Fall 2016

UNDERSTANDING THE EVOLUTION OF PATHOGENICITY WITHIN GEOSMITHIA

Taruna Aggarwal
University of New Hampshire, Durham

UNDERSTANDING THE EVOLUTION OF PATHOGENICITY WITHIN *GEOSMITHIA*

BY

TARUNA AGGARWAL
Bachelor of Sciences, University of California, Davis, 2010

THESIS

Submitted to the University of New Hampshire
In Partial Fulfillment of
The Requirements for the Degree of

Master of Science

in

Genetics

September 2016

This thesis has been examined and approved in partial fulfillment of the requirements for the degree of Master of Science in Genetics by:

Thesis Director
Matthew MacManes, Assistant Professor (Molecular, Cellular, & Biomedical Sciences)

Jeffrey Foster, Assistant Professor (Molecular, Cellular, & Biomedical Sciences)

Kirk Broders, Assistant Professor (Colorado State University, Bioagricultural Sciences & Pest Management)

June 29, 2016___________________________
Date

Original approval signatures are on file with the University of New Hampshire Graduate School.

For mom, dad, Himani, Mary, David, Greg, and James Schuelke

Acknowledgments

I want to thank my advisor, Matt MacManes, for inspiring and nurturing my newfound love for bioinformatics. Matt always encouraged me to do better, and I will remember all the lessons he has taught me during my two years at UNH. He is an exceptional advisor! I am also grateful to Kirk Broders who welcomed me into his lab, allowing me to take the reins of his *Geosmithia* project. I give many thanks to my committee member, Jeff Foster, whose guidance has been indispensable in this work. He was always available to meet with me and discuss my work, even on short notice. A special thanks to Kelley Thomas, Dan Bergeron, and David Plachetzki—these three faculty members have each trained me to be an independent scientist and thinker.

I also want to thank all my friends and colleagues who alleviated the stress of being a graduate student. Finally, I could not have succeeded in graduate school without the support of my family and James Schuelke. I am and always will be grateful to have such an amazing family and life partner.

Table of Contents

Dedication……….…………….……………………………………………….…..………..........iii

Acknowledgements…………………………………………...………………..……..…..……....iv

Table of contents…..…………………………………….………...……...……...………….....…..v

List of tables………………………………………………..…………………..…………….....viii

List of figures…………………………………………..……….………………………..….….......ix

Abstract………………………………...…………………………………...……………..……....xi

Chapter 1 Introduction…..…....………………...………...……………………………….………1

1.1 Brief summary of fungal evolution………………..……………………………..………….1

1.2 Coevolution of bark beetles and their fungal symbionts……..………………..…………3

1.3 Thousand Cankers Disease………………………..………………..……………….…….4

1.4 Population distribution of *Geosmithia morbida*…………………………………...……5

1.5 Thesis objectives……………………………..……………………………………...………6

1.6 Tables and figures…………………………………………………………..……………..……7

1.7 References……………………………………………………………..............……….9

Chapter 2: *De novo* genome assembly of *Geosmithia morbida*, the causal agent of thousand cankers disease…...……………………………………………………………………………….……14

2.1 Introduction……………………………………………………………………………..……14

2.2 Methods……………………...………………………………………………………...……16

2.2.1 DNA extraction and library preparation…………………………….....….16

2.2.2 Preprocessing sequence data…………………..……………………..…........17

2.2.3 *De novo* genome assembly and evaluation………………………..……......17

2.2.4 Structural and functional annotation of *G. morbida* genome……………….18

2.2.5 Assessing repetitive elements profiles…………..…………….............…...18

2.2.6 Identifying putative proteins contributing to pathogenicity……………...…..18

2.3 Results and discussion……………………....……………………………………...……19

2.3.1 Data processing………………....………………………….....…………..….19

2.3.2 Assembly features………………....……………………............…………..…19

2.3.3 Gene annotation……………….....…………….…......……………….……..20

2.3.4 Repetitive elements………………....……………...….......…………………...20

2.3.5 Identifying putative pathogenicity genes……......….......………………..……..20

2.3.6 Identifying putative secreted proteins……..…......……………………..…...21

2.3.7 Identifying protein domains……..….....................……………………...……21

2.4 Conclusion……...…....................……………….................……………...……22

2.5 Tables and figures..................………..…..……23

2.6 References..................………...……………………25

Chapter 3: Comparing *Geosmithia* species to uncover positively selected genes contributing to pathogenicity in *G. morbida*.……………...…………………………………………………...…30

