Coexistence of multiple endemic and pandemic lineages of the rice blast pathogen

Pierre Gladieux[1], Sébastien Ravel[2], Adrien Rieux[3], Sandrine Cros-Arteil[1], Henri Adreit[2], Joëlle Milazzo[2], Maud Thierry[2], Elisabeth Fournier[1], Ryohei Terauchi[4], Didier Tharreau[2]

[1]INRA, UMR BGPI, Montpellier, France; [2]CIRAD, UMR BGPI, Montpellier, France; [3]CIRAD, UMR PVBMT, St Pierre de la Reunion; [4]Iwate Biotechnology Research Center, Kitakami, Iwate, Japan

Abstract

The rice blast fungus *Magnaporthe oryzae* (syn. *Pyricularia oryzae*) causes the most damaging rice disease. Yet, little is known about the genetic makeup of this important system for plant pathology. Using whole genome resequencing of a worldwide collection of isolates, we identified four major *M. oryzae* lineages mainly associated with japonica or indica rice, including one pandemic lineage on each rice subspecies. Tip-dating calibration indicated that *M. oryzae* lineages separated about a millenium ago, much later than the initial domestication of rice. The major lineage endemic to Southeast continental Asia displayed signatures of sexual recombination and evidence for having received DNA from multiple lineages.Tests of weak selection revealed that the pandemic spread of clonal lineages entailed an evolutionary 'cost', in terms of an accumulation of deleterious mutations. Our work reveals the coexistence of multiple endemic and pandemic lineages with contrasting population and genetic characteristics within a widely distributed pathogen.

Introduction

Fungal plant pathogens provide many examples of geographically widespread, often clonal, lineages displaying strong ability to adapt rapidly to anthropogenic changes, such as fungicides or resistant varieties, despite extremely low levels of population genetic diversity [1, 2]. An accurate characterization of the population biology and evolutionary history of these organisms is crucial for understanding the factors underpinning their emergence and spread, and for providing new,

powerful and enduring solutions to control these factors. First, knowledge of the origins and lines of descent connecting extant pathogen population informs our understanding of the tempo and mode of disease emergence and subsequent dispersal [2, 3]. Second, inferring the history and structure of pathogen populations can enable the identification of disease reservoirs and improve our understanding of the transmissibility and longevity of populations [4, 5]. Last, quantification of the amount and distribution of genetic variation across space and time provides a population-level genomic framework for defining molecular markers to assist in the control of pathogens and for investigating the molecular underpinnings of phenotypic and fitness differences between divergent pathogen lineages.

Rice blast is the most damaging rice disease worldwide [6, 7]. Rice blast is caused by the ascomycete fungus *Magnaporthe oryzae* (syn. *Pyricularia oryzae*), which became a model for plant pathology in parallel to the development of rice as a model crop species [6, 8-10]. The rice-infecting lineage of *M. oryzae* coexists with multiple host-specialized and genetically divergent lineages infecting other cereals and grasses [11-13]. The lineage infecting foxtail millet (*Setaria italica*, hereinafter « Setaria ») is the closest relative to the rice-infecting lineage and it was therefore postulated that rice blast has emerged following a host-shift from Setaria ca. 2500 to 7500 years ago [14], at a time when Setaria was the most favored staple food in East Asia [15, 16]. *Magnaporthe oryzae* infects the two major subspecies of rice, *Oryza sativa* ssp. *indica* and *japonica* (hereinafter indica and japonica, respectively). Population genomics studies support a model in which de novo domestication occurred only once in japonica, with later divergence into temperate and tropical japonica, and introgressive hybridization from japonica leading to domesticated indica [17-21]. Using microsatellite markers, Saleh et al. [22] identified multiple endemic and pandemic genetic pools of rice infecting strains, but could not resolve evolutionary relationships among them. At the level of rice agrosystems, rice blast is a dynamic threat to varietal resistance, which provided many examples of rapid adaptation to varietal resistance [23]. This is despite the fact that rice blast has little diversity and is infertile and/or asexual in most rice-growing areas [23, 24], making the pathogen particularly exposed to the « cost of pestification » (by analogy with the cost of domestication [25-28]) that posits that the combination of small effective

population size, strong selection on pestification genes and the lack of recombination causes the accumulation of deleterious mutations [29]. Potential limitations to adaptation could be counterbalanced in *M. oryzae* by boom-and-bust cycles, allowing adaptation to occur during the boom phases when short-term effective population size is large [2, 30]. Adaptive mutations can also be introduced through cryptic genetic exchanges with conspecifics or heterospecifics [31-34], but these mechanisms remain to be investigated in natural populations of *M. oryzae* [35]. An accurate understanding of the population genetics of successful clonal fungal pathogens such as *M. oryzae* can provide important insights into the genomic and eco-evolutionary processes underlying pathogen emergence and adaptation to anthropogenic changes.

Here, we used whole-genome re-sequencing data of time- and space-distributed samples of *M. oryzae* and population genomic analyses to address the following questions: What is the population structure of *M. oryzae* and is it comprised of relatively ancient or recent clonal lineages? What is the colonization history of temperate japonica, tropical japonica and indica japonica rice by *M. oryzae*? Do *M. oryzae* lineages display differences in pathogenicity toward rice subspecies?Can we identify genetic exchanges among rice-infecting lineages and what were the genomic regions exchanged? Is there evidence for a cost of pestification in terms of an accumulation of deleterious mutations?

Results

Genome sequencing and SNP calling. To elucidate the emergence, diversification and spread of *M. oryzae* in rice agrosystems, we interrogated genome-wide variation across geographically widespread samples. Twenty-five and eighteen genomes sequenced using single-end reads and paired-end reads Illumina technology, respectively, were combined with seven published genomes sequenced using Solexa and mate-pair Titanium [9, 11], resulting in a total of fifty genomes available for analysis (Supplementary Table S1). Fourty-five isolates originated from cultivated rice (*Oryza sativa*), 4 isolates from cultivated barley (*Hordeum vulgare*), and one from foxtail millet (*Setaria italica*). The sample set included multiple samples from geographically separated areas (North and South America, South, Southeast and East Asia, sub-Saharan Africa, Europe and the

Mediterranean area), the reference lab strain 70-15 and its parent GY11 from French Guyana. Nine samples were collected on tropical japonica rice, seven on temperate japonica, fifteen on indica, and three on hybrid elite varieties. Sequencing reads were aligned to the 41.1Mb reference genome 70-15. Average sequencing depth ranged from 5X to 64X for genomes sequenced using single-end reads, and 5X to 10X for genomes sequenced using paired-end reads (Supplementary Table S1). SNP calling identified 182,804 biallelic single nucleotide polymorphisms (SNPs) distributed along seven chromosomes (95,925 SNPs excluding the Setaria-infecting lineage).

Population subdivision, genealogical relationships and levels of genetic variation. We used Discriminant Analysis of Principal Components to determine the number of lineages represented in our dataset. Increasing the number of cluster K from 2 to 5 progressively individualized the four lineages previously identified in Asia by Saleh et al. [22] using microsatellite data, and individualized a cluster of three strains collected in Yunnan and Hunan Provinces of China (Figure 1). Further increasing K subdivided this latter Yunnan/Hunan cluster. Barley-infecting isolates clustered within rice-infecting lineage 1, confirming previous phylogenetic work [11, 12]. Barley is a universal suscept to rice-infecting isolates at least in lab conditions. However, the barley isolates included in this study were collected in Thailand and no major blast epidemic has been reported on this host since then in this area, indicating that barley is a minor host for rice-infecting populations.

To determine if the clusters observed at K>4 in the DAPC represented new independent lineages or subdivision of the main clusters, we inferred a genome genealogy using RAxML [36]. We based the analysis on a dataset combining the full set of SNPs and monomorphic sites, rather than just SNPs, to increase topological and branch lengths accuracy [37]. The total evidence genealogy revealed four lineages corresponding to lineages 1 to 4 in Saleh et al. [22] and two new lineages (named lineages 5 and 6) corresponding to the three-individuals cluster observed at K=5 in the DAPC (Figure 1).

Absolute divergence (dxy) between pairs of lineages was on the order of 10^{-4} differences per base pair and was highest in comparisons with lineage 6 (Supplementary table S2). Nucleotide diversity within lineages was an order of magnitude smaller than divergence in lineages 2 to 4 (θ_w per site:

5.2e-5 to 7.2e-5, π per site: 4.5e-5 to 4.9e-5) and it was highest in lineage 1 (θ_w per site: 2.3e-4; π per site: 2.1e-4) (Table 1). Tajima's D was negative in all lineages, indicating an excess of low frequency polymorphism, and values were closer to zero in lineage 1 and 4 (D=-0.56 and D=-0.82, respectively) than in lineages 2 and 3 (D=-1.45 and D=-1.72).

