



OPEN ACCESS

EDITED BY

Chongjing Xia,
Southwest University of Science and
Technology, China

REVIEWED BY

Abhay K. Pandey,
North Bengal Regional R & D Center,
India
Vinay Kumar,
ICAR-National Institute of Biotic Stress
Management, India

*CORRESPONDENCE

Martin I. Chilvers,
✉ chilvers@msu.edu

RECEIVED 21 November 2022

ACCEPTED 20 April 2023

PUBLISHED 07 June 2023

CITATION

Ortiz V, Chang H-X, Sang H, Jacobs S,
Malvick DK, Baird R, Mathew FM,
Estévez de Jensen C, Wise KA,
Mosquera GM and Chilvers MI (2023),
Population genomic analysis reveals
geographic structure and climatic
diversification for *Macrophomina
phaseolina* isolated from soybean and dry
bean across the United States, Puerto
Rico, and Colombia.
Front. Genet. 14:1103969.
doi: 10.3389/fgene.2023.1103969

COPYRIGHT

© 2023 Ortiz, Chang, Sang, Jacobs,
Malvick, Baird, Mathew, Estévez de
Jensen, Wise, Mosquera and Chilvers.
This is an open-access article distributed
under the terms of the [Creative
Commons Attribution License \(CC BY\)](#).
The use, distribution or reproduction in
other forums is permitted, provided the
original author(s) and the copyright
owner(s) are credited and that the original
publication in this journal is cited, in
accordance with accepted academic
practice. No use, distribution or
reproduction is permitted which does not
comply with these terms.

Population genomic analysis reveals geographic structure and climatic diversification for *Macrophomina phaseolina* isolated from soybean and dry bean across the United States, Puerto Rico, and Colombia

Viviana Ortiz^{1,2}, Hao-Xun Chang³, Hyunkyu Sang⁴,
Janette Jacobs¹, Dean K. Malvick⁵, Richard Baird⁶,
Febina M. Mathew⁷, Consuelo Estévez de Jensen⁸,
Kiersten A. Wise⁹, Gloria M. Mosquera¹⁰ and Martin I. Chilvers^{1,2*}

¹Department of Plant, Soil and Microbial Sciences, College of Agriculture and Natural Resources, Michigan State University, East Lansing, MI, United States, ²Ecology, Evolution and Behavior Program, Michigan State University, East Lansing, MI, United States, ³Department of Plant Pathology and Microbiology, National Taiwan University, Taipei, Taiwan, ⁴Department of Integrative Food, Bioscience and Biotechnology, Chonnam National University, Gwangju, Republic of Korea, ⁵Department of Plant Pathology, University of Minnesota, St. Paul, MN, United States, ⁶BCH-EPP Department, Mississippi State University, Mississippi State, MS, United States, ⁷Department of Plant Pathology, North Dakota State University, Fargo, ND, United States, ⁸Department of Agroenvironmental Sciences, University of Puerto Rico, Mayagüez, PR, United States, ⁹Department of Plant Pathology, College of Agriculture, Food and Environment, University of Kentucky, Princeton, KY, United States, ¹⁰Plant Pathology, Crops for Nutrition and Health, International Center for Tropical Agriculture (CIAT), The Americas Hub, Palmira, Colombia

Macrophomina phaseolina causes charcoal rot, which can significantly reduce yield and seed quality of soybean and dry bean resulting from primarily environmental stressors. Although charcoal rot has been recognized as a warm climate-driven disease of increasing concern under global climate change, knowledge regarding population genetics and climatic variables contributing to the genetic diversity of *M. phaseolina* is limited. This study conducted genome sequencing for 95 *M. phaseolina* isolates from soybean and dry bean across the continental United States, Puerto Rico, and Colombia. Inference on the population structure using 76,981 single nucleotide polymorphisms (SNPs) revealed that the isolates exhibited a discrete genetic clustering at the continental level and a continuous genetic differentiation regionally. A majority of isolates from the United States (96%) grouped in a clade with a predominantly clonal genetic structure, while 88% of Puerto Rican and Colombian isolates from dry bean were assigned to a separate clade with higher genetic diversity. A redundancy analysis (RDA) was used to estimate the contributions of climate and spatial structure to genomic variation (11,421 unlinked SNPs). Climate significantly contributed to genomic variation at a continental level with temperature seasonality explaining the most variation while precipitation of warmest quarter explaining the most when spatial structure was accounted for. The loci significantly associated with multivariate climate were found closely to the genes related to fungal stress responses, including transmembrane transport,

glycoside hydrolase activity and a heat-shock protein, which may mediate climatic adaptation for *M. phaseolina*. On the contrary, limited genome-wide differentiation among populations by hosts was observed. These findings highlight the importance of population genetics and identify candidate genes of *M. phaseolina* that can be used to elucidate the molecular mechanisms that underly climatic adaptation to the changing climate.

KEYWORDS

charcoal rot, climate change, genotype–environment associations, pathogen adaptation, redundancy analysis, phylogenomics, landscape genetics

1 Introduction

Understanding the genetic diversity of plant pathogen populations and the factors influencing it allow for inferences about their evolutionary potential and identification of the molecular bases of adaptation. Plant pathogens are often genetically structured in different agricultural landscapes as a result of geographic and environmental differences (Gladieux et al., 2014; McDonald and Stukenbrock, 2016). Among different environments, agroecosystems provide remarkable conditions for rapid adaptation of plant-pathogenic fungi due to various abiotic and biotic factors such as genetic crop uniformity of monocultures, the prevalent occurrence of human-mediated migration (Wingfield et al., 2015; Crous et al., 2017), and intrinsic characteristics of fungi such as their mode of reproduction (McDonald and Stukenbrock, 2016). These factors are strong drivers of genomic divergence and adaptation in plant pathogenic fungi (Stukenbrock et al., 2011; Savolainen et al., 2013; Croll and McDonald, 2016). However, characterizing how selective pressures of abiotic and biotic factors contribute to population genetics of plant-pathogenic fungi remains challenging.

Macrophomina phaseolina is a seed- and soil-borne fungal pathogen that infects more than 500 host species (Batista et al., 2021), and causes dry root rot, seedling blight, crown rot and charcoal rot in many important economic and subsistence crops worldwide, including soybean (*Glycine max*) and dry bean (*Phaseolus vulgaris*) (Dhingra and Sinclair, 1978; Yang and Navi, 2005; Saleh et al., 2010; Jacobs et al., 2019). During host infection, *M. phaseolina* invades the xylem preventing water uptake, causing wilting and premature plant death (Mengistu et al., 2007; Romero Luna et al., 2017). These symptoms can develop rapidly causing extensive yield loss and grain or seed quality reduction (Smith and Carvil, 1997). Charcoal rot of soybean has been ranked seventh out of 25 pests and pathogens causing global yield losses (Savary et al., 2019), with the potential for yield reductions within individual fields of up to 50% (Wrather et al., 2001). In the United States, charcoal rot ranked among the top seven most destructive diseases with economic losses totaling 220 billion dollars from 2010 to 2014 (Allen et al., 2017). Disease development is favored by hot and dry conditions (Dhingra and Sinclair, 1974), with colonization in the soybean and dry bean tap root and lower stem being greatest under high temperatures (28°C–35°C) and low soil moisture (Dhingra and Sinclair, 1974; Meyer and Sinclair, 1974; Kendig et al., 2000; Mengistu et al., 2011; Reznikov et al., 2018).

Macrophomina phaseolina is haploid, reproduces asexually, and overwinters in soil and crop residue as abundant, melanized microsclerotia that serve as the primary inoculum to initiate infection in subsequent seasons (Gupta et al., 2012; Islam et al., 2012). Pycnidia are occasionally produced on soybean and other host plants, however, their epidemiological significance has yet to be fully defined (Knox-Davies, 1965; Dhingra and Sinclair, 1978; Mihail and Taylor, 1995). Depending on environmental conditions, *M. phaseolina* may survive as microsclerotia in soil for up to 15 years (Short et al., 1980; Baird et al., 2003), and for up to 3 years as microsclerotia in symptomatic seeds or as mycelium in asymptomatic seeds (Hartman et al., 2015). To date, no clonal lineages or pathotypes have been identified for *M. phaseolina*, despite reports of within-species variation in morphology and pathogenicity (Dhingra and Sinclair, 1973; Dhingra and Sinclair, 1978; Sexton et al., 2016). Population genetic studies based on microsatellite markers of isolates representing different geographic regions and hosts across the United States have found moderate to high genetic diversity and mixed evidence of population structure by host or geography. Although considerable efforts have been focused on ascertaining host specialization, it is generally concluded that there is no strong evidence of this specificity, in which isolates from one plant species can often cause disease in other plant species (Su et al., 2001; Zveibil et al., 2012; Romero Luna et al., 2017). Nevertheless, genetic similarity of isolates according to host and United States regions and some degree of host preference have been noted (Su et al., 2001; Jana et al., 2005; Baird et al., 2010; Saleh et al., 2010; Arias et al., 2011). Notably, a group of *M. phaseolina* isolates obtained from strawberry in California were found to form a species-specific cluster, exhibiting strong host preference for strawberry over other hosts around California (Koike et al., 2016; Burkhardt et al., 2019).

Studying population genetics using statistical methods that leverage genomic, geographic and environmental data can account for continuous and discrete genetic variation and provide insights into the genetic basis underlying environmental adaptation (Hoban et al., 2016; Bontrager and Angert, 2018; Bradburd et al., 2018). These approaches may be used to identify environmental factors driving selection and provide an understanding of how and why pathogen populations vary across space. Population genomics and genotype–environment associations have been applied in numerous studies to resolve the basis of rapid adaptation and identify candidate adaptive loci associated with environmental variation (Lasky et al., 2012; Forester et al., 2018; Xuereb et al., 2018; Gibson and Moyle, 2020; Capblancq and Forester, 2021). However, characterizing population structure and

unravelling the effects of continuous or discrete processes on the genetic differentiation remains challenging for many plant-pathogenic fungi.

A major challenge arises because continuous geographic differentiation (e.g., isolation by distance or climatic variation along a gradient) can be confounded with discrete processes such as admixture and long-distance migration (human-mediated migration) which are commonly observed in plant pathogens (Wingfield et al., 2015; Crous et al., 2017; Tabima et al., 2019; LeBlanc et al., 2021). In addition, collinearity between spatial and environmental variables makes it difficult to elucidate to what extent geographic and environmental differences may be contributing to genetic differentiation. To address these issues, multivariate statistical methods, specifically redundancy analysis (RDA), have been increasingly used to disentangle the effects of environmental factors in shaping genetic variation. RDA is a type of constrained ordination in which a set of SNPs are modeled as responses in a function of combinations of environmental predictors. Because of its ability to evaluate many loci simultaneously, RDA has been found to be superior to traditional mixed-models associations methods in identifying weak, multilocus selection (Forester et al., 2018), suggestive of polygenic adaptation. Furthermore, partial RDA models, in which covariables can be included, has been used to account for underlying population structure in the identification of loci associated with environmental factors for climate adaptation in a variety of systems including plant and animal species (Lasky et al., 2012; Forester et al., 2018; Xuereb et al., 2018; Gibson and Moyle, 2020; Capblancq and Forester, 2021).

Climate fluctuation and temperature in particular are important abiotic factors leading to local adaptation of plant-associated fungi (Savolainen et al., 2013; Croll and McDonald, 2016), especially in species occupying spatially and climatically heterogeneous environments (Ellison et al., 2011; Branco et al., 2015; Branco et al., 2016; Fitzpatrick and Keller, 2015). *Macrophomina phaseolina* is recognized for its different ecological roles as an endophyte, saprotroph, and latent or opportunistic pathogen with broad geographic distribution (Dhingra and Sinclair, 1974; Slippers and Wingfield, 2007; Slippers et al., 2013; Parsa et al., 2016; Crous et al., 2017). Worldwide diseases caused by *M. phaseolina* have re-emerged in recent decades, with outbreaks occurring in tropical and subtropical regions as well as temperate regions (Leyva-Mir et al., 2015; Casano et al., 2018; Koehler and Shew, 2018; Meena et al., 2018; Nishad et al., 2018; Tančić Živanov et al., 2018; Wang et al., 2020). In the United States, charcoal rot of soybean has been a primary issue in southern and central states historically. However, charcoal rot has been noted to occur in northern states such as Wisconsin, New York, Minnesota, and Michigan (Yang and Navi, 2005; Bradley and del Río, 2007; ElAraby et al., 2007; Baird et al., 2010; Cummings and Bergstrom, 2013) and currently it is a consistent threat to soybean production across southern and northern states (Bradley et al., 2021; Roth et al., 2021). Although many factors may influence disease incidence, greater disease and yield losses have been observed in years with high temperature and low soil moisture (Bradley and Allen, 2014; Allen et al., 2017). A recent study concluded that *M. phaseolina* isolates were regionally adapted when comparing isolates from the northern and southern United States states (Sexton et al., 2016). Investigations in the context of species within the Botryosphaeriaceae family suggest

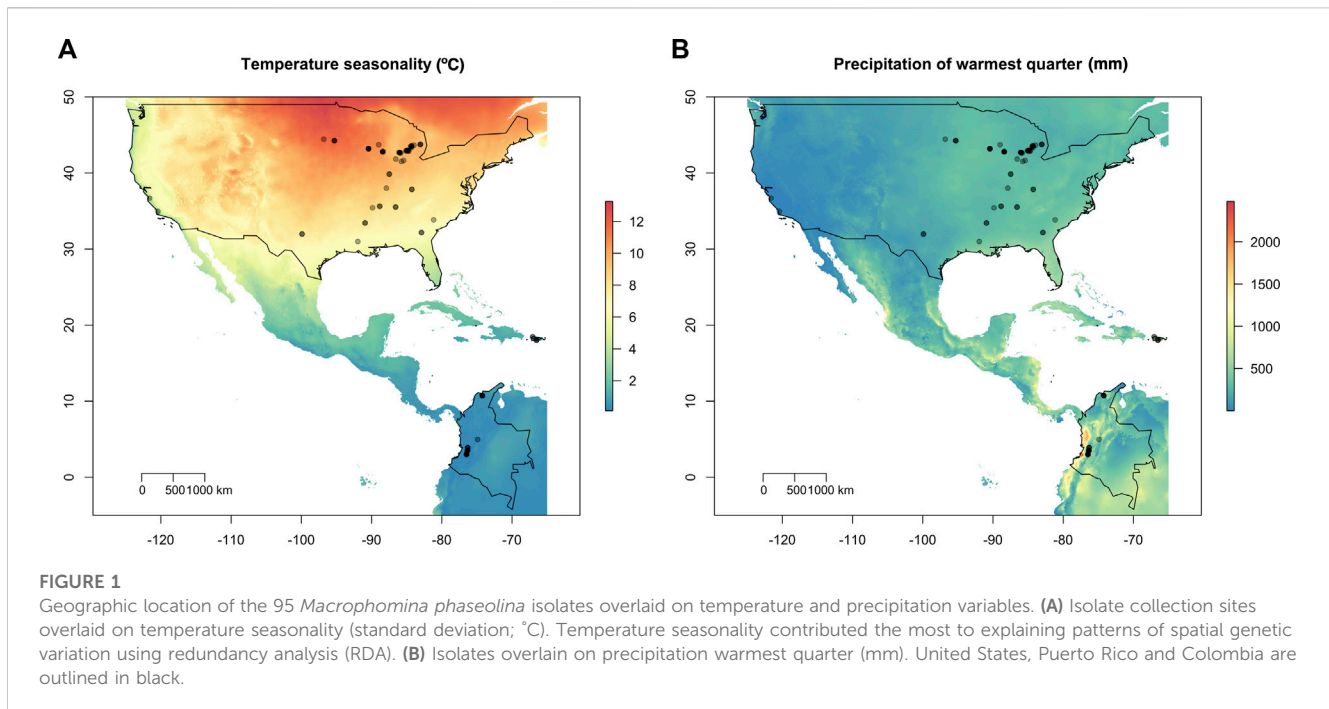
that geographical distribution and host affinity dynamics in *M. phaseolina* are strongly influenced by climate due to its broad host range and ecologically diverse roles (Slippers and Wingfield, 2007; Batista et al., 2021), while recent reviews have indicated that the impact of charcoal rot on crop losses may intensify in the face of global warming (Basandrai et al., 2021; Pandey and Basandrai, 2021; Cohen et al., 2022). These factors, together with predicted increases in temperature and extreme rainfall variation as projected in the climatic change models (IPCC, 2022), make it critical to better understand genetic structure and climatic factors as potential selection agents of *M. phaseolina*.

