Adaptive evolution in virulence effectors of the rice blast fungus *Pyricularia oryzae*

Le Naour--Vernet Marie[1*], Charriat Florian[1*], Gracy Jérôme[2], Cros-Arteil Sandrine[1], Ravel Sébastien[1,3], Veillet Florian[1], Meusnier Isabelle[1], Padilla André[2], Kroj Thomas[1], Cesari Stella[1**], Gladieux Pierre[1**]

[1]PHIM Plant Health Institute, Univ Montpellier, INRAE, CIRAD, Institut Agro, IRD, Montpellier, France

[2]Centre de Biologie Structurale (CBS), Univ Montpellier, INSERM, CNRS, 29 rue de Navacelles, 34090 Montpellier, France.

[3]CIRAD, UMR PHIM, 34090 Montpellier, France

* These authors contributed equally to this work

** Corresponding authors:

Pierre Gladieux

Tel: +33 4 99 62 48 17

Email: pierre.gladieux@inrae.fr

Stella Cesari

Tel: +33 4 99 62 48 62

Email: stella.cesari@inrae.fr

Short title: MAX effector evolution in the rice blast fungus

ABSTRACT

Plant pathogens secrete proteins called effectors that target host cellular processes to promote disease. Recently, structural genomics has identified several families of fungal effectors that share a similar three-dimensional structure despite remarkably variable amino-acid sequences and surface properties. To explore the selective forces that underlie the sequence variability of structurally-analogous effectors, we focused on MAX effectors, a structural family of effectors that are major determinants of virulence in the rice blast fungus *Pyricularia oryzae*. Using structure-informed gene annotation, we identified 58 to 78 MAX effector genes per genome in a set of 120 isolates representing seven host-associated lineages. The expression of MAX effector genes was primarily restricted to the early biotrophic phase of infection and strongly influenced by the host plant. Pangenome analyses of MAX effectors demonstrated extensive presence/absence polymorphism and identified gene loss events possibly involved in host range adaptation. However, gene knock-in experiments did not reveal a strong effect on virulence phenotypes suggesting that other evolutionary mechanisms are the main drivers of MAX effector losses. MAX effectors displayed high levels of standing variation and high rates of non-synonymous substitutions, pointing to widespread positive selection shaping the molecular diversity of MAX effectors. The combination of these analyses with structural data revealed that positive selection acts mostly on residues located in particular structural elements and at specific positions. By providing a comprehensive catalog of amino acid polymorphism, and by identifying the structural determinants of the sequence diversity, our work will inform future studies aimed at elucidating the function and mode of action of MAX effectors.

AUTHOR SUMMARY

Fungal plant pathogens use small secreted proteins, called effectors, to manipulate to their own advantage their host's physiology and immunity. The evolution of these effectors, whether spontaneously or in response to human actions, can lead to epidemics or the emergence of new diseases. It is therefore crucial to understand the mechanisms underlying this evolution. In this article, we report on the evolution of effectors in one of the prime experimental model systems of plant pathology, *Pyricularia oryzae*, the fungus causing blast diseases in rice, wheat, and other cereals or grasses. We further characterize in this fungus a particular class of effectors, called MAX effectors, using structural models based on experimental protein structures of effectors. We show that this class of effector is produced by the pathogen during the early stages of infection, when plant cells are still alive. By comparing the gene content of isolates infecting different plant species, we show that the MAX effector arsenal is highly variable from one isolate to another. Finally, using the inferential framework of population genetics, we demonstrate that MAX effectors exhibit very high genetic variability and that this results from the action of natural selection.

INTRODUCTION

Plant pathogens secrete effector proteins to manipulate the physiology and metabolism of their host and to suppress its immunity. Consequently, effectors are expected to engage in coevolutionary interactions with plant defense molecules. The proximate mechanisms of effector-driven adaptation are relatively well-characterized: plant pathogens adapt to new hosts through changes in effector repertoires and effector sequences [1, 2]. However, the ultimate (eco-evolutionary) mechanisms underlying effector diversification have remained elusive. The concept of coevolution posits that adaptation in one partner drives counter-adaptations in the coevolving partner [3-5]. Under the co-evolutionary arms race model, variation for disease resistance and pathogen virulence is transient, resulting in a turnover of sequence variation through repeated episodes of strong directional selection [6]. In agricultural systems, because pathogens tend to be ahead of their hosts in the arms race owing to their larger populations and shorter generation times, the co-evolutionary arms race tends to result in so-called boom and bust cycles [7]. Under the alternative, 'trench warfare' hypothesis, advances and retreats of resistance or virulence genes frequencies maintain variation as dynamic polymorphisms [8, 9]. The maintenance of genetic polymorphisms is called 'balancing selection', a process by which different alleles or haplotypes are favored in different places (via population subdivision) and/or different times (via frequency-dependent negative selection). While there is a growing body of data demonstrating the nature and prevalence of the selective pressures that shape the diversity of immune systems in plants [6, 10-12], we still lack a clear picture of the co-evolutionary mechanisms underlying the molecular evolution of virulence factors in their interacting antagonists [13].

Effectors from plant pathogenic fungi are typically cysteine-rich secreted proteins smaller than 200 amino acids with an infection-specific expression pattern. Effectors are numerous in fungal genomes (several hundred to more than a thousand per genome), and rarely show homologies with known proteins or domains. They are also highly variable in sequence and do not form large families of sequence homologs. Based on similarity analyses, fungal

effectors can form small groups of paralogs (typically with less than five members), but they are most often singletons. This apparent lack of larger effector families has hindered attempts to probe into the evolutionary factors underlying their diversification. In addition, the high diversity of fungal effectors has hampered functional analyses due to the lack of good criteria for prioritizing them and our inability to predict their physiological role. Consequently, the virulence function and evolutionary history of most fungal effectors remain unknown.

Recently, the resolution of the three-dimensional (3D) structure of fungal effectors combined with Hidden Markov Model (HMM) pattern searches and structure modeling revealed that fungal effector repertoires are, despite their hyper-variability, dominated by a limited number of families gathering highly sequence-diverse proteins with shared structures and, presumably, common ancestry [14-17]. One such structurally-conserved but sequence-diverse fungal effector family is the MAX (*Magnaporthe* Avrs and ToxB-like) effector family. MAX effectors are specific to ascomycete fungi and show massive expansion in *Pyricularia oryzae* (synonym: *Magnaporthe oryzae*) [17], the fungus causing rice blast disease, one of the most damaging diseases of rice [18, 19]. MAX effectors are characterized by a conserved structure composed of six β-strands organized into two antiparallel β-sheets that are stabilized in most cases by one or two disulfide bridges. The amino acid sequence of MAX effectors is very diverse and they generally have less than 15% identity, which suggests that they are a family of analogous, rather than homologous, effector proteins. In other words, they share a similar three-dimensional structure, but their level of amino-acid identity does not allow to conclude about their shared ancestry. MAX effectors are massively expressed during the biotrophic phase of infection, suggesting an important role in disease development and fungal virulence [17]. Remarkably, about 50% of the known avirulence (AVR) effectors of *P. oryzae* belong to the MAX family, indicating that these effectors are closely monitored by the host plant immune system [17].

Pyricularia oryzae is a multi-host, poly-specialist pathogen that infects more than 50 monocotyledonous plants, including major cereal crops such as rice, maize, wheat, or barley [20-

120 23]. *Pyricularia oryzae* has repeatedly emerged on new hosts [21, 24], in new geographical areas
121 [25, 26], and phylogenomic analyses have revealed that it can be subdivided into several genetic
122 lineages, each preferentially associated with a specific or restricted set of host plant genera [27].
123 In *P. oryzae*, effectors can play a major role in host-shifts or host-range expansions [28-30]. For
124 example, loss of function of the PWT3 effector in *Lolium*-infecting strains contributed to gain of
125 virulence on wheat [29]. Similarly, loss of the MAX effector AVR1-CO39 is thought to have
126 contributed to the emergence of rice blast from foxtail-millet infecting isolates [20, 31]. This
127 indicates that MAX effectors may be important determinants of host specificity in *P. oryzae*.

128 In this study, we characterized the genetic diversity of MAX effectors in *P. oryzae* and
129 within its different host-specific lineages. We explored the evolutionary drivers of the
130 diversification of MAX effectors and tested whether MAX effectors represent important
131 determinants of *P. oryzae* host specificity. To this aim, we assembled and annotated 120 high-
132 quality *P. oryzae* genomes from isolates representing seven main host-specific lineages. We
133 mined these genomes for putative effectors and used hidden Markov models based on fold-
134 informed protein alignments to annotate putative MAX effectors. We identified 58 to 78 putative
135 MAX effector genes per individual genome distributed in 80 different groups of MAX homologs.
136 We showed that the expression of *MAX* effector genes is largely restricted to the early biotrophic
137 phase of infection and strongly influenced by the host plant. Our evolutionary analyses showed
138 that MAX effectors harbor more standing genetic variation than other secreted proteins and
139 non-effector genes, and high rates of non-synonymous substitutions, pointing to positive
140 selection as a potent evolutionary force shaping their sequence diversity. Pangenome analyses of
141 MAX effectors demonstrated extensive presence/absence polymorphism and identified several
142 candidate gene loss events possibly involved in host range adaptation. Our work demonstrates
143 that MAX effectors represent a highly dynamic compartment of the genome of *P. oryzae*, likely
144 reflecting intense co-evolutionary interactions with host molecules.

145 **RESULTS**

146 *Genome assembly and prediction of MAX effector genes.*

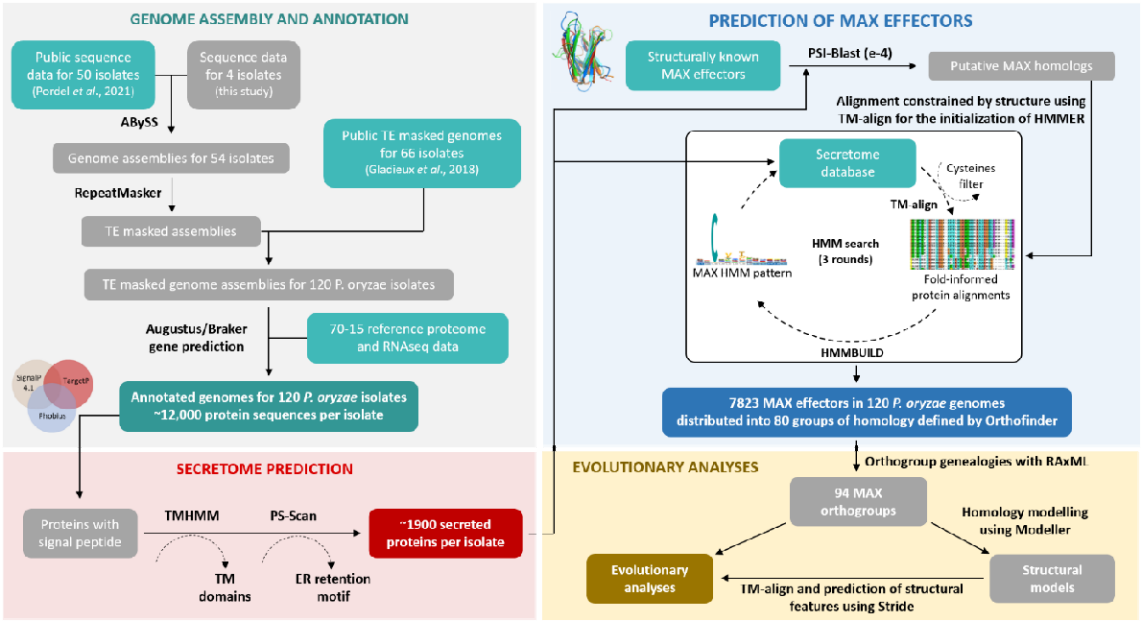
147 We used a collection of genome assemblies that included 120 haploid isolates of *Pyricularia*
148 *oryzae* fungi from 14 different host genera: *Oryza* (n=52), *Triticum* (n=21), *Lolium* (n=12),
149 *Setaria* (n=8), *Eleusine* (n=8), *Echinochloa* (n=4), *Zea* (n=4), *Bromus* (n=2), *Brachiaria* (n=2),
150 *Festuca* (n=2), *Stenotaphrum* (n=2), *Eragrostis* (n=1), *Hordeum* (n=1), and *Avena* (n=1) (S1
151 Table). Assembly size ranged from 37Mb to 43.2Mb, with an average size of 40.2 Mb (standard
152 deviation [s.d.]: 1.9Mb). L50 ranged from five to 411 contigs (mean: 97.1; s.d.: 83.2) and N50
153 from 28Kb to 4.0Mb (mean: 238.6Kb; s.d.: 43.8Kb; S1 Table). Gene prediction based on protein
154 sequences from reference 70-15 and RNAseq data identified 11,520 to 12,055 genes per isolate
155 (mean: 11,763.2; s.d.: 103.7). The completeness of assemblies, as estimated using BUSCO [32],
156 ranged between 93.4 and 97.0% (mean: 96.4%; s.d.: 0.6%; S1 Table).

157 MAX effectors were identified among predicted secreted proteins using a combination of
158 similarity searches [33, 34] and structure-guided alignments [35] as summarized in Figure 1. To
159 assess variation in the MAX effector content of *P. oryzae*, we constructed groups of homologous
160 genes (i.e., "orthogroups" or OG) using the clustering algorithm implemented in ORTHOFINDER
161 [36]. A given orthogroup was classified as secreted proteins or MAX effectors if 10% of
162 sequences in the group were identified as such by functional annotation. Sequences were
163 grouped in 14,767 orthogroups, of which 80 were classified as encoding MAX effectors and
164 3,283 as encoding other types of secreted proteins (Figure 1). The number of MAX orthogroups
165 per isolate ranged from 58 to 73 (average: 65.8; s.d.: 2.8), representing between 58 to 78 *MAX*
166 genes per isolate (average: 68.4; s.d.: 3.6). The 80 orthogroups of MAX effectors were further
167 split into 94 groups of orthologs, by identifying paralogs using gene genealogies inferred with
168 RAxML v8 [37] (S2 Table). Comparison of these 94 groups of orthologs with MAX effectors
169 predicted by previous studies [15-17, 38] revealed that 19 were not predicted by any other
170 study, while 75 were predicted by at least one other study (including nine predicted by all

171 studies). Twenty-five MAX effectors predicted by other analyses, were not identified by our
172 prediction pipeline and were, therefore, not considered in our study (S2 Table).

