**Emergence of semi and mostly clonal lineages of the soybean pathogen *Xylaria necrophora***

Teddy Garcia-Arocaa\*, Jonathan Richardsb, Jeremy Brownb, Paul “Trey” Pricec, Cheryl P. Andamd, and Vinson P. Doyleb\*

a Department of Plant Pathology, University of Nebraska-Lincoln, Lincoln, NE 68583.

b Department of Plant Pathology and Crop Physiology, Louisiana State University, Baton Rouge, Louisiana 70803.

c LSU AgCenter, Macon Ridge Research Station, Winnsboro, Louisiana 71295.

d Department of Biological Sciences, University at Albany, State University of New York, Albany, NY 12222.

\*Corresponding authors:

Vinson P. Doyle ([VDoyle@agcenter.lsu.edu](mailto:VDoyle@agcenter.lsu.edu))

Teddy Garcia-Aroca ([teddy.garcia@unl.edu](mailto:teddy.garcia@unl.edu))

**ORCID**

Teddy Garcia-Aroca: [0000-0002-7567-4363](http://orcid.org/0000-0002-7567-4363)

Paul “Trey” Price: [0000-0002-1004-3616](http://orcid.org/0000-0002-1004-3616)

Jonathan K. Richards: [0000-0001-9342-3595](https://orcid.org/0000-0001-9342-3595)

Jeremy Brown: [0000-0002-1447-8633](https://orcid.org/0000-0002-1447-8633)

Cheryl P. Andam: [0000-0003-4428-0924](https://orcid.org/0000-0003-4428-0924)

Vinson P. Doyle: [0000-0002-2350-782X](http://orcid.org/0000-0002-2350-782X)

**ABSTRACT**

*Xylaria necrophora* is an emerging pathogen that causes a disease known as Taproot Decline (TRD) of soybean [*Glycine max* (L.) Merr], representing a new threat to the soybean industry in the southern United States. *X. necrophora* infects soybean roots, causing necrosis and foliar symptoms such as interveinal chlorosis followed by necrosis. This pathogen has not been reported outside of the southern US, but some specimens found in the 1920’s-1940’s and 2005-2010 as saprophytes in forests and agricultural systems, within and outside the known range of the disease, suggest an extended geographical distribution. Furthermore, those specimens suggested sexual recombination could have occurred in the ancestral populations, because sexual spores (ascospores) were observed, whereas extants with a pathogenic lifestyle are sterile. The genetic diversity of *X. necrophora* populations and its implications on the emergence of the pathogen, its evolutionary potential, and reproductive strategies remain unknown. Here, we focused on addressing these gaps in our knowledge by studying the genetic diversity of *X. necrophora* inthe region by sequencing 162 specimens collected within- and -outside its known distribution using short-read Illumina sequencing. We obtained ~11, 700 single nucleotide polymorphisms (SNPs) and three lineages were differentiated. The levels of recombination among lineages, assessed using the index of association (*IA*) among alleles, suggested one semi-clonal and two mostly clonal lineages of *X. necrophora* are present. Lineage 1, with the lowest *IA* was the most frequent in the region, resembling a semi-clonal population and suggesting higher levels of sexual recombination, whereas the *IA*’s detected in lineages 2 and 3 resembled mostly and entirely clonal populations, respectively. Altogether, these results represent unexpected diversity within the pathogen populations in the southern United States, likely influenced by the levels of either cryptic or historical sexual recombination of this pathogen.

**Key words:** plantdisease, genomics**,** population, recombination.

## **INTRODUCTION**

The appearance of plant pathogens in agricultural systems has been traditionally explained by a combination of different approaches, from anecdotal evidence to patterns of genome evolution 1–3. A significant number of pathogens have co-evolved with their host(s) in long-term interactions that eventually lead to pathogenicity4. As pathogens spread to new geographical regions and infect naïve host species or hosts with varying levels of susceptibility, they continue evolving and adapting to host defenses in order to survive (CITATIONS). These concepts are explained as the endemic and the novel (or introduced) pathogen hypotheses, respectively5. The geographical distribution of emerging pathogens is limited, among other factors, by the distribution of their host(s) or alternative host(s), and their levels of susceptibility/resistance (CITATIONS). From a pathogen developing more effectors or virulence factors to a host or host population becoming more susceptible or developing less effective resistance genes. These host-pathogen interactions determine pathogen distribution in space and time (CITATIONS). Furthermore, pathogen fitness, influenced by its ability to produce viable inoculum and spread to complete its life cycle, and the environmental conditions that are favorable for disease development are known factors that influence the geographical distribution of plant pathogens (CITATIONS). The biogeographical and genetic factors leading to pathogen appearance are often overlooked, limiting our knowledge of the life history and evolution of these pathogens overtime.

Predictions of pathogen spread are impacted by our knowledge of its geographical origin and distribution, and management practices vary greatly depending on these inferences6–8. When managing a novel pathogen, the focus relies on identification and control of the main pathways of spread, highlighting the importance of understanding its migration pathways; An endemic pathogen would be managed by investigating and managing co-factors, synergies, and context dependence6,8–10. For instance, sources of resistance for plant breeding could be found or tested at the center of biodiversity, where differences in pathogenicity levels exist in the source populations. Furthermore, management by quarantine, monitoring, and intercepting new variants can be effective management strategies when center of origin and migration pathways are known6,11–14. High migration rates between populations increase the gene flow and decreases the population diversity, likely reducing the phenotypic variation, including pathogenicity levels (CITATIONS). Therefore, management strategies developed in the region do not have to be tested everywhere, because the pathogen populations would be nearly identical. Alternatively, low migration rates could lead to higher genetic diversity and population structure (CITATIONS). Both population diversity and migration rates remain unknown in *X. necrophora* populations from soybean*.*

*Xylaria necrophora* is an emerging soil-borne fungal pathogen responsible for a disease known as Taproot Decline of soybean (TRD)15,16. For many years, the disease was either not noticed or misdiagnosed because of its resemblance to other plant diseases and physiological disturbances17. But recent discoveries from genetic studies of historical and current specimens suggest *X. necrophora* may have been present in the region for nearly a century. The pathogen has not been reported in soybean fields outside of the United States, despite two collections in the Caribbean Island of Martinique and the Yunnan region of China16. Therefore, the known distribution of *X. necrophora* seems restricted to southern United States, but questions regarding its evolutionary history such as geographical origin, its migration and reproductive strategies remain. When comparing multi-locus DNA sequences of *X. necrophora* isolates from soybean collected across five states in the southern United States, no patterns of genetic structure were observed. This suggests the genetic diversityof the pathogen population is very low when assessed using this method16, limiting our capability to determine variations in pathogen populations or its center of biodiversity, its dispersal mechanism, or its patterns of reproduction. These represents important gaps in our knowledge, limiting our understanding of the evolutionary history and biology of this pathogen. In the context of population genomics, the number of individuals representing populations and the number of loci sampled from the *X. necrophora* genomes were considerably small to determine true levels of genetic diversity, relative to the size of the population and the estimated size of the *X. necrophora* genome (56.9 Mb18). Therefore, more robust genomic data might be needed to address these questions about the evolutionary history, population diversity, and reproductive strategies of this pathogen.

It has been demonstrated that recombination strategies of a certain fungal species can greatly influence the long-term population composition, contributing to both phenotypic and genotypic diversity 19,20. Sexual recombination is relatively complicated in most fungal species, but generally, when sexual recombination is not evident (by observation of sexual spores or detection of mating types), signatures of cryptic or past recombination events can be detected in the pathogen populations21. The low genetic diversity observed in *X. necrophora*, suggest potentially clonal reproduction mechanisms21. Moreover, most of the specimens recovered in the field and pure cultures of the fungus are sterile across different types of media16. The evidence for sexual recombination in *X. necrophora* is limited, but observations of sexual reproductive structures (fertile stromata and ascospores) in recently sequenced saprophytic-specimens found in the forest in the past century, suggest potential for cryptic recombination, which is known to occur in Xylariales16,22. However, asexual structures (sterile stromata, mycelia, and conidia) have been observed in soybean fields almost exclusively (Garcia-Aroca et al. 2021). Except for one historical specimen collected on sugarcane debris, sexual structures of *X. necrophora* have not been observed in agricultural systems (Garcia-Aroca et al. 2021). This is not uncommon, since many pathogens that have been introduced to agricultural systems maintain clonal populations because monoculture allows the pathogen haplotypes to rapidly reproduce clonally, overwhelming the host 23,24. These aspects of pathogen biology and life history are relevant when designing long-term management strategies. For instance, low rates of sexual recombination would result in low levels population diversity, resembling clonal populations with low phenotypic diversity (CITATIONS). This can lead to either specialized strains that are adapted to specific host genotypes or opportunistic lineages that rely on host availability and susceptibility to survive (CITATIONS).