The broad geographic distribution and population dynamics of *M. phaseolina* suggest that populations in the continental United States, Puerto Rico and Colombia might have been influenced by a complex environmental and agricultural landscape and may be structured and differentially adapted at a continental or regional level. In the present study, the first aim was to better understand the genetic structure in *M. phaseolina* populations isolated from soybean and dry bean across the United States, Puerto Rico and Colombia using genome-wide single nucleotide polymorphisms (SNPs). Specifically, the contribution of discrete vs continuous genetic differentiation was assessed by inferring *M. phaseolina* genetic groups while accounting for geographic isolation by distance. The hypotheses tested were *M. phaseolina* populations differentiated i) between countries and ii) between hosts within the United States. The second aim was to investigate whether climatic variables contribute to patterns of adaptive genetic variation in *M. phaseolina*. Using RDA, the hypotheses were i) specific climatic variables contribute to genetic variation, ii) climatic variables independently contribute to patterns of genetic variation when accounting for underlying spatial and population structure, and iii) loci in strong association with multivariate climate can be identified and have roles in driving local adaptation to climate.

2 Results

2.1 Whole-genome sequencing for 95 *Macrophomina phaseolina* isolates

Whole-genome resequencing was completed for 95 *M. phaseolina* isolates collected across the United States, Puerto Rico, and Colombia, including 52 soybean isolates, 40 dry bean isolates, two strawberry isolates, and one Ethiopian mustard isolate (Figure 1; Supplementary Table S1). Sequence coverage varied across individual isolates from 5X to 85X, across 93% of the Macpha1 reference genome (JGI MycoCosm, MPI-SDFR-AT-0080 v1.0). A total of 2.8 million SNPs were identified across all isolates, and a mean read depth (DP) of 12X was obtained for all SNPs after filtering. Most SNPs had a mapping quality (MQ) value equal to 60 (94%) and SNPs with MQ values <60 were removed. The distribution of missing data across the isolates and across the variants was even, with most individuals representing similar missing data (0%–0.006%), and all variants containing missing data were removed. The final data set contained 76,981 high-quality biallelic SNPs in all isolates, and the data set was retained for all analyses.



2.2 Phylogenomics differentiated 95 isolates into two main clades of the United States and Colombian–Puerto Rican origins

To infer the genetic similarity in *M. phaseolina* isolates across the continental United States, Colombia and Puerto Rico, a maximum-likelihood (ML) phylogenetic tree based on the 76,981 SNPs was constructed. Five genetic clusters were identified across the United States ($n = 3$), Colombia and Puerto Rico ($n = 2$). Furthermore, a pattern of hierarchical structure differentiating the United States and Colombian–Puerto Rican isolates was observed. The ML tree provided strong support (100% bootstrap) for two main clades, hereafter referred to as US and COLPR, and five well-supported clades within the main clades (Figure 2A). The United States isolates M11–12 and M13–26 from California, and TN501 from Louisiana clustered in the COLPR clade, while the Colombian isolates Mph-22, Mph-23, and Mph-49 fell within the US clade (Figure 2A). Other than these six isolates, all isolates from the United States were placed in the US clade, and all isolates from Colombia and Puerto Rico were grouped in the COLPR clade.

There were three subclades (US1A, US1B and US2) within the US clade and two subclades (COLPR1 and COLPR2) within the COLPR clade. The PCA clustered isolates in five distinct groups in agreement with phylogenetic analysis, with little evidence of within group differentiation (Figure 2B). The first PC explains most of the variance (50.6%) and separates out isolates in the US clade from the isolates in the COLPR clade, while the second PC explains 15.5% of the variance dividing isolates into the five groups in the phylogenetic analysis (Figure 2B). An exception was isolate MP258 which in the PCA was grouped in US1B instead of US1A. Since the phylogenetic and PCA clustering

revealed essentially the same hierarchical groupings, they were named genetic clusters US1A, US1B, US2, COLPR1 and COLPR2.

US1A isolates represented the predominant group in the United States, with most isolates collected in the East North Central and Central regions in the states of Michigan (29), followed by Wisconsin (11), Indiana (5), Tennessee (5) and Kentucky (2). Cluster US1B was represented by isolates from Mississippi (2) and South Carolina (1). US2 isolates represented the second largest group in the United States and were mostly collected in the Northern Great Plains [Minnesota (3), South Dakota (1)] and South [Texas (2) and Georgia (2)] regions. Also, within this cluster were isolates from Wisconsin (1), Michigan (4), and Kentucky (1). In contrast, the COLPR1 cluster grouped most isolates from Colombia (11) and Puerto Rico (4) while COLPR2 grouped isolates from Colombia (5), one isolate from Puerto Rico, and three isolates from the United States. No evidence of population structure by states was found, which indicated that states do not represent genetic groups and *M. phaseolina* is genetically structured at a broader subcontinental regional extent.

A ML phylogeny rooted with the *M. phaseolina* Macpha1 reference genome was reconstructed using the set of high-quality SNPs. The Macpha1 reference genome was considered as a suitable outgroup based on its European and *Arabidopsis thaliana* origin. The phylogenetic reconstruction with Macpha1 as a root revealed the COLPR2 clade as an outgroup to all other clades, while the United States clades were reconstructed as terminal clades (Supplementary Figure S1). The topology of the rooted ML phylogeny indicated the COLPR clades as more diverse than the major US terminal clades (US1A and US2). This higher diversity in COLPR clades was indicated by longer average branch length than in the US clades, representing a higher average number of substitutions per site.

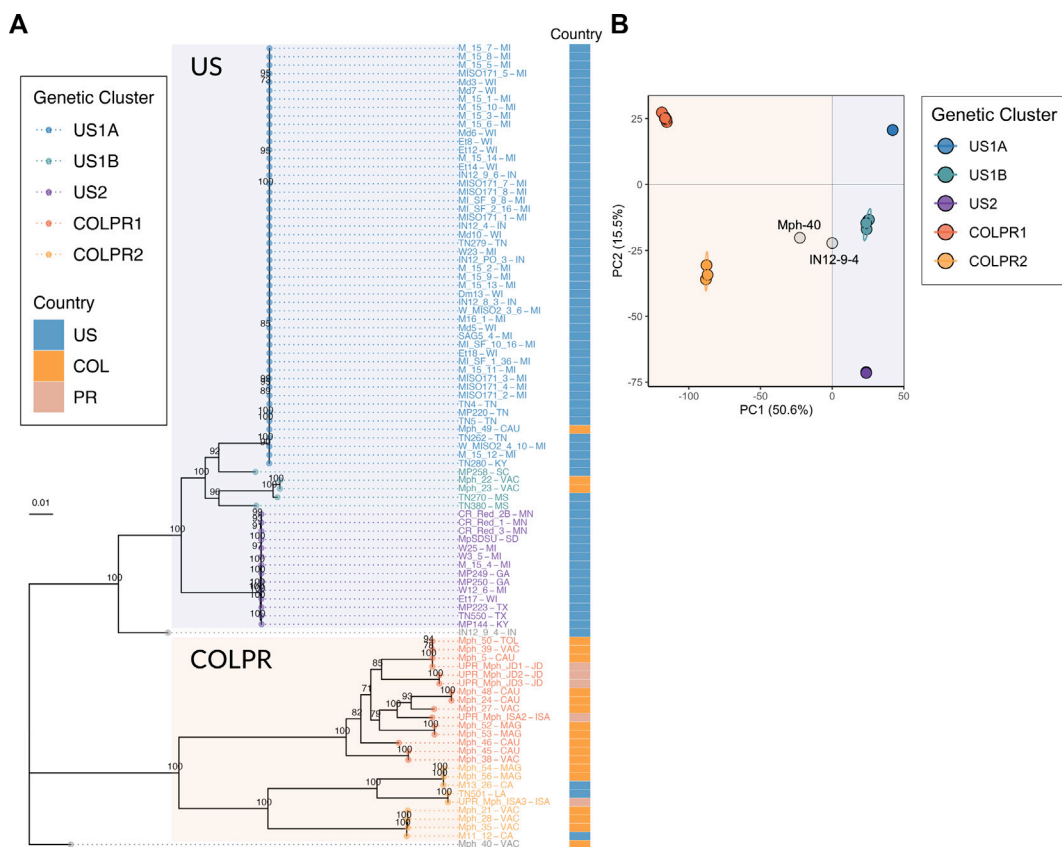


FIGURE 2

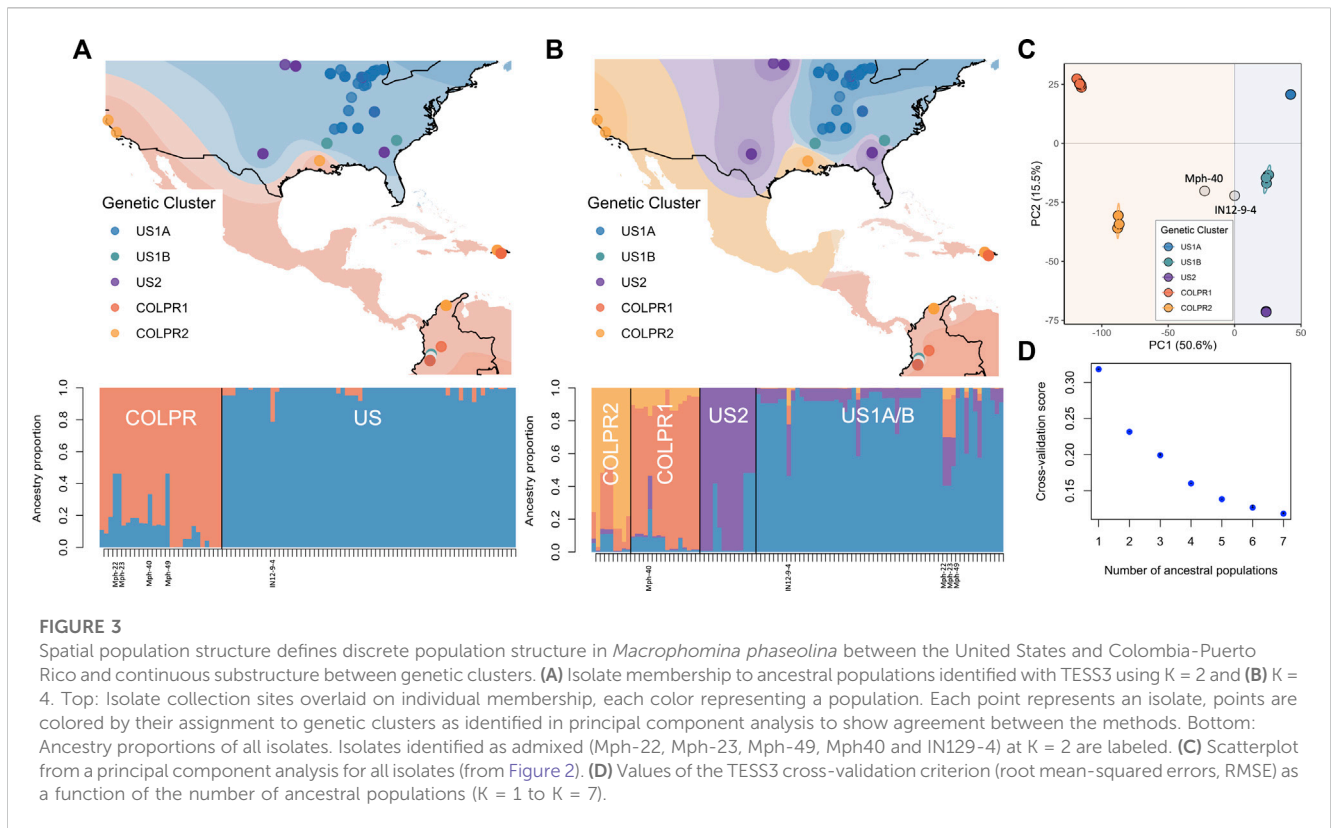
Population structure of *Macrophomina phaseolina* in the United States, Colombia and Puerto Rico reveals five genetic clusters in a pattern of hierarchical structure. (A) Maximum-likelihood phylogeny reconstructed using 76,981 high-quality SNPs. Bootstrap support values over 70 are shown at nodes. Bootstrapping converged after 400 replicates. Colored tips represent the genetic cluster for each isolate as defined by principal components analysis. The two main clades, US and COLPR, are highlighted by rectangular shading. The country of collection for each isolate is denoted by colored squares at the right bar. (B) Scatterplot from a principal component analysis based on the two first PCs (the eigenvectors of the SNP dataset) for all isolates. Points are colored by membership in the five genetic clusters. Isolate names include states/municipalities codes: CA: California, CAU: Cauca, GA: Georgia, IN: Indiana, ISA: Isabela, JD: Juana Diaz, KY: Kentucky, LA: Louisiana, MAG: Magdalena, MI: Michigan, MN: Minnesota, MS: Mississippi, SC: South Carolina, SD: South Dakota, TN: Tennessee, TOL: Tolima, TX: Texas, VAC: Valle del Cauca, WI: Wisconsin. Country codes: US: United States, COL: Colombia and PR: Puerto Rico.

Differences in diversity can also be inferred from the PCA clustering. In PC space, 50 isolates in US1A and 14 isolates in US2 genetic clusters clustered effectively on top of each other, while isolates in US1B, COLPR1 and COLPR2, although projected near each other, clustered distinctively more dispersed (Figure 2B). The placement of COLPR genetic clusters and their higher diversity as compared to US genetic clusters indicates them as potential sources to the US clusters.

To test the relatedness of *M. phaseolina* isolates from soybean and dry bean in United States, the host information was mapped to the ML tree (Supplementary Figure S1A). Generally, isolates that shared a common host did not cluster within genetic clusters in the United States. Isolates collected from soybean and dry bean grouped together in the two larger United States genetic clusters (US1A and US2; Supplementary Figure S1A). This lack of structure was further supported in a PCA showing overlapping ellipses representing 95% of the isolates from each of the hosts (Supplementary Figure S2).

2.3 Spatial population structure defines discrete population structure in *Macrophomina phaseolina* between the United States and Colombia-Puerto Rico and continuous substructure between genetic clusters within US and COLPR clades

To infer the number of distinct genetic groups in *M. phaseolina* while accounting for continuous geographic differentiation, spatial analysis of population structure was conducted using a Bayesian (conStruct) and a model-free matrix factorization (TESS3) framework. Spatial analysis of population structure incorporates geographic distance in the estimation of ancestry coefficients (the proportion of individual isolate’s genome originating from the ancestral genetic group, K). The genetic structure of the 95 isolates was explained better by a spatial model of admixture between discrete genetic groups, where isolation by distance was accounted for



rather than the non-spatial model. This was indicated by the increase in predictive accuracy in the conStruct spatial models for all tested values of K (referred hereafter as layers in conStruct framework; Supplementary Figure S3B). This suggests that isolation by distance or climatic gradients likely play a role in shaping patterns of genetic variation in the sampled isolates.

Spatial population structure description using TESS3 returned the greatest decrease in root mean-squared errors at $K = 2$ (0.087, from 0.318 at $K = 1$ to 0.232 at $K = 2$; Figure 3D) and detected the US and COLPR clades. At $K = 2$, TESS3 spatial estimation strongly assigned 95% of isolates to a single ancestral population (ancestry proportion $Q > 0.8$; Figure 3A). All isolates in the US clade, except for the three isolates collected in Colombia, were identified as being derived from a single ancestral population (represented by blue; Figure 3A, bottom). Likewise, all COLPR isolates are estimated to have a majority component of ancestry from a single source population (represented by orange; Figure 3A, bottom) including the three isolates collected in the United States (M11–12 and M13–26 from California, and TN501 from Louisiana). The three isolates collected in Colombia grouping in the United States clade (Mph-22, Mph-23 and Mph-49) were identified as admixed (i.e., to have ancestry from more than one population instead of drawing ancestry mostly [$Q > 0.8$] from a single ancestral population) between the two ancestral groups (Figure 3A, bottom) as well as the two isolates (IN12-9-4 from Indiana and Mph-40 from Colombia) placed outside the supported clusters in the ML tree and PCA. At $K = 4$, further substructure was detected that generally reflect the genetic

clusters within the US and COLPR clades; except that an ancestral population for US1B isolates was not inferred (Figure 3B). The decrease in root mean-squared errors at $K = 4$ (0.04; from 0.20 at $K = 3$ to 0.16 at $K = 4$; Figure 3D) was the second largest value after that at $K = 2$, reflecting the hierarchical structure observed in previous analyses. However, although isolates in each genetic cluster (except US1B) were inferred as drawing the most ancestry from their own ancestral population, only 76% of isolates had an ancestry proportion (Q) > 0.80 to a single ancestral population (Figure 3B, bottom), demonstrating weaker assignments than those at $K = 2$.