Footprints of natural selection and the cost of pestification. The standard neutral model of molecular evolution was tested using the McDonald-Kreitman method based on genome-wide patterns of synonymous and non-synonymous variation (Table 2). The null hypothesis could be rejected for all four lineages ($p<0.0001$). The neutrality index, which quantifies the direction and degree of departure from neutrality, was above 1, indicating an excess of amino acid polymorphism. This pattern suggests that lineages 1 to 4 have accumulated slightly deleterious mutations during divergence from the Setaria-infecting lineage. Under near neutrality, the ratio of non-synonymous to synonymous nucleotide diversity π_N/π_S gives an estimate of the proportion of effectively neutral mutations that are strongly dependent on the effective population size N_e [38]. The π_N/π_S ratios ranged from 0.43 in lineage 1 to 0.61 in lineage 4, and were intermediate in lineages 2 and 3 (π_N/π_S=0.49), and the ratio of non-sense (i.e premature stop codons) to sense non-synonymous mutations ($P_{nonsense}/P_{sense}$) followed the same pattern. Altogether, π_N/π_S and $P_{nonsense}/P_{sense}$ ratios suggest a greater proportion of slightly deleterious mutations segregating in lineage 4, and to a lesser extent in lineages 2 and 3, than in lineage 1. Assuming identical mutation rates, one can also estimate that the long-term population size of lineage 1 (π_S=0.00018/bp) was 2.5 to 3 times higher than other lineages, consistent with the N_e effect on the efficacy of negative selection predicted under near neutrality.

Distribution and reproductive biology of *M. oryzae* lineages. Strains in lineage 1 and 2 originated from rainfed upland rice, including experimental fields. Lineage 2 was exclusively associated with tropical and temperate japonica, while Lineage 1 was sampled on barley, tropical japonica and hybrid rice varieties (Supplementary Table S1; Figure 1). Lineage 1 was restricted to Southeast continental Asia (Laos, Thailand, Yunnan). The reference lab strain GY-11 (=Guy11) was collected in French Guyana in fields cultivated by Hmong refugees who fled Laos in the 1970s. Lineage 2 had a pandemic distribution, and included all European samples.

Lineage 3 and 4 samples originated from irrigated or rainfed upland/lowland rice, and they were associated with indica rice, with the addition of two samples collected on hybrid varieties and one sample originating from tropical japonica (Supplementary Table S1; Figure 1). Lineage 3 had a pandemic distribution and included all sub-Saharan Africa samples, while lineage 4 was found in the Indian subcontinent, Zhejiang (China) and USA. Lineages 5 and 6 were collected on rainfed upland rice in Yunnan and Hunan (China), respectively on indica and tropical japonica varieties.

Lineage 2, 3 and 4, had low rates of female fertility (20%, 0%, 0%, respectively), and showed significant imbalance in the mating type ratios (frequency of Mat-1:100%, 14.3% and 100%, respectively; Chi^2 test, $P<0.001$), while lineage 1 had 88.9% female fertility and non-significant imbalance in its mating type ratio (frequency of Mat-1:33.3%; Chi^2 test, $P=0.083$). Lineage 5 was Mat-1 and only one of the two strains was female fertile (no data for lineage 6).

Pathogen compatibility range. Forty-six isolates representing lineages 1 to 4 - including 31 isolates published by Gallet et al. [39] - were inoculated onto 38 tropical japonica, temperate japonica, and indica and 19 differential varieties with known resistance genes to test for differences in virulence spectra (i.e. qualitative pathogenicity) between lineages (Supplementary table S3; Supplementary table S4).

The success of infection (binary response) was analyzed using a generalized linear model. Analysis of the proportion of compatible interactions revealed significant effects of rice subspecies, pathogen lineage and the interaction between them (Supplementary table S5). The lineage effect could be explained by lineage 2 having a lower infection frequency than lineage 1 (contrast: lineage 1 vs lineage 2; z-value=-2.779; p-value=0.005) and lineage 3 having a higher infection frequency than lineage 1 (contrast: lineage 3 vs lineage 1; z-value=2.683; p-value=0.007), while the infection frequency of lineage 4 was not different from lineage 1 (contrast: lineage 4 vs lineage 1; z-value=1.121; p-value=0.262). The rice subspecies effect could be attributed to tropical japonica varieties having a wider compatibility range than indica varieties (contrast: tropical japonica vs indica; z-value=1.793; p-value=0.073), and temperate japonica having a wider compatibility range than indica varieties (contrast: temperate japonica vs indica; z-value=1.830; p-

value=0.067). The significant interaction between rice subspecies and pathogen lineages indicates that the effect of the lineage of origin of isolates on the proportion of compatible interactions was different for the three rice subspecies. This interaction effect can be attributed to pathogen specialization to indica and tropical japonica, with lineage 1 (mostly originating from tropical japonica and/or tropical japonica-growing areas) infecting more frequently tropical japonica varieties than indica varieties, lineage 2 (the lineage sampled on temperate japonica) infecting more frequently temperate japonica varieties than other varieties, lineages 3 and 4 (mostly originating from indica varieties) infecting more frequently indica varieties than tropical japonica varieties, and all four lineages infecting temperate japonica varieties at relatively high frequencies (Supplementary table S6; Figure 2A).

Major resistance (R) genes can be a major determinant of pathogen host range, and promote divergence among pathogen lineages by exerting strong divergent selection on a limited number of pathogenicity-related genes [40-42]. To investigate the possible role of major resistance (R) genes in the observed differences in compatibility between rice subspecies and pathogen lineages, we challenged 19 differential varieties with the 46 isolates assigned to lineages 1 to 4. Analysis of the number of R genes overcome revealed a significant effect of pathogen lineage (Supplementary table S7). This effect was mostly driven by lineage 2, which could overcome fewer R genes than other lineages (Supplementary table S8; Figure 2B).

Recombination within and between lineages. To visualize evolutionary relationships while taking into account the possibility of recombination within or between lineages, we used the phylogenetic network approach neighbor-net as implemented in SPLITSTREE 4.13 [43]. This method allows visualizing conflicting phylogenetic signals, indicated by the presence of reticulations in the network. The role of recombination was evident from the non tree-like structure of the inner connections between lineages, hinting at genetic exchanges among unrelated isolates (Figure 3). There was greater reticulation of the network between lineages 1, 5 or 6 and the other lineages than among these other lineages themselves, the presumed consequence of higher levels of genetic exchanges involving lineages 1, 5 and 6. Lineages 2 to 4 displayed long interior branches and star-like topologies consistent with long-term clonality. Terminal branches appeared to be

shorter for lineages 2 to 4 than for lineage 1, consistent with the lower levels of variability detected in these lineages.

To further study the amount of recombination within lineages, we estimated the population recombination parameter (rho=$2N_e r$) and tested for the presence of recombination using a likelihood permutation test implemented in the PAIRWISE program in LDHAT. Recombination analyses confirmed the heterogeneity across lineages in the contribution of recombination to genomic variation, with recombination rates averaged across chromosomes being more than two to three orders of magnitude higher in lineage 1 (10.57 crossovers/Mb/generation) than in other lineages (lineage 2: 0.28; lineage 3: 0.01; lineage 4: 0.33 crossovers/Mb/generation) (Table 3). Analyses using SPLITSTREE to visualize reticulations within each lineage and to test for recombination using the PHI test were consistent with this pattern (Table 3; Figure 3; Supplementary Figure S2). The null hypothesis of no recombination was rejected only for lineage 1 and 4 [44]. Differences in recombination-based variation among lineages were confirmed by analyses of homoplasy (Supplementary text S1).

To gain insight into the genomic impact of recombination, we analyzed patterns of linkage disequilibrium (LD), which represents the tendency for different alleles to co-occur non-randomly. For lineage 1, LD decayed smoothly with physical distance, reaching half of its maximum value at about 10 kb, while for lineages 2, 3 and 4, no pattern of LD decay was observed (Supplementary Figure S3). These analyses also revealed that background LD levels were not higher in lineage 2, 3, or 4, which appear to be largely clonal, than in lineage 1. However, both simulation work and empirical data have shown that population history, including bottlenecks or admixture, have a strong impact on population background LD levels [45].

Genome scan for genetic exchanges among lineages.