In the current study, we present data describing the evolutionary history of *X. necrophora* in the southern United States. Using whole genome sequences of axenic cultures isolated from diseased plants collected across the region (e.g. AL, AR, LA, MS, MO, and TN), we sought to determine the following: i) Can we differentiate populations of *X. necrophora* in the southern United States or are we facing a single population of the pathogen? ii) If more than one population is found, is the southern USA the center of biodiversity for the taproot decline pathogen?, and iii) What are the recombination strategies across populations of *X. necrophora*? In the context of pathogen biology and evolution, the answers to these questions can contribute to larger efforts to determine the best approaches for developing effective, long-lasting management strategies for TRD.

## **MATERIALS AND METHOD**

**Sampling of symptomatic plants and isolation of *Xylaria necrophora***. Symptomatic soybean plants were collected across the southern United States where the disease has been reported: AL, AR, LA, MS, MO, and TN. The presence of the pathogen was suspected when plants exhibiting foliar symptoms such as interveinal chlorosis followed by necrosis on leaves were pulled from the ground and black, necrotic tissue (fungal stroma) was observed along the main root. Plants were brought to the laboratory in zip-loc bags inside a cooler to maintain the temperatures for fungal survival. Soybean roots were surface-sterilized with a 0.6 percent sodium hypochlorite solution for 10 min, cut into small pieces in the areas where black fungal tissue was observed, then again for 2 mins, rinsed with distilled water and placed on Difco potato dextrose agar (PDA) plates amended with 0.75 mg of streptomycin sulfate and 0.25 mg of chloramphenicol per liter (PDA-CS). When fungal mycelia was found inside the root pith, direct inoculation to PDA-CS plates was performed with a sterilized needle. Plates were incubated at room temperature and monitored daily for fungal growth. Re-isolations or subcuwere conducted when multiple colonies or contamination fungi were observed15,16.

A total of 162 *X. necrophora* isolates from diseased soybean, representing one isolate per plant sampled, were obtained for high throughput sequencing. In addition, a pure culture of a saprophytic *X. necrophora* recovered from the forests in the Caribbean island of Martinique; four historical, saprophytic *X. necrophora* from the forest collected in the 1920’s-1940’s; three *X. arbuscula*’s*,* onecollected in 2020 and two collected before the 2000’s;one *X. arbuscula* var. *plenofisura*, one *X. striata,* one *X. venosula,* and one *X. bambusicola* were included, for a total of 175 isolates/specimens (TABLE 1). Metadata for all isolates/specimens sequenced is provided in supplementary TABLE S1.

**Isolation of nucleic acids**. Fungal tissue from pure cultures was harvested using a sterilized spatula and placed in microcentrifuge safe-lock tubes containing zirconium oxide beads (~0.35 gr of 2.3 mm and ~0.26 gr of 0.5 mm diameter) and 400 uL of the AP1 buffer from a Qiagen DNeasy® Plant Mini DNA extraction kit (Qiagen USA, Germantown, Maryland). The safe-lock tubes were placed in a bullet blender (Next Advance, New York, New York) for 3 min at a speed of 7000 rpm to grind tissue and release genomic DNA into the AP1 buffer. For “historical” specimens, dried tissue from fungal stromata was placed in safe-lock tubes containing zirconium oxide beads (~0.35 gr of 2.3 mm and ~0.26 gr of 0.5 mm diameter), ground to a powder in a bullet blender (Next Advance, New York, New York) for 3 minutes at speed 7. The rest of the protocol for DNA isolation followed the manufacturer recommendations.

Genomic DNA quality and concentrations were estimated using a NanoDrop ND-1000 Spectrophotometer (Thermo Fisher, Waltham, Massachusetts) and a Qubit® 3.0 Fluorometer (Thermo Fisher, Waltham, Massachusetts), following the manufacturers’ protocols.

**Library preparation and next-generation sequencing**. Genomic DNA libraries were prepared using the New England Biolabs (NEB) Next® Ultra™ II FS DNA library prep kit for Illumina, following the manufacturer’s protocol. For DNA inputs higher than 100 ng, DNA was diluted based on the Qubit concentrations to the working concentration of ~15 ng/uL for an input volume of 13 uL (total DNA ~200 ng), using nucleoside free water. Library quality was assessed using a 2100 Agilent Bioanalyzer (Agilent, Santa Clara, CA). Pooled libraries were sent to Admera Health, LLC (South Plainfield, NJ) for WGS sequencing in a MiSeq sequencing platform and paired-end raw reads were obtained from Genohub (<https://genohub.com/>).

**Quality control, variant calling, filtering, and single nucleotide polymorphism (SNP) discovery**. Raw data quality from fastQ files was assessed using FASTQC v0.11.7 25,26 and MULTIQC v1.10.1 27 and the mean quality scores, sequence lengths, percentage GC-content, percentage of duplicate sequences, and percentage adapter content, were used as indicators to assess the overall quality of libraries. Sequences with a quality score lower than 25 and high percentage of adapter sequences for Illumina were trimmed using a sliding window approach in subsequent analyses using TRIMMOMATIC v0.39 28 in the LSU High Performance Computer Clusters (LSU HPC). Trimmed reads were aligned to the MSU\_SB201401 reference genome, GenBank project NPFG01.1 (file: “GCA\_002288965.1\_ASM228896v1\_genomic.fna), because this is the most reliable genome of *X. necrophora*18 that is available at the time of this study, using the Burrows-Wheeler short read alignment component (BWA) v0.7.17 29. Mapped reads were indexed and sorted with SAMTOOLS v1.9 30,31 and duplicates were marked with GATK v4.0.2.1. Variant calling of genotypes was performed on GATK “HaplotypeCaller” function with sorted BAM files, using reference genome sequence index and dictionary created with SAMTOOLS and GATK “CreateSequenceDictionary”, respectively; producing individual genotyped variant calling files (“g.vcf”). Genotyping was done by combining the latter g.vcf files with the “CombineGVCFs” and genotyping with the “GenotypeGVCFs” functions of GATK.

Filtering for missing data and reads in the variant calling file (VCF) was done in VCFTOOLS v0.1.1732, by sequentially applying a minimum genotype quality (“minGQ”) of 20-40, minimum read depth of 2-5 (“minDP”), and maximum of 30-70 percent missing data (“max-missing”). Homozygous sites were extracted using TASSEL v5.20210623 33 by changing the heterozygous sites to unknown (N) with the “homozygous” flag. The recoded dataset was exported as a new VCF file and filtered again under the same VCFTOOLS parameters described above, to remove the newly added missing data. Further filtering for minor allele frequency (“MAF”) of 0.1 remove monomorphic sites and SNPs correlated by more than 10 percent in 50 SNP windows (“indep-pairwise”) to preserve neutral SNPs using PLINK v1.9 34. The filtered dataset with less than 10 percent missing data was chosen for further analyses to reduce computational times and the effect of missing data. The same dataset was used for analyses index of association described below.

**Population structure and genetic diversity analyses**. In order to determine population connectivity and structure, we used a model-based Bayesian clustering method to infer the correct number of clusters, consistent with the number of populations or lineages 8,35,36. The filtered VCF files with SNP datasets were analyzed with STRUCTURE v2.3.4 35,37–39 with 1-20 population clusters (*K*) for initial assessments, and 1-7 K for in-depth analyses to reduce the computational time. In-depth analyses were ran with 1,000,000 iterations of burn-in followed by 10 million iterations of a Markov Chain Monte Carlo (MCMC), using the correlated allele frequencies with no-admixture model 40,41. A more exhaustive analysis was run using PARALLEL-STRUCTURE 42,43 with the same parameters mentioned above for *K* values from 1 to 10, with 3 replicates each. The results were summarized using STRUCTURE HARVESTERv0.6.9444 and the optimal value of *K* was evaluated using the Evanno (ΔK) method 45. These results were also confirmed by running analyses under the same parameters on FASTSTRUCTURE v1.0, useful for the largest SNP datasets, and choosing the number of clusters (K) based on the model that maximizes marginal likelihoods 46. The best number of cluster to divide our SNP datasets obtained from the Evanno method was confirmed with ADMIXTURE v1.3 47. Cross-validation analyses were run in the R package ADEGENET v.2.1.1 48,49 to assess the optimal value of *K* and confirm the values found in STRUCTURE and ADMIXTURE.