Consistently, the results from conStruct spatial model with $K = 2$ returned the greatest increase in predictive accuracy and primarily partitioned the isolates in two main groups mostly in line with US and COLPR clades (Supplementary Figure S3A). Based on cross-validation results, the predictive accuracy increased with increasing values of K (Supplementary Figure S3B), however additional layers beyond $K = 2$ contribute little to total covariance (Supplementary Figure S3C). Therefore, supporting two discrete ancestral populations while population substructure can be explained by continuous genetic differentiation.

Taken together conStruct and TESS3 results supported two discrete genetic groups for the US and COLPR main clades and suggested that most isolates within US and COLPR clades can be better described to have ancestry mainly from each single ancestral population. It may therefore be reasonable that the evolutionary processes leading to divergence between genetic clusters within the US (US1A, US1B, US2) and COLPR

TABLE 1 Summary statistics for genetic diversity of *Macrophomina phaseolina* clades and genetic clusters. N is number of isolates (sample size); MLG is number of observed multilocus genotypes; eMLG is the number of expected MLG at a sample size of 25 for clades and 5 for genetic clusters based on rarefaction. MLL is number of observed multilocus lineages by population using a bitwise cutoff distance of 0.0001; CF is clonal fraction (1 - (MLL/N)). Clone corrected values are shown and indicated by asterisks for indices of genotypic diversity: Shannon-Wiener Index (H*), Stoddart and Taylor's Index (G*), Simpson's index (lambda*) and evenness (E5*).

Clade, genetic cluster	N	Gene diversity (He)	Median pairwise genetic distance	MLG	eMLG	MLL	eMLL	CF	H*	G*	Lambda*	E5*
US	70	0.0680	0.0890	54	21.75	19	10.70	0.73	1.990	5.53	8	0.702
US1A	50	0.0001	0.0002	34	4.71	6	2.87	0.88	0.935	2.48	0.561	0.883
US1B	5	0.0540	0.0900	5	5	4	4	0.20	1.332	3.57	0.720	0.922
US2	14	0.0010	0.0008	14	5	8	4.26	0.43	1.4	3.99	0.738	0.950
COLPR	25	0.2360	0.3320	25	25	15	15	0.40	2.640	13.30	0.925	0.942
COLPR1	15	0.0840	0.0960	15	5	9	4.43	0.40	1.456	4.24	0.756	0.962
COLPR2	9	0.1510	0.2650	9	5	5	3.73	0.44	1.242	3.35	0.688	0.924

Note: Summary statistics were calculated using the clone-corrected data at 79 MLGs

(COLPR1 COLPR2) clades were associated to isolation by distance or climatic differences rather than different discrete ancestry.

2.4 Genetic diversity and differentiation between the US and COLPR clades and genetic clusters of *Macrophomina phaseolina*

To examine genome-wide diversity of *M. phaseolina* within and among clades and genetic clusters, we estimated gene diversity (He) and median pairwise genetic distance for each of the clades and genetic clusters. Pairwise genetic distance showed that COLPR isolates had greater genetic distances among isolates than those in the US clade, with a gene diversity (He) significantly higher in the COLPR clade (0.236) than the US clade (0.068; Table 1) (Hs.test, $p = 0.002$). Among clusters, the COLPR2 cluster has the highest genetic diversity, considering both gene diversity and pairwise genetic distance, followed by COLPR1, US1B, US2, and the US1A cluster has the lowest values (Table 1). The higher genetic distance among isolates in the US1B cluster as compared to other US clusters, likely reflects that the cluster is only represented by five isolates of which two were collected in Mississippi, two in Colombia and one in South Carolina.

To evaluate genotypic diversity both in terms of genotypic richness (the number of observed genotypes) and evenness of distribution of genotypes, the number of multilocus genotypes (MLG) was calculated for each clade and genetic cluster. A MLG was defined as a unique combination of SNPs. Given the large number of 76,981 SNPs and genotyping error rate from NGS data, it is unlikely that a true clone will be represented by an MLG. Thus, to better represent clones, closely related genotypes were collapsed into multilocus lineages (MLLs) based on a Prevosti's genetic distance threshold of 0.0001 (8 SNPs). Of the 95 isolates, 79 had unique genotypes (MLGs) corresponding to 34 MLLs (Table 1). eMLG and eMLL are the number of expected MLGs and MLLs based on rarefaction at the lowest common sample size between clades and genetic clusters and were used to allow comparisons across them given their unequal sample sizes.

Genotypic richness was highest in the COLPR clade (15 eMLLs) as compared to the US clade (10.7 eMLLs). Among genetic clusters, the COLPR1 cluster had the highest number of eMLLs, followed by US2, US1B, COLPR2 and US1A. This indicates genotypic richness is highest in COLPR1 and lowest in the US1A genetic cluster, in which more than 80% of the isolates were clonal (Table 1, CF). Although, lower genotypic richness is inferred in COLPR2 and US1B as compared to the gene diversity pattern, this may be due to their low sample size. Evenness and the corrected Shannon-Wiener's index, Stoddart and Taylor's index and Simpson's Index, were all highest in the COLPR clade than in the US clade and followed the same pattern among genetic clusters as with genotypic richness (Table 1). Finally, there were no shared MLGs or MLLs among genetic clusters.

Similarly, between countries, significantly higher gene diversity in Colombia (0.263) compared with the United States (0.104) (Hs.test, $p = 0.002$). Gene diversity in Puerto Rico (0.163) was intermediate and not significantly different from the United States (Hs.test, $p = 0.218$) or Colombia (Hs.test, $p = 0.396$). Pairwise genetic distances, corrected genotypic diversity indices and evenness calculated for each country follow the same pattern of gene diversity (Supplementary Table S3). To infer migration among countries by tracking genotype flow, MLLs shared among countries were identified. A total of three MLLs were shared among countries. The MLL with one isolate from Colombia (Mph-5) and one from Puerto Rico (UPR-Mph-JD1) clustering in COLPR1, the MLL with one isolate from Puerto Rico (UPR-Mph-ISA3) and one from Louisiana (TN501) clustering in COLPR2, and the MLL with one isolate from Colombia (Mph-49) and 19 isolates from the United States clustering in US1A (Supplementary Figure S4). In addition, all populations clustering approaches indicated that Colombian isolates Mph-22 and Mph-23 are the most closely related to the United States isolates clustered in US1B, and Californian isolates M13-26 and M11-12 are the most closely related to Colombian isolates clustering in COLPR2. The rooted ML tree indicated isolate Mph-40 (from Colombia) as an outgroup to US clusters and discriminatory analysis of principal components (DAPC) clustered this isolate along with IN12-9-4 (from Indiana) with US1B isolates (Supplementary Figure S1). Overall,

TABLE 2 Population differentiation using Nei's G_{ST} pairwise genetic dissimilarity between genetic clusters identified in *Macrophomina phaseolina*.

Genetic cluster	US1A	US1B	US2	COLPR1
US1B	0.54			
US2	0.99	0.64		
COLPR1	0.80	0.69	0.81	
COLPR2	0.68	0.59	0.69	0.50

migration between Colombia, Puerto Rico and United States is a likely scenario.

To test the hypothesis that genetic clusters of *M. phaseolina* are differentiated, we used hierarchical analysis of molecular variance (AMOVA) and Nei's G_{ST} (an F_{ST} -analogous genetic differentiation measure applicable to haploids). Populations were significantly differentiated among clades, genetic clusters, as well as within genetic clusters ($p < 0.001$; Supplementary Table S2). AMOVA revealed that most of the total genetic variance was partitioned among US and COLPR clades (47%) and among genetic clusters (42%), and only 11% within genetic clusters. Consistently, very high genetic differentiation was found between US and COLPR clades ($G_{ST} = 0.45$) and among genetic clusters ($G_{ST} = 0.50$ – 0.99 ; Table 2). The COLPR2 ($G_{ST} = 0.50$ – 0.69) and US1B ($G_{ST} = 0.54$ – 0.69) clusters had the lowest G_{ST} when compared with any other cluster. Differentiation was lowest between COLPR1 – COLPR2 ($G_{ST} = 0.50$) clusters, and US1A – US1B ($G_{ST} = 0.54$) and highest between COLPR1 – US1A ($G_{ST} = 0.80$), COLPR1 – US2 ($G_{ST} = 0.81$) and US1A – US2 ($G_{ST} = 0.99$).

All other pairwise comparisons had similar intermediate levels of differentiation when compared to any other genetic cluster ($G_{ST} = 0.63$ – 0.69). The high values of G_{ST} in all pairwise comparisons suggest very high differentiation and little migration between genetic clusters. However, US1A – US2 G_{ST} estimation, which is notably high, was limited in power due to the low levels of gene diversity (H_{exp}) within these clusters. Across the 76,981 loci, there were only 76 and 255 polymorphic loci within US1A and US2 clusters, respectively. Thus, low gene diversity (H_{exp}) in US1A and US2 subpopulations likely resulted in overestimation of G_{ST} in pairwise comparisons of US1A and US2 with all other clusters.

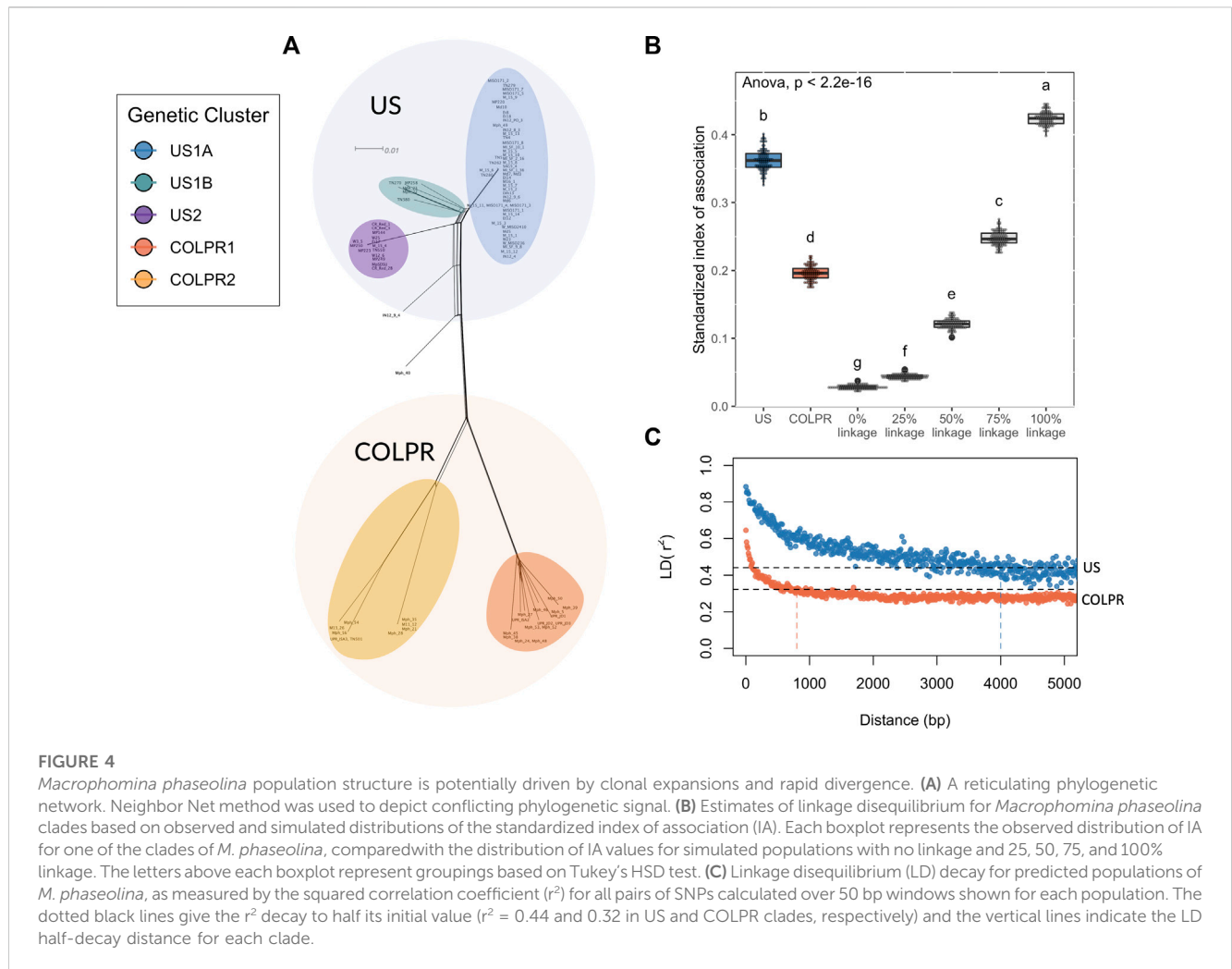
2.5 *Macrophomina phaseolina* is predominantly clonal in the United States and semi-clonal to mostly-clonal in Colombia and Puerto Rico

The predominantly star-like topology with little reticulation, in the Neighbor-Net network analysis, is consistent with a clonally reproducing population (Figure 4A). The standardized index of association (I_A) (Brown et al., 1980) was used to estimate the degree of clonality for each of the *M. phaseolina* main populations (US and COLPR clades). The observed I_A distributions for each population were compared to I_A distributions for simulated populations with no linkage, 25%, 50%, 75% and 100% linkage. A predominantly clonal mode of reproduction was inferred in the US and COLPR populations of *M.*

phaseolina. The simulated distributions and the different populations were significantly different from each other using ANOVA ($F = 25,287$, $df = 6$, $p < 0.001$). The distribution of the standardized I_A for the US population fell within the 75%–100% range of the linkage simulation (Figure 4B). This indicates a mostly clonal mode of reproduction with little potential for recombination. The distribution of the standardized I_A for the COLPR population fell within the 50%–75% range of the linkage simulation, indicating semi-clonal to mostly clonal reproduction in COLPR clades (Figure 4B). To further investigate the extent to which populations reproduce clonally, the linkage disequilibrium (LD) decay, as measured by the squared correlation coefficient (r^2) was calculated across pairs of loci for each of the clades. LD extends across a much larger distance in the US clade than in the COLPR clade, decaying over the first thousand base pairs, while in the COLPR clade LD decayed over the first hundreds of bases. LD half-decay distance, calculated as the average physical distance over which r^2 decays to half of its initial value was ~4,000 bp for US clade and ~800 bp for COLPR clade (Figure 4C). This indicates a high level of linkage occurs over larger regions of the genome in the US clade versus the COLPR clade. Importantly, although this may provide evidence for less clonal reproduction and higher recombination rates in the COLPR population, interpretation of standardized I_A and LD decay as associated with the frequency of recombination should be done with caution. It is possible that higher LD values did not reflect greater recombination; instead, it may be affected by lower sample size in COLPR and lower diversity in the US clade.

2.6 Climate contributes to SNP variation between *Macrophomina phaseolina* genetic clusters

To test the hypothesis that climate variation contributes to genetic variation across *M. phaseolina* genetic clusters a redundancy analysis (RDA) was employed. Four climatic variables were identified as significantly predictive of genetic variation using the simple RDA model with forward variable selection. Temperature seasonality (TSsd) was the strongest predictor, explaining 28% of the variation, followed by precipitation of warmest quarter (Pwq), precipitation seasonality (PScv) and mean temperature of warmest quarter (mTwq) (Table 3). Importantly, the climatic variables included in the RDA model were selected by their presumed biological significance and to avoid collinearity with other climatic variables and thus represent a subset of the variables possibly contributing to variation based on climate. The correlation of these variables with the first two RDA axes suggests their differential contribution to SNP variation among genetic clusters (Figure 5). Spatial structure, represented as distance-based Moran's eigenvectors maps (dbMEM), was used to identify climatic variables that are structured in space and to account for the effect of space in variance partitioning of total genomic variation. A total of three spatial variables were identified (dbMEM1-3; Supplementary Figure S5). Notably, when accounting for spatial structure (dbMEM1-3 variables), only Pwq, mTwq and precipitation of driest quarter (Pdq) were significant and accounted for 6% of SNP variation across isolates as determined with forward selection (Table 3), indicating collinearity between TSsd, PScv and space (i.e., spatially structured TSsd and PScv variation). To identify the spatial variables significantly contributing to genomic variation forward selection was



used. Of the three spatial variables, only dbMEM3 significantly explained 4% of the genomic variation and described broad-scale spatial structure (Supplementary Figure S5).

