Genomic comparison of *Geosmithia morbida* isolates from genetically distinct clusters

Denita Hadziabdic[1], Romina Gazis[1], Thomas Lane[1], William Klingeman[2], Bonnie Ownley[1], and Margaret Staton[1]

[1]University of Tennessee, Department of Entomology and Plant Pathology, 370 Plant Biotechnology Building, Knoxville, TN 37996.

[2]University of Tennessee, Department of Plant Sciences, 2431 Joe Johnson Dr., 252 Ellington Plant Sciences Building, Knoxville, TN 37996.

Geosmithia morbida **is a plant pathogenic fungus associated with a disease complex affecting walnuts,** ***Juglans*** **spp.,** **known as thousand cankers disease (TCD). The fungus can move through the phloem, yet it is most effectively spread throughout the tree by its vector,** ***Pityophthorus juglandis***. **TCD was originally described from the western U.S. and now has expanded to the native, eastern range of black walnut in the U.S. TCD has recently been discovered in northwestern Italy on both black and native** ***J. regia***. **If TCD becomes established throughout the native range of black walnut, economic, environmental and social consequences could be significant as the current estimated value of standing** ***J. nigra*** **stock in the U.S. alone is $568 billion. Currently, there is limited knowledge regarding the mechanisms of host-pathogen interactions, and there are no known host resistance mechanisms to TCD. To further our understanding of the pathogenicity of** ***G. morbida***, **we used a comparative genomic approach to characterize regions within the** ***G. morbida*** **genome that have experienced selection and rapid evolution since diverging from a recent non-pathogenic ancestor. The genomes of five** ***G. morbida*** **isolates were sequenced at a depth of 23X-35X coverage and trimmed reads were mapped to the** ***G. morbida*** **reference genome with >93% success rate. More than 127,000 variants were identified across five isolates, with 25,066 in coding regions and over 12,800 variants within putatively pathogenic genes. Combined with a recently published reference genome, the re-sequencing and assembly of representative isolates from our previously identified genetic clusters will enhance the utility of genomic data for comparative research within the genus and across fungal lineages.**

INTRODUCTION

The fungus, *Geosmithia morbida,* vectored by the walnut twig beetle, *Pityophthorus juglandis*, has been associated with devastating disease outbreaks in black walnut, *Juglans nigra,* known as Thousand Cankers Disease (TCD). As beetles and the pathogen move within the phloem and spread throughout the tree, multiple dark brown to black cankers form, coalesce, and girdle twigs and branches, hence the name "thousand cankers" to describe the disease [1]. Fungal colonization eventually destroys the phloem and cambium of the branches and main stem, resulting in nutrient depletion and dieback within three to four years after initial symptoms are detected [1, 2]. Although the genus *Geosmithia* is distributed worldwide, the species, *G. morbida*, and most recently, *G. pallida,* are the first members of the genus to be described as plant pathogens [1, 3, 4].

In their recent study, Zerrillo et al. [5] identified four genetically distinct groups of *G. morbida* that clustered into three geographic regions, with no evidence of sexual reproduction or genetic recombination [5]. This is supported by our recent findings suggesting that *G. morbida* was disseminated to different regions of the U.S. multiple times from different sources [5, 6]. Currently, there is limited scientific knowledge regarding mechanisms of host-pathogen-insect interactions, and there are no known host plant resistance mechanisms for TCD. In this preliminary assessment, we report results of a genome-enabled research approach that is focused on *G. morbida* isolates from five distinct genetic clusters identified with the program STRUCTURE (Fig. 1).

