A THOUSAND-GENOME PANEL RETRACES THE GLOBAL SPREAD AND CLIMATIC ADAPTATION OF A MAJOR CROP PATHOGEN

Alice Feurtey[1,2], Cécile Lorrain[2], Megan C. McDonald[3,4], Andrew Milgate[5], Peter Solomon[3], Rachael Warren[6], Guido Puccetti[1,7], Gabriel Scalliet[7], Stefano F. F. Torriani[7], Lilian Gout[8], Thierry C. Marcel[8], Frédéric Suffert[8], Julien Alassimone[2], Anna Lipzen[9], Yuko Yoshinaga[9], Christopher Daum[9], Kerrie Barry[9], Igor V. Grigoriev[9,10], Stephen B. Goodwin[11], Anne Genissel[8], Marc-Henri Lebrun[8], Gert H. J. Kema[12], Bruce A. McDonald[2], Daniel Croll[1,*]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland
[2] Plant Pathology, D-USYS, ETH Zurich, CH-8092 Zurich, Switzerland
[3] Division of Plant Science, Research School of Biology, The Australian National University, Canberra, ACT, Australia
[4] School of Biosciences, Institute of Microbiology and Infection, University of Birmingham, Birmingham, United Kingdom
[5] NSW Department of Primary Industries, Wagga Wagga Agricultural Institute, Pine Gully Road, Wagga Wagga, NSW 2650, Australia
[6] The New Zealand Institute for Plant and Food Research Limited, Lincoln, New Zealand
[7] Syngenta Crop Protection AG, CH-4332 Stein, Switzerland
[8] Université Paris Saclay, INRAE, UR BIOGER, 78850 Thiverval-Grignon, France
[9] US Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[10] Department of Plant and Microbial Biology, University of California Berkeley, Berkeley, CA 9472
[11] USDA-Agricultural Research Service, West Lafayette, IN, United States
[12] Wageningen University and Research, Laboratory of Phytopathology, Wageningen, The Netherlands

[*] Correspondence: daniel.croll@unine.ch

Keywords: crop pathogens, climate adaptation, domestication, genome evolution, population genomics, transposable elements

Running title: Genomic basis of global pathogen adaptation

Human activity impacts the evolutionary trajectories of many species worldwide. Global trade of agricultural goods contributes to the dispersal of pathogens reshaping their genetic makeup and providing opportunities for virulence gains. Understanding how pathogens surmount control strategies and cope with new climates is crucial to predicting the future impact of crop pathogens. Here, we address this by assembling a global thousand-genome panel of *Zymoseptoria tritici*, a major fungal pathogen of wheat reported in all production areas worldwide. We identify the global invasion routes and ongoing genetic exchange of the pathogen among wheat-growing regions. We find that the global expansion was accompanied by increased activity of transposable elements and weakened genomic defenses. Finally, we find significant standing variation for adaptation to new climates encountered during the global spread. Our work shows how large population genomic panels enable deep insights into the evolutionary trajectory of a major crop pathogen.

Main text

Human activity has broken down natural barriers to gene flow for a large number of species through trade and travel. Reshaped species distributions helped spread invasive plants and pathogens[1]. A major contributor to the range expansion of pathogens is the distribution of suitable host species. Pathogens and their hosts often share a common evolutionary history either through co-evolution or through shared constraints of their common environment[2,3]. Discrepancies in the evolutionary history of hosts and their pathogens can be caused by host jumps, significant differences in gene flow or local adaptation. Agricultural pathogens have often emerged during the domestication of their host species[4] causing significant threats to food production. Increased global trade of agricultural products has precipitated serious disease outbreaks over the past decades[5]. Crop pathogens are exposed to globally homogeneous host conditions created by planting genetically similar crop cultivars and application of similar pesticidal compounds to control diseases[6,7]. Furthermore, climate change reshapes the geographic distribution of pathogen species, with poleward range expansions being suspected since the 1960s[8]. Range expansions may lead to significant changes in the genetic make-up of pathogen species by founder effects and shifting barriers to gene flow[9]. Understanding how emerging pathogens surmount control strategies and cope with climate adaptation is crucial to predict the future impact of crop pathogens in a changing world.

Outbreaks of fungal diseases on crops are reported regularly across continents[5]. In addition to episodic damage, most crop pathogens are endemic and continuously reduce yields. The ascomycete *Zymoseptoria tritici* is a major pathogen of bread and durum wheat, causing the disease Septoria tritici blotch, which is now reported in most wheat-growing regions and causes significant damage[10]. The center of origin of *Z. tritici* is located in the Middle East, where sister species are found to infect wild grasses[11]. The emergence of *Z. tritici* was concomitant with the domestication of wheat[11]. The timing

and shared geographic origin of the pathogen and domesticated wheat strongly suggests coevolution between the two species. The pathogen harbors extensive standing variation from individual infected leaves to large agricultural regions[12,13]. As a consequence, the pathogen showed rapid responses across all major wheat-producing areas to overcome host resistance and gain tolerance to fungicides in less than a decade[10]. Population genomic analyses showed that rapid adaptation of the pathogen was facilitated by parallel evolution across geographic regions[14,15]. However, a comprehensive picture of pathogen dispersal and adaptation across the global distribution range is lacking.

Here, we assembled over one thousand genomes of the fungal crop pathogen *Z. tritici* to retrace worldwide invasion routes out of its Middle Eastern origin and identify ongoing genetic exchange among major wheat-producing regions. We show that the global expansion was accompanied by increased activity of transposable elements and weakened genomic defenses. Finally, we identify standing genetic variation for adaptation to new climates encountered during the global spread.

Global genetic structure of the pathogen tracks the historical spread of wheat

We assessed the evolutionary trajectory of the pathogen in conjunction with the history of global wheat cultivation (Fig. 1A). For this, we assembled a worldwide collection of *Z. tritici* isolates from naturally infected fields (Fig. 1B). We selected isolates covering most wheat production areas, both in the center of origin of the crop (*i.e.,* the Fertile Crescent in the Middle-East), and in areas where wheat was introduced during the last millennia (*i.e.,* Europe and North Africa), or last centuries (*i.e.*, the Americas and Oceania; Fig. 1C). We gathered 1109 high-quality, whole-genome, short-read sequencing datasets (Table S1) covering 42 countries and a broad range of climates. Using a joint genotyping approach, we produced raw variant calls for further inspection. To assess genotyping accuracy, we used eight isolates with replicate sequencing data to analyze discrepancies. We adjusted quality thresholds targeting specifically the type of genotyping errors observed in our data set (Fig. S1). The improved filtering yielded 8,406,818 high-confidence short variants (short indels and SNPs). The final variant set included 5,578,488 biallelic SNPs corresponding to 14.1% of the genome.

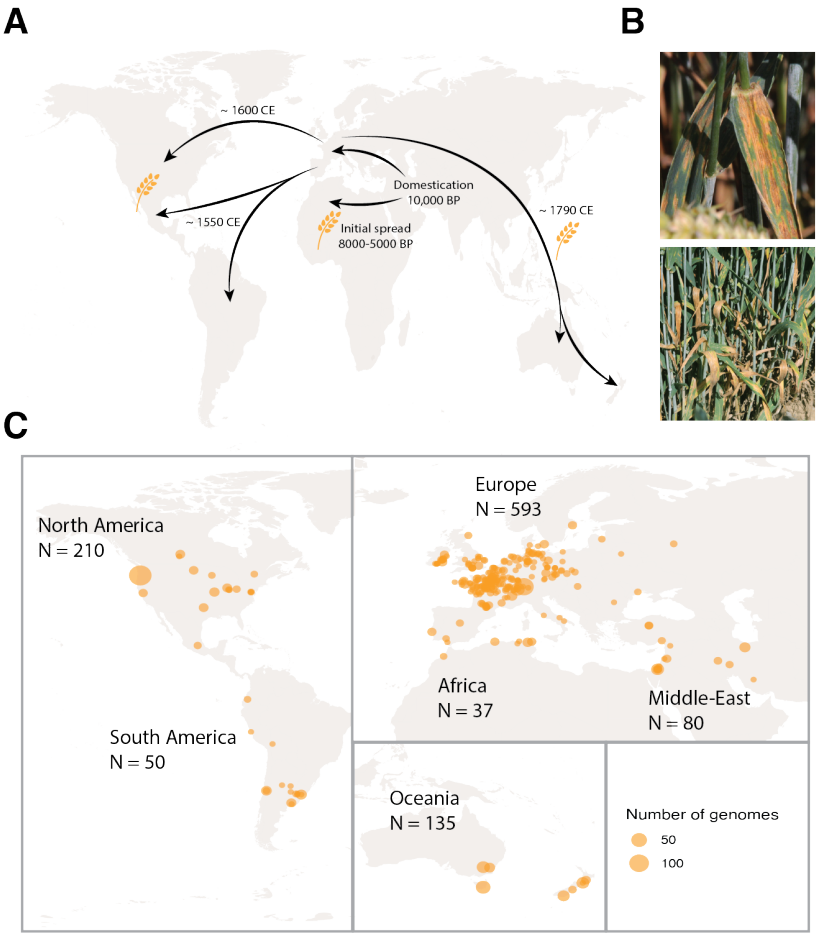

Figure 1: Global sampling of the wheat pathogen *Zymoseptoria tritici* retracing the historical spread of its host. A. Schematic representation of the introduction of wheat across continents. B. Septoria tritici blotch symptoms caused by *Z. tritici* on wheat leaves. C. Map of the sampling scheme for the global collection of 1109 isolates for whole-genome sequencing.