The genetic diversity was mapped against the levels of metadata for each isolate/specimen, including year of collection, substrate [although the sampling for different lifestyles (e.g. pathogenic, saprophytic) was limited to historical specimens], and distinct putative populations based on geographical origin and potential biogeographical barriers (i.e. country, state, east or west of the Mississippi river, Miss. river vs non-Miss. River, isolated populations vs. Miss river). Discriminant analyses of principal component (DAPC) were performed using ADEGENET to depict the genetic divergence between populations and subpopulations based on the aforementioned strata. Euclidean genetic distances were determined in POPPR v2.9.0 and ADEGENET, as well as statistics of genetic diversity for initial assessments. Further filtering of missing data was done in POPPR to assess the effect of missing data when removing SNPs and genotypes with more than 10 and 50 percent missing data. After initial assessments to determine the number of clusters with STRUCTURE and ADMIXTURE, the analyses for genetic diversity and principal component analyses were re-run for the newly found lineages. Euclidean genetic distance-based trees for the unweighted pair group method with arithmetic mean (UPGMA) were produced with POPPR, with 100 bootstrap repetitions and a cutoff of 50 for support values. Analyses of molecular variance (AMOVA) by state and by lineages were conducted using the R packages POPPR and ADE4 v1.7-16, with 1000 permutations and the “quasieuclid” method 50.

To estimate population differentiation, pairwise Nei’s GST 51 and Jost’s *D* 52–54 were calculated between hypothesized populations based on geographical origin [i.e. states, east or west of the Mississippi River, isolated populations in Red River (LA), Mississippi River (AR, LA, and MS) AL, MS, and TN, etc), collection year, and lineages determined by STRUCTURE, ADMIXTURE, and post assessments by PCA and DAPC. GST is equivalent of FST for haploid organisms 8,51. Populations with a low number of genotypes (n=1; i.e. MO and Martinique, and historical specimens from LA and FL, n=2 for each state), were removed because of the effect of sample size on these calculations.

**Detecting recombination**. The index of association (*IA*) (Brown, 1980) is used to test if populations are in linkage disequilibrium (LD) or not. This test is useful to determine if populations are clonal (where significant disequilibrium is expected due to linkage among loci) or sexual (where linkage among loci is expected to be minimal). The populations were separated in lineages defined by STRUCTURE v2.3.4 35,37–39 and ADMIXTURE v1.3 47 and *IA* simulated under different expectations was compared to observed *IA* for each lineage, using a similar approach as 6. Initial calculations for each lineage were made using the R package POPPR v2.9.0 55,56. Simulated values of *IA*under clonal (90 percent linked SNPs), partially-clonal (50 percent linked), and sexual (10 percent linked SNPs) were created with ADEGENET v.2.1.3 48,49. Comparisons between observed and simulated *IA* values were made using the R packages DPLYR 0.8.4 57 and AGRICOLAE 1.3-2 58, using ANOVA (*P* ≤ *0.*05) and Tukey’s Honest Significant Difference (HSD) test 59. Simultaneously, we used d as a less biased estimate that accounts for the number of loci sampled in clone corrected datasets 6,55,60. The null hypothesis tested is that alleles observed at different loci are not linked if populations are sexual because alleles recombine freely into new genotypes during the process of sexual reproduction. The *IA* and d were calculated for each population using POPPR.

**Data availability**. Raw, untrimmed sequence reads were deposited to the National Center for Biotechnology (NCBI) Sequence Read Archive (SRA) under the BioProject [PRJNA895869](https://dataview.ncbi.nlm.nih.gov/object/PRJNA895869?reviewer=i4mhq8njt90u8v8rg04b71tnpp) (Supp. TABLE S1). Datasets, metadata, and processing scripts are available in the GitHub repository: <https://github.com/teddyaroca/X_necrophora_pop_gen.git>

## **RESULTS**

**Quality control, SNP discovery and filtering for missing data**. The overall raw data quality assessed from fastQ files showed high (Phred>30) quality scores per sequence most genome files except for specimens 1956\_JDR, BPI584125, and BPI584151, for which Phred was lower than 28 around the 80 bp position. The GC-content was lower than 50 percent for all but 28 read files and the majority did not follow a normal distribution. Sequence duplication levels were high for only 4 read files (1956\_JDR and BPI 584151) and the percentage of overrepresented sequences was also higher than 10 percent the total number of sequences for 1956\_JDR, 1960\_JDR, BPI 584125, and BPI 584189. The percentage of adapter sequences (adapter content) was higher than 20 percent for the vast majority of read files (344/350), but this number was reduced to less than 6 percent for the majority of read files (332/350) after trimming and mapping to the reference genome.

Genomes were mapped to the MSU\_SB201401 reference genome, project NPFG0118, because this is the most reliable genome assembly of *X. necrophora* at the time of this study. A gradient of filtering parameters was used to determine the effect of missing data. Two datasets were assembled after trimming and mapping to the reference genome, one containing 175 individuals, representing 162 *X. necrophora* (DMCC 1467 – DMCC 3866, except DMCC 3828) from soybean collected from 2016-2020 and one specimen collected as a saprophyte in Martinique in 2005 (DMCC 3828), four *X. necrophora* from the forest collected in the 1920’s-1940’s (BPI 583548, from *Gardenia* sp.; BPI 584026 from sugarcane; BPI 584151 and BPI 584151 from undetermined substrates in the forest), and as outgroup species, three *X. arbuscula,* onecollected in 2020 and two collected before the 2000’s*,* one *X. arbuscula* var. *plenofisura*, one *X. striata,* one *X. venosula,* and one *X. bambusicola* (TABLE B.1). The other dataset contained 162 *X. necrophora* from soybean collected from 2016-2020 and one collected as a saprophyte in Martinique in 2005, four *X. necrophora* from the forest collected in the 1920’s-1940’s.

In the dataset containing 175 genotypes, the number of SNPs ranged from 6,256,805 under the least stringent filtering parameters to 1,196 under the most stringent filtering parameters used on VCFTOOLS and PLINK. Whereas the dataset containing 166 *X. necrophora* genotypes contained 2,302,604 SNPs under the least stringent parameters and 437 SNPs under the most stringent parameters. These filtering parameters did not influence the total number of clusters defined by STRUCTURE and ADMIXTURE (Figure 1C).

**Population structure and within population diversity**. No population structure was detected when defining populations *a priori* based on their geographical origin. However, two clusters (K=2) were consistently detected in both STRUCTURE and ADMIXTURE, regardless of filtering parameters or hypothesized population origin, for the most filtered dataset containing 175 genotypes (TABLE 3; FIGURE 7A-B), one cluster containing outgroup genotypes and the other containing 166 *X. necrophora* genotypes*.* When removing outgroup sequences, for the dataset with 166 *X. necrophora* genotypes, two clusters (K=2) were consistently observed across filtering parameters and regardless of hypothesized geographical population origin in STRUCTURE and ADMIXTURE analyses (FIGURE 7).

The results from summaries of STRUCTURE probability outputs from K=1 to K=10 obtained using the Evanno method in STRUCTURE HARVESTER suggested the highest mean log probability (LnP) and delta K (ΔK) were found when K=2, for both the dataset containing 175 genotypes (166 *X. necrophora* + 9 outgroups) and the dataset containing 166 *X. necrophora* genotypes only, regardless of filtering parameters (FIGURE 7). Two clusters were consistently observed for both datasets, regardless of filtering parameters. This result was consistently observed across datasets with different filtering parameters in ADMIXTURE outputs. Each isolate was grouped in its respective cluster when probability of belonging to the assigned cluster, calculated by ADMIXTURE, was higher than 0.5. Further analyses to assess the genetic variation and post assessments of genotype clustering with ADEGENET were run using this assignment to each cluster as another level of the strata. These were considered Lineage 1 and Lineage 2 (cluster 1 and cluster 2, respectively; FIGURE 8). Clustering of genotypes at additional values of K (K=3 and K=4) are provided in FIGURE B.2.

One dataset filtered for minimum genotype quality (min GQ) of 40, minimum read depth (minimum SNP positions of reads supporting a given SNP; min DP) of 5, maximum missing data (max-missing) of 0.3, and minor allele frequency (maf) of 0.1 was used in subsequent analyses (this dataset was also included in STRUCTURE and ADMIXTURE analyses). The raw dataset contained 175 samples, 11,708 variants (SNPs) present in 1570 contigs with 48.88 percent overall missing data. To reduce the overall amount of missing data, the dataset was further filtered in R for loci that were missing more than 10 percent data (meaning that the SNP was present in at least 90 percent of the individuals). The filtered dataset contained 175 individuals, 1570 contigs, and 4,806 variants (SNPs), reducing the dataset to 3,252 SNPs for the dataset containing 160 *X. necrophora* genotypes only (removing populations with n=1 and 9 outgroup genotypes), and the overall amount of missing data to less than 5 percent. The effect of missing data was assessed by comparing filtered and unfiltered datasets (FIGURE B.3).