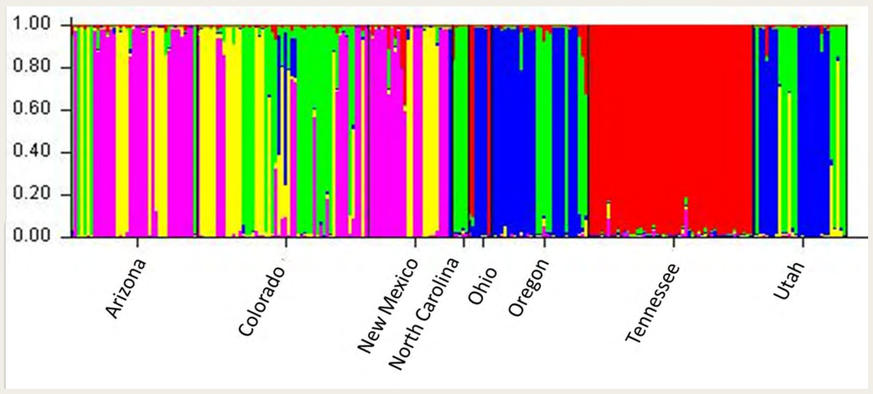

Figure 1. Bar plot with Bayesian assignment probabilities for samples from eight ***Geosmithia morbida*** **subpopulations using program STRUCTURE. The program identified five genetic clusters (k=5), which were used to select a total of 50** ***G. morbida*** **isolates – 10 from each genetic cluster (only 5 isolates presented in the current study).** The portion of each bar that is blue, green, pink, red, or yellow indicates the assignment probability of *G. morbida* individuals to each of the five estimated clusters.

Our long-term project objective will be to sequence, assemble, and annotate additional 45 isolates of *G. morbida* for comparative genomics. The results obtained in this study will provide a better understanding of the biology of the pathogen involved in TCD and identify candidate genes involved in pathogenicity.

MATERIALS AND METHODS

Genomic DNA Extraction and Sequencing. Five out of fifty *G. morbida* isolates were used in this study: GM17 and GM20 from Tennessee, GM108 from Arizona, and GM188 and GM205 from Colorado. Isolates were grown in potato dextrose broth at 25˚C for two months. DNA was extracted with the Qiagen Blood & Cell Culture DNA kit Maxi according to the manufacturer's instructions with the following modifications: (1) incubation time was increased to 3 hours and (2) RNase A and Proteinase K quantities were increased to 100 ul and 3 ml per sample, respectively. Before samples were sequenced, their taxonomic identity was confirmed with the Internal Transcribed Spacer region (ITS) and ITS1 and ITS4 primers [7]. Samples were sequenced at the Genomic Sequencing and Analysis Facility at the University of Texas. Optimal DNA dilutions ranged from 0.2 to 0.6 ng/ul. DNA libraries were prepared using Illumina Bead-In High-Throughput Library Prep Protocol, based on Broad Pond method (PMCID: PMC3091298) and sequenced with the pair-end Illumina HiSeq platform.

Table 1. Mapping results of paired end reads from five isolates of ***Geosmithia morbida*** **against a** ***G. morbida*** **reference genome using Bowtie2.**

Geosmithia morbida Isolate [Location]	Trimmed Read Pairs	Bases	Genome Coverage	# Pairs that Aligned Concordantly exactly 1 time	# Pairs that Aligned Concordantly >1 time	Overall Alignment Rate
GM17 [TN]	3,218,410	936,268,156	35	2,871,697	24,019	94.14%
GM20 [TN]	2,552,274	741,242,553	28	2,324,603	15,837	94.69%
GM108 [AZ]	2,652,848	770,234,802	29	2,374,196	15,986	94.79%
GM188 [CO]	2,108,577	612,441,506	23	1,896,900	9,549	95.66%
GM205 [CO]	2,178,136	634,011,898	24	1,937,449	19,146	93.88%

Data Analyses. The sequence reads were aligned to a California-sourced reference genome of *G. morbida* isolate 1262 [8]. The quality was determined using FastQC version 0.11.4 [9]. Adapter trimming was performed using Trimmomatic version 0.35, and reads shorter than 30 base pairs were discarded [10]. Cleaned reads were mapped to the *G. morbida* reference genome using Bowtie2 version 2.2.7, using the sensitive parameter [8, 11]. The resulting alignment files were sorted using SAMtools version 1.3 [12]. The Genome Analysis Toolkit (GATK) version 3.5-0 best practices were followed for realigning single nucleotide variants (SNVs), realigning indels, and marking duplicates with Picard Tools version 2.1.0 [13, 14]. Variants were called using FreeBayes version 1.0.2-15-g357f175 using a ploidy level of one [15]. Low quality (Phred quality scores lower than 10) variants calls were marked using BCFtools version 1.2 and removed from further analysis [16]. Variants were analyzed using SnpEff version 4.2 in order to determine the impact of each variant on the gene product [17]. Overlaps between known *G. morbida* genes and variants were found using BEDTools version 2.25.0 [18]. A multi-dimensional scaling (MDS) plot was constructed using plink version 1.07 in order to assess variance across the five isolates [19].

