Population genomics of the maize pathogen *Ustilago maydis*: demographic history and role of virulence clusters in adaptation

Gabriel Schweizer^{1,2*}, Muhammad Bilal Haider³, Gustavo V. Barroso^{3,4}, Nicole Rössel¹, Karin Münch¹, Regine Kahmann¹, and Julien Y. Dutheil^{1,3,5*}

¹Max-Planck-Institute for Terrestrial Microbiology, Department of Organismic Interactions, Karl-von-Frisch-Straße 10, 35043 Marburg, Germany.

²Present address: University of Zürich, Department of Evolutionary Biology and Environmental Studies, Winterthurerstrasse 190, 8057 Zürich, Switzerland.

³Max-Planck-Institute for Evolutionary Biology, Research Group Molecular Systems Evolution, August-Thienemann-Straße 2, 24306 Plön, Germany.

⁴Present address: University of California Los Angeles, Department of Ecology and Evolutionary Biology, 621 Charles E. Young Drive South 90095, 1606 Los Angeles CA, USA.

⁵University of Montpellier 2, Institute of Evolutionary Sciences of Montpellier, Place Eugène Bataillon, 34095 Montpellier, France.

* Authors for correspondence:

Gabriel Schweizer
University of Zürich
Department of Evolutionary Biology and Environmental Studies
Winterthurerstrasse 190
8057 Zürich
Switzerland
Phone: +41 44 635 6143
Fax: +41 44 635 6144
Email: gabriel.schweizer@ieu.uzh.ch

Julien Y. Dutheil
Max-Planck-Institute for Evolutionary Biology
Research Group Molecular Systems Evolution
August-Thienemann-Straße 2
24306 Plön
Germany
Phone: +49 4522 763 298
Fax: +49 49 4522 763 281
Email: dutheil@evolbio.mpg.de

45 Abstract

46 The tight interaction between pathogens and their hosts results in reciprocal selective forces that impact
 47 the genetic diversity of the interacting species. The footprints of this selection differ between
 48 pathosystems because of distinct life-history traits, demographic histories, or genome architectures. Here,
 49 we studied the genome-wide patterns of genetic diversity of 22 isolates of the causative agent of the corn
 50 smut disease, *Ustilago maydis*, originating from five locations in Mexico, the presumed center of origin
 51 of this species. In this species, many genes encoding secreted effector proteins reside in so-called
 52 virulence clusters in the genome, an arrangement that is so far not found in other filamentous plant
 53 pathogens. Using a combination of population genomic statistical analyses, we assessed the geographical,
 54 historical and genome-wide variation of genetic diversity in this fungal pathogen.

55 We report evidence of two partially admixed subpopulations that are only loosely associated
 56 with geographic origin. Using the multiple sequentially Markov coalescent model, we inferred the
 57 demographic history of the two pathogen subpopulations over the last 0.5 million years. We show that
 58 both populations experienced a recent strong bottleneck starting around 10,000 years ago, coinciding
 59 with the assumed time of maize domestication. While the genome average genetic diversity is low
 60 compared to other fungal pathogens, we estimated that the rate of non-synonymous adaptive substitutions
 61 is three times higher in genes located within virulence clusters compared to non-clustered genes,
 62 including non-clustered effector genes. These results highlight the role that these singular genomic
 63 regions play in the evolution of this pathogen.

64

65

66 Key words

67 *Ustilago maydis*, population genomics, demographic history, adaptive mutation rate, virulence clusters,
 68 sequentially Markov coalescent

69 **Significance statement**

70 The maize pathogen *Ustilago maydis* is a model species to study fungal cell biology and biotrophic host-
 71 pathogen interactions. Population genetic studies of this species, however, were so far restricted to using
 72 a few molecular markers, and genome-wide comparisons involved species that diverged more than 20
 73 million years ago. Here, we sequenced the genomes of 22 Mexican *U. maydis* isolates to study the recent
 74 evolutionary history of this species. We identified two co-existing populations that went through a recent
 75 bottleneck and whose divergence date overlaps with the time of maize domestication. Contrasting the
 76 patterns of genetic diversity in different categories of genes, we further showed that effector genes in
 77 virulence clusters display a high rate of adaptive mutations, highlighting the importance of these effector
 78 arrangements for the adaptation of *U. maydis* to its host.

81 **Introduction**

82 The coevolution between plant pathogens and their hosts impacts the genetic diversity of the interacting
 83 species. The response to these reciprocal selective forces depends on multiple factors, including the
 84 genome architecture of the organisms (e.g. genome size and karyotype structure), their life history traits
 85 (e.g. importance and frequency of sexual reproduction), but also stochastic factors in relation to
 86 demography (e.g. variable population size and population structure). Consequently, plant domestication
 87 and the subsequent emergence of agriculture, which typically results in a population bottleneck and a
 88 strong directional selection, had a strong impact on the selected organisms (Tang *et al.*, 2010; Milla *et*
 89 *al.*, 2015). It also affected the evolution of the associated pathogens, because domestication resulted in
 90 significant losses of genetic variation and strong selection on a few genes (Glemin and Bataillon, 2009).
 91 For example, speciation of the rice blast pathogen *Magnaporthe oryzae*, the wheat pathogen
 92 *Zymoseptoria tritici* and the barley pathogen *Rhynchosporium secalis* from their wild relatives was

93 associated with the domestication of their host plants (Couch *et al.*, 2005; Stukenbrock *et al.*, 2007;
 94 Zaffarano *et al.*, 2008). Here, we investigate the evolutionary history of the maize pathogen *Ustilago*
 95 *maydis*, a basidiomycete from the group of smut fungi (family: Ustilaginaceae). This family comprises
 96 about 550 described species (Begerow *et al.* 2014), among which are pathogenic species of grasses,
 97 including crops like maize, sorghum, wheat, barley, and sugarcane (Agrios, 2005). The genomes of
 98 several crop pathogens have been sequenced (Kämper *et al.*, 2006; Schirawski *et al.*, 2010; Laurie *et al.*,
 99 2012; Que *et al.*, 2014; Taniguti *et al.*, 2015; Dutheil *et al.*, 2016; Benevenuto *et al.*, 2018) as well as
 100 some species parasitizing wild grasses or dicot plants (Sharma *et al.*, 2014; Rabe *et al.*, 2016; Ye *et al.*,
 101 2017). Dating of speciation events between these species suggested that their divergence predates the
 102 domestication of their hosts and therefore occurred in their wild ancestors (Munkacsı *et al.*, 2007;
 103 Schweizer *et al.*, 2018). Among these species, *U. maydis* is the best studied and serves as a model for
 104 elucidating the molecular basis of biotrophic host-pathogen interactions (Matei and Doehlemann, 2016;
 105 Lanver *et al.*, 2017). These studies showed that the interaction with the host plant maize is largely
 106 controlled by secreted effector proteins of which about half lack known functional domains (Lanver *et*
 107 *al.*, 2017). Genome comparisons of *U. maydis* and related species revealed that many effector genes
 108 reside in gene clusters in the genome, an arrangement that is so far not described in other filamentous
 109 plant pathogens. Despite the high divergence level of these clustered effectors, homology between
 110 clusters of distinct species was established due to conserved synteny between genomes (Schirawski *et*
 111 *al.*, 2010; Dutheil *et al.*, 2016). Functional analyses of such clusters showed that they contain important
 112 virulence determinants in the barley pathogen *Ustilago hordei* (Ali *et al.*, 2014), in *U. maydis* (Kämper
 113 *et al.*, 2006; Schirawski *et al.*, 2010; Brefort *et al.*, 2014; Navarrete *et al.*, 2019), and in the maize
 114 pathogen *Sporisorium reilianum* (Ghareeb *et al.*, 2018).

115 It is hypothesized that the center of origin of *U. maydis* lies in Mexico from where it spread
 116 following the domestication of maize from teosinte (Munkacsı *et al.*, 2008), starting 6,000 to 10,000

117 years ago (Matsuoka *et al.*, 2002; Hake and Ross-Ibarra, 2015). Investigating infected maize fields
118 demonstrated that *U. maydis* shows a single generation and limited spreading between host plants in one
119 growing season (Baumgarten *et al.*, 2007). Moreover, an analysis of amplified fragment length
120 polymorphism markers of isolates sampled in the USA and Uruguay showed that *U. maydis* reproduces
121 predominantly by out-crossing, and this finding was independent of differences in agricultural practice
122 at the sampling sites (Barnes *et al.*, 2004).

123 Munkacsi *et al.* (2008) investigated the impact of maize domestication on the evolution of *U.*
124 *maydis*, using ten microsatellite markers. Samples from different locations revealed that subpopulations
125 in Mexico diverged within a time window that is consistent with the domestication and cultivation of
126 maize in the Americas. Moreover, genetic diversity of *U. maydis* was not found to be greater in Mexico
127 (the presumed origin of the species) than in other parts of the Americas, suggesting that the domestication
128 of maize from teosinte imposed a bottleneck that reduced the ancient genetic diversity in *U. maydis*
129 (Munkacsi *et al.*, 2008). Furthermore, analyses of sequence polymorphisms in 18 *U. maydis* isolates
130 originating from 11 locations in Europe, North America, and South America with a focus on the virulence
131 clusters 2A and 19A as well as the single effector *pep1* demonstrated low genetic variation in these
132 regions and uncovered three subpopulations based on geographic origin (Kellner *et al.*, 2014). While
133 these studies of individual genomic loci highlighted the effect of domestication on the evolutionary
134 history of *U. maydis*, they did not allow the detailed inference of the demographic history and genome-
135 wide patterns of selection in this species. Such studies require the availability of full genome sequences
136 (Stukenbrock *et al.*, 2011; Grünwald *et al.*, 2016).

137 To extend our understanding of the evolutionary history of *U. maydis*, we employed a
138 population genomics approach and sequenced 22 isolates originating from five different regions in
139 Mexico (Valverde *et al.*, 2000). We used this data set to investigate the population structure and the
140 demographic history of the sampled isolates. We further assessed patterns of genome-wide nucleotide

diversity and inferred the rate of adaptive substitutions in distinct categories of genes, allowing us to highlight the unique role of virulence clusters in the adaptive evolution of this fungal plant pathogen.