We scanned the genomes for mutations exchanged among lineages using a method based on lineage-diagnostic SNPs and a probabilistic method of ‘chromosome painting’ (Figure 4). In the lineage-diagnostic SNPs approach, each isolate is removed from the dataset, in turn, lineage-diagnostic SNPs are identified (i.e. biallelic sites at which a mutation is specific to a given lineage),

each focal isolate is then added back to the dataset and scanned for the presence of lineage-diagnostic SNPs identified in lineages that are not its lineage of origin. Using this approach, we identified 515 lineage-diagnostic singletons with 276, 96 and 140 singletons in lineage 1, 5 and 6, respectively, and only a single singleton in lineages 2, 3 and 4. Putatively migrant singletons were assigned to all other lineages for lineage 1 and 5, and to all other lineages except lineage 1 for lineage 6 (Supplementary Table S9). 'Chromosome painting' is a probabilistic method of reconstructing the chromosomes of each individual sample as a combination of all other homologous sequences. For each isolate, we identified migrant mutations as mutations having a probability above 90% of having been copied from a lineage that is not the lineage of origin of the focal isolate. The method only takes population data as recipient populations, and thus only lineages 1 to 4 could be included in the analysis. Using chromosome painting we identified 464 migrant mutations, all of which segregated in lineage 1. Putatively migrant mutations were assigned to all five other lineages (92.8 mutations per lineage on average), with greatest contribution of lineage 2 (165 mutations) and smallest contribution of lineage 4 (39 mutations).

The sets of putative migrant mutations identified using the two methods matched different sets of genes that showed enrichment in NOD-like Receptors [46], HET-domain genes [47] or the GO term 'lipid catabolic process' (Supplementary Table S10; Supplementary Table S11). However, the presence of false positive caused by the random sorting of ancestral polymorphism in lineage 1 and other lineages cannot be excluded. To reduce the impact of the retention of ancestral mutations, we reasoned that series of adjacent mutations are more likely to represent genuine genetic exchange events. We identified all genomic regions defined by three adjacent putative migrant mutations stemming from the same donor lineage. We searched for such mutations in the set of putative migrant mutations identified by the two methods. In total we identified 12 such regions, matching 1917 genes. Functional enrichment tests conducted for each recipient isolates revealed enrichment in GO term 'pathogenesis' for isolate CH999, GO term 'phosphatidylinositol biosynthetic process' for isolate TH17 and GO term 'telomere maintenance' for isolate CH1019 (Supplementary table S12).

Molecular dating. To investigate the timing of rice blast emergence and diversification, we performed Bayesian phylogenetic analyses using BEAST. Isolates were collected from 1967 to 2009 (Supplementary table S1), which allowed us to use a tip-based calibration approach to coestimate evolutionary rates and ancestral divergence times. Fitting a linear regression between the age of samples and their root-to-tip distance (i.e. the number of substitution separating each sample from the hypothetical ancestor at the root of the tree) revealed enough temporal signal to perform thorough tip-dating inferences (Supplementary Figure S4) [48]. We therefore used tip-dating to estimate jointly the rate at which mutations accumulate (i.e. the substitution rate) and the age of every node in the tree including the root (i.e. time to the most recent common ancestor). At the genome scale, the average substitution rate estimate was 1.98e-8 substitution/site/year (Supplementary Figure S5). We inferred that the ancestor of rice-infecting and Setaria-infecting lineages lived ~14,000 years before present (ybp) but credibility intervals were relatively large (95% HPD [6500-30,000] ybp). The six rice-infecting lineages were estimated to have diversified ~900 to ~1300 ybp (95% HPD [175-2700] ybp) (Figure 5). Bootstrap node support were high and similar node age estimates were obtained when excluding recombining lineage 1 and potentially recombining lineages 5 and 6 (not shown), indicating that recombination has limited effect on our inferences.

Discussion

We have performed a whole-genome sequence analysis of time- and space-distributed samples to elucidate the emergence, diversification and spread of *M. oryzae* as a rapidly evolving pathogen with devastating impact on rice agrosystems. Analyses of population subdivision and tip-dating analyses of a collection of samples covering over 40 years revealed that the six lineages identified within *M. oryzae* diversified ~900 to ~1300 ybp (95% credible intervals [175-2700] ybp). Lineage 1 (which includes the reference strains GY11 and 70-15) was found in mainland Southeast Asia and originated from barley, tropical japonica or from undetermined varieties. All isolates from lineage 1, 5 and 6 were collected in rainfed upland agrosystems that are typical for japonica rice, and pathogenicity tests were consistent with local adaptation of lineage 1 to tropical japonica rice. Lineage 2 was pandemic in irrigated fields of temperate japonica outside Asia, and cross inoculation experiments revealed specialization to this host as well as an ability to overcome fewer R genes, on average, than other lineages. Lineages 3 and 4 were associated with indica. Lineage 3 is pandemic, and cross inoculation indicated local adaptation to this host, relative to tropical japonica, although in general terms lineages 3 and 4 had relatively wide compatibility ranges consistent with generalism. A possible explanation to both the wide compatibility range of temperate japonica varieties and the narrow compatibility range of lineage 2 is that temperate japonica varieties display smaller repertoires of R genes, as resistance to blast is less of a concern to breeders in temperate irrigated conditions that are less conducive to epidemics [39].

The Southeast continental Asian lineage was the most basal in total evidence genome genealogies, mirroring an evolutionary pathway to domesticated Asian rice [15, 17] in which the de novo domestication of rice occurred only once, in japonica. However, the diversification of *M. oryzae* into multiple rice-infecting lineages (point estimates ranging from ~900 to ~1300 ybp) appears to be much more recent than the de novo domestication of rice (8500-6500 ybp [15, 49, 50]), the spread of rice cultivation in paddy fields, and the domestication of indica in South Asia, following introgressive hybridization from the early japonica gene pool into ‘proto-indica’ rice (ca. 4000 ybp [15, 51]). At the time corresponding to the upper bound of 95% credible intervals (2700 ybp), japonica rice and paddy-field cultivation had spread to most areas of continental and insular

South, East and Southeast Asia [15, 52], and indica rice was beginning to spread out of the Ganges plains [15]. The point estimates for the splitting of *M. oryzae* lineages in fact corresponds with the Tang dynasty ('the Golden Age') in China, and the late classical period in India, during which food production became more rational and scientific and intensive irrigated systems of cultivation were developed, bringing about economic, demographic and material growth [53].

Genome scans based on polymorphism and divergence revealed heterogeneity in the genomic and life-history changes associated with the emergence and spread of the different lineages. The Southeast continental Asia lineage 1 is recombining, fertile, genetically diverse and has both mating types, suggesting sexual reproduction and relatively high long-term population size. The pandemic lineages 2 and 3, as well as the South Asia/US lineage 4, were largely clonal, genetically depauperate, sterile, and single mating type, suggesting a lack of sexual reproduction and demographic bottlenecks associated with their emergence in agrosystems. Patterns of polymorphism and diversity at non-synonymous and synonymous sites indicated that deleterious mutations were enriched in clonal lineages of *M. oryzae* with smaller long-term population size, consistent with a higher cost of pestification in these lineages. Introgression of genetic elements from clonal lineages harboring greater loads of deleterious mutations may be counteracting efficient purging of deleterious mutations in the recombining lineage 1 from mainland Southeast Asia, and contribute to reducing differences in the proportion of nonsynonymous mutations between recombining and clonal lineages. However, the extensive variability detected in the origin and genomic distribution of the detected putative migrant mutations suggested that most of them were false positives and that only the series of adjacent such mutations originating from the same donor lineage represented genuine genetic exchange events. Field-scale studies in areas where different lineages coexist should provide more detailed insights into the relative importance of interlineage recombination and allow testing whether genetic exchanges are driven by positive selection or whether they represent an incidental byproduct of the sympatric coexistence of interfertile lineages. We predict that the accumulation of deleterious mutations in pandemic clonal complexes and gene flow into sexual lineages during disease emergence and spread are widespread phenomena, not due to idiosyncrasies of *M. oryzae*, and we expect patterns to hold true in other invasive fungal plant pathogens.

Examination of additional isolates from under-sampled geographical regions (including Africa and South-America) using sequencing approaches and sampling schemed tailored to detecting adaptation from de novo mutations will be required to enhance our understanding of the biogeography of *M. oryzae* and the genetic basis of adaptation in the different *M. oryzae* lineages. Nevertheless, the catalogue of variants detected in our study provides a solid foundation for future research on the population genomics of adaptation in *M. oryzae*. Our work also provides a population-level genomic framework for defining molecular markers to assist in the control of rice blast and for investigating the molecular underpinnings of phenotypic and fitness differences between divergent lineages.

