Population history of the Northern corn leaf blight fungal pathogen *Setosphaeria turcica* in Europe

Mireia Vidal-Villarejo[a,1], **Fabian Freund**[a,1], **Hendrik Hanekamp**[b], **Andreas von Tiedemann**[b], **and Karl Schmid**[a]

[a]University of Hohenheim, Stuttgart, Germany; [b]University of Göttingen, Göttingen, Germany

[1]M.V-V. and F.F. contributed equally to this work.

[2]To whom correspondence should be addressed. E-mail: karl.schmid@uni-hohenheim.de

FF, AvT and KS designed the study. AvT contributed materials. HH phenotyped the races. MV-V, FF, and KS analysed the data. MV-V, FF and KS wrote the manuscript.

We declare no conflict of interest.

Abstract

Setosphaeria turcica is a major fungal pathogen of maize and causes the foliar disease Northern corn leaf blight (NCLB). It originates from tropical regions and expanded into Central Europe since the 1980s, simultaneously with a rapid increase of maize cultivation area in this region. To investigate evolutionary processes influencing the rapid expansion of *S. turcica* we sequenced 121 isolates from Central Europe, Western Europe and Kenya. Population genetic inference revealed five genetically distinct clusters that differ by their geographic distribution and emergence dates. One genetically diverse cluster is restricted to Kenya, and the four European clusters consist of three distinct clonal lineages with low genetic diversity and one genetically diverse cluster with several clonal sublineages. A comparison of two different coalescent models for genetic diversity in the most frequent and geographically widespread clonal lineage in Europe supported a model of neutral, strongly exponential population growth over models accounting for different types of selection. In contrast to Kenyan isolates, European isolates did not show sexual recombination despite the presence of both mating types *MAT1-1* and *MAT1-2* in Europe. Within clonal lineages phenotypic variation in virulence to different monogenic resistances likely originated from repeated *de novo* mutations in virulence genes of *S. turcica*. *k*-mer based association mapping between genetic clusters did not identify genomic regions associated with pathogen races but few genomic regions that are significantly differentiated between two clonal lineages and contain putative effector genes. Our results suggest that the rapid colonization of Europe by different clonal lineages of *S. turcica* was not driven by selection of virulent races but reflects a neutral demographic process of fast pathogen population growth fostered by a rapid expansion of the maize cultivation area in this region.

Keywords: Demographic history | Population structure | Coalescent theory | *Setospheria turcia* | Maize

Introduction

Modern agricultural practice is characterized by reduced crop rotation, large field sizes of monocultures, high chemical inputs and cultivation of resistant varieties. These factors influence both short-term epidemics and a long-term evolution of resistant pathogen strains that may rapidly expand over large geographic areas (1). In addition, climate warming favors the spread and adaptation of pathogen species to new environments and geographic regions (2). These factors contribute to rapid crop-pathogen co-evolution, whose understanding is essential to improve

management practices and plant breeding to maintain food security in a rapidly changing world (3). Global pathogen monitoring systems for plant pathogens identify the origin and expansion of new pathogen strains (4) to support resistance breeding and adaptation of crop management practices. Disease monitoring is greatly facilitated by genome sequencing to characterize pathogen diversity (5) although a sequence-based prediction of virulence types remains challenging due to a rapid evolution of pathogen genomes (6–10). Sequencing data were used to track the epidemiology and demographic history of pathogens (e.g., 11–13) and to reconstruct introductions (14). However, the relative importance of demographic effects versus selection-driven adaptation to cultivation conditions or plant resistance genes is still little understood. Therefore, a characterization of demography and selection to evaluate the evolutionary potential of pathogen species (15) will contribute to developing evolution-informed, durable crop management strategies to avoid rapid breaking of host resistance genes and reduce chemical inputs in plant protection (16).

The hemibiotrophic fungal pathogen *Setosphaeria turcica* (Luttrell) Leonard and Suggs (teleomorph *Exserohilum turcicum*, formerly known as *Helminthosporium turcicum*) is the most important leaf pathogen of maize. It causes Northern corn leaf blight (NCLB), whose symptoms are long, elliptical stripes of necrotic tissues (lesions) on maize leaves, which limits the photosynthetic productivity and causes yield reduction (17). NCLB is a worldwide disease with a high incidence in the tropics, where it is a major cause of yield loss in maize. The most important methods for controlling the disease are breeding resistant varieties (18) and adapted management practices including fungicide applications. Additional management practices, such as biological control, are being studied (19, 20). *S. turcica* shows asexual and sexual reproduction, which requires mating of two strains with different *MAT1-1* and *MAT1-2* alleles at the *MAT1* mating type locus. Worldwide surveys of genetic diversity of *S. turcica* showed that sexual reproduction is restricted to regions with a warm climate (17). Genetic diversity was higher in populations from Mexico in comparison to Kenya, China and Europe suggesting that *S. turcica* originated in Mexico and recently arrived in Europe (21). NCLB was first reported in Italy in 1876, followed by South-Western France around 1900. Until the 1980s, NCLB was mainly restricted to the warmer regions of Southern Europe and the Balkans, but between 1988 and 1992 the disease crossed the Alps, and in 1995 it was reported in the Upper Rhine Valley in South Germany. Afterwards it rapidly expanded throughout the maize cultivation regions in Northwestern Europe. In response to the expansion of NCLB in Europe, breeders improved commercial varieties by selecting for polygenic, quantitative resistances and by introgression of monogenic, race-specific resistance genes from genetic resources. The four main resistance genes introgressed are *Ht1, Ht2, Ht3 and Htn1* (22). Different races of *S. turcica* are defined by their infection ability of a differentiation set of varieties harboring one of the four *Ht* genes. Race monitoring of more than 500 isolates revealed that *S. turcica* races are unequally distributed throughout Europe (23). Such a distribution raises the question whether the rapid expansion reflects a neutral demographic process like a repeated and independent introduction of different strains that were rapidly distributed by seed trade and agricultural practices, or a selection-driven adaptation to resistant host varieties that favored the rapid expansion of novel, virulent pathogen strains throughout Europe.

We investigated both hypotheses by characterizing the genomic diversity of *S. turcica* isolates collected from natural infections of different susceptible maize varieties (lacking any known *Ht* genes) that were cultivated throughout Europe in 2011 and 2012. Using phylogenetic analyses and coalescence models we identify different clonal lineages

throughout Central and Western Europe that are distinct from Kenyan isolates used for comparison. The overall genetic diversity of the most widespread European clonal lineage was not shaped by strong selection exerted by host resistance genes, but reflects a neutral, exponential growth.

Results

Read mapping and variant discovery. We sequenced a sample of 166 isolates (157.2 GB raw sequence) from 11 different countries (Dataset S1) and subsequently removed 37 isolates because of low coverage or a high proportion of reads not mapping to the reference genome. Eight samples were technical replicates of the same isolate to estimate the sequencing error rate. After excluding low quality samples and replicates we obtained a final sample of 121 isolates with an average read coverage of 14.9x and a range from 5.5x to 44.7x coverage. After mapping and calling SNPs with both GATK and samtools-bcftools, we identified 55,534 SNPs by both methods and retained 23,209 SNPs after filtering (Materials and Methods). SNPs with a maximum of 35% of missing data were imputed by multiple correspondence analysis (MCA) (24). The median number of SNPs differing between the eight technical replicates was 9.5 corresponding to 99.96% identity between replicates (*SI Appendix*, Table S1). To polarize SNPs into ancestral and derived variants we included *Bipolaris sorokiniana* and *Bipolaris maydis* as outgroups (25, 26). This data set was expanded by two *Setosphaeria turcica* reference genomes obtained from isolates Et28A and NY001 collected in the United States, which resulted in a total sample of 123 isolates. The data derived from this sample consisted of 4,257 polarized SNPs, corresponding to 18.3% of non-polarized SNP data.

Presence of different clonal lineages. To determine the genetic relationship of *S. turcica* isolates we clustered the original 121 samples with ADMIXTURE into $K = 5$ clusters (Fig. 1B). Five isolates had ancestry coefficients of $<$70% and were not assigned to clusters. All clusters defined by ADMIXTURE were supported by a rooted Neighbor-Joining tree based on polarized SNPs, a principal component analysis (PCA) and Community Oriented Network Estimation ((CONE), 27) (Fig. 1A-D). Three of the five ancestral clusters, which we named 'Big Clonal' (47 isolates), 'Small Clonal' (16 isolates) and 'French Clonal' (9 isolates), showed very short internal branches and the two remaining clusters, 'Diverse' (17 isolates) and 'Kenyan' (27 isolates), showed long internal branches in the phylogenetic tree. The NJ tree, PCA and Neighbor-Net reveal a close relationship of the French Clonal cluster with the Kenyan isolates and a strong differentiation from the other three European clusters (Fig. 1A,C,E and *SI Appendix*, Fig. S1). All five clusters, however, appear to have arisen by sexual recombination as indicated by reticulate patterns at the base of each clade in the Neighbor-Net (Fig. 1E).

We also observed genetic differentiation within clusters. ADMIXTURE identified two distinct subclusters within the Kenyan cluster ($K = 7$), the Diverse cluster ($K = 6$), and the Big Clonal cluster ($K = 8$; *SI Appendix*, Fig. S1). CONE identified four connected subclusters within the Big Clonal cluster and two disconnected subclusters in the Diverse cluster (Fig. 1D). The latter may consist of distinct clonal lineages that originated by recombination as shown by the Neighbor-Net. In contrast, no recombination is evident within the Big Clonal cluster and its subclusters, which therefore reflect evolutionary lineages of independent mutations.