Materials and Methods

Origin, genomic DNA extraction, and sequencing of haploid U. maydis isolates

We sequenced the genome of 22 Mexican *U. maydis* isolates that are part of an earlier described isolate collection (Valverde *et al.*, 2000; supplementary table S1). The isolates were stored as haploid sporidia in glycerol stocks at -80°C. Isolates were thawed by plating them on Potato-Dextrose (PD) plates (3.9% [w/v] Potato-Dextrose Agar, 1% [v/v] Tris-HCl [1 M, pH 8.0]) and incubating them for two days at 28°C. Next, fungal cells were scratched off the PD plates and ground together with glass beads in liquid nitrogen. Genomic DNA was extracted by adding 500 µl TE-phenol/chloroform (1:1) and 500 µl lysis buffer (100 mM NaCl, 10 mM Tris-HCl (pH 8.0), 1 mM EDTA, 2 % Triton X-100, and 1 % SDS) followed by precipitation in 70% Ethanol. RNA was removed from the samples with the Master Pure Complete DNA & RNA Purification Kit (Biozym Scientific, Hessisch Oldendorf, Germany). DNA concentration was adjusted to about 150 to 400 ng/µl and about 1 µg of DNA was used for sequencing. After fragmentation of the genomic DNA, sequencing libraries were prepared using the TruSeq DNA LT Kit (Illumina, San Diego, USA) and sequenced at the Max Planck Genome Centre (Cologne, Germany) using the HiSeq sequencing kit on a HiSeq2000 cyciler (Illumina). Paired-end sequencing was performed with a 100 base pair read length and for each sequencing library, at least 21.6 million reads were generated, corresponding to a 100-fold average coverage (supplementary table S1). All Illumina paired-end reads were deposited at NCBI (BioProject ID: PRJNA561077).

165 *De novo genome assemblies*

166 A *de novo* assembly of the haploid genome was generated individually for each Mexican isolate with
 167 SOAPdenovo2 (Luo *et al.*, 2012) as follows: all odd kmer lengths ranging from 51 to 83 were tested and
 168 the kmer length yielding the highest N₅₀ was selected for each library. The estimated genome size (option
 169 -z) was set in all cases to 20,000,000 base pairs, the genome size of the reference genome for *U. maydis*
 170 isolate 521 (Kämper *et al.*, 2006). Resulting genome assembly statistics, including N₅₀ contig lengths for
 171 each kmer and isolate are summarized in supplementary table S2. Next, SOAPdenovo2 was used to get
 172 a genome for each isolate with the determined optimal kmer length. A sparse pregraph was built, and
 173 contigs were then computed, mapped and assembled into scaffolds. Finally, the GapCloser program was
 174 used to close remaining assembly gaps. The assembled genome sequences were deposited at NCBI
 175 (BioProject ID: PRJNA561077).

176

177 *Multiple genome alignment*

178 We generated a multiple genome alignment that comprised the *de novo* assembled genomes of the 22
 179 Mexican *U. maydis* isolates together with the *U. maydis* reference genome of the isolate 521 (Kämper *et*
 180 *al.*, 2006). The reference genome sequence was obtained from Mycocosm of the Joint Genome Institute
 181 (Grigoriev *et al.*, 2014) in version 2_2 from December 5, 2017. The 23 genomes served as input for the
 182 Multiz genome aligner from the Threaded Blockset Aligner package (Blanchette *et al.*, 2004). The
 183 resulting alignment was then projected on the reference genome, yielding an alignment length of
 184 20,028,090 base pairs. This alignment was then processed using MafFilter (Dutheil *et al.*, 2014). First,
 185 all synteny blocks were realigned using Mafft (Katoh and Standley, 2013), with blocks of a length greater
 186 than 10 kb being first split before alignment for computational efficiency. This unfiltered alignment was
 187 then subjected to two pipelines. The first pipeline focused on protein coding genes and extracted all exons
 188 from the unfiltered alignment (see below, *Building gene families*). In the second pipeline, the realigned

189 synteny blocks were filtered to remove ambiguously aligned regions. This was achieved in two steps:
 190 first, only blocks that comprised sequences from all 23 isolates are kept and alignment blocks with
 191 multiple “paralogous” sequences per species were discarded. Second, alignment blocks were further
 192 processed with a sliding window approach. Within 10 bp windows slid by 1 nucleotide, short indels were
 193 identified and the window was discarded if it contained at least one indel shared by at least two isolates,
 194 or, alternatively, if the quantity of gap characters in the 23 species was higher than 100 gaps in the
 195 window. In these two steps, unresolved base positions were assigned as gaps. These filtering steps yielded
 196 a final alignment with a length of 19,224,664 bp in 2,676 blocks. Alignment lengths and number of blocks
 197 resulting from each filtering step are summarized in supplementary table S3. We generated a second
 198 alignment using the same protocol, including this time all 23 *U. maydis* sequences together with the
 199 genome sequence of *S. reilianum* SRZ2 (version 2) which we obtained from the PEDANT database
 200 (Walter *et al.*, 2009). The corresponding alignment statistics for each filtering step are provided in
 201 supplementary table S3. Pairwise similarity distances were computed using MafFilter and a global tree
 202 was constructed using the FastME software (Lefort *et al.*, 2015), with default nucleotide model, nearest
 203 neighbor interchange (NNI) and subtree pruning regrafting (SPR) topology optimization. One thousand
 204 bootstrap replicates were performed in order to assess the support of each clade.

205

206 *Analyses of population structure*

207 Single nucleotides polymorphisms (SNPs) were called from the filtered alignment using MafFilter and
 208 exported to a file in the Variant Call Format (VCF). The set of SNPs was thinned according to linkage
 209 disequilibrium using the bcftools (Li, 2011), and only pairs with $r^2 < 0.6$ in 250 kb windows were kept.
 210 The resulting “unlinked” SNP set was exported to a file in PLINK format using plink 1.9 (Purcell *et al.*,
 211 2007). The smartPCA software (Patterson *et al.*, 2006) was used to compute principal components from
 212 the unlinked SNP set, and results were plotted with the R statistical environment (Ortutay and Ortutay,

2017). Model-based inference of population structure was conducted using the ADMIXTURE software (Alexander *et al.*, 2009) on SNPs filtered as for the PCA analysis. We performed a cross-validation analysis for a range of models with one to six genetic components. Each model was rerun from ten random initial conditions and results were summarized using the PONG software (Behr *et al.*, 2016).

Analyses of nucleotide diversity

The mean number of nucleotide difference between all pairs of sequences (π), GC content and fixation index F_{ST} were computed in non-overlapping windows of 10 kb from the multiple genome alignment of *U. maydis* isolates. The divergence between *U. maydis* and *S. reilianum* reference genomes was computed from the multiple genome alignment with outgroup in non-overlapping windows of 10 kb. A Tamura 92 model (Tamura, 1992) was fitted independently in each window in order to account for multiple substitutions while accounting for variable ratios of transitions over transversions, as well as non-homogeneous GC content. F_{ST} values were calculated using Hudson's 1992 estimator (Hudson *et al.*, 1992). All calculations were performed using the MafFilter program (Dutheil *et al.*, 2014). The distribution of F_{ST} values showed a tail of extreme F_{ST} values (supplementary figure S1) and was best fitted with a mixture of normal distributions using the 'fitdistr' function from the MASS package (Venables and Ripley, 2002) for R (<https://www.R-project.org>). To assess the significance of high F_{ST} values, we computed the probability that the F_{ST} value belonged to the lower mode of the distribution, using the estimated parameters of the two normal distributions. The 191 regions for which this probability was lower than 1% were considered as high F_{ST} regions and were subsequently scanned for genes, resulting in 751 candidate genes.

237 *Detection of Gene Ontology term enrichments*

238 All annotated *U. maydis* proteins were used as input for an Interpro search with version 5.35-74.0-64
 239 (Mitchell *et al.*, 2019), and mapped Interpro domains for each protein are listed in supplementary table
 240 S4. Next, the Interpro domains were linked to Gene Ontology (GO) Terms with the file “interpro2go”,
 241 which is provided by the GO consortium (version 2019/05/02 15:27:19;
 242 <http://current.geneontology.org/ontology/external2go/interpro2go>). In this way, 1,948 unique GO terms
 243 could be assigned to 4,147 *U. maydis* proteins (supplementary table S4). Each GO Term was associated
 244 with one of the three major subontologies “Cellular Component,” “Biological Process,” or “Molecular
 245 Function” with the Bioconductor package topGO (Alexa *et al.*, 2006). Enriched GO Terms were then
 246 identified by computing *P*-values for each GO term using Fisher’s classic test with parent-child
 247 correction (Grossmann *et al.*, 2007). For this analysis, genes with a high F_{ST} value were compared to all
 248 genes for which F_{ST} values could be computed, and results were considered to be significant at the 5%
 249 level.

250

251 *Inference of demography using the multiple sequentially Markov coalescent*

252 We used MSMC2, a re-implementation of the multiple sequentially Markov coalescent (Malaspinas *et al.*
 253 *et al.*, 2016) to estimate the time variation of coalescence rates. The filtered alignment was converted to
 254 MSMC input format using the MafFilter program (Dutheil *et al.*, 2014). MSMC2 was then run with
 255 default options. In order to convert time estimates from coalescent units to years and coalescence rates
 256 into effective population sizes, measures of generation time and average mutation rate are needed.
 257 Munkacsi and colleagues provided estimates of the synonymous mutation rate in smut fungi using six
 258 species comparisons, in four different genes, leading to 24 estimates $\{x_i\}$ of the mutation rate (table 4 in
 259 Munkacsi *et al.*, 2007). These estimates are exponentially distributed, and we therefore computed a
 260 genome geometric mean (u) using the formula $u = \exp(\sum_i \log(x_i)/24)$, leading to a value of $u = 5.23 \times 10^{-9}$

mutations per site per generation. We further considered a generation time of one year (Munkacsı *et al.*, 2008). Cross-coalescence rate analyses were conducted following the protocol described in the documentation of the MSMC2 program. In order to assess the significance of the pattern of cross-coalescence rate between two samples, we randomly permuted genomes in samples and recomputed the rates of cross-coalescence. A total of 10 permutations was conducted.