Methods

Genome sequencing and SNP calling. Preparation of sequencing libraries and Illumina HiSeq 2500 sequencing was performed either at Beckman Coulters Genomics (Danvers, USA) or at Iwate Biotechnological Research Center (Supplementary table S1). Genomic DNA for sequencing at BCG was isolated from 100 mg of fresh mycelium grown in liquid medium. Mycelium was pre-treated by enzymatic activities that degrade the cell walls (mainly betaglucanase) followed by an incubation in a lysis buffer (Triton 2 X - SDS 1% - NaCl 100 mM - TrisHCl 10 mM - EDTA 1 mM). Nucleic acids were extracted using a chloroform:isoamyl alcohol 24:1 treatment, overnight isopropanol precipitation, and rinsing with 70% ethanol. The nucleic acid extract was treated with RNase A (0.2mg/mL final) to remove RNA and purified with another chloroform:isoamyl alcohol 24:1 treatment. Genomic DNA for sequencing at IBRC was isolated using a protocol adapted from the animal tissue (Mouse tail) protocol available in the Promega Wizard® Genomic DNA Purification Kit. Nucleic acids were extracted from 20 mg of fresh mycelium grown in liquid medium, grind into powder using liquid nitrogen and pre-chilled mortar and pestle. Centrifugation time specified in the mouse tail protocol was increased to 15 min and centrifugation was carried out at 4°C after the precipitation step of 3 hours at -20°C. Nucleic acids were resuspended in

water, treated with RNase A (0.2mg/mL final), purified with a chloroform:isoamyl alcohol 24:1 treatment, precipitated overnight using isopropanol with 0.1 volume of sodium acetate (3 M pH = 5), and rinsed using 70% ethanol rinsing.

Sequencing reads were either paired-end (read length 100 nucleotides, insert size ~ 500bp, DNAs sequenced at IBI) or single-end (read length 100 nucleotides, DNAs sequenced by BCG). Reads were trimmed to remove barcodes, adapters, and filtered to eliminate sequences containing ambiguous base calls. Reads were mapped against the 70-15 reference genome version 8 [9] using BWA [54] (sub-command alb, option -n 5; sub-command sampe option -a 500). Alignments were sorted with SAMTOOLS [55], and reads with mapping quality below 30 were removed. Duplicates were removed with PICARD (http://broadinstitute.github.io/picard/). We used REALIGNER-TARGETCREATOR, TARGETCREATOR and INDELREALIGNER within the GENOME ANALYSES TOOLKIT (GATK) [56] to define intervals to target for local realignment and to perform local realignment of reads around indels, respectively, and UNIFIED GENOTYPER to call SNPs. High-confidence SNPs were identified using GATK's VARIANTFILTRATION with the following parameters: MQ0< 3.0 (Total Mapping Quality Zero Reads), prof >= 15.0 (Number of reference alleles + number of alternative alleles, computed as the sum of allelic depths for the reference and alternative alleles in the order listed), and RA <= 0.1 (Number of reference alleles / number of alternative alleles). High-confidence SNPs were annotated using SNPEFF v4.3 [57].

Mating type and female fertility assays. Mating type and female fertility had previously been determined [24], or were determined as previously described [58].

Genealogical relationships and Population subdivision. Total evidence genealogy was inferred using RAxML based on pseudo-assembled genomic sequences (i.e. tables of SNPs converted into a fasta file using the reference sequence as a template), assuming a General Time-Reversible model of nucleotide substitution with the Γ model of rate heterogeneity. Bootstrap confidence levels were determined using 100 replicates. DAPC was performed using the ADEGENET package in R [59]. Sites with missing data were excluded. We retained the first 20 principal components, and the first 6 discriminant functions.

Diversity and divergence. Polymorphism and divergence statistics were computed using EGGLIB 3.0.0b10 [60] excluding sites with >30% missing data. The neutrality index was calculated as $(P_n/P_s)/(D_n/D_s)$, with P_n and P_s the numbers of non-synonymous and synonymous polymorphisms, D_n and D_s the numbers of non-synonymous and synonymous substitutions. D_n and D_s were computed in GESTIMATOR [61] using the Setaria-infecting lineage as an outgroup. P_n and P_s were computed using EGGLIB.

Linkage disequilibrium and Recombination. The coefficient of linkage disequilibrium (r2) [62] was computed using VCFTOOLS [63] with missing data and sites with minor allele frequency below 10% excluded. For all lineages we computed r2 between all pairs of SNPs less than 100kb apart and averaged LD values in distance classes of 1kb for lineages 1 and 4, and 10kb for lineages 2 and 3 to minimize noise due to their low genetic diversity. Only sites without missing data and at which the minor allele frequency was above 10% were included, to minimize the dependency of r2 to levels of minor allele frequency [64]. Recombination rates were estimated for each chromosome using the PAIRWISE program in LDHAT version 2.2 [65]. Singletons and sites with missing data were excluded.

Pathogenicity tests. We used pathotyping data for 31 isolates published by Gallet et al. [39]. We supplemented this dataset with pathotyping data for 17 isolates, produced by the same authors using the same protocol, but not included in their study due to uncertainty in the nature of their rice subspecies of origin. We used a combination of multilocus microsatellite and SNP data to assign the 58 pathotyped isolates to the six lineages. To improve accuracy of assignment tests, we also included SNP and microsatellite data for 19 non-pathotyped isolates, and the full dataset included 77 multilocus genotypes. Multilocus microsatellite genotyping at 12 loci was carried out as in Saleh et al. [22]. For the 49 isolates for which genomic data were available, we kept 1% of the SNP loci that had no missing data (i.e. 164 SNPs). Missing data were introduced at SNP and microsatellite loci for the 27 non-sequenced isolates and the 4 sequenced isolates without microsatellite data.

Assignments were carried out using the STRUCTURE 2.3.1 program [66-68]. The model implemented allowed admixture and correlation in allele frequencies. Burn-in length was set at 10 000 iterations followed by a run phase of 40 000 iterations. Four independent runs were performed to check for

convergence. At K=6 the four main clusters identified with the full genomic dataset were recovered, although 15 genotypes out of 77 could not be assigned because of admixture or a lack of power. Finally, out of the 58 inoculated isolates, 46 could be assigned to lineages 1 to 4; the remaining twelve isolates could not be divided between lineages 1, 5 or 6 and were not analyzed further (Supplementary Figure S1)The success of infection was analyzed using a generalized linear model with a binomial error structure and logit link function. Treatment contrasts were used to assess the specific degrees of freedom of main effects and interactions.

Genome scan for genetic exchanges. Probabilistic chromosome painting was carried out using the CHROMOPAINTER program version 0.0.4 [69]**.** This method ‘‘paints’’ individuals in ‘‘recipient’’ populations as a combination of segments from ‘‘donor’’ populations using linkage information in probability computation on the basis that linked alleles are more likely to be exchanged together during recombination events. All lineages were used as donors, while lineage 1-4 only were used as recipients (sample size too small for lineages 5 and 6). We initially ran the model using increments of 50 expectation-maximization iterations, starting at 10 iterations, and examined convergence of parameter estimates to determine how many iterations to use. Hence, the recombination scaling constant N_e and emission probabilities (μ) were estimated in lineages 1-4 by running the expectation-maximization algorithm with 200 iterations for each lineage and chromosome. Estimates of N_e and μ were then computed as averages weighted by chromosome length (N_e=8160 for all lineages, lineage 1; μ=0.0000506; lineage 2, μ=0.0000171; lineage 3, μ=0.000021; lineage 4, μ=0.000011). These parameter values and the per-chromosome recombination rates estimated using LDHAT were then used to paint the chromosome of each lineage, considering the remaining lineages as donors, using 200 expectation-maximization iterations. We used a probability threshold of 0.9 to assign mutations in a recipient lineage to a donor lineage.

Tip-calibrated phylogenetic analysis. Tip-calibrated phylogenetic inferences were performed using only the 48 isolates with recorded sampling dates, i.e. all isolates except the reference 70-15 and PH0018. We investigated whether our dataset contained sufficient signal to perform thorough tip-dating inferences by building a phylogenetic tree using PHYML [71] without constraining tip-

heights to isolate's sampling times and then fitting root-to-tip distances (a proxy for the number of accumulated substitutions since the most recent common ancestor, TMRCA) to collection dates using TEMPEST [70]. We observed a positive and significant correlation (Supplementary Figure S4), demonstrating sufficient temporal signal to perform thorough tip-dating inferences at this evolutionary scale.The tip-calibrated inferences were then ran using Markov chain Monte Carlo sampling in BEAST 1.8.2 [72]. The topology was fixed as the total evidence genome genealogy inferred using RAXML. We used an annotation of the SNPs with SNPEFF [57] to partition Bayesian inference (i.e. several substitution models and rates of evolution were fitted to the different sets of SNPs during a single analysis). The optimal partitioning scheme and the best-fit nucleotide substitution model for each partition of the genome were estimated using the software PARTITIONFINDER [73]. The best partitioning was obtained for K=3 schemes (synonymous: HKY, non-synonymous: GTR and non-exonic SNPs: GTR) and was used henceforth. Nodes age were then estimated using this optimal partitioning scheme. Rate variation among sites was modeled with a discrete gamma distribution with four rate categories. We assumed an uncorrelated lognormal relaxed clock to account for rate variation among lineages. To minimize prior assumptions about demographic history, we adopted an extended Bayesian skyline plot approach in order to integrate data over different coalescent histories. The tree was calibrated using tip-dates only. To do so, we applied flat priors (i.e., uniform distributions) for the substitution rate (1.10^{-12} - 1.10^{-2} substitutions/site/year) as well as for the age of any internal node than in the tree (including the root). We ran five independent chains in which samples were drawn every 5,000 MCMC steps from a total of 50,000,000 steps, after a discarded burn-in of 5,000,000 steps. Convergence to the stationary distribution and sufficient sampling and mixing were checked by inspection of posterior samples (effective sample size >200). Parameter estimation was based on the samples combined from the different chains. The best-supported tree was estimated from the combined samples using the maximum clade credibility method implemented in TREEANNOTATOR.