Partial redundancy analysis (pRDA) was used to estimate the partial contribution of each set of explanatory variables (e.g., climate) while removing the effect of the remaining variable sets (e.g., neutral population structure and space). Variance partitioning with pRDA, using the LD-filtered set of 11,421 SNPs, revealed that climate (TSsd, Pwq, PScv and mTwq identified by forward selection), neutral population structure (isolate PC scores for the first three axes of a PCA using intergenic SNPs) and space (dbMEM3 variable identified by forward selection) together significantly explained 72% of the total SNP variance. Nearly half of this variance was uniquely attributable to neutral genetic structure (32%), climate (4%), or space (1%), while the other half of the SNP variation was explained jointly between the three sets of variables (Table 4). The effect of climate alone was highly significant and explained 4% of the total genetic variance after removing the effects of neutral population structure and space (Table 4). These results support the hypothesis that climate significantly contributes to genetic variation and importantly, suggests that migration, drift, and potentially additional demographic and spatially structured processes (e.g., isolation by distance), represented by neutral population structure,

play a major role in shaping genomic variation in *M. phaseolina*. Moreover, the large fraction of variation common to climate, population structure and space, emphasizes the importance of accounting for confounded effects in genotype-environment associations, particularly when inferring causal associations.

2.7 Genotype-environment associations identify candidate SNPs for climatic adaptation

To identify loci that are potentially involved in local adaptation to climatic conditions, SNPs strongly associated with climatic variables were identified using RDA with and without accounting for population structure. Neutral population structure was used as it uniquely contributed the most to genetic variation. The RDA models, whether accounting for population structure (partial RDA) or not (simple RDA), were globally significant ($p < 0.001$) and the first three RDA axes explained most of the genomic variation associated with climate.

The candidate adaptive loci were identified based on extreme SNPs loadings, ± 3 or ± 4 SD from the mean, on each of the first three axes (Forester et al., 2018). In the partial RDA models, in which the effects of

TABLE 3 Climatic variables significantly contributing to SNP variation as determined by forward variable selection with simple RDA (redundancy analysis) and partial RDA conditioned on space.

Simple RDA (unconditioned)					Partial RDA (conditioned on space)					
Variable	R^2	Cum R^2	Cum R^2_{adj}	F-value	p -value	Variable	Cum R^2_{adj}	AIC	F-value	p -value
TSsd	0.28	0.28	0.27	36.63	0.001***	Pwq	0.03	638.88	4.96	0.002**
Pwq	0.05	0.33	0.31	6.28	0.001***	mTwq	0.04	637.67	3.06	0.004**
PScv	0.03	0.36	0.34	4.42	0.001***	Pdq	0.06	636.50	2.98	0.010**
mTwq	0.02	0.38	0.36	3.52	0.005**					

*** $p \leq 0.001$.

** $p \leq 0.01$.

AIC = Akaike information criterion

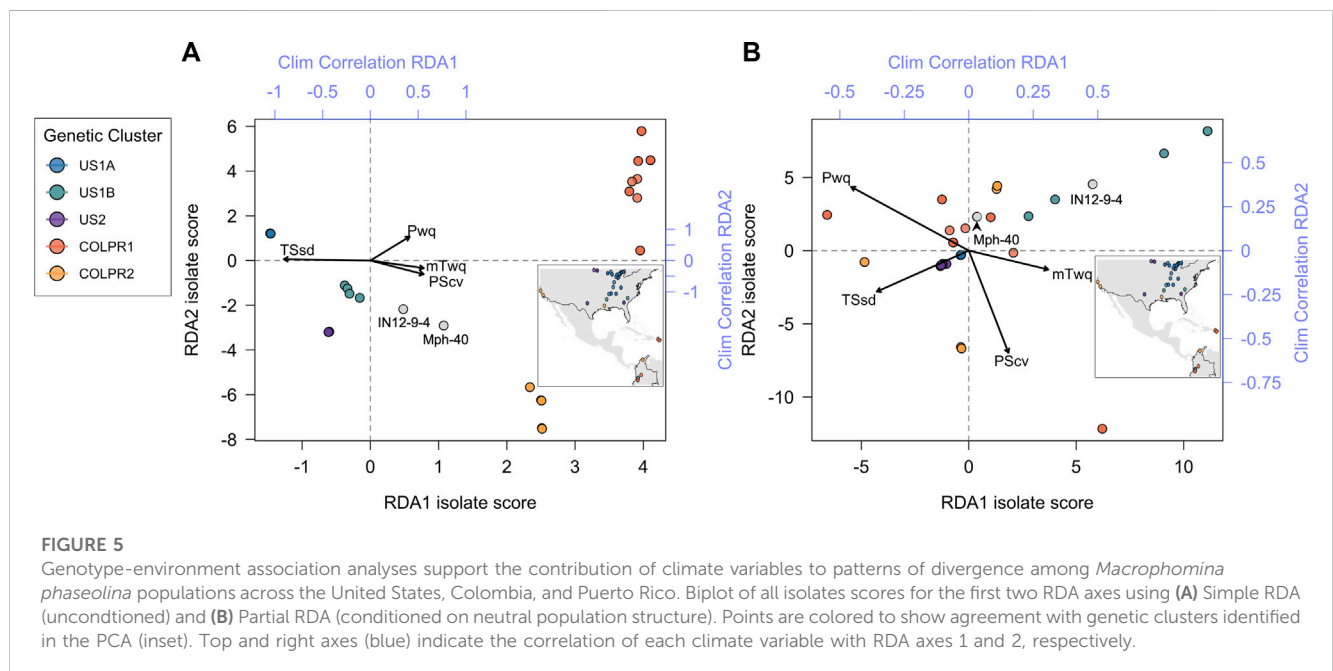


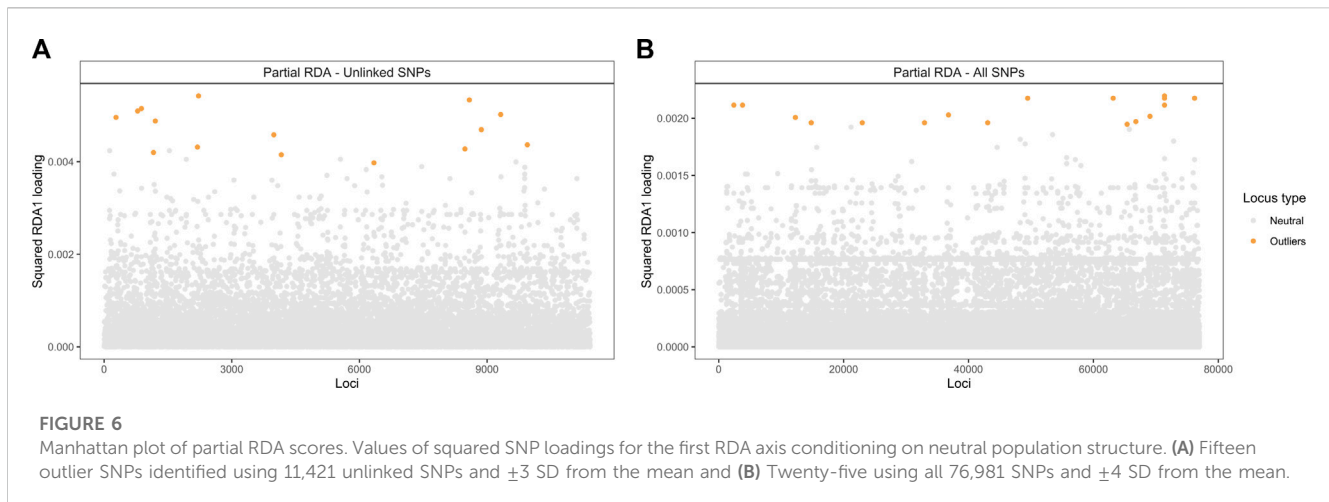
FIGURE 5 Genotype-environment association analyses support the contribution of climate variables to patterns of divergence among *Macrophomina phaseolina* populations across the United States, Colombia, and Puerto Rico. Biplot of all isolates scores for the first two RDA axes using (A) Simple RDA (unconditioned) and (B) Partial RDA (conditioned on neutral population structure). Points are colored to show agreement with genetic clusters identified in the PCA (inset). Top and right axes (blue) indicate the correlation of each climate variable with RDA axes 1 and 2, respectively.

TABLE 4 Contribution of climate, neutral population structure and space to SNP variation (11,421 unlinked SNPs) as determined by variance partitioning with partial RDA (redundancy analysis).

Partial RDA model	Inertia (variance)	R^2	p -value	Proportion of explainable variance	Proportion of total variance
Full model: $G \sim \text{clim.} + \text{sp.} + \text{struct}$	863.8	0.725	0.001***	1.00	0.72
Pure climate: $G \sim \text{clim.} \mid (\text{sp.} + \text{struct.})$	46.2	0.039	0.001***	0.05	0.04
Pure structure: $G \sim \text{struct.} \mid (\text{clim.} + \text{sp.})$	387.0	0.325	0.001***	0.45	0.32
Pure space: $G \sim \text{sp.} \mid (\text{clim.} + \text{struct.})$	6.3	0.005	0.001***	0.01	0.01
Confounded climate/structure/space	424.4			0.49	0.36
Total unexplained	327.9				0.28
Total inertia	1,191.7				1.00

*** $p \leq 0.001$.

Note: Climate variables are temperature seasonality (TSsd), precipitation of warmest quarter (Pwq), precipitation seasonality (PScv) and mean temperature of warmest quarter (mTwq) as identified with forward selection.



population structure were removed, 49 unlinked SNPs (when using the LD-filtered set and ± 3 SD from the mean; [Supplementary Figure S4](#)) and 75 SNPs (using all SNPs and ± 4 SD from the mean; [Supplementary Figure S5](#)) strongly associated with climatic variables were identified along the first three RDA axes. Of these SNPs, 15 and 25 (outliers in [Figure 6](#)) were identified in the first RDA axis when using the LD-filtered set or all SNPs, respectively, and 20 (19%) in both partial models. The strongest associations include SNPs with predicted effects in the membrane-associated 753,275-ankyrin, the 681,752-Ksh1 and the 241,776-protoporphyrinogen oxidase proteins. Structural modeling of the 753,275-ankyrin protein revealed that 598 residues (96% of the sequence) was modelled with 100% homology confidence to the transient receptor potential (TRP) NOMPC (No mechanoreceptor potential C) mechanotransduction channel protein in *Drosophila melanogaster* (chain C, highest scoring template; PDB ID: 5VKQ; Extended Data [Supplementary File 1](#)). Other SNPs with top associations are located within or in physical proximity to genes related to transmembrane transport, glycoside hydrolase activity, DNA binding and the gene encoding the 28,417-heat shock protein ([Table 5](#); [Supplementary Table S6](#)).

Because population structure could not be fully disentangled from climate (and space), as revealed in variance partitioning, the candidate loci obtained with population structure correction represent a conservative set subjected to a reduction in the detection of SNPs truly associated with climate. In the simple RDA model, without correcting for population structure, 91 candidate unlinked SNPs were identified ([Supplementary Table S7](#)). Only two SNPs were identified by both partial RDA and simple RDA models using unlinked SNPs ([Supplementary Figure S6](#)). This is in line with the high level of collinearity observed between genetic, space and climate ([Table 4](#)), and highlights the importance of accounting for confounded effects when identifying candidate loci under selection with genotype-environment associations.

3 Discussion

In this study, we describe the population structure of *M. phaseolina* in the continental United States, Puerto Rico and

Colombia collected from soybean and dry bean fields and the contributions of climatic factors to patterns of genomic diversity among populations. We found that five distinct genetic clusters of *M. phaseolina* evolved across the United States, Colombia and Puerto Rico and evidence suggests migration between genetic clusters and countries. To date, population genetic studies in *M. phaseolina* have performed their analyses at the resolution of microsatellites molecular markers and have provided important information on genetic diversity, host and geographic associations in the United States ([Baird et al., 2010](#); [Arias et al., 2011](#); [Koike et al., 2016](#)). However, no population-level genomic studies have been conducted to investigate population structure in this widespread pathogen. To our knowledge, we present the first population genomics study to investigate population dynamics and the role of climate in shaping patterns of genomic variation in *M. phaseolina* at a continental and regional scale. This study uses population genomics data to identify multiple strongly differentiated genetic lineages in the United States and demonstrated novel population structure in Colombia and Puerto Rico, which previously remained unstudied. Furthermore, our results highlight the importance of within-species genetic variation in understanding pathogens adaptive response to a changing climate and offers new insight with respect to the functional roles of genomic regions potentially underlying adaptation to climate. Notably, this research provides a practical framework for genotype-environment associations studies in *M. phaseolina* and other plant pathogens with complex evolutionary and demographic histories.

The influence of the low number of loci on limiting inferences about *M. phaseolina* population structure is emphasized by recent studies that used microsatellites markers ([Baird et al., 2010](#); [Arias et al., 2011](#)). These studies identified genetic groups in the United States; however, the genetic groups did not represent lineages (i.e., genetic groups and supported phylogenetic clades). Using population genomics, we provided strong evidence for five distinct genetic clusters of *M. phaseolina* and revealed that genomic variation in this globally distributed pathogen was consistent with a population hierarchically structured at a broad subcontinental regional extent. Two genetically differentiated *M. phaseolina*

TABLE 5 Candidate SNPs and gene models along the first RDA axis, after accounting for neutral population structure using the LD-filtered set of 11,421 SNPs.

SNP position	RDA1 loading	Climate variable	Correlation	SNP category	Distance from locus (bp)	Mycocosm gene location	Mycocosm protein ID	InterPro/KOG Desc	KOG Class/Putative function
scaffold_13:111451	-0.068	TSsd	0.90	Intergenic	3150	scaffold_13:108068-110230	753275	Ankyrin repeat	Cell wall/membrane/envelope biogenesis
scaffold_44:156076	-0.068	TSsd	0.89	Intergenic	3974	scaffold_44:152589-153590	681752	Ksh1(Protein kish)	Involved in the early part of the secretory pathway
scaffold_3:1825169	0.070	TSsd	0.62	Missense	0	scaffold_3:1824304-1825956	241776	Protoporphyrinogen oxidase	Coenzyme transport and metabolism/Heme biosynthesis
scaffold_3:1762959	-0.065	TSsd	0.62	Intergenic	1106	scaffold_3:1763965-1765572	763945	None	Unknown
scaffold_6:1672095	-0.074	TSsd	0.61	Intergenic	755	scaffold_6:1672555-1674150	726065	None	Unknown
scaffold_13:484791	-0.064	TSsd	0.59	Intergenic	563	scaffold_13:485238-487511	88264	Transcription factor domain, fungi	DNA binding/Zinc ion binding (Zn(II) ₂ Cys ₆ transcription factor-related)
scaffold_1:2854576	-0.070	TSsd	0.56	Synonymous	0	scaffold_1:2854139-2855202	36408	GXWXG domain	Unknown/Putative transcription factor <i>Cmr1</i> homolog
scaffold_59:25310	-0.066	TSsd	0.51	Synonymous	0	scaffold_59:24613-27244	365205	AMP-dependent synthetase/ligase	Lipid transport and metabolism
scaffold_22:420202	-0.063	TSsd	0.50	Intergenic	4892	scaffold_22:414577-415638	787628	DUF1772 family	Unknown
scaffold_48:217394	-0.071	TSsd	0.45	Intergenic	4970	scaffold_48:211321-213431	633816	Glycoside hydrolase, family 1	Carbohydrate transport and metabolism
scaffold_3:91936	0.072	TSsd	0.39	Intergenic	1176	scaffold_3:93112-93948	735628	Allergen V5/Tpx-1-related, conserved site	Unknown
scaffold_6:1470614	0.066	TSsd	0.37	Intergenic	14457	scaffold_6:1454967-1456373	382155	Ribonuclease T2-like	RNA processing and modification
scaffold_2:1936912	0.071	TSsd	0.35	Intergenic	443	scaffold_2:1936361-1937105	643334	Thioesterase superfamily	Unknown
scaffold_42:203393	0.073	mTwq	0.29	Synonymous	0	scaffold_42:202488-204519	608404	Cytochrome P450, E-class, group I	Lipid transport and metabolism
scaffold_41:309748	0.065	mTwq	0.29	Intergenic	4841	scaffold_41:303593-305479	582790	Flavin-containing monooxygenase	Secondary metabolites biosynthesis, transport and catabolism

clades at the United States and Colombian-Puerto Rican geographical level and five distinct genetic clusters representing finer population structure within each of these clades were identified. These genetic clusters, except for US1B, represent strongly supported phylogenetic clades and likely represent different evolutionary lineages of *M. phaseolina*. This distinction is important because the identification of lineages allows for the inference of ecological and evolutionary processes in a population-specific manner and underscores the potential for local adaptation in *M. phaseolina* populations.