Building gene families

We used the multiple genome alignment of *U. maydis* isolates to extract nucleotide sequences of protein coding genes according to the annotation of the *U. maydis* reference genome, obtained from MycoCosm of the Joint Genome Institute (Grigoriev *et al.*, 2014). This annotation encompassed 6,785 protein coding genes, of which we discarded 36 genes with splice variants. We extracted nucleotide sequences for 6,742 genes (supplementary table S4). An outgroup sequence from the related species *Sporisorium reilianum* f. sp. *zeae* was further added using the procedure described below. We obtained the proteome of *S. reilianum* (Schirawski *et al.*, 2010) from the protein data base PEDANT (Walter *et al.*, 2009) with 6,676 proteins. The proteome of the *U. maydis* isolate 521 was searched against the *S. reilianum* proteome using blastp, and the result was used as input for the SiLiX algorithm in order to reconstruct gene families (Miele *et al.*, 2011). This software infers homologous relationships based on two criteria: the percent identity between two sequences and the coverage, defined as the relative length of a hit compared to the total length of the two sequences. We used a range for coverage and identity thresholds between 5% and 95% in 5% steps to identify values that result in the largest number of families with one-to-one homologs. The thresholds of 45 % identity and 55 % coverage were selected, because they lead to the maximum number (5,685) of families comprising one gene each in *U. maydis* and *S. reilianum* (supplementary table S5). This allowed us to map a single *S. reilianum* ortholog to 5,678 genes that were extracted from the multiple genome alignment. Several proteins in virulence clusters, however, could not be assigned to

285 families of 1:1 orthologs predicted by SiLiX, because such genes evolved by duplication (Dutheil *et al.*,
 286 2016). In order to identify an outgroup sequence for these genes, we conducted a second blastp search
 287 with, as query, all *U. maydis* genes that could be extracted from the multiple genome alignment but had
 288 not been mapped to a *S. reilianum* ortholog. We used all *S. reilianum* proteins that are not mapped to a
 289 *U. maydis* ortholog by SiLiX as target. We considered only hits with an E-value $< 10^{-6}$ for further analyses
 290 and found 404 cases where one *U. maydis* gene mapped to one *S. reilianum* gene and vice versa. In
 291 summary, we could assign a *S. reilianum* outgroup sequence to 6,082 *U. maydis* genes out of 6,742 genes
 292 that could be extracted from the multiple genome alignment (supplementary table S4). This set of gene
 293 families was aligned at the codon level using MACSE (Ranwez *et al.*, 2011), with the exception of one
 294 gene (*UMAG_10543*) for which the program failed to output an alignment. Two genes (*UMAG_00001*
 295 and *UMAG_10807*) were additionally discarded as their annotation changed since the previous version
 296 of the genome used in Dutheil *et al.* (2016) and they were, therefore, not included in the prediction of
 297 effector clusters. 6,051 genes were detected in all 22 Mexican isolates and were selected for subsequent
 298 analyses. Of these genes, 56 have at least one predicted in-frame stop codon before the end of the coding
 299 sequence in the reference genome and were discarded, as they could correspond to polymorphic open
 300 reading frames or contain sequencing errors. The final data set contained a total of 5,993 genes.
 301 Annotations of candidate effector genes were taken from the “strict” prediction described in Dutheil *et*
 302 *al.* (2016), which contained 553 genes predicted to encode secreted effector proteins. In the same study,
 303 156 effector genes were found in gene clusters. The filtered dataset studied here contained 95 genes in
 304 virulence clusters, 344 non-clustered effectors, and 5,554 other non-effectors, non-clustered genes.

305

306 *Reconstruction of site frequency spectra*

307 Filtered alignments were analyzed with the bppPopStat program from the bppSuite software (Guéguen
 308 *et al.*, 2013) in order to compute the unfolded synonymous and non-synonymous site frequency spectra

309 (SFS) for each gene of the 22 Mexican isolates. Ancestral alleles were inferred using a marginal
310 reconstruction after fitting a codon model (Yang and Nielsen's model with F3X4 frequencies; Yang and
311 Nielsen, 1998), including the outgroup sequence. When computing the site frequency spectra, positions
312 with more than two alleles were ignored, as well as positions where the outgroup displayed an allele
313 distinct from the set of alleles in the ingroup.

314

315 *Estimation of the distribution of fitness effects (DFE) of mutations and rate of adaptive substitutions*

316 Gene-specific SFS for each gene category (gene in virulence cluster, non-clustered effectors, or non-
317 clustered, non-effector genes) set were obtained by pooling the SFS for each constitutive gene to generate
318 the input file for the Grapes program (Galtier, 2016). A model selection procedure was conducted for
319 each set separately, using the '-m all' option of Grapes. We further used the 'no_div_param' option to use
320 divergence predictions from the polymorphism data instead of using the sequence of the outgroup to
321 estimate the corresponding parameters, relaxing the assumption that the DFE was constant since the
322 divergence with the outgroup. The best model according to the Akaike's information criterion (AIC) was
323 found to be the scaled-beta distribution for the non-clustered effector genes, while the gamma-
324 exponential distribution and gamma-gamma distributions best fitted clustered genes and non-effector
325 non-clustered genes, respectively (see supplementary table S6). In all categories of genes, however, these
326 three models had very similar AIC values, and we performed a model averaging procedure by weighting
327 each model with its relative likelihood (Dormann *et al.*, 2018). In order to assess the sampling variance
328 of the inferred DFE and rate of adaptive substitutions for each set of genes, a bootstrap procedure was
329 conducted by re-sampling genes in each category 100 times. Parameters of the DFE as well as rates of
330 adaptive substitutions were estimated for each bootstrap replicate, using the model with the best fit for
331 each gene category, and their distribution used to compute confidence intervals. To assess the significance
332 of observed differences in parameters between gene categories, we performed a permutation test, re-

sampling all genes between the three categories a 1,000 times and rerunning Grapes with the selected model on each sample. For each pair of gene sets, P -values were then computed using the formula $(\sum_i [S_i \geq S_{obs}] + 1) / 1001$ where S_i denotes the difference in parameter estimates for replicate i and $\sum_i [S_i \geq S_{obs}]$ is the number of replicates for which the difference is greater or equal to the observed difference. The following comparisons were performed: clustered genes *vs.* non-clustered effectors, clustered genes *vs.* non-clustered non-effector genes, and non-clustered effectors *vs.* non-clustered non-effector genes. The parameters tested included non-synonymous diversity (π_N), synonymous diversity (π_S), and their ratio (π_N/π_S), ratio of non-synonymous to synonymous divergences ($d_N/d_S = \omega$), the rate of non-adaptive non-synonymous substitutions (ω_{NA}), the rate of adaptive non-synonymous substitutions (ω_A), and the proportion of adaptive non-synonymous substitutions (α). In order to control for protein length, a similar analysis was conducted after selecting a subset of non-clustered effector genes and non-clustered, non-effector genes with protein length similar to that of clustered genes. This was achieved by selecting, for each clustered gene, the corresponding gene with the most similar protein length in the non-clustered effector (resp. non-clustered, non-effector) gene sets. The generated sets have, therefore, the same number of genes as the clustered set, and their average length did not differ significantly (Kruskal-Wallis rank sum test, P -value = 1, supplementary file S1).

349

350 *Analysis of mating type loci*

Sequences from the *a* and *b* mating type loci were extracted from the genome alignment using annotations from the *U. maydis* reference genome. The corresponding genes are *UMAG_02382* (1 exon), *UMAG_02383* (4 exons), and *UMAG_02384* (3 exons) for the *a*-locus on chromosome 5, and *UMAG_12052* (2 exons) and *UMAG_00578* (2 exons) for the *b*-locus on chromosome 1. The complete region was extracted for each locus. Sequences of the two genes of the *b*-locus were combined with publicly available sequences (Kämper *et al.*, 2020), and phylogenetic trees were reconstructed for each

gene using the PhyML program (Guindon *et al.*, 2010), using a Le and Gascuel model of protein evolution (Le and Gascuel, 2008) with a 4-classes discrete Gamma distribution of rates. The “Best of NNI and SPR” topology search option was selected, and 100 non-parametric bootstraps were generated. Nodes with a bootstrap value below 60% were set as unresolved. Trees were plotted using the ‘ggtree’ package for R (Yu *et al.*, 2016).

362

363 *Verification of the a mating type locus of isolate A*

All reads were mapped on the assembly of the isolate A using bwa mem (Li and Durbin, 2009). Site-specific coverage and read heterogeneity were computed using the mpileup program from the samtools (Li *et al.*, 2009) and further processed by python and R scripts. We performed a second independent genomic DNA extraction for isolates A and B. Isolates were streaked out from a glycerol stock and were cultivated for two days on PD plates at 28°C. Next, a single colony was inoculated in YEPS-light liquid medium (1 % w/v yeast extract, 1 % w/v peptone, 1 % w/v sucrose) and grown over night at 28°C. Cells were pelleted and lysed with ca. 0.3 g glass beads, 500 µl TE-phenol/chloroform (1:1) and 500 µl lysis buffer for 15 minutes on a Vibrax shaker. The supernatant was precipitated with 70 % (v/v) ethanol and dissolved in TE-buffer (1 mM Na₂-EDTA, 10 mM Tris-HCl with pH 8.0) supplemented with RNaseA.

The Phusion high fidelity PCR master mix (New England BioLabs) was prepared with 0.2µM primers (forward sequence: CTAGCTACACCAGCGAGGACGATA; reverse sequence: TCATTCCTAGCTCTTCTTGCGTTGA) and 100 ng genomic DNA as template in a 20 µl reaction volume. The PCR was performed as follows: initial denaturation at 98°C for 1 minute; 30 cycles with 10 seconds denaturation at 98°C, 15 seconds annealing at 67°C, 5 minutes extension at 72°C and 7 minutes final extension at 72°C. Products from the PCR were analyzed by gel electrophoresis. The entire volume of the samples was mixed with 3 µl of bromophenol blue dye and loaded on 1% agarose gel prepared in 1X TAE buffer (40 mM Tris, 20 mM acetic acid, 1 mM EDTA with pH 8.0).