**Genetic diversity of lineages found in the southern region**. From the above results, the datasets containing 166 *X. necrophora* genotypes were determined to contain 2 main lineages (Lineage 1 and Lineage 2) and these new hypothesized populations were added as additional strata to group indivudals. Euclidean genetic distances suggested additional genetic variation is present in isolates of *X. necrophora*, with potentially 9-14 divergent sub-lineages (FIGURE 10). However, this genetic variation does not correlate directly with the geographic origin of the isolates when compared at the state level (FIGURE 10A-B). Similarly, other ways to *a priori* stratify populations did not correlate with these lineages, including comparisons between east and west of the Mississippi River, Red River vs Mississippi River vs isolated populations in AL, MS, and TN. The outgroup genotypes, *X. arbuscula* (1956\_JDR, 1960\_JDR, BPI 584189, DMCC 3879, and DMCC 4042), *X. arbuscula* var. plenofissura (DMCC 4041), *X. bambusicola* (DMCC 4044), *X. striata* (DMCC 4043), and *X. venosula* (DMCC 4045) formed a distinct cluster from the *X. necrophora* genotypes as expected, providing direction to our analyses. The only historical *X. necrophora* present after filtering (BPI 584026) clustered within lineage 1 determined by STRUCTURE and ADMIXTURE consensus, but unfortunately the genotype was removed when filtering to remove genotypes with more than 10 percent missing data. In the unfiltered dataset, the genetic distance from another historical specimen BPI 584189 was very small (close to zero) compared to the distance from known pathogenic isolates of *X. necrophora.*

The distribution of these lineages in southern United States overlapped in the following ratios: Lineage 1, 33/52 isolates from Louisiana, 28/35 from Arkansas, 24/37 from Mississippi, 16/26 from Alabama, and all 10 from Tennessee. Lineage 2, 19/52 from Louisiana, 7/35 from Arkansas, 13/37 from Mississippi, 10/26 from Alabama, and 0/10 from Tennessee (FIGURE 8; FIGURE 9; TABLE 3). When MO and the putative isolate from the island of Martinique where present in these analyses, both genotypes were assigned to Lineage 1 (FIGURE 8). The consensus of these results for each genotype are also provided in TABLE B.2 (column Hyp\_STRUCTURE/ADM1).

Two clades were observed, one containing outgroup genotypes and the other containing all *X. necrophora* genotypes were observed when comparing genetic distances in the unfiltered dataset containing all genotypes (166 *X. necrophora* and 9 outgroups) (FIGURE B.4)*.* When the same analysis was conducted on the filtered dataset containing 166 *X. necrophora* genotypes only, no clear clustering pattern was observed when mapping the genotypes based on their geographical origin with hypothesized populations being state, east vs west side of the Mississippi River, Red River (LA) vs all isolates near the Miss. River, and isolated populations in MS, AL, and TN. Similarly, no clustering of genotypes based on Euclidean genetic distance was observed by state (FIGURE 10A). The genotypes of *X. necrophora* clustered closer together within lineages as defined by STRUCTURE and ADMIXTURE than by any other hypothesized population level based on geographical origin (FIGURE 10B). More than one clade of genotypes was observed within lineages in these distance-based phylogenies (FIGURE 10B).

**Table 2.** Summary of polymorphic, multi-locus genotypes of X. necrophora by state and lineages determined by STRUCTURE and ADMIXTURE (excluding historical specimens, n=4, and MO and Martinique, n=1).

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **State** | **Lineage (AMD1)** | **Count** |
| 1 | LA | Lin1 | 33 |
| 2 | LA | Lin2 | 19 |
| 3 | AR | Lin1 | 28 |
| 4 | AR | Lin2 | 7 |
| 5 | MS | Lin1 | 24 |
| 6 | MS | Lin2 | 13 |
| 7 | AL | Lin1 | 16 |
| 8 | AL | Lin2 | 10 |
| 9 | TN | Lin1 | 10 |
|  |  | Total= | 160 |

In genetic diversity analyses where sample sizes influenced the calculated statistics, populations of *X. necrophora* were rarefied to a sample size equal to the population with the lowest number of individuals when comparing among states, including populations of *X. necrophora* from the continental United States and discarding one isolate from MO. A total of 160 genotypes were compared for clone correction and the results suggested 160 multi-locus genotypes, all containing polymorphic loci were present in our dataset. However, in a rarefaction curve (FIGURE B.5), a minimum of 10 MLGs were observed for the TN population (all belonging to Lineage 1), representing the minimum number of MLGs that were used in subsequent analyses when sampling size represented a major limitation (e.g. diversity index and simulations of recombination by population). A summary of the number polymorphic MLGs per population is provided in TABLE 3 (excluding MO and Martinique, because n=1 in those populations, and 4 historical genotypes).

Principal component analyses supported the hypothesis of two clusters when strata were separated by lineages defined by ADMIXTURE, but when stratified based on the geographic origin of the genotypes/isolates, no clear clustering in two groups was observed (FIGURE 11A-B). However, both principal component analysis and discriminant analyses of principal components supported the hypothesis of 2 clusters, and showed a clear separation between lineages, with some genotypes assigned to Lineage 1 present in the cluster of Lineage 2 and vice-versa (FIGURE 11C). A clear differentiation between lineages defined by ADMIXTURE was observed when comparing the density of discriminant function 1 and posterior probabilities of membership by lineage (FIGURE 11D). However, 17 genotypes initially defined as Lineage 2 in ADMIXTURE were found with posterior probabilities (PPM) of belonging to Lineage 1 greater than 0.75, including DMCC 2103, DMCC 2108, DMCC2113, DMCC 2124, DMCC 2530, DMCC 2617, DMCC 2624, DMCC 3164, DMCC 3166, DMCC 3168, DMCC 3180, DMCC 3274, DMCC 3429, DMCC 3430, DMCC 3434, DMCC 3445, and DMCC 3829 (FIGURE 11D). Similarly, one individual genotype (DMCC 3230) assigned to Lineage 1 by ADMIXTURE was found with more than 0.5 PPM of belonging to Lineage 2 (FIGURE 11D).

The divergence between lineages was also supported by the results from AMOVA analyses because the variation among lineages was significant (P=0.01), whereas among states was not significantly different (P=0.89) (TABLE 4).

Population differentiation as measured by pairwise comparisons of GST and Jost’s *D* statistics were close to zero between states in both filtered and unfiltered datasets (FIGURE B.6 and B.7). With the exception of TN in the unfiltered dataset containing 160 *Xylaria necrophora* genotypes (FIGURE B.6. A and B.7. B), no significant differentiation was observed among hypothesized populations by geographic origin, year, or populations determined by ADMIXTURE or in post assessments by PCA and DAPC (FIGURE B.6 and B.7). The results from these analyses suggested no population differentiation between states and lineages in filtered datasets. Overall, the values of Jost’s *D* were higher in pairwise comparisons, but ultimately, did not show differentiation between states or lineages, because the values were close to zero.

**Recombination analyses within linages.** The null hypothesis of no association among loci was rejected when comparing d within lineages. As expected, the null distributions of simulated alleles were normal for each lineage, under the hypothesis of sexual recombination. When randomly sampling 999 loci for Lineages 1 and 2 separately, the observed d value was 0.1716576 (P=0.001) for Lineage 1 and 0.072497 (P=0.001) for Lineage 2, with both observed d and P-values falling outside of the normal distribution of the simulated distribution under sexual recombination (FIGURE 12A-B). The null hypothesis of sexual recombination was rejected for both lienages. However, when comparing index of association (*IA*) between each lineage and simulated values for clonal, partially clonal (semi- and mostly-clonal), the *IA*of each lineage was significantly different than the simulated values for partially clonal populations (simulated separately for each lineage). However, in both cases, the lineages were also significantly different than the simulated values for clonal populations (FIGURE 12 C-D). These values were calculated for both unfiltered and minor allele frequency (MAF) of 0.01 filtered datasets (both clone-corrected) to assess the effect of missing data.