Functional enrichment. Gene enrichment analysis was conducted using R package TOPGO for GO terms, and fisher exact tests for enrichment in HET-domain genes, NLRs, small-secreted proteins and MAX-effector genes. MAX-effector genes were obtained from de Guillen et al. [70],

NLRs were identified in Dyrka et al. [46], and small-secreted proteins and HET-domain were identified using Ensembl's Biomart.

Acknowledgements.

We thank the scholars who contributed samples, François Bonnot and Romain Gallet for help with statistics, and the Southgreen and Migale computing facilities. Sequencing of 25 strains was supported by a grant of Bayer Crop Science to DT.

Contributions.

PG analyzed genomic data and wrote the manuscript, with input from all co-authors; AR conducted tip-dating inferences; SR processed and annotated genomic data; HA and SCA developed DNA extraction protocol and extracted DNA; DT, EF, HA and JM conducted inoculation experiments; PG and MT analyzed pathotyping data; EF, DT, RT designed research; HA, DT and RT provided samples; DT and RT obtained funding.

Competing financial interests.

None

Materials & Correspondence.

Correspondence and material requests should be addressed to PG (pierre.gladieux@inra.fr), DT (tharreau@cirad.fr) and RT (terauchi@ibrc.or.jp).

Bibliography

1. Taylor JW, Jacobson D, Fisher M. The Evolution of Asexual Fungi: Reproduction, Speciation and Classification. Annual Review of Phytopathology. 1999;37:197-246. PubMed PMID: 266.

2. Gladieux P, Feurtey A, Hood ME, Snirc A, Clavel J, Dutech C, et al. The population biology of fungal invasions. Molecular Ecology. 2015:n/a-n/a. doi: 10.1111/mec.13028.

3. Martin MD, Vieira FG, Ho SYW, Wales N, Schubert M, Seguin-Orlando A, et al. Genomic characterization of a South American Phytophthora hybrid mandates reassessment of the geographic origins of Phytophthora infestans. Molecular biology and evolution. 2016;33(2):478-91.

4. Simwami SP, Khayhan K, Henk DA, Aanensen DM, Boekhout T, Hagen F, et al. Low diversity Cryptococcus neoformans variety grubii multilocus sequence types from Thailand are consistent with an ancestral African origin. PLoS Pathog. 2011;7(4):e1001343.

5. Ali S, Gladieux P, Rahman H, Saqib MS, Fiaz M, Ahmad H, et al. Inferring the contribution of sexual reproduction, migration and off-season survival to the temporal maintenance of microbial populations: a case study on the wheat fungal pathogen Puccinia striiformis f.sp tritici. Molecular Ecology. 2014;23(3):603-17. doi: Doi 10.1111/Mec.12629. PubMed PMID: WOS:000329980000010.

6. Talbot NJ. On the trail of a cereal killer: exploring the biology of Magnaporthe grisea. Annual Reviews in Microbiology. 2003;57(1):177-202.

7. Gurr S, Samalova M, Fisher M. The rise and rise of emerging infectious fungi challenges food security and ecosystem health. Fungal Biology Reviews. 2011;25(4):181-8.

8. Valent B. Rice blast as a model system for plant pathology. Phytopathology. 1990;80(1):33-6.

9. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, et al. The genome sequence of the rice blast fungus Magnaporthe grisea. Nature. 2005;434(7036):980-6.

10. Ebbole DJ. Magnaporthe as a model for understanding host-pathogen interactions. Annu Rev Phytopathol. 2007;45:437-56.

11. Chiapello H, Mallet L, Guerin C, Aguileta G, Amselem J, Kroj T, et al. Deciphering Genome Content and Evolutionary Relationships of Isolates from the Fungus Magnaporthe oryzae Attacking Different Host Plants. Genome Biol Evol. 2015;7(10):2896-912. Epub 2015/10/11. doi: 10.1093/gbe/evv187. PubMed PMID: 26454013; PubMed Central PMCID: PMCPMC4684704.

12. 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.

13. Yoshida K, Saunders DGO, Mitsuoka C, Natsume S, Kosugi S, Saitoh H, et al. Host specialization of the blast fungus Magnaporthe oryzae is associated with dynamic gain and loss of genes linked to transposable elements. BMC genomics. 2016;17(1):1.

14. Couch BC, Fudal I, Lebrun MH, 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. Epub 2005/04/02. doi: 10.1534/genetics.105.041780. PubMed PMID: 15802503; PubMed Central PMCID: PMCPMC1450392.

15. Fuller DQ, Sato Y-I, Castillo C, Qin L, Weisskopf AR, Kingwell-Banham EJ, et al. Consilience of genetics and archaeobotany in the entangled history of rice. Archaeological and Anthropological Sciences. 2010;2(2):115-31.

16. Diao X, Jia G. Origin and Domestication of Foxtail Millet. Genetics and Genomics of Setaria: Springer; 2017. p. 61-72.

17. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490(7421):497-501.

18. Civáň P, Craig H, Cox CJ, Brown TA. Three geographically separate domestications of Asian rice. Nature plants. 2015;1:15164.

19. Huang X, Han B. Rice domestication occurred through single origin and multiple introgressions. Nature plants. 2015;2:15207.

20. Civán P, Craig H, Cox CJ, Brown TA. Multiple domestications of Asian rice. Nature plants. 2016;2:16037.

21. Choi JY, Platts AE, Fuller DQ, Wing RA, Purugganan MD. The rice paradox: Multiple origins but single domestication in Asian rice. Molecular biology and evolution. 2017;34(4):969-79.

22. Saleh D, Milazzo J, Adreit H, Fournier E, Tharreau D. South-East Asia is the center of origin, diversity and dispersion of the rice blast fungus, Magnaporthe oryzae. The New phytologist. 2014;201(4):1440-56. Epub 2013/12/11. doi: 10.1111/nph.12627. PubMed PMID: 24320224; PubMed Central PMCID: PMCPMC4265293.

23. Zeigler RS. Recombination in*Magnaporthe Grisea*. Annual Review of Phytopathology. 1998;36(1):249-75. PubMed PMID: 301.

24. Saleh D, Xu P, Shen Y, Li C, Adreit H, Milazzo J, et al. Sex at the origin: an Asian population of the rice blast fungus Magnaporthe oryzae reproduces sexually. Molecular Ecology. 2012;21(6):1330-44.

25. Lu J, Tang T, Tang H, Huang J, Shi S, Wu C-I. The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends in Genetics. 2006;22(3):126-31.

26. Glémin S, Bataillon T. A comparative view of the evolution of grasses under domestication. New phytologist. 2009;183(2):273-90.

27. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, et al. The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome research. 2011;21(12):2157-66.

28. Stukenbrock EH, Bataillon T. A population genomics perspective on the emergence and adaptation of new plant pathogens in agro-ecosystems. PLoS pathogens. 2012;8(9):e1002893.

29. Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737-56.

30. Karasov T, Messer PW, Petrov DA. Evidence that adaptation in Drosophila is not limited by mutation at single sites. PLoS genetics. 2010;6(6):e1000924.

31. Ellison CE, Hall C, Kowbel D, Welch J, Brem RB, Glass NL, et al. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. Proceedings of the National Academy of Sciences. 2011;108(7):2831-6.

32. Roper M, Ellison C, Taylor JW, Glass NL. Nuclear and genome dynamics in multinucleate ascomycete fungi. Current biology. 2011;21(18):R786-R93.

33. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, et al. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nature communications. 2014;5.

34. Gladieux P, Ropars J, Badouin H, Branca A, Aguileta G, De Vienne DM, et al. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Molecular Ecology. 2014;23(4):753-73. doi: Doi 10.1111/Mec.12631. PubMed PMID: WOS:000330264000003.

35. Noguchi MT, Yasuda N, Fujita Y. Evidence of genetic exchange by parasexual recombination and genetic analysis of pathogenicity and mating type of parasexual recombinants in rice blast fungus, Magnaporthe oryzae. Phytopathology. 2006;96(7):746-50.

36. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30(9):1312-3.