381 Results and Discussion

382 We sequenced 22 haploid isolates of *U. maydis* originating from five locations in Mexico (Valverde *et*
383 *al.*, 2000; figure 1A and supplementary table S1). We performed a *de novo* genome assembly for each
384 isolate and computed a multiple genome alignment including the previously sequenced reference isolate
385 521 (Kämper *et al.*, 2006). After filtering for alignment uncertainty, the length of the alignment totaled
386 19.2 Mb, covering 97.76 % of the *U. maydis* reference genome. A total of 61,745 SNPs was called from
387 the newly sequenced isolates, corresponding to a mean number of nucleotide differences of 9×10^{-4} per
388 nucleotide. This level of diversity is comparable to that found in populations of great apes or of
389 *Drosophila melanogaster* (Nam *et al.*, 2015; Haudry *et al.*, 2020), but remarkably low for a fungal
390 pathogen (Zheng *et al.*, 2013; McMullan *et al.*, 2018; Stukenbrock and Dutheil, 2018). Moreover, 6,742
391 genes out of the 6,785 annotated genes in the reference genome had a homologous sequence in all
392 Mexican isolates, and of these, 5,993 had both a coding sequence without predicted in-frame stop codon
393 and an identifiable homolog in *S. reilianum*. This high level of sequence conservation shows (1) that the
394 generated dataset is of high quality and encompasses a large proportion of the genome of the pathogen
395 and (2) that the genetic diversity of the population at the center of origin is very low.

396

397 *The investigated isolates represent two subpopulations*

398 To infer the level of similarity between the sequenced isolates, we constructed a global tree (figure 1B).
399 We found that all Mexican isolates (Valverde *et al.*, 2000) are more similar to each other than they are to
400 the genome of the reference isolate 521, which was collected from a corn field near St. Paul/Minnesota,
401 USA (Holliday, 1961). Within the Mexican isolates, isolates P, Q, and R were all collected from Sinaloa
402 and cluster in one group. Moreover, the isolates S, T, and U from Toluca clustered independently of all
403 other isolates. Isolates M and K appeared to be very similar, displaying less than one SNP per
404 chromosome on average after quality filtering, suggesting that the two individuals are very closely

related. Overall, we detected only a loose association between sampling origin and genome similarity, suggesting that population structure, if any, is not induced by geography. To further investigate the population structure, we performed a principal component analysis (PCA) (Patterson *et al.*, 2006). The results of the PCA were consistent with those of the global genome similarity tree (figure 2A and figure 2B): the first principal components distinguished three groups of isolates. The first set, consisting of isolates D, E, F, G, I, K, M, N, and O (referred below as the ‘DEFGIKMNO’ or ‘DEFGIKNO’ population, depending on whether the M isolate was included in the analysis, see below) formed a well-supported group in the similarity tree containing four of the five isolates from Pachuca, three of the five isolates from Oaxaca, and two of the five isolates from Irapuato. The second set, consisting of isolates A, B, C, P, Q, R, S, T, U, and V (referred below as the ‘ABCPQRSTUV’ population), grouped the four isolates from Toluca, the three isolates from Sinaloa, and three of the five isolates from Irapuato. The third group was found in between these two populations and consisted of only three isolates: isolates J and H from Oaxaca and isolate L from Pachuca. To corroborate this finding, we employed the ADMIXTURE program (Alexander *et al.*, 2009). Cross-validation favored a model with two subpopulations. Rerunning the model estimation procedure with distinct initial conditions showed that this model is also the most consistent, that is, 10 replicates out of 10 agreed on the population partitioning (supplementary figure S2). The two inferred subpopulations matched the grouping of the PCA analysis and the similarity tree, showing that the isolates H, L and J are a mixture of these two subpopulations with a ratio of roughly 70% to 30% (figure 2C). Our findings are in line with previous studies that also reported the presence of subpopulations of *U. maydis* in Mexico (Munkacsy *et al.*, 2008). Interestingly, the reference isolate 521 grouped together with the ABCPQRSTUV population, while showing some low degree of admixture with the DEFGIKMNO group. Deep sampling outside of Mexico is required to confirm the relationship of the reference isolate to the Mexican populations. The mechanisms of divergence of the two populations remain to be elucidated. Teosinte occurs in Mexico with two subspecies, *Zea mays* ssp. *parviglumis* and

429 *Zea mays* ssp. *mexicana* (Fukunaga *et al.*, 2005; Ross-Ibarra *et al.*, 2009). An interesting hypothesis is
430 that the two populations of *U. maydis*, while being able to infect maize, could represent *formae speciales*
431 primarily adapted to each of these two subspecies of teosinte.

432

433 *Both subpopulations experienced strong bottlenecks*

434 We elucidated the demographic history of the two subpopulations using a multiple sequentially
435 Markovian coalescent (MSMC2) approach (Malaspinas *et al.*, 2016), which we applied to the two
436 population components that we inferred from the population structure analysis. We found that the
437 ABCPQRSTUV population displayed a relatively constant effective population size between 1,000,000
438 and 7,000 years ago, but experienced a strong bottleneck ending about 500 years ago (figure 3A). The
439 second subpopulation DEFGIKMNO showed a similar trend, although experiencing first a population
440 increase before going through a stronger bottleneck starting 10,000 and ending 500 years ago. The time
441 frame of these bottlenecks coincides with the beginning of maize domestication 6,000 to 10,000 years
442 before present (Matsuoka *et al.*, 2002; Hake and Ross-Ibarra, 2015), indicating that the demography of
443 *U. maydis* was most likely affected by the domestication of its host plant, although to a different extent
444 in the two populations. To assess the timing of differentiation between the two populations, we conducted
445 a cross-coalescence analysis (figure 3B). This analysis showed that the differentiation of the two
446 populations occurred rapidly between 5,000 and 1,000 years before present. Permutations of samples
447 between the two populations confirmed that this pattern is very well supported by the data.

448 The MSMC approach infers variation of the coalescence rate in the past. Under a standard
449 coalescent model, the coalescence rate is inversely proportional to the effective population size; thus, it
450 provides a snapshot of the demographic history. However, assuming a standard coalescent may lead to
451 the inference of an artificial bottleneck when the sampled population is structured (Mazet *et al.*, 2016).
452 Furthermore, the presence of loci under purifying selection can result in the inference of a recent

expansion (Platt and Harris, 2020). The dataset presented here is potentially subject to these issues, given the presence of population structure and the high density of protein coding regions. While in line with the expectation that the domestication of maize impacted the evolution of the pathogen *U. maydis*, the inferred demographic scenarios should, therefore, be taken with necessary caution.

We searched for regions putatively involved in the divergence of the two populations (Wolf and Ellegren, 2017) by computing F_{ST} values in 10 kb windows along the genome (figure 4A), with the aim to highlight genes with a role in the adaptation of the pathogen during the domestication of the host. The genome-wide distribution of F_{ST} values had a mode around 0.2, and a skew toward high positive values. This distribution is well modeled by a mixture of two normal distributions (supplementary figure S1). One hundred ninety-one regions were found to be within the high- F_{ST} component with a 1% significance level, and contained 751 genes. These genes were not found to be enriched in candidate effector genes (Fisher's exact test, P -value = 0.86), and GO-term enrichment analysis of these genes only exhibited top-level metabolic categories (supplementary table S7). In summary, these results provide no support for the existence of highly differentiated loci that may drive the divergence of the two subpopulations. Further analyses, notably accounting for recombination rate variation, are required to identify loci possibly involved in population differentiation (Booker *et al.*, 2020).

The a mating type locus carries alleles that are conserved within the Mexican population but different from the reference isolate

U. maydis possesses a tetrapolar mating system where two loci, *a* and *b*, are involved (Bakkeren *et al.*, 2008). Our data set contains two idiomorphs of the *a*-locus: isolates B, D, J, K, L, M, N, O, Q, R, and U share the same *a1* idiomorph as the reference isolate, while the other isolates have another one, similar to *a2* (Bölker *et al.*, 1992; Urban *et al.*, 1996). Because distinct idiomorphs share no homology (Froeliger and Leong, 1991), the *a2* mating type regions could not be aligned with the reference genome and are

477 represented as gaps in our genome alignment (supplementary file 1). The *a1* and *a2* idiomorphs, however,
478 differ from the previously published sequences of *a1* in the 521 isolate (accession U37795) and *a2* in the
479 RK32 isolate (accession U37796; Bölker *et al.*, 1992; Urban *et al.*, 1996). All Mexican *a1* alleles are
480 100% identical to each-other, but display a few SNPs, and a (non-coding) deletion of thymine at position
481 5,556 with respect to the *a1* sequence of isolate 521 (supplementary file 1). The SNPs are distributed as
482 follows: one non-synonymous SNP is located at position eight in the *mfa1* gene (leading to asparagine
483 instead of threonine in the *a1* pheromone precursor protein). The *pra1* genes contain four SNPs in their
484 coding regions, of which two are non-synonymous (resulting in glycine instead of alanine at position 63
485 and leucine instead of proline at position 103 of the amino acid sequence). The Mexican *a2* alleles differ
486 from each other only at position 62 in the *rga2* gene, where isolates E, F, G and I encode a glycine instead
487 of a serine. The Mexican alleles all differ, however, from the published RK32 sequence of *a2* (accession
488 U37796). They feature a non-coding insertion of 20 nucleotides at position 1,222 and a three-nucleotide
489 insertion at position 1,300 of the *a2* sequence from isolate RK32 (supplementary file 1), more than 2.5
490 kb downstream the *pra2* gene. SNPs in coding regions are distributed as follows: three in gene *pra2*,
491 among which two non-synonymous at positions 218 (leading to alanine instead of arginine) and 294
492 (threonine instead of alanine), and three in *lga2*, among which one non-synonymous at position 63
493 (threonine instead of alanine). In addition to position 62, the *rga2* gene differs from the *a2* sequence
494 found in isolate RK32 by two other SNPs, including a non-synonymous one at positions 45 (serine
495 instead of isoleucine). The region between positions 1,981 to 3,735 of the U37796 sequence is notably
496 absent in all Mexican isolates (supplementary file 1). This region contains the pseudogenized *mfa* gene
497 copy in *a2* of RK32 reported in Urban *et al.*, 1995, which was hypothesized to be a remnant of a
498 multiallelic ancestor containing two pheromone genes. This result suggests that either the ancestor of the
499 Mexican isolates lost this remnant, or that the degenerated pseudogenized *mfa* gene was not ancestral but
500 inserted in the ancestor of the RK32 isolate originating from Germany.