Our results provide support for regional clustering within the United States and a lack of strong grouping at a state level, also observed in previous studies based on microsatellite data (Baird et al., 2010; Arias et al., 2011). The US1A cluster, found in the East North Central and Central region, expands previous studies confirming that isolates collected from soybean in these regions represent a largely homogeneous population (Arias et al., 2011). This is supported by low gene diversity and pairwise genetic distances found in the US1A genetic cluster in agreement with low diversity detected with microsatellite markers in soybean isolates collected mostly in Tennessee and Missouri (Arias et al., 2011) and midwestern states (group III; Baird et al., 2010). The US2 genetic cluster found in Northern Great Plains and South United States regions grouping isolates from Minnesota, South Dakota, Texas, and Georgia is partially consistent with Baird et al. study. Isolates from these states along with isolates from North Dakota represent the majority of a subcluster of group I in Baird et al. Like in the US clusters, grouping at broad geographic regions was observed in COLPR1 and COLPR2 clusters. Both COLPR1 and COLPR2 clusters grouped isolates from locations across Colombia and Puerto Rico. In COLPR2, isolates from California and Louisiana grouped closely to isolates from Colombia and Puerto Rico. Although the small sample size from these states (only two isolates collected from strawberry in California and one isolate from soybean in Louisiana) demands that this grouping be reassessed once more isolates are included from these states and hosts in future studies. The clustering of isolates from widespread geographic regions observed in COLPR2, as well as in US1A and US1B clusters, suggests a role for migration in structuring *M. phaseolina* populations. These results better align our understanding of *M. phaseolina* population structure with a metapopulation model, that predicts regional persistence of populations while local populations are unstable and connected by some level of migration (Hanski, 1998; Milgroom, 2015). The metapopulation dynamics view expands the interpretation of past *M. phaseolina* population structure studies while providing a conceptual basis for the design of future studies.

The presence of multiple distinct genetic clusters in the United States and higher genetic diversity in COLPR clusters led us to inquire about whether Colombia and Puerto Rico may serve as potential source populations for US populations. In the rooted ML phylogeny, the reconstruction of COLPR clusters as outgroups to US clusters support this hypothesis. Furthermore, across all analyses we found indications that US1B may serve as a sink population for Colombia and Puerto Rico populations. The US1B genetic cluster grouped isolates from Mississippi and South Carolina along with two Colombian isolates and was the most genetically diverse of the US clusters. Further, US1B was positioned centrally in PCA space,

basal to US1A cluster in the rooted ML phylogeny and was less differentiated, along with COLPR2, from all other clusters based on G_{ST} values. Finally, in DAPC analysis, US1B isolates clustered with IN12-9-4 and Mph-40 isolates, which are reconstructed intermediate between US and COLPR clades in the rooted ML phylogeny and as admixed in spatial population structure analyses. Although, the high diversity in US1B may be reflective of the grouping of comparatively few isolates from different geographic regions in this cluster. However, when all data are considered, it suggests the US1B cluster geographic region as a potential route of introduction of isolates from Colombia or Puerto Rico to the United States. More isolates from the United States and other countries would need to be included in future studies to test this hypothesis.

The discrete population structure observed between US and COLPR clades, provides compelling evidence for isolates in each clade drawing ancestry from different ancestral populations. A plausible explanation, supported by our results, for this different ancestry would be a demographic event such as a rare long-distance migration (e.g., introduction event) from the COLPR clusters, leading to a recent bottleneck in the US populations. The high probability assignments observed in US clusters may be consistent with the expected strong recent genetic drift in bottlenecked populations (Lawson et al., 2018). In this scenario, we speculate that the diversity in US clusters represent a subset of the diversity of the COLPR genotypes found in Colombia and Puerto Rico. At the finer genetic cluster population structure, isolation by distance provided a potential explanation for the continuous genetic differentiation in spatial population structure analyses. Although, isolation by distance patterns may be observed as part of a variety of underlying biological processes and demographic scenarios (Sexton et al., 2014; Milgroom, 2015), it is possible that these patterns reflect a scenario of restricted dispersal in the context of divergence following clonal expansions in the US genetic clusters. For example, both US1A and US2 genetic clusters are found in Michigan, Wisconsin, and Kentucky, supporting dispersal of isolates among these states. However, high population differentiation indicated by high G_{ST} values between genetic clusters, suggest substantial restriction to gene flow. Given the soilborne nature of *M. phaseolina* and limited natural dispersal ability but high potential for anthropogenic mediated dispersal, restricted events of dispersal associated to seed, plant material or farm equipment at limited distances relative to the geographic range of the genetic clusters, seems a likely occurrence (Baird et al., 2010). Similar isolation by distance patterns has been observed in other soilborne fungal and oomycete pathogens with restricted long-distance dispersal (Grünwald and Hoheisel, 2006; Milgroom et al., 2016).

Diversity was found to be further reduced in US1A genetic cluster as compared to all other clusters. Low diversity and high differentiation are signatures of genetic drift but also selection. If reduced diversity in the US1A genetic cluster was consistent with a clonal expansion following a bottleneck, the divergence and marked low diversity could reflect both genetic drift and selection. Genetic drift is expected to have substantial effects on pathogen populations, because migrations resulting in founder effects and reduced population sizes associated with pathogens survival in soil (Milgroom, 2015). Additionally, we speculate that climatic

conditions, particularly strong fluctuations in temperature in the northern United States, could impose strong selection on *M. phaseolina* populations in this region. Overall, we believe the genomic signals of discrete and continuous structure that differentiate *M. phaseolina* populations could be reflective of a complex demographic and evolutionary history. Therefore, alternative demographic scenarios, including one of multiple independent introductions, should be considered in future studies ideally applying demographic modelling with a broad geographic and temporal distribution of isolates.

Across all analyses we found support for Colombia and Puerto Rico as potential sources for United States *M. phaseolina* populations. Genetic diversity between countries also supported this hypothesis. Whereas Colombian isolates were significantly more diverse than United States isolates, diversity in Puerto Rico was intermediate and not significantly different from United States or Colombia. These findings may be consistent with the idea of Middle or South America as putative centers of origin for *M. phaseolina* and with its introduction to North America as part of historical crop migrations. For example, common bean Middle American origin, domestication centers in Middle America and South America (Bitocchi et al., 2017) and later movement to the United States via the Caribbean, Central and Eastern United States (Kelly, 2010), makes likely an explanation for *M. phaseolina* introduction to the United States in bean seeds. Pathogen geographic origins have been associated with the centers of diversity of their major crop host. Nonetheless, pathogen origin associated with their hosts' wild relatives, have been also observed in some plant pathogens. For example, a genetically diverse and sexually reproducing population of *Phytophthora infestans* was found in central Mexico consistent with this pathogen's origin in a secondary center of potato (*Solanum tuberosum*) diversity and potentially involved in a host jump from native *Solanum* species (Goss et al., 2014). Given *M. phaseolina* host generalist nature, a strict host-pathogen coevolution scenario is not expected (Slippers and Wingfield, 2007), obscuring inferences about its center of origin. In Kansas, isolates collected from wild tallgrass prairie were found more diverse than isolates from maize, soybean and sorghum crops (Saleh et al., 2010). This finding may indicate *M. phaseolina* presence in the United States precedes the introduction of agriculture or it may be explained by connectivity dynamics between natural and agricultural ecosystems contributing to patterns of diversity in *M. phaseolina* populations from these ecosystems (Saleh et al., 2010). Thus, the origin and evolutionary history of *M. phaseolina* is likely more ancient and complex than could be tested with the isolates included in this study, and future studies may benefit from considering the potential involvement of host adaptation from wild hosts.

Genotype tracking provided compelling evidence for migration among the United States, Colombia, and Puerto Rico. The MLL consisting of the Colombian isolate Mph-49 and several isolates from the United States clustering in US1A, along with the high clonality found in this cluster and the significantly high diversity in Colombia, makes a Colombian source likely. Similarly, the MLL shared between Colombia and Puerto Rico and the MLL between Puerto Rico and Louisiana support migration between countries. Alternatively, the same MLLs could have been introduced independently to United States, Puerto Rico, and Colombia,

potentially from an ancestral and more diverse population not included in this study. Although this scenario seems less likely, it remains a possibility. Given that besides historical crop migrations, migration as part of international seed exchange is a likely occurrence in *M. phaseolina*, as in other seedborne species and latent pathogens of the Botryosphaeriaceae family (Sakalidis et al., 2013; Crous et al., 2017), we believe that *M. phaseolina* has been spread at least intercontinentally, possibly globally, through seed. However, time, frequency, and directionality of migration between United States, Colombia, and Puerto Rico, and the potential for multiple introductions would need to be examined in future studies.

Although various population genetic studies in *M. phaseolina* have found patterns of host associations (Jana et al., 2005; Baird et al., 2010; Arias et al., 2011; Koike et al., 2016; Reznikov et al., 2018; Burkhardt et al., 2019), our results did not find that genetic variation is associated with host in the two major US clusters. Soybean and dry bean isolates grouped together in US1A and US2 clusters. Given that most previous studies support some degree of host preference, and genomic evidence for genes uniquely present in the *M. phaseolina* strawberry genotype further support host preference (Burkhardt et al., 2019), we suspect that our sampling scheme was not enough to capture clear associations to a plant host. A clear limitation in our study was that the host origin was confounded with geographic origin, except for Michigan where isolates were sampled from both soybean and dry bean. The independent grouping of host might also reflect crop rotation and equipment practices implemented in fields. Additionally, our data may reflect that the sampled hosts are both legumes. Genetic similarity has been found to be greater among isolates collected from the same host than from hosts in different families (Su et al., 2001; Saleh et al., 2010). These results do, nonetheless, have important practical implications for soybean breeding resistance to charcoal rot. In the US1A cluster, the high genetic similarity of isolates collected from soybean and dry bean, may indicate that the use of one or few isolates collected from these crops throughout East North Central and Central United States regions may suffice for resistance screening of soybean breeding material. An important limitation to this assumption is that we use a single reference genome approach to characterize genetic diversity and thus accessory genes and other structural variation potentially involved in pathogenesis are not considered (Bertazzoni et al., 2018).

Importantly, the dry bean diversity in research plots from which Colombian and Puerto Rican isolates were collected is a factor likely contributing to their higher genetic diversity as compared to United States isolates. In research plots, multiple lines are continually evaluated as part of breeding programs, in contrast to commercial fields in which a single or few varieties are used. This coupled with climatic conditions in Colombia and Puerto Rico that favor year-round inoculum presence in crop residue represent important considerations when interpreting isolate genetic diversity in relation to host origin.

The population structure results suggest that *M. phaseolina* populations lay in-between the clonality-recombination spectrum (Smith et al., 1993). Furthermore, our results suggest that this may occur in a population-specific manner. On one side of the spectrum, we found *M. phaseolina* to have a markedly clonal population structure (Milgroom, 2015). First, most of the intraspecific genetic variation in *M. phaseolina* is explained by differences between clades and genetic clusters, while low genetic variation

was observed within genetic clusters. Second, the occurrence of nearly identical genotypes (i.e., MLLs) from widespread geographic locations found in *M. phaseolina* is in line with a markedly clonal population structure (Milgroom, 1996; Milgroom, 2015). On the other end of the spectrum genotypic diversity, network analyses and measures of linkage among loci provided support for recombination within some of the genetic clusters. High levels of genotypic diversity is one of the characteristics reflective of recombination in fungal populations (Milgroom, 1996). The higher genotypic diversity (eMLLs) in US1B, US2, and COLPR clusters, may be consistent with the occurrence of recombination in these clusters. Network analyses account for recombination by allowing to infer homoplasmy caused by recombination. The boxes between isolates within genetic clusters in the network and the PHI test supporting recombination within all clusters except for US1A, strengthen this hypothesis. The index of association, I_A , revealed an overall high degree of linkage among SNP markers, in line with a pathogen that reproduces clonally. However, the observed I_A values in the COLPR clade and LD decaying faster in COLPR than in US populations, support the potential occurrence of recombination among isolates within COLPR clusters. Although the problem of smaller sample size in COLPR clusters should be at least partially accounted for by using simulations in IA analysis and clone-corrected data in LD-decay analysis, particularly half-decay LD values should be interpreted with caution and examined in future studies to determine the extent of recombination in *M. phaseolina* populations.

These results are consistent with the population structure model that lays in between the “strictly clonal” and “epidemic” structure proposed by Smith et al. (1993), in which frequent recombination does not occur between isolates in separate branches of an evolutionary tree but it occurs between isolates within a given branch (Smith et al., 1993). These models have been used to describe the population structure of plant pathogens with mixed modes of reproduction or inferred recombination (Grünwald and Hoheisel, 2006; Milgroom et al., 2014; 2016; Milgroom, 2015). While little is known about the occurrence of recombination in *M. phaseolina*, recent studies have started to shed light on potential recombination mechanisms involving parasexuality (Pereira et al., 2018) and horizontal gene transfer mediated by giant mobile genetic elements (Gluck-Thaler et al., 2022). Whether other potential recombination mechanisms occur, and the frequency of recombination in *M. phaseolina* remains an important and exciting area of study.

Partial RDA revealed that nearly half of the SNP variance is confounded between neutral genetic structure, climate, and space. This means that this fraction of the variance cannot be statistically associated to a direct effect of any single set of variables. Importantly, the effects of population structure and space often cannot be independently disentangled from spatially structured process (e.g., IBD) or spatially structured environmental variables (Lasky et al., 2015). This study, while highlighting the challenges in assessing genotype-environmental associations, provided an assessment of the fraction of confounded variance and allowed us to start disentangling the effects of climate, spatial, and population structure on genomic variation in *M. phaseolina* populations. The genotype-environment association analyses using partial RDA support our hypothesis that local climatic differences contribute

to patterns of adaptive divergence among *M. phaseolina* populations across the United States, Colombia, and Puerto Rico. Seasonal variation in temperature and precipitation of warmest quarter, were the primary climatic variables associated with variation of candidate adaptive loci without and after accounting for neutral genetic population structure, respectively. We found SNPs within or in physical proximity to genes with functional annotations related to transmembrane transport, glycoside hydrolase activity and DNA binding. In fungi, genes involved in these activities are known to be important in responses to environmental stressors including temperature, water availability, and oxidative stress (Aguilera et al., 2007; Gasch, 2007; Branco et al., 2016). Similarly, among the candidates, we found the 241,776-protoporphyrinogen oxidase protein, involved in heme biosynthesis and the putative small heat shock protein 28,417-*Hsp20*. Heme has been shown to regulate several mechanisms during cold shock in *Saccharomyces cerevisiae* (Abramova et al., 2001) while *Hsp20* proteins have been found involved in fungal thermal stress response to both heat and cold (Wu et al., 2016; Wang et al., 2021).

The SNP with the highest correlation with temperature seasonality was located upstream to the 753,275-ankyrin repeat protein (Table 5). We found that *M. phaseolina* 753,275-ankyrin protein is homologous to the TRP NOMPC mechanotransduction channel in *D. melanogaster* (Jin et al., 2017). Ankyrin family proteins link membrane proteins, including ion channels, to microtubules of the cytoskeleton by binding of its ankyrin repeat domain. The ankyrin proteins in the NOMPC channel links a displacement of the cytoskeleton to the channel opening, translating external stimuli into intracellular signals (Jin et al., 2017). Moreover, the TRP1 (transient receptor potential 1) ion channel from the alga *Chlamydomonas reinhardtii*, which shares structural homology to the TRP NOMPC channel, was found to act as thermal sensor, with ankyrin proteins mediating the channel opening in response to increased temperature (McGoldrick et al., 2019). Although there is no structural or functional characterization of the *M. phaseolina* 753,275-ankyrin protein, it represents a promising candidate to investigate a potential temperature-related mechanism for environmental stimuli transduction. These findings are consistent with the established roles of proteins in environmental stress responses both specific to fungi and conserved across the tree of life. Although our results cannot confirm whether SNPs are the causal mechanism, the candidate genes could be used in future functional studies. Additionally, common garden experiments could provide support for local adaptation to climate in *M. phaseolina*.

Overall, our observations point to a scenario in which *M. phaseolina*, as other plant pathogens with clonal population structures, is structured in a subcontinental regional stable manner in the face of instability at local scales in line with the metapopulation dynamics perspective. These results are consistent with a scenario of evolution after migrations driven by divergence following clonal expansions. The presence of MLLs across countries underscores the potential for a large influence of anthropogenic migration introducing *M. phaseolina* to new environments. The association of genetic divergence with climatic variables and putatively adaptive functions of the genes with SNPs strongly associated that would hypothetically benefit *M. phaseolina* in specific environments, is consistent with potential selection imposed by specific climatic variables. Future studies will be needed to identify the degree to which

distinct genetic groups reflect their adaptation to host and climate. Such analyses will benefit from a global sampling collected from diverse hosts in conjunction with multiple reference genomes sequenced with long-read technologies that will allow further characterization of the role of genomic variation, including structural variation, in *M. phaseolina* adaptation to host and the climatic environment.