We tested whether global diversity patterns of pathogen populations are likely a consequence of the history of wheat cultivation. We first performed an unsupervised clustering of genotypes and identified eleven well-supported clusters (Fig. 2A, Fig. S2-3). Over 90% of the genotypes were clearly assigned to a single cluster (Fig. 2A, Table S3). Two clusters were identified among genotypes originating from the pathogen center of origin, distinguishing collections from Iran and Middle Eastern regions. Genotypes from Africa and Europe split into two distinct genetic clusters without any apparent secondary structure within clusters. This lack of any fine-scale structure is remarkable given the extensive geographic sampling of European genotypes and suggests extensive gene flow within the continent. Genotypes from Oceania grouped into three distinct clusters marked by collections from Tasmania, the Australian mainland and New Zealand. Genotypes from North America formed two clusters along a North-South separation. Finally, South American genotypes formed two clusters split along the Andes (Chile versus Argentina and Uruguay). A principal component analysis of all genotypes confirmed the nested genetic structure with differentiation at the continent level, subdivisions within some continents and the existence of admixed genotypes (Fig. 2B, Fig. S4).

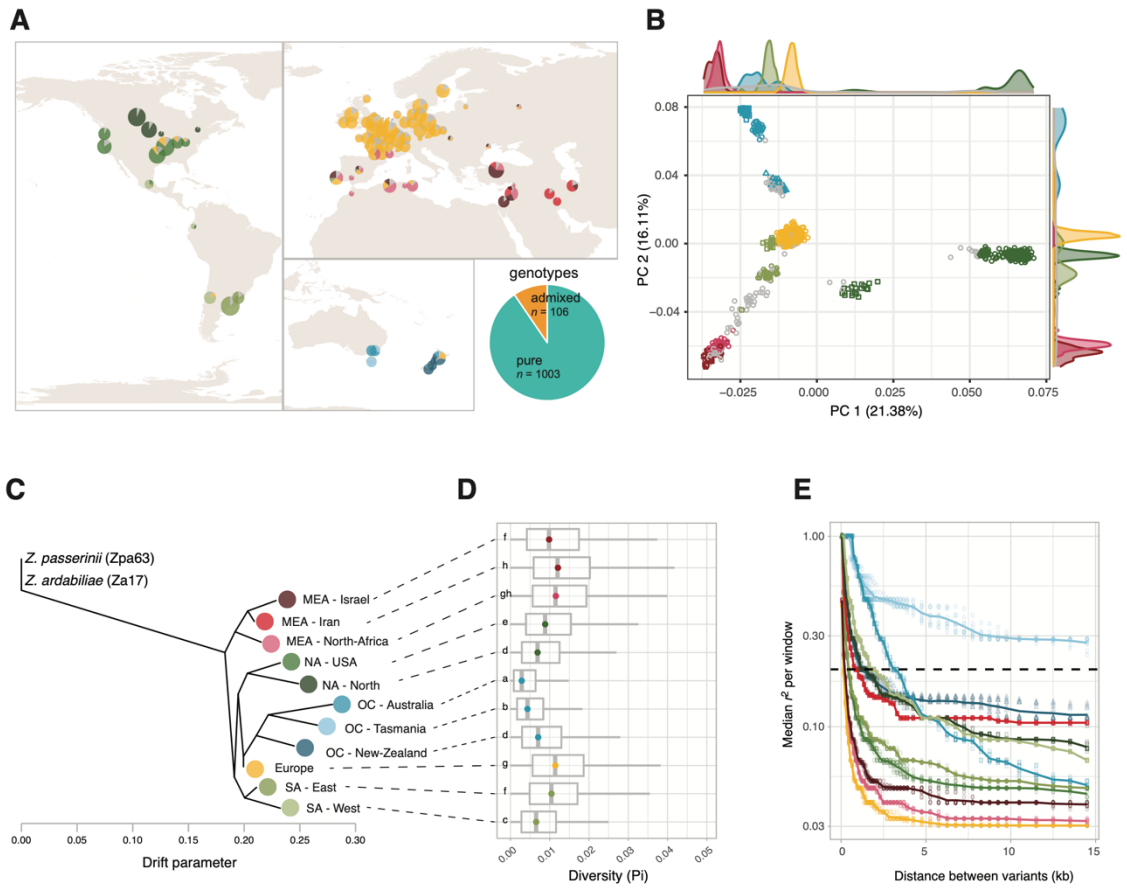

Figure 2: Global genetic structure based on 1109 genomes. A. Map of the genetic clustering based on a thinned genome-wide SNP dataset using sNMF. Each color represents a different genetic cluster, and the sizes of the slices represent the average attribution to the cluster across the isolates from each location. Fractions representing less than 10% of all genotypes of a location were colored in grey to improve clarity. The large pie chart outside of the map represents the proportion of isolates assigned clearly (≥75%) to a single genetic cluster (pure; in teal) and isolates identified as hybrids (admixed) between clusters (in yellow). B. Principal component analysis, showing the first and second component (PCs) based on a subset of variants. Colors and shapes indicate the genomic clusters identified with the sNMF method (with hybrids in grey). The marginal distributions represent the distribution for each PC. PCs 1 to 8 are shown in Fig S4. C. Population tree based on Treemix, rooted using two genomes from the sister species *Z. passerinii* and *Z. ardabiliae*. The colors are the same as in the previous panels and only samples which were fully assigned to a cluster were used. D. Diversity estimated with using *pi* per genetic cluster. The boxplots are ordered according to the tree of panel C. E. Linkage disequilibrium (r^2) between variants per genetic cluster. The color are identical that that of the other panels.

We analyzed the history of population splits and admixture using allele frequency information (Fig. 2C). The analyses largely supported a genetic structure shaped by the introduction of wheat across continents. The historical relationships between clusters show an early divergence of the Middle Eastern and North African clusters matching the early introduction of agriculture in these regions. Populations in Europe and the Americas share a similar time point of divergence consistent with extensive contributions of European genotypes to the Western hemisphere. Oceanian groups have diverged as a single branch from genotypes most closely related to extant European populations. Matching the

144 introduction of wheat to Oceania from the European continent, the Australian and New Zealand
145 pathogen populations share a common origin rooted in European genetic diversity. Populations from
146 Australia show also a striking loss of diversity and higher linkage disequilibrium compared to European
147 diversity consistent with a significant founder effect (Fig. 2D-E). Similarly, populations in South and
148 North America have reduced genetic diversity compared to extant European populations. The highest
149 diversity was found in populations from Africa and the Middle East closest to the center of origin.
150 Overall, the global genetic structure of the pathogen reveals multiple founder events associated with the
151 introduction of wheat to new continents.
152
153 Ongoing gene flow among regions should lead to admixed genotypes. We found that nearly 10% of all
154 analyzed genotypes showed contributions from at least two clusters. The most significant recent gene
155 flow was detected between Middle Eastern/North African clusters and European clusters in North
156 Africa (*i.e.*, Algeria and Tunisia) as well as Southern and Eastern Europe (*i.e.*, France, Italy, Hungary,
157 Ukraine, Portugal, and Spain; Table S3). We found a particularly high incidence of recent immigration
158 in a durum wheat population in the south of France. The population consisted only of hybrids or atypical
159 genotypes suggesting either recent migration from North Africa or host specialization on durum wheat
160 varieties. Additionally, we found hybrid genotypes with European ancestry in both North America and
161 in Oceania. The relatively balanced ancestry proportions in these hybrids suggest very recent gene flow
162 dating back to only a few generations. We further investigated past gene flow between clusters by
163 allowing Treemix to infer migration events, thus creating a population network (Fig. S5A-D). Three
164 distinct recent migration events were best explaining the data with specific migration routes from the
165 Middle East/African clusters to North America, from an Australian cluster to South America and
166 between two Oceanian clusters (Fig. S5D). However, the migration events did not affect the overall
167 shape of the inferred population tree (Fig. 2C, Fig. S5B-D). To better understand effects of long-distance
168 gene flow, we investigated the relationship between relatedness among genotypes (*i.e.,* identity-by-
169 state) and geographic distance. At the continent level, we observed a negative relationship between
170 identity-by-state and geographic distance (Fig. S6). The wide distribution of identity-by-state values
171 shows that although closely related isolates tend to be found at closer geographic distance, distantly
172 related isolates can be found at both far and close geographic distances. In combination, our findings
173 show an important role of long-distance dispersal impacting the genetic make-up of populations from
174 individual fields to continental scale genetic diversity.
175
176 ***Relaxation of genomic defenses against transposable elements concurrent with global spread***
177 Transposable elements (TEs) are drivers of genome evolution. In *Z. tritici*, TE activity created beneficial
178 mutations for fungicide resistance and virulence on the wheat host[16,17]. Due to the deleterious nature of
179 TEs, genomic defenses have evolved an array of mechanisms to counteract their activity. Hence, rapid
180 recent adaptation of the pathogen has benefitted from the activity of TEs with consequences for genome