501 We further observed that the assembly of the A isolate contained two complete *a1* and *a2* loci,
502 only separated by 13 nucleotides and identical to the ones found in other Mexican isolates. We note that
503 these 13 nucleotides include a N character, indicating that the two loci were on separate contigs and only
504 assembled at the scaffolding stage. In order to investigate the possibility of wrong assembly or
505 contamination, we performed a second independent DNA extraction and amplified the mating type region
506 in isolate A, using isolate B (containing the *a1* idiomorph) as a comparison. The result revealed the
507 presence of two amplified segments matching the sizes of the Mexican *a1* and *a2* idiomorphs, confirming
508 their presence in the isolate, while from isolate B only one segment of the expected size for *a1* could be
509 amplified (supplementary figure S3). This result is incompatible with the reconstructed assembly and
510 shows that the two loci are not consecutive in the genome. To investigate possible coverage variation,
511 we mapped all reads along the scaffold containing the *a* mating type locus in isolate A. We show that
512 each idiomorph is supported by a coverage lower than the average of the rest of the scaffold, roughly 2/3
513 for the *a1* idiomorph and 1/3 for the *a2* idiomorph (supplementary figure S4). Altogether, these results
514 suggest that the region containing the *a* mating type locus of the A isolate may have been diploid, and
515 that the A isolate might be aneuploid. The amplified fragment corresponding to the *a2* locus is less strong
516 than the amplified fragment representing the *a1* locus (supplementary figure 3). While this could be due
517 to the larger size of the *a2* segment impeding the amplification efficacy, we note that it mirrors the lower
518 coverage of the *a2* sequence in the genome sequencing (supplementary figure 4). An intriguing
519 possibility could be that the presumed aneuploid strain is unstable and lost one of the idiomorphs during
520 subsequent mitotic divisions. If the *a2* idiomorph is preferentially lost or if cells carrying the *a2*
521 idiomorph are dividing more slowly than cells carrying *a1* because of autocrine pheromone stimulation,
522 the final cell culture may contain unequal proportion of *a1* and *a2* sequences. Further genetic and
523 cytological investigations are needed to confirm this hypothesis.

524 In order to assess whether the presumed diploid status of the A isolate could impact our
525 population genomic analyses, we examined the read mapping to search for possible heterozygous
526 positions. After mapping all reads on the assembled genome, we counted the number of mismatches in
527 reads for each assembled position. If the genome was diploid, we would expect heterozygous positions
528 to have alternative states in the reads, in a proportion of 50% on average. We counted the proportions of
529 sites with an alternative state present in at least 20% of the reads: isolate A had 0.15% of such positions,
530 while isolate B had 0.21% and isolate C had 0.20% of such sites. Scaffold 79 of strain A, which contains
531 the *a* locus, only had 0.05% of sites with more than 20% of read supporting an alternative state. This
532 suggests that either the potential aneuploidy of isolate A is restricted to the genomic region containing
533 the *a* locus, or that the aneuploid region is highly homozygous. In either case, aneuploidy did not have a
534 significant impact on the genome sequencing of isolate A outside the mating type locus.

535

536 *Genetic diversity is higher at the b mating type loci*

537 The *b* mating type locus located on chromosome 1 represents a hotspot of diversity (figure 4C),
538 presumably because the underlying genes evolve under balancing selection (May *et al.*, 1999).
539 Reconstructing the phylogenetic relationships between all previously sequenced alleles (described in
540 Kämper *et al.*, 2020) and alleles that could be extracted from the multiple genome alignment showed that
541 the allele in isolate T is very similar to the *b2* allele. The *b7* allele is identified in isolates B and U. The
542 *b14* allele is detected in isolates G, I, K, M, N, and Q. The *b15* allele is present in isolates P and R. The
543 *b17* allele is found in isolates H, J, and L, and the *b18* allele is present in isolate V (supplementary figure
544 S5). Interestingly, two alleles, one from isolates A and C and one from isolates D, E, F, O, and S did not
545 cluster with any previously identified allele and may therefore be potentially novel. This finding needs
546 to be corroborated with *in vitro* mating assays and successful plant infections.

547

548 *The GC content of U. maydis correlates with chromosome size and is lower in virulence clusters*

549 We investigated several patterns of genetic diversity in windows of 10 kb along the genome of *U. maydis*.
 550 The GC content, divergence with *S. reilianum* and mean number of nucleotide differences within the two
 551 subpopulations all appeared to be homogeneous at the chromosome scale (figures 4B-D). We report a
 552 slightly significant negative correlation between the average GC content per chromosome and the
 553 chromosome length (Kendall's rank correlation test, $\tau = -0.296$, $P\text{-value} = 0.04984$). A negative
 554 correlation between chromosome size and average recombination rate is frequently observed in
 555 eukaryotes and usually explained by small chromosomes having a higher recombination rate (Kong *et al.*,
 556 2002; Jensen-Seaman *et al.*, 2004). A positive correlation between GC content and recombination
 557 rate could be the result of GC-biased gene conversion (Lesecque *et al.*, 2013), and this could drive the
 558 higher GC content in smaller chromosomes. Finally, we report that the GC-content is significantly lower
 559 in regions encompassing clusters of effectors (Wilcoxon's rank test, $P\text{-value} = 1.372 \times 10^{-14}$), in agreement
 560 with the previous report that virulence clusters are associated with AT-rich repeat elements (Dutheil *et al.*,
 561 2016). Below, we further discuss the link between repeat elements and virulence clusters.

562

563 *Genetic diversity is higher around clusters of effector genes*

564 In *U. maydis* and related species, several effector genes are organized in gene clusters, and many of them
 565 have a role in virulence (Kämper *et al.*, 2006; Schirawski *et al.*, 2010). We previously reported that, when
 566 comparing multiple related species, these gene clusters are enriched in genes evolving under positive
 567 selection (Schweizer *et al.*, 2018). Here, we find that the mean number of nucleotide differences was on
 568 average higher in windows overlapping with a virulence cluster: the median across 10 kb windows of the
 569 mean number of nucleotide differences is 0.0011 in virulence cluster regions, compared to 0.0008 for the
 570 rest of the genome in population ABCPQRSTUV (Wilcoxon test, $P\text{-value} = 5.131 \times 10^{-7}$), and 0.0007 vs.
 571 0.0004 in population DEFGIKMNO (Wilcoxon test, $P\text{-value} = 0.02154$), consistent with a previous

report based on a smaller set of genes (Kellner *et al.*, 2014). This higher diversity may result from a locally higher mutation rate and / or distinct selection regime. A higher mutation rate can be caused by the activity of transposable elements (TEs), directly or indirectly: active TEs may locally introduce mutations, the generated occurrence of repeated sequence may lead to increased slippage of the DNA polymerase, and mechanisms protecting the genome against TEs may be leaky and affect surrounding regions (Horns *et al.*, 2012; Laurie *et al.*, 2012; Dutheil *et al.*, 2016). In addition, selection may impact diversity along the genome (Charlesworth, 2009; Gossmann *et al.*, 2011; Grandaubert *et al.*, 2019). All these mechanisms may act simultaneously and are difficult to disentangle. Additional insights may be obtained by contrasting different classes of polymorphisms in protein coding regions.

581

582 *Clustered effector genes show a high rate of adaptive mutations*

To elucidate the selection regime under which different categories of genes evolve, we assigned each gene in *U. maydis* to one of the three categories “clustered genes” (genes in predicted virulence clusters), “non-clustered effectors” (genes predicted to encode an effector protein, but not located in a predicted cluster) or “remaining genome” as defined earlier (Dutheil *et al.*, 2016). We then computed the amount of synonymous and non-synonymous divergence, which we compared to the amount of synonymous and non-synonymous polymorphism for each category of genes. We find no significant difference of synonymous nucleotide diversity (π_s , figure 5A), non-synonymous diversity (π_n , figure 5B), and ratio of non-synonymous to synonymous diversity (π_n/π_s , figure 5C) between genes in the three categories. This contrasts with results obtained in other fungal pathogens such as *Z. tritici*, where genes encoding effector proteins are evolving under relaxed purifying selection evidenced by a higher π_n/π_s ratio (Grandaubert *et al.*, 2019). The ratio of non-synonymous to synonymous divergences ($d_n/d_s = \omega$), however, was found to be significantly higher in clustered genes than in non-clustered effectors, which have a significantly higher d_n/d_s than non-effector genes (figure 5D; P -value = 0.00599). In order to disentangle the adaptive

596 and non-adaptive part of the d_N/d_S ratio, we fitted models of distribution of fitness effects on polymorphic
597 data for genes in each category (Galtier, 2016), and estimated the rate of non-adaptive non-synonymous
598 substitutions (ω_{NA}), the rate of adaptive non-synonymous substitutions (ω_A), and the proportion of non-
599 synonymous substitutions that are adaptive (α). We find that genes located in clusters of effectors have a
600 lower rate of non-adaptive non-synonymous mutations (model averaged $\omega_{NA} = 0.06$, Figure 5E and
601 supplementary table S6) compared to non-clustered effector genes ($\omega_{NA} = 0.19$; P -value = 0.042) and
602 non-clustered non-effector genes ($\omega_{NA} = 0.12$). The rate of adaptive mutations was found to be
603 significantly higher in genes located in effector clusters ($\omega_A = 0.38$, Figure 5F), compared to non-
604 clustered effectors ($\omega_A = 0.11$; P -value = 0.00599) and non-clustered, non-effector genes ($\omega_A = 0.11$; P -
605 value = 0.00599). The proportion of adaptive non-synonymous substitutions ($\alpha = \omega_A / \omega$) was found to
606 differ marginally between genes in effector clusters ($\alpha = 86\%$, Figure 5G), non-clustered effector genes
607 ($\alpha = 46\%$; P -value = 0.0599) and non-clustered, non-effector genes ($\alpha = 48\%$; P -value = 0.0879). These
608 results are consistent with previous reports that genes encoding effector proteins undergo a higher rate of
609 adaptive evolution (Stukenbrock and Bataillon, 2012; Grandaubert *et al.*, 2019). Differences in
610 evolutionary rates can be explained by differences in the effective population size, N_e , or by the selection
611 coefficient, s , of the mutations. Consequently, two non-exclusive hypotheses for this observation can be
612 invoked. First, because some effector genes are essential for virulence, the fitness effect of mutations at
613 these genes might be larger. Under an arms race scenario, mutations creating new alleles that prevent
614 recognition by the host will have a large positive effect (large positive value of s , leading to a high value
615 of ω_A). When established, however, successful effector alleles are under strong negative selection so that
616 most mutations are highly deleterious (larger negative value of s , impeding the accumulation of non-
617 adaptive mutations and leading to a low value of ω_{NA}), until the plant target protein evolves a sequence
618 that prevents interaction with the effector. Examples for such conserved effectors include Pep1
619 (Doehlemann *et al.*, 2009) and Pit2 (Doehlemann *et al.*, 2011), which are inhibitors of the maize