37. Leaché AD, Banbury BL, Felsenstein J, Stamatakis A. Short Tree, Long Tree, Right Tree, Wrong Tree: New Acquisition Bias Corrections for Inferring SNP Phylogenies. Systematic Biology. 2015;64(6):1032-47.

38. Akashi H, Osada N, Ohta T. Weak selection and protein evolution. Genetics. 2012;192(1):15-31.

39. Gallet R, Fontaine C, Bonnot F, Milazzo J, Tertois C, Adreit H, et al. Evolution of Compatibility Range in the Rice-Magnaporthe oryzae System: An Uneven Distribution of R Genes Between Rice Subspecies. Phytopathology. 2016;106(4):348-54. Epub 2015/12/17. doi: 10.1094/phyto-07-15-0169-r. PubMed PMID: 26667186.

40. Giraud T, Gladieux P, Gavrilets S. Linking emergence of fungal plant diseases and ecological speciation. Trends in Ecology and Evolution. 2010;In press.

41. Schulze-Lefert P, Panstruga R. A molecular evolutionary concept connecting nonhost resistance, pathogen host range, and pathogen speciation. Trends Plant Sci. 2011;16(3):117-25. Epub 2011/02/15. doi: 10.1016/j.tplants.2011.01.001. PubMed PMID: 21317020.

42. Liao J, Huang H, Meusnier I, Adreit H, Ducasse A, Bonnot F, et al. Pathogen effectors and plant immunity determine specialization of the blast fungus to rice subspecies. eLife. 2016;5:e19377.

43. Bryant D, Moulton V. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. Molecular Biology and Evolution. 2004;21(2):255-65. doi: 10.1093/molbev/msh018. PubMed PMID: WOS:000220083300008.

44. Bruen TC, Philippe H, Bryant D. A simple and robust statistical test for detecting the presence of recombination. Genetics. 2006;172(4):2665-81.

45. Ardlie KG, Kruglyak L, Seielstad M. Patterns of linkage disequilibrium in the human genome. Nature Reviews Genetics. 2002;3(4):299-309.

46. Dyrka W, Lamacchia M, Durrens P, Kobe B, Daskalov A, Paoletti M, et al. Diversity and variability of NOD-like receptors in fungi. Genome Biology and Evolution. 2014:3137-58.

47. Saupe SJ. Molecular genetics of heterokaryon incompatibility in filamentous ascomycetes. Microbiology and molecular biology reviews. 2000;64(3):489-502.

48. Rieux A, Balloux F. Inferences from tip□calibrated phylogenies: a review and a practical guide. Molecular ecology. 2016;25(9):1911-24.

49. Zhao Z. New archaeobotanic data for the study of the origins of agriculture in China. Current Anthropology. 2011;52(S4):S295-S306.

50. Castillo CC, Bellina B, Fuller DQ. Rice, beans and trade crops on the early maritime Silk Route in Southeast Asia. Antiquity. 2016;90(353):1255-69.

51. Fujiwara H. Search for the Origin of Rice Cultivation: The Ancient Rice Cultivation in Paddy Fields at the Cao Xie Shan Site in China. Society for Scientific Studies on Cultural Property, Miyazaki. 1996.

52. Fuller DQ, Qin L. Water management and labour in the origins and dispersal of Asian rice. World Archaeology. 2009;41(1):88-111.

53. Siddiqui IH. Water works and irrigation system in India during pre-Mughal times. Roots and Routes of Development in China and India: Brill; 2008. p. 429-54.

54. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

55. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

56. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome research. 2010;20(9):1297-303.

57. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80-92.

58. Notteghem JL, Silue D. Distribution of the mating type alleles in Magnaporthe grisea populations pathogenic on rice. Phytopathology. 1992;82(4):421-4.

59. Jombart T, Ahmed I. adegenet 1.3-1: new tools for the analysis of genome-wide SNP data. Bioinformatics. 2011;27(21):3070-1.

60. De Mita S, Siol M. EggLib: processing, analysis and simulation tools for population genetics and genomics. BMC genetics. 2012;13(1):1.

61. Thornton K. libsequence: a C++ class library for evolutionary genetic analysis. Bioinformatics. 2003;19(17):2325-7. doi: 10.1093/bioinformatics/btg316. PubMed PMID: ISI:000186919200024.

62. Hill WG, Robertson A. Linkage disequilibrium in finite populations. Theoretical and Applied Genetics. 1968;38(6):226-31.

63. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.

64. Amaral AJ, Megens H-J, Crooijmans RPMA, Heuven HCM, Groenen MAM. Linkage disequilibrium decay and haplotype block structure in the pig. Genetics. 2008;179(1):569-79.

65. Auton A, McVean G. Recombination rate estimation in the presence of hotspots. Genome research. 2007;17(8):1219-27.

66. Pritchard JK, Stephens M, Donnelly P. Inference of Population Structure Using Multilocus Genotype Data. Genetics. 2000;155(2):945-59. PubMed PMID: 226.

67. Falush D, Stephens M, Pritchard JK. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics. 2003;164(4):1567-87. PubMed PMID: 88.

68. Hubisz MJ, Falush D, Stephens M, Pritchard JK. Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources. 2009;9(5):1322-32. doi: 10.1111/j.1755-0998.2009.02591.x. PubMed PMID: ISI:000268855000004.

69. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8(1):e1002453.

70. 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.

71. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC evolutionary biology. 2007;7(1):214.

72. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic biology. 2010;59(3):307-21.

73. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular biology and evolution. 2012;29(6):1695-701.

74. Rambaut A, Lam TT, Carvalho LM, Pybus OG. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evolution. 2016;2(1):vew007.

Tables.

Table 1. Summary of population genomic variation in non-overlapping 100kb window

Lineage	n	S	K	H_e	θ_w	π	D
1	10	57.6	3.6	0.31	2.2456E-04	2.1087E-04	-0.558
2	14	19.7	7.2	0.20	6.9427E-05	4.9195E-05	-1.454
3	16	21.5	7.9	0.17	7.2400E-05	4.5304E-05	-1.718
4	6	10.5	4.3	0.38	5.1492E-05	4.5220E-05	-0.824

Lineages 5 and 6 were not included in calculations because of too small a sample size (n=2 and n=1, respectively); n is sample size; θ_w is Watterson's θ per bp; π is nucleotide diversity per bp; H_e is haplotype diversity; K is the number of haplotypes; D is Tajima's neutrality statistic.

Table 2. McDonald-Kreitman tests based on genome-wide patterns of synonymous and non-synonymous variation, and measures of the genome-wide intensity of purifying selection. Divergence was measured against predicted gene sequences of *Magnaporthe oryzae* setaria-infecting isolates US71.

Lineage	π_N/π_S	$P_{nonsense}/P_{sense}$	P_n/P_s	D_n/D_s	NI
1	0.43 (0.00041/0.00018)	0.011 (49/4244)	1.23 (4293/3492)	0.70 (16444/23656)	1.77*
2	0.49 (0.00012/0.00006)	0.022 (36/1622)	1.52 (1658/1088)	0.72 (15565/21745)	2.13*
3	0.49 (0.00015/0.00007)	0.018 (32/1814)	1.17 (1846/1578)	0.97 (14789/15293)	1.21*
4	0.61 (0.00012/0.00007)	0.034 (31/914)	1.59 (945/593)	0.72 (15302/21347)	2.22*

π_N/π_S is the ratio of non-synonymous to synonymous nucleotide diversity. Under near neutrality, π_N/π_S gives an estimate of the proportion of effectively neutral mutations that are strongly dependent on the effective population size N_e. π_S is a proxy for N_e. $P_{nonsense}/P_{sense}$ is the number of nonsynonymous nonsense mutations (i.e., 'premature' stop codon) divided by the number of sense nonsynonymous mutations. *P-value<0.0001, chi-square independence tests. Neutrality index NI=(Pn/Ps)/(Dn/Ds), quantifies the direction and degree of departure from neutrality. NI=1 if nonsynonymous mutations are either neutral or strongly deleterious. NI < 1 indicates an amino acid substitutions and implies the fixation of advantageous mutations. NI > 1 indicates an excess of amino acid polymorphism, as expected when there are slightly deleterious mutations.

Table 3. Estimates of the population recombination rate ρ, tests of recombination based on homoplasy and linkage disequilibrium, proportion of homoplastic SNPs.