To test whether the four European genetic clusters are geographically clustered, we analysed the spatial auto-correlation with Moran's I using ADMIXTURE ancestry coefficients ($K = 5$; Fig. 2A). Correlograms of Moran's I

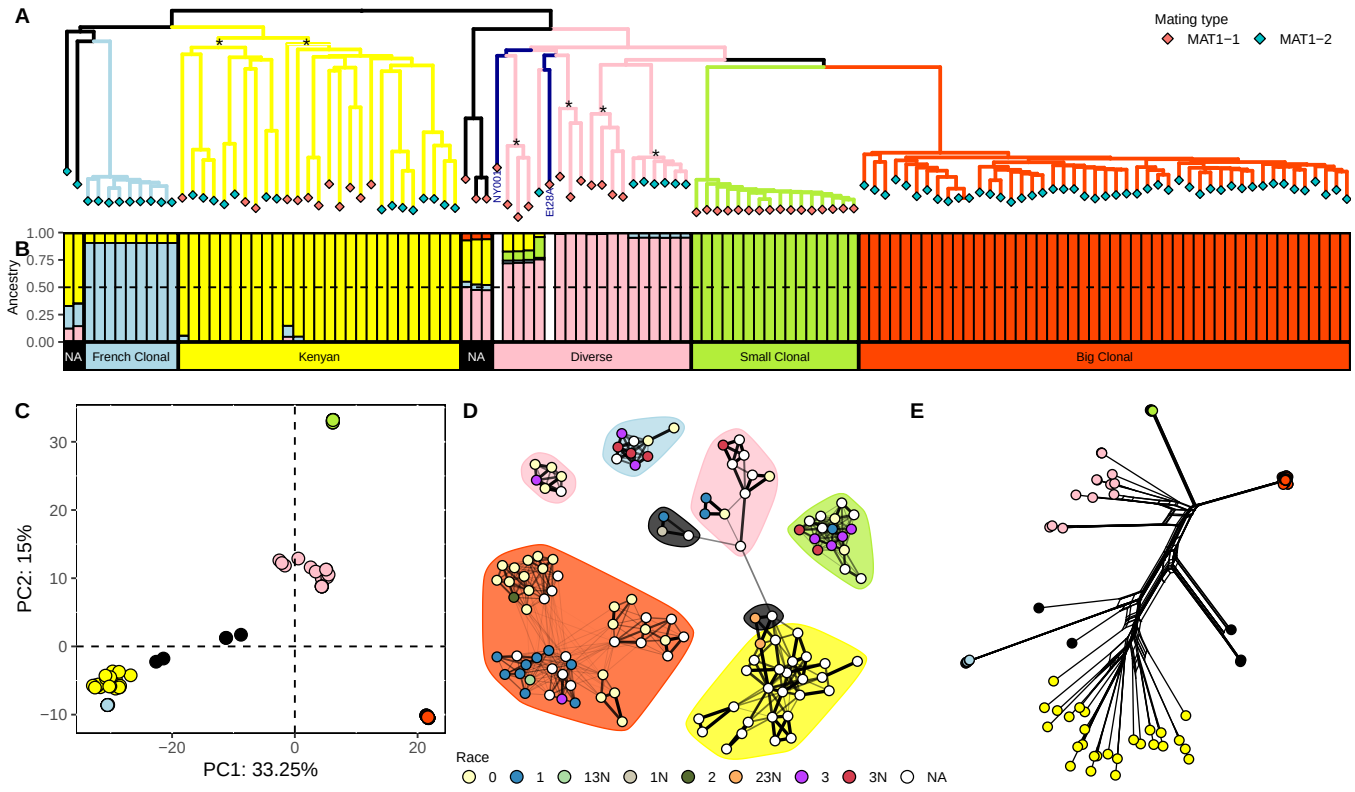

Fig. 1. A) Rooted Neighbor-Joining tree from the polarized SNP dataset, branches are colored according to the observed ADMIXTURE clusters, dark blue indicates the two reference genomes (NY001 and Et28A), * sign indicates the subclusters within the major clusters (two subclusters in Kenyan and four in the Diverse cluster). Rhombuses in tip nodes are colored according to the mating type. **B)** Individual ancestry coefficients from ADMIXTURE for K=5 in the same order as the rooted NJ Tree. White gaps correspond to the two reference genomes which were not analysied in ADMIXTURE. NA show admixed individuals with no cluster assigned. **C)** First two axes of a PCA colored according to the five observed ADMIXTURE clusters. **D)** Population network created with CONE colored according the phenotyped race (NA in white for unknown race). Background color highlights the five ADMIXTURE classification clusters. **E)** Neighbor-Net created with SplitsTree colored according to the five observed ADMIXTURE clusters.

indicate a wide geographic distribution and absence of geographic clustering of the Diverse and Big Clonal clusters (Fig. 2B). In contrast, the French Clonal cluster is strongly clustered in France and the Small Clonal cluster at sampling locations within and between the Upper Rhine Valley and the border between Northwestern Austria and Southeastern Germany.

Mating type and recombination. Sexual reproduction in *S. turcica* is controlled by the *MAT1* locus with the *MAT1-1* and *MAT1-2* ideomorphs (28, 29), which are highly dissimilar alleles. Sequence reads from *MAT1-2* isolates do not map to a *MAT1-1* reference (and vice versa) resulting in an alignment gap. To determine the mating type of isolates we assembled all unmapped reads *de novo* into contigs and compared them with BLAST to a database of *S. turcica* sequences that included both *MAT1-1* and *MAT1-2* alleles. Isolates were classified as either *MAT1-1* or *MAT1-2* because all reads and contigs mapped to only one of the two mating types. Three clusters (Big Clonal, Small Clonal and French Clonal) are fixed for one mating type, whereas the Kenyan and Diverse clusters each have approximately 1:1 ratios of the two mating types, consistent with a history of sexual reproduction (Table 1). The presence of different mating types as indicator of sexual reproduction is supported by the Phi recombination test, which identifed past recombination events in the Kenyan and Diverse, but not in the other three clusters (Table

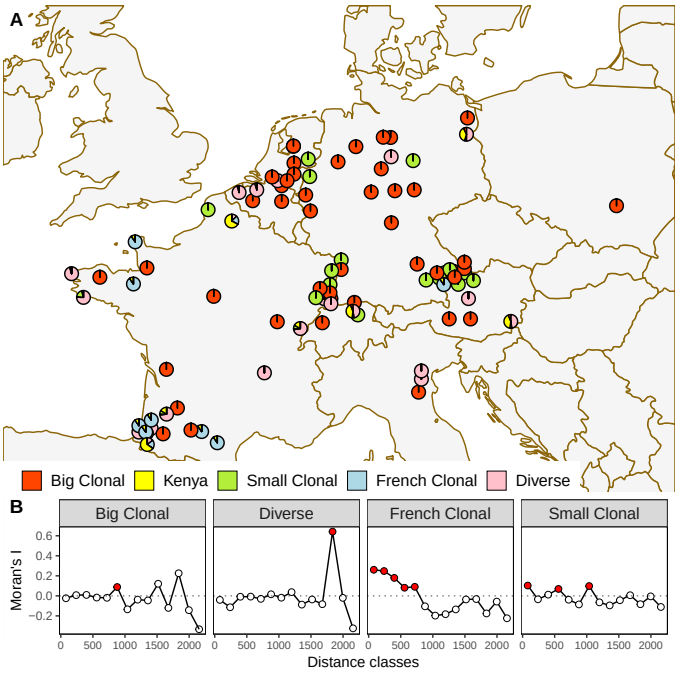

Fig. 2. A) Geographic origin of European isolates. Pie charts indicate ancestry coefficients for $K = 5$ to show the geographic distribution of the five major genetic clusters. Geographically close isolates are shifted to avoid overlaping of pie charts. **B)** Correlogram of Moran's I of the European ancestry coefficient along different distance classes. In red, p-value of Moran's I < 0.05

1). The test also detected recombination within the two Kenyan subclusters, of which each harbors both mating types in roughly equal proportions (Fig. 1A). We found no recombination within the four lineages of the Diverse cluster, consistent with the fixation one mating type within each lineage of this cluster. A permutation test on the standardized index of association, $\bar{r}_d$, rejected the null hypothesis of random association of alleles in all five clusters, suggesting that despite past episodes of sexual reproduction, the Diverse and Kenyan clusters also show high rates of asexual reproduction in recent time.

The geographic distribution of the two mating types is correlated with the geographic distribution of the four European clusters as the mating type is fixed within each of the three European clonal lineages. However, mating types of the Diverse cluster are unequally distributed with a higher proportion of *MAT1-1* in the Southeastern part and a higher proportion of *MAT1-2* in the Northwestern part of its sampling area (*SI Appendix*, Fig. S2)

Differences in genetic diversity between clusters. Consistent with their different histories of sexual and asexual reproduction, the five clusters also differ by their level of nucleotide variation (Table 1, Fig. 3A). Nucleotide diversity, π and Watterson's estimator, θ_W, are higher among the 26 isolates from Kenya (Genome-wide $\pi = 6.631 \times 10^{-5}$, per base pair) than among the 94 isolates from Europe (5.365×10^{-5}). Both clusters harbor a high proportion of SNPs not present in the other cluster because only 4,647 SNPs segregate in both clusters, corresponding to 33% and 39% of the SNPs of the Kenyan and European clusters, respectively.

SNP-based genetic diversity differs between the four European clusters (Table 1A). The Diverse cluster shows 10 to 30 fold higher genetic diversity compared to the three clonal lineages. Its genetic diversity is 82% of the total European and 65% of Kenyan samples, respectively. Similar differences between the clusters are observed with

Table 1. Diversity and reproduction type statistics of Kenyan and European isolates

Cluster	n	S	π	θ_W	Tajima's D	*MAT1-1*	:	*MAT1-2*	Phi (p-value)	$\bar{r}_d$ (p-value)
Kenya	26	11,880	6.631×10^{-5}	7.852×10^{-5}	-0.62	11	:	15	0.0000	0.001
Europe	94	14,094	5.365×10^{-5}	6.949×10^{-5}	-0.78	29	:	65	-	-
Small Clonal	16	393	1.52×10^{-6}	2.99×10^{-6}	-2.15	16	:	0	0.4873	0.001
French Clonal	9	215	1.47×10^{-6}	2×10^{-6}	-1.38	0	:	9	0.1034	0.001
Diverse	17	5,631	4.445×10^{-5}	4.201×10^{-5}	0.25	10	:	7	0.0000	0.001
Big Clonal	47	1,514	3.11×10^{-6}	8.65×10^{-6}	-2.36	0	:	47	0.9229	0.001

Diversity statistics: S: number of segregating sites, π: nucleotide diversity per bp, θ_W: Watterson's estimator per bp, D: Tajima's D. Results are rounded to the number of presented digits. Reproduction type statistics: *MAT1-1* : *MAT1-2* as matying type counts. Phi (p-value): p-value of the Phi recombination test. $\bar{r}_d$ (p-value): p-value of the standarized test of random association of alleles.

