the number of genes in each category.



971 **Supplemental Figure 19.** Rarefaction analysis of the number of unique eccDNA forming regions at
972 different subsamples of eccDNA forming regions across all samples for A. LTR-eccDNAs, B. large eccDNAs
973 and C. microDNAs.



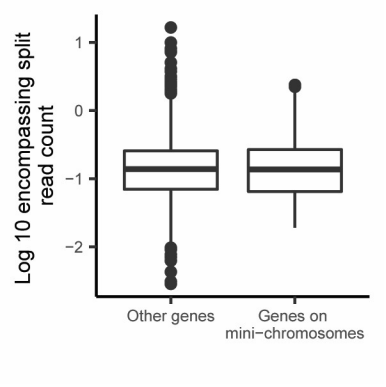
A



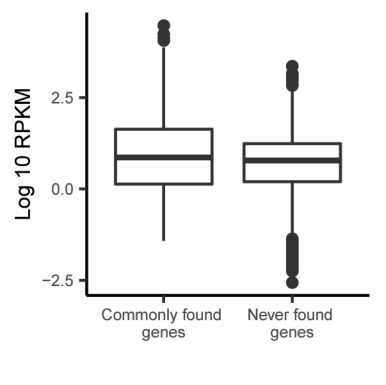
B

974

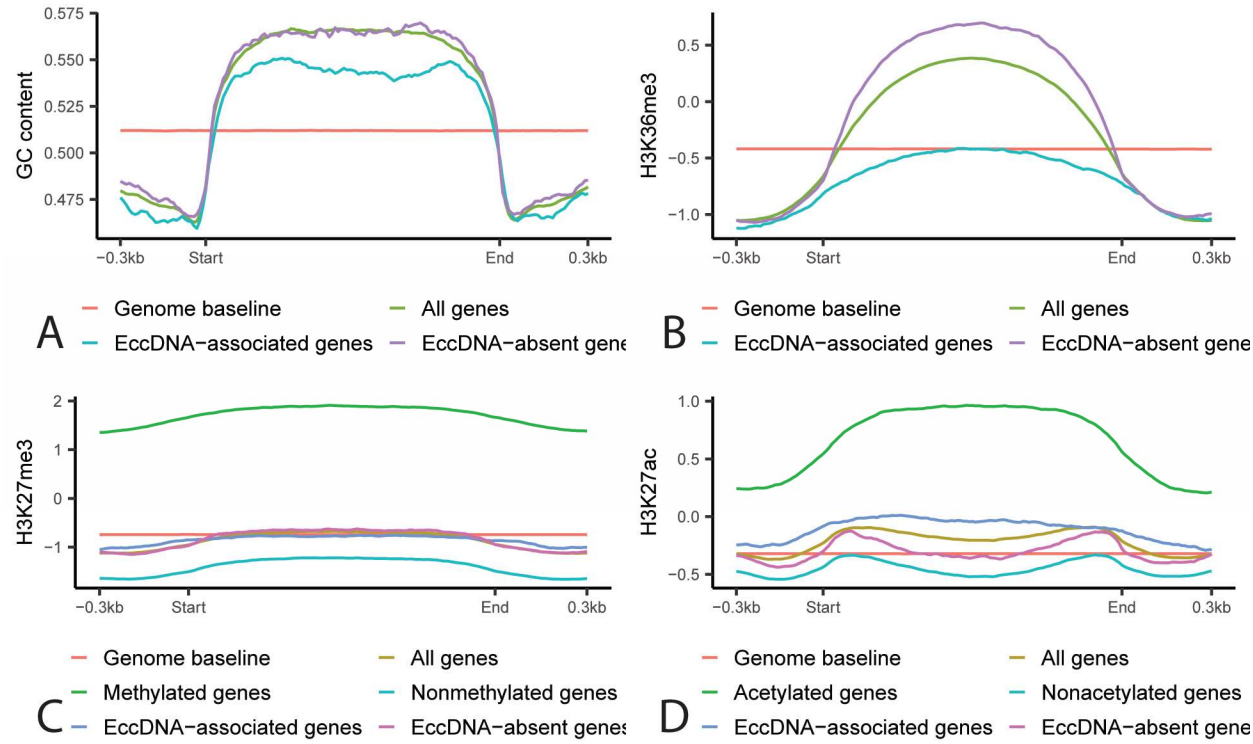
975 **Supplemental Figure 20.** Dot plot alignments between *S. cerevisiae* M22 and *S. cerevisiae* EC1118
976 genomes showing a DNA translocation likely caused by an eccDNA intermediate in yeast. A. A scaffold of
977 the EC1118 genome aligns to two different scaffolds of the M22 genome. B. A scaffold of the M22
978 genome aligns to two different scaffolds of the EC1118 genome.