181 size[18]. To analyze the effectiveness of genomic defenses against active TEs, we screened all genomes
182 for evidence of TE insertions. We found that the frequency spectrum of TE insertions is heavily skewed
183 towards low frequencies with 77% of TE insertions being found in single isolates (~0.1% frequency)
184 and 96% of insertions were found in ten or fewer isolates (<1%; Fig S7A) consistent with strong
185 purifying selection. The *Z. tritici* genome contains both core and accessory chromosomes (*i.e.,*
186 chromosomes not shared among all isolates of the species; Fig. 3A) [19]. Accessory chromosomes have
187 higher TE densities (Fig. 3B) reflecting lower selection pressure on accessory chromosomes[20]. Beyond
188 this, accessory chromosomal regions are broadly differentiated from core regions based on sequence,
189 transcription and epigenetic feature sets (Fig. 3C). The primary differentiator (*i.e.*, the first principal
190 component) separated euchromatic and heterochromatic (H3K9me2) chromosomal segments. These
191 differences in epigenetic marks were matching TE density variation, consequences of genomic defenses
192 (*i.e.,* repeat-induced point mutations) and GC content. The number of TE insertions was not correlated
193 with GC content, but was positively correlated with the presence of facultative heterochromatin
194 (H3K27me3) and negatively correlated with euchromatin marks. H3K27me3 is also a hallmark of
195 accessory chromosome segments consistent with previous findings[21–23]. Short insertion/deletion (indel)
196 polymorphism was positively correlated with the heterochromatic accessory regions, consistent with
197 purifying selection acting against indels mostly in gene-dense regions.
198
199 The pathogen genome shows variation in TE content among populations from different continents,
200 suggesting that the TE content was influenced by the colonization history of the species[18,24]. Hence, we
201 tested whether recently established populations were differentiated from center-of-origin populations.

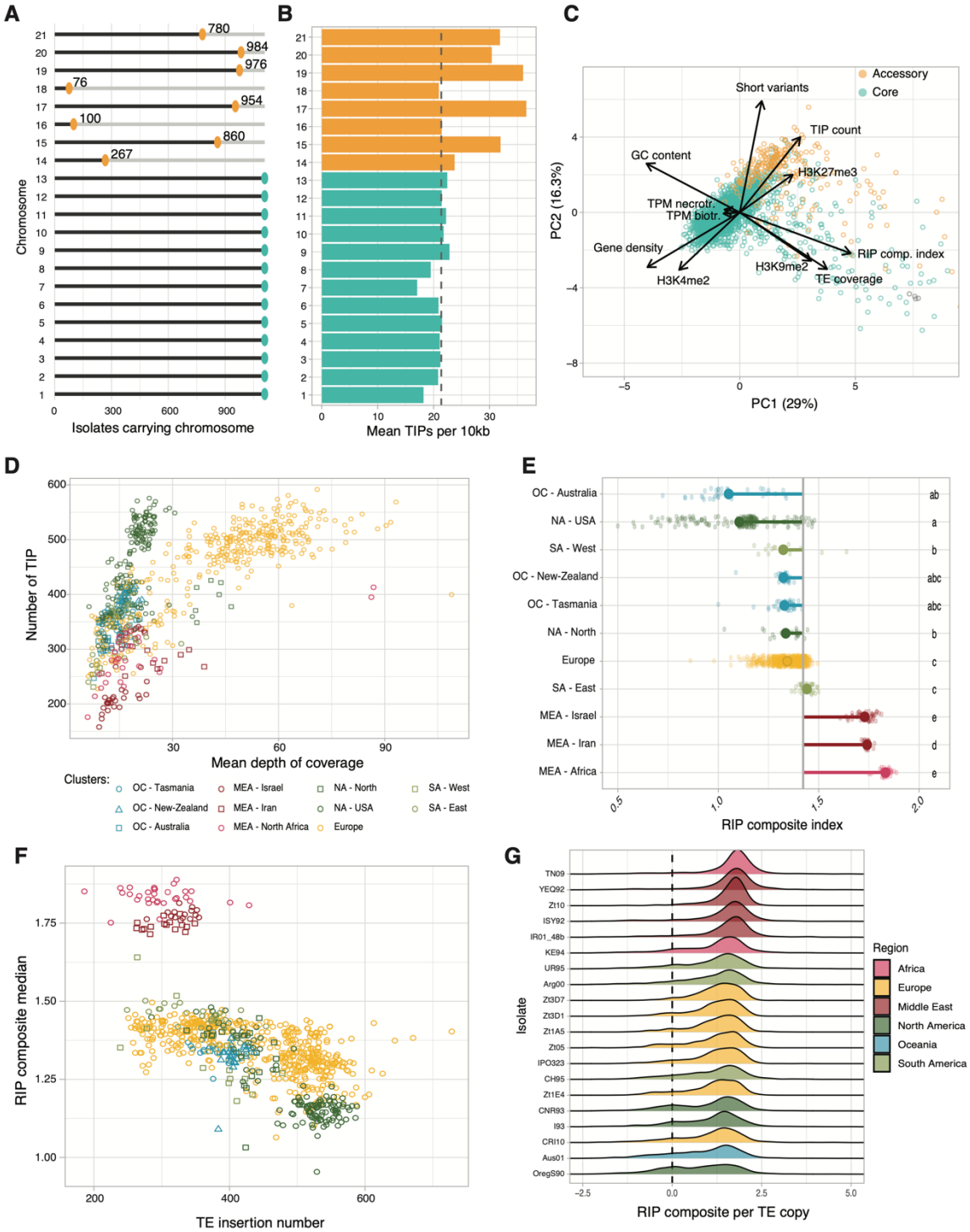

Figure 3: Relaxation of genomic defenses against transposable elements. A. Presence of each chromosome in the 1109 isolates assessed by depth of coverage. Yellow colors in panels A, B, and C represent the known accessory chromosomes while the core chromosomes are shown in teal. B. Mean number of transposable element insertion polymorphisms (TIPs) per 10-kb window for each chromosome. The dotted line represents the genome-wide mean. C. Principal component analysis using genomic, epigenomic and transcriptomic information per 10-kb window as well as number of variants (short variants and TIPs). D. Dot plot showing the number of TIPs and the genome-wide depth of coverage for each sequenced isolate. The color and shape differentiate the genetic clusters identified based on the genome-wide short variants. E. RIP composite index in the TEs for each isolate (small transparent dot) and as an average per genetic cluster (large opaque dot). The grey vertical line represents the overall average. F. Dot plot representing the number of TIPs per genome compared to the median of the RIP

composite in reads mapping on TE consensus sequences. The colors/shapes are the same as in panel D. G. Distribution of RIP composite index per TE copy in 20 high-quality genome assemblies for the species. The dotted line at value 0 corresponds to no detected RIP signal.

We found that TE insertion polymorphisms are most often specific to a genetic cluster with only 31% of insertions shared by two or more populations (Fig. S7B). We used TE insertion polymorphisms as a genetic marker and found that population differentiation was matching differentiation assessed using short variants (Fig. S7C). Hence, the population history of the pathogen was an important factor shaping TE content. We also found that the per-individual TE content has increased in most areas outside of the center of origin. Accounting for depth of sequencing, the TE content of the Middle Eastern and African genomes was lower than in any other region (Fig. 3D). In contrast, the Oceanian and American genomes contained among the highest TE numbers. We identified multiple-step increases in TE content in genomes sampled from outside of the center of origin, suggesting a relaxation of genomic defenses over the evolutionary history of the pathogen on wheat.

Many ascomycetes share a genome defense mechanism against TEs that can rapidly introduce targeted mutations into newly duplicated sequences, called repeat-induced point mutation (RIP)[25,26]. RIP machinery is active in genomes of *Z. tritici*, with high levels of RIP-like mutations identified in genomes from the center of origin and wild-grass-infecting sister species[24,27]. We analyzed the global panel for evidence of RIP-like mutations reporting the RIP composite index. The median index is above 1 in TE sequences across all genetic clusters. However, we found that RIP strength varies considerably among genetic clusters with the strongest signatures found in genomes from Middle Eastern and African isolates (Fig. 3E). The Middle Eastern and African clusters tend to include genomes with both low TE content and strong RIP signatures (Fig. 3F, S8A-B). In more recently colonized regions, genomes showed a negative correlation between the strength of RIP signatures and the amount of TEs per genome (p value $< 2.2e-16$; Fig. 3F & S8A). The association between TE content and strength of RIP signature is consistent with genomic defenses being less capable to prevent new TE insertions following migration out of the center of origin.