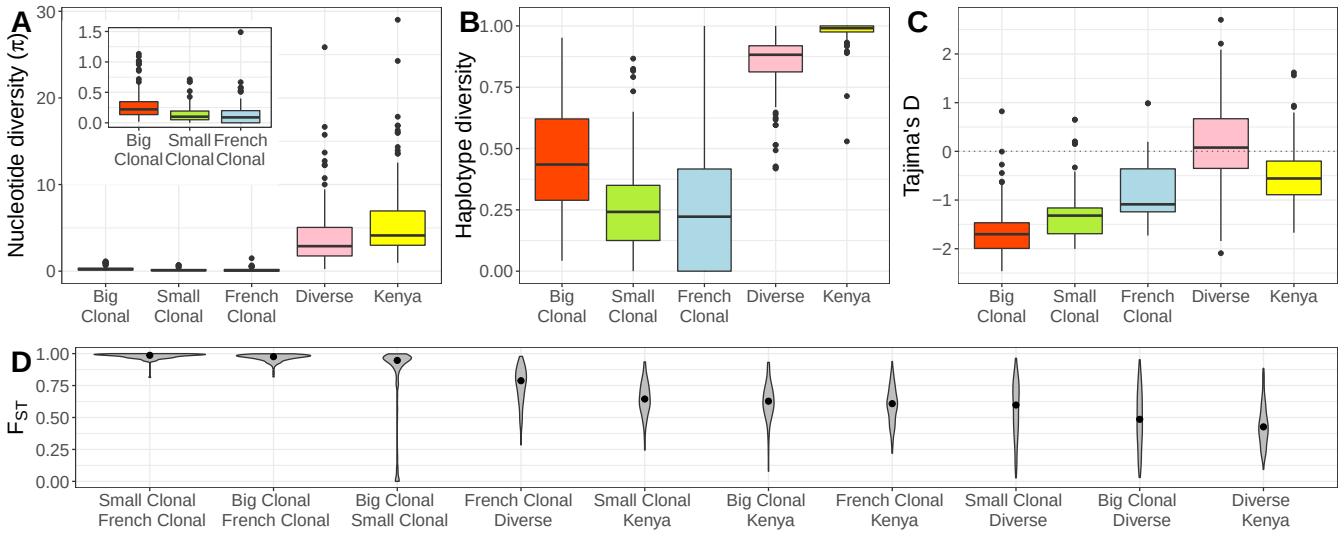

Fig. 3. Levels of genetic diversity in five different genetic clusters defined by ADMIXTURE $k = 5$. **A)** Nucleotide diversity π per bp (in units of 10^{-6}), **B)** haplotype diversity, **C)** Tajima's D, and **D)** pairwise Fst calculated in windows of 250kb. The inset plot in (A) zooms into the y axis for the three first clusters (same units).

haplotype diversity (Fig. 3B). Tajima's D values of the Big and Small Clonal clusters are highly negative (< 2) and less negative in the French Clonal cluster (-1.4; Table 1, Fig. 3C). The negative Tajima's D values of the clonal lineages indicate a genome-wide excess of rare alleles that may be caused by demographic effects like population growth following recent emergence or genome-wide purifying selection. The first explanation was proposed for similar patterns in clonal lineages of other plant pathogens (e.g., 13). Genetic differentiation of SNPs was measured as F_{st} and was highest between clonal lineages and smaller between the clonal lineages and the Diverse and Kenyan clusters, respectively (Fig. 3D).

The five clusters also differ in the distribution of genetic diversity along the genome. The Big Clonal, Small Clonal and French Clonal clusters have numerous genomic regions devoid of any genetic variation, whereas variation is more uniformly distributed in the Diverse and Kenyan clusters (*SI Appendix*, Figs. S3, S4 and S5). The lack of diversity is particularly strong for the Small Clonal and French Clonal clusters, because only 3% (Small Clonal)

Table 2. McDonald-Kreitman test

Cluster	π_N/π_S	P_n/P_s	D_n/D_s	NI	p-value
Big Clonal	0.747 ($1.30\times10^{-6}/1.80\times10^{-6}$)	1.648 (201/122)	1.366 (168/123)	1.206	0.283429
Small Clonal	2.123 ($8.00\times10^{-7}/4.00\times10^{-7}$)	5.455 (60/11)	1.480 (182/123)	3.686	0.000055
French Clonal	0.299 ($4.00\times10^{-7}/1.30\times10^{-6}$)	2.818 (31/11)	1.653 (390/236)	1.705	0.185753
Kenya	0.506 ($2.47\times10^{-5}/4.88\times10^{-5}$)	1.734 (1278/737)	1.395 (113/81)	1.243	0.161458

π_N/π_S is the ratio of non-synonymous to synonymous nucleotide diversity; P: population polymorphisms; D: fixed derived mutations (reference Et28A as outgroup); n: non-synonymous mutations, s: synonymous mutations; NI: Neutrality Index, calculated as $(P_n/P_s)/(D_n/D_s)$; p-value: Fisher's exact test p-value

and 1% (French Clonal) of all 100 kb windows on the 15 longest scaffolds of the reference genome segregate for five or more SNPs. For the clonal lineages, most windows reflect the genome-wide negative Tajima's D values and there are no visible outliers with highly negative Tajima's D values that may reflect strong localized selective sweeps (*SI Appendix*, Fig. S6). In contrast, both the Big and Small Clonal clusters have windows with highly positive Tajima's D values (e.g. on scaffolds 4 and 10), which may indicate mapping errors caused by structural variants or strong balancing selection (*SI Appendix*, Table S2). However, these regions contain only very few (≤ 5) SNPs and, for the Big Clonal cluster, outlier Tajima's D values do not deviate significantly from a neutral model of a constant or exponentially growing population (*SI Appendix*, Text A, Table S3).

Tests of selection. To investigate whether the genetic clusters were affected by positive or purifying selection, we applied the McDonald-Kreitman (MK) test and compared synonymous and non-synonymous variation among isolates relative to the reference genome Et28A (Table 2). Although ratios of non-synonymous and synonymous substitutions (D_n/D_s) are usually used for interspecific comparisons, they can also be interpreted for well separated clonal lineages (30). The Et28A reference clusters with the Diverse cluster, thus they are not well separated and we did not perform the analysis for the Diverse cluster. The ratio of synonymous to non-synonymous nucleotide diversity (π_N/π_S) estimates the fraction of effectively neutral mutations among all mutations (31) under Ohtas's nearly neutral model (32). The Big Clonal, French Clonal and Kenyan clusters show π_N/π_S ratios below 1 indicating that a majority of mutations are non-neutral or nearly neutral. Variation in the Small Clonal cluster differs from a nearly neutral model with a ratio $\pi_N/\pi_S = 2.1$ and a much higher ratio of non-synonymous to synonymous mutations, $P_n/P_s = 5.5$ than the other clusters (Table 2). However, with the exception of the Small Clonal Cluster ($p < 0.0001$), a MK test does not reject the null hypothesis of neutral evolution indicating that purifying selection has no significant effect on the fate of mutations in four of the five genetic clusters of our sample, which is unexpected given the π_N/π_S ratios observed.

Inference of split times. To investigate the demographic history of European isolates we included the two North American isolates Et28A and NY001 and used the polarized SNP data.A rooted tree revealed a close relationship of the American and European isolates (Fig. 1A), which was independently confirmed by merging our resequencing data with genotyping by sequencing (GBS) data of 13 North American isolates (33) resulting in a set of 280 genome-wide SNPs (*SI Appendix*, Fig. S7). The resulting phylogenetic tree and PCA plot (*SI Appendix*, Fig. S7) of the merged dataset are essentially identical to the analyses of European isolates based on the complete sequencing data. Both

methods group the North American isolates with the Diverse cluster, consistent with the tree in Fig. 1A.

To test whether European clonal lineages split before or after their introduction to Europe we estimated divergence times between the five clusters (times back to the most recent common ancestor, MRCA, of a pair of clusters) and emergence times of clonal lineages within clusters (times back to the MRCA of each cluster) using BEAST (Fig. 4A and *SI Appendix*, S8). The three clonal clusters diversified quite recently with posterior mean emergence of the most recent common ancestor in the year 1985 for the Big Clonal (1978-1990 include $\geq$ 95% posterior mass with highest posterior density, 95%HPI), 1998 for the Small Clonal (1993-2001, 95%HPI) and 1999 for the French Clonal (1995-2002, 95%HPI) clusters. Split times between clusters are more distant and range from the year 1609 between Small Clonal and Big Clonal (1480-1809, 95%HPI), 1503 between the ancestors of Big Clonal, Small Clonal and the Diverse cluster (1456-1667, 95%HPI) to 1198 between the ancestors of the Big Clonal, Small Clonal, North American reference isolates and French Clonal (975-1368, 95%HPI). The Diverse cluster emerged much later than the clonal clusters in 1520 (1386-1624, 95%HPI) and the nodes of its genealogical tree are more spread over time. Split times and tree topology of the BEAST analysis agree with the phylogeny in Fig. 1A and support a much closer pairwise relationship of Big Clonal and Small Clonal than to the French Clonal cluster. Including non-clonal lineages in analyses to estimate split and emergence times may introduce a bias due to reticulate events (13). To test for such a bias, we included only the Big Clonal, Small Clonal, and French Clonal clusters with and reference genome in a BEAST analysis. We obtained the same phylogeny as in Figure 4A and split time estimates that are slightly more in the past (*SI Appendix*, Table S4 and *SI Appendix*, Figs. S9 and S10). This comparison shows that time estimates are robust with respect to the mode of reproduction. We then investigated whether the global expansion of maize cultivation after the beginning of the Columbian exchange in 1492 and the strong increase of maize cultivation in Europe during recent decades was accompanied by an increase of the effective pathogen population size, N_e. After adding global population size as parameter to the phylogentic model for BEAST median, posterior estimates of N_e changed substantially over time (Fig. 4B and *SI Appendix*, Figure S8). Estimates of N_e based on the European samples, three samples from Kenyan cluster and the North American reference sequence indicate a long phase of population growth since the time of the most recent common ancestor (MRCA) of the European samples about 825 years ago until a period between 1859 and 1900, followed by a population decline until 1999, when population size N_e was lower than at the time of the MRCA. This decline was then followed by a very recent epoch of a strong population growth for 20 years until the last sampling date 2012. A recent, rapid growth is consistent with strongly negative genome-wide Tajima's D values within the three European clonal clusters. A decline of N_e followed by recent strong growth was confirmed by analysing only Big Clonal, Small Clonal, and French Clonal clusters together with the reference genome (*SI Appendix*, Figs. S9 and S10).