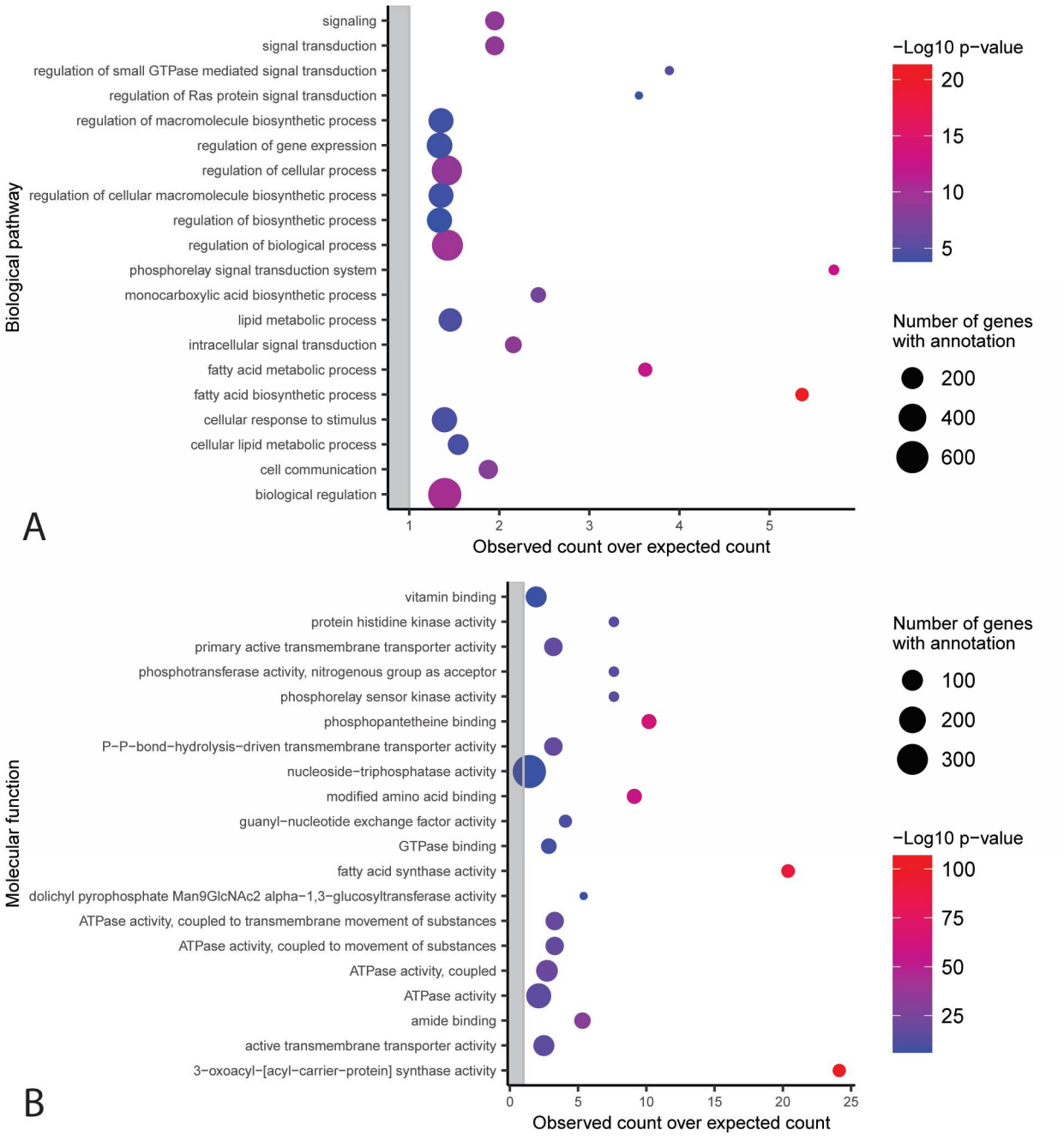
Supplemental Figure 21. Box plot showing the log 10 of the number of junction split reads per million reads averaged across all sequenced samples that fully encompass genes previously found on mini-chromosomes in other strains of *M. oryzae* and other genes.