The values of *IA*calculated separately for each population by state and lineage, based on the total number of individual genotypes per population differed substantially per lineage, whereas the value of d for each lineage remained the same (TABLE 5), suggesting sample sizes have an important effect on these statistics.

**Table 1**. Statistics calculated with POPPR v2.9.0 for each population by geographical origin (state) and lineages defined by STRUCTURE and ADMIXTURE.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **No** | **Pop** | **N** | **MLG** | **eMLG** | **SE** | **H** | **G** | **lambda** | **E.5** | **Hexp** | **Ia** | **rbarD** |
| 1 | LA | 52 | 52 | 10 | 2.16E-06 | 3.95 | 52 | 0.981 | 1 | 0 | 140.6 | 0.1275 |
| 2 | AR | 35 | 35 | 10 | 8.89E-07 | 3.56 | 35 | 0.971 | 1 | 0 | 27.4 | 0.0903 |
| 3 | MS | 37 | 37 | 10 | 0.00E+00 | 3.61 | 37 | 0.973 | 1 | 0 | 216.6 | 0.2267 |
| 4 | AL | 26 | 26 | 10 | 1.09E-06 | 3.26 | 26 | 0.962 | 1 | 0 | 30 | 0.0749 |
| 5 | TN | 10 | 10 | 10 | 0.00E+00 | 2.3 | 10 | 0.9 | 1 | 0 | 35.2 | 0.2129 |
| 6 | Total | 160 | 160 | 10 | 0.00E+00 | 5.08 | 160 | 0.994 | 1 | 0 | 133 | 0.1107 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| **No** | **Pop** | **N** | **MLG** | **eMLG** | **SE** | **H** | **G** | **lambda** | **E.5** | **Hexp** | **Ia** | **rbarD** |
| 1 | Lin1 | 111 | 111 | 49 | 0.00E+00 | 4.71 | 111 | 0.991 | 1 | 0 | 188.9 | 0.1717 |
| 2 | Lin2 | 49 | 49 | 49 | 0.00E+00 | 3.89 | 49 | 0.98 | 1 | 0 | 31.8 | 0.0725 |
| 3 | Total | 160 | 160 | 49 | 1.55E-05 | 5.08 | 160 | 0.994 | 1 | 0 | 133 | 0.1107 |

Pop= original population or hypothesized lineage

N= number of individuals/genotypes

MLG = multilocus genotypes

eMLG= expected number of multilocus genotypes

SE= standard error

H= Shannon-Weiner Diversity index

G= Stoddard and Taylor's index

Lambda= Simpson's index

E.5= Evenness

Hexp= Nei's gene diversity (expected heterozygosity)

Ia= Index of Association

rbarD= value of the Standardized Index of Association (d)

## **DISCUSSION**

Emerging and re-emerging plant pathogens represent a threat to global food security 10,61. The invasion of exotic pathogens into new environments may be an indirect effect of globalization 62,63. The rate of geographical expansion of a plant pathogen is oftentimes unknown, and potential outbreaks may be avoided with a better understanding of populations dynamics (Fisher et al. 2012). Most recently, the availability of molecular data has provided insights into the life history of many plant pathogens 7. However, in many instances, these studies are not carried out for known plant pathogens; therefore, geographic origin, diversity, migration strategies, recombination rates and evolution potential, genomic regions associated with pathogenicity, and migration across host species remain unknown 8. As discussed earlier, the two competing hypotheses (EPH and NPH) of pathogen emergence are not necessarily exclusive, making these assessments more complicated.

The appearance of invasive plant and fungal species in agricultural ecosystems often involves founder effects, when the invasion event is caused by a single genotype that infects a new host 24,64. These introductions may happen through different mechanisms, including host jumps and tracking, horizontal gene transfer, hybridization, and other evolutionary processes that influence host-pathogen interactions 1,65,66. In most cases, the center of biodiversity offers insight into pathogen life history, infection potential, epidemic propensity, emergence timeline, and potential sources of resistance 7,8,67–69. The origin, distribution of lineages, migration, and recombination strategies represent important aspects of of plant pathogen biology. The center of origin serves as the main geographical area where most of the biodiversity and both, pathogenic and putative, non-pathogenic lineages can be found 8. In plant pathogens such as *Phytophtora infestans*, the assessment of its center of origin has been investigated over time and updated when more sensitive data and evidence became available. Until recently, two competing hypotheses argued that either the South American Andes or central Mexico was its center of origin, ultimately demonstrating the latter to be the most likely scenario 70–73. In this case, population diversity and sexual reproduction in the Mexican isolates helped determine the center of origin 70. For other pathogens, the spread and center of origin has been determined using a combination of historical records, anecdotal evidence, and molecular-level data, such is the case for the grape-vine powdery mildew fungus, *Erysiphe necator* 8,74.

In the case of *Xylaria necrophora*,the pathogen associated with taproot decline of soybean, it represents a non-model organism where genetic diversity has not been studied, except for the phylogeographic evidence presented in Chapter 2 16. The emergence of this pathogen appears to be recent compared to other systems, and a potential explanation for this could be the recent introduction of soybean to North America (1940’s) and the release of modern soybean cultivars (1970’s) 75,76. In the current study, we used a whole genome sequencing approach to compare the genetic variation among *X. necrophora* genotypes collected in the region in an attempt to determine its geographic origin, diversity, main lineages, and recombination potential.

**No population structure observed, but two main lineages detected in the region**

The hypothesized populations based on the geographical origin and terrain landscape did not directly correlate with the distribution of genetic diversity in the region. However, two main lineages determined by STRUCTURE and ADMIXTURE without *a priori* hypotheses were applied to datasets tested in the currentstudy. We named these lineages, Lineage 1 and Lineage 2 to differentiate between these hypothesized lineages. Lineage 1 was more dominant in the region compared to Lineage 2, indicating this lineage either is spread more easily or has become dominant because of some underlying adaptive advantage. Lineage 1 was found more frequently than Lineage 2 in all states, suggesting Lineage 1 has been present in the region longer. For instance, Lineage 2 was not recovered from TN, granted this population was also the one with the lowest number of isolates and multi-locus genotypes (n=10). Despite being the smallest population, the ratio of Lineage 1 to Lineage 2 seemed to be higher than 60% for all other states, suggesting there were chances as high as 40% of collecting one isolate from Lineage 2 in TN. The most balanced ratios of Lineage 1 to Lineage 2 were found in Alabama (62% lineage 1, 38% lineage 2) and Louisiana (63% Lineage 1 and 37% Lineage 2), suggesting either of these states could be where the pathogen started infecting soybean. On the other hand, these lineages do not represent mating types, meaning the ratio of lineages simply provide an estimate of the current diversity levels in the region.

The results from STRUCTURE suggest ongoing gene flow among the population. The directionality of this gene flow is not known but can be futher investigated with these datasets. One plausible hypothesis is that one lineage could contribute to the gene pool of the other lineage. The results from ADMIXTURE supported this hypothesis, since many admixed genotypes were found in our datasets. The majority of lineages showed high probabilities (close to 1) of belonging to the lineage they were assigned. Another indicator that this could be the case is the fact that independent DAPC analyses showed high posterior probability of membership (PPM) to Lineage 1 for 17 genotypes classified as Lineage 2, and high PPM for 1 genotype classified as Lineage 1, belonging to Lineage 2. In this analysis, the majority of genotypes exhibited at least some probability of belonging to the other lineage than the lineage assigned by STRUCTURE and ADMIXTURE, supporting the hypothesis of gene flow present in our dataset. Overall, these results could indicate high migration rates resulting in high gene flow between states and lineages in the region, but specific hypotheses testing the directionality of these migration events should be tested. A potential explanation for the high rate of gene flow could be pathogen spread by infected soybean seeds, but no empirical evidence supporting this hypothesis exist to date.

The results from AMOVA supported the hypothesis of no differentiation among states, but significant differentiation among lineages. In contrast, comparisons of population differentiation statistics GST and Jost’s *D* did not support genetic differentiation within and between lineages as well as states, except for TN in the unfiltered dataset, suggesting potential effects of sample sizes and missing data in our analyses. Regardless of the differences in sample sizes across populations, it has been demonstrated that when calculated separately, GST or Jost’s *D* cannot estimate the levels of true differentiation in models with high mutation rates, and a more reliable approach is to supplement GST with either G’ST or Jost’s *D* 77. This approach was not tested in the current study, but further research should address the differences in filtering algorithms, samples sizes, and supplementing GST with either G’ST or Jost’s *D* for these datasets. Another potential approach that can be used is to compare the observed values of GST with the values produced in permutation tests where random sampling is applied to produce a null distribution 8.