620 peroxidase POX12 and a group of papain-like cysteine proteases, respectively (Hemetsberger *et al.*,
621 2012; Müller *et al.*, 2013; Misas Villamil *et al.*, 2019). Other examples of important effectors are ApB73
622 and Cce1, but their molecular function remains to be elucidated (Stirnberg and Djamei, 2016; Seitner *et al.*,
623 2018). Second, the pattern of selection may result from the genomic context of the underlying genes.
624 A higher recombination rate in the region of effector clusters is predicted to locally increase N_e by
625 reducing linkage disequilibrium and, therefore, to increase the efficacy of both negative and positive
626 selection (Charlesworth and Campos, 2014; Croll *et al.*, 2015; Grandaubert *et al.*, 2019). However, we
627 could not infer a recombination map based on the 22 Mexican *U. maydis* isolates, because the small
628 sample size and low genetic diversity prevented the application of linkage disequilibrium-based methods
629 (Baudat *et al.*, 2013). Moreover, the small chromosome and genome size as well as the high density of
630 protein coding genes (about 61.2 % of the genome) did not allow us to obtain reliable results with
631 approaches based on sequentially Markovian coalescent processes like iSMC (Barroso *et al.*, 2019).
632 Constructing a recombination map for *U. maydis*, therefore, would require the sampling of more isolates
633 from diverse geographic origins.

634 Effector genes are characterized by their relatively short size, and genes in virulence clusters
635 are, therefore, significantly shorter than non-clustered genes (317 amino-acids on average *vs.* 396 amino-
636 acids for non-clustered genes, Wilcoxon test, P -value $< 2.2 \times 10^{-16}$). Short genes have been shown to
637 evolve faster, as the encoded proteins have a higher proportion of exposed residues, which tend to
638 experience a higher frequency of adaptive substitutions (Moutinho *et al.*, 2019). In order to test whether
639 the shorter protein size of genes in virulence clusters explains their higher rate of adaptive evolution, we
640 created samples of genes with similar lengths to that of genes in virulence clusters (see Methods and
641 supplementary file S1). We then conducted the same analyses on the size-restricted data sets. The results,
642 including estimates of ω_A and ω_{NA} are very similar when controlling for protein length, suggesting that it
643 does not account for the observed differences (supplementary figure S6).

644 Lastly, we note that predicted effector genes that are not located in virulence clusters do not
 645 show an increased rate of adaptive evolution, contrasting with other pathogens where effector-encoding
 646 genes show an increased evolutionary rate (Stukenbrock *et al.*, 2011; Hacquard *et al.*, 2012; Huang *et*
 647 *al.*, 2014; Sperschneider *et al.*, 2014; Sharma *et al.*, 2014). The *U. maydis* genome is almost devoid of
 648 transposable elements (TEs), which are mostly concentrated in the region of virulence clusters. This
 649 distribution may result from TEs being locally tolerated, as they could be providing a supply of mutations
 650 in rapidly evolving regions. Another explanation is that negative selection against TEs is locally not
 651 strong enough, owing to recurrent positive selection on genes in virulence clusters, which reduces the
 652 local effective population size because of linkage. These two explanations are not exclusive and can both
 653 contribute positively to the local accumulation of TEs. Altogether, these results highlight the singular
 654 role that virulence clusters play in the adaptation of *U. maydis* to its host.

655

656

657 **Acknowledgements**

658 This work was generously supported by funds from the Max Planck Society. We thank Octavio Paredes-
 659 López for providing the Mexican isolates of *U. maydis*.

660

661

662 **Author contributions**

663 Conceptualization: GS (equal), JYD (equal), RK (equal); Formal Analysis: GS (equal), JYD (equal);
 664 Funding Acquisition: RK (lead); Investigation: GS (equal), JYD (equal), MBH (supporting), GVB
 665 (supporting), NR (supporting), KM (supporting), RK (supporting); Project Administration: GS (lead),
 666 JYD (supporting), RK (supporting); Supervision: JYD (lead), RK (supporting); Visualization: JYD
 667 (lead), GS (supporting); Writing – Original Draft Preparation: GS (lead), JYD (supporting), NR

(supporting), KM (supporting), MBH (supporting); Writing – Review and Editing: GS (equal), JYD (equal), RK (equal), GVB (supporting)

670

671

672 **Data availability**

673 Illumina paired-end reads and assembled genome sequences were deposited at NCBI (BioProject Id:
674 PRJNA561077). The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under
675 the accessions WEIZ000000000 to WEJU000000000. The versions described in this paper are versions
676 WEIZ010000000 to WEJU010000000. Multiple genome alignments, sequences of open reading frames
677 that could be extracted from the multiple genome alignment, and alignments of *U. maydis* gene sequences
678 with their outgroup in *S. reilianum* are available from supplementary file 1. Scripts necessary to
679 reproduce the analyses are available at <http://gitlab.gwdg.de/molsysevol/UmaydisPopGen> and in
680 supplementary file 1.

681

682

683 **Literature cited**

- 684 Agrios G. 2005. Plant Pathology. Elsevier/Academic Press, 5th Edition.
685
686 Alexa A, Rahnenführer J, Lengauer T. 2006. Improved scoring of functional groups from gene expression
687 data by decorrelating GO graph structure. *Bioinformatics* 22(13): 1600–1607.
688
689 Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated
690 individuals. *Genome Res* 19: 1655-1664.
691
692 Ali S et al. 2014. An immunity-triggering effector from the barley smut fungus *Ustilago hordei* resides
693 in an Ustilaginaceae-specific cluster bearing signs of transposable element-assisted evolution. *PLoS*
694 *Pathog* 10(7): e1004223.
695
696 Bakkeren G, Kämper J, Schirawski J. 2008. Sex in smut fungi: Structure, function and evolution of
697 mating-type complexes. *Fungal Genet Biol* 45: S15-S21.
698

699 Barnes CW, Szabo LJ, May G, Groth JV. 2004. Inbreeding levels of two *Ustilago maydis* populations.
700 *Mycologia* 96: 1236-1244.
701

702 Barroso GV, Puzović N, Dutheil JY. 2019. Inference of recombination maps from a single pair of
703 genomes and its application to ancient samples. *PLoS Genet* 15(11): e1008449.
704

705 Baudat F, Imai Y, de Massy B. Meiotic recombination in mammals: localization and regulation. *Nat Rev*
706 *Genet* 14: 794-806.
707

708 Baumgarten AM, Suresh J, May G, Phillips RL. 2007. Mapping QTLs contributing to *Ustilago maydis*
709 resistance in specific plant tissues of maize. *Theo Appl Genet* 114: 1229-1238.
710

711 Begerow D et al. 2014. Ustilaginomycotina. In: Esser K (ed). The Mycota - A comprehensive treatise on
712 fungi as experimental systems for basic and applied research. Systematics and evolution (Part A).
713 Springer, Heidelberg.
714

715 Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. PONG: fast analysis and visualization
716 of latent clusters in population genetic data. *Bioinformatics* 32: 2817-2823.
717

718 Benevenuto J, Teixeira-Silva NS, Kuramae EE, Croll D, Monteiro-Vitorello CB. 2018. Comparative
719 genomics of smut pathogens: insights from orphans and positively selected genes into host specialization.
720 *Front Microbiol* 9: article 660.
721

722 Blanchette M et al. 2004. Aligning multiple genomic sequences with the Threaded Blockset Aligner.
723 *Genome Res* 14: 708-715.
724

725 Bölker M, Urban M, Kahmann R. 1992. The a mating type locus of *U. maydis* specifies cell signaling
726 components. *Cell* 68: 441-450.
727

728 Brefort T et al. 2014. Characterization of the largest effector gene cluster of *Ustilago maydis*. *PLoS*
729 *Pathog* 10(7): e1003866.
730

731 Charlesworth B. 2009. Effective population size and patterns of molecular evolution and variation. *Nat*
732 *Rev Genet* 10: 195-205.
733

734 Charlesworth B, Campos JL. 2014. The relations between recombination rate and patterns of molecular
735 variation and evolution in *Drosophila*. *Annu Rev Genet* 48: 383-403.
736

737 Couch BC et al. 2005. Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in
738 crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. *Genetics*
739 170(2): 613-630.
740

741 Croll D, Lendenmann MH, Stewart E, McDonald BA. 2015. The impact of recombination hotspots on
742 genome evolution of a fungal plant pathogen. *Genetics* 201(3): 1213-1228.

743
744 Doehlemann G et al. 2009. Pep1, a secreted effector protein of *Ustilago maydis*, is required for successful
745 invasion of plant cells. *PLoS Pathog* 5(2): e1000290.
746
747 Doehlemann G, Reissmann S, Assmann D, Fleckstein M, Kahmann R. 2011. Two linked genes encoding
748 a secreted effector and a membrane protein are essential for *Ustilago maydis*-induced tumor formation.
749 *Mol Microbiol* 81(3): 751-766.
750
751 Dormann CF et al. 2018. Model averaging in ecology: a review of Bayesian, information-theoretic, and
752 tactical approaches for predictive inference. *Ecol Monogr* 88(4): 485-504.
753
754 Dutheil JY, Gaillard S, Stukenbrock EH. 2014. MafFilter: a highly flexible and extensible multiple
755 genome alignment files processor. *BMC Genomics* 15: 53.
756
757 Dutheil JY et al. 2016. A tale of genome compartmentalization: the evolution of virulence clusters in
758 smut fungi. *Genome Biol Evol* 83: 681-704.
759
760 Froeliger EH, Leong SA. 1991. The a mating-type alleles of *Ustilago maydis* are idiomorphs. *Gene* 100:
761 113-122.
762
763 Fukunaga K et al. 2005. Genetic diversity and population structure of teosinte. *Genetics* 169: 2241-2254.
764
765 Galtier N. 2016. Adaptive protein evolution in animals and the effective population size hypothesis. *PLoS*
766 *Genet* 12: e1005774.
767
768 Ghareeb H, Zhao Y, Schirawski J. 2018. *Sporisorium reilianum* possesses a pool of effector proteins that
769 modulate virulence on maize. *Mol Plant Pathol* 20(1): 124-136.
770
771 Glemin S, Bataillon T. 2009. A comparative view of the evolution of grasses under domestication. *New*
772 *Phytologist* 183: 273-290.
773
774 Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size
775 within a genome. *Genetics* 189(4): 1389-1402.
776
777 Grandaubert J, Dutheil JY, Stukenbrock EH. 2019. The genomic determinants of adaptive evolution in a
778 fungal pathogen. *Evolution Letters* 3: 299-312.
779
780 Grigoriev IV et al. 2014. MycoCosm portal: gearing up for 1000 fungal genomes. *Nucl Acid Res* 42:
781 D699-D704.
782
783 Grossmann S, Bauer S, Robinson PN, Vingron M. 2007. Improved detection of overrepresentation of
784 Gene-Ontology annotations with parent child analysis. *Bioinformatics* 2322: 3024-3031.
785

786 Grünwald NJ, McDonald BA, Milgroom MG. 2016. Population genomics of fungal and oomycete
787 pathogens. *Annu Rev Phytopathol* 54: 323-46.