Supplemental Figure 22. Box plot showing the log 10 reads per kilobase million (RPKM) averaged over 12 previously published RNAseq samples for eccDNA-associated genes and eccDNA-absent genes.

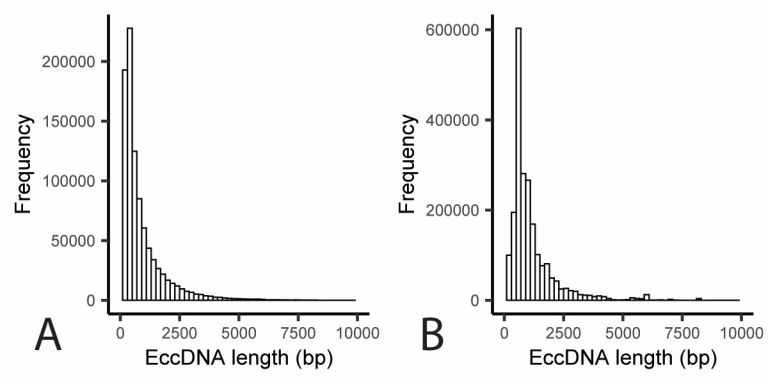


Supplemental Figure 23. Profile plots showing the average A. percent GC content, B., log2 ratio between read coverage for H3K36me3 chromatin immunoprecipitation and input control, C. log2 ratio between read coverage for H3K27me3 chromatin immunoprecipitation and input control and D. log2 ratio between 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 are also represented in C and D, respectively.



994

995 **Supplemental Figure 24.** Functional categories in the A. molecular function and B. biological pathway
996 Gene Ontology with an observed number of eccDNA-absent genes that is significantly different from the
997 expected number with correction for gene length bias (Chi-square test, $p < 0.05$). The y-axis shows the
998 different functional categories, and the x-axis represents the observed number of genes divided by the
999 expected number of genes in this group. Dots outside of the grey rectangle represent functional
1000 categories that are observed more often than expected. The size of dots indicates the total number of
1001 genes in the *M. oryzae* genome that belong to each functional category. Only the 20 categories with the
1002 largest -log10 p-values are shown.



Supplemental Figure 25. Histogram 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.

1008 **Supplemental Tables**

Accessions	Read count	EccDNA forming regions	Junction split reads	False positives per million reads
ERR2660591	222500000.	12.00	195.0	0.8765
ERR5060418	73760000.	15.00	210.0	2.847
SRR11528297	168600000.	5.000	66.00	0.3915

1009

1010 **Supplemental Table 1.** Read count, eccDNA forming regions inferred and number of junction split reads
1011 found using our pipeline on three previously published whole genome sequencing datasets for *M.*
1012 *oryzae*.

1013

Study	DNA extraction	Column purification	Linear DNA degradation	Circular DNA amplification
Moller 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)
Moller 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 (Magnaporthe 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 (Oryza 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)

Supplemental Table 2. 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.

Supplemental Data Files

Supplemental Data File 1. 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.

Supplemental Data File 2. 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.

Supplemental Data File 3. 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.

Supplemental Data File 4. List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* leaf tissue published by Lanciano et al[27]. 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.

Supplemental Data File 5. List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* seed tissue published by Lanciano et al[27]. 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.

Supplemental Data File 6. List of eccDNA forming regions called using Illumina circularome sequencing data for *O. sativa* callus tissue published by Lanciano et al[27]. 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.

Supplemental Data File 7. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* WT tissue published by Lanciano et al[27]. 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.

Supplemental Data File 8. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana epi12* mutant tissue published by Lanciano et al[27]. 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.

Supplemental Data File 9. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* leaf tissue published by Wang et al[32]. 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.

Supplemental Data File 10. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* root tissue published by Wang et al[32]. 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.

Supplemental Data File 11. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* stem tissue published by Wang et al[32]. 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.

Supplemental Data File 12. List of eccDNA forming regions called using Illumina circularome sequencing data for *A. thaliana* flower tissue published by Wang et al[32]. 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.