Neutral versus selection-driven population dynamics. The low genetic diversity and genome-wide excess of rare polymorphisms within clonal lineages may reflect rapid population growth or result from recurrent, short phases in which newly emerged genotypes with a skewed offspring distribution become dominant. Among predominately asexually reproducing fungal pathogens, following processes may lead to a skewed offspring distribution even without population size changes: (*i*) rapid selection of newly emerged genotypes with a very high fitness coefficient (34), (*ii*) a large number of offspring originating by chance from a single parental genotype analogous to sweepstake

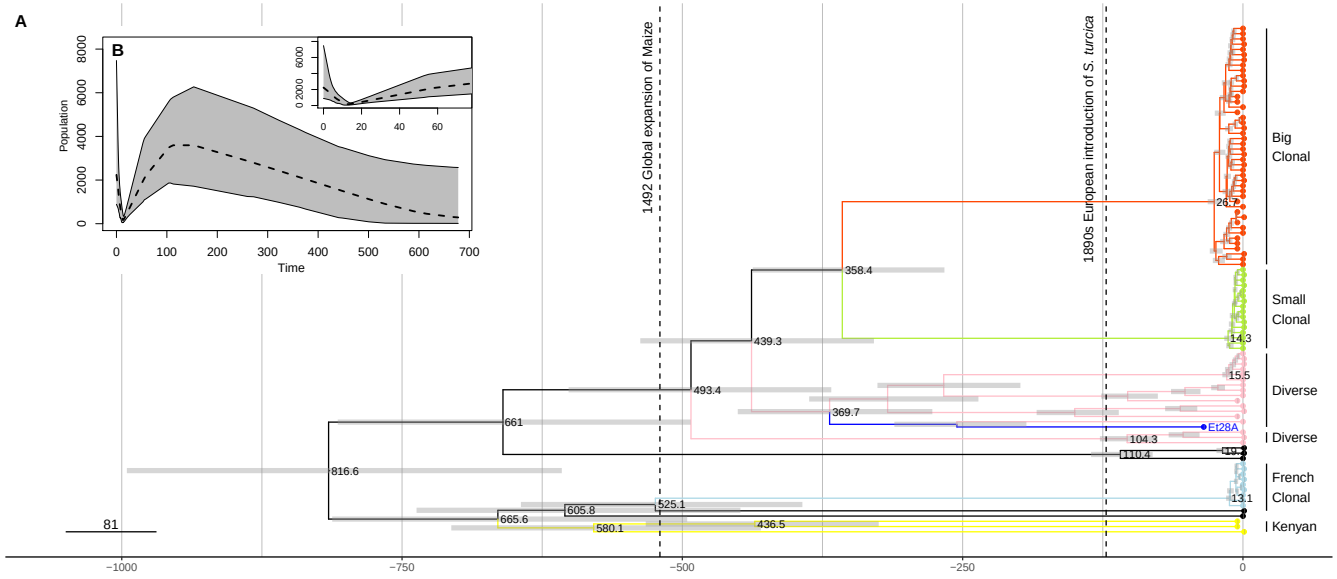

Fig. 4. A) Dated phylogeny obtained with BEAST, using all European isolates, reference genome Et218A (in dark blue) and three samples from Kenyan cluster. Time is given as years before 2012, the year of the most recent sampling. Horizontal gray bars show the 95 % highest posterior density intervals (95% HPI) for split times. **B)** Extended Bayesian skyline plot obtained with BEAST for the analysis from (A). The inset zooms into the most recent past. Time runs backwards from 2012. The dashed line shows the median posterior population size, while the gray area shows the 95% HPI.

reproduction in marine species (35, 36), or (*iii*) a large number of offspring from genotypes that evolved virulence against monogenic resistance genes present in maize varieties (boom-bust cycles) (37). Genealogies in these cases can be modeled as multiple-merger coalescents. We compared these models with a standard Wright-Fisher type reproduction with growing population sizes, modeled via a bifurcating Kingman coalescent with exponential growth. For the Big Clonal and Kenyan clusters we compared both coalescent models using SNPs segregating within this cluster on the five largest scaffolds of the current *S. turcica* reference genome using a Random Forest Approximate Bayesian Computation (RF-ABC) approach for model comparison and parameter estimation. In the other clusters we observed high prior error rates and low posterior probabilities and considered these results not as robust (*SI Appendix*, Table S5). Table 3 shows the results of the RF-ABC analysis. For the Big Clonal cluster, it provides strong support for a bifurcating Kingman coalescent with strong exponential growth over a multiple merger coalescent, which refutes strong selection without growth or sweepstake reproduction. The Kingman coalescent is also preferred for the Kenyan cluster, but with much smaller growth rates. All scaffolds in the Big Clonal and in the Kenyan cluster show 'positive' ($\geq$ 3 odds ratio) to 'strong' support ($\geq$ 20 odds ratio, only for Big Clonal) for an exponential growth model over a multiple merger model according to the Kass-Raftery scale (38). Simulations showed that the observed genetic diversity in both the Kenyan and Big Clonal clusters are obtained with the best-fitting model (*SI Appendix*, Text B, Figs. S11 and S12).

Although our analyses reject the hypothesis that skewed offspring distribution alone shapes genetic diversity for the Big Clonal and Kenyan clusters, a combination of exponential growth with skewed offspring distributions explains the data similarly well as neutral population growth (*SI Appendix*, Text C).

Table 3. RF-ABC model selection results

Cluster	SNP count	Mean OOB	Post. prob. Exp. growth	Odds ratio	Fitted rate g
Big Clonal	92-128	21-22%	83 - 98%	4.9-49	181-630
Kenya	683-1007	20%	75.2 - 84.3%	3-5.4	5.5-8.5

Mean OOB is the out-of-bag prior error rate for model classes, averaged over all model classes. Post. prob. exp. growth gives the posteror probability of the exponential growth model. Fitted parameter g: Posterior median of the exponential growth parameter in coalescent units, where one unit represents $2N$ generations. For each variable, we report the range of values for the five biggest scaffolds. For the transformation of posterior probabilities into odds ratios for a comparison of exponential growth vs. any other genealogy model used, see Materials and Methods.

Different pathogen races within clonal lineages. To test whether isolates within clonal lineages belong to the same or different races, we identified 62 isolates in our sample whose race was determined in a race monitoring of 542 European *S. turcica* isolates collected in 2011 and 2012 from the major maize growing regions in Europe (23). The monitoring revealed that race 0 was the most dominant with 45% of isolates, followed by race 1 (22%), 3 (15%) and 3N (14%). Only 4% of isolates were virulent against two or more resistance genes (races 13, 123, 23, 2, 23N, 12, 1N and 13N). Mapping the race type of the 62 isolates onto the CONE network reveals that Big and Small Clonal clusters harbor four races each and the French Clonal cluster three races (Fig. 1D). Single, independent *de novo* mutations in pathogen effector genes are sufficient to create new races and may explain the diversity of races within clonal lineages. Alternatively, the presence of the same races in different lineages may reflect shared polymorphisms that originated in ancestral populations although such an explanation seems unlikely given the low genetic diversity within clonal lineages (Table 1).

Identification of divergent regions and structural variants. The differences in SNP allele frequencies between clonal lineages suggests that highly divergent genomic regions and presence absence structural variants (PAVs) also contribute to genomic differentiation. We therefore used sequence read coverage and k-mer frequencies to identify highly divergent regions and PAVs. First, we calculated for each isolate its sequence coverage of the reference genome in 43,443 windows of 1 kb length and expressed coverage as percent bases covered by at least one sequence read in each window. Windows with low coverage indicate a high proportion of mapping gaps in the reference and windows with a highly variable coverage between isolates pinpoint structural variants. Using the top 2.5% windows ($n = 1,012$) with the most variable sequence coverage between isolates, we constructed a NJ tree from a pairwise Euclidean distance matrix of reference sequence coverage (*SI Appendix*, Fig. S13) to cluster isolates with similar variation in coverage. The topology of the resulting tree is highly similar to the rooted SNP-based tree indicating that highly variable regions and PAVs reflect similar genealogical process than SNP allelic variation.

To identify genomic regions that differentiate pairs of clonal clusters we used HAWK (39), which identifies k-mers whose frequency differs between clusters. Among all pairwise comparisons (see Methods), we obtained different k-mer frequencies only between the Big Clonal vs. Small Clonal clusters. Among 6,341 k-mers that differentiate the two clusters, 3,048 are associated with the Big and 3,293 with the Small Clonal cluster. We *de novo* assembled both k-mer clusters independently into longer sequence contigs and found that 93% of assembled k-mers mapped

to few, distinct regions of the reference genome, suggesting that a small number of genomic regions contribute to genomic differences between the two clusters. Assembled *k*-mers mapped to only 1,167 (12.73 %) 10kb windows of the reference genome. There were only 30 windows (0.33% of all windows) that collected the top 2.5% *k*-mer counts with at least 9.85 mapped *k*-mers per window. Among all mapped *k*-mers, 25% map to these 30 windows, which tend to be highly repetitive. A majority of 22 out of 30 windows (73%) is highly repetitive with $\geq$ 50% repetitive elements and no window contains gene-rich regions.