RESULTS AND DISCUSSION

We have sequenced the genomes of five *G. morbida* isolates at a depth of 23X-35X coverage (Table 1). An average of 739 megabases of aligned sequence was generated for each library. Trimmed reads were mapped to the *G. morbida* reference genome with >93% success rate (Table 1). Over 127,000 variants were identified across five *G. morbida* isolates, with 25,066 in coding regions. By comparing the list of pathogenic genes provided by Schuelke et al. [8], over 12,800 variants were found within putatively pathogenic genes. Additional profiling of variation in putatively pathogenicity-associated genes revealed a set of 21 mutations that are likely to result in early termination of translation in at least one isolate (Table 2). Further work is ongoing to identify and classify all nonsense and missense mutations in the population. Using all variation, a multi-dimensional scaling (MDS) plot was constructed to assess variance across the five isolates. The two Tennessee isolates (GM 17 and GM 20) clustered closely with Colorado isolate GM 188; while the other Colorado isolate (GM205) and the Arizona isolate (GM108) did not cluster with any other samples.

Table 2. ***Geosmithia morbida*** **variants in putative pathogenicity genes that are predicted to cause an early termination of the coding region.** A '0' denotes that the variant was identical to the reference genome, a '1' denotes that the alternate variant was detected, and NED denotes not enough information to call SNP. PhiBase best matches were determined using BLAST searches of the 3.8 PhiBase database, using an E-value of 1e-6. Information on gene function and effect on pathogenicity are based on the pathogenic species reported in PhiBase.

Gene	GM108	GM17	GM188	GM205	GM20	PhiBase Best Match / Gene name / Pathogenic species	Gene Function	Effect on Pathogenicity
1	1	1	1	1	1	PHI:1784 / GzZC099 / *Fusarium graminearum*	Transcription Factor	Unaffected
2	0	0	1	0	0	PHI:244 / CLAP1 / *Colletotrichum lindemuthianum*	Copper Transporting ATPase	Loss of Pathogenicity
3	1	0	0	0	0	PHI:26 / CaMDR1 / *Candida albicans*	Fungal Efflux Pump	Reduced Virulence
4	1	0	0	0	0	PHI:1238 / FGSG_06970 / *Fusarium graminearum*	Protein Kinase	Lethal
5	1	0	0	0	0	PHI:1681 / GzNF001 / *Fusarium graminearum*	Transcription Factor	Unaffected
6	NED	1	1	NED	1	PHI:1057 / MTP1 / *Magnaporthe oryzae*	Type II Integral Membrane Protein	Unaffected
7	0	0	0	1	0	PHI:1851 / GzZC166 / *Fusarium graminearum*	Transcription Factor	Lethal
8	0	1	0	0	0	PHI:1662 / GzCCHC002 / *Fusarium graminearum*	Transcription Factor	Unaffected
9	0	0	1	0	0	PHI:1057 / MTP1 / *Magnaporthe oryzae*	Type II Integral Membrane Protein	Unaffected
10	1	0	0	1	0	PHI:2654 / DUR1,2 / *Candida albicans*	Urea Metabolism	Reduced Virulence
11	1	0	1	1	0	PHI:2267 / Mls1 / *Parastagonospora nodorum*	Malate Synthase	Loss of Pathogenicity
12	1	1	1	1	1	PHI:2491 / FgPTC1 / *Fusarium graminearum*	Type 2C Protein Phosphatase	Reduced Virulence
13	0	1	0	0	1	PHI:1796 / GzZC111 / *Fusarium graminearum*	Transcription Factor	Unaffected
14	1	0	0	0	0	PHI:3076 / CpATC1 / *Candida parapsilosis*	Acid Trehalase	Reduced Virulence
15	1	0	0	0	0	PHI:1555 / GzMyb019 / *Fusarium graminearum*	Transcription Factor	Unaffected
16	1	0	0	1	0	PHI:2290 / BcBOA6 / *Botrytis cinerea*	Polyketide Synthases	Reduced Virulence
17	1	0	1	1	0	PHI:1681 / GzNF001 / *Fusarium graminearum*	Transcription Factor	Unaffected
18	0	0	1	0	0	PHI:2748 / snf1 / *Ustilago maydis*	Dual Regulator Of Cell Wall Degrading Enzymes	Reduced Virulence
19	1	1	1	1	1	PHI:1180 / Sc_Sat4 / *Fusarium graminearum*	Protein Kinase	Reduced Virulence
20	0	0	0	0	1	PHI:2251 / Gox1 / *Parastagonospora nodorum*	Glyoxalase	Unaffected
21	1	0	0	0	0	PHI:1348 / GzC2H008 / *Fusarium graminearum*	Transcription Factor	Reduced Virulence
Total	**14**	**6**	**9**	**8**	**6**			