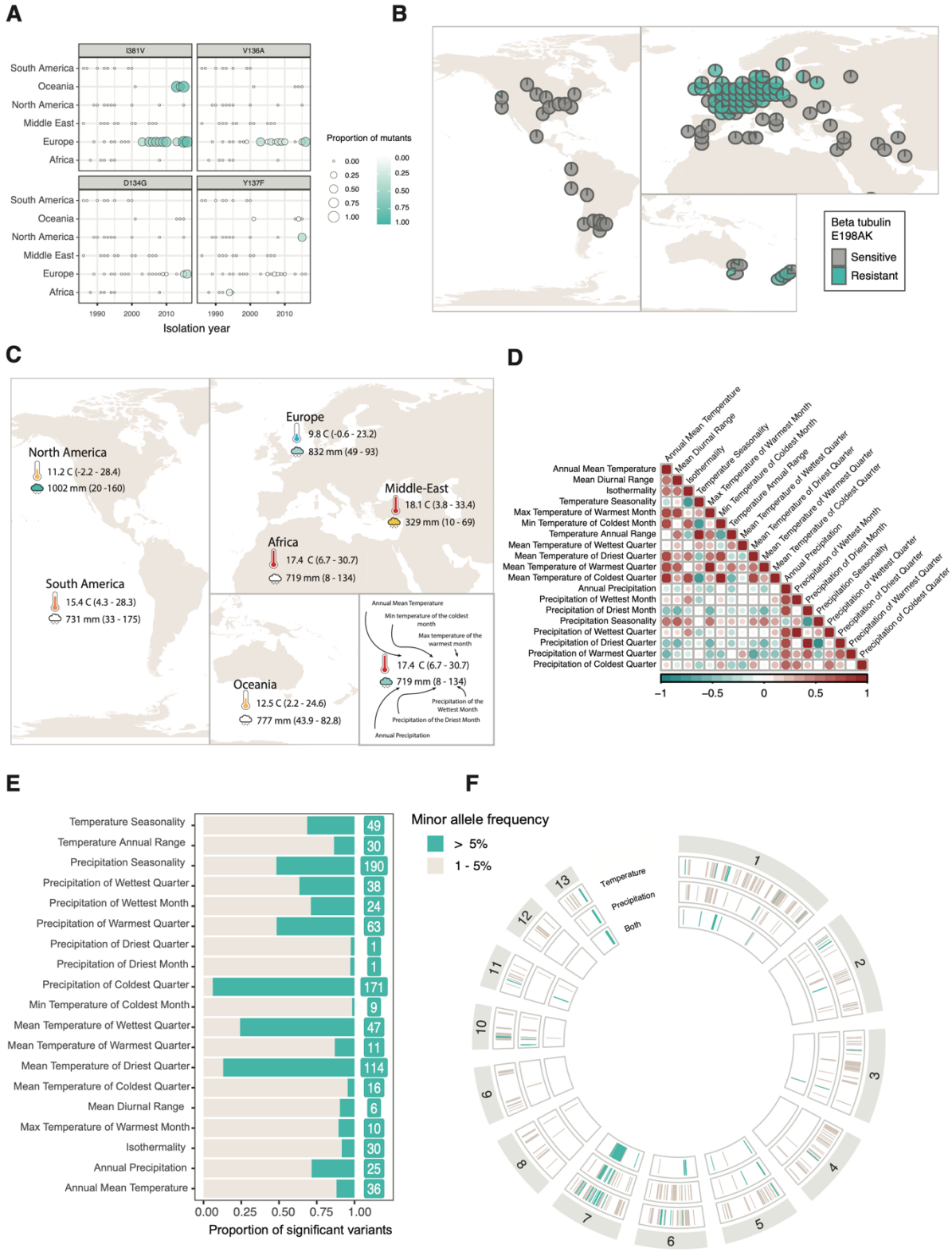

Figure 4: Adaptation to fungicides and climatic gradients. A. Mutations underlying azole fungicide resistance across time and regions. B. Map representing the allele frequency of known beta-tubulin resistance mutations (E198A/K) against benzimidazole fungicides. C. Examples of bioclimatic variables used for the genotype-environment association (GEA) analyses among regions. The values represent the average for each continent. D. Correlation plot between the 19 analyzed bioclimatic variables. E. Proportion and number of variants significantly associated with each bioclimatic variable and with a minor allele frequency (MAF) higher than 5%. F. Genomic

249 location of variants significantly associated with the bioclimatic variables grouped into three main categories
250 (variables describing temperature, precipitation levels, and variables describing combined measures of
251 temperature and precipitation).
252
253 How genomic defenses against TEs are modulated in fungi remains largely unknown. The RIP
254 machinery is activated during sexual recombination in *Neurospora*[28], thus suggesting that reduced
255 sexual reproduction could lower the efficacy of RIP. To assess links with rates of sex, we compared the
256 ratio of mating types in populations and found that the ratio was always close to 50:50, consistent with
257 frequent sexual reproduction (Fig. S8C). Alternatively, the RIP machinery could have lost function in
258 some populations. In this case, TEs should show bimodal signatures of RIP with old TEs carrying
259 signatures of historical RIP activity and recent insertions with no or weak RIP signatures. Indeed, a high
260 percentage of TEs shows only weak evidence for RIP (composite index <0.5) in genetic clusters outside
261 of the Middle East and Africa (mean = 20.1%, range from 11.3 - 34.6%; Fig. S8D). By contrast, in the
262 center of origin all genomes had less than 10% of TEs showing weak evidence for RIP (mean = 7.48%,
263 min = 5.18%). We confirmed the bimodal RIP signatures by analyzing individual TEs in 20
264 chromosome-level assembled genomes (Fig. 3G). All genomes shared a major RIP composite index
265 peak of ~2. A secondary peak indicative of TEs without RIP was found in genomes from the Americas,
266 Oceania and Europe. Despite the higher TE activity, we found no evidence that the loss in RIP
267 functionality also led to high gene duplication events. In *N. crassa,* two pathways mediate RIP with one
268 dependent on RID and one on Dim2[28]. In *Z. tritici*, the Dim2 methyltransferase was functionally linked
269 to the occurrence of RIP-like mutations in repeats[27]. However, the *dim2* gene was duplicated multiple
270 times in some genomes causing loss-of-function mutation in *dim2* caused by the RIP mechanism itself
271 [27,29]. We found that the ancestral copy of *dim2* shared higher identity with the functional copy of *dim2*
272 in the genomes of the Middle Eastern isolates than other populations (Fig. S8E). The European
273 populations had both the largest range in the number of detected paralogs and the highest copy numbers
274 overall (Fig. S8F). This is consistent with a deleterious runaway gene-duplication process affecting a
275 molecular component of the RIP machinery, explaining the loss of RIP efficacy within the species.
276
277 ***Adaptation to fungicides and changing climates along continental gradients***
278 Globally distributed pathogens experience significant environmental heterogeneity that potentially
279 constrains or facilitates future range expansions and adaptation. The use of pesticides across the globe
280 to combat agricultural pathogens has triggered the parallel emergence of resistance with significant
281 economic consequences[6]. To retrace the global emergence of fungicide resistance in *Z. tritici,* we
282 analyzed mutations in resistance genes using isolates collected over three decades (1986 to 2016). This
283 time span covers the introduction of several major fungicides to agricultural fields. Resistance to the
284 ubiquitously used azole fungicides is often mediated by mutations in the *CYP51* gene [30]. Recent North
285 American populations gained the Y137F mutation but not the I381V or V136A mutations rising in
286 frequency in Europe over the same time period (Fig. 4A). European populations harbored the most

287 diverse set of azole resistance mutations, consistent with the early and intense applications of this
288 fungicide class. The I381V and V136A mutations occurred at high frequency since the early 2000s,
289 whereas the S524T mutation was only observed at a low frequency with a delayed onset. Resistance
290 arose later in Oceania and North America, consistent with the later application of azoles in those
291 locations. No resistance mutations were detected in the Middle Eastern or African populations,
292 matching the absence of azole treatments in those regions. We found similar geographic patterns for the
293 E198AK mutation in the beta-tubulin gene associated with benzimidazole resistance as well as for the
294 G143A mutation in the mitochondrial gene *cytb*, known to cause resistance to Quinone outside
295 inhibitors fungicides (Fig. 4B). As expected from their more recent introduction, mutations related to
296 resistance to succinate dehydrogenase inhibitors (SDHI) were only observed in the most recently
297 sampled populations in Europe (Fig. S9). Overall, the global analyses of fungicide resistance signatures
298 show how European populations consistently developed the first known mutations to newly introduced
299 fungicide.
300
301 Changes in climatic conditions create complex challenges for plant breeders to create resilient crops[31].
302 Concurrently, pathogen populations are exposed to changes in temperature and humidity patterns. The
303 historic spread of *Z. tritici* has likely created significant selection pressure to adapt to climates
304 associated with the global range of wheat cultivation. Here, we analyze the genetic architecture of
305 climate adaptation by mapping standing variation along climatic clines. The pathogen is endemic to
306 regions with distinct climates, from the dry and warm conditions in the Middle East to the temperate
307 oceanic climate of Western Europe and the humid continental climates of some North American
308 locations. We performed genotype-environment association (GEA) mapping based on the climatic
309 conditions of the sampling locations. We analyzed a total of 19 bioclimatic variables covering annual
310 trends, seasonality and extreme environmental factors, such as the maximum temperature of the
311 warmest month or the precipitation of the driest quarter year (Fig. 4CD, Table S4).