This knowledge expands the impact that spatial population genomics and genotype-environment associations can have on our ability to characterize adaptive potential in plant pathogens by identifying candidate genes and presents a preliminary and complementary approach to the forward-genetics and phenotypic characterization approaches. The ability to identify candidate genes at a population specific level in a clonal pathogen presents an opportunity to evaluate candidate genes in a population specific manner, which represents a powerful approach specially in clonal pathogens in which unusually high levels of linkage prevent the application of genome scan methods. Additionally, the RDA approach could be applied using candidate adaptive genetic markers to predict pathogens' "adaptive landscape" representing its adaptive variation for any environment across a geographic range (Capblancq and Forester, 2021). As climate and agricultural challenges become more demanding, the characterization of pathogen adaptation capabilities enabled by population genomics should become increasingly utilized for plant disease risk prediction models specially under adverse future climate scenarios. In conclusion, our study emphasizes the importance of population genomics for identifying distinct genetic groups and uncovering potential recombination and population-specific adaptation related to climate in fungal plant pathogens, specifically in *M. phaseolina*. The spread of these adaptations among populations through gene flow at local scales or transcontinental introductions could create challenges in managing charcoal rot, as selection pressure could lead to the emergence of highly adapted and virulent pathogen populations. Hence, it is crucial to monitor the emergence and introduction of pathogens with novel adaptations. Our findings provide a foundation for future research with practical applications for management of charcoal rot, such as developing plant disease risk prediction models, informing the development of resistant crop varieties, and regulating the global movement of *M. phaseolina*.

4 Methods

4.1 Isolate collection and DNA preparation

A total of 95 *M. phaseolina* isolates were obtained from culture collections, as well as roots or lower stems of soybean and dry bean plants in production fields (Supplementary Table S1). Isolates were originally collected in the United States, 70; Colombia, 20; and Puerto Rico, 5 and selected to reflect a latitudinal range and climatic variation (Figure 1). There were 52 isolates collected from soybean across a latitudinal range in 13 US states, including 38 isolates from a previous study (Sexton et al., 2016). Forty isolates were collected from dry bean grown in Michigan, Puerto Rico and Colombia. Isolates from Michigan were collected from 2011 to 2017 as part of root rot surveys conducted in soybean and dry bean (Rojas et al., 2016; Jacobs et al., 2019). Isolates from Puerto Rico and Colombia were collected from research plots at the University of Puerto Rico (UPR) and at the International Center for Tropical Agriculture

(CIAT) and are currently maintained at the UPR and CIAT Plant Pathology Laboratory culture collections. Two strawberry isolates collected from California and one isolate from Ethiopian mustard (*Brassica carinata*) were included as host outgroups. Cultures were routinely grown on potato dextrose agar (PDA; Acumedia, Lansing, MI) medium.

For genomic DNA extraction, four 5-mm plugs taken from the edge of a single hyphal tip culture were used to inoculate 50 mL of potato dextrose broth amended with chloramphenicol (50 mg/L). The broth was incubated for 7–9 days at room temperature. Mycelia were harvested, lyophilized for 24 h and ground using a FastPrep FP120 homogenizer (BIO 101 Savant Instruments, Hobrook, NY). Genomic DNA was extracted from the lyophilized tissue using a modified SDS-based method; briefly, 50 mg of ground mycelia were mixed in lysis buffer (3% SDS (w/v); 100 mM Tris-HCl, pH 8.0; 50 mM EDTA, pH 8.0) followed by phenol/chloroform DNA extraction. The identity of all isolates was confirmed by multigene DNA analysis of the Internal Transcribed Spacer regions for the nuclear rDNA operon (ITS), part of the Translation Elongation Factor (TEF-1 α) gene region, and part of the actin (ACT) gene region according to (Sarr et al., 2014). Maximum likelihood analysis of the combined sequence alignment placed all the isolates tested in the *M. phaseolina* cluster. A full heuristic search using the first ten most parsimonious trees and the Neighbor-joining tree as starting trees with 100 random sequence additions was performed in PAUP v4.0b10 (Swofford, 2003), to find the maximum likelihood tree (Supplementary Figure S7).

4.2 Whole genome sequencing and variant calling

Genomic libraries were constructed and each of the isolates were whole-genome sequenced to 23X coverage using a 150 base-pair paired-end strategy on the Illumina HiSeq 4,000 platform at the Michigan State University Research Technology Support Facility Genomics Core (East Lansing, MI). The libraries were prepared using the Illumina TruSeq Nano DNA Library Preparation Kit HT. The resulting sequences were quality assessed using FastQC (Andrews, 2010) and cleaned using Cutadapt v1.16 (Martin, 2011), with the following parameters: f: fastq, -q 20,20, --trim-n, -m 30, -n 3, -a AGATCGGAAGAGCACACGTCT GAACTCCAGTCAC, -A AGATCGGAAGAGCGTCGTGTAGGG AAAGAGTGTAGATCTCGGTGGTCGCCGTATCATT. After initial quality filtering, the remaining sequences were aligned to the *M. phaseolina* reference genome Macpha1 (JGI Mycosm, MPI-SDFR-AT-0080 v1.0) using bwa-mem (Li, 2013). The isolate used for the Macpha1 reference genome was collected from natural *A. thaliana* populations in France (Mesny et al., 2021). The mapping statistics, genome alignment rate and genome coverage were assessed with SAMtools flagstat (Li et al., 2009). Alignments were sorted and indexed using SAMtools (Li et al., 2009). After mapping, duplicate reads were identified using MarkDuplicates and removed during the variant calling step.

Single nucleotide polymorphisms (SNPs) of all 95 isolates were predicted using the Genome Analysis Toolkit (GATK) v4.0 (McKenna et al., 2010). Initially, SNPs were called individually with GATK's HaplotypeCaller. GVCF files were combined, and common SNPs jointly identified using CombineGVCFs and GenotypeGVCFs programs. The later using the -new-qual

parameter. The combined vcf file was quality filtered using vcfR v1.10.0 package (Knaus and Grünwald, 2017) in R v4.0.0 (R Core Team, 2019). To be included in the high-quality set, SNPs were filtered to remove SNPs with a minimum read depth (DP) of <4x and greater than the 95th percentile of each sample DP distribution and exclude SNPs with minimum threshold mapping quality (MQ < 60) and minimum allele frequency (MAF < 0.02) which corresponds to the allele presence in at least two isolates. Only variants with no missing data were retained, which corresponds to positions with 0 missing data for all the sequenced isolates. The final high-quality dataset was used in all subsequent analysis. The final vcf was annotated using SnpEff v5.0c (Cingolani et al., 2012b) and a vcf containing only SNPs in intergenic regions was created using SnpSift v5.0c (Cingolani et al., 2012a).

4.3 Phylogenomics and population genetic structure

The population structure was inferred according to the results from both model-based and model-free clustering methods and phylogenetic inference. The phylogenetic tree was inferred from the full set of high-quality SNPs among the 95 *M. phaseolina* isolates in RAXML-NG v1.01 (Kozlov et al., 2019). The RAXML analysis was performed using the “-all” option which conducted 20 maximum likelihood inferences on the original SNP alignment, standard bootstrapping with automatic determination of the number of replicates (Felsenstein’s bootstrap, FBP; MRE-based bootstrapping test) and the subsequent maximum likelihood search. The General-Time-Reversible (GTR) model of nucleotide substitution with GAMMA model of rate heterogeneity and correction for ascertainment bias (GTR + G + ASC_LEWIS) was used. The best-scoring ML tree was used for optimizing all model and branch length parameters and model evaluation. A model-free dimensionality-reduction approach, principal component analysis (PCA), and discriminatory analysis of principal components (DAPC) were also conducted on the full set of SNPs using adegenet package (Jombart, 2008; Jombart and Ahmed, 2011) in R 4.0.0 (R Core Team, 2019). To infer population dynamics and reconstruct a rooted *M. phaseolina* phylogeny, the Macphal1 (JGI Mycocosm, MPI-SDFR-AT-0080 v1.0) reference genome was used as outgroup taxon. Maximum likelihood analysis was run in RAXML-NG v1.01 using the “-all” option with automatic bootstrap replicates and the GTR + G + ASC_LEWIS substitution model.

4.4 Spatial genetic structure

Bayesian clustering of allele frequencies was implemented in conStruct (Bradburd et al., 2018). To assess whether population structure was well described by modelling isolates as admixtures between multiple discrete genetic groups or by both discrete and continuous genetic structure, spatial analysis of population structure was conducted using conStruct (Bradburd et al., 2018). Spatial analysis in conStruct, accounts for isolation by distance by allowing genetic differentiation to increase with geographic distance within discrete genetic groups (layers, K). The data was analyzed treating individual isolates as the unit of analysis, using the spatial models setting K

between 1 and 7 with 20,000 iterations, and compared these models using cross-validation with 10 replicates. For cross-validation, 90% of loci were used to fit the model and the remaining loci for model evaluation. A geographically constrained least-squares method as implemented in TESS3 (Caye et al., 2016), was used to estimate ancestry coefficients and create interpolation maps based on the coefficients. TESS3 uses a spatially explicit algorithm that can be considered model-free. The algorithm was run using the function “tess3” with K between 1 and 7 and 10 replicates.

4.5 Population genetic and genotypic diversity

For each clade and genetic cluster, gene diversity (Nei, 1978) was calculated using the Hs function in the adegenet package (Jombart, 2008; Jombart and Ahmed, 2011). The median estimates of pairwise genetic distance and genotypic diversity indices were calculated within each clade and genetic cluster using the R package poppr v2.9.0 (Kamvar et al., 2014). Genotypic diversity was assessed by calculating the number of multilocus genotypes (MLGs). A MLG was defined as a unique combination of the 76,981 SNPs. MLGs were collapsed into larger groups called multilocus lineages using the average neighbor algorithm and a Prevosti’s distance threshold of 0.0001 (bitwise.dist function; Kamvar et al., 2015). Rarefaction was used to correct for uneven sample sizes using the R package vegan v2.5–6 (Oksanen et al., 2019) and obtain the number of expected MLGs and MLLs (eMLG and eMLL) at the lowest common sample size (i.e. 25 for clades and 5 for genetic clusters). Genotypic diversity indices, Shannon-Wiener Index (H*), Stoddart and Taylor’s Index (G*), Simpson’s index (lambda*) and evenness (E5*) (Grünwald et al., 2003), were calculated using the R package poppr v2.9.0 (diversity_ci function; Kamvar et al., 2014) based on the number of MLLs in each clade and genetic cluster and correcting for unequal sample sizes based on rarefaction. The function mlg_crosspop in poppr was used to detect the presence of MLGs occurring across populations. Migration was inferred by tracking MLGs across genetic clusters, referred here as genotype flow (McDonald and Linde, 2002).

4.6 Population differentiation between genetic clusters and countries

The F_{ST} analog, G_{ST} (Nei, 1973) was calculated from clone-corrected data using vcfR (Knaus and Grünwald, 2017) to infer differentiation among genetic clusters. To describe the population dynamics between the United States, Puerto Rico and Colombia, the degree of genetic differentiation across *M. phaseolina* samples was measured hierarchically by genetic clusters within clades. Analysis of molecular variance (AMOVA) based on the quasi-Euclidean distance matrix was conducted in poppr v2.9.0 (Kamvar et al., 2014). AMOVA estimates the number of differences summed over loci based on a matrix of distances between individuals and covariance components are used to calculate Φ fixation indices for each hierarchical level, among clades, among genetic clusters and within genetic clusters. Significant differences of fixation indices were determined by 1,000 random permutations (Grünwald and Hoheisel, 2006).

4.7 Recombination and clonality

To account for potential intraspecific recombination among *M. phaseolina* isolates, a phylogenetic network was built using the Neighbor-Net algorithm as implemented in SplitsTree4 v4.16.1. The extent of clonality was tested by calculating the proportion of significant linkage between pairs of loci, by computing the standardized index of association (I_A , Brown et al., 1980) for each of the main populations (US and COLPR) using poppr v2.9.0 (Kamvar et al., 2014). Linkage disequilibrium is expected in asexual or inbreeding populations and I_A values close to zero are expected for outcrossing populations (Burt et al., 1996). The observed I_A distributions for each population were compared to five simulated recombined distributions (0%, 25%, 50%, 75% and 100% linkage) generated among 76,981 loci and 48 samples (corresponding to the median population size of the two clusters). The observed and simulated I_A values were tested for normality using the Shapiro-Wilk's normality test and an analysis of variance (ANOVA) was conducted to test for significant differences among the distributions. Pairwise comparisons between the I_A simulated distributions and for each population were tested for difference with Tukey's HSD test in R. The extent of clonality was correlated to clonal (100%), mostly clonal (75%), semiclinal (50%, 25%) or sexual (0%) modes of reproduction. Linkage disequilibrium (LD) decay rate was estimated using the physical distance over which LD decays to half its initial value, as measured by the squared correlation coefficient (r^2). The linkage disequilibrium decay was calculated for each clade using the correlation coefficient (r^2) in TASSEL v5 (Bradbury et al., 2007) within a window of 50 sites among SNPs using the clone-corrected dataset (79 MLGs). The mean r^2 values, representing the correlation between alleles at two loci within 10 bp of physical distance, were then plotted in R 4.0.0 (R Core Team, 2019).

4.8 Climatic data

For each isolate, the 19 standard bioclimatic variables available at the WorldClim2 database (Fick and Hijmans, 2017) were obtained using 'getData' function from raster R package (Hijmans, 2022). All variables are the average for the years 1970–2000 and were obtained at a spatial resolution of 2.5 min (~21.5 km²). We used data at a resolution of 2.5 min (~21.5 km²), because it corresponds with our sampling design (single isolate samples rather than populations) being at a field or county scale. Coarser resolutions could combine multiple sampling locations into a single spatial grid and finer resolutions (30-s or <30-s), while this may be important for structuring patterns of genetic variation within populations, these data are less suitable for our sampling design and focus on regional to continental-wide patterns. We reduced the number of climatic variables from 19 to five to account for collinearity among them ($|r| > 0.7$) and to represent our hypothesis about the most important factors potentially driving selection. Diseases caused by *M. phaseolina* are more prevalent during hot and dry conditions, therefore temperature and precipitation variables were included. The selected climatic variables were: BIO18 = Precipitation of Warmest Quarter,

BIO15 = Precipitation Seasonality (Coefficient of Variation), BIO17 = Precipitation of Driest Quarter, BIO10 = Mean Temperature of Warmest Quarter and BIO4 = Temperature Seasonality (standard deviation *100). Each bioclimatic variable was scaled, centered, and evaluated for inclusion using forward selection with 10,000 permutations using adespatial R package (Dray et al., 2022).

To account for underlying spatial structure (autocorrelation) and reduce spurious GEA, distance-based Moran's eigenvector maps (dbMEM) were generated using sample coordinates in the quickMEM R function (Borcard et al., 2018). The dbMEMs are a matrix of axes that capture spatial patterns from multiple angles rather than just a latitudinal or longitudinal vector. Only significant dbMEM axes were selected using forward selection with 1,000 permutations. A simple RDA model and partial RDA model conditioning on space, using only significant dbMEMs, were used to identify the climatic variables significantly contributing to genomic variation and those structured in space.

4.9 Variance partitioning and outlier loci identification

To identify potentially adaptive loci, associations between genetic data (loci) and climatic variables hypothesized to drive selection were evaluated using a multivariate method, redundancy analysis (RDA, as implemented by (Capblancq and Forester, 2021)). RDA simultaneously tests multiple loci that covary in response to climatic variables. Partial RDA models were used for variance partitioning and outlier loci identification while correcting for neutral genetic population structure. Variance partitioning analysis was performed with linkage-disequilibrium (LD)-filtered ($r^2 > 0.9$) dataset of 11,421 SNPs. The independent contribution of each set of explanatory variables: climate, neutral population structure or space, was assessed while removing the effect of the remaining variable sets using partial RDA. In outlier loci identification, using a partial RDA is recommended to reduce the number of false-positive detections particularly in scenarios of multilocus adaptation when selective agents are unknown (Forester et al., 2018). On the other hand, partial RDA can lead to high false-negative detections when variance is confounded between climatic variables and neutral population structure (Capblancq and Forester, 2021). Candidate adaptive loci were identified using simple and partial RDA models to examine the extent of this issue. A partial RDA model conditioning on neutral population genetic structure was used for candidate outlier SNPs detection. Outlier loci were identified in the three significant constrained axes as the SNPs having loadings ± 3 or ± 4 SD from the mean score of each constrained axis using both the LD-filtered set of 11,421 SNPs and the full set of 76,981 SNPs, respectively (Lasky et al., 2012; Forester et al., 2018). A simple RDA model, without correcting for population structure, using the LD-filtered set of 11,421 SNPs and outlier loci were identified in the three significant constrained axes as the SNPs having loadings ± 3 SD from the mean score. Gene annotations for the significant candidate SNPs were used to investigate putative adaptive functions, using the Macphal1-annotated vcf.