Supplemental Data File 13. List of eccDNA forming regions called using Illumina circularome sequencing data for *H. sapiens* muscle tissue published by Møller et al[11]. 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.

Supplemental Data File 14. List of eccDNA forming regions called using Illumina circularome sequencing data for *H. sapiens* leukocytes published by Møller et al[11]. 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.

Supplemental Data File 15. List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* WT cells published by Møller et al[11]. 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.

Supplemental Data File 16. List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* GAP1^{circle} cells published by Møller et al[28]. 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.

Supplemental Data File 17. List of eccDNA forming regions called using Illumina circularome sequencing data for *S. cerevisiae* cells from the deletion collection published by Møller et al[28]. 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.

Supplemental Data File 18. 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[28]. 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.

Supplemental Data File 19. 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.

Supplemental Data File 20. 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.

Supplemental Data File 21. 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 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.

Supplemental Data File 22. 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.

Supplemental Data File 23. List of GenBank accessions for the genomes of rice-infecting *M. oryzae* isolates used in this study for gene annotation.

Supplemental Data File 24. 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.

Supplemental Data File 25. List of GenBank accessions for the genomes of *M. oryzae* used in this study to search for eccDNA-mediated translocations.

Supplemental Data File 26. 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.

Supplemental Data File 27. 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.

Supplemental Data File 28. 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.

Supplemental Data File 29. Consensus sequences of LTR retrotransposons in the *M. oryzae* Guy11 genome in FASTA format.

Supplemental Data File 30. Genome, gene annotation, and transposable element annotation files used for comparative circularome analysis.

Supplemental Data File 31. List of SRA accessions for RNAseq data used in this study.

References

1. Paulsen, T., Kumar, P., Koseoglu, M. M. & Dutta, A. Discoveries of Extrachromosomal Circles of DNA in Normal and Tumor Cells. *Trends Genet.* **34**, 270–278 (2018).
2. Kilzer, J. M. *et al.* Roles of host cell factors in circularization of retroviral DNA. *Virology* **314**, 460–467 (2003).
3. Garfinkel, D. J. *et al.* Retrotransposon Suicide: Formation of Ty1 Circles and Autointegration via a Central DNA Flap. *J. Virol.* **80**, 11920–11934 (2006).
4. Møller, H. D. *et al.* Formation of extrachromosomal circular DNA from long terminal repeats of retrotransposons in Saccharomyces cerevisiae. *G3 Genes, Genomes, Genet.* **6**, 453–462 (2016).
5. Gresham, D. *et al.* Adaptation to diverse nitrogen-limited environments by deletion or extrachromosomal element formation of the GAP1 locus. *Proc. Natl. Acad. Sci.* **107**, 18551–18556 (2010).
6. Koo, D.-H. *et al.* Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed *Amaranthus palmeri*. *Proc. Natl. Acad. Sci.* **115**, 3332–3337 (2018).
7. Hull, R. *et al.* Transcription-induced formation of extrachromosomal DNA during yeast ageing. *PLoS Biol.* **17**, (2019).
8. Shcheprova, Z., Baldi, S., Frei, S. B., Gonnet, G. & Barral, Y. A mechanism for asymmetric segregation of age during yeast budding. *Nature* **454**, 728–734 (2008).
9. Nathanson, D. A. *et al.* Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science (80-.).* **343**, 72–76 (2014).
10. Turner, K. M. *et al.* Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. *Nature* **543**, 122–125 (2017).
11. Møller, H. D. *et al.* Circular DNA elements of chromosomal origin are common in healthy human somatic tissue. *Nat. Commun.* **9**, 1–12 (2018).
12. Shibata, Y. *et al.* Extrachromosomal MicroDNAs and Chromosomal Microdeletions in Normal Tissues. *Science (80-.).* **336**, 82–86 (2012).
13. Durkin, K. *et al.* Serial translocation by means of circular intermediates underlies colour sidedness in cattle. *Nature* **482**, 81–84 (2012).
14. Galeote, V. *et al.* Amplification of a Zygosaccharomyces bailii DNA segment in wine yeast genomes by extrachromosomal circular DNA formation. *PLoS One* **6**, 1–10 (2011).
15. Selin, C., de Kievit, T. R., Belmonte, M. F. & Fernando, W. G. D. Elucidating the role of effectors in plant-fungal interactions: Progress and challenges. *Front. Microbiol.* **7**, 1–21 (2016).
16. Dong, O. X. & Ronald, P. C. Genetic engineering for disease resistance in plants: Recent progress and future perspectives. *Plant Physiol.* **180**, 26–38 (2019).
17. Sánchez-Vallet, A. *et al.* The Genome Biology of Effector Gene Evolution in Filamentous Plant Pathogens. *Annu. Rev. Phytopathol.* **56**, 21–40 (2018).