788

789 Guéguen L et al. 2013. Bio++: efficient extensible libraries and tools for computational molecular
790 evolution. *Mol Biol Evol* 30: 1745-1750.

791

792 Guindon S et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies:
793 assessing the performance of PhyML 3.0. *Syst Biol* 59(3): 307-321.

794

795 Hacquard S et al. 2012. A comprehensive analysis of genes encoding small secreted proteins identifies
796 candidate effectors in *Melampsora larici-populina* (Poplar Leaf Rust). *Mol Plant Microbe In* 253: 279-
797 293.

798

799 Hake S, Ross-Ibarra J. 2015. The natural history of model organisms: genetic, evolutionary and plant
800 breeding insights from the domestication of maize. *eLife* 4: e05861.

801

802 Haudry A, Laurent S, Kapun M. 2020. Population genomics on the fly: Recent advances in *Drosophila*.
803 In: Dutheil JY, editor. Statistical Population Genomics. New York Springer. p. 357-396.

804

805 Hemetsberger C, Herrberger C, Zechmann B, Hillmer M, Doehlemann G. 2012. The *Ustilago maydis*
806 effector Pep1 suppresses plant immunity by inhibition of host peroxidase activity. *PLoS Pathog* 8(5):
807 e1002684.

808

809 Holliday R. 1961. The genetics of *Ustilago maydis*. *Genet Res* 2: 204-230.

810

811 Horns F, Petit E, Yockteng R, Hood ME. 2012. Patterns of repeat-induced point mutation in transposable
812 elements of basidiomycete fungi. *Genome Biol Evol* 4: 240-247.

813

814 Huang J, Si W, Deng Q, Li P, Yang S. 2014. Rapid evolution of avirulence genes in rice blast fungus
815 *Magnaporthe oryzae*. *BMC Genet* 15: 45.

816

817 Hudson RR, Slatkin M, Maddison WP. 1992. Estimation of levels of gene flow from DNA sequence data.
818 *Genetics* 132(2): 583-589.

819

820 Jensen-Seaman MI et al. 2004. Comparative recombination rates in the rat, mouse, and human genomes.
821 *Genome Res* 14: 528-538.

822

823 Kämper J et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*.
824 *Nature* 444: 97-101.

825

826 Kämper J, Friedrich MW, Kahmann R. 2020. Creating novel specificities in a fungal nonself recognition
827 system by single step homologous recombination events. *New Phytologist* 228: 1001-1010.

828

829 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements
830 in performance and usability. *Mol Biol Evol* 30: 772-780.

831

832 Kellner R, Hanschke C, Begerow D. 2014. Patterns of variation at *Ustilago maydis* virulence clusters 2A
833 and 19A largely reflect the demographic history of its populations. *PLoS One* 9: e98837.

834

835 Kong A et al. 2002. A high-resolution recombination map of the human genome. *Nat Genet* 31: 241-247.

836

837 Lanver D et al. 2017. *Ustilago maydis* effectors and their impact on virulence. *Nat Rev Microbiol* 15:
838 409-421.

839

840 Laurie JD et al. 2012. Genome comparison of barley and maize smut fungi reveals targeted loss of RNA
841 silencing components and species-specific presence of transposable elements. *Plant Cell* 245: 1733-
842 1745.

843

844 Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol* 25(7):
845 1307-1320.

846

847 Lefort V, Desper R, Gascuel O. 2015. FastME 2.0: A comprehensive, accurate, and fast distance-based
848 phylogeny inference program. *Mol Biol Evol* 32(10): 2798-2800.

849

850 Lesecque Y, Mouchiroud D, Duret L. 2013. GC-Biased gene conversion in yeast is specifically associated
851 with crossovers: Molecular mechanisms and evolutionary significance. *Mol Biol Evol* 30(6): 1409-1419.

852

853 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform.
854 *Bioinformatics* 25: 1754-1760.

855

856 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000
857 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and
858 SAMtools. *Bioinformatics* 25: 2078-2079.

859

860 Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and
861 population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987-2993.

862

863 Luo R et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo
864 assembler. *Gigascience* 1: 18.

865

866 Malaspinas AS et al. 2016. A genomic history of Aboriginal Australia. *Nature* 538: 207-214.

867

868 Matei A, Doehlemann G. 2016. Cell biology of corn smut disease - *Ustilago maydis* as a model for
869 biotrophic interactions. *Curr Opin Microbiol* 34: 60-66.

870

871 Matsuoka Y et al. 2002. A single domestication for maize shown by multilocus microsatellite genotyping.
872 *P Natl Acad Sci USA* 99 (9): 6080-6084.

873
874 May G, Shaw F, Badrane H, Vekemans X. 1999. The signature of balancing selection: Fungal mating
875 compatibility gene evolution. *P Natl Acad Sci USA* 96(16): 9172-9177.
876
877 Mazet O, Rodríguez W, Grusea S, Boitard S, Chikhi L. 2016. On the importance of being structured:
878 instantaneous coalescence rates and human evolution – lessons for ancestral population size inference?
879 *Heredity* 116: 362-371.
880
881 McMullan M et al. 2018. The ash dieback invasion of Europe was founded by two genetically divergent
882 individuals. *Nat Ecol Evol* 2: 1000-1008.
883
884 Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX.
885 *BMC Bioinformatics* 12: 116.
886
887 Milla R, Osborne CP, Turcotte MM, Violle C. 2015. Plant domestication through an ecological lens.
888 *Trends Ecol Evol* 30(8): 463-469.
889
890 Mitchell AL et al. 2019. InterPro in 2019: improving coverage, classification and access to protein
891 sequence annotations. *Nucleic Acids Res* 47(D1): D351–D360.
892
893 Misas Villamil J et al. 2019. A fungal substrate mimicking molecule suppresses plant immunity via an
894 inter-kingdom conserved motif. *Nat Commun* 10: article 1576.
895
896 Moutinho AF, Fontes Trancoso F, Dutheil JY. 2019. The impact of protein architecture on adaptive
897 evolution. *Mol Biol Evol* 36(9): 2013-2028.
898
899 Müller AN, Ziemann S, Treitschke S, Assmann D, Doeblemann G. 2013. Compatibility in the *Ustilago*
900 *maydis*–maize interaction requires inhibition of host cysteine proteases by the fungal effector Pit2. *PLoS*
901 *Pathog* 9(2): e1003177.
902
903 Munkacsı AB, Stoxen S, May G. 2007. Domestication of maize, sorghum, and sugarcane did not drive
904 the divergence of their smut pathogens. *Evolution* 612: 388-403.
905
906 Munkacsı AB, Stoxen S, May G. 2008. *Ustilago maydis* populations tracked maize through
907 domestication and cultivation in the Americas. *Proc Royal Soc B* 275: 1037-1046.
908
909 Nam K et al. 2015. Extreme selective sweeps independently targeted the X chromosomes of the great
910 apes. *P Natl Acad Sci USA* 112(20): 6413-6418.
911
912 Navarrete F et al. 2019. The Pleiades cluster of fungal effector genes inhibit host defenses. *bioRxiv* (doi:
913 10.1101/827600).
914
915 Ortutay C, Ortutay Z. 2017. Introduction to R statistical environment. In: Ortutay C, Ortutay Z. Molecular
916 data analysis using R. John Wiley & Sons, Inc., Hoboken, New Jersey USA.

917 Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* 2: e190.
918
919 Platt A, Harris DN. 2020. Phantom histories of misspecified pasts. *bioRxiv* (doi:
920 10.1101/2020.06.26.173963).
921
922 Purcell S et al. 2007. PLINK: a toolset for whole-genome association and population-based linkage
923 analysis. *Am J Hum Genet* 81: 559-575.
924
925 Que Y et al. 2014. Genome sequencing of *Sporisorium scitamineum* provides insights into the pathogenic
926 mechanisms of sugarcane smut. *BMC Genomics* 15: 996.
927
928 Rabe F et al. 2016. A complete toolset for the study of *Ustilago bromivora* and *Brachypodium* sp. as a
929 fungal-temperate grass pathosystem. *eLife* 5: e20522.
930
931 Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences
932 accounting for frameshifts and stop codons. *PLoS One* 6: e22594.
933
934 Ross-Ibarra J, Tenaillon M, Gaut BS. 2009. Historical divergence and gene flow in the genus *Zea*.
935 *Genetics* 181: 1399-1413.
936
937 Schirawski J et al. 2010. Pathogenicity determinants in smut fungi revealed by genome comparison.
938 *Science* 330: 1546-1548.
939
940 Schweizer G et al. 2018. Positively selected effector genes and their contribution to virulence in the smut
941 fungus *Sporisorium reilianum*. *Genome Biol Evol* 10(2): 629-645.
942
943 Seitner D, Uhse S, Gallei M, Djamei A. 2018. The core effector Cce1 is required for early infection of
944 maize by *Ustilago maydis*. *Mol Plant Pathol* 19(10): 2277-2287.
945
946 Sharma R, Mishra B, Runge F, Thines M. 2014. Gene loss rather than gene gain is associated with a host
947 jump from monocots to dicots in the smut fungus *Melanopsichium pennsylvanicum*. *Genome Biol Evol*
948 6(8): 2034-2049.
949
950 Sperschneider J et al. 2014. Diversifying selection in the wheat stem rust fungus acts predominantly on
951 pathogen-associated gene families and reveals candidate effectors. *Front Plant Sci* 5: article 372.
952
953 Stirnberg A, Djamei A. 2016. Characterization of ApB73, a virulence factor important for colonization
954 of *Zea mays* by the smut *Ustilago maydis*. *Mol Plant Pathol* 17(9):1467-1479.
955
956 Stukenbrock EH, Banke S, Javan-Nikkah M, McDonald BA. 2007. Origin and domestication of the
957 fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. *Mol Biol Evol* 24(2): 398-
958 411.
959

960 Stukenbrock EH et al. 2011. The making of a new pathogen: Insights from comparative population
961 genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species.
962 *Genome Research* 21: 2157-2166.