Data availability statement

Raw sequencing data for the 95 *M. phaseolina* WGS samples can be accessed at the NCBI SRA accessions SAMN34106355 to SAMN34106449 (<http://www.ncbi.nlm.nih.gov/bioproject/953167>). The accession number(s) can be found in the article [Supplementary Material](#). The code and data for this study are available at https://github.com/vivianaortiz/Macrophomina_popen.

Author contributions

VO: Conceptualization, Methodology, Investigation, Data curation, Formal analysis, Visualization, Writing - original draft. H-XC: Conceptualization, Supervision, Writing - review and editing. HS: Methodology, Writing - review and editing. JJ: Methodology, Resources, Data curation, Project administration, Writing - review and editing. DM, RB, FM, CE, KW, GM: Resources, Data curation, Writing - review and editing. MC: Conceptualization, Supervision, Resources, Project administration, Funding acquisition, Writing - review and editing. All authors contributed to the article and approved the submitted version.

Funding

This work was supported in part by the United Soybean Board, Michigan Soybean Committee, North Central Soybean Research Program, Colombian Department of Science, Technology and Innovation (COLCIENCIAS), and Project GREEN.

References

- Abramova, N. E., Cohen, B. D., Sertil, O., Kapoor, R., Davies, K. J. A., and Lowry, C. V. (2001). Regulatory mechanisms controlling expression of the DAN/TIR mannoprotein genes during anaerobic remodeling of the cell wall in *Saccharomyces cerevisiae*. *Genetics* 157, 1169–1177. doi:10.1093/genetics/157.3.1169
- Aguilera, J., Ranz-Gil, F., and Prieto, J. A. (2007). Cold response in *Saccharomyces cerevisiae*: New functions for old mechanisms. *FEMS Microbiol. Rev.* 31, 327–341. doi:10.1111/j.1574-6976.2007.00066.x
- Allen, T. W., Bradley, C. A., Sisson, A. J., Byamukama, E., Chilvers, M. I., Coker, C. M., et al. (2017). Soybean yield loss estimates due to diseases in the United States and Ontario, Canada, from 2010 to 2014. *Plant Health Prog.* 18, 19–27. doi:10.1094/php-rs-16-0066
- Andrews, S. (2010). *FastQC: A quality control tool for high throughput sequence data*. Babraham, UK: Babraham Institute.
- Arias, R. S., Ray, J. D., Mengistu, A., and Scheffler, B. E. (2011). Discriminating microsatellites from *Macrophomina phaseolina* and their potential association to biological functions. *Plant Pathol.* 60, 709–718. doi:10.1111/j.1365-3059.2010.02421.x
- Baird, R. E., Watson, C. E., and Scroggs, M. (2003). Relative longevity of *Macrophomina phaseolina* and associated mycobiota on residual soybean roots in soil. *Plant Dis.* 87, 563–566. doi:10.1094/PDIS.2003.87.5.563
- Baird, R. E., Wadl, P. A., Allen, T., McNeill, D., Wang, X., Moulton, J. K., et al. (2010). Variability of United States isolates of *Macrophomina phaseolina* based on simple sequence repeats and cross genus transferability to related genera within botryosphaeriaceae. *Mycopathologia* 170, 169–180. doi:10.1007/s11046-010-9308-3
- Basandrai, A. K., Pandey, A. K., Somta, P., and Basandrai, D. (2021). *Macrophomina phaseolina*-host interface: Insights into an emerging dry root rot pathogen of mungbean and urdbean, and its mitigation strategies. *Plant Pathol.* 70, 1263–1275. doi:10.1111/ppa.13378
- Batista, E., Lopes, A., and Alves, A. (2021). What do we know about botryosphaeriaceae? An overview of a worldwide curated dataset. *Forests* 12, 313. doi:10.3390/f12030313
- Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., and Hane, J. K. (2018). Accessories make the outfit: Accessory chromosomes and other dispensable

Acknowledgments

We thank our colleagues, Teresa Hughes, Frank Martin, and Alemu Mengistu listed in [Supplementary Table S1](#), for collecting and providing *Macrophomina phaseolina* isolates. We thank Austin McCoy, Emily Roggenkamp, Greg Bonito and Gideon Bradburd for helpful comments on this manuscript.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The supplementary material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fgene.2023.1103969/full#supplementary-material>

- DNA regions in plant-pathogenic fungi. *Mol. Plant-Microbe Interact.* 31, 779–788. doi:10.1094/MPMI-06-17-0135-FI
- Bitocchi, E., Rau, D., Bellucci, E., Rodriguez, M., Murgia, M. L., Gioia, T., et al. (2017). Beans (*Phaseolus* spp) as a model for understanding crop evolution. *Front. Plant Sci.* 8, 722–821. doi:10.3389/fpls.2017.00722
- Bontrager, M., and Angert, A. L. (2018). Genetic differentiation is determined by geographic distance in *Clarkia pulchella*. *bioRxiv*, 1–28. doi:10.1101/374454
- Borcard, D., Gillet, F., and Legendre, P. (2018). "Spatial analysis of ecological data," in *Numerical ecology with R*, 299–367. doi:10.1007/978-3-319-71404-2_7
- Bradburd, G. S., Coop, G. M., and Ralph, P. L. (2018). Inferring continuous and discrete population genetic structure across space. *Genetics* 210, 33–52. doi:10.1534/genetics.118.301333
- Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., and Buckler, E. S. (2007). Tassel: Software for association mapping of complex traits in diverse samples. *Bioinformatics* 23, 2633–2635. doi:10.1093/bioinformatics/btm308
- Bradley, C., and Allen, T. (2014). *Estimates of soybean yield reductions caused by diseases in the United States*. University of Illinois Outreach and Extension 2014.
- Bradley, C. A., and del Río, L. E. (2007). First report of charcoal rot on soybean caused by *Macrophomina phaseolina* in north Dakota. *Plant Dis.* 87, 601. doi:10.1094/PDIS.2003.87.5.601B
- Bradley, C. A., Allen, T. W., Sisson, A. J., Bergstrom, G. C., Bissonnette, K. M., Bond, J., et al. (2021). Soybean yield loss estimates due to diseases in the United States and Ontario, Canada, from 2015 to 2019. *Plant Health Prog.* 22, 483–495. doi:10.1094/php-01-21-0013-rs
- Branco, S., Gladieux, P., Ellison, C. E., Kuo, A., LaButti, K., Lipzen, A., et al. (2015). Genetic isolation between two recently diverged populations of a symbiotic fungus. *Mol. Ecol.* 24, 2747–2758. doi:10.1111/mec.13132
- Branco, S., Bi, K., Liao, H.-L., Gladieux, P., Badouin, H., Ellison, C. E., et al. (2016). Continental-level population differentiation and environmental adaptation in the mushroom *Suillus brevipes*. *Mol. Ecol.* 26, 2063–2076. doi:10.1111/mec.13892

- Brown, A. H. D., Feldman, M. W., and Nevo, E. (1980). Multilocus structure of natural populations of *Hordeum spontaneum*. *Genetics* 96, 523–536. doi:10.1093/genetics/96.2.523
- Burkhardt, A. K., Childs, K. L., Wang, J., Ramon, M. L., and Martin, F. N. (2019). Assembly, annotation, and comparison of *Macrophomina phaseolina* isolates from strawberry and other hosts. *BMC Genomics* 20, 802. doi:10.1186/s12864-019-6168-1
- Burt, A., Carter, D. A., Koenig, G. L., White, T. J., and Taylor, J. W. (1996). Molecular markers reveal cryptic sex in the human pathogen *Coccidioides immitis*. *Proc. Natl. Acad. Sci. U. S. A.* 93, 770–773. doi:10.1073/pnas.93.2.770
- Capblancq, T., and Forester, B. R. (2021). Redundancy analysis: A Swiss army knife for landscape genomics. *Methods Ecol. Evol.* 12, 2298–2309. doi:10.1111/2041-210x.13722
- Casano, S., Cotan, A. H., Delgado, M. M., García-Tejero, I. F., Saavedra, O. G., Puig, A. A., et al. (2018). First report of charcoal rot caused by *Macrophomina phaseolina* on hemp (*Cannabis sativa*) varieties cultivated in southern Spain. *Plant Dis.* 102, 1665. doi:10.1094/pdis-02-18-0208-pdn
- Caye, K., Deist, T. M., Martins, H., Michel, O., and François, O. (2016). TESS3: Fast inference of spatial population structure and genome scans for selection. *Mol. Ecol. Resour.* 16, 540–548. doi:10.1111/1755-0998.12471
- Cingolani, P., Patel, V. M., Coon, M., Nguyen, T., Land, S. J., Ruden, D. M., et al. (2012a). Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Front. Genet.* 3, 35–39. doi:10.3389/fgene.2012.00035
- Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012b). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly. (Austin)* 6, 80–92. doi:10.4161/fly.19695
- Cohen, R., Elkabetz, M., Paris, H. S., Gur, A., Dai, N., Rabinovitz, O., et al. (2022). Occurrence of *Macrophomina phaseolina* in Israel: Challenges for disease management and crop germplasm enhancement. *Plant Dis.* 106, 15–25. doi:10.1094/PDIS-07-21-1390-FE
- Croll, D., and McDonald, B. A. (2016). The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems. *Mol. Ecol.* 26, 2027–2040. doi:10.1111/mec.13870
- Crous, P. W., Slippers, B., Groenewald, J. Z., and Wingfield, M. J. (2017). Botryosphaeriaceae: Systematics, pathology, and genetics. *Fungal Biol.* 121, 305–306. doi:10.1016/j.funbio.2017.01.003
- Cummings, J. A., and Bergstrom, G. C. (2013). First report of charcoal rot caused by *Macrophomina phaseolina* in soybean in New York. *Plant Dis.* 97, 1506. doi:10.1094/PDIS-03-13-0318-PDN
- Dhingra, O. D., and Sinclair, J. B. (1973). Location of *Macrophomina phaseoli* on soybean plants related to culture characteristics and virulence. *Phytopathology* 63, 934–936. doi:10.1094/Phyto-63-934
- Dhingra, O. D., and Sinclair, J. B. (1974). Survival of *Macrophomina phaseolina* sclerotia in soil: Effects of soil moisture, carbon:nitrogen ratios, carbon sources, and nitrogen concentrations. *Phytopathology* 65, 236–240. doi:10.1094/phyto-65-236
- Dhingra, O. D., and Sinclair, J. B. (1978). *Biology and pathology of Macrophomina phaseolina*. Minas Gerais: Universidade Federal de Vicosa.
- Dray, S., Bauman, D., Blanchet, G., Borcard, D., Clappe, S., Guenard, G., et al. (2022). Package ‘adespatial’. *Ecol. Monogr.* 82, 1–138. doi:10.1890/11-1183.1
- ElAraby, M. E., Kurl, J. E., and Stetina, S. R. (2007). First report of charcoal rot (*Macrophomina phaseolina*) on soybean in Minnesota. *Plant Dis.* 87, 202. doi:10.1094/PDIS.2003.87.2.202C
- Ellison, C. E., Hall, C., Kowbel, D., Welch, J., Brem, R. B., Glass, N. L., et al. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proc. Natl. Acad. Sci.* 108, 2831–2836. doi:10.1073/pnas.1014971108
- Fick, S. E., and Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *Int. J. Climatol.* 37, 4302–4315. doi:10.1002/joc.5086
- Fitzpatrick, M. C., and Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. *Ecol. Lett.* 18, 1–16. doi:10.1111/ele.12376
- Forester, B. R., Lasky, J. R., Wagner, H. H., and Urban, D. L. (2018). Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. *Mol. Ecol.* 27, 2215–2233. doi:10.1111/mec.14584
- Gasch, A. P. (2007). Comparative genomics of the environmental stress response in ascomycete fungi. *Yeast* 24, 961–976. doi:10.1002/yea
- Gibson, M. J. S., and Moyle, L. C. (2020). Regional differences in the abiotic environment contribute to genomic divergence within a wild tomato species. *Mol. Ecol.* 29, 2204–2217. doi:10.1111/mec.15477
- Gladieux, P., Ropars, J., Badouin, H., Branca, A., Aguilera, G., de Vienne, D. M., et al. (2014). Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. *Mol. Ecol.* 23, 753–773. doi:10.1111/mec.12631
- Gluck-Thaler, E., Ralston, T., Konkell, Z., Ocampos, C. G., Ganeshan, V. D., Dorrance, A. E., et al. (2022). Giant starship elements mobilize accessory genes in fungal genomes. *Mol. Biol. Evol.* 39, msac109. doi:10.1093/molbev/msac109
- Goss, E. M., Tabima, J. F., Cooke, D. E. L., Restrepo, S., Fry, W. E., Forbes, G. A., et al. (2014). The Irish potato famine pathogen *Phytophthora infestans* originated in central Mexico rather than the Andes. *Proc. Natl. Acad. Sci.* 111, 8791–8796. doi:10.1073/pnas.1401884111
- Grünwald, N. J., and Hoheisel, G. A. (2006). Hierarchical analysis of diversity, selfing, and genetic differentiation in populations of the oomycete *Aphanomyces euteiches*. *Phytopathology* 96, 1134–1141. doi:10.1094/PHYTO-96-1134
- Grünwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of genotypic diversity data for populations of microorganisms. *Phytopathology* 93, 738–746. doi:10.1094/PHYTO.2003.93.6.738
- Gupta, G. K., Sharma, S. K., and Ramteke, R. (2012). Biology, epidemiology and management of the pathogenic fungus *Macrophomina phaseolina* (tassi) goid with special reference to charcoal rot of soybean (*Glycine max* (L) merrill). *J. Phytopathology* 160, 167–180. doi:10.1111/j.1439-0434.2012.01884.x
- Hanski, I. (1998). Metapopulation dynamics. *Nature* 396, 41–49. doi:10.1038/23876
- Hartman, G. L., Rupe, J. C., Sikora, E. J., Domier, L. L., Davis, J. A., and Steffey, K. L. (2015). “Part I. Infectious diseases,” in *Compendium of soybean diseases and pests*. Fifth Edition, 15–119.
- Hijmans, R. (2022). *raster: Geographic data analysis and modeling*. CRAN. R package version 3.6-20. <https://rspatial.org/raster>.
- Hoban, S., Kelley, J. L., Lotterhos, K. E., Antolin, M. F., Bradburd, G., Lowry, D. B., et al. (2016). Finding the genomic basis of local adaptation: Pitfalls, practical solutions, and future directions. *Am. Nat.* 188, 379–397. doi:10.1086/688018
- IPCC (2021). “Climate change 2022: Impacts, adaptation, and vulnerability,” in *Contribution of working group II to the sixth assessment report of the intergovernmental panel on climate change*. Editors H.-O. Pörtner, D. C. Roberts, M. Tignor, E. S. Poloczanska, K. Mintenbeck, A. Alegría, et al. (UK and New York, NY, USA: Cambridge).
- Islam, M., Haque, M., Islam, M., Emdad, E., Halim, A., Hossen, Q. M., et al. (2012). Tools to kill: Genome of one of the most destructive plant pathogenic fungi *Macrophomina phaseolina*. *BMC Genomics* 13, 493. doi:10.1186/1471-2164-13-493
- Jacobs, J. L., Kelly, J. D., Wright, E. M., Varner, G., and Chilvers, M. I. (2019). Determining the soilborne pathogens associated with root rot disease complex of dry bean in Michigan. *Plant Health Prog.* 20, 122–127. doi:10.1094/php-11-18-0076-s
- Jana, T., Sharma, T. R., and Singh, N. K. (2005). SSR-based detection of genetic variability in the charcoal root rot pathogen *Macrophomina phaseolina*. *Mycol. Res.* 109, 81–86. doi:10.1017/s0953756204001364
- Jin, P., Bulkley, D., Guo, Y., Zhang, W., Guo, Z., Huynh, W., et al. (2017). Electron cryo-microscopy structure of the mechanotransduction channel NOMPC. *Nature* 547, 118–122. doi:10.1038/nature22981
- Jombart, T., and Ahmed, I. (2011). ADEGENET 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* 27, 3070–3071. doi:10.1093/bioinformatics/btr521
- Jombart, T. (2008). ADEGENET: A R package for the multivariate analysis of genetic markers. *Bioinformatics* 24, 1403–1405. doi:10.1093/bioinformatics/btn129
- Kamvar, Z. N., Tabima, J. F., and Grünwald, N. J. (2014). Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2014, 2811–e314. doi:10.7717/peerj.281
- Kamvar, Z. N., Brooks, J. C., and Grünwald, N. J. (2015). Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* 6, 208. doi:10.3389/fgene.2015.00208
- Kelly, J. D. (2010). *The story of bean breeding*. East Lansing, Michigan: Michigan State University, 1–30. BeanCAP & PBG.
- Kendig, S. R., Rupe, J. C., and Scott, H. D. (2000). Effect of irrigation and soil water stress on densities of *Macrophomina phaseolina* in soil and roots of two soybean cultivars. *Plant Dis.* 84, 895–900. doi:10.1094/PDIS.2000.84.8.895
- Knaus, B. J., and Grünwald, N. J. (2017). vcfR: a package to manipulate and visualize variant call format data in R. *Mol. Ecol. Resour.* 17, 44–53. doi:10.1111/1755-0998.12549
- Knox-Davies, P. S. (1965). Pycnidium production by *Macrophomina*. *S. Afr. J. Agric. Sci.* 8, 205–218.
- Koehler, A. M., and Shew, H. D. (2018). First report of charcoal rot of stevia caused by *Macrophomina phaseolina* in North Carolina. *Plant Dis.* 102, 241. doi:10.1094/pdis-05-17-0693-pdn
- Koike, S. T., Arias, R. S., Hogan, C. S., Martin, F. N., and Gordon, T. R. (2016). Status of *Macrophomina phaseolina* on strawberry in California and preliminary characterization of the pathogen. *Int. J. Fruit Sci.* 16, 148–159. doi:10.1080/15538362.2016.1195313
- Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., and Stamatakis, A. (2019). RAxML-NG: A fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* 35, 4453–4455. doi:10.1093/bioinformatics/btz305
- Lasky, J. R., Des Marais, D. L., McKay, J. K., Richards, J. H., Juenger, T. E., and Keitt, T. H. (2012). Characterizing genomic variation of *Arabidopsis thaliana*: The roles of