**Genetic distance and principal component analyses suggested more lineages could be present**

The number of lineages detected in STRUCTURE and ADMIXTURE analyses was confirmed by PCA and DAPC analyses. However, our PCA and DAPC suggested more than 2 lineages (or sub-lineages) could be present in the region, because as many as 4 clusters could be observed in PCA clustering. The results obtained in PCA and DAPC analyses could be considered more reliable to determine subclades and provide better insights into the number of clusters present in our genotypic data. The main reason for this assertion is related to the assumptions made by STRUCTURE that are not made by PCA and DAPC analyses. For instance, STRUCTURE assumes Hardy-Weinberg equilibrium and lack of linkage disequilibrium 8, which does not apply for many plant pathogens, including *X. necrophora.* PCA does not make any assumptions and has recently become more popular because it can handle larger SNP datasets 8. On the other hand, both analyses cannot detect FST values lower than 0.05. 8,78–80. FST was not considered a relevant metric in our study, though it was used by STRUCTURE as a distance-based method with the assumption of no admixture to determine the best number of clusters in our dataset, representing another assumption that could have impacted our estimates.

The genetic diversity assessed as the Euclidean genetic distance between genotypes also supported the hypothesis of more clusters than those defined by STRUCTURE and ADMIXTURE analyses. The genetic distance between specimens classified as Lineages 1 and 2 was consistent with placement in STRUCTURE and ADMIXTURE, since the genetic distances were lower within lineage than between lineages (FIGURE 10B). The number of smaller clades observed within lineages suggested more genetic diversity in the region. Based on these analyses 9-14 sub-clades could be determined, suggesting more genetic variation within linages.

**Two main semi-clonal lineages are present in the region**

Permutation tests for d calculated separately for each lineage were used to reject the hypothesis of no association (sexual reproduction), suggesting neither of these lineages are reproducing sexually. Initially, the value, position, and probability of d were used as the main indicator to reject the hypothesis of sexual reproduction, because d was found outside of the null distribution of loci, simulated under the assumption of sexual reproduction (no association among markers). This approach is considered more robust because d is less biased when it comes to differences in sample sizes in the number of loci 60. However, in separate analyses for each lineage, *IA* calculated in permutation tests resembled that of a partially-clonal population. The mean calculated values for Lineages 1 and 2 were significantly lower than that of simulated clonal populations. However, the observed values were not as low as those for simulated semi-clonal and mostly-clonal (FIGURE 12C-D). Simulated values for sexual, semi-clonal, mostly-clonal, and clonal indexes of association were simulated separately for each lineage, based on the number of individuals (Lineage 1=111, Lineage 2=49) and randomly sampling 300 SNPs, with 100 permutations. Since the values were calculated separately for each lineage, the effects of differences in sample sizes were reduced. The effects of missing data were not estimated in this part of the study, but further testing with other datasets could help determine if more variation is observed when other datasets are subjected to the same simulations.

Despite not being within the levels of *IA* considered as partially-clonal (semi-clonal and mostly-clonal) in other studies 6, main lineages of *X. necrophora* appeared to show a semi-clonal mode of reproduction. One historical and one specimen collected in the Caribbean Island of Martinique were found to belong to lineage 1 in STRUCTURE and Euclidean distance analyses. These two specimens exhibited sexual structures in the past, suggesting potential for sexual reproduction. The fact that lineages are not completely clonal suggest that either cryptic sexual reproduction is present in both lineages, which is known to happen in members of the order Xylariales 22, or signs of past sexual recombination were detected in these datasets.

In general, asexual reproduction is considered an evolutionary dead end, because deleterious mutations cannot be removed when recombination is not happening, ultimately reducing the fitness of the organism 8,81–84. Asexual reproduction is a negative description based on the apparent absence of sex, but absence of evidence does not equal evidence for absence 85. In many cases, cryptic sex and recombination occur to purge deleterious mutations 8,19,85. In the case of *X. necrophora* we have both physical and molecular evidence of past sexual recombination, but it appears that sexual reproduction is uncommon or absent in soybean pathogen populations. However, if sexually recombining individuals exist in the population, they could contribute to the observed genetic variation.

Under the NPH, a single introduction of *X. necrophora* could have led to a mostly clonal population.

**On the effect of missing data**

The effects of missing data were tested for the majority of our analyses. When a gradient of filtering parameters for different levels of minimum genotype quality (“minGQ”), minimum SNP positions supporting a SNP at a given position (“min DP”), and maximum missing data per genotype in VCFTOOLS, no effect on the number of clusters determined by the Evanno method from STRUCTURE (based on FST) outputs was observed. However, the analyses were run for 1 million Markov Chain Monte Carlo (MCMC) generations maximum, and the split between clusters after 1 million generations was not tested for any dataset. More combinations of these parameters were not tested because of computational constraints.

The effect of missing data was evident in calculations of population differentiation by GST and Jost’s *D* statistics. The dataset containing *X. necrophora* 162 genotypes, including one genotype from MO and one from the island of Martinique showed high genetic differentiation from the smallest populations (TN, MO, and Martinique) in the unfiltered dataset, highlighting the importance of comparing datasets with the same number of individuals or multi-locus genotypes. When the same dataset was filtered to remove the smallest populations (MO and Martinique), some differentiation could be seen from TN against all populations. But when the same dataset was filtered to remove SNPs with more than 10% missing data across genotypes (keeping all genotypes), no differentiation between states or among lineages was observed, suggesting missing data can influence our inferences.

**Overall conclusions**

Two lineages of *X. necrophora* were detected from populations of the pathogen found in southern United States, with the potential for more sub-populations within these lineages. The center of origin of this pathogen remains unknown, but the presence of two lineages in the southern United States narrows down possible explanations for its appearance. For instance, the pathogen could be novel to soybean and endemic to forests in the region, supporting the EPH, consistent with historical records and specimens associated with TRD collected in the forest before the introduction of modern soybean cultivars. The two lineages could represent two separate adaptations/introductions to soybean from ancestral populations of the pathogen that survived on plant species found in the forest until soybean was introduced as a crop. Alternatively, the pathogen could have recently undergone two different population bottlenecks that reduced the number of lineages and population diversity in the region. Both lineages of *X. necrophora* found in the southern United States were not completely clonal and resembled the expected index of association (*IA*) for partially sexual (mostly clonal) populations, suggesting either cryptic or past sexual recombination, also supporting the EPH. The EPH would be challenged if more genetically diverse populations with higher rates of sexual recombination are reported in soybean pathogen populations found outside of the region covered in this study.

The genetic makeup of the pathogen population(s) could be used to address questions regarding its origin. For instance, the genotypic diversity observed within populations could help determine the center of biodiversity because source pathogen populations are expected to be more diverse than introduced or newly established populations, simply because introduced populations would have a smaller effective population size 74. This is sometimes referred to as genetic bottlenecks or founder effects 8,86,87. For that reason, the genetic diversity is expected to be higher at the center of biodiversity compared to other geographic areas where the pathogen is found, because it contains a genetic pool that accounts for most of the overall variation observed 8,88. This highlights the importance of sampling across the known distribution of the pathogen, to account for most of the genetic variation. In the case of *X. necrophora,* the genetic diversity in the overall population of the pathogen is unknown. One single introduction of *X. necrophora* to an agricultural setting (soybean or any of the other crops from which the pathogen has been recovered, such as corn or cotton) would have led to the spread of the pathogen throughout the region, leading to limited population structure among geographic regions. Under the NPH, the genetic diversity of the pathogen population would be low across the region under study, because a novel pathogen has been introduced to a new host or a new geographic area. Whereas under the EPH, a single pathogen origin is possible, but the genetic diversity would be high at center of origin, which could be found within the region being studied. Since *X. necrophora* has only been reported as as pathogen of soybean in southern United States, the NPH can only be tested within this region. Given the known history of the pathogen, the EPH provides the best explanation at this point. However, if ancient populations of the pathogen were to be found outside of the US, then the NPH hypothesis could be more approtiate to explain the appearance of TRD on soybean in the US.