To identify proteins that may differentiate the Big and Small Clonal clusters, we conducted a BLASTX analysis against a non-redundant BLAST protein database with the remaining unmapped *k*-mers. For both clusters, 'hypothetical protein' was the most frequent annotation of proteins among the five best hits with a cutoff *e*-value of < 0.001, followed by the mating type *MAT1-2* for the Big Clonal cluster. The latter finding is a positive control of the *k*-mer mapping approach because Big Clonal has *MAT1-2*, which does not map to the reference genome Et28A, and Small Clonal has *MAT1-1*, which maps to the reference genome. For the Small clonal cluster, the second most frequent BLAST hit was 'polyketide syntase protein', which is potentially associated with pathogen virulence (25). We also used the race assignment to identify *k*-mers associated with race-specific genes, however no significant and robust outcome was found (Table S6). This negative result may either reflect a too small sample size or genetic differences of single or few variants that are not uncovered by the analysis of *k*-mers.

Discussion

Our work confirms earlier studies of *S. turcica* genetic diversity and mode of reproduction in Europe and Africa (21, 40). Isolates originating from Kenya form a single cluster with high genetic diversity, equal frequency of both mating types and genomic patterns of recombination consistent with a higher rate of sexual reproduction of *S. turcica* in tropical climates. In contrast, European isolates are composed of four distinct clusters, which differ by their relative frequency and geographic distribution. Three clusters (Big Clonal, Small Clonal, French Clonal) represent single clonal lineages, whose genetic diversity is very low, do not show evidence of recent recombination, and are fixed for one of the two mating types. The fourth cluster (Diverse) consists of diverse clonal lineages that, taken together, have a high level of genetic diversity, evidence of past recombination and an equal frequency of both mating types. These three characteristics in combination with low F_{ST} values between the Diverse vs. the Big Clonal and Small Clonal clusters, respectively, suggest that the Diverse cluster is a source of genetic diversity from which clonal lineages emerged previous to the arrival of *S. turcica* in Europe. The two North American isolates Et28A and NY001 cluster with the Diverse cluster and are highly similar to different European isolates indicating the close connection between European and American samples that may reflect an American origin of the Big Clonal, Small Clonal and Diverse clusters. In contrast, the French Clonal cluster is closely related to the Kenyan cluster and therefore likely has the same origin. A previous study interpreted the presence of African alleles in an isolate from Southwestern France of *S. turcica* as recent migration (21). This is not supported by our analysis because the split time of Kenyan and French Clonal clusters predates the arrival of *S. turcica* in Europe, and their close relationship reflects a common ancestry instead of recent migration.

Our divergence time estimates suggest that individual lineages within each of the three Big Clonal, Small Clonal

and French Clonal clusters emerged less than 40 years ago. Although these very recent emerge times are based on a limited sample we consider them reliable (41) (*SI Appendix*, Text D). In contrast, the divergence times of the five clusters identified in our sample are more distant and range between 816 to 360 years ago (Fig. 4), which predates the introduction of *S. turcica* into Europe and strongly suggests these clusters originated outside Europe and were independently introduced. The clonal sublineages within the Diverse cluster originated between 370 to 50 years ago and are separated by sexual recombination events, which are unlikely under European climatic conditions. For this reason, they were likely independently introduced into Europe.

Evolutionary forces determining pathogen demography. Although multiple crop pathogens expanded globally in short time, only few studies analysed the evolutionary forces determining expansions, in particular the role of selection on plant pathogens, using explicit population genetic modeling (42). We employed Approximate Bayesian Computation (ABC) to compare two coalescent models and to differentiate between a neutral model of exponential growth and a selection-based model of the *S. turcica* expansion in Europe. Simulations of models with asexual reproduction demonstrate a high power of ABC to differentiate between neutral and selection-driven demographies with suitable summary statistics (43). The Big Clonal cluster is particularly interesting for such an analysis because it is the most successful cluster in terms of sample frequency and geographic distribution in Europe. Its large sample size provides better statistical power and a restriction of ABC to clusters without a strong internal population structure removes a bias in distinguishing among genealogy models (44). The ABC analysis of the Big Clonal cluster (Table 3) reveals a recent population size increase and in addition that observed genetic diversity in this cluster is not consistent with a history of rapid selection or boom-bust cycles caused by host-pathogen coevolution without population growth. In other fungal crop pathogens such as *Zymoseptoria tritici* random fluctuations in fecundity and a potential for very large offspring numbers per individual have been proposed (36), which should lead to a multiple merger genealogy if it is strong enough. Our results exclude such a model as sole explanation for the observed diversity in the Big Clonal Cluster or indicate that fecundity differences in the pathogen are too small to affect the shape of the genealogy. Instead, observed diversity within this cluster can be explained by just assuming neutral population growth. These analyses do not exclude the possibility that an exponential increase of N_e in the Big Clonal cluster results from a relative fitness advantage caused by adaptive *de novo* mutations or a favourable combination of adaptive mutations achieved via sexual recombination in the founders of the cluster. In addition, a more complex pattern of neutral population growth on top of selection processes or sweepstake reproduction can also not be ruled out. Since the Kenyan cluster also supports a neutral coalescent model with a low rate of population growth there is no reason to expect multiple mergers as standard gene genealogies in *S. turcica*. The absence of interpretable results for the ABC analyses for the Small Clonal and French Clonal clusters likely results from too small sample sizes.

Tests of neutrality based on comparisons of non-synonymous and synonymous genetic diversity (Table 2) do not contradict a model of neutral evolution as main driver of genetic diversity for the Big Clonal, French Clonal and Kenyan clusters.Although there is an excess of non-neutral diversity in these clusters, it is not strongly selected against as indicated by non-significant MK test results. This seems contradictory at first, but may have several explanations: clonality preventing efficient selection against deleterious mutants, a surplus of beneficial founder mutations offsetting the effect of purifying selection (this may need many such mutations and thus be unlikely) or

simply an underpowered MK test. Nevertheless, it shows that even if there is purifying selection, it is not a main driver of genetic diversity within these clusters. The Small Clonal cluster, however, may have a different evolutionary history because the significant MK test result for purifying selection contrasts with an excess of non-synonymous diversity, which suggests that evolution in this cluster does not follow a nearly-neutral model. Overall *S. turcica* genetic clusters do not provide evidence for strong purifying selection, which is in contrast to the rice fungal pathogen *Magnaporthe oryzae* (11). Taken together, absence of selection as main driver for the genetic diversity of the Big Clonal cluster and a close temporal coincidence of *S. turcica* population growth with an expansion of maize cultivation in Europe leads us to propose that the expansion of this cluster was not driven by rapid evolutionary adaptation to European maize varieties or the environment.

Limited evidence for host-pathogen co-evolution in Europe. Our sample of isolates was collected in 2011 and 2012 and represents a snapshot in time and space that is restricted to Europe and Kenya. Both factors limit further interpretations of our results and lead to questions about the role of *S. turcica* - maize coevolution within and outside Europe. First, the demographic analysis suggests an independent and recent single introduction of the French Clonal, Small Clonal and Big Clonal clusters into Europe(Table 3), although our results are also consistent with independent, repeated introductions of the same clonal lineages. For example, the clonal lineages within the Diverse cluster originated by sexual recombination over an extended period of time (Fig. 4). Since sexual recombination is unlikely under European climatic conditions, the lineages likely originated outside Europe and were then subsequently introduced. Further evidence for repeated introductions is the high genetic similarity among European and North American isolates suggesting recent exchange or a common origin in a different region, such as Mexico, because European isolates were more similar to Mexican than to Kenyan isolates (21). Additional samples from putative regions of origin such as Central America and tropical Africa are required to resolve this issue.

A second question refers to the effects of maize resistance genes on *S. turcica* evolution and epidemiology. There is no association between the five genetic clusters and the distribution of *S. turcica* races among these clusters. This observation and a high proportion of race 0 (i.e., non-virulent against four tested *Ht* genes) isolates in all five clusters shows that race-specific virulence did not generate new pathogen lineages with a strongly increased fitness. In combination with the evidence for neutral evolution of genetic variation in the European isolates, we conclude that strong selection against qualitative or quantitative maize resistances had very little or no effect on genetic diversity in Europe. However, future studies should associate the genetic diversity of host and pathogen genomes using joint association analysis (e.g., 45) to elucidate the role of genotype by genotype (GxG) effects in the spatial and temporal dynamics of host-pathogen interactions. Such information will contribute to avoid breakdown of resistance genes and achieve long-term resistance management (46).

A third question refers to the evolution or new races, because the presence of multiple races within the five *S. turcica* clusters suggests a rapid and repeated breakdown of *Ht*-based monogenic resistances in maize varieties (Fig. 1D). Since selection against host resistance does not seem to affect the evolutionary dynamics of *S. turcica*, a frequent origin of new races may be facilitated by a high mutation rate, which we estimated as posterior mean substitution rate of 10^{-4} substitutions per year per site using BEAST. This rate is much higher than in *Magnaporthe oryzae*, where it was estimated to be in the order of 10^{-8} (11)). We used *k*-mer based association analysis to identify genomic regions

that may contribute to resistance breakdown, but did not find *k*-mers that are significantly associated with race type, possibly because of a small sample size for each race. In contrast, a *k*-mer analysis of the complete sample unambiguously identified the mating type gene and several genomic regions that differentiate clonal groups and harbor genes with putative roles in pathogenicity.

In conclusion, our analyses indicates a rapid spread of different *S. turcica* clonal lineages in Central and Western Europe in the absence of both recombination and strong selection for pathogen virulence. Monitoring of pathogen diversity on larger geographical scales and over time is required to fully understand forces influencing pathogen epidemiology and evolution, and the evolution of pathogen races. However, our work shows that large scale sequencing and population genomic analysis provide useful information to develop breeding programs informed by host-pathogen evolution and to control plant pathogens by improved agricultural management.

Materials and Methods

Cultivation of fungal isolates. The origin and sampling information of isolates is described in Dataset S1 Lyophilized isolates were transferred to Becton Dickinson BBD Potato Dextrose Agar plates and incubated for at least 10 days at 25°C and a 12h light / 12hr dark cycle until plates were completely covered by mycelia. This fungal tissue was scraped from the surface with a spatulum and collected in a 2 ml plastic reaction tube.