- geography and climate. *Mol. Ecol.* 21, 5512–5529. doi:10.1111/j.1365-294X.2012.05709.x
- Lasky, J. R., Upadhyaya, H. D., Ramu, P., Deshpande, S., Hash, C. T., Bonnette, J., et al. (2015). Genome-environment associations in sorghum landraces predict adaptive traits. *Sci. Adv.* 1, 14002188–e1400314. doi:10.1126/sciadv.1400218
- Lawson, D. J., van Dorp, L., and Falush, D. (2018). A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nat. Commun.* 9, 3258–3311. doi:10.1038/s41467-018-05257-7
- LeBlanc, N., Cubeta, M. A., and Crouch, J. A. (2021). Population genomics trace clonal diversification and intercontinental migration of an emerging fungal pathogen of boxwood. *Phytopathology* 111, 184–193. doi:10.1094/PHYTO-06-20-0219-FI
- Leyva-Mir, S. G., Velázquez-Martínez, G. C., Tlalpal-Bolaños, B., Alvarado-Gómez, O. G., Tovar-Pedraza, J. M., and Hernández-Arenas, M. (2015). First report of charcoal rot of sugarcane caused by *Macrophomina phaseolina* in Mexico. *Plant Dis.* 99, 553. doi:10.1094/pdis-06-14-0652-pdn
- Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079. doi:10.1093/bioinformatics/btp352
- Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*. doi:10.48550/arXiv.1303.3997
- Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10. doi:10.14806/ej.17.1.200
- McDonald, B. A., and Linde, C. (2002). Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* 40, 349–379. doi:10.1146/annurev.phyto.40.120501.101443
- McDonald, B. A., and Stukenbrock, E. H. (2016). Rapid emergence of pathogens in agro-ecosystems: Global threats to agricultural sustainability and food security. *Philos. Trans. R. Soc. B Biol. Sci.* 371, 20160026. doi:10.1098/rstb.2016.0026
- McGoldrick, L. L., Singh, A. K., Demirkhanyan, L., Lin, T. Y., Casner, R. G., Zakharian, E., et al. (2019). Structure of the thermo-sensitive TRP channel TRP1 from the alga *Chlamydomonas reinhardtii*. *Nat. Commun.* 10, 4180–4212. doi:10.1038/s41467-019-12121-9
- McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The genome analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20, 1297–1303. doi:10.1101/gr.107524.110
- Meena, B. R., Nagendran, K., Tripathi, A. N., Kumari, S., Sagar, V., Gupta, N., et al. (2018). First report of charcoal rot caused by *Macrophomina phaseolina* in Basella alba in India. *Plant Dis.* 102, 1669. doi:10.1094/pdis-01-18-0152-pdn
- Mengistu, A., Ray, J. D., Smith, J. R., and Paris, R. L. (2007). Charcoal rot disease assessment of soybean genotypes using a colony-forming unit index. *Crop Sci.* 47, 2453–2461. doi:10.2135/cropsci2007.04.0186
- Mengistu, A., Smith, J. R., Ray, J. D., and Bellaloui, N. (2011). Seasonal progress of charcoal rot and its impact on soybean productivity. *Plant Dis.* 95, 1159–1166. doi:10.1094/PDIS-02-11-0100
- Mesny, F., Miyauchi, S., Thiergart, T., Pickel, B., Atanasova, L., Karlsson, M., et al. (2021). Genetic determinants of endophytem in the Arabidopsis root mycobiome. *Nat. Commun.* 12, 7227–7315. doi:10.1038/s41467-021-27479-y
- Meyer, W. A., and Sinclair, J. B. (1974). Factors affecting charcoal rot of soybean seedlings. *Phytopathology* 64, 845–849. doi:10.1094/phyto-64-845
- Mihail, J. D., and Taylor, S. J. (1995). Interpreting variability among isolates of *Macrophomina phaseolina* in pathogenicity, pycnidium production, and chlorate utilization. *Can. J. Bot.* 73, 1596–1603. doi:10.1139/b95-172
- Milgroom, M. G., Jiménez-Gasco, M., del, M., Olivares-García, C., Drott, M. T., and Jiménez-Dacutegiaz, R. M. (2014). Recombination between clonal lineages of the asexual fungus *Verticillium dahliae* detected by genotyping by sequencing. *PLoS One* 9, e106740. doi:10.1371/journal.pone.0106740
- Milgroom, M. G., Jiménez-Gasco, M. D. M., Olivares-García, C., and Jiménez-Díaz, R. M. (2016). Clonal expansion and migration of a highly virulent, defoliating lineage of *Verticillium dahliae*. *Phytopathology* 106, 1038–1046. doi:10.1094/PHYTO-11-15-0300-R
- Milgroom, M. G. (1996). Recombination and the multilocus structure of fungal populations. *Annu. Rev. Phytopathol.* 34, 457–477. doi:10.1146/annurev.phyto.34.1.457
- Milgroom, M. G. (2015). *Population biology of plant pathogens: Genetics, ecology, and evolution*. St. Paul, MN, USA: APS Press, The American Phytopathological Society.
- Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci.* 70, 3321–3323. doi:10.1073/pnas.70.12.3321
- Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583–590. doi:10.1093/genetics/89.3.583
- Nishad, I., Srivastava, A. K., Saroj, A., Babu, B. K., and Samad, A. (2018). First report of root rot of *Nepeta cataria* caused by *Macrophomina phaseolina* in India. *Plant Dis.* 102, 2380. doi:10.1094/PDIS-04-18-0558-PDN
- Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al. (2019). *vegan: Community ecology package*. CRAN, 2395–2396.
- Pandey, A. K., and Basandrai, A. K. (2021). Will *Macrophomina phaseolina* spread in legumes due to climate change? A critical review of current knowledge. *J. Plant Dis. Prot.* 128, 9–18. doi:10.1007/s41348-020-00374-2
- Parsa, S., García-Lemos, A. M., Castillo, K., Ortiz, V., López-Lavalle, L. A., Braun, J., et al. (2016). Fungal endophytes in germinated seeds of the common bean, *Phaseolus vulgaris*. *Fungal Biol.* 120, 783–790. doi:10.1016/j.funbio.2016.01.017
- Pereira, T. S., Machado Esquissato, G. N., da Silva Franco, C. C., de Freitas Mathias, P. C., Soares, D. J., Bock, C. H., et al. (2018). Heterokaryosis and diploid formation among Brazilian isolates of *Macrophomina phaseolina*. *Plant Pathol.* 67, 1857–1864. doi:10.1111/ppa.12919
- R Core Team (2019). *R: A language and environment for statistical computing*. Available at: <https://www.R-project.org/>.
- Reznikov, S., Vellicce, G. R., Mengistu, A., Arias, R. S., Gonzalez, V., LisiDe, V., et al. (2018). Disease incidence of charcoal rot (*Macrophomina phaseolina*) on soybean in north-western Argentina and genetic characteristics of the pathogen. *Can. J. Plant Pathology* 40, 423–433. doi:10.1080/07060661.2018.1484390
- Rojas, A., Jacobs, J. L., Napieralski, S., Karaj, B., Bradley, C. A., Chase, T., et al. (2016). Oomycete species associated with soybean seedlings in north America—Part II: Diversity and ecology in relation to environmental and edaphic factors. *Phytopathology* 107, 293–304. doi:10.1094/phyto-04-16-0176-r
- Romero Luna, M. P., Mueller, D., Mengistu, A., Singh, A. K., Hartman, G. L., and Wise, K. A. (2017). Advancing our understanding of charcoal rot in soybeans. *J. Integr. Pest Manag.* 8, 526. doi:10.1093/jipm/pmw020
- Roth, M. G., Webster, R. W., Mueller, D. S., Chilvers, M. I., Faske, T. R., Mathew, F. M., et al. (2021). Integrated management of important soybean pathogens of the United States in changing climate. *J. Integr. Pest Manag.* 11. doi:10.1093/jipm/pmaa013
- Sakalidis, M. L., Slippers, B., Wingfield, B. D., St J Hardy, G. E., and Burgess, T. I. (2013). The challenge of understanding the origin, pathways and extent of fungal invasions: Global populations of the *Neofusicoccum parvum*-*N. Ribis* species complex. *Divers. Distrib.* 19, 873–883. doi:10.1111/ddi.12030
- Saleh, A. A., Ahmed, H. U., Todd, T. C., Travers, S. E., Zeller, K. A., Leslie, J. F., et al. (2010). Relatedness of *Macrophomina phaseolina* isolates from tallgrass prairie, maize, soybean and sorghum. *Mol. Ecol.* 19, 79–91. doi:10.1111/j.1365-294X.2009.04433.x
- Sarr, M. P., Ndiaye, M., Groenewald, J. Z., and Crous, P. W. (2014). Genetic diversity in *Macrophomina phaseolina*, the causal agent of charcoal rot. *Phytopathol. Mediterr.* 53, 559–564. doi:10.14601/Phytopathol
- Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 3, 430–439. doi:10.1038/s41559-018-0793-y
- Savolainen, O., Lascoux, M., and Merilä, J. (2013). Ecological genomics of local adaptation. *Nat. Rev. Genet.* 14, 807–820. doi:10.1038/nrg3522
- Sexton, J. P., Hangartner, S. B., and Hoffmann, A. A. (2014). Genetic isolation by environment or distance: Which pattern of gene flow is most common? *Evol. (N Y)* 68, 1–15. doi:10.1111/evo.12258
- Sexton, Z. F., Hughes, T. J., and Wise, K. A. (2016). Analyzing isolate variability of *Macrophomina phaseolina* from a regional perspective. *Crop Prot.* 81, 9–13. doi:10.1016/j.cropro.2015.11.012
- Short, G. E., Wylie, T. D., and Bristow, P. R. (1980). Survival of *Macrophomina phaseolina* in soil and in residue of soybean. *Phytopathology* 70, 13–17. doi:10.1094/phyto-70-13
- Slippers, B., and Wingfield, M. J. (2007). Botryosphaeriaceae as endophytes and latent pathogens of woody plants: Diversity, ecology and impact. *Fungal Biol. Rev.* 21, 90–106. doi:10.1016/j.fbr.2007.06.002
- Slippers, B., Boissin, E., Phillips, A. J. L., Groenewald, J. Z., Lombard, L., Wingfield, M. J., et al. (2013). Phylogenetic lineages in the botryosphaeriaceae: A systematic and evolutionary framework. *Stud. Mycol.* 76, 31–49. doi:10.3114/sim0020
- Smith, G. S., and Carvil, O. N. (1997). Field screening of commercial and experimental soybean cultivars for their reaction to *Macrophomina phaseolina*. *Plant Dis.* 81, 363–368. doi:10.1094/PDIS.1997.81.4.363
- Smith, J. M., Smith, N. H., O'Rourke, M., and Spratt, B. G. (1993). How clonal are bacteria? *Proc. Natl. Acad. Sci. U. S. A.* 90, 4384–4388. doi:10.1073/pnas.90.10.4384
- Stukenbrock, E. H., Bataillon, T., Dutheil, J. Y., Hansen, T. T., Li, R., Zala, M., et al. (2011). The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Res.* 21, 2157–2166. doi:10.1101/gr.118851.110
- Su, G., Suh, S.-O., Schneider, R. W., and Russin, J. S. (2001). Host specialization in the charcoal rot fungus, *Macrophomina phaseolina*. *Phytopathology* 91, 120–126. doi:10.1094/PHYTO.2001.91.2.120
- Swofford, D. L. (2003). *PAUP*. Phylogenetic analysis using parsimony (*and other methods)*. Sunderland, Massachusetts: Sinauer Associates. Version 4.
- Tabima, J. F., Sondrei, K., Kerio, S., Feau, N., Sakalidis, M. L., Hamelin, R., et al. (2019). Population genomic analyses reveal connectivity via human-mediated transport across *Populus*

plantations in North America and an undescribed sub-population of *Sphaerulina musiva*. *Mol. Plant-Microbe Interact.* 33, 189–199. doi:10.1094/mpmi-05-19-0131-r

Tančić Živanov, S., Dedić, B., Dimitrijević, A., Dušanić, N., Mikić, S., Jocić, S., et al. (2018). First report of charcoal rot on zebra plant (*Aphelandra squarrosa*) caused by *Macrophomina phaseolina*. *Plant Dis.* 102, 2377. doi:10.1094/pdis-03-18-0480-pdn

Wang, W., Liu, Z., Wang, W., and Song, X. (2020). First report of *Macrophomina phaseolina* causing stalk rot of sacha inchi (*Plukenetia volubilis*) in China. *Plant Dis.* 104, 570. doi:10.1094/pdis-06-19-1312-pdn

Wang, L., Liao, B., Gong, L., Xiao, S., and Huang, Z. (2021). Haploid genome analysis reveals a tandem cluster of four HSP20 genes involved in the high-temperature adaptation of *Coriopsis troglodytes*. *Microbiol. Spectr.* 9, 00287211–e28814. doi:10.1128/Spectrum.00287-21

Wingfield, M. J., Brockerhoff, E. G., Wingfield, B. D., and Slippers, B. (2015). Planted forest health: The need for a global strategy. *Science* 349, 832–836. doi:10.1126/science.aac6674

Wrather, J. A., Anderson, T. R., Arsyad, D. M., Tan, Y., Ploper, L. D., Porta-Puglia, A., et al. (2001). Soybean disease loss estimates for the top ten soybean-producing countries in 1998. *Can. J. Plant Pathol.* 23, 115–121. doi:10.1080/07060660109506918

Wu, J., Wang, M., Zhou, L., and Yu, D. (2016). Small heat shock proteins, phylogeny in filamentous fungi and expression analyses in *Aspergillus nidulans*. *Gene* 575, 675–679. doi:10.1016/j.gene.2015.09.044

Xuereb, A., Kimber, C. M., Curtis, J. M. R., Bernatchez, L., and Fortin, M. J. (2018). Putatively adaptive genetic variation in the giant California sea cucumber (*Parastichopus californicus*) as revealed by environmental association analysis of restriction-site associated DNA sequencing data. *Mol. Ecol.* 27, 5035–5048. doi:10.1111/mec.14942

Yang, X. B., and Navi, S. S. (2005). First report of charcoal rot epidemics caused by *Macrophomina phaseolina* in soybean in Iowa. *Plant Dis.* 89, 526. doi:10.1094/PD-89-0526B

Zveibil, A., Mor, N., Gnyem, N., and Freeman, S. (2012). Survival, host-pathogen interaction, and management of *Macrophomina phaseolina* on strawberry in Israel. *Plant Dis.* 96, 265–272. doi:10.1094/PDIS-04-11-0299