963
964 Stukenbrock EH, Bataillon T. 2012. A population genomics perspective on the emergence and adaptation
965 of new plant pathogens in agro-ecosystems. *PLoS Pathog* 8: e1002893.

966
967 Stukenbrock EH, Dutheil JY. 2018. Fine-scale recombination maps of fungal plant pathogens reveal
968 dynamic recombination landscapes and intragenic hotspots. *Genetics* 208(3): 1209-1229.

969
970 Tamura K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-
971 transversion and G+C-content biases. *Mol Biol Evol* 9(4): 678-687.

972
973 Tang H, Sezen U, Paterson AH. 2010. Domestication and plant genomes. *Curr Op Plant Biol* 13(2): 160-
974 166.

975
976 Taniguti LM et al. 2015. Complete genome sequence of *Sporisorium scitamineum* and biotrophic
977 interaction transcriptome with sugarcane. *PLoS One* 106: e0129318.

978
979 Urban M, Kahmann R, Bölker M. 1996. The biallelic a mating type locus of *Ustilago maydis*: remnants
980 of an additional pheromone gene indicate evolution from a multiallelic ancestor. *Mol Gen Genet* 250:
981 414-420.

982
983 Valverde MA, Vandemark GJ, Martinez O, Paredes-Lopez O. 2000. Genetic diversity of *Ustilago maydis*
984 isolates. *World J Microb Biot* 16: 49-55.

985
986 Venables WN, Ripley BD. 2002. Modern applied statistics with S. Fourth Edition. Springer, New York.

987
988 Walter MC et al. 2009. PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res* 37: D408-
989 D411.

990
991 Wolf JB, Ellegren H. 2017. Making sense of genomic islands of differentiation in light of speciation. *Nat*
992 *Rev Genet* 18: 87-100.

993
994 Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals.
995 *J Mol Evol* 46: 409-418.

996
997 Ye Z et al. 2017. Comparative whole-genome analysis reveals artificial selection effects on *Ustilago*
998 *esculenta* genome. *DNA Research* 24(6): 635-648.

999
1000 Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2016. ggtree : an R package for visualization and annotation
1001 of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 8(1): 28-36.

1002

1003 Zaffarano PL, McDonald BA, Linde CC. 2008. Rapid speciation following recent host shifts in the plant
1004 pathogenic fungus *Rhynchosporium*. *Evolution* 62(6): 1418-1436.

1005
1006 Zheng W et al. 2013. High genome heterozygosity and endemic genetic recombination in the wheat stripe
1007 rust fungus. *Nat Commun* 4: article 2673.

1008

1009

1010 **Figures**

1011 Figure 1: Genome similarity of isolates is only loosely associated with their geographic origin. (A)
1012 Mexican *U. maydis* isolates (named with letters from A to V) originated from the five regions Irapuato,
1013 Oaxaca, Pachuca, Sinaloa, and Toluca. The reference isolate 521 was collected in the USA. (B) Tree
1014 showing the genome-wide similarity levels between isolates. Isolates and branches are colored according
1015 to the sampling location as shown in A. Numbers indicate bootstrap support values as percentage from
1016 1,000 replicates, and branch lengths are the mean number of substitutions per site.

1017

1018 Figure 2: Inference of population structure reveals two subpopulations. (A) The relative inertia (vertical
1019 axis) are shown for the first 20 principal components (horizontal axis). (B) The first four principal
1020 components (PC) of a PC analysis based on genotypes are shown. The left panel shows the first and
1021 second PC, and the right panel depicts the third and fourth PC. Colors indicate the geographic origin of
1022 each isolate as shown in figure 1. (C) Population components as inferred by the ADMIXTURE model.
1023 Orange and purple colors represent the ancestry proportion of the two subpopulations (vertical axis) in
1024 each isolate (horizontal axis).

1025

1026 Figure 3: Both subpopulations experienced a recent bottleneck that started at a time overlapping with the
1027 supposed start of the domestication of maize, between 6,000 and 10,000 years ago (gray area) (A) The
1028 effective population size N_e (vertical axis) is shown over the last 1 million years (horizontal axis). The
1029 green line shows the result that is obtained when considering all isolates, whereas the purple, orange and

1030 yellow lines represent results for a subset of isolates as indicated. (B) A cross-coalescence analysis
1031 reveals the timing of the differentiation between the two subpopulations and shows that divergence
1032 happened rapidly within a time frame that coincides with the domestication of teosinte.

1033
1034 Figure 4: Values for different measures of genetic diversity vary at the fine (but not chromosomal) scale.
1035 The horizontal axis depicts the chromosome number, and the vertical axes show values of the fixation
1036 index (A), the nucleotide diversity in both subpopulations (B), the level of divergence between *U. maydis*
1037 and *S. reilianum* (C), and the GC content (D). Blue and gray colors of data points indicate different
1038 chromosomes. Orange lines indicate the localization of gene clusters as defined in Dutheil *et al.* (2016),
1039 and green lines show the localization of the *a* mating type locus on chromosome 5 and the *b* mating type
1040 locus on chromosome 1, respectively. Pop₁ includes isolates A, B, C, P, Q, R, S, T, U and V and Pop₂ the
1041 isolates D, E, F, G, I, K, M, N, and O.

1042
1043 Figure 5: Genes in virulence clusters display a higher rate of adaptive substitutions. *P*-values for pairwise
1044 comparisons between clusters of effector genes, unclustered effector genes and non-clustered, non-
1045 effector genes are computed as described in Methods. (A) Mean number of synonymous nucleotide
1046 differences (π_S), (B) mean number of non-synonymous nucleotide differences (π_N), (C) ratio of the mean
1047 number of non-synonymous nucleotide differences and mean number of synonymous nucleotide
1048 differences (π_N / π_S), (D) ratio of non-synonymous over synonymous divergence (d_N / d_S), (E) rate of non-
1049 adaptive non-synonymous substitutions (ω_{NA}), (F) rate of adaptive non-synonymous substitutions (ω_A),
1050 (G) proportion of non-synonymous adaptive substitutions (α). Box-and-whiskers plots indicate the
1051 median, first and third quantiles over 100 bootstrap replicates performed using the best model according
1052 to Akaike's information criterion. Grey diamonds indicate the corresponding value when averaging over
1053 all models of distribution of fitness effects (see Methods for details).

1054

1055

1056 **Supplementary figures**

1057 Supplementary figure S1: Distribution of F_{ST} values in 10 kb windows. The orange curve shows the fit
1058 of a two-normal mixture distribution.

1059

1060 Supplementary figure S2: Detailed results from the ADMIXTURE analysis, as represented by the PONG
1061 software. Each row represents the most likely population structure for a given number of subpopulations
1062 (K). Model fitting was repeated 10 times in each case with different initial random parameter values, and
1063 the most frequently inferred scenario was depicted. Ratios in blue indicate the number of replicates
1064 supporting the represented scenario in each case.

1065

1066 Supplementary figure S3: Amplification of the *a* mating type locus. (A) Expected sizes in the reference
1067 isolates for the two idiomorphs *a1* and *a2*, the Mexican isolates B and C (having the Mexican *a1* and *a2*
1068 idiomorphs) and isolate A, predicted to carry both idiomorphs. Expected sizes of the Mexican isolates
1069 are based on the genome assemblies. (B) Amplification of the corresponding locus in isolates A (lane 1)
1070 and B (lane 2). M, size marker.

1071

1072 Supplementary figure S4: Sequencing coverage along scaffold 79 of the A isolate assembly. The region
1073 containing the *a1* idiomorph is depicted in red, the region containing the *a2* locus is plotted in green, and
1074 the central region of the scaffold (N) in blue. Coverage is measured in number of reads mapped to each
1075 position. (A) Per nucleotide coverage. Straight lines represent the linear regression on the corresponding
1076 data points. (B) Coverage distribution per region.

1077

1078 Supplementary figure S5: Phylogenetic tree of the mating type region of the b-locus. Left tree: *bW* alleles
 1079 (*UMAG_00578*). Right tree: *bE* alleles (*UMAG_12052*). Nodes with a bootstrap value lower than 60%
 1080 have been unresolved and displayed as multifurcations. Colors of the isolate names distinguish the
 1081 ABCPQRSTUV and DEFGIKMNO subpopulations, as well as the three admixed individuals H, J, and
 1082 L. Alleles of the reference strain 521 are displayed in black, and all previously published alleles are in
 1083 grey boxes.

1084

1085 Supplementary figure S6: Comparison of genetic diversity and rate of adaptive substitutions in clusters
 1086 of effector genes, unclustered effectors and non-clustered, non-effector genes after controlling for protein
 1087 length. Legend as in Figure 5.

1088

1089

Supplementary tables

1091 Supplementary table S1: Overview of the 22 Mexican *U. maydis* isolates

1092 Supplementary table S2: Overview of the assembly results with different kmer lengths

1093 Supplementary table S3: Intermediate alignment lengths and number of blocks

1094 Supplementary table S4: List of all *U. maydis* genes with mapped Interpro domains, Gene Ontology
 1095 Terms, F_{ST} values, and outgroup sequences

1096 Supplementary table S5: SiLiX results for different identity and coverage settings

1097 Supplementary table S6: Estimates of the rate of adaptive substitutions for various models of distributions
 1098 of fitness effects.

1099 Supplementary table S7: Gene Ontology Term enrichment analysis for genes with high F_{ST} values

1100

1101

1102 **Supplementary files**

1103 Supplementary file S1: Data set and scripts necessary to reproduce the analyses in this work.

A **B**

A **B**