173 The number of MAX effectors per isolate was primarily determined by the host of origin.
174 We did not observe any significant relationship between the number of MAX effectors and
175 assembly properties (assembly size and N50), unlike the number of secreted proteins and total
176 number of genes (S1 Figure). Analysis of variance revealed that the host of origin had a
177 significant impact on the number of MAX effectors ($F_{13,95}$=10.33; p=3e^{-8}), while the origin of
178 genomic data did not have an effect (i.e., the study in which genomic data were initially
179 described; $F_{11,95}$=1.07; p=0.39; S1 Figure). These analyses indicate that the observed variation
180 in the size of the MAX effector repertoire is primarily biological in origin, not technical.

181



182

183 Figure 1. Schematic representation of the main steps of the bioinformatic pipeline used to predict genes in 120
184 genomes of *Pyricularia oryzae*, to identify genes encoding MAX effectors, and to measure their genetic variability.
185 References: Gladieux et al. 2018 [27]; Pordel et al. 2021 [21]. Genome assembly and annotation: using RNAseq and
186 reference proteins, we predicted genes with AUGUSTUS 3.4.0 [56] and BRAKER 1 [39] in 54 genomes that were
187 assembled in this study with ABYSS 1.9.0 [40], as well as in 66 assemblies that were published earlier; repeated
188 sequences were masked using REPEATMASKER 4.1.0 (http://www.repeatmasker.org/). Secretome prediction: we
189 identify signal peptides in the predicted genes by running SIGNALP 4.1 [41], TARGETP 1.1 [42], and PHOBIUS 1.01 [62],
190 removing genes encoding proteins with a predicted transmembrane domain (identified using TMHMM [43]), as well

as endoplasmic reticulum proteins (identified with PS-SCAN from expasy.org). Prediction of MAX effectors: MAX effectors were identified using PSI-BLAST [33] to search for homologs of known MAX effectors (AVR1-CO39, AVR-Pia, AvrPiz-t, AVR-PikD, and ToxB) in the predicted secretome, followed by structure-guided alignment of PSI-BLAST hits using TM-ALIGN [35], and three iterative rounds of HMMER [34] searches based on HMMBUILD models built on TM-ALIGN alignments of significant hits. Only proteins with two expected conserved cysteines less than 33-48 amino acids apart were retained [17]. Evolutionary analyses: the 11 orthogroups that included paralogous copies of MAX effectors were split into sets of orthologous sequences using genealogies inferred using RAxML v8 [37], yielding a total of 94 single-copy MAX orthologs for evolutionary analyses; protein structure models were inferred using homology modeling in MODELLER [44], and structural features were identified with STRIDE [45].

MAX effectors are massively deployed during rice infection.

To determine whether these putative MAX effectors are deployed by *P. oryzae* during plant infection, we analyzed the expression patterns of the 67 *MAX* genes predicted in the genome of the reference isolate Guy11 by qRT-PCR (Figure 2A). Using RNA samples from Guy11 mycelium grown on artificial media, we found that 94% of the *MAX* genes (63 genes) were not, or very weakly expressed during axenic culture, and only four (i.e., *MAX24*, *MAX29, MAX59,* and *MAX66*) showed weak, medium or strong constitutive expression (Figure 2A). *MAX* genes were, therefore, predominantly repressed in the mycelium of *P. oryzae*.

We analyzed the expression of the *MAX* genes 2, 3, 4, and 7 days following spray inoculation of Guy11 on the rice cultivar Maratelli, which is highly susceptible to *P. oryzae*. We found that 67% of the *MAX* genes (45 genes) were expressed (Figure 2A). Among them, three were also expressed in the mycelium (i.e., *MAX31, MAX59*, and *MAX94*). *MAX31* was over-expressed under infection conditions, whereas the other two showed similar expression levels *in vitro* and during infection. 64% of the *MAX* genes (42 genes) showed an infection-specific expression profile with relative expression levels ranging from very low (0.008-0.04) to very high (>5). Like the *Bas3* gene, encoding a *P. oryzae* effector specifically induced during the biotrophic phase of infection, all *MAX* genes showed maximal expression between the second and fourth day post-inoculation (Figure 2A; S2 Figure).

219 To test whether the genotype of the host plant could influence the expression of *MAX*
220 genes, we analyzed their expression patterns upon infection of the rice cultivar Kitaake, which
221 has a higher basal resistance to *P. oryzae* than Maratelli. We computed the median relative
222 expression across three independent experiments with five biological replicates each (Figure
223 2A; Figure 2B). During Kitaake infection, 78% of the *MAX* genes (52 genes) were upregulated
224 compared to the *in vitro* condition, while 64% (43 genes) were induced upon infection of
225 Maratelli (Figure 2A). Some *MAX* genes not expressed in Maratelli were induced in Kitaake (e.g.,
226 *MAX24, MAX30, MAX32, MAX43, MAX71,* and *MAX73*) (Figure 2B, S3 Figure). Others were
227 significantly upregulated in Kitaake compared to Maratelli (i.e., *MAX22, MAX44, MAX55, MAX57,*
228 *MAX69*, and *MAX91*). However, a few genes, such as *MAX15, MAX37* and *MAX62*, among the
229 most strongly expressed effectors in Maratelli, showed weaker levels of expression in Kitaake.
230 These results show that Guy11 deploys a wider diversity of MAX effectors during the infection of
231 Kitaake compared to that of Maratelli, and that MAX effectors are subject to host-dependent
232 expression polymorphism.
233 Taken together, our data indicate that during the biotrophic phase of rice infection, *P.*
234 *oryzae* actively expresses a significant portion of its MAX effector repertoire in a host-dependent
235 manner, which suggests that these effectors have an important function in fungal virulence.
236
237

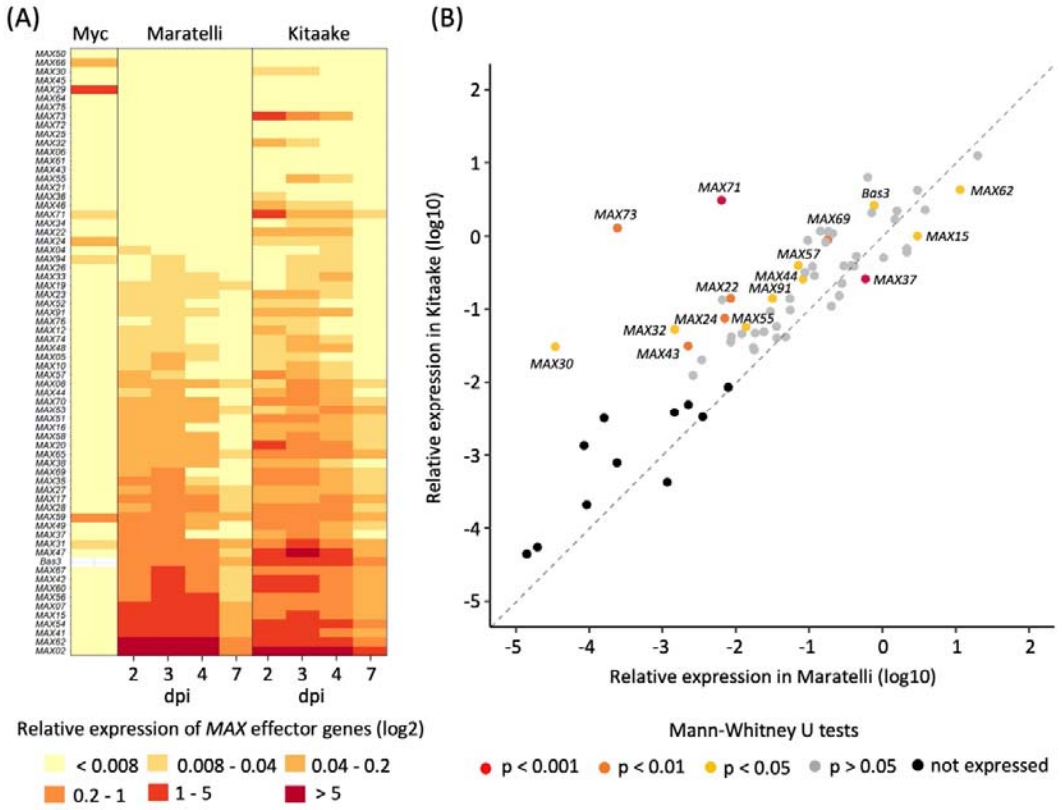


Figure 2. The expression of *MAX* genes is biotrophy-specific and is influenced by the host plant. (A) Transcript levels of *MAX* genes and the biotrophy marker gene *Bas3* were determined by qRT-PCR in the mycelium (Myc) of the *P. oryzae* isolate Guy11 grown for 5 days in liquid culture, and in infected leaves of the rice cultivars Maratelli and Kitaake at 2, 3, 4, and 7 days post inoculation (dpi) with Guy11. Relative expression levels were calculated using the constitutively expressed *MoEF1α* (*Elongation Factor 1α*) gene as a reference. The heatmap shows the median relative expression value for each gene (in log2 scale), calculated from 6 independent biological samples for the Myc condition, and 3 independent inoculation experiments (each with 5 independent leaf samples per time point) for each rice cultivar. Effectors were ranked from top to bottom by increasing relative expression values in Maratelli. Relative expression values were assigned to six categories: not expressed (<0.008), very weakly (0.008-0.04), weakly (0.04-0.2), moderately (0.2-1), strongly (1-5) and very strongly expressed (>5). (B) Scatter plot comparing the relative expression levels of *MAX* genes in Guy11-infected Maratelli and Kitaake cultivars. Each point shows the maximum median relative expression value (in log10 scale) calculated in the infection kinetics described in (A). Difference in effector relative expression levels between the two conditions was assessed by Mann-Whitney U tests and dots were colored according to significance results: grey (p>0.05), yellow (p<0.05), orange (p<0.001), red (p<0.0001), black (effectors not expressed in both conditions).

256 *The MAX effector repertoire is highly variable.*

257 To investigate the genetic diversity of MAX effectors in *P. oryzae*, we analyzed their nucleotide
258 diversity per base pair (π), their ratio of non-synonymous to synonymous nucleotide diversity
259 (π_N/π_S; page 226 in ref. [46]), and their presence-absence polymorphism. Compared to other
260 secreted proteins or other genes, MAX effector orthogroups had higher π, and π_N/π_S values, and
261 lower presence frequency (S4 Figure). Orthogroups including known avirulence genes like
262 *AVR1-CO39*, *AvrPiz-t* and *AVR-Pik* featured among the most diverse orthogroups of MAX
263 effectors (S1 Data).

264 We categorized genes in the pangenome according to their presence frequencies [36],
265 with core genes present in all isolates, softcore genes present in >99% isolates, shell genes
266 present in 1-99% isolates and cloud genes present in <1% isolates. The majority of MAX effector
267 genes were classified as shell (64/80 [80%] orthogroups), while the majority of other secreted
268 proteins or other genes were classified as core or softcore (1650/3283 [50.2%] and
269 6714/11404 [58.9%] orthogroups, respectively) (Figure 3A). Only a minority of genes were
270 present in multiple copies (MAX: 15/80 [18.8%]; other effectors: 746/3283 [22.7%]; other
271 genes: 1439/11404 [12.6%]; Figure 3A). Assessment of the openness of the pan-genome by
272 iteratively subsampling isolates revealed a closed pangenome with a limited number of pan and
273 core genes for MAX effectors, other secreted proteins and the remainder of the gene space
274 (Figure 3B). Nucleotide diversity differed significantly between categories of the pangenome for
275 non-MAX effectors (Kruskal-Wallis test: H=181.17, d.f.=2, $p<0.001$) and other genes (Kruskal-
276 Wallis test: H=225.25, d.f.=2, $p<0.001$), but not for MAX effectors (Kruskal-Wallis test: H=2.50,
277 d.f.=2, $p>0.05$). For non-MAX effectors and other genes, nucleotide diversity π was significantly
278 higher in the shell genes than in softcore genes and core genes (Post-hoc Mann-Whitney U-tests,
279 $p<0.001$; Figure 3C). The frequency of MAX orthogroups was positively and significantly
280 correlated with the frequency of neighboring orthogroups at the species-wide level and at the
281 level of Oryza and Setaria lineages, which indicates that non-core MAX effectors tend to be
282 located in regions with presence-absence variation (S5 Figure).

Together, these analyses show that the MAX effector repertoire is highly plastic compared to other gene categories, both in terms of the presence/absence of orthogroups and the sequence variability within orthogroups.