18. Hollomon, D. W. Fungicide Resistance: 40 Years on and Still a Major Problem. *Springer* 3–11 (2015). doi:10.1007/978-4-431-55642-8

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

20. Bertazzoni, S. *et al.* Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Mol. Plant-Microbe Interact.* **31**, 779–788 (2018).

21. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, 8–10 (2012).

22. Soanes, D. & Richards, T. A. Horizontal Gene Transfer in Eukaryotic Plant Pathogens. *Annu. Rev. Phytopathol.* **52**, 583–614 (2014).

23. Dong, S., Raffaele, S. & Kamoun, S. The two-speed genomes of filamentous pathogens: Waltz with plants. *Curr. Opin. Genet. Dev.* **35**, 57–65 (2015).

24. Dean, R. *et al.* The Top 10 fungal pathogens in molecular plant pathology. *Mol. Plant Pathol.* **13**, 414–430 (2012).

25. Fernandez, J. & Orth, K. Rise of a Cereal Killer: The Biology of Magnaporthe oryzae Biotrophic Growth. *Trends Microbiol.* **26**, 582–597 (2018).

26. Foster, A. J. *et al.* CRISPR-Cas9 ribonucleoprotein-mediated co-editing and counterselection in the rice blast fungus. *Sci. Rep.* **8**, 1–12 (2018).

27. Lanciano, S. *et al.* Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet.* **13**, 1–20 (2017).

28. Møller, H. D., Parsons, L., Jørgensen, T. S., Botstein, D. & Regenberg, B. Extrachromosomal circular DNA is common in yeast. *Proc. Natl. Acad. Sci.* **112**, E3114–E3122 (2015).

29. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. High levels of profilin suppress the lethality caused by overproduction of actin in yeast cells. *FEBS Lett.* **316**, 41–47 (1993).

30. Gladieux, P. *et al.* Coexistence of multiple endemic and pandemic lineages of the rice blast pathogen. *MBio* **9**, (2018).

31. Zhong, Z. *et al.* Genetic Variation Bias toward Noncoding Regions and Secreted Proteins in the Rice Blast Fungus Magnaporthe oryzae. *mSystems* **5**, (2020).

32. Wang, K. *et al.* Deciphering extrachromosomal circular DNA in Arabidopsis. *Comput. Struct. Biotechnol. J.* **19**, 1176–1183 (2021).

33. Havecker, E. R., Gao, X. & Voytas, D. F. The diversity of LTR retrotransposons. *Genome Biol.* **5**, (2004).

34. Dillon, L. W. *et al.* Production of Extrachromosomal MicroDNAs Is Linked to Mismatch Repair Pathways and Transcriptional Activity. *Cell Rep.* **11**, 1749–1759 (2015).

35. Breier, A. M., Chatterji, S. & Cozzarelli, N. R. Prediction of Saccharomyces cerevisiae replication origins. *Genome Biol.* **5**, (2004).

36. Wang, Z.-Q. *et al.* A Putative Zn2Cys6 Transcription Factor Is Associated With Isoprothiolane

Resistance in Magnaporthe oryzae. *Front. Microbiol.* **9**, 2608 (2018).

37. Bohnert, S. *et al.* Fungicide resistance toward fludioxonil conferred by overexpression of the phosphatase gene MoPTP2 in Magnaporthe oryzae. *Mol. Microbiol.* **111**, 662–677 (2019).

38. Kim, K. *et al.* Evolution of the Genes Encoding Effector Candidates Within Multiple Pathotypes of Magnaporthe oryzae. **10**, 1–15 (2019).

39. Latorre, S. M., Reyes-avila, C. S., Malmgren, A. & Win, J. Recently expanded clonal lineages of the rice blast fungus display distinct patterns of presence / absence of effector genes. 1–31 (2020).

40. Borneman, A. R. *et al.* Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of saccharomyces cerevisiae. *PLoS Genet.* **7**, (2011).