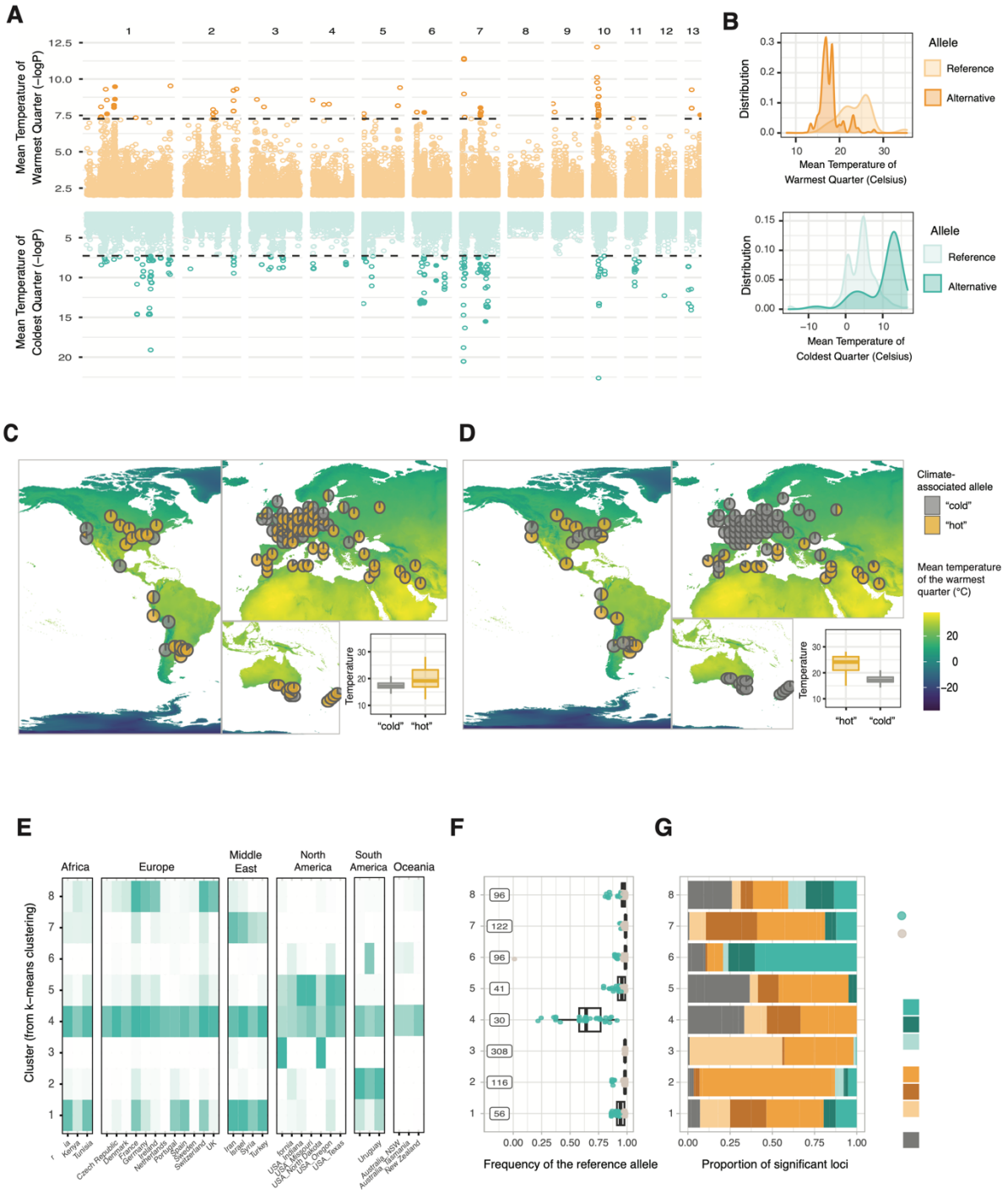

Figure 5: Pathogen adaptation along global climatic gradients. A. Manhattan plot for genotype-climate associations for two bioclimatic variables related to high and low temperatures, respectively. B. Density plots for two examples of significantly associated variants from panel A. The lightest-colored curve represents the climatic values of the sampling site for the isolates carrying the reference allele and the darkest color for the isolates carrying the alternative allele. The variant shown in yellow is located on chromosome 10 at position 452,864 bp. The variant in teal is located on chromosome 6 at position 1,686,518 bp. C. Map showing allele frequencies of a variant significantly associated with the mean temperature of the warmest quarter (chromosome 1, 2,090,068 bp). The associated box plots represent the distribution of the mean temperature of the warmest cluster at the sampling location of isolate carrying the two alleles. D. Identical to panel C for the variant at position 452864 bp on chromosome 10. E. Heat map representing the proportion of minor alleles for all the variants in each k-means cluster which are present in a given country (or state). Dark colored cells indicate that the minor allele is found at least once in the corresponding country (or state) for all the variants grouped in the corresponding k-means cluster.

A white colored cell indicates that the minor allele is absent for all the variants classified in the k-mean cluster. F. Minor allele frequency and number of variants classified in each k-mean cluster. G. Bioclimatic variables associated with the variants classified in each k-mean cluster.

We identified 1956 variants significantly associated with at least one climatic variable and 640 variants with a minor allele frequency (MAF) higher than 5%. The number of associated variants per climatic variable ranged between 36 and 541 (for BIO9 and BIO6 respectively), including 1-190 significant SNPs with a MAF > =5% per climatic variable (Fig. 4E). We investigated whether particular variant classes were enriched in the set of significantly associated SNPs. Using permutations, we found that both nonsynonymous and intergenic variants were more frequently associated with climatic variables than expected randomly while synonymous variants were significantly depleted (Fig. S10B). We found 187 genes that were in proximity or directly affected by variants associated with bioclimatic variables and with a MAF >= 5%, including 65 containing non-synonymous variants (Table S5). For each GEA, we retrieved significantly associated loci by clustering significant variants within a distance of 10 kb. The significant variants clustered into 5-27 distinct loci per GEA consistent with a polygenic basis for most climate adaptation (Fig. S11-15; Table S4). A large number of associated loci were shared between GEA of different climatic variables. Highly correlated climatic variables including BIO5 and BIO10, the maximum temperature of the warmest month and the mean temperature of the warmest quarter, respectively, shared also higher numbers of significantly associated SNPs (Fig. 4D, Fig, S10A). However, we also identified hotspots of climatic adaptation loci for largely independent climatic variables (Fig. 4F). We identified a large segment of chromosome 7 and a telomeric region of chromosome 13 to be hotspots for climate associations. The chromosome 7 locus overlaps with the *Cyp51* gene involved in azole fungicide resistance. Hence, the association could be due to correlations in the application of fungicides and climatic factors. Temperate regions such as Europe show higher azole resistance (see above) than the Middle East, thus leading to an indirect association between fungicide resistance genes and climatic variables. However, at the same location on chromosome 7 is a quantitative trait locus (QTL) for growth at suboptimal temperature[32], so that locus could well underpin climatic adaptation independent of *Cyp51*. Further analyses of temperature sensitivity QTLs showed that three out of the four previously described loci are overlapping with loci associated with climatic variables (Table S6). For example, the temperature-sensitivity QTL on chromosome 1 overlaps with loci associated with multiple temperature-related climatic variables such as the mean temperature of the warmest quarter (Fig. 5A-C). A variant associated with the mean temperature of the warmest quarter and overlapping with a QTL previously discovered in a central European cross (chr 1 at 2,090,068 bp) shows a global distribution of both alleles. By contrast, a second variant (chr 10 at 452,864 bp) associated with the mean temperature of the warmest quarter but not overlapping the previously discovered QTL (Fig. 5B and D), shows no allelic variation within central European populations. This highlights the power of covering global genetic diversity to gain comprehensive inisghts into the genetic architecture of recent adaptation in species.

To identify whether adaptive mutations tend to arise locally or occur across large geographic areas, we clustered significant loci based on the presence or absence of adaptive variants per country. We identified eight clusters characterized by shared adaptive variants with a similar geographic distribution (Fig. 5E, Fig. S16). Some clusters of adaptive variants are highly geographically localized (*i.e.,* clusters 3, 5 and 6) while other clusters are widespread (*i.e.*, cluster 4). Most adaptive alleles for extreme cold conditions were found in the populations subjected to the harsh winters of continental North America (see cluster 3; Fig. 5E-G). Clusters of adaptive alleles were geographically widespread such as the distributions along the Mediterranean coast (see cluster 1). Taken together, the global genome panel revealed substantial standing variation for environmental adaptation with complex geographic patterns of local adaptation evolution.