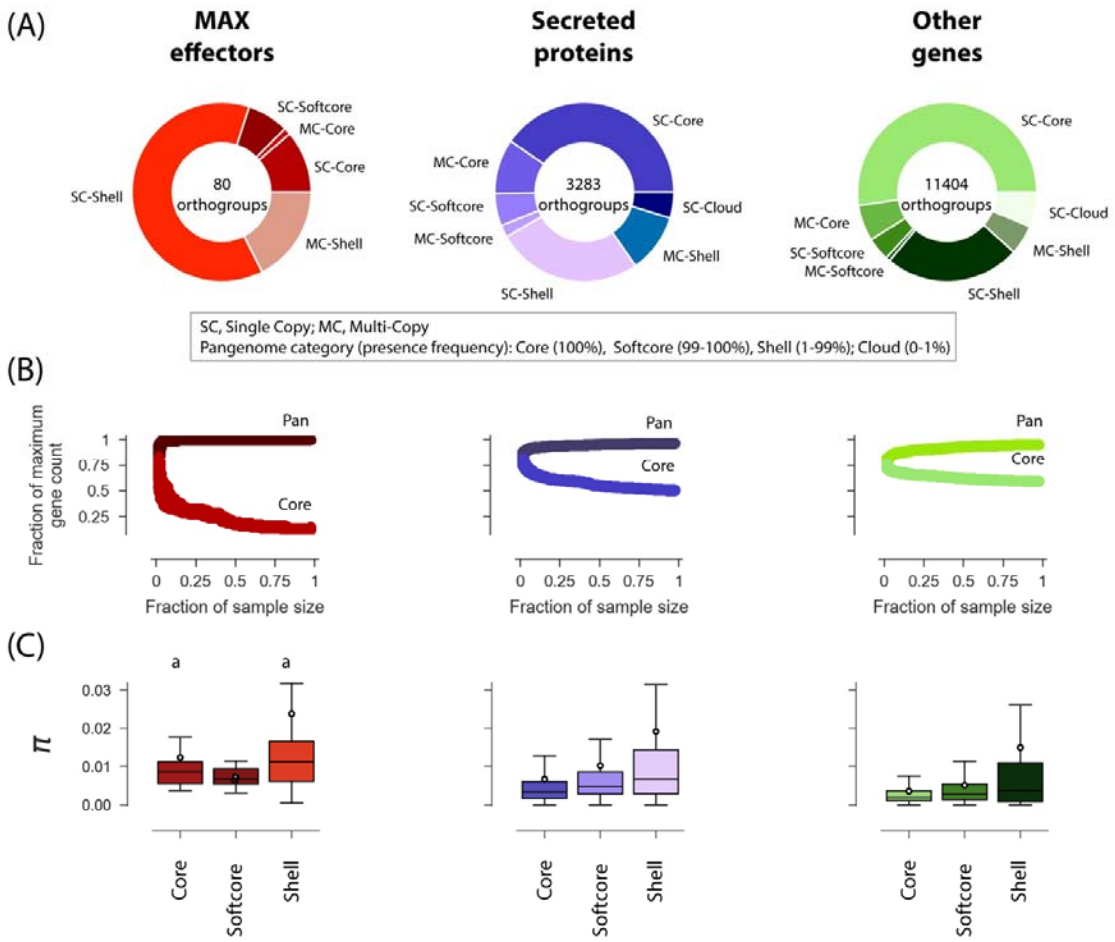


Figure 3. The pan-genome of *P. oryzae*. (A) Composition of the pangenome of MAX effectors, other secreted proteins, and other genes. (B) Rarefaction analysis of the size of pan- and core-genomes. For k ranging from two to the sample size minus one, pan- and core-genome sizes were computed for 1000 random combinations of k isolates. Subsample size is represented as a fraction of the sample size (n=120), and pan- and core-genome sizes as a fraction of maximum gene counts (reported at the center of donut plots in panel A). "core" genes are present in all isolates of a pseudo-sample of size k; "pan" qualifies genes that are not "core". (C) Nucleotide diversity per base pair (π) in core, softcore, and shell genes. A number of data points were cropped from the nucleotide diversity plot for visually optimal presentation but included in statistical tests. In box plots, the black circle is the mean, the black line is the median. Cloud genes were not included in the nucleotide diversity plot because it was not computable due to the small sample

size or lack of sequence after filtering for missing data. Shared superscripts indicate non-significant differences (Post-hoc Mann-Whitney U-tests).

Population subdivision.

To investigate signatures of positive selection in the genome of *P. oryzae*, and identify candidate loci involved in host specificity, we first identified the divergent lineages represented in our dataset. We inferred population structure from 6780 SNPs in single-copy core orthologs to minimize the potential impact of natural selection on our findings. We used complementary approaches that make no assumption about random mating or linkage equilibrium. Both clustering analyses with the sNMF software [47] (Figure 4A) and neighbor-net phylogenetic networks [48] (Figure 4B) revealed consistent patterns that split genetic variation primarily by host of origin. Although the lowest cross-entropy value was observed at K=11 in the sNMF analysis, we chose to represent K=8 because the cross-entropy was only slightly higher and K=8 did not split the *Triticum*- and *Eleusine*-associated clusters (S6 Figure). Lineage-level analyses were limited to the lineages with the largest sample size, associated with rice (*Oryza*), foxtail millet (*Setaria*), wheat (*Triticum*), ryegrass (*Lolium*), and goosegrass (*Eleusine*).

Population subdivision inferred from MAX effectors using either 130 SNPs without missing data in single-copy core MAX effectors (Figure 4C) or presence/absence variation of all 80 MAX effector orthogroups (Figure 4D) revealed essentially the same groups as the analysis of the single-copy core orthologs. This indicates that genome-wide nucleotide variation, variation in MAX effector content, and nucleotide variation at MAX effectors reflected similar genealogical processes. The *Oryza* and *Setaria* lineages displayed exceptionally high presence/absence variation of MAX effectors (average Hamming distance between pairs of isolates: 0.123 and 0.095; Figure 4D), but only limited sequence variation at single-copy core MAX effectors (average Hamming distance between pairs of isolates: 0.017 and 0.012; Mann-Whitney U-tests, $p<0.05$; Figure 4C).

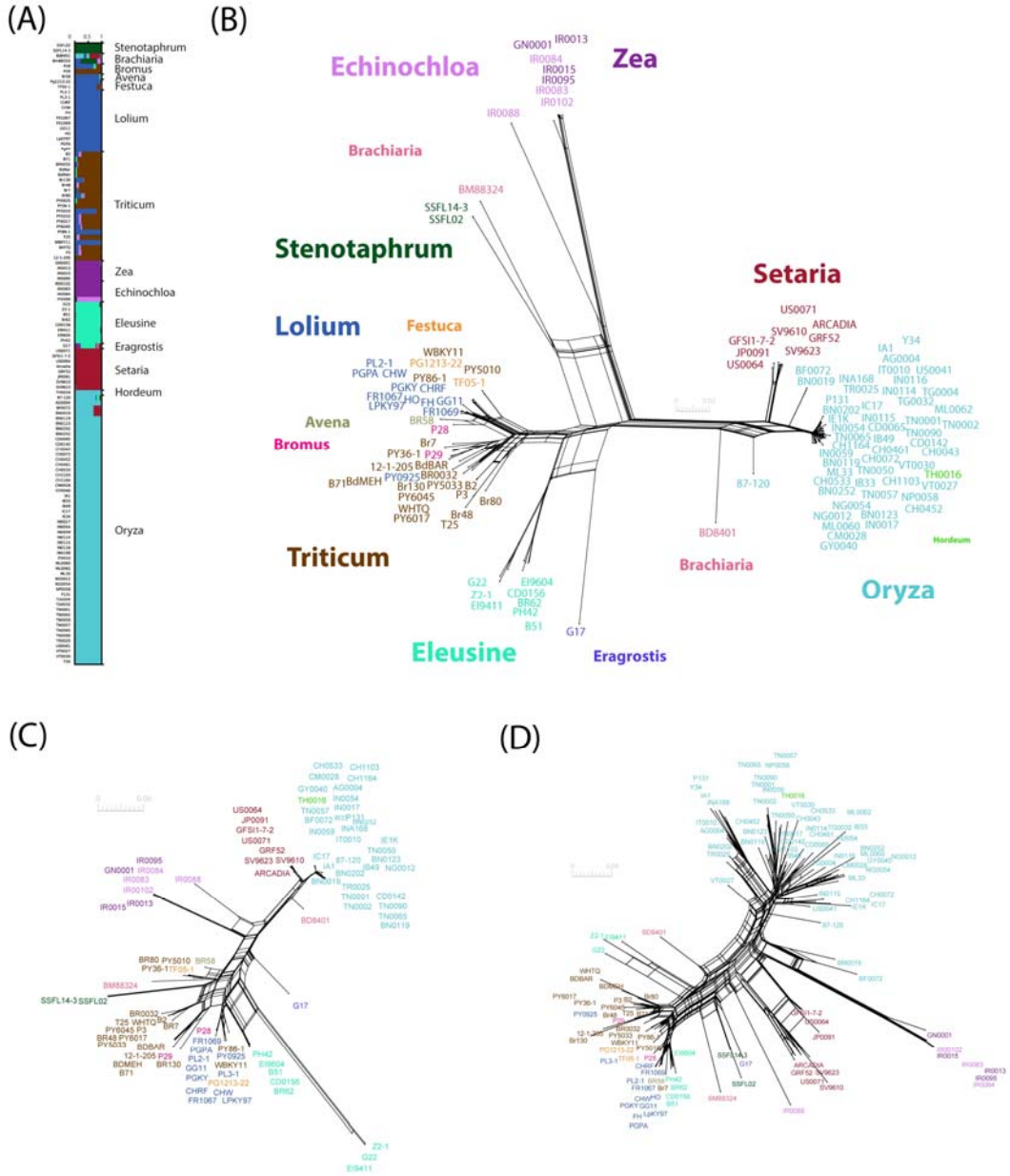


Figure 4. Population subdivision in 120 isolates of *Pyricularia oryzae*. Population subdivision was inferred from (A-B) 6780 polymorphisms at four-fold degenerate synonymous sites identified in coding sequences of single-copy core orthologs (one polymorphism without missing data randomly chosen per ortholog), (C) 130 SNPs without missing data identified in coding sequences of single-copy core MAX effectors, and (D) a table of presence/absence data (coded as 0 and 1) for all 80 MAX effector orthogroups. (A) Genetic ancestry proportions for individual isolates in K=8 ancestral populations, as estimated using the sNMF clustering algorithm [47]; each isolate is depicted as a horizontal bar divided into K segments representing the proportion of ancestry in K=8 inferred ancestral populations; host genus of origin is indicated on the right side. (B-D) Neighbor-net phylogenetic networks estimated with SplitsTree [48], with isolate names colored according to their host of origin.

333
334 *Loss of MAX effectors in specific lineages does not appear to be associated with host specificity.*
335 The comparison of the MAX effector content in the genomes of 120 *P. oryzae* isolates revealed
336 extensive presence/absence polymorphism between host-specific groups (S3 Table). To address
337 the underlying evolutionary mechanisms, we tested experimentally the hypothesis that MAX
338 effector losses are massively related to escape from receptor-mediated non-host resistance.
339 Indeed, the loss of MAX effectors in specific lineages of *P. oryzae* could primarily serve to escape
340 from non-host resistance during infections of novel plant species carrying immune receptors
341 specifically recognizing these effectors. To test this hypothesis, we focused on the *Oryza*- and
342 *Setaria*-infecting lineages, as previous investigations suggested that the *Oryza*-infecting lineage
343 emerged by a host shift from *Setaria* and we found both groups to be closely related (Figure 4)
344 [21, 49]. Our strategy was to introduce into the *Oryza*-isolate Guy11 MAX effectors absent from
345 the *Oryza* lineage but present in the *Setaria* lineage, and to assess the ability of these transgenic
346 isolates to infect rice.
347 We identified three MAX orthogroups that were largely or completely absent from the
348 *Oryza* lineage, but present in the majority of isolates of the other lineages (S3 Table).
349 Orthogroup MAX79 (OG0011591-1) was absent in all 52 *Oryza*-infecting isolates, while MAX83
350 (OG0011907), and MAX89 (OG0012141) were absent in 50 and 46 of them, respectively (S3
351 Table). Constructs carrying the genomic sequence of *MAX79*, *MAX83* or *MAX89* derived from the
352 *Setaria* isolate US0071 and under the control of the strong infection specific promoter of the
353 effector *AVR-Pia* were generated and stably introduced into Guy11. For each construct, three
354 independent transgenic lines were selected. Transgene insertion was verified by PCR and the
355 expression of transgenes was measured by qRT-PCR (S7 Figure). To test whether the selected
356 MAX effectors trigger immunity in rice, the transgenic isolates were spray-inoculated onto a
357 panel of 22 cultivars representative of the worldwide diversity of rice (S4 Table).
358 As controls, we used the MAX effectors *AVR-Pia*, which is rare outside the *Oryza* and
359 *Setaria* lineages, and *AVR1-CO39*, which is absent or pseudogenized in the *Oryza* lineage, but
360 present in all other host-specific lineages including *Setaria*. Both effectors are detected in rice by

361 the paired NLR immune receptors RGA4 and RGA5 from the *Pi-a/Pi-CO39* locus and thereby
362 contribute, respectively, to host or non-host resistance in this plant species [50, 51].
363 As expected, isolates expressing *AVR1-CO39* or *AVR-Pia* triggered resistance in the rice
364 variety Aichi Asahi that carries *Pi-a*, but caused disease on Nipponbare (*pi-a*`) and other varieties
365 lacking this *R* locus (Figure 5; S4 Table). Unlike the positive controls, the effectors MAX79,
366 MAX83 and MAX89 were not recognized and did not induce resistance in any of the tested rice
367 cultivars (Figure 5; S4 Table). The disease symptoms caused by the transgenic isolates carrying
368 these effectors were similar to those observed for wild-type Guy11 or Guy11 isolates carrying an
369 *RFP* (red fluorescent protein) construct. This suggests that these effectors do not significantly
370 increase the virulence of Guy11.
371 These experiments show that despite their loss in the *Oryza*-infecting lineage of *P.*
372 *oryzae,* and unlike AVR1-CO39, the effectors MAX79, MAX83, and MAX89 do not seem to induce
373 non-host resistance in rice. Consequently, other mechanisms than escape from host immunity
374 contributed to the loss of these MAX effectors during the putative host shift of *P. oryzae* from
375 *Setaria* to *Oryza*.
376