DNA extraction and NGS sequencing. After adding six ceramic beads (2.8 mm diameter; MoBio, USA) to each tube, the tissue was ground in a Retsch mixer mill (MM400) for 30 sec at a speed of 30 sec^{-1}. The DNA was then extracted with the Micro AX Blood Gravity KI (A&A Biotechnology, Poland; Cat. No. 101-100) according to manufaturer's instructions and diluted to a concentration of 2.5 ng μl^{-1} EB buffer. Whole genome equencing libraries were generated using a multiplex tagmentation protocol (47) with minor modifications. Our detailed protocol is available at protocols.io at and the assignment of barcodes to isolates as detailed in Dataset S1. The libraries were paired-end sequenced (2 x 100 bp) on a HiSeq 2500 Illumina sequencer (Macrogen, Korea) in three batches of 24, 96 and 46 isolates, respectively.

Read mapping and variant calling. Raw Illumina reads were processed for sequence quality using Trimmomatic v0.36 (48) with arguments `ILLUMINACLIP:NexteraPE-PE2:30:10:8 CROP:98 HEADCROP:8 LEADING:28 TRAILING:28 SLIDINGWINDOW:15:28 MINLEN:40 AVGQUAL:30`. Read pairs for which both forward and reverse passed quality control were kept for further analysis. Trimmed reads were mapped with BWA v0.7.12-r1039 (49) against the *Setosphaeria turcica* reference genome Et28A v1.0 (race 23N strain 28A) (25, 26). The reference consists of 403 scaffolds and was obtained from EnsemblFungi version 39 (50). PCR duplicates were removed with MarkDuplicates from Picard tools (http://broadinstitute.github.io/picard/) and mapped reads locally realigned with GATK v3.7-0 (51). Mean percentage of mapped reads and mean coverage were calculated with Qualimap v2.2.1 (52). Samples with low percentage of mapping reads (<83%) and/or low coverage (<5X) were excluded to avoid the analysis of contaminated samples. For variant calling we used two different methods an only kept variants identified by both methods. Genotypes were called for each sample eith GATK HaplotypeCaller (`-emitRefConfidence GVCF -min_base_quality_score 28 -min_mapping_quality_score 20 -ploidy 1`), and genotypes were merged with

GenotypeGVCFs (`-ploidy 1`). As second method samtools mpileup (53, 54) (`-t DP -t SP -t AD -g -C 50 -A -E -q 20 -Q 28`) was used with bcftools (55) (`-mv -ploidy 1`). Insertion-deletion (INDEL) variants were excluded with vcftools (56) (`-remove-indels`). We kept overlapping SNPs between GATK and samtools-bcftools methods using bcftools `isec`. SNPs were filtered for minimum read depth of 3 and a maximum of 100, minimum proportion of reads supporting a genotype call of 0.8, maximum percentage of missing data per SNP of 35%. Monomorphic positions within the sample set and also all non bi-allelic SNPs were excluded. Missing genotypes were imputed with multiple correspondance analysis (MCA) using five components with the function 'imputeMCA' of the missMDA (57) R package. The number of differences between imputed replicates were counted to estimate the error in the final SNP dataset.

Variant polarization. To polarize alleles we used the reference genomes of two closely related species, *Bipolaris sorokiniana* ND90Pr (25, 26) and *Bipolaris maydis* ATCC 48331 (25, 26) that were both obtained from EnsemblFungi. Outgroup genomes were aligned with *Setosphaeria turcica* reference genome using TBA (58). Genotype calling from the alignment was done with MafFilter (59) with VcfOutput option using *Setosphaeria turcica* scaffolds for variant location and alignments larger than 500 bp. Only bi-allelic variants shared with both outgroup species were kept. Ancestrality of the alleles was assigned to the allele of the outgroup species and genotypes of the 130 samples (129 isolates and *S. turcica* reference genome) were polarized accordingly. Additionally, we included the draft genome of *S. turcica* (race 1 strain NY001; JGI Fungal Program (60, 61) under GOLD Project ID Gp0110874), originally collected in Freeville, New York in 1983 (62) as additional sample to the polarized SNP dataset.

Mating type assignment. The *S. turcica* Et28A reference genome is of mating type *MAT1-1*. Since alleles for the *MAT1-1* and *MAT1-2* mating types are highly divergent, *MAT1-2* reads did not map to the *MAT1* locus of the reference genome. For this reason, isolates with a mapping gap on the *MAT1* locus were candidates for *MAT1-2* type. Confirmation of the mating type was done with a *de-novo* alignment of the unmapped reads with MegaHit (63) and posterior blasting with BLAST (64, 65) to a nucleotide database of *S. turcica* (which included the sequences of MAT1-1 and MAT1-2). *MAT1-1* type was assigned to samples that had reads mapping to the *MAT1-1* sequence from Et28A reference genome and *MAT1-2* type was assigned to samples that had a mapping gap at the *MAT1-1* sequence and that the unmapped reads had a blast hit for the *MAT1-2* sequence.

Population structure. To asses population structure we conducted Principal Component Analysis (PCA), maximum likelihood estimation of individual ancestries with ADMIXTURE (66) and population network estimation with community detection using neighborhood selection implemented in CONE (27). PCA was calculated with the glPca function of the `adegenet` (67, 68) R package. To run ADMIXTURE we used a pruned SNP dataset (LD cutoff r^2 >0.2 and minor allele frequency >0.05) and performed 20 independent ADMIXTURE runs with K from one to 15 and setting the `-cv` argument to compute cross validation errors. We used $K = 5$ clusters to explain the data because higher values did not significantly reduce cross-validation error (Fig. S14) and produced always the same composition of clusters with $K = 5$ in 20 independent runs. To merge the different admixture runs we used the CLUMPAK main pipeline (69) and kept the MajorGcluster output to plot the results. Population networks were estimated using R scripts for haploid data provided by CONE authors (27). The optimal value for the tuning

parameter was chosen with StARS using 40 subsamples of each of 1,777 SNPs, and 40 different values for the tuning parameter ranging from 0.5 to 0.005. Neighborhood selection was carried out with the whole dataset and selected tuning parameters. A population graph was constructed with the Fruchterman-Reingold algorithm and the Walktrap algorithm was used for community detection. An unrooted Neighbor-joining tree was built from Euclidean distances and Neighbor-net was calculated with SplitsTree v4.14.6 (70) using Hamming distances calculated from the unpolarized SNP dataset. A rooted Neighbor-joining tree was constructed with polarized SNP data that included the two references genomes of US-American origin. A correlogram on Moran's I (71) was used to test the European spatial autocorrelation along different distance classes of equal frequency. As quantitative variable we used the individual's ancestry coefficient of each ADMIXTURE cluster with $K = 5$. To calculate the number of distance bins, Sturges method was used as implemented in the `correlog` function (used to calculate the correlogram) from the `pgirmess` R package (72, 73), which uses `moran.test` function from `spdep` R package (74).

Diversity statistics. Numbers of segregating sites S, genome-wide nucleotide diversity π, Watterson's estimator θ_W and Tajima's D were calculated for the two sets of European and of Kenyan isolates, as well as for the subpopulations within them. The single isolate from Turkey within the Kenyan cluster (WGRS-Test_23) has a strong effect on its diversity measures because it contributes 886 additional SNPs (7% of the total). Since this sample is geographically separated from Kenya, we excluded it from the subsequent analysis of the Kenyan cluster. Both π and θ_W are reported per base pair by dividing genomewide values by the maximum number of bases aligned to the reference across all sampled isolates (which are 39,649,104 of 43,013,545). For the 15 biggest scaffolds, an additional sliding window analysis for windows of size $100k$ bp was performed. We computed π, θ_W, Tajima's D and the haplotype diversity per window for all groups with the R package `PopGenome` (75).

MK test. To calculate the McDonald–Kreitman (MK) test we first ran SnpEff version 4.3t (76) with the $-classic$ output style and Setosphaeria_turcica_et28a genome version on vcf subsets that included i) only population polymorphisms and ii) only fixed derived mutations. We used reference Et28A as outcluster, and excluded positions were Et28A and the sample WGRS_62 (closest sample to Et28A) were different . We counted the number of non-synonymous mutations as those classified as NON_SYNONYMOUS, STOP_GAINED, START_LOST or STOP_LOST, and the number of synonymous mutations as those classified as SYNONYMOUS_CODING, SYNONYMOUS_START or SYNONYMOUS_STOP. Thereafter Neutrality Index (NI) was calculated as $\frac{(P_n/P_s)}{(D_n/D_s)}$, where P are polymorphisms, D substitutions, s synonymous mutations and n non-synonymous mutations. Fisher's exact test P-value was computed using the 2x2 contingency table of the four type of mutations. π_N/π_S ratio was calulated as $(\sum_{i=1}^{I} \pi_{ni} N_i / \sum_{i=1}^{I} N_i)/(\sum_{i=1}^{I} \pi_{si} S_i / \sum_{i=1}^{I} S_i)$, where I is the number of scaffolds, N_i is the number of non-synonymous sites in scaffold i, S_i the number synonymous sites in scaffold i and π_{ni}, π_{si} are the non-synonymous or synonymous nucleotide diversities per non-synonymous or synonymous site in scaffold i. All π_{ni}, π_{si}, N_i and S_i were obtained from *population_summary.txt* output file after running SnpGenie (77) .

Analysis of reproduction type. The reproduction type (clonal vs. sexual) was analyzed with three approaches. First, we calculated the mating type ratio for each population. A 1:1 ratio of the mating type is a strong indicator for sexual reproduction whereas a significant skewed ratio indicates clonal reproduction (78, 79). Second, we tested

for recombination using the Phi test (80) as implemented in SplitsTree and third, we tested the null hypothesis of random association of alleles by 999 permutation tests of the standarized index of association ($r\bar{d}$) with `poppr` (81).