Conclusions

Analysis of a thousand-genome panel recapitulated the spread of a major fungal crop pathogen revealing tight links to the history of global wheat cultivation. The early divergence between Middle Eastern and African genotypes from those collected in the rest of the world is consistent with a single expansion event from the center of origin dating back millennia. The extant genetic variation was strongly shaped by successive colonization bottlenecks during the introduction of the pathogen to the Americas and Oceania. The distinct loss of genetic diversity and increased linkage disequilibrium likely caused the loss of adaptive genetic variation and reduced evolutionary potential in the most recently colonized regions. TE activity has underpinned the rise of major adaptive mutations in the species[16,33,34]. Remarkably, the TE activity is underpinned by a marked relaxation or even loss of genomic defenses following population bottlenecks during global colonization. The higher activity of TEs is likely a direct consequence of reduced control and may have long-term consequences for the pathogen. The relaxation of genomic defenses in populations from the American, Oceanian and European continents could have been selected as an evolvability trait, thus increasing variance in fitness in populations. The relaxation of genomic defenses likely underpins incipient genome expansion within the species while increasing the risk of mutational meltdown. The resilience of crops and agricultural ecosystems is threatened by the changing climate. The ability of pathogens to adapt and expand their range under altered humidity and temperature regimes as well as changes in seasonal patterns is a major concern. The identified genomic regions associated with adaptation to environmental conditions highlight how a global pathogen carries extensive variation to cope with climate change. Integrating genetic and population genetic information of pathogen adaptation to climatic gradients is a powerful asset for risk models of future pathogen spread.

398 **Methods**

399

400 ***Sample collection, culturing and sequencing***

401 Fungal isolates were grown on V8, yeast sucrose broth (YSB) or yeast malt agar (YMA) plates or liquid
402 culture prior to DNA extraction. Complete information about the geographic origin, date of collection,
403 available sequencing datasets and references are given in Table S1. Lyophilized or frozen fungal tissue
404 was used for DNA extraction with QIAGEN kits (DNeasy Plant Mini Kit, QIAcube HT). Sequencing
405 libraries were prepared from sheared DNA on an Illumina sequencing platform following TruSeq
406 library preparations. For the collection from the Joint Genome Institute (see Table S1), DNA was
407 extracted from single-spore isolates, the fragments were treated with end-repair, A-tailing, and ligation
408 of Illumina-compatible adapters (IDT, Inc) using the KAPA-Illumina library creation kit (KAPA
409 biosystems). Plate-based DNA library preparation for Illumina sequencing was performed on the
410 PerkinElmer Sciclone NGS robotic liquid handling system using Kapa Biosystems library preparation
411 kit. 200 ng of sample DNA was sheared to 600 bp using a Covaris LE220 focused-ultrasonicator. The
412 sheared DNA fragments were size selected by double-SPRI and then the selected fragments were end-
413 repaired, A-tailed, and ligated with Illumina-compatible sequencing adaptors from IDT containing a
414 unique molecular index barcode for each sample library. The prepared libraries were quantified using
415 KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480
416 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina
417 HiSeq sequencing platform utilizing a TruSeq paired-end cluster kit, v4. Sequencing of the flow cell
418 was performed on the Illumina HiSeq2500 sequencer using HiSeq TruSeq SBS sequencing kits, v4,
419 following a 2x150 indexed run recipe. Overall, we obtained 1368 Illumina resequencing datasets for
420 quality evaluation.

421

422 ***Draft de novo assembly and variant calling procedures***

423 We created *de novo* draft assemblies using the software SPADES v.3.14.1[35] with the "careful" method.
424 We used only the assemblies (filtered to remove any contigs shorter than 1 kb) with fewer than 1600
425 contigs and a total length between 30 and 4 Mb.
426 We trimmed and filtered the reads with Trimmomatic v.0.39, thereby removing adapter sequences,
427 trimming leading and trailing bases with a quality lower than 15 for the resequencing of 2020 and 10
428 for all previous resequencing, and removing sequences shorter than 50 bp[36]. The trimmed reads were
429 mapped to the reference genome of IPO323[19] using bowtie2 v.2.4.1[37]. GATK v4.1.4.1 was used for
430 short-variant calling with the commands HaplotypeCaller, CombineGVCFs, and GenotypeGVCFs,
431 setting the ploidy to 1 and the maximum number of alternative alleles to 2[38].

432 To filter out erroneously called short indels and SNPs, we started with a standard set of hard filters
433 using the GATK quality metrics, for which the thresholds were set based on visualization of the metrics
434 across the called variants. The per-site filters included: FS > 10, MQ < 20, QD < 20, ReadPosRankSum
435 between -2 and 2, MQRankSum between -2 and 2, and BaseQRankSum between -2 and 2. We also
436 included a per-genotype filter, removing any genotype with a depth lower than 3.
437 We further assessed the quality of our SNP calling using 8 isolates which were sequenced two times
438 (including in some cases in different sequencing datasets). We made the assumption that the real
439 variation between these pairs should be close to 0, although we cannot completely exclude the
440 possibility of a small number of mutations happening during culturing or maintaining of these isolates
441 in collection. We used the differences, i.e., erroneously called variants, between the resequencing pairs
442 as an estimation of genotyping errors and to identify the causes of genotyping errors. Most of the
443 erroneously called variants remaining after the hard filtering were related to genotypes with near-equal
444 numbers of reads supporting the reference allele and an alternative allele, i.e.., "heterozygous" alleles.
445 Such genotypes were called with a high confidence despite the fact that such a heterozygous-like pattern
446 should be recognized as errors in a haploid organism and could be due to misalignment or repeated
447 sequences in the genomes. We consequently implemented an allelic balance custom script to recognize
448 such positions and to filter out any genotype that had fewer than 90% of reads supporting the called
449 allele. This filter removed 75% of the erroneous variants left between the resequenced pairs after the
450 hard filtering. As the rest of the erroneous variants were related to low sequencing depth, we further
451 implemented a per-sample missing data and low-depth filtering, removing any sample with more than
452 20% of missing data and a mean depth of coverage lower than 6 on the core chromosomes (based on
453 vcftools –missing-indv and --depth options)[39].
454 In the next filtering step, we removed the samples that were clones or near-clones. To identify these
455 isolates, we created a network of isolates with an identity-by-state value superior to 0.99 (as measured
456 by plink v1.9[40]) and extracted the subgraphs designating groups of clones (with the R packages
457 tidygraph and ggraphs for visualization). In each group of clones, we filtered out all isolates except for
458 the isolate with the lowest amount of missing data. These per-sample filters resulted in a final isolate
459 count of 1109.
460 The final filtering step was a per-site filter based on the number of missing genotypes. Considering that
461 *Z. tritici* contains accessory chromosomes that are expected to be present in some isolates and absent in
462 others, the relevant threshold of missing data has to be adapted per chromosome. To identify the
463 presence-absence of accessory chromosomes, we assessed the depth of coverage in windows across all
464 chromosomes with bedtools v.2.29.2 (option coverage followed by the option groupby to calculate the
465 median per window)[41]. We normalized the depth estimates using the median depth over all windows of
466 the core chromosomes for the per-window depth. The normalized depth was then used to infer presence-
467 absence variants or copy-number variation of chromosomes, in which we considered that any
468 chromosome with a normalized depth lower than 0.2 was absent. Based on the estimated number of

469 chromosomes present in the dataset, we calculated missing data thresholds at 80% of genotyped isolates
470 with the --max-missing-count option of vcftools (NA_{max} = 222 for the core chromosomes and between
471 328 and 1048 for the accessory chromosomes)[39].
472
473 ***Population structure and population-level statistics***
474 We used a subset of the filtered biallelic SNPs (one SNP every 1 kb, no missing data and a minor allele
475 frequency of 0.05) to estimate the population structure of our worldwide *Z. tritici* collection. This was
476 done separately with a principal component analysis (R package SNPRelate[42]) and with a snmf
477 clustering method from the LEA package[43]. The clustering analysis ran for a value of K (i.e., the number
478 of clusters) ranging from 1 to 15 and with 10 repetitions per K. To identify the best K, we used the
479 entropy method implemented in snmf which evaluates the quality of fit of the model to the data, as well
480 as the smallest cluster size and the number of isolates assigned to any cluster with a coefficient higher
481 than 0.75 (considered as non-admixed).
482
483 For the analyses relying on the comparison of distinct groups, we discretized the populations by using
484 only isolates belonging to one of the populations with a proportion higher than 0.75 at K=11, the inferred
485 best number of clusters. These genotypes were then used as input for treemix which infers splits between
486 populations and creates a population tree[44]. We ensured that the tree shape was consistent regardless of
487 possible migration events by running treemix with a number of possible migration events ranging from
488 0 to 6. We used the assembled genomes of *Z. ardabiliae* and *Z. passerinii* as outgroups to root the tree[45].
489 We used the scikit-allel python package to measure genetic diversity (pi), taking into account only the
490 non-admixed isolates from each cluster, in non-overlapping windows of 1 kb[46]. To remove windows
491 with too much missing data (i.e., those that would artifactually lower the diversity), we selected only
492 windows in which less than 20% of the variants were filtered out. We controlled for the variation in
493 isolate numbers between clusters by subsampling each cluster 10 times to the smallest cluster size (N=
494 16) and averaging the obtained diversity estimates per window over the 10 subsamples.
495
496 ***Transposable elements and repeat-induced point mutations***
497 We called the TE insertions with the software ngs-te-mapper2[47]. In this process, the sequencing reads
498 are first queried against a library of TE sequences, for which we used the TE consensus sequences
499 obtained from 20 fully assembled genomes of 19 global isolates[48]. The "junction reads" that align both
500 on a TE consensus sequence and on the flanking genome are used to determine the site of insertion of
501 reference and non-reference TEs. To take into account any inaccuracies in the detection of the insertion
502 sites, the positions of the insertions were rounded to 100 bp, so that insertions of the same element in a
503 short window were considered to be the same insertion for further analyses. The insertions found in
504 more than 10 samples were used to create a PCA (prcomp function from the stats R package), clustering
505 the isolates based on their shared TE insertions.