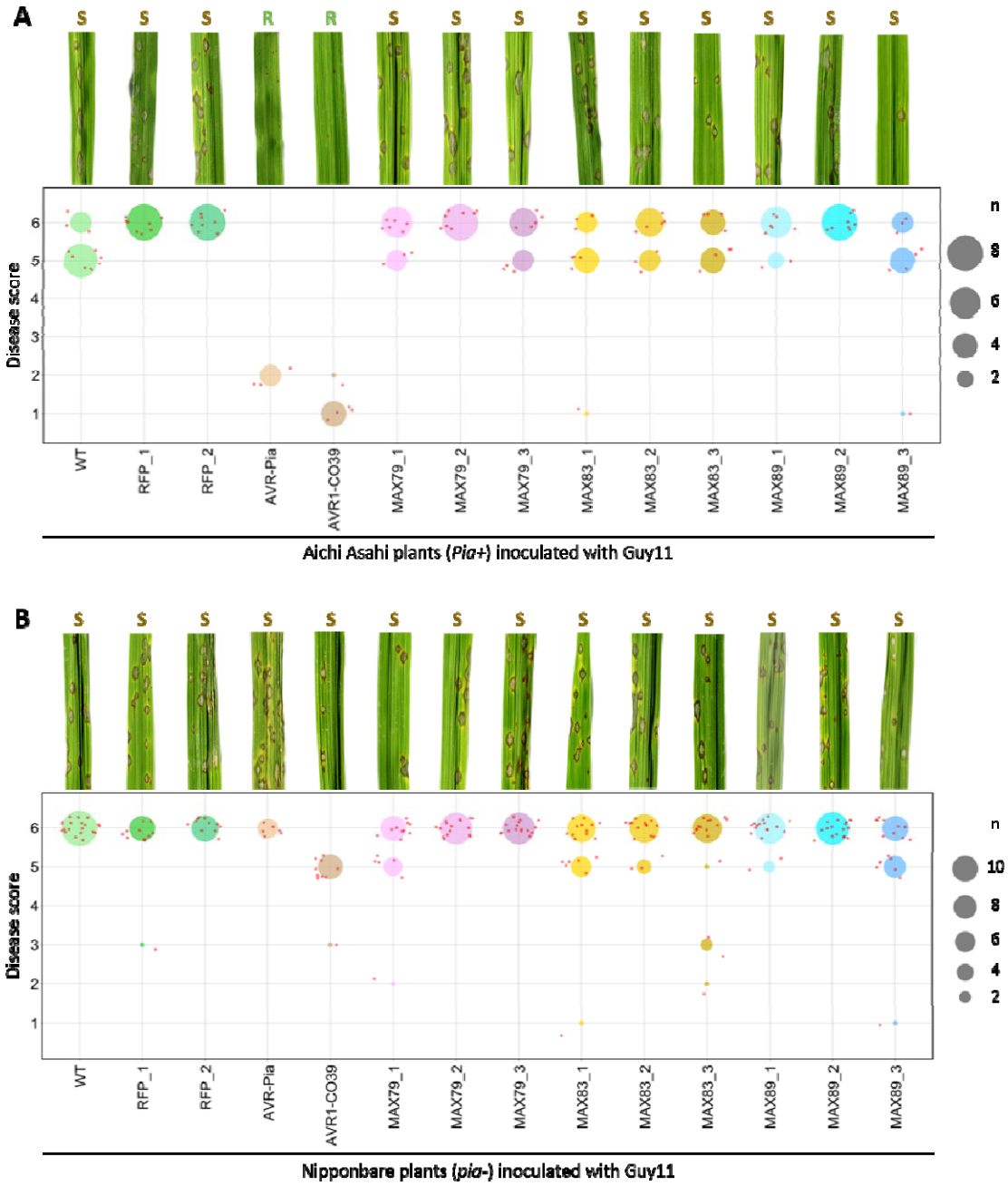


Figure 5. AVR1-CO39 contributes to non-host specificity in rice but not MAX79, MAX83 or MAX89. Wild type and transgenic isolates of *P. oryzae* Guy11 expressing the *RFP* (*red fluorescent protein*), *AVR-Pia, AVR1-CO39, MAX79*, *MAX83* or *MAX89* gene were spray-inoculated at 40 000 spores/ml on three-week-old rice plants of the cultivars Aichi Asahi (A) and Nipponbare (B). For each condition, representative disease phenotypes on rice leaves at seven days post-inoculation are shown (top panels, R: resistance, S: susceptibility). Disease phenotypes were also scored (from 1 [complete resistance] to 6 [high susceptibility]) on leaves from three to five individual rice plants and data are shown as dot plots (bottom panels). The size of each circle is proportional to the number of replicates (n) matching the corresponding score for each condition. Small red dots correspond to individual measurements. The experiment was

performed twice for Aichi Asahi and four times for Nipponbare for all isolates except for WT, AVR-Pia, and AVR1-CO39 control isolates. For these isolates, experiments were performed once on Aichi Asahi and twice on Nipponbare because disease phenotypes are well characterized in the literature.

MAX effectors display signatures of balancing selection.

To investigate the impact of balancing selection on MAX effector evolution, we focused on single-copy core, softcore, and shell orthogroups to avoid the possible effect of gene paralogy. We then computed π (nucleotide diversity per bp), F_{ST} (the amount of differentiation among lineages [52]), π_N (non-synonymous nucleotide diversity), π_S (synonymous nucleotide diversity), and π_N/π_S (the ratio of non-synonymous to synonymous nucleotide diversity). Large values of π and π_N/π_S, in particular, are possible indicators of a gene being under balancing selection.

Nucleotide diversity (π) differed significantly between groups of genes (Kruskal-Wallis test, H=509.9, d.f.=2, p <0.001; Figure 6A; S5 Table). π was significantly higher for the set of MAX effectors (average π: 0.0104, standard deviation: 0.0137), than for other secreted proteins (average π: 0.0079, standard deviation: 0.020), and other genes (average π: 0.0049, standard deviation: 0.014; Mann-Whitney U-tests, p<0.05), showing that MAX effectors, and to a smaller extent other secreted proteins, are more variable than a typical gene. At the lineage level, however, nucleotide diversity at MAX effectors tended to not significantly differ from other putative effectors, or other genes (S5 Table).

In addition to having greater nucleotide variation than other genes at the species level, MAX effectors also displayed a higher ratio of non-synonymous to synonymous nucleotide diversity (Figure 6B; S5 Table). The π_N/π_S ratio differed significantly between groups of genes (Kruskal-Wallis tests H=101.4, $d.f.$=2, p<0.001), and the excess of non-synonymous diversity was significantly, and markedly, higher for MAX effectors (average π_N/π_S: 1.826, standard deviation: 3.847) than for other effectors (average π_N/π_S: 0.461, standard deviation: 1.600), and other genes (average π_N/π_S: 0.448, standard deviation: 1.463; Mann-Whitney U-tests, p<0.05). The higher π_N/π_S of MAX effectors was mostly driven by differences in π_N (Figure 6A; S5 Table). Twenty MAX effectors displayed values in the top 5% percentile of non-effector genes, far

413 exceeding the four genes expected by chance ($p<0.05$). More specifically, 26 MAX effectors
414 displayed π_N/π_S values greater than 1, which is the value expected under neutrality. This
415 included three well-known avirulence genes: *AVR1-CO39* ($\pi_N/\pi_S=2.564$), *AVR-Pik*
416 ($\pi_N/\pi_S=15.574$), and *AvrPiz-t* ($\pi_N/\pi_S=1.431$). The average π_N/π_S ratio was also higher at MAX
417 effectors than other secreted proteins and other genes in all lineages, with significant differences
418 in four lineages (Mann-Whitney U-tests, $p<0.05$), and the average π_N/π_S was greater than one in
419 the *Oryza* lineage (Figure 6B; S5 Table). Seven to eleven MAX effectors had $\pi_N/\pi s>1$ at the
420 lineage level, representing 8% (*Setaria*-infecting lineage) to 41% (*Lolium*-infecting lineage) of
421 MAX effectors with a defined $\pi_N/\pi s$ ratio (S2 Data).

422 $\pi_N/\pi_S>1$ is a strong indication of multiallelic balancing selection (*i.e.*, multiple alleles at
423 multiple sites are balanced), as single sites under very strong balancing selection cannot
424 contribute enough non-synonymous variability to push the π_N/π_S ratio above one [53]. To assess
425 whether the adaptation of lineages to their respective hosts may contribute to the species-wide
426 excess of non-synonymous diversity detected at MAX effectors, we estimated population
427 differentiation. The differentiation statistic F_{ST} differed significantly between groups of genes
428 (Kruskal-Wallis tests $H=8.731$, $d.f.=2$, $p=0.013$), and differentiation was significantly higher for
429 MAX effectors than for other secreted proteins and other genes (Figure 6A; S5 Table). F_{ST} was
430 also significantly, albeit relatively weakly, correlated with π_N/π_S at MAX effectors (Spearman's ρ:
431 0.304, $p=0.007$; S8 Figure). These observations indicate that between-lineages differences in
432 allele frequencies are greater for MAX effectors than for other secreted proteins or other genes,
433 which may result from divergent selection exerted by hosts.
434
435

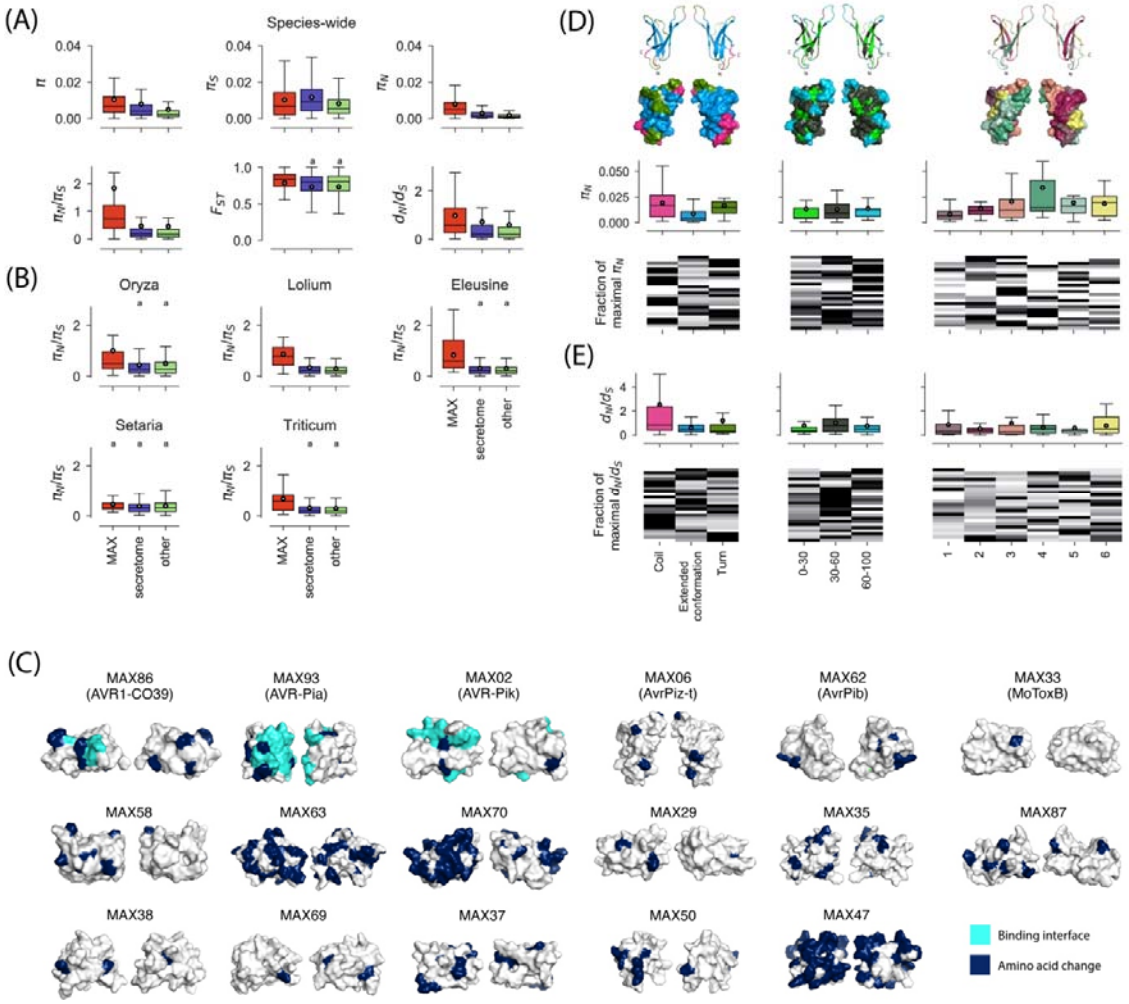


Figure 6. Summary statistics of polymorphism and divergence at MAX effectors, other secreted proteins (i.e., secretome), and other genes of *P. oryzae*. (A) Species-wide estimates of π (nucleotide diversity per bp), F_{ST} (the amount of differentiation among lineages), π_N (non-synonymous nucleotide diversity per bp), π_S (synonymous nucleotide diversity per bp), π_N/π_S (the ratio of non-synonymous to synonymous nucleotide diversity), d_N/d_S (the ratio of non-synonymous to synonymous rates of substitutions). (B) Lineage-specific estimates of π_N/π_S. (C) Amino acid changes segregating in *P. oryzae* at MAX effectors with an avirulence function and MoToxB (first row) and MAX effectors with $\pi_N/\pi_S>2$ (next rows); amino acid changes are shown in dark blue and known binding interfaces in light blue. Note that the π_N/π_S ratio is >2 for the avirulence genes *AvrPiz-t* and AVR-Pik. Proteins are displayed twice in panels C and E, with one copy rotated 180 degrees around a vertical axis. The interface involved in binding with host proteins is known for AVR1-CO39, AVR-Pia, and AVR-Pik only ([54-57]). (D) and (E) Species-wide estimates of π_N and d_N/d_S computed at MAX effectors with signatures of balancing selection ($\pi_N/\pi_S>1$; panel D) and signatures of directional selection ($d_N/d_S>1$; panel E) for different classes of structural features highlighted on the three-dimensional structure of *AvrPiz-t* above panel D: (i) secondary structure elements, with three subclasses: "coils", "extended conformation", and "turns"; (ii) solvent accessibility percentage of the Van der Waals surface of the amino acid side chain, with three sub-classes: 0-30% (buried), 30-60% (intermediate), and 60-100% (exposed); (iii) structural domains, with six subclasses

that grouped the coil, extended conformation and turn residues that define the six beta strands characteristic of MAX effectors. In the heatmaps, each line represents a MAX effector. For a given MAX effector and a given structural feature, the darkest color indicates the class of the structural feature for which the summary statistic is the highest. Only single-copy core, softcore, and shell groups of orthologous genes were included in the calculations. Shared superscripts indicate non-significant differences (post-hoc Mann-Whitney U-tests, $p>0.05$). A number of data points were cropped from plots in (A) and (B) for visually optimal presentation but included in statistical tests. In box plots, the black circle is the mean, the black line is the median.