Population structure, defined as the genetic structure or differentiation between groups of individuals that share similar levels of within-group genetic variation, can be influenced by many factors, including physical separation by distance or both physical and biological barriers, such as oceans, lakes, mountains, rivers, forest, or large agricultural areas 7,8,69,74. The rates of migration between populations would also influence population structure, determining the amount of gene flow between populations 8. Both population structure and levels of migration between *X. necrophora* populations are not known. Population structure for *X. necrophora* could be influenced by factors such as the phenotypic variation among the known host (soybean cultivars) and potential alternative host species, the geographic distribution of the pathogen, and the isolation by distance caused by physical barriers that separate soybean fields. Determining the effect of geographic distance on the levels of genetic differentiation is important for the purposes of elucidating potential mechanisms of migration 8,88. The impact of gene flow between populations is influenced by the rate of migration between populations, influencing genetic diversity 8. Therefore, genetic diversity levels would also be influenced by the migration rates between populations. In principle, low migration rates would increase diversity because of the isolation by distance effect and the opposite would happen if the migration rates are high 89,90. The study of migration in the context of population genetics refers to the historical migrations that can be traced back using genetic signatures 8,21,90. In the case of *X. necrophora* historical and current migrations are unknown, but the potential paths for pathogen transmission are limited to soil or soybean debris movement caused by agricultural practices, such as movement of machinery from infested fields to un-infested fields, because of the lifestyle of the pathogen. Other hypotheses regarding the migration pathways and spread of the pathogen throughout the region remain to be tested.

**ACKNOWLEDGEMENTS**

The authors are grateful for all the support and resources provided by the LSU High Performance Computers and Information Technology personnel.

**FUNDING**

Vinson and Trey should know where most of the funding for this project came from.

**LITERATURE CITED**

1. Stukenbrock, E. H. & McDonald, B. A. The origins of plant pathogens in agro-ecosystems. *Annu. Rev. Phytopathol.* **46**, 75–100 (2008).

2. McDonald, B. A. & Stukenbrock, E. H. Rapid emergence of pathogens in agro-ecosystems: global threats to agricultural sustainability and food security. *Philos. Trans. R. Soc. B Biol. Sci.* **371**, 20160026 (2016).

3. Anderson, P. K. *et al.* Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends Ecol. Evol.* **19**, 535–544 (2004).

4. Stukenbrock, E. H. *et al.* The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. *Genome Res.* **21**, 2157–2166 (2011).

5. Rachowicz, L. J. *et al.* The Novel and Endemic Pathogen Hypotheses: Competing Explanations for the Origin of Emerging Infectious Diseases of Wildlife. *Conserv. Biol.* **19**, 1441–1448 (2005).

6. Shakya, S. K. *et al.* Phylogeography of the wide-host range panglobal plant pathogen Phytophthora cinnamomi. *Mol. Ecol.* (2021) doi:https://doi.org/10.1111/mec.16109.

7. Grünwald, N. J., McDonald, B. A. & Milgroom, M. G. Population Genomics of Fungal and Oomycete Pathogens. *Annu. Rev. Phytopathol.* **54**, 323–346 (2016).

8. Milgroom, M. G. *Population biology of plant pathogens: Genetics, ecology, and evolution*. *APS Press* (2015). doi:10.1094/9780890544525.

9. Almeida, R. P. P. Emerging plant disease epidemics: Biological research is key but not enough. *PLoS Biol.* **16**, 1–5 (2018).

10. Ristaino, J. B. *et al.* The persistent threat of emerging plant disease pandemics to global food security. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–9 (2021).

11. CDC. Addressing emerging infectious disease threats: a prevention strategy for the United States (Executive Summary). *Morb. Mortal. Wkly. Rep. MMWR* **43**, 19 (1994).

12. Goodwin, S. B. Clonal Diversity and Genetic Differentiation of Phytophthora infestans Populations in Northern and Central Mexico . *Phytopathology* vol. 82 955 Preprint at https://doi.org/10.1094/phyto-82-955 (1992).

13. Tooley, P. W. Fitness and Virulence of Phytophthora infestans Isolates from Sexual and Asexual Populations . *Phytopathology* vol. 76 1209 Preprint at https://doi.org/10.1094/phyto-76-1209 (1986).

14. McDonald, B. A. & Linde, C. Pathogen population genetics, evolutionary potential, and durable resistance. *Annu. Rev. Phytopathol.* **40**, 349–379 (2002).

15. Allen, T. *et al.* First description of the causal agent of taproot decline of soybean, an emerging disease in the southern United States. *Plant Health Prog.* **18**, 35–40 (2017).

16. Garcia-Aroca, T. *et al.* *Xylaria necrophora*, sp. nov., is an emerging root-associated pathogen responsible for taproot decline of soybean in the southern United States. *Mycologia* **113**, 326–347 (2021).

17. Garcia-Aroca, T. *et al.* Secondary metabolites produced by Xylaria necrophora are responsible for foliar symptoms associated with taproot decline of soybean. *Plant Health Prog.* (2022) doi:10.1094/PHP-03-22-0021-RS.

18. Sharma, S. *et al.* Draft genome sequence of Xylaria sp., the causal agent of taproot decline of soybean in the southern United States. *Data Brief* **17**, 129–133 (2018).

19. Taylor, J. W., Jacobson, D. J. & Fisher, M. C. The evolution of asexual fungi: Reproduction, speciation and classification. *Annu. Rev. Phytopathol.* **37**, 197–246 (1999).

20. Taylor, J. W., Hann-Soden, C., Branco, S., Sylvain, I. & Ellison, C. E. Clonal reproduction in fungi. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 8901–8908 (2015).

21. Milgroom, M. G., Jiménez-Gasco, M. D. M., Olivares-García, C., Drott, M. T. & Jiménez-D́iaz, R. M. Recombination between clonal lineages of the asexual fungus *Verticillium dahliae* detected by genotyping by sequencing. *PLoS ONE* **9**, (2014).

22. Robinson, A. J. & Natvig, D. O. Diverse members of the Xylariales lack canonical mating-type regions. *Fungal Genet. Biol.* **122**, 47–52 (2019).

23. Gladieux, P. *et al.* The population biology of fungal invasions. *Mol. Ecol.* **24**, 1969–1986 (2015).

24. Drenth, A., McTaggart, A. R. & Wingfield, B. D. Fungal clones win the battle, but recombination wins the war. *IMA Fungus* **10**, 1–6 (2019).

25. Wingett, S. W. & Andrews, S. Fastq screen: A tool for multi-genome mapping and quality control [version 1; referees: 3 approved, 1 approved with reservations]. *F1000Research* **7**, 1–13 (2018).

26. Andrews, S. FASTQC.: A quality control tool for high throughput sequence data. Preprint at (2010).

27. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).

28. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

29. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

31. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

32. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

33. Bradbury, P. J. *et al.* TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**, 2633–2635 (2007).

34. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

35. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).

36. Stift, M., Kolář, F. & Meirmans, P. G. Structure is more robust than other clustering methods in simulated mixed-ploidy populations. *Heredity* **123**, 429–441 (2019).

37. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).

38. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Dominant markers and null alleles. *Mol. Ecol. Notes* **7**, 574–578 (2007).

39. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).

40. Holland, B. R., Ketelaar-Jones, S., O’Mara, A. R., Woodhams, M. D. & Jordan, G. J. Accuracy of ancestral state reconstruction for non-neutral traits. *Sci. Rep.* **10**, 1–10 (2020).

41. Revell, L. J. Ancestral character estimation under the threshold model from quantitative genetics. *Evolution* **68**, 743–759 (2014).

42. Koc, A. alkc/parallel-structure: (Version v0.6.1). Zenodo. (2021) doi:http://doi.org/10.5281/zenodo.4697229.

43. Tange, O. GNU Parallel 20210622 ('Protasevich’). *Zenodo* doi:https://doi.org/10.5281/zenodo.5013933.

44. Earl, D. A. & vonHoldt, B. M. STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv. Genet. Resour.* **4**, 359–361 (2012).

45. Evanno, G., Regnaut, S. & Goudet, J. Detecting the number of clusters of individuals using the software STRUCTURE: A simulation study. *Mol. Ecol.* **14**, 2611–2620 (2005).

46. Raj, A., Stephens, M. & Pritchard, J. K. FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197**, 573–589 (2014).

47. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

48. Jombart, T. & Ahmed, I. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).

49. Tabima, J. F., Coffey, M. D., Zazada, I. A. & Grünwald, N. J. Populations of phytophthora rubi show little differentiation and high rates of migration among states in the western United States. *Mol. Plant. Microbe Interact.* **31**, 614–622 (2018).

50. Excoffier, L., Smouse, P. E. & Quattro, J. M. Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes: Application. *Genetics* **131**, 479–491 (1992).

51. Nei, M. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci. U. S. A.* **70**, 3321–3323 (1973).

52. Jost, L. GST and its relatives do not measure differentiation. *Mol. Ecol.* **17**, 4015–4026 (2008).

53. Jost, L. *et al.* Differentiation measures for conservation genetics. *Evol. Appl.* **11**, 1139–1148 (2018).

54. Alcala, N. & Rosenberg, N. A. GST′, Jost’s D, and FST are similarly constrained by allele frequencies: A mathematical, simulation, and empirical study. *Mol. Ecol.* **28**, 1624–1636 (2019).