506 We investigated the genomic distribution of variants, in relation to genomics, transcriptomics and
507 epigenomics estimates. We gathered information from several sources and aggregated the data in 10-
508 kb windows. We used transcriptomic data produced previously[49] and analyzed to calculate TPM[45],
509 representing the gene expression during the necrotrophic and the asymptomatic phases of infection for
510 the reference isolate. To include epigenomics data, we used previously published histone mark ChIP-
511 seq data[21] and identified the peaks for several histone marks: H3K4me2, H3K27me3, and H3K9me2.
512 We trimmed and mapped the reads using trimmomatic as above and bowtie, and filtered the mapped
513 reads to keep only those aligning with a quality higher than 30 using samtools. The histone mark peaks
514 were called using macs2 (option —no-model)[50], and only the regions appearing in both repeats were
515 kept (bedtools intersect). To remove windows with low variant counts due to a low mappability, we
516 used genmap to estimate mappability and removed windows with a value lower than 0.85 (threshold
517 estimated visually based on a density plot of windows across the genome)[51].
518
519 To analyze the repeat-induced point mutations in the collection of genomes, we used the same
520 consensus TE sequences[48] as a reference genome for mapping of reads (as single reads) with bowtie2.
521 Using a custom python script based on the biopython library[52], we also estimated the GC content and
522 the RIP composite index[53], an estimate created to detect ratios of dimers indicative of mutations typical
523 of the repeat-induced point mutation process (RIP). This was done on the reads aligning to the TE
524 consensus sequences, thus providing an estimation of RIP in all transposable elements as well as per
525 TE consensus. We also used previously computed values of the RIP composite index for 20
526 chromosome-level assemblies[24].
527
528 One of the most important enzymes for RIP is dim2. Based on previous knowledge that the strain Zt10
529 contains a functional copy of dim2[27], we extracted the sequence of the gene (Zt10_unitig_006_0417)
530 as well as its two flanking genes (Zt10_unitig_006_0416, and Zt10_unitig_006_0418). These were then
531 used as query sequences for the software *blast* to detect the presence and location of the 3 genes in *de*
532 *novo* draft assemblies based on the Illumina resequencing. In many isolates dim2 is found in multiple
533 copies. We thus identified the native copy as the copy found between the two flanking genes or, when
534 the assemblies did not include all three genes on one contig, the copy found within less than 10 kb of
535 one of the flanking genes. We then considered the percentage of identity between the native copies and
536 the functional copy of Zt10.
537
538 Statistical differences between geographical groups were assessed using a one-way ANOVA with
539 blocks, with the sequencing batch considered as the confounding block. The sequencing batch effect
540 was especially strong for the genomes from Hartmann et al.[54] probably due to a strong GC bias in the
541 sequencing. As a post-hoc analysis, we used mean separation tests (least-square means) and displayed
542 the results as letters on the corresponding plots.

543

544 ***Adaptation and selection***

545 We obtained geographical coordinates from the metadata attached to the isolates or inferred them based
546 on the most precise sampling location available. Coordinates inferred can be found in Table S1. We
547 downloaded gridded weather and climate data at the 10' resolution from the WorldClim database
548 version 2 (WorldClim.org). The geographical coordinates were used to approximate the environmental
549 conditions of origin for each isolate from all bioclimatic variables, which include for example the mean
550 diurnal range and the annual precipitation as an average between 1970 and 2000. Based on these
551 environmental estimates and the genomic variants, we identified genotype-to-environment associations
552 with the software GEMMA `0.98.3`[55]. We used a LOCO (Leave-One-Chromosome-Out) approach to
553 estimate the kinship matrix on the genome excluding the particular chromosome on which we were
554 estimating the associations. Significance threshold was set using the Bonferroni correction method, i.e.,
555 by dividing the traditional threshold of significance of 0.05 by the numbers of variants that were tested.
556 Nearby significant SNPs were grouped together in "significant loci" if they were closer than 10 kb.
557 To identify the genes which are potentially causal for adaptation to climatic conditions, we predicted
558 the effect of the significant variants on the genes and proteins using SnpEff[56]. We created a custom
559 SnpEff database so that the predictions would match the gene annotation we are using[57] and setting the
560 upstream/downstream interval length to 1 kb. We also used the SnpEff predictions to compare the
561 distribution of effects (synonymous, non-synonymous and modifier) in the significantly associated
562 variants and in all the variants using 200 random draws.

563

564 We investigated the geographic distribution of the potentially adaptive alleles we found through k-
565 means clustering of each allele's presence-absence per country. Per significant locus and per bioclimatic
566 variable, we selected the variant with the lowest p value (i.e., the top variant in each "peak"), and
567 identified whether the minor allele was present or absent per country/state including more than 5
568 isolates. The matrix of presence/absence was then used for the clustering. This analysis did not reveal
569 a clear-cut pattern of adaptation sharing a set number of geographic distributions. Although there was
570 no clear-cut best number of clusters even when testing up to 30 clusters, we chose a number of 8 clusters
571 for graphical representation, using the elbow method. To investigate fungicide resistance, we identified
572 the presence/absence of known resistance alleles in the isolates of the dataset, following the method in
573 [15]. We then compared the frequency of these alleles in the different geographic locations, and through
574 time.

575

576 **Code availability**

577 To ensure reproducibility of the analyses presented in this manuscript, all custom scripts are available
578 at https://github.com/afeurtey/WW_PopGen. Post-processing and visualization of the data were done
579 in R, bash and python, available as R markdown reports in the github repository.

580

581 **Data availability**

582 All sequencing data is available from the NCBI Sequence Read Archive. Individual accession numbers
583 can be retrieved from Table S1.

584

585 **Author contributions**

586 AF and DC conceived the study; AF and CL performed analyses; all co-authors contributed samples
587 and/or datasets; BAM and DC supervised the work; AF and DC wrote the manuscript with input from
588 all co-authors.

589

590 **Competing interests**

591 We declare that we have no competing interests.

592

593 **Funding**

594 DC and GS were supported by the Swiss Innovation Agency Innosuisse. The work (proposal:
595 10.46936/10.25585/60000699) conducted by the U.S. Department of Energy Joint Genome Institute
596 (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of
597 Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231 and lab
598 work was supported by the USDA-Agricultural Research Service project 3602-22000-015-00D.

599

600 **Acknowledgements**

601 Data produced and analyzed in this paper were generated in collaboration with the Genetic Diversity
602 Centre (GDC), ETH Zurich. We would like to thank Lucio Garcia and the NGS platform at
603 Syngenta for assistance with sequencing.