MAX effectors display signatures of recurrent directional selection.

To detect adaptive molecular evolution, we collected orthologous sequences from outgroup *Pyricularia sp.* LS [58, 59] and estimated the d_N/d_S ratio (the ratio of non-synonymous to synonymous substitution rates) using a maximum likelihood method [60]. Note that d_N/d_S was not computed at the intra-specific level, but computed along the branches connecting the outgroup and isolates from the ingroup. Outgroup sequences could be retrieved for 10,174 out of 14,664 single-copy orthogroups, including 66 out of 94 single-copy orthologs of MAX effectors. The d_N/d_S ratio differed significantly between groups of genes (Kruskal-Walli. s tests $H=45.812$, $d.f.=2$, $p<0.001$; Figure 6A; S5 Table), and was higher for MAX effectors (average d_N/d_S: 0.977, s.d.: 1.316) than for other secreted proteins (average d_N/d_S: 0.711, s.d.: 1.722), and other genes (average d_N/d_S: 0.584, s.d.: 1.584; Mann-Whitney U-tests, $p<0.05$). The same pattern of higher d_N/d_S for MAX effectors was observed at the lineage level (S5 Table). Twenty-four of the 66 MAX effectors with outgroup sequence (i.e., 36.4%) showed $d_N/d_S>1$ (S2 Data), which is a strong indication of directional selection. $d_N/d_S>1$ is only expected for genes that have experienced repeated bouts of directional selection which led to repeated fixations of amino-acid substitutions [53]. Eleven MAX effectors displayed signatures of both multiallelic balancing selection ($\pi_N/\pi_S>1$) and multiallelic directional selection ($d_N/d_S>1$).

The divergence data, therefore, indicate that a scenario of molecular co-evolution involving repeated selective sweeps may apply to a substantial fraction (at least one-third) of MAX effectors.

484 *Structural determinants of polymorphism and divergence at MAX effectors.*

485 Different parts of proteins can be under different selective forces. To investigate if this is the
486 case in MAX effectors, we examined the relationship between three different measures of
487 diversity and three structural features. The measures of diversity were (1) the probability of an
488 amino acid being polymorphic, (2) the non-synonymous nucleotide diversity π_N, and (3) the
489 d_N/d_Sratio. The analyzed structural features were: (1) secondary structure annotations with the
490 three subclasses "extended conformation", "coils", and "turns"; (2) solvent accessibility
491 percentage of the Van der Waals surface of the amino acid side chain, with the three sub-classes:
492 0-30% (buried), 30-60% (intermediate), and 60-100% (exposed); (3) structural domains, with
493 six subclasses that grouped the coil, extended conformation and turn residues that define the six
494 beta strands characteristic of MAX effectors. Structural features were determined using MAX
495 effector structures predicted by homology modeling and computing with STRIDE.

496 For the relationship between the probability of residues being polymorphic and the
497 structural features, we used a generalized linear mixed model with a set of predictor variables.
498 The fixed effects were the structural features and the model was fitted using maximum
499 likelihood estimation, with MAX effector modeled as a random effect. Only a single structural
500 feature, the solvent accessibility had a statistically significant effect on the probability of amino
501 acid polymorphism (S1 Text). Predicted probabilities of amino acid change were higher for
502 accessibility class 60-100% (95% prediction interval: 0.1294-0.1295), than for accessibility
503 classes 30-60% (95% prediction interval: 0.1029-0.1049) and 0-30% (95% prediction interval:
504 0.0834-0.0835). A major factor explaining the variability of the response variable turned out to
505 be the identity of MAX effectors. Indeed, the random effect (σ: 1.17) had a larger standard
506 deviation than the largest fixed-effect factor (coefficient for accessibility class 60-100%: 0.64)
507 (S1 Text).

508 To visualize the localization of polymorphisms, we projected the distribution of amino
509 acid changes on the surface of protein structure models of two types of MAX effectors: those
510 with an avirulence function and MAX effectors with the strongest signatures of multiallelic

511 balancing selection (π_N/π_S>2; Figure 6C; S9 Figure). For the three effectors whose binding
512 interfaces have been experimentally characterized (AVR1-CO39, AVR-Pia, AVR-Pik [54-57]), a
513 substantial fraction of amino acid changes co-localized with residues interacting with immune
514 receptors and, presumably, also with their host target proteins. Polymorphic residues are,
515 therefore, potentially good predictors for binding interfaces in MAX effectors and the specific
516 surface regions, where polymorphisms cluster in several MAX effectors (such as MoToxB,
517 MAX58, MAX87, MAX69, or MAX50) could correspond to interfaces that bind immune receptors
518 and/or host target proteins (Figure 6C).

519 To determine which parts of the MAX structure is most responsible for the high level of
520 standing variation in these effectors, we calculated for the three different structural features and
521 their subclasses the non-synonymous nucleotide diversity π_N (S10 Figure). We restricted this
522 analysis to the 25 MAX effectors exhibiting balancing selection (π_N/π_S >1) and we used π_N and
523 not π_N/π_S because the latter tended to be undefined due to relatively short sequence lengths.
524 Non-synonymous nucleotide diversity π_N differed between subclasses (Kruskal-Wallis tests
525 H=8.504, $d.f.$=2, p=0.014; Figure 6D; S6 Table) and was higher at coils and turns, than at
526 extended conformations (coils: π_N=0.0191; turns: π_N=0.0167; extended conformations:
527 π_N=0.0086; posthoc Mann-Whitney U-tests, p<0.05). Ten and nine MAX effectors displayed
528 their highest values of π_N in coils and turns, respectively. π_N did not significantly differ between
529 relative solvent accessibility subclasses (Kruskal-Wallis tests H=2.308, $d.f.$=2, p=0.315), but
530 differences were marginally significant between structural domains (Kruskal-Wallis tests
531 H=11.035, $d.f.$=5, p=0.051). The third, fourth, and fifth beta strands displayed the highest levels
532 of non-synonymous diversity (π_N=0.0341, π_N=0.0341, and π_N=0.0341, respectively), and 18 out
533 of 25 MAX effectors displayed their highest values of π_N at one of these three beta strands (S6
534 Table).

535 To identify the parts of MAX effector structures that experience multiallelic directional
536 selection, we analyzed the 23 proteins with d_N/d_S >1. This showed that differences in d_N/d_Swere
537 most pronounced between subclasses of structural features (Kruskal-Wallis tests H=5.499,

538 *d.f.*=2, p=0.064), with higher average d_N/d_S values for coils and turns (d_N/d_S=2.490 and
539 d_N/d_S=1.184, respectively) than extended conformations (d_N/d_S=0.573) (Figure 6E; S6 Table).
540 The average d_N/d_S was also close to one for the 30-60% subclass of relative solvent accessibility
541 (d_N/d_S=0.994), and 12 MAX effectors with signatures of directional selection had their highest
542 d_N/d_S values for this subclass, although differences were not significant.

543 Overall, these analyses show that multiallelic balancing and directional selection acted
544 preferentially on coils and turns, but that the impact of two forms of selection on structural
545 domains and solvent accessibility subclasses differs.

546

547

548

549 **DISCUSSION**

550 *MAX effectors as model systems to investigate effector evolution.*

551 Effectors involved in coevolutionary interactions with host-derived molecules are expected to
552 undergo non-neutral evolution. Yet, the role of natural selection in shaping polymorphism and
553 divergence at effectors has remained largely elusive [2]. Despite the prediction of large and
554 molecularly diversified repertoires of effector genes in many fungal genomes, attempts to probe
555 into the evolutionary drivers of effector diversification in plant pathogenic fungi have been
556 hindered by the fact that, until recently, no large effector families had been identified. In this
557 study, we overcome the methodological and conceptual barrier imposed by effector hyper-
558 diversity by building on our previous discovery [17] of an important, structurally-similar, but
559 sequence-diverse family of fungal effectors called MAX. We used a combination of structural
560 modeling, evolutionary analyses, and molecular plant pathology experiments to provide a
561 comprehensive overview of polymorphism, divergence, gene expression, and presence/absence
562 at MAX effectors. When analyzed species-wide or at the level of sub-specific lineages, ratios of
563 non-synonymous to synonymous nucleotide diversity, as well as ratios of non-synonymous to
564 synonymous substitutions, were consistently higher at MAX effectors than at other loci. At the

565 species level, the two ratios were also significantly higher than expected under the standard
566 neutral model for a large fraction of MAX effectors. The signatures of adaptive evolution
567 detected at MAX effectors, combined with their extensive presence/absence variation, are
568 consistent with their central role in coevolutionary interactions with host-derived ligands that
569 impose strong selection on virulence effectors.
570
571 *Adaptive evolution at MAX effectors*
572 Rates of evolution determined from orthologous comparisons with outgroup sequences revealed
573 that, for a large fraction of MAX effectors, non-synonymous changes have accumulated faster
574 than synonymous changes. The fast rate of amino-acid change at MAX effectors is consistent
575 with a classic arms race scenario, which entails a series of selective sweeps as new virulent
576 haplotypes - capable of avoiding recognition by plant immune receptors that previously
577 prevented pathogen multiplication - spread to high frequency [61, 62]. Furthermore, it is
578 important to note that although large values of the d_N/d_S ratio provide strong evidence for
579 directional selection, small values do not necessarily indicate the lack thereof, as d_N/d_S ratios
580 represent the integration of genetic drift, constraint, and adaptive evolution [50][63]. Much of
581 the adaptive changes at MAX effectors probably took place before the radiation of *P. oryzae* on
582 its various hosts. However, the observation that d_N/d_S values determined from orthologous
583 comparisons with outgroup are higher at the species level than at the sub-specific lineage level
584 indicates that part of the signal of directional selection derives from inter-lineage amino acid
585 differences associated with host-specialization. Our structural modeling indicates that it is
586 preferentially "turns" and "coils", but also residues with intermediate solvent accessibility,
587 which often evolve at an unusually fast rate, and therefore that these are probably the residues
588 of MAX proteins preferentially involved in coevolutionary interactions with host-derived
589 molecules.
590 MAX effectors are characterized by a remarkable excess of non-synonymous
591 polymorphism, compared to synonymous polymorphism, at the species level, but also - albeit to

592 a lesser extent - at the sub-specific lineage level. This raises the question of how polymorphisms
593 are maintained in the face of adaptive evolution, given that selective sweeps under a classic arms
594 race scenario are expected to erase variation [6, 9]. Directional selection restricted to host-
595 specific lineages - i.e., local adaptation - may contribute to the signature of multiallelic balancing
596 selection observed at the species level. The observation of a positive correlation between π_N/π_S
597 and the differentiation statistic F_{ST}, together with the fact that most MAX effectors are
598 monomorphic at the lineage level, are consistent with a role of divergent selection exerted by
599 hosts in the maintenance of species-wide diversity at MAX effectors. However, the finding that
600 MAX effectors with a defined π_N/π_S at the sub-specific lineage level (i.e., MAX effectors with
601 $\pi_S \neq 0$) present a higher ratio than the other genes also indicates that the adaptive evolution
602 process is not simply one of successive selective sweeps. This is consistent with balancing
603 selection acting at the lineage level, through which polymorphisms in MAX virulence effectors
604 are maintained due to spatiotemporal variation in selection pressures posed by the hosts – a
605 process known as the trench-warfare model [61]. Our structural modeling suggests in particular
606 that the "coils" and "turns" are the preferred substrate of these coevolutionary interactions
607 leading to the maintenance of elevated polymorphism at MAX virulence effectors. Mirroring the
608 existence of hypervariable MAX effectors, we also detect a substantial proportion of effectors
609 that show no variability, either at the species level or at the lineage level. However, the lack of
610 variability does not necessarily mean they have no impact on virulence. It is possible that their
611 role in virulence is associated with evolutionary constraints that restrict their variability to a
612 limited region of sequence space. Core, monomorphic MAX effectors could be prime targets for
613 genetically-engineered NLRs [64].
614
615 *Expression kinetics of MAX effectors*
616 Expression profiling showed that the MAX effector repertoire was induced specifically and
617 massively during infection. Depending on the host genotype, between 64 and 78% of the MAX
618 effectors were expressed and expression was particularly strong during the early stages of

619 infection. These findings are consistent with previous studies that analyzed genome-wide gene
620 expression during rice infection or specifically addressed MAX effector expression, and they
621 reinforce the hypothesis that MAX effectors are crucial for fungal virulence and specifically
622 involved in the biotrophic phase of infection [17, 38].