55. Kamvar, Z. N., Tabima, J. F. & Gr̈unwald, N. J. Poppr: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* **2014**, 1–14 (2014).

56. Kamvar, Z. N., Brooks, J. C. & Grünwald, N. J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* **6**, 1–10 (2015).

57. Wickham, H., François, R., Henry, L. & Müller, K. dplyr: A grammar of data manipulation. Preprint at (2018).

58. de Mendiburu, F. agricolae: statistical procedures for agricultural research. Preprint at (2020).

59. Tukey, J. W. Exploratory Data Analysis. *Biometrics* vol. 33 768 Preprint at (1977).

60. Agapow, P. M. & Burt, A. Indices of multilocus linkage disequilibrium. *Mol. Ecol. Notes* **1**, 101–102 (2001).

61. Ristaino, J. B. & Records, A. *Emerging plant diseases and global food security*. *APS Press* (2020). doi:10.1016/B978-0-444-52512-3.00209-6.

62. Fisher, M. C., Garner, T. W. J. & Walker, S. F. Global emergence of Batrachochytrium dendrobatidis and amphibian chytridiomycosis in space, time, and host. *Annu. Rev. Microbiol.* **63**, 291–310 (2009).

63. Fisher, M. C. *et al.* Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**, 186–194 (2012).

64. Kendig, A. E., Svahnström, V. J., Adhikari, A., Harmon, P. F. & Flory, S. L. Emerging fungal pathogen of an invasive grass: Implications for competition with native plant species. *PLoS ONE* **16**, 1–17 (2021).

65. Stukenbrock, E. H. & Croll, D. The evolving fungal genome. *Fungal Biol. Rev.* **28**, 1–12 (2014).

66. Möller, M. & Stukenbrock, E. H. Evolution and genome architecture in fungal plant pathogens. *Nat. Rev. Microbiol.* **15**, 756–771 (2017).

67. O’Hanlon, S. J. *et al.* Recent Asian origin of chytrid fungi causing global amphibian declines. *Science* **360**, 621–627 (2018).

68. Taylor, D. L. *et al.* Accurate Estimation of Fungal Diversity and Abundance Through Improved Lineage-Specific Primers Optimized for Illumina Amplicon Sequencing. *Appl. Environ. Microbiol.* **82**, 78–83 (2016).

69. Milgroom, M. G. & Cortesi, P. Analysis of population structure of the chestnut blight fungus based on vegetative incompatibility genotypes. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10518–10523 (1999).

70. Goss, E. M. *et al.* The Irish potato famine pathogen Phytophthora infestans originated in central Mexico rather than the Andes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 8791–8796 (2014).

71. Yoshida, K. *et al.* The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine. *eLife* **2013**, 1–25 (2013).

72. Grünwald, N. J. *et al.* Population structure of phytophthora infestans in the toluca valley region of central mexico. *Phytopathology* **91**, 882–890 (2001).

73. Grünwald, N. J. & Flier, W. G. The biology of Phytophthora infestans at its center of origin. *Annu. Rev. Phytopathol.* **43**, 171–190 (2005).

74. Brewer, M. T. & Milgroom, M. G. Phylogeography and population structure of the grape powdery mildew fungus, Erysiphe necator, from diverse Vitis species. *BMC Evol. Biol.* **10**, 268 (2010).

75. Gizlice, Z., Carter Jnr, T. E. & Burton, J. W. Genetic base for North American public soybean cultivars released between 1947 and 1988. *Crop Sci.* **34**, 1143–1151 (1994).

76. Gizlice, Z., Carter, T. E. & Burton, J. W. Genetic Diversity in North American Soybean: I. Multivariate Analysis of Founding Stock and Relation to Coefficient of Parentage. *Crop Sci.* **33**, 614–620 (1993).

77. Verity, R. & Nichols, R. A. What is genetic differentiation, and how should we measure it - G ST, D, neither or both? *Mol. Ecol.* **23**, 4216–4225 (2014).

78. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, 2074–2093 (2006).

79. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, (2009).

80. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured. *BMC Genet.* **11**, 1–15 (2010).

81. Gibson, A. K., Delph, L. F. & Lively, C. M. The two-fold cost of sex: Experimental evidence from a natural system. *Evol. Lett.* **1**, 6–15 (2017).

82. Felsenstein, J. The evolutionary advantage of recombination. *Genetics* **78**, 737–756 (1974).

83. Anderson, J. B. & Kohn, L. M. Clonality in Soilborne, Plant-Pathogenic Fungi. *Annu. Rev. Phytopathol.* **33**, 369–391 (1995).

84. Andersson, D. I. & Hughes, D. Muller’s ratchet decreases fitness of a DNA-based microbe. *Proc. Natl. Acad. Sci. U. S. A.* **93**, 906–907 (1996).

85. Schurko, A. M., Neiman, M. & Logsdon, J. M. Signs of sex: what we know and how we know it. *Trends Ecol. Evol.* **24**, 208–217 (2009).

86. Nei, M., Maruyama, T. & Chakraborty, R. The bottleneck effect and genetic variability in populations. *Evolution* **29**, 1–10 (1975).

87. Dlugosch, K. M. M. & Parker, I. M. Founding events in species invasions: Genetic variation, adaptive evolution, and the role of multiple introductions. *Mol. Ecol.* **17**, 431–449 (2008).

88. Grünwald, N. J., Goodwin, S. B., Milgroom, M. G. & Fry, W. E. Analysis of genotypic diversity data for populations of microorganisms. *Phytopathology* **93**, 738–746 (2003).

89. Wright, S. Isolation by distance. *Genetics* **28**, 114–138 (1943).

90. Brewer, M. T., Frenkel, O. & Milgroom, M. G. Linkage Disequilibrium and Spatial Aggregation of Genotypes in Sexually Reproducing Populations of Erysiphe necator. **102**, 997–1005 (2012).

**FIGURES**

**Figure 1.** Results of sensitivity analysis in STRUCTURE v2.3.4 to determine the best number of clusters (K) for the dataset containing 175 genotypes, including 166 *Xylaria necrophora* and 9 outgroups. The dots connected by lines represent the change in Delta K based on the log likelihood when different clusters (K) are used.

**Figure 2.** Clustering of genotypes with ADMIXTURE based on the probability of group membership (Pr) and admixture percentage. Bars represent each genotype sequenced (161 *Xylaria necrophora* from soybean and 1 saprophytic *X. necrophora* from Martinique). On the y-axis, the fill for each bar represents the probability of group membership determined by admixture indexes calculated in ADMIXTURE for K=2. On the x-axis, AL=Alabama, AR=Arkansas, LA= Louisiana, MART= Martinique, MO= Missouri, MS= Mississippi, and TN=Tennessee. n= number of genotypes per original population. Custering for additional values of K are provided in FIGURE B.2.

**Figure 3.** Distribution of pathogenic *Xylaria necrophora* lineages in southern United States and one saprophytic *X. necrophora* from the Caribbean Island of Martinique, assessed with STRUCTURE and ADMIXTURE for the best number of clusters (K=2). Green circles represent Lineage 1 and orange circles represent Lineage 2. Black diamonds represent historical, saprophytic *X. necrophora* specimens also genotyped in this part of the study.

**Figure 4.** Principal component analyses (PCA) and discriminant analysis of principal components (DAPC) for 160 Xylaria necrophora genotypes. A. PCA with clustering by states. B. DAPC with clustering by states. C. PCA with clustering of lineages as defined by lineages found in STRUCTURE and ADMIXTURE. D. Summary of DAPC posterior probability of membership by lineages defined by STRUCTURE and ADMIXTURE. The inertia elipses on panels A-C represent the a priori populations hypothesized outside of PCA and DAPC analyses (states), including lineages determined by the software ADMIXTURE (C). Bars on panel D represent the posterior probability of membership for each genotype calculated under the hypothesis of two lineages.

**Figure 5.** Index of association (IA) and d calculations for 160 Xylaria necrophora genotypes. A. Distribution and calculations of d and its probability in 999 permutations for Lineage 1. B. Distribution and calculations of d and its probability in 999 permutations for Lineage 2. C-D. Simulated IA for sexual, semi-clonal and mostly clonal compared to the IA calculated for lineage 1 (C) and Lineage 2 (D). Simulated values are presented in greyscale.