41. Foster, A. J., Jenkinson, J. M. & Talbot, N. J. Trehalose synthesis and metabolism are required at different stages of plant infection by Magnaporthe grisea. *EMBO J.* **22**, 225–235 (2003).

42. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 10–12 (2011).

43. Bao, J. *et al.* PacBio Sequencing Reveals Transposable Elements as a Key Contributor to Genomic Plasticity and Virulence Variation in Magnaporthe oryzae. *Mol. Plant* **10**, 1465–1468 (2017).

44. Dean, R. A. *et al.* The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature* **434**, 980–986 (2005).

45. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

46. Li, H. Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094–3100 (2018).

47. Min, B., Grigoriev, I. V. & Choi, I. G. FunGAP: Fungal Genome Annotation Pipeline using evidence-based gene model evaluation. *Bioinformatics* **33**, 2936–2937 (2017).

48. Bao, W., Kojima, K. K. & Kohany, O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA* **6**, 4–9 (2015).

49. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).

50. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).

51. Breen, J. *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.* **10**, 98 (2010).

52. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**, 4673–4680 (1994).

53. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0.

54. Huang, W., Li, L., Myers, J. R. & Marth, G. T. ART: a next-generation sequencing read simulator. *Bioinformatics* **28**, 593–594 (2012).

55. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

56. Rice, P., Longden, L. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).

57. Zhang, W., Huang, J. & Cook, D. E. Histone modification dynamics at H3K27 are associated with altered transcription of in planta induced genes in Magnaporthe oryzae. *PLoS Genet.* **17**, 1–29 (2021).

58. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

59. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).

60. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

61. Törönen, P., Medlar, A. & Holm, L. PANNZER2: a rapid functional annotation web server. *Nucleic Acids Res.* **46**, W84–W88 (2018).

62. Alexa, A. & Rahnenfuhrer, J. topGO: Enrichment Analysis for Gene Ontology. (2019).

63. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 1–14 (2019).

64. Almagro Armenteros, J. J. *et al.* SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nat. Biotechnol.* **37**, 420–423 (2019).

65. Krogh, A., Larsson, B., Von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).

66. Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B. & Taylor, J. M. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Mol. Plant Pathol.* **19**, 2094–2110 (2018).

67. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **18**, 366–368 (2021).

68. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, 1–14 (2018).

69. Langner, T., Harant, A., Gomez-luciano, L. B., Shrestha, R. K. & Win, J. Genomic rearrangements generate hypervariable mini- chromosomes in host-specific lineages of the blast fungus. 1–34 (2020).

70. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6**, 1–11 (2005).

71. Tange, O. *GNU Parallel 2018*. (Ole Tange, 2018). doi:10.5281/zenodo.1146014

72. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

1319 73. pandas development team, T. pandas-dev/pandas: Pandas. (2020). doi:10.5281/zenodo.3509134

1320 74. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).

1321 75. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat.*
1322 *Methods* **17**, 261–272 (2020).

1323 76. Dowle, M. & Srinivasan, A. data.table: Extension of `data.frame`. (2020).

1324 77. Wickham, H. tidyr: Tidy Messy Data. (2021).

1325 78. Wickham, H. Reshaping Data with the reshape Package. *J. Stat. Softw.* **21**, 1–20 (2007).

1326 79. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A Grammar of Data Manipulation. (2021).

1327 80. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer-Verlag New York, 2016).

1328 81. Neuwirth, E. RColorBrewer: ColorBrewer Palettes. (2014).

1329 82. Wickham, H. & Seidel, D. scales: Scale Functions for Visualization. (2020).

1330 83. Wilke, C. O. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2'. (2020).

1331 84. Slowikowski, K. ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'.
1332 (2021).

1333 85. Kassambara, A. ggpubr: 'ggplot2' Based Publication Ready Plots. (2020).

1334 86. Hahne, F. & Ivanek, R. Statistical Genomics: Methods and Protocols. in (eds. Mathé, E. & Davis, S.)
1335 335–351 (Springer New York, 2016). doi:10.1007/978-1-4939-3578-9_16

1336 87. Iannone, R., Cheng, J. & Schloerke, B. gt: Easily Create Presentation-Ready Display Tables. (2021).

1337 88. Hancks, D. C. & Kazazian, H. H. Roles for retrotransposon insertions in human disease. *Mob. DNA*
1338 **7**, (2016).

1339