623 How this coordinated deployment of the MAX effectors is regulated remains largely
624 unknown. Genome organization does not seem to be a major factor, since MAX effectors do not
625 colocalize and more generally, there is no clustering of effectors in the *P. oryzae* genome, only a
626 slight enrichment in subtelomeric regions [38, 65]. This differs from other pathogenic fungi,
627 such as *Leptosphaeria maculans*, for which early-expressed effectors are clustered in AT-rich
628 isochores, and co-regulated by epigenetic mechanisms [66]. Analysis of promoter regions of
629 MAX effectors did not identify common DNA motifs that may be targeted by transcription
630 factors, and no such transcriptional regulators that would directly regulate large fractions of the
631 effector complement of *P. oryzae* have been identified yet. The few known transcriptional
632 networks controlled by regulators of *P. oryzae* pathogenicity generally comprise different
633 classes of fungal virulence genes, such as secondary metabolism genes or carbohydrate-active
634 enzymes; they are not restricted to effectors. Recently, it was shown that Rgs1, a regulator of G-
635 protein signaling necessary for appressorium development, represses the expression of 60
636 temporally co-regulated effectors in axenic culture and during the pre-penetration stage of plant
637 infection [67]. Of these, six belong to the MAX family and their expression is affected in *cer7* and
638 Δ*rgs1* mutants: MGG_1004T0 [15, 38], MGG_15443T0 [16, 38], MGG_08817T0 [15],
639 MGG_17266T0 [15-17, 38] and MAX15 (MGG_05424T0) and MAX67 (MGG_16175T0), both
640 identified in our study. This represents only 5% of the MAX effectors predicted to date (S2
641 Table) and suggests that multiple complementary mechanisms contribute to the precise
642 coordination of MAX effector expression during rice invasion.

643 Expression profiling also revealed that the plant host genotype strongly influenced the
644 expression of the MAX effector repertoire, suggesting that plasticity in effector expression may
645 contribute to the adaptation of *P. oryzae* to its hosts. MAX effectors were stronger expressed in

646 the more resistant Kitaake rice variety than in highly susceptible Maratelli rice. This is
647 reminiscent of other pathogenic fungi, such as *Fusarium graminearum* and *L. maculans,* for
648 which a relationship between host resistance levels and effector expression was established [68,
649 69]. An expression analysis of MAX effectors in isolates infecting a wider range of host plants
650 with varying resistance levels could be conducted to further investigate the connection between
651 plant resistance and MAX effectors' expression.
652
653 *Presence/absence polymorphism of MAX effectors*
654 Pangenome analyses demonstrated extensive variability in the MAX effector repertoire. In cases
655 where MAX effectors are specifically absent from some lineages, but present in most or all
656 others, it is tempting to hypothesize that they experienced immune-escape loss-of-function
657 mutations that directly contributed to host range expansion or host shifts. A possible example of
658 such a mechanism is the non-MAX effector *PWT3* of *P. oryzae* that is specifically absent from the
659 *Triticum*-infecting lineage [29]. PWT3 triggers resistance in wheat cultivars possessing the
660 *RWT3* resistance gene [70], and its loss was shown to coincide with the widespread deployment
661 of *RWT3* wheat. Similarly, the loss of the effector *AVR1-CO39* (*MAX86*), which is specifically
662 absent from the *Oryza*-infecting lineage and that is detected by the rice NLR immune receptor
663 complex RGA4/RGA5, has been suggested to have contributed to the initial colonization of rice
664 by the *Setaria*-infecting lineage [20, 31, 71]. Two other orthologous *P. oryzae* effectors, *PWL1*
665 and *PWL2*, exclude Eleusine and rice-associated isolates from infecting *Eragrostis curvula*, and
666 can, therefore, also be considered as host-specificity determinants [28, 72]. Interestingly,
667 ALPHAFOLD predicts PWL2 to adopt a MAX effector fold [73]. In our study, however, gene knock-
668 in experiments with *MAX79*, *MAX83,* and *MAX89* - specifically absent from the *Oryza*-infecting
669 lineage - did not reveal a strong effect on virulence towards a large panel of rice varieties. Hence,
670 unlike AVR1-CO39, these effectors are not key determinants of host-specificity. This suggests
671 that overcoming non-host resistance is not the only and maybe not the main evolutionary
672 scenario behind the specific loss of MAX effectors in the *Oryza*-infecting lineage. A possible

alternative mechanism that can explain massive MAX effector loss during host shifts is a lack of functionality in the novel host. Some MAX effectors from a given lineage may have no function in the novel host, simply because their molecular targets are absent or too divergent in the novel host. Cellular targets of fungal effectors remain unknown for the most part, but knowledge of the molecular interactors of MAX effectors may help shed light on the drivers of their presence/absence polymorphism.

Concluding remarks

The discovery of large, structurally-similar, effector families in pathogenic fungi and the increasing availability of high-quality whole genome assemblies and high-confidence annotation tools, pave the way for in-depth investigations of the evolution of fungal effectors by interdisciplinary approaches combining state-of-the-art population genomics, protein structure analysis, and functional approaches. Our study on MAX effectors in the model fungus and infamous cereal killer *P. oryzae* demonstrates the power of such an approach. Our investigations reveal the fundamental role of directional and balancing selection in shaping the diversity of MAX effector genes and pinpoint specific positions in the proteins that are targeted by these evolutionary forces. This type of knowledge is still very limited on plant pathogens, and there are very few studies compared to the plethoric literature on the evolution of virulence factors in human pathogens. Moreover, by revealing the concerted and plastic deployment of the MAX effector repertoire, our study highlights the current lack of knowledge on the regulation of these processes. A major challenge will now be to identify the regulators, target proteins and mode of action of MAX effectors, in order to achieve a detailed understanding of the relationships between the structure, function and evolution of these proteins.

METHODS

Genome assemblies, gene prediction, and pan-genome analyses

Among the 120 genome assemblies included in our study, 66 were already assembled and publicly available, and 54 were newly assembled (S1 Table). For the 54 newly assembled genomes, reads were publicly available for 50 isolates, and four additional isolates were sequenced (available under BioProject PRJEB47684). For the four sequenced isolates, DNA was extracted using the same protocol as in ref. [74]. TruSeq nano kits were used to prepare DNA libraries with insert size of ~500bp for 150 nucleotide paired-end indexed sequencing with Illumina HiSeq 3000. For the 54 newly generated assemblies, CUTADAPT [75] was used for trimming and removing low-quality reads, reads were assembled with ABYSS 1.9.0 [40] using eight different K-mer sizes, and we chose the assembly produced with the K-mer size that yielded the largest N50. For all 120 genome assemblies, genes were predicted by BRAKER 1 [39] using RNAseq data from ref. [21] and protein sequences of isolate 70-15 (Ensembl Fungi release 43). To complement predictions from BRAKER, we also predicted genes using AUGUSTUS 3.4.0 [56] with RNAseq data from ref. [21] , protein sequences of isolate 70-15 (Ensembl Fungi release 43), and *Magnaporthe grisea* as the training set. Gene predictions from BRAKER and AUGUSTUS were merged by removing the genes predicted by AUGUSTUS that overlapped with genes predicted by BRAKER. Repeated elements were masked with REPEATMASKER 4.1.0 (http://www.repeatmasker.org/). The quality of genome assembly and gene prediction was checked using BUSCO 4.0.4 [32]. The homology relationships among predicted genes were identified using ORTHOFINDER v2.4.0 [36]. The size of pan- and core-genomes was estimated using rarefaction, by resampling combinations of one to 119 genomes, taking a maximum of 100 resamples by pseudo-sample size. Sequences for each orthogroup were aligned at the codon level (i.e., keeping sequences in coding reading frame) with TRANSLATORX 1.1 [76], using MAFFT v7 [77] as the aligner and default parameters for GBLOCKS 0.91B [78]. The effect of assembly properties, host of origin, and study of origin on the number of predicted genes computed from

723 the orthology table was analyzed in PYTHON 3.7 using the function in PEARSONR in SCIPY.STATS
724 1.10.1, and functions FORMULA.API.OLS and STATS.ANOVA.ANOVA_LM in STATSMODELS 0.15.0.
725
726
727 *Identification of effectors sensu lato, and MAX effectors*
728 We predicted the secretome by running SIGNALP 4.1 [41], TARGETP 1.1 [42], and PHOBIUS 1.01 [62]
729 to identify signal peptides in the translated coding sequences of 12000 orthogroups. Only
730 proteins predicted to be secreted by at least two methods were retained. Transmembrane
731 domains were identified using TMHMM [43] and proteins with a transmembrane domain
732 outside the 30 first amino acids were excluded from the predicted secretome. Endoplasmic
733 reticulum proteins were identified with PS-SCAN
734 (https://ftp.expasy.org/databases/prosite/ps_scan/), and excluded.
735 To identify MAX effectors, we used the same approach as in the original study that
736 described MAX effectors [17]. We first used PSI-BLAST 2.6.0 [33] to search for homologs of
737 known MAX effectors (AVR1-CO39, AVR-Pia, AvrPiz-t, AVR-PikD, and ToxB) in the predicted
738 secretome. Significant PSI-BLAST hits (e-value < e-4) were aligned using a structural alignment
739 procedure implemented in TM-ALIGN [35]. Three rounds of HMMER [34] searches were then
740 carried out, each round consisting of alignment using TM-ALIGN version 20140601, model
741 building using HMMBUILD, and HMM search using HMMSEARCH (e-value < e-3). Only proteins with
742 two expected conserved cysteines less than 33-48 amino acids apart were retained in the first
743 two rounds of HMMER searches, as described in ref. [17].
744 Subsequent evolutionary analyses were conducted on three sets of orthogroups: MAX
745 effectors, putative effectors, and other genes. The "MAX" group corresponded to 80 orthogroups
746 for which at least 10% of sequences were identified as MAX effectors. The "secreted proteins"
747 groups corresponded to 3283 orthogroups that were not included in the MAX group, and for
748 which at least 10% of sequences were predicted to be secreted proteins. The last group included
749 the remaining 11404 orthogroups.

750 For missing MAX effector sequences, we conducted an additional similarity search to
751 correct for gene prediction errors. For a given MAX orthogroup and a given isolate, if a MAX
752 effector was missing, we used BLAST-N to search for significant hits using the longest sequence of
753 the orthogroup as the query sequence, and the isolate's genome assembly as the subject
754 sequence (S3 Table). We also corrected annotation errors, such as the presence of very short
755 (typically <50bp) or very long (typically >500bp) introns, missing terminal exons associated
756 with premature stops, or frameshifts caused by indels. All these annotation errors were checked,
757 and corrected manually if needed, using the RNAseq data used in gene prediction in the
758 INTEGRATIVE GENOME VIEWER [79, 80]. We also found that some orthogroups included chimeric
759 genes resulting from the erroneous merging of two genes that were adjacent in assemblies. This
760 was the case for orthogroups OG0000093 and OG0010985, and we used RNA-seq data in the
761 INTEGRATIVE GENOME VIEWER to split the merged genic sequences and keep only the sequence
762 corresponding to a MAX effector.

763 For evolutionary analyses conducted on single-copy orthologs, the 11 orthogroups that
764 included paralogous copies of MAX effectors were split into sets of orthologous sequences using
765 genealogies inferred using RAXML v8 [37], yielding a total of 94 single-copy MAX orthologs, of
766 which 90 orthologs passed our filters on length and sample size to be included in evolutionary
767 analyses (see below). For each split orthogroup, sets of orthologous sequences were assigned a
768 number that was added to the orthogroup's identifier as a suffix (for instance paralogous
769 sequences of orthogroup OG0000244 were split into orthogroups OG0000244_1 and
770 OG0000244_2). Sequences were re-aligned using TRANSLATORX (see above) after splitting
771 orthogroups.

772 All genome assemblies, gene models, aligned coding sequences for all orthogroups, and
773 single-copy orthologs, are available in Zenodo, doi: 10.5281/zenodo.7689273 and doi:
774 10.5281/zenodo.8052494.

775

776 *Analysis of population subdivision*

Population structure was inferred from SNPs identified in GBLOCKS-cleaned alignments of coding sequences at 7317 single-copy core orthologs (described in section *Genome assemblies, gene prediction, and pan-genome analyses*). We kept only one randomly chosen four-fold degenerate synonymous site per single-copy core ortholog. We used the sNMF method from the LEA package in R [47] to infer individual ancestry coefficients in *K* ancestral populations. We used SPLITSTREE version 5.3 [48] to visualize relationships between genotypes in a phylogenetic network, with reticulations to represent the conflicting phylogenetic signals caused by homoplasy.

Homology modeling of MAX effectors

To check that orthogroups predicted to be MAX effectors had the typical 3D structure of MAX effectors with two beta sheets of three beta strands each, eight experimental structures with MAX-like folds were selected as 3D templates for homology modeling (PDB identifiers of the templates: 6R5J, 2MM0, 2MM2, 2MYW, 2LW6, 5A6W, 5Z1V, 5ZNG). For each of the 94 MAX orthologous groups, one representative protein was selected and homology models of this 1D query relative to each 3D template were built using MODELLER [44] with many alternative query-template threading alignments. The structural models generated using the alternative alignments were evaluated using a combination of six structural scores (DFIRE [81], GOAP [82], and QMEAN's E_{1D}, E_{2D}, E_{3D} scores [83]). A detailed description of the homology modeling procedure is provided in S2 Text. The best structural models for the 94 representative sequences of each group of MAX orthologs are available at https://pat.cbs.cnrs.fr/magmax/model/. The correspondence between MAX orthogroups identifiers used in homology modeling and MAX orthogroups identifiers resulting from gene prediction is given in S2 Table. Protein models were visualized with PYMOL 2.5 [84].