Demographic signals in clonal subpopulation. For the five clusters Big Clonal, Small Clonal, French Clonal, Diverse and Kenyan, we performed model selection between sweepstake reproduction (genealogies modelled by Dirac- and Beta-n-coalescents) or rapid selection (Bolthausen-Sznitman n-coalescent) in a fixed-size population and standard reproduction (Kingman's n-coalescent) in a fixed-size or an exponentially growing population. Additionally, we performed parameter estimation within the best-fitting model class. Model selection and parameter estimation is performed via random-forest based Approximate Bayesian Computation (82, 83) using quantiles of summary statistics for unpolarized data as described in (84). For the analysis, we treat each scaffold as a single non-recombining locus and run it on the 5 biggest scaffolds. We consider Beta$(2-\alpha,\alpha)$-n-coalescents with $\alpha \in [1,2)$ ($\alpha = 1$ denotes the Bolthausen-Sznitman n-coalescent), Dirac n-coalescents with parameter $p \in (0,1)$ and, for Kingman's n-coalescent, exponential growth rates in $[0,2500)$. We set a uniform prior on p for Dirac-n-coalescents, while for Beta-n-coalescents, we set $\alpha = 1$ with a probability of 5% and in all other case draw α uniformly from $(1,2)$. For Kingman's n-coalescent with exponential growth, the prior distribution on the parameters chooses growth rate $g = 0$ with probability 0.02, and with probability 0.98 an auxiliary paramter g' is chosen uniformly from $(\log(0.5), \log(2500))$ which is then transformed to $g = \exp(g')$. In other words, we use an uniform prior on the log scale on $(0, 2500)$ with an additional spike at $g = 0$.

The scaled mutation rate θ is set to the generalized Watterson estimator $\theta_W = 2S/E(L_n)$, where L_n is the expected total length of the underlying genealogy model, but with a random fluctuation around this estimate, as in Scenario 3 in (84) (binomial prior with 11 steps in $[\theta_W/5, 5\theta_W]$ with log-equidistant). As statistics, we use the $(.1, .3, .5, .7, .9)$-quantiles of the branch length of the neighbor-joining tree reconstructed from the genetic data, of the Hamming distances and of the linkage disequilibrium statistic r^2, as well as the number of segregating sites S, nucleotide diversity π and the folded site frequency spectrum, where all minor allele counts above 15 are summed up as a single statistic. We do not correct for unequal isolation times. While the effect of serial sampling may affect distinguishing multipler merger coalescents from Kingman's coalescent with exponential growth (85), its effect is neglectable if the model selection points to Kingman's coalescent with exponential growth. Each model class is simulated 175,000 times and the random forest is built from 500 trees. All simulations are performed as described in (84), the ABC parameter estimation and model selection are performed using the R package `abcrf`. An estimated lower bound for odds ratio/Bayes factors of the best fitting model to any other model (equivalent here due to a flat prior on the three model classes) is given by $BF = \frac{P(model|Data)}{1-P(model|Data)}$, i.e. we treat the posterior probability of any one other model as the posterior probability that the best fitting model is not the true model.

Phylogenetic dating with BEAST. We ran BEAST2 (86) on the non-polarized variants for all European isolates but WGRS_5, three Kenyan isolates (WGRS_26, WGRS_29 and WGRS-Test_23) and the American reference genome (setup 'full') and separately for the clonal Big Clonal (excluding WGRS_5), Small Clonal and French Clonal clusters, and the American reference genome (setup 'clonal'). Sample WGRS_5 was excluded because according to given information it is a sample from Kenya, but according to population structure analyses that seem not possible as it is

clearly a clonal isolate from Big Clonal cluster. Therefore, because of the uncertanty of the origin of this sample, we decided to exclude it from any analysis using any geographic information.

The Kenyan and Diverse clusters were excluded from the second analysis to omit most recombination signals that were not incorporated in the BEAST approach, following (13). However, we added the reference, because the clonal lineages covered only a narrow window of sampling time differences (in time and mutations) between individuals and showed more noisy posterior estimates for split times, see *SI Appendix* Table S4 and Text E for more details. Each isolate was timed relative to its time of isolation. As site model, we used a Γ model with four categories and estimated the proportion of invariant site, starting with a proportion of 0.8. We used the HKY model for mutation, estimating the frequencies and assumed used a strict molecular clock. Test runs with a relaxed exponential molecular clock with two discrete rates and with the different mutation model GTR showed only very small changes, which indicated that potentially shorter generation times of *S. turcica* in warmer climates need not to be accounted for. Since only in-species samples are included, we the tree modeled with a coalescent, i.e the Coalescent Extended Bayesian Skyline. As starting tree, we used the cluster tree estimated via `NJ2`. All other model settings were kept at the default values. The MCMC parameters were 225 million cycles (every 100th trace and 1,000th tree stored) with a 10 million pre-burnin period (300 million for running on only clonal lineages and reference). For tree annotation, we used a 10 % burn-in. After several pilot runs to adjust parameters, we conducted analyses with the above parameters. Effective Sample Size (ESS) scores that describes the correspondence the posterior distribution approximated by BEAST to the number of independent posterior, $>$100 for all non-population size parameters for the 'full' setup and $>$200 for the 'clonal' setup. Several population size parameter scored between 45 and 100 for the 'full' setup and for the 'clonal setup' all but three sizes were $>$200 with a minimum ESS of 172.

Variation in sequence coverage. To use variation in sequence coverage as phylogenetic signal, we calculated variation of coverage in 1kb windows in each isolate. The most variable 2.5% windows were used to calculate a pairwise distance matrix of variance in coverage between all samples, from which a Neighbor-Joining tree was constructed.

Reference-free association mapping. To characterize sequence reads that did not map to the reference, we conducted reference-free association study based on *k*-mers using the HAWK (Hitting Associations With *K*-mers) pipeline (39). It identifies *k*-mers with significantly different frequencies between two clusters, and then assembles significant *k*-mers into longer sequences. We ran HAWK between pairs of different clusters identified by admixture, and between pairs of races independent of their assignment to populations: race 1 vs. race 0, race 0 vs. all, race 1 vs. all, race 3 vs. all, race 3N vs. race 0, race 3N vs. all, race 3 vs. race 1 and 0, race 3N vs. race 1 and 0, race 3 and 3N vs race 0 and 1. Race 1 vs. race 0 was also analysed for samples from the Big Clonal cluster only. Races included in the analysis were race 0, 1, 3 and 3N. Numbers of other races were too low for meaningful comparisons. Significantly differentiated *k*-mers were mapped against the reference genome to test for their presence in the reference genome and to characterize the extent of clustering in some regions. Repetitive elements in windows with high number of *k*-mers mapped were searched with the protein-based RepeatMasking (87). Gene-rich or gene-poor regions were determined for windows with high numbers of mapped *k*-mers by counting genes in these regions. Remaining

unmapped assembled *k*-mers were compared against the NCBI non-redundant protein database using BLASTX to identify putative protein sequences.

Data Availability. Raw sequence data of the 121 isolates generated in this study is available in the European Nucleotide Archive (ENA) under the project ID PRJEB37432. Scripts for analysing the data can be downloaded at DOI: 10.5281/zenodo.4036236. Geographic and phenotypic information of the isolates is in the *SI Appendix* (Dataset S1).

ACKNOWLEDGMENTS. We thank Elisabeth Kokai-Kota for cultivating pathogen isolates, DNA extraction and sequencing preparation. We are grateful to Ana Galiano and Thomas Miedaner for comments on the manuscript. This work was funded by Deutsche Forschungsgemeinschaft (DFG) Priority program SPP1819 Rapid Evolutionary Adaptation - Potential and Constraints (Grant SCHM1354/11-1) to K. J. S. and SPP1590 Probabilistic Structures in Evolution (Grant FR3633/2-1) to F. F. The authors acknowledge the additional support by the state of Baden-Württemberg through bwHPC.

Bibliography

1. Papaïx J, Rimbaud L, Burdon JJ, Zhan J, Thrall PH (2017) Differential impact of landscape-scale strategies for crop cultivar deployment on disease dynamics, resistance durability and long-term evolutionary control. *Evolutionary Applications* 11(5):705–717.
2. Bebber DP, Ramotowski MAT, Gurr SJ (2013) Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*.
3. McDonald BA, Stukenbrock EH (2016) Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371(1709):20160026.
4. Islam MT, et al. (2016) Emergence of wheat blast in Bangladesh was caused by a South American lineage of Magnaporthe oryzae. *BMC Biology* 14(1).
5. Hubbard A, et al. (2015) Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology* 16(1).
6. Lamour KH, et al. (2012) Genome Sequencing and Mapping Reveal Loss of Heterozygosity as a Mechanism for Rapid Adaptation in the Vegetable Pathogen *Phytophthora capsici*. *Molecular Plant-Microbe Interactions* 25(10):1350–1360. 00000.
7. Raffaele S, et al. (2010) Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. *Science* 330(6010):1540–1543.
8. Dong S, Raffaele S, Kamoun S (2015) The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics & Development* 35:57–65. 00043.
9. Thordal-Christensen H, Birch PRJ, Spanu PD, Panstruga R (2018) Why did filamentous plant pathogens evolve the potential to secrete hundreds of effectors to enable disease?: Why so many effectors? *Molecular Plant Pathology* 19(4):781–785.
10. Frantzeskakis L, Kusch S, Panstruga R (2019) The need for speed: compartmentalized genome evolution in filamentous phytopathogens. *Molecular Plant Pathology* 20(1):3–7.
11. Gladieux P, et al. (2018) Coexistence of Multiple Endemic and Pandemic Lineages of the Rice Blast Pathogen. *mBio* 9(2):e01806–17.
12. Stam R, Sghyer H, Tellier A, Hess M, Huckelhoven R (2019) The current epidemic of the barley pathogen Ramularia collo-cygni derives from a recent population expansion and shows global admixture. *Phytopathology* pp. PHYTO–04–19–0117–R.
13. Latorre SM, et al. (2020) Recently expanded clonal lineages of the rice blast fungus display distinct patterns of presence/absence of effector genes. *bioRxiv*.
14. Yoshida K, et al. (2013) The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *Elife* 2:e00731.
15. McDonald BA, Linde C (2002) Pathogen Population Genetics, Evolutionary Potential, and Durable Resistance. *Annual Review of Phytopathology* 40(1):349–379.
16. Burdon JJ, Barrett LG, Rebetzke G, Thrall PH (2014) Guiding deployment of resistance in cereals using evolutionary principles. *Evolutionary Applications* 7(6):609–624.