In addition to the whole genome sequence of *G. morbida,* sequencing of multiple *G. morbida* isolates with known differential genetic background will enable the detection of structural and functional differences and will help to assess their relation to their virulence. Preliminary results presented here confirm that there is sufficient genomic variation among *G. morbida* isolates to support the hypothesis that pathogenicity varies by genotype and that it is feasible to correlate pathogenicity (phenotype) with genotype. However, additional sequencing of isolates and a robust phenotyping of pathogenicity of those same isolates is needed to correlate gene with function. Completion of this project will provide insight into the evolution and ecology of *G. morbida,* elucidate the genes involved in its pathogenicity, and identify potential mechanisms utilized to overcome host defenses.

THE UNIVERSITY OF TENNESSEE KNOXVILLE

FOREST SERVICE US DEPARTMENT OF AGRICULTURE

UTIA INSTITUTE OF AGRICULTURE THE UNIVERSITY OF TENNESSEE Real. Life. Solutions.

References

1. Kolarik, M., et al., Mycologia, 2011. **103**(2): p. 325-32; 2. Tisserat, N., et al., Plant Health Progress, 2009; 3. Kolarik, M. and L.R. Kirkendall, Fungal Biology, 2010. **114**(8): p. 676-89; 4. Lynch, S.C., et al., Plant Disease, 2014. **98**(9): p. 1276-1276. 5. Zerillo, M., et al., PLoS One, 2014. **9**(11): p. e112847; 6. Hadziabdic, D., et al., Curr Genet, 2014. **60**(2): p. 75-87; 7. White, T.J., et al., 1990, Academic Press: San Diego. 315-322; 8. Schuelke, T.A., et al., PeerJ, 2016. **4**: p. e1952; 9. Andrews, S., Reference Source, 2010. 10. Bolger, A.M., et al. Bioinformatics, 2014; 11. Langmead, B. and S.L. Salzberg, Nature Methods, 2012. **9**(4): p. 357-359. 12. Li, H., et al., Bioinformatics, 2009. **25**(16): p. 2078-2079; 13. Van der Auwera, G.A., et al., 2002, John Wiley & Sons, Inc.; 14. *Picard Tools: A set of tools for working with high-throughput sequencing data;* 15. Garrison, E. and G. Marth, arXiv preprint arXiv:1207.3907, 2012; 16. Danecek, P., et al., Bioinformatics, 2011. **27**(15): p. 2156-2158; 17. Cingolani, P., et al., Fly, 2012 **6**(2): p. 80-92; 18. Quinlan, A.R. and I.M. Hall, Bioinformatics, 2010. **26**(6): p. 841-842; 19. Purcell, S., et al., The American Journal of Human Genetics, 2007. **81**(3): p. 559-575.