Evolutionary analyses

803 Lineage-level analyses were conducted on a dataset from which divergent or introgressed
804 isolates were removed (G17 from *Eragrostis*, Bm88324 & Bd8401 from *Setaria*, 87-120; BF0072
805 and BN0019 from *Oryza*; IR0088 from *Echinochloa*), to limit the impact of population
806 subdivision within lineages. The *Stenotaphrum*-infecting lineage was not included in lineage-
807 level analyses due to the small sample size.

808 Nucleotide diversity [85], synonymous and non-synonymous nucleotide diversity, and
809 population differentiation [52] were estimated using EGGLIB v3 [86] using classes COMPUTESTATS
810 and CODINGDIVERSITY. Sites with more than 30% missing data were excluded. Orthogroups with
811 less than 10 sequences (*nseff*<10, *nseff* being the average number of used samples among sites
812 that passed the missing data filter) or shorter than 30bp (*lseff*<30, *lseff* being the number of sites
813 used for analysis after filtering out sites with too many missing data) were excluded from
814 computations. For analyses of polymorphism at secondary structure annotations, the cutoff on
815 *lseff* was set at 10bp.

816 For the computation of d_N/d_S and quantification of adaptive evolution, we used isolate
817 NI919 of *Pyricularia* sp. LS [58, 59] as the outgroup (assembly GCA_004337975.1, European
818 Nucleotide Archive). Genes were predicted in the outgroup assembly using EXONERATE v2.2
819 CODING2GENOME [87]. For each gene, the query sequence was a *P. oryzae* sequence randomly
820 selected among sequences with the fewest missing data. In parsing EXONERATE output, we
821 selected the sequence with the highest score, with a length greater than half the length of the
822 query sequence.

823 The d_N/d_S ratio was estimated using a maximum likelihood approach (runmode=-2,
824 CodonFreq=2 in CODEML [88]), in pairwise comparisons of protein coding sequences (*i.e.*,
825 without using a phylogeny). For each d_N/d_S we randomly selected 12 ingroup sequences and
826 computed the average d_N/d_S across the 12 ingroup/outgroup pairs.

827 Kruskal-Wallis tests were performed using the SCIPY.STATS.KRUSKAL library in PYTHON 3.7.
828 Posthoc Mann-Whitney U-tests were performed using the SCIKIT_POSTHOCS library in PYTHON 3.7,
829 with p-values adjusted using the Bonferroni-Holm method.

830 Amino acid change data was modeled using a binomial generalized linear mixed model
831 with function GLMER with package LME4 version 1.1-32 in R version 4.1.2. Interactions between
832 predictor variables were not significant and thus not included in the model presented in the
833 Results section.
834
835 *Constructs for the transformation of fungal isolates*
836 PCR products used for cloning were generated using the Phusion High-Fidelity DNA Polymerase
837 (Thermo Fisher) and the primers listed in S7 Table. Details of the constructs are given in S8
838 Table. Briefly, the pSC642 plasmid (derived from the pCB1004 vector), containing a cassette for
839 the expression of a gene of interest under the control of the *AVR-Pia* promoter (*pAVR-Pia*) and
840 the *Neurospora crassa β-tubulin* terminator (*t-tub*), was amplified by PCR with primers oML001
841 and oTK609 for the insertion of MAX genes listed in S9 Table. The MAX genes
842 *Mo_US0071_000070* (*MAX79*), *Mo_US0071_046730* (*MAX89*) and *Mo_US0071_115900*
843 (*MAX83*), amplified by PCR from genomic DNA of the *P. oryzae* isolate US0071, were inserted
844 into this vector using the Gibson Assembly Cloning Kit (New England BioLabs). The final
845 constructs were linearized using the KpnI restriction enzyme (Promega) before *P. oryzae*
846 transformation.
847
848 *Plant and fungal growth conditions*
849 Rice plants (*Oryza sativa*) were grown in a glasshouse in a substrate of 31% coconut peat, 30%
850 Baltic blond peat, 15% Baltic black peat, 10% perlite, 9% volcanic sand, and 5% clay,
851 supplemented with 3.5 $g.L^{-1}$ of fertilizer (Basacote® High K 6M, NPK 13-5-18). Plants were
852 grown under a 12h-light photoperiod with a day-time temperature of 27°C, night-time
853 temperature of 21°C, and 70% humidity. For spore production, the wild-type and transgenic
854 isolates of *P. oryzae* Guy11 were grown for 14 days at 25°C under a 12h-light photoperiod on
855 rice flour agar medium (20 $g.L^{-1}$ rice seed flour, 2.5 $g.L^{-1}$ yeast extract, 1.5% agar, 500.000U
856 penicillin g), supplemented with 240 $\mu g.ml^{-1}$ hygromycin for transgenic isolates. For mycelium

857 production, plugs of mycelium of *P. oryzae* Guy11 were grown in liquid medium (10 $g.L^{-1}$
858 glucose, 3 $g.L^{-1}$ KNO_3, 2 $g.L^{-1}$ KH_2PO_4, 2,5 $g.L^{-1}$ yeast extract, 500 000U penicillin g) for 5 days at
859 25°C in the dark under agitation.
860
861 *Fungal transformation*
862 Protoplasts from the isolate Guy11 of *P. oryzae* were transformed by heat shock with 10µg of
863 KpnI-linearized plasmids for the expression of MAX effectors or RFP as described previously
864 [89]. After two rounds of antibiotic selection and isolation of monospores, transformed isolates
865 were genotyped by Phire Plant Direct PCR (Thermo Scientific) using primers described in S7
866 Table. The Guy11 transgenic isolates expressing *AVR-Pia* and *AVR1-CO39* were previously
867 generated [54, 90].
868
869 *Fungal growth and infection assays*
870 For the analysis of interaction phenotypes, leaves of three-week-old rice plants were spray-
871 inoculated with conidial suspensions (40 000 $conidia.ml^{-1}$ in water with 0.5% gelatin). Plants
872 were incubated for 16 hours in the dark at 25°C and 95% relative humidity, and then grown for
873 six days in regular growth conditions. Seven days after inoculation, the youngest leaf that was
874 fully expanded at the time of inoculation was collected and scanned (Scanner Epson Perfection
875 V370) for further symptoms analyses. Phenotypes were qualitatively classified according to
876 lesion types: no lesion or small brown spots (resistance), small lesions with a pronounced
877 brown border and a small gray center (partial resistance), and larger lesions with a large gray
878 center or dried leaves (susceptibility). For the analysis of gene expression, plants were spray-
879 inoculated with conidial suspensions at 50 000 $conidia.ml^{-1}$ (in water with 0.5% gelatin), and
880 leaves were collected three days after inoculation.
881
882 *RNA extraction and qRT-PCR analysis*

883 Total RNA extraction from rice leaves or Guy11 mycelium and reverse transcription were
884 performed as described by ref. [91]. Briefly, frozen leaves and mycelium were mechanically
885 ground. RNA was extracted using TRI-reagent® (Sigma-Aldrich) and chloroform separation.
886 Denaturated RNA (5µg) was retrotranscribed and used for quantitative PCR using GoTaq qPCR
887 Master Mix according to the manufacturer's instructions (Promega) at a dilution of 1/10 for
888 mycelium and 1/7 for rice leaves. The primers used are described in S7 Table. Amplification was
889 performed as described by ref. [91] using a LightCycler480 instrument (Roche), and data were
890 extracted using the instrument software. To calculate *MAX* gene expressions, the $2^{-\Delta\Delta CT}$ method
891 and primers measured efficiency were used. Gene expression levels are expressed relative to the
892 expression of constitutive reference gene *MoEF1α*.
893
894 *Statistical analyses of phenotypic data*
895 For expression comparison between Kitaake and Maratelli infection, all analyses were
896 performed using R (www.r-project.org). The entire kinetic experiment was repeated three times
897 with five biological replicates for each time point. For each variety, gene, and experimental
898 replicate, values corresponding to the day post-inoculation with the highest median expression
899 were extracted for statistical analyses. Expression data were not normally distributed so for
900 each gene, differences between varieties were evaluated using non-parametric Mann-Whitney
901 U-tests.
902
903 **SUPPLEMENTARY MATERIALS**
904 S1 Table. Genomic assemblies with metadata.
905 S2 Table. Nomenclature of MAX effectors predicted in this study and in previous reports.
906 S3 Table. Presence/absence of MAX effector orthologs.
907 S4 Table. The expression of *MAX79*, *MAX83* and *MAX89* in Guy11 does not trigger recognition in a panel of
908 rice varieties.
909 S5 Table. Gene average of summary statistics of polymorphism, differentiation and divergence.

S6 Table. π_N and d_N/d_S in different classes of secondary structure annotations for MAX effectors with $\pi_N/\pi_S>1$ and $d_N/d_S>1$, respectively.

S7 Table. Primers for cloning and expression analyses.

S8 Table. Vector constructs.

S9 Table. Sequences of the MAX effectors in the isolate US0071 that were used for the complementation of Guy11.

S1 Figure. Effect of assembly properties on the number of genes.

S2 Figure. Expression patterns of MAX effectors during rice infection.

S3 Figure. Differential expression levels of MAX effectors upon infection of two different rice cultivars.

S4 Figure. Nucleotide diversity (π), ratio of non-synonymous to synonymous nucleotide diversity (π_N/π_S), orthogroup frequency for MAX effectors, other secreted proteins, and other genes.

S5 Figure. Frequency of MAX effector orthogroups as a function of the frequency of the adjacent orthogroups in the genome.

S6 Figure. Analyses of population subdivision with sNMF.

S7 Figure. *MAX79*, *MAX83* and *MAX89* are expressed in the transgenic Guy11 isolates upon rice inoculation.

S8 Figure. F_{ST} versus π_N/π_S at MAX effectors

S9 Figure. Amino acid changes segregating in *P. oryzae* at MAX effectors with an avirulence function and MoToxB (first row), and MAX effectors with $\pi_N/\pi_S>2$ (next rows); amino acid changes are shown in dark blue and known binding interfaces in light blue. Note that the π_N/π_S ratio is >2 for the avirulence genes *AvrPiz-t* and AVR-Pik. Proteins are displayed twice, with one copy rotated 180 degrees around a vertical axis. The interface involved in binding with host proteins is known for AVR1-CO39, AVR-Pia, and AVR-Pik only ([54-57]).

S10 Figure. Secondary structure annotations of MAX effectors aligned with TM-ALIGN.

S1 Data. Summary statistics per orthogroup.

S2 Data. Summary statistics per MAX effector ortholog, species wide, and per lineage.

S3 Data. Structural properties and polymorphism of amino acids in MAX effectors.

940 S1 Text. Fitting a generalized linear model to amino acid polymorphism data.

941 S2 Text. Homology modeling procedure.

REFERENCES

1. Schulze-Lefert P, Panstruga R. A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends in Plant Science. 2011;16(3):117-25. doi: 10.1016/j.tplants.2011.01.001.
2. Sánchez-Vallet A, Fouché S, Fudal I, Hartmann FE, Soyer JL, Tellier A, et al. The genome biology of effector gene evolution in filamentous plant pathogens. Annual review of phytopathology. 2018;56:21-40.
3. Haldane JBS. Disease and evolution. Ricerca Scient 1949;19:68-76.
4. Flor HH. Inheritance of pathogenicity in Melampsora lini. Phytopathology. 1942;32:653-69.
5. Barrett JA. Frequency-dependent selection in plant-fungal interactions. Philosophical Transactions of the Royal Society of London B, Biological Sciences. 1988;319(1196):473-83.
6. Bergelson J, Kreitman M, Stahl EA, Tian D. Evolutionary dynamics of plant R-genes. Science. 2001;292(5525):2281-5.
7. Brown JKM. Chance and selection in the evolution of barley mildew. Trends in Microbiology. 1994;2(12):470-5. PubMed PMID: 29.
8. Brown JKM. Durable resistance of crops to disease: a Darwinian perspective. Annual review of phytopathology. 2015;53:513-39.
9. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature. 1999;400(6745):667-71.
10. Gladieux P, van Oosterhout C, Fairhead S, Jouet A, Ortiz D, Ravel S, et al. Extensive immune receptor repertoire diversity in disease-resistant rice landraces. bioRxiv. 2022:2022-12.
11. Bakker EG, Toomajian C, Kreitman M, Bergelson J. A genome-wide survey of R gene polymorphisms in Arabidopsis. The Plant cell. 2006;18(8):1803-18.
12. Bakker EG, Traw MB, Toomajian C, Kreitman M, Bergelson J. Low levels of polymorphism in genes that control the activation of defense response in Arabidopsis thaliana. Genetics. 2008;178(4):2031-43.
13. Ebert D, Fields PD. Host–parasite co-evolution and its genomic signature. Nature Reviews Genetics. 2020;21(12):754-68.
14. Ebbole DJ, Chen M, Zhong Z, Farmer N, Zheng W, Han Y, et al. Evolution and Regulation of a Large Effector Family of Pyricularia oryzae. Molecular Plant-Microbe Interactions. 2021;34(3):255-69.
15. Seong K, Krasileva KV. Computational structural genomics unravels common folds and novel families in the secretome of fungal phytopathogen Magnaporthe oryzae. Molecular Plant-Microbe Interactions. 2021;34(11):1267-80.
16. Seong K, Krasileva KV. Prediction of effector protein structures from fungal phytopathogens enables evolutionary analyses. Nature Microbiology. 2023;8(1):174-87.
17. de Guillen K, Ortiz-Vallejo D, Gracy J, Fournier E, Kroj T, Padilla A. Structure analysis uncovers a highly diverse but structurally conserved effector family in phytopathogenic fungi. PLoS Pathog. 2015;11(10):e1005228.
18. Savary S, Willocquet L, Pethybridge SJ, Esker P, McRoberts N, Nelson A. The global burden of pathogens and pests on major food crops. Nature ecology & evolution. 2019;3(3):430.
19. Fernandez J, Orth K. Rise of a cereal killer: the biology of Magnaporthe oryzae biotrophic growth. Trends in microbiology. 2018;26(7):582-97.
20. Couch BC, Fudal I, Lebrun M-H, Tharreau D, Valent B, van Kim P, et al. Origins of Host-Specific Populations of the Blast Pathogen *Magnaporthe oryzae* in Crop Domestication With Subsequent Expansion of Pandemic Clones on Rice and Weeds of Rice. Genetics. 2005;170(2):613-30. PubMed PMID: 61.
21. Pordel A, Ravel S, Charriat F, Gladieux P, Cros-Arteil S, Milazzo J, et al. Tracing the origin and evolutionary history of Pyricularia oryzae infecting maize and barnyard grass. Phytopathology. 2021;111(1):128-36.
22. Kato H, Yamamoto M, Yamaguchi-Ozaki T, Kadouchi H, Iwamoto Y, Nakayashiki H, et al. Pathogenicity, mating ability and DNA restriction fragment length polymorphisms of Pyricularia populations isolated from Gramineae, Bambusideae and Zingiberaceae plants. Journal of General Plant Pathology. 2000;66:30-47.
23. Urashima AS, Igarashi S, Kato H. Host range, mating type, and fertility of Pyricularia grisea from wheat in Brazil. Plant Disease. 1993;77(12):1211-6.
24. Igarashi S. Pyricularia em trigo. 1. Ocorrencia de Pyricularia sp noestado do Parana. Fitopatol Bras. 1986;11:351-2.
25. Milazzo J, Pordel A, Ravel S, Tharreau D. First scientific report of Pyricularia oryzae causing gray leaf spot disease on perennial ryegrass (Lolium perenne) in France. Plant Disease. 2019;103(5):1024-.