Lineage	ρ (crossovers/Mb/generation)								Homoplastic SNPs	PHI test (p-value)
	Chromosomes									
	1	2	3	4	5	6	7	Average		
1	8.6*	3.8*	15.1*	1.4	8.5*	10.6*	13.5*	10.57	34.56 %	0.0000
2	0.0	0.2	0.2*	0.3	0.6	0.5	0.3	0.28	0.09 %	0.0944
3	0.4*	0.2*	0.0	0.0	0.0	0.0	0.0	0.01	0.47 %	0.0535
4	0.2	0.2*	0.3	0.4	0.4	0.4	0.4	0.33	0.40 %	0.0014

*P<0.05. PHI test is the Pairwise Homoplasy Test. The null hypothesis of no recombination was tested, for the PHI test and for ρ, using random permutations of the positions of the SNPs based on the expectation that sites are exchangeable if there is no recombination. For the ρ test, significance was computed using the distribution of maximum composite likelihood values calculated on permuted data.

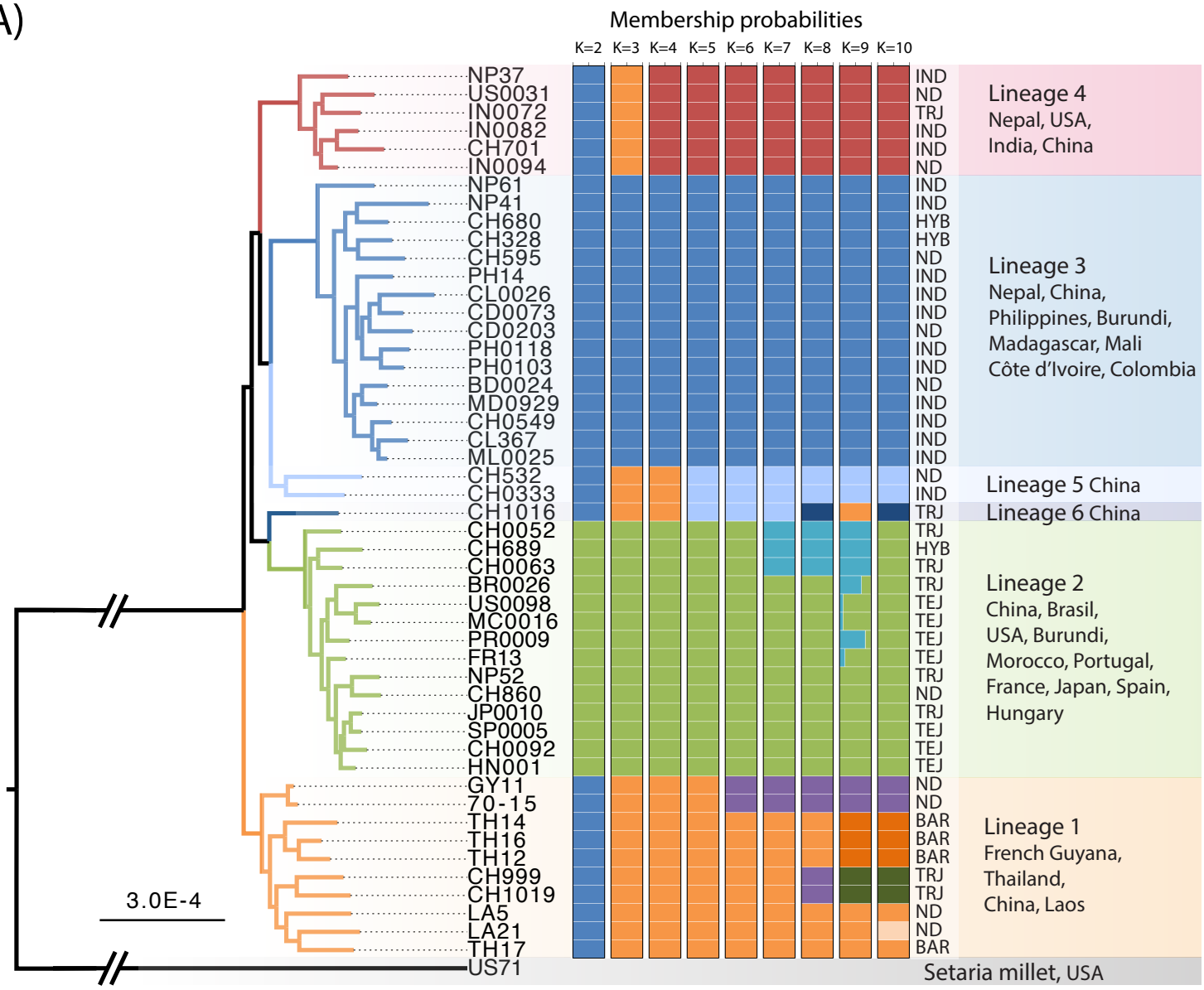
Figures.

Figure 1. Population subdivision in the sample set analyzed. (A) Total evidence maximum likelihood genome genealogy and Discriminant Analysis of Principal Components, (B) geographic distribution of the six lineages identified based on results presented in (A). In panel A, all nodes had bootstrap supports above 95% (100 resamplings), except the node carrying isolates BR0026, US0098, PR0009 and MC0016 (support: 72%); in the barplot, each isolate is represented by a thick vertical line divided into K segments that represent the isolate's estimated membership probabilities in the K clusters assumed. In panel B, diameters are proportional to the number of isolates collected per site (smallest diameter represents 1 isolate). TRJ, tropical japonica; TEJ, temperate japonica; IND, indica; HYB, hybrid; BAR, barley; ND, no data.

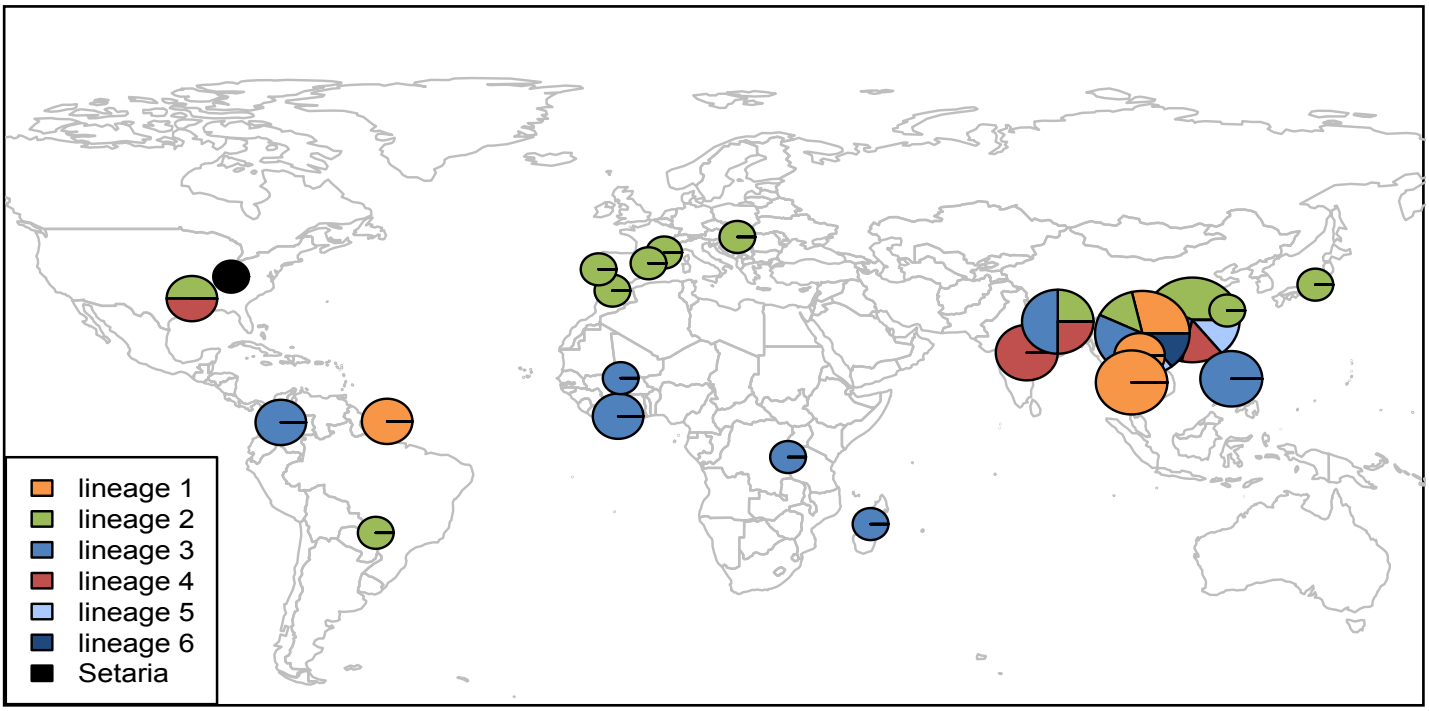
Figure 2. Neighbor-Net networks showing relationships between haplotypes identified based on the full set of 16,370 SNPs without missing data, (A) in the whole sample set, (B) in lineage 1, (C) in lineage 2, (D) in lineage 3, (E) in lineage 4.