17. Galiano-Carneiro AL, Miedaner T (2017) Genetics of Resistance and Pathogenicity in the Maize/Setosphaeria turcica Pathosystem and Implications for Breeding. *Frontiers in Plant Science* 8.
18. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ (2011) Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. *Proceedings of the National Academy of Sciences* 108(17):6893–6898.
19. Sartoria M, Nescia A, Formento Á, Etcheverry M (2015) Selección de microorganismos epifíticos de maíz como potenciales agentes debiocontrolde Exserohilum turcicum. *Revista Argentina de Microbiologia* 47(1):62–71.
20. Sartori M, et al. (2017) Efficacy of epiphytic bacteria to prevent northern leaf blight caused by Exserohilum turcicum in maize. *Revista Argentina de Microbiologia* 49(1):75–82.
21. Borchardt DS, Welz HG, Geiger HH (1998) Genetic Structure of Setosphaeria turcica Populations in Tropical and Temperate Climates. *Phytopathology* 88(4):322–329.
22. Welz HG, Geiger HH (2000) Genes for resistance to northern corn leaf blight in diverse maize populations. *Plant Breeding* 119(1):1–14.
23. Hanekamp, Hendrik (2016) Ph.D. Thesis (University of Göttingen, Göttingen).
24. Josse J, Husson F (2016) missMDA : A Package for Handling Missing Values in Multivariate Data Analysis. *Journal of Statistical Software* 70(1).
25. Ohm RA, et al. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen dothideomycetes fungi. *PLOS Pathogens* 8(12):1–26.
26. Condon BJ, et al. (2013) Comparative genome structure, secondary metabolite, and effector coding capacity across cochliobolus pathogens. *PLOS Genetics* 9(1):1–29.
27. Kuismin MO, Ahlinder J, Sillanpää MJ (2017) CONE: Community Oriented Network Estimation Is a Versatile Framework for Inferring Population Structure in Large-Scale Sequencing Data. *G3: Genes, Genomes, Genetics* 7(10):3359–3377.
28. Nelson MA (1996) Mating systems in ascomycetes: a romp in the sac. *Trends in Genetics* 12(2):69 – 74.
29. Turgeon BG (1998) Application of mating type gene technology to problems in fungal biology. *Annual Review of Phytopathology* 36(1):115–137. PMID: 15012495.
30. Kryazhimskiy S, Plotkin JB (2008) The population genetics of dn/ds. *PLoS genetics* 4(12):e1000304.
31. Akashi H, Osada N, Ohta T (2012) Weak selection and protein evolution. *Genetics* 192(1):15–31.
32. Ohta T (1973) Slightly deleterious mutant substitutions in evolution. *Nature* 246(5428):96–98.
33. Mideros SX, et al. (2018) Determinants of Virulence and In Vitro Development Colocalize on a Genetic Map of *Setosphaeria turcica*. *Phytopathology* 108(2):254–263.
34. Neher RA, Hallatschek O (2013) Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences USA* 110(2):437–442.
35. Steinrücken M, Birkner M, Blath J (2013) Analysis of dna sequence variation within marine species using beta-coalescents. *Theoretical Population Piology* 87:15–24.
36. Dutta A, Croll D, McDonald BA, Barrett LG (2020) Maintenance of variation in virulence and reproduction in populations of an agricultural plant pathogen. *Evolutionary Applications* n/a(n/a).
37. Tellier A, Lemaire C (2014) Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular ecology* 23(11):2637–2652.
38. Kass RE, Raftery AE (1995) Bayes factors. *Journal of the american statistical association* 90(430):773–795.
39. Rahman A, Hallgrímsdóttir I, Eisen M, Pachter L (2018) Association mapping from sequencing reads using k-mers. *eLife* 7:e32920.
40. Borchardt DS, Welz HG, Geiger HH (1998) Molecular marker analysis of European Setosphaeria turcica populations. *European Journal of Plant Pathology* 104(6):611–617.
41. Saunders IW, Tavaré S, Watterson G (1984) On the genealogy of nested subsamples from a haploid population. *Advances in Applied probability* 16(3):471–491.
42. Croll D, McDonald BA (2017) The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems. *Molecular Ecology* 26(7):2027–2040.

43. Freund F, Siri-Jégousse A (2020) The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis* p. 107055.
44. Koskela J, Wilke Berenguer M (2019) Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences* 311:1–12.
45. Wang M, et al. (2018) Two-way mixed-effects methods for joint association analysis using both host and pathogen genomes. *Proceedings of the National Academy of Sciences* 115(24):E5440–E5449.
46. Nelson R, Wiesner-Hanks T, Wisser R, Balint-Kurti P (2018) Navigating complexity to breed disease-resistant crops. *Nature Reviews Genetics* 19(1):21–33.
47. Baym M, et al. (2015) Inexpensive Multiplexed Library Preparation for Megabase-Sized Genomes. *PLOS ONE* 10(5):e0128036.
48. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for illumina sequence data. *Bioinformatics* 30(15):2114–2120.
49. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 25(14):1754–1760.
50. Kersey PJ, et al. (2016) Ensembl genomes 2016: more genomes, more complexity. *Nucleic Acids Research* 44(D1):D574–D580.
51. McKenna A, et al. (2010) The genome analysis toolkit: A mapreduce framework for analyzing next-generation dna sequencing data. *Genome Research* 20(9):1297–1303.
52. Okonechnikov K, Conesa A, García-Alcalde F (2016) Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* 32(2):292–294.
53. Li H, et al. (2009) The sequence alignment/map format and samtools. *Bioinformatics* 25(16):2078–2079.
54. Li H (2011) A statistical framework for snp calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27(21):2987–2993.
55. Narasimhan V, et al. (2016) Bcftools/roh: a hidden markov model approach for detecting autozygosity from next-generation sequencing data. *Bioinformatics* 32(11):1749–1751.
56. Danecek P, et al. (2011) The variant call format and vcftools. *Bioinformatics* 27(15):2156–2158.
57. Josse J, Husson F (2016) missmda: A package for handling missing values in multivariate data analysis. *Journal of Statistical Software, Articles* 70(1):1–31.
58. Blanchette M, et al. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Research* 14(4):708–715.
59. Dutheil JY, Gaillard S, Stukenbrock EH (2014) Maffilter: a highly flexible and extensible multiple genome alignment files processor. *BMC Genomics* 15(1):53.
60. Grigoriev IV, et al. (2012) The genome portal of the department of energy joint genome institute. *Nucleic Acids Research* 40(D1):D26–D32.
61. Nordberg H, et al. (2014) The genome portal of the department of energy joint genome institute: 2014 updates. *Nucleic Acids Research* 42(D1):D26–D31.
62. Chung CL, Jamann T, Longfellow J, Nelson R (2010) Characterization and fine-mapping of a resistance locus for northern leaf blight in maize bin 8.06. *Theoretical and Applied Genetics* 121(2):205–227.
63. Li D, Liu CM, Luo R, Sadakane K, Lam TW (2015) Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics* 31(10):1674–1676.
64. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology* 215(3):403–410.
65. Camacho C, et al. (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10(1):421.
66. Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19(9):1655–1664.
67. Jombart T (2008) adegenet: a r package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405.
68. Jombart T, Ahmed I (2011) adegenet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics*.
69. Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources* 15(5):1179–1191.

673 70. Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* 23(2):254–267.
675 71. Cliff AD, Ord JK, Cliff AD (1981) *Spatial processes: models & applications*. (Pion London), p. 266 p. :.
676 72. Bivand R, Wong DWS (2018) Comparing implementations of global and local indicators of spatial association. *TEST* 27(3):716–748.
677 73. Bivand RS, Pebesma E, Gomez-Rubio V (2013) *Applied spatial data analysis with R, Second edition*. (Springer, NY).
678 74. Giraudoux P (2018) *pgirmess: Spatial Analysis and Data Mining for Field Ecologists*. R package version 1.6.9.
679 75. Pfeifer B, Wittelsbürger U, Ramos-Onsins SE, Lercher MJ (2014) Popgenome: an efficient swiss army knife for population genomic analyses in r. *Molecular biology and evolution* 31(7):1929–1936.
681 76. Cingolani P, et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff. *Fly* 6(2):80–92.
683 77. Nelson CW, Moncla LH, Hughes AL (2015) SNPGenie: estimating evolutionary parameters to detect natural selection using pooled next-generation sequencing data. *Bioinformatics* 31(22):3709–3711.
685 78. Sommerhalder RJ, McDonald BA, Zhan J (2006) The Frequencies and Spatial Distribution of Mating Types in *Stagonospora nodorum* are Consistent with Recurring Sexual Reproduction. *Phytopathology* 96(3):234–239.
687 79. Milgroom MG (1996) Recombination and the Multilocus Structure of Fungal Populations. *Annual Review of Phytopathology* 34(1):457–477.
689 80. Bruen TC, Philippe H, Bryant D (2006) A simple and robust statistical test for detecting the presence of recombination. *Genetics* 172(4):2665–2681.
691 81. Kamvar ZN, Tabima JF, Grünwald NJ (2014) Poppr: an r package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2:e281.
693 82. Pudlo P, et al. (2015) Reliable abc model choice via random forests. *Bioinformatics* 32(6):859–866.
694 83. Raynal L, et al. (2019) Abc random forests for bayesian parameter inference. *Bioinformatics* 35(10):1720–1728.
695 84. Freund F, Siri-Jégousse A (2020) The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis* p. 107055.
697 85. Menardo F, Gagneux S, Freund F (2020) Multiple merger genealogies in outbreaks of Mycobacterium tuberculosis. *Molecular Biology and Evolution*. msaa179.
699 86. Bouckaert R, et al. (2014) Beast 2: a software platform for bayesian evolutionary analysis. *PLoS computational biology* 10(4):e1003537.
701 87. A.F.A. Smit RHPG (2013-2015) Repeatmasker open-4.0.