604

605

606 **References**
607

608 1. Santini, A., Liebhold, A., Migliorini, D. & Woodward, S. Tracing the role of human civilization in
609 the globalization of plant pathogens. *ISME J.* **12**, 647–652 (2018).
610 2. Feurtey, A. *et al.* Strong phylogeographic co-structure between the anther-smut fungus and its
611 white campion host. *New Phytol.* 668–679 (2016) doi:10.1111/nph.14125.
612 3. Hartmann, F. E. *et al.* Congruent population genetic structures and divergence histories in anther-
613 smut fungi and their host plants Silene italica and the Silene nutans species complex. *Mol. Ecol.*
614 **29**, 1154–1172 (2020).
615 4. Stukenbrock, E. H. & McDonald, B. A. The Origins of Plant Pathogens in Agro-Ecosystems.
616 *Annu. Rev. Phytopathol.* **46**, 75–100 (2008).
617 5. Fones, H. N. *et al.* Threats to global food security from emerging fungal and oomycete crop
618 pathogens. *Nat. Food* **1**, 332–342 (2020).
619 6. Fisher, M. C., Hawkins, N. J., Sanglard, D. & Gurr, S. J. Worldwide emergence of resistance to
620 antifungal drugs challenges human health and food security. *Science* **360**, 739–742 (2018).
621 7. Khoury, C. K. *et al.* Increasing homogeneity in global food supplies and the implications for food
622 security. *Proc. Natl. Acad. Sci.* **111**, 4001–4006 (2014).
623 8. Bebber, D. P., Ramotowski, M. A. T. & Gurr, S. J. Crop pests and pathogens move polewards in a
624 warming world. *Nat. Clim. Change* **3**, 985–988 (2013).
625 9. Excoffier, L., Foll, M. & Petit, R. J. Genetic Consequences of Range Expansions. *Annu. Rev. Ecol.*
626 *Evol. Syst.* **40**, 481–501 (2009).
627 10. Petit-Houdenot, Y., Lebrun, M.-H. & Scalliet, G. Understanding plant-pathogen interactions in
628 Septoria tritici blotch infection of cereals. in *Burleigh Dodds Series in Agricultural Science* (ed.
629 Oliver, R.) 263–302 (Burleigh Dodds Science Publishing, 2021). doi:10.19103/AS.2021.0092.10.
630 11. Stukenbrock, E. H., Banke, S., Javan-Nikkhah, M. & McDonald, B. A. Origin and domestication
631 of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol. Biol.*
632 *Evol.* **24**, 398–411 (2007).
633 12. Singh, N. K., Karisto, P. & Croll, D. Population-level deep sequencing reveals the interplay of
634 clonal and sexual reproduction in the fungal wheat pathogen *Zymoseptoria tritici*. *Microb.*
635 *Genomics* **7**, 000678 (2021).
636 13. McDonald, B. A., Suffert, F., Bernasconi, A. & Mikaberidze, A. How large and diverse are field
637 populations of fungal plant pathogens? The case of *Zymoseptoria tritici*. 2022.03.13.484150
638 Preprint at https://doi.org/10.1101/2022.03.13.484150 (2022).
639 14. McDonald, M. C. *et al.* Rapid Parallel Evolution of Azole Fungicide Resistance in Australian
640 Populations of the Wheat Pathogen *Zymoseptoria tritici*. *Appl. Environ. Microbiol.* **85**, e01908-
641 18 (2019).
642 15. Hartmann, F. E. *et al.* The complex genomic basis of rapid convergent adaptation to pesticides
643 across continents in a fungal plant pathogen. *Mol. Ecol.* **30**, 5390–5405 (2021).
644 16. Omrane, S. *et al.* Plasticity of the MFS1 Promoter Leads to Multidrug Resistance in the Wheat
645 Pathogen *Zymoseptoria tritici*. *mSphere* **2**, e00393-17 (2017).
646 17. Meile, L. *et al.* A fungal avirulence factor encoded in a highly plastic genomic region triggers
647 partial resistance to septoria tritici blotch. *New Phytol.* **219**, 1048–1061 (2018).
648 18. Oggenfuss, U. *et al.* A population-level invasion by transposable elements triggers genome
649 expansion in a fungal pathogen. *eLife* **10**, e69249 (2021).
650 19. Goodwin, S. B. *et al.* Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella*
651 *graminicola* Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis.
652 *PLoS Genet.* **7**, e1002070 (2011).

653 20. Grandaubert, J., Dutheil, J. Y. & Stukenbrock, E. H. The genomic determinants of adaptive
654 evolution in a fungal pathogen. *Evol. Lett.* **3**, 299–312 (2019).
655 21. Schotanus, K. *et al.* Histone modifications rather than the novel regional centromeres of
656 *Zymoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics Chromatin* **8**, 41
657 (2015).
658 22. Galazka, J. M. & Freitag, M. Variability of chromosome structure in pathogenic fungi – of "ends
659 and odds". *Curr. Opin. Microbiol.* **0**, 19–26 (2014).
660 23. Möller, M. *et al.* Destabilization of chromosome structure by histone H3 lysine 27 methylation.
661 *PLOS Genet.* **15**, e1008093 (2019).
662 24. Lorrain, C., Feurtey, A., Möller, M., Haueisen, J. & Stukenbrock, E. Dynamics of transposable
663 elements in recently diverged fungal pathogens: lineage-specific transposable element content
664 and efficiency of genome defenses. *G3 GenesGenomesGenetics* **11**, (2021).
665 25. van Wyk, S., Wingfield, B. D., De Vos, L., van der Merwe, N. A. & Steenkamp, E. T. Genome-
666 Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. *Front. Microbiol.* **11**,
667 (2021).
668 26. Testa, A. C., Oliver, R. P. & Hane, J. K. OcculterCut: A Comprehensive Survey of AT-Rich
669 Regions in Fungal Genomes. *Genome Biol. Evol.* **8**, 2044–2064 (2016).
670 27. Möller, M. *et al.* Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in
671 repeats and changes evolutionary trajectory in a fungal pathogen. *PLOS Genet.* **17**, e1009448
672 (2021).
673 28. Gladyshev, E. & Kleckner, N. DNA sequence homology induces cytosine-to-thymine mutation
674 by a heterochromatin-related pathway in *Neurospora*. *Nat. Genet.* **49**, 887–894 (2017).
675 29. Dhillon, B., Cavaletto, J. R., Wood, K. V. & Goodwin, S. B. Accidental Amplification and
676 Inactivation of a Methyltransferase Gene Eliminates Cytosine Methylation in *Mycosphaerella*
677 *graminicola*. *Genetics* **186**, 67–77 (2010).
678 30. Hawkins, N. J. & Fraaije, B. A. Fitness Penalties in the Evolution of Fungicide Resistance. *Annu.*
679 *Rev. Phytopathol.* **56**, 339–360 (2018).
680 31. Ceccarelli, S. *et al.* Plant breeding and climate changes. *J. Agric. Sci.* **148**, 627–637 (2010).
681 32. Lendenmann, M. H., Croll, D., Palma-Guerrero, J., Stewart, E. L. & McDonald, B. A. QTL
682 mapping of temperature sensitivity reveals candidate genes for thermal adaptation and growth
683 morphology in the plant pathogenic fungus *Zymoseptoria tritici*. *Heredity* **116**, 384–394 (2016).
684 33. Wang, C., Milgate, A. W., Solomon, P. S. & McDonald, M. C. The identification of a transposon
685 affecting the asexual reproduction of the wheat pathogen *Zymoseptoria tritici*. *Mol. Plant Pathol.*
686 **22**, 800–816 (2021).
687 34. Krishnan, P. *et al.* Transposable element insertions shape gene regulation and melanin production
688 in a fungal pathogen of wheat. *BMC Biol.* **16**, 78 (2018).
689 35. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-
690 cell sequencing. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **19**, 455–77 (2012).
691 36. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence
692 data. *Bioinformatics* **30**, 2114–2120 (2014).
693 37. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**,
694 357–359 (2012).
695 38. Auwera, G. van der & O'Connor, B. D. *Genomics in the cloud: using Docker, GATK, and WDL*
696 *in Terra*. (O'Reilly Media, 2020).
697 39. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
698 40. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage
699 analyses. *Am. J. Hum. Genet.* **81**, 559–75 (2007).

41. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Zheng, X. *et al.* A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinforma. Oxf. Engl.* **28**, 3326–8 (2012).
43. Frichot, E. & François, O. LEA: An R package for landscape and ecological association studies. *Methods Ecol. Evol.* **6**, 925–929 (2015).
44. Pickrell, J. K. & Pritchard, J. K. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genet.* **8**, e1002967 (2012).
45. Feurtey, A. *et al.* Genome compartmentalization predates species divergence in the plant pathogen genus *Zymoseptoria*. *BMC Genomics* **21**, 588 (2020).
46. Miles, A. *et al.* cggh/scikit-allel: v1.3.3. (2021) doi:10.5281/zenodo.4759368.
47. Linheiro, R. S. & Bergman, C. M. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in *Drosophila melanogaster*. *PLOS ONE* **7**, e30008 (2012).
48. Badet, T., Oggenfuss, U., Abraham, L., McDonald, B. A. & Croll, D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. *BMC Biol.* **18**, 12 (2020).
49. Haueisen, J. *et al.* Highly flexible infection programs in a specialized wheat pathogen. *Ecol. Evol.* (2018) doi:10.1002/ece3.4724.
50. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
51. Pockrandt, C., Alzamel, M., Iliopoulos, C. S. & Reinert, K. GenMap: ultra-fast computation of genome mappability. *Bioinformatics* **36**, 3687–3692 (2020).
52. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
53. Lewis, Z. A. *et al.* Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. *Genome Res.* **19**, 427–437 (2009).
54. Hartmann, F. E., McDonald, B. A. & Croll, D. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. *Mol. Ecol.* **27**, 2725–2741 (2018).
55. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
56. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**, 80–92 (2012).
57. Grandaubert, J., Bhattacharyya, A. & Stukenbrock, E. H. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus *Zymoseptoria* Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. *G3 Bethesda Md* **5**, 1323–33 (2015).