26. Islam MT, Croll D, Gladieux P, Soanes DM, Persoons A, Bhattacharjee P, et al. Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. BMC biology. 2016;14(1):84.
27. Gladieux P, Condon B, Ravel S, Soanes D, Maciel JLN, Nhani A, et al. Gene Flow between Divergent Cereal- and Grass-Specific Lineages of the Rice Blast Fungus Magnaporthe oryzae. mBio. 2018;9(1).
28. Sweigard JA, Carroll AM, Kang S, Farrall L, Chumley FG, Valent B. Identification, Cloning, and Characterization of *PWL2*, a Gene for Host Species Specificity in the Rice Blast Fungus. The Plant cell. 1995;7(8):1221-33. PubMed PMID: 260.
29. Inoue Y, Vy TTP, Yoshida K, Asano H, Mitsuoka C, Asuke S, et al. Evolution of the wheat blast fungus through functional losses in a host specificity determinant. Science. 2017;357(6346):80-3.
30. Asuke S, Tanaka M, Hyon G-S, Inoue Y, Vy TTP, Niwamoto D, et al. Evolution of an Eleusine-specific subgroup of Pyricularia oryzae through a gain of an avirulence gene. Molecular Plant-Microbe Interactions. 2020;33(2):153-65.
31. Zheng Y, Zheng W, Lin F, Zhang Y, Yi Y, Wang B, et al. AVR1-CO39 is a predominant locus governing the broad avirulence of Magnaporthe oryzae 2539 on cultivated rice (Oryza sativa L.). Molecular plant-microbe interactions. 2011;24(1):13-7.
32. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.
33. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research. 1997;25(17):3389-402.
34. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic acids research. 2011;39(suppl_2):29-37.
35. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic acids research. 2005;33(7):2302-9.
36. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome biology. 2015;16(1):157.
37. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.
38. Yan X, Tang B, Ryder LS, MacLean D, Were VM, Eseola AB, et al. The transcriptional landscape of plant infection by the rice blast fungus Magnaporthe oryzae reveals distinct families of temporally co-regulated and structurally conserved effectors. The Plant cell. 2023;35(5):1360-85.
39. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2015;32(5):767-9.
40. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. ABySS: a parallel assembler for short read sequence data. Genome research. 2009;19(6):1117-23.
41. Petersen TN, Brunak S, Von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature methods. 2011;8(10):785-6.
42. Emanuelsson O, Nielsen H, Brunak S, Von Heijne G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. Journal of molecular biology. 2000;300(4):1005-16.
43. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. Journal of molecular biology. 2001;305(3):567-80.
44. Webb B, Sali A. Protein Structure Modeling with MODELLER. Methods Mol Biol. 2021;2199(1940-6029 (Electronic)):239–55.
45. Frishman D, Argos P. Knowledge-based protein secondary structure assignment. Proteins: Structure, Function, and Bioinformatics. 1995;23(4):566-79.
46. Fay JC, Wu C-I. Sequence divergence, functional constraint, and selection in protein evolution. Annual review of genomics and human genetics. 2003;4(1):213-35.
47. Frichot E, Mathieu F, Trouillon T, Bouchard G, François O. Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014;196(4):973-83.
48. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Molecular Biology and Evolution. 2006;23(2):254-67.
49. Tosa Y, Osue J, Eto Y, Oh H-S, Nakayashiki H, Mayama S, et al. Evolution of an avirulence gene, AVR1-CO39, concomitant with the evolution and differentiation of Magnaporthe oryzae. Molecular Plant-Microbe Interactions. 2005;18(11):1148-60.
50. Cesari S, Thilliez G, Ribot C, Chalvon V, Michel C, Jauneau A, et al. The rice resistance protein pair RGA4/RGA5 recognizes the Magnaporthe oryzae effectors AVR-Pia and AVR1-CO39 by direct binding. The

1056 Plant cell. 2013;25(4):1463-81. Epub 2013/04/04. doi: 10.1105/tpc.112.107201. PubMed PMID:
1057 23548743; PubMed Central PMCID: PMCPMC3663280.
1058 51. Okuyama Y, Kanzaki H, Abe A, Yoshida K, Tamiru M, Saitoh H, et al. A multifaceted genomics
1059 approach allows the isolation of the rice Pia-blast resistance gene consisting of two adjacent NBS-LRR
1060 protein genes. The Plant Journal. 2011;66(3):467-79.
1061 52. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. Evolution.
1062 1984;38:1358-70. PubMed PMID: 283.
1063 53. Hahn MW. Molecular population genetics: Oxford University Press; 2018.
1064 54. Ortiz D, De Guillen K, Cesari S, Chalvon V, Gracy J, Padilla A, et al. Recognition of the Magnaporthe
1065 oryzae effector AVR-Pia by the decoy domain of the rice NLR immune receptor RGA5. The Plant cell.
1066 2017;29(1):156-68.
1067 55. Maqbool A, Saitoh H, Franceschetti M, Stevenson CEM, Uemura A, Kanzaki H, et al. Structural basis
1068 of pathogen recognition by an integrated HMA domain in a plant NLR immune receptor. Elife. 2015;4.
1069 56. Guo L, Cesari S, de Guillen K, Chalvon V, Mammri L, Ma M, et al. Specific recognition of two MAX
1070 effectors by integrated HMA domains in plant immune receptors involves distinct binding surfaces.
1071 Proceedings of the National Academy of Sciences. 2018;115(45):11637-42.
1072 57. Bentham AR, Petit-Houdenot Y, Win J, Chuma I, Terauchi R, Banfield MJ, et al. A single amino acid
1073 polymorphism in a conserved effector of the multihost blast fungus pathogen expands host-target binding
1074 spectrum. PLoS Pathogens. 2021;17(11):e1009957.
1075 58. Hirata K, Kusaba M, Chuma I, Osue J, Nakayashiki H, Mayama S, et al. Speciation in Pyricularia
1076 inferred from multilocus phylogenetic analysis. Mycological Research. 2007;111(7):799-808.
1077 59. Gómez Luciano LB, Tsai IJ, Chuma I, Tosa Y, Chen Y-H, Li J-Y, et al. Blast fungal genomes show
1078 frequent chromosomal changes, gene gains and losses, and effector gene turnover. Molecular Biology and
1079 Evolution. 2019;36(6):1148-61.
1080 60. Yang Z. Likelihood ratio tests for detecting positive selection and application to primate lysozyme
1081 evolution. Molecular biology and evolution. 1998;15(5):568-73.
1082 61. Clay K, Kover PX. The Red Queen hypothesis and plant/pathogen interactions. Annual review of
1083 Phytopathology. 1996;34(1):29-50.
1084 62. Van Valen L. A new evolutionary law. 1973.
1085 63. Yang Z, Bielawski JP. Statistical methods for detecting molecular adaptation. Trends in ecology &
1086 evolution. 2000;15(12):496-503.
1087 64. Kourelis J, Marchal C, Posbeyikian A, Harant A, Kamoun S. NLR immune receptor–nanobody
1088 fusions confer plant disease resistance. Science. 2023;379(6635):934-9.
1089 65. Chiapello H, Mallet L, Guerin C, Aguileta G, Amselem J, Kroj T, et al. Deciphering Genome Content
1090 and Evolutionary Relationships of Isolates from the Fungus Magnaporthe oryzae Attacking Different Host
1091 Plants. Genome Biol Evol. 2015;7(10):2896-912. Epub 2015/10/11. doi: 10.1093/gbe/evv187. PubMed
1092 PMID: 26454013; PubMed Central PMCID: PMCPMC4684704.
1093 66. Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, et al. Epigenetic control of
1094 effector gene expression in the plant pathogenic fungus Leptosphaeria maculans. PLoS genetics.
1095 2014;10(3):e1004227.
1096 67. Tang B, Yan X, Ryder LS, Cruz-Mireles N, Soanes DM, Molinari C, et al. Rgs1 is a regulator of
1097 effector gene expression during plant infection by the rice blast fungus Magnaporthe oryzae. bioRxiv.
1098 2022:2022-09.
1099 68. Fall LA, Salazar MM, Drnevich J, Holmes JR, Tseng M-C, Kolb FL, et al. Field pathogenomics of
1100 Fusarium head blight reveals pathogen transcriptome differences due to host resistance. Mycologia.
1101 2019;111(4):563-73.
1102 69. Sonah H, Zhang X, Deshmukh RK, Borhan MH, Fernando WGD, Belanger RR. Comparative
1103 transcriptomic analysis of virulence factors in Leptosphaeria maculans during compatible and
1104 incompatible interactions with canola. Frontiers in plant science. 2016;7:1784.
1105 70. Arora S, Steed A, Goddard R, Gaurav K, O'Hara T, Schoen A, et al. A wheat kinase and immune
1106 receptor form host-specificity barriers against the blast fungus. Nature Plants. 2023:1-8.
1107 71. Farman ML, Eto Y, Nakao T, Tosa Y, Nakayashiki H, Mayama S, et al. Analysis of the structure of the
1108 *AVR1-Co39* avirulence locus in virulent rice-infecting isolates of *Magnaporthe grisea*. Molecular Plant-
1109 Microbe Interactions. 2002;15:6-16. PubMed PMID: 90.
1110 72. Kang S, Sweigard Ja Fau - Valent B, Valent B. The PWL host specificity gene family in the blast
1111 fungus Magnaporthe grisea. Molecular Plant Microbe Interactions. 1995;(0894-0282 (Print)).
1112 73. Brabham HJ, Gómez De La Cruz D, Were V, Shimizu M, Saitoh H, Hernández-Pinzón I, et al. Barley
1113 MLA3 recognizes the host-specificity determinant PWL2 from rice blast (M. oryzae). bioRxiv. 2022:2022-
1114 10.

74. Thierry M, Charriat F, Milazzo J, Adreit H, Ravel S, Cros-Arteil S, et al. Maintenance of divergent lineages of the Rice Blast Fungus Pyricularia oryzae through niche separation, loss of sex and post-mating genetic incompatibilities. PLoS pathogens. 2022;18(7):e1010687.
75. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):10-2.
76. Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic acids research. 2010;38(suppl_2):W7-W13.
77. Katoh K, Toh H. Recent developments in the MAFFT multiple sequence alignment program. Briefings in Bioinformatics. 2008;9(4):286-98. doi: 10.1093/bib/bbn013. PubMed PMID: ISI:000256756400004.
78. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. 2000;17(4):540-52.
79. Robinson JT, Thorvaldsdóttir H, Turner D, Mesirov JP. igv. js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Bioinformatics. 2023;39(1):btac830.
80. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nature biotechnology. 2011;29(1):24-6.
81. Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein science. 2002;11(11):2714-26.
82. Zhou H, Skolnick J. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. Biophysical journal. 2011;101(8):2043-52.
83. Benkert P, Tosatto SCE, Schomburg D. QMEAN: A comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics. 2008;71(1):261-77.
84. Schrödinger L, DeLano W. PyMOL. 2020.
85. Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437-60.
86. Siol M, Coudoux T, Ravel S, De Mita S. EggLib 3: A python package for population genetics and genomics. Molecular Ecology Resources. 2022;22(8):3176-87.
87. Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. BMC bioinformatics. 2005;6(1):1-11.
88. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Computer applications in the biosciences. 1997;13(5):555-6.
89. Villalba F, Collemare J, Landraud P, Lambou K, Brozek V, Cirer B, et al. Improved gene targeting in Magnaporthe grisea by inactivation of MgKU80 required for non-homologous end joining. Fungal Genetics and Biology. 2008;45(1):68-75.
90. Ribot C, Césari S, Abidi I, Chalvon V, Bournaud C, Vallet J, et al. The M agnaporthe oryzae effector AVR 1–CO 39 is translocated into rice cells independently of a fungal-derived machinery. The Plant Journal. 2013;74(1):1-12.
91. Pélissier R, Buendia L, Brousse A, Temple C, Ballini E, Fort F, et al. Plant neighbour-modulated susceptibility to pathogens in intraspecific mixtures. Journal of experimental botany. 2021;72(18):6570-80.