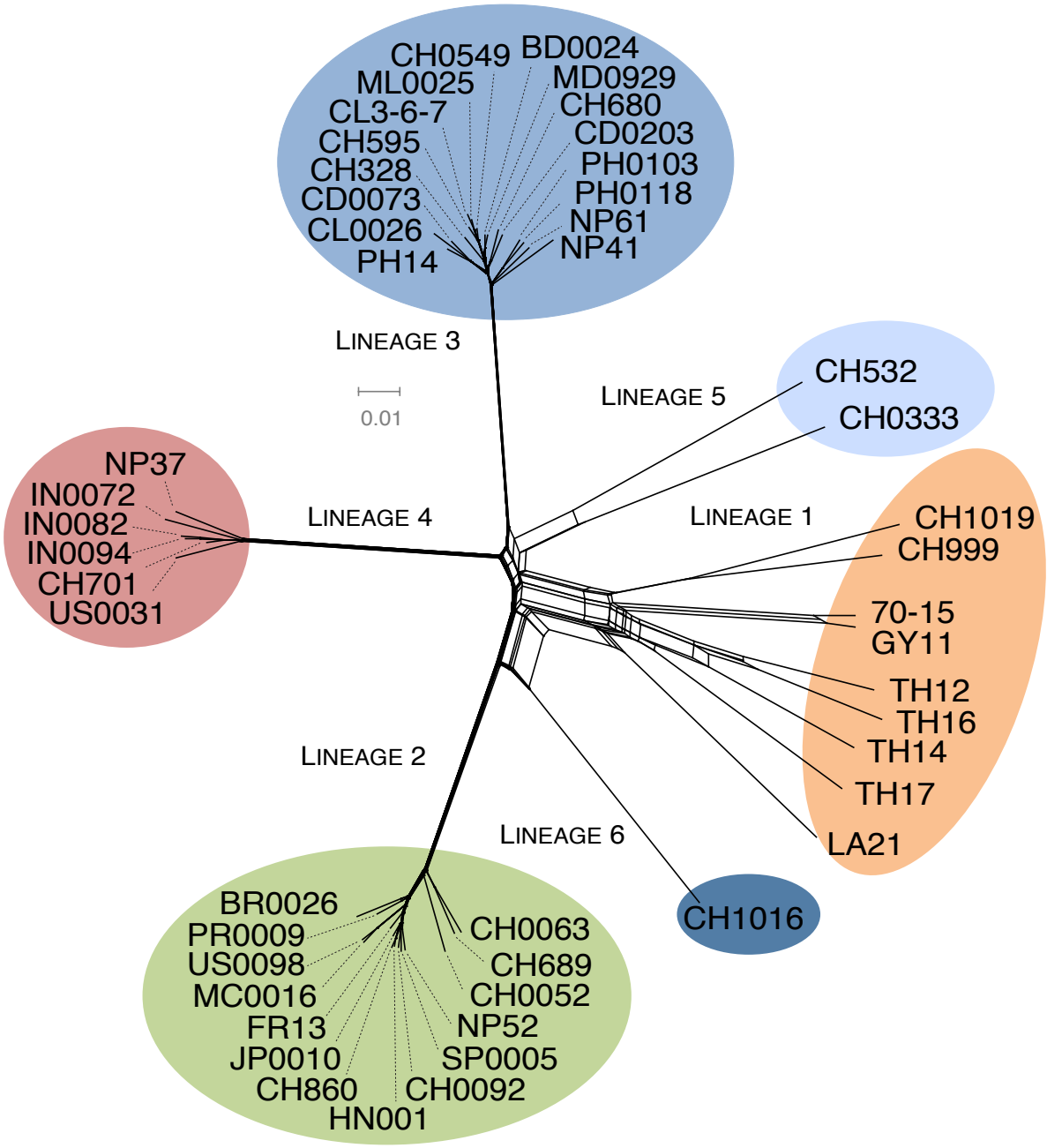
Figure 3. Genomic distribution of candidate immigrant mutations in lineages 1, 5 and 6. (Left) Lineage-diagnostic mutations segregating as singletons in other lineages. (center) Lineage 1 mutations for which the most probable donor is lineage 2, 3, 4 in probabilistic Chromosome painting analysis. Lineages 5 and 6 (n=2 and n=1, respectively) could not be included as recipient populations in Chromosome painting analysis of small sample size. No candidate immigrant mutations were identified in lineages 2, 3, and 4. (right) Genomic regions corresponding to series of adjacent putative migrant mutations identified using lineage-diagnostic singletons in lineage 1. Chromosomes 1 to 7 in clockwise order, with ticks every megabase.

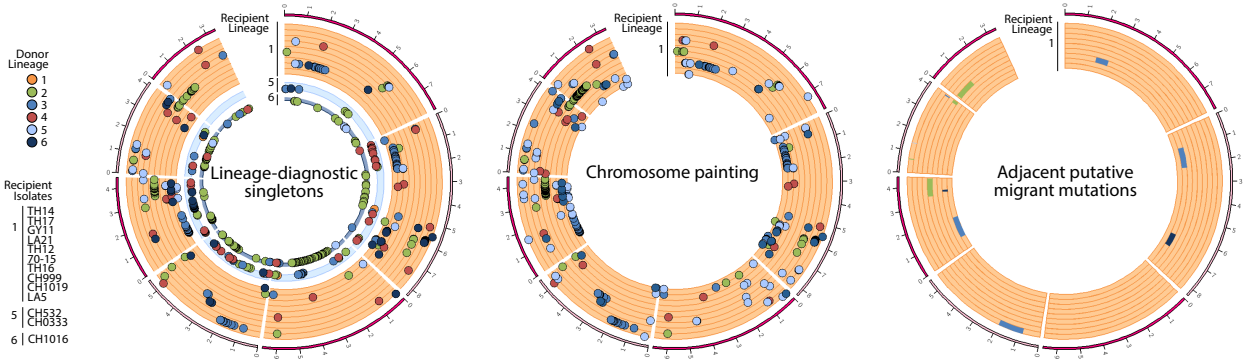
(A)


Membership probabilities
K=2 K=3 K=4 K=5 K=6 K=7 K=8 K=9 K=10
NP37 IND
US0031 ND
IN0072 TRJ
IN0082 IND
CH701 IND
IN0094 ND
NP61 IND
NP41 IND
CH680 HYB
CH328 HYB
CH595 ND
PH14 IND
CL0026 IND
CD0073 IND
CD0203 ND
PH0118 IND
PH0103 IND
BD0024 ND
MD0929 IND
CH0549 IND
CL367 IND
ML0025 IND
CH532 ND
CH0333 IND
CH1016 TRJ
CH0052 TRJ
CH689 HYB
CH0063 TRJ
BR0026 TRJ
US0098 TEJ
MC0016 TEJ
PR0009 TEJ
FR13 TEJ
NP52 TRJ
CH860 ND
JP0010 TRJ
SP0005 TEJ
CH0092 TEJ
HN001 TEJ
GY11 ND
70-15 ND
TH14 BAR
TH16 BAR
TH12 BAR
CH999 TRJ
CH1019 TRJ
LA5 ND
LA21 ND
TH17 BAR
US71
Setaria millet, USA
Lineage 4
Nepal, USA,
India, China
Lineage 3
Nepal, China,
Philippines, Burundi,
Madagascar, Mali
Côte d'Ivoire, Colombia
Lineage 5 China
Lineage 6 China
Lineage 2
China, Brasil,
USA, Burundi,
Morocco, Portugal,
France, Japan, Spain,
Hungary
Lineage 1
French Guyana,
Thailand,
China, Laos
3.0E-4

(B)


lineage 1
lineage 2
lineage 3
lineage 4
lineage 5
lineage 6
Setaria


CH0549
BD0024
ML0025
MD0929
CL3-6-7
CH680
CH595
CD0203
CH328
PH0103
CD0073
PH0118
CL0026
NP61
PH14
NP41
LINEAGE 3
0.01
LINEAGE 5
CH532
CH0333
NP37
IN0072
IN0082
IN0094
CH701
US0031
LINEAGE 4
LINEAGE 1
CH1019
CH999
70-15
GY11
TH12
TH16
TH14
TH17
LA21
LINEAGE 2
LINEAGE 6
CH1016
BR0026
PR0009
US0098
MC0016
FR13
JP0010
CH860
HN001
CH0063
CH689
CH0052
NP52
SP0005
CH0092


Donor Lineage
1
2
3
4
5
6
Recipient Isolates
1
TH14
TH17
GV11
LA21
TH12
70-15
TH16
CH999
CH1019
LA5
5
CH532
CH0333
6
CH1016
Recipient Lineage
Lineage-diagnostic singletons
Recipient Lineage
Chromosome painting
Recipient Lineage
Adjacent putative migrant mutations