3.1 Introduction..................………..………30

3.2 Methods..................………..………32

3.2.1 DNA extraction and sequencing......................……………………..........…32

3.2.2 Preprocessing sequence data...................………….............................…32

3.2.3 Assembly construction……………………..…………………………….……33

3.2.4 Structural and Functional Annotation.................……….......................…33

3.2.5 Assessing repetitive elements profile.................……….......................…33

3.2.6 Identifying putative genes involved in host-pathogen interactions………...33

3.2.7 Identifying carbohydrate-active proteins and peptidases……..………….....34

3.2.8 Phylogenetic analysis...........................…………..................................…34

3.2.9 Measuring genomic distances using MinHash technique...……………...….35

3.2.10 Detecting genes under positive selection.................………….........…36

3.3 Code availability..................……......................……..……37

3.4 Results and discussion...............................……......……......…...................…37

3.4.1 Assembly features..................…........….………………………………........…37

3.4.2 Identifying putative genes involved in pathogenicity................................…38

3.4.3 Identifying putative secreted proteins..…39

3.4.4 Profiling carbohydrate active enzymes and peptidases...........................…39

3.4.5 Inferring phylogeny...……………........…40

3.4.6 Genomic distances..…........…41

3.4.7 Genes under positive selection...………….........…41

3.4.8 Transmembrane protein and effector genes...…43

3.4.9 Genes under adaptive evolution in beetle-vectored fungal pathogens..…...........…44

3.5 Conclusion...…..........…44

3.6 Tables and Figures..…..........…46

3.7 References...………….........…58

Closing remarks...….........…67

List of tables

Chapter 2: *De novo* genome assembly of *Geosmithia morbida*, the causal agent of thousand cankers disease

Table 2.1
Statistics for *Geosmithia morbida* sequence data……………………………….……………..23

Table 2.2
Geosmithia morbida reference genome assembly statistics generated using QUAST (v2.3)…………………………………………………………………………..……………………..23

Table 2.3
Repetitive elements profile for *Geosmithia morbida*, *Grosmannia clavigera* and *Fusarium solani*……………………………………..…………………………………………..……24

Chapter 3: Comparing *Geosmithia* species to uncover positively selected genes contributing to pathogenicity in *G. morbida*

Table 3.1
Species, geographic regions, *Juglans* host information for *Geosmithia morbida*, *Geosmithia flava*, and *Geosmithia putterillii*……………………………………………………………..……46

Table 3.2
Fungal species used for phylogenetic analysis…………………………………….…………..47

Table 3.3
Statistics for *Geosmithia morbida* isolates, *Geosmithia flava* and *Geosmithia putterillii* sequence data……………………………………….…………………………...………………………………49

Table 3.4
Length-based statistics for *Geosmithia morbida* isolates, *Geosmithia flava*, and *Geosmithia putterillii* generated with QUAST v2.3…………………………………………..……………50

Table 3.5
Repetitive elements profile of *Geosmithia* and 9 additional fungal species generated with RepeatMasker v4.0.5………………………….……………………………………………….…51

Table 3.6
Gene count and PHI-base results for *Geosmithia morbida*, *Geosmithia flava*, *Geosmithia putterillii*, and *Acremonium chrysogenum*……………………..……………………………………51

Table 3.7
Functions of 14 PHI-base proteins that are present only in *Geosmithia morbida*……….............52

List of figures

Chapter 1 Introduction

Figure 1.1
Map illustrating TCD distribution in the United States as of April 2015. The lined states depict regions where TCD has been confirmed and quarantine has been issue in the tan shaded areas. The states that are lined and tan shaded represent areas where disease was detected and quarantine was issued (www.thousandcankers.com)……........……………………………..…….7

Figure 1.2
Map illustrating *Geosmithia morbida* haplotype distribution in the United States. The callouts correspond to three grouped geographic regions (blue=NW_AZ, central CA, northern CA and CO, TN; green=central AZ; red=southwestern CA, OR_WA, southern CO). The shaded wedges in each pie chart represent four genetic clusters (1=blue; 2=red/brown; 3=yellow; 4=green) (Zerillo, et al. 2014)…..……........…………………………..…………………………………………8

Chapter 3: Comparing *Geosmithia* species to uncover positively selected genes contributing to pathogenicity in *G. morbida*

Figure 3.1
Carbohydrate active enzymes (CAZymes) distribution for *Geosmithia* species, other Hypocreales, and *Ceratocystis platani*. The species in red are pathogens, while the names in black are nonpathogens. CAZymes were identified with HMMer searches of dbCAN peptide models. GH: glycoside hydrolases, GT: glycosyltransferases, PL: polysaccharide lyases, CE: carbohydrate esterases, AA: auxiliary activities enzymes, and CBM: carbohydrate-binding molecules. ……………………………………………………………………………..………………53

Figure 3.2
Proteolytic enzymes distribution for *Geosmithia* species, other Hypocreales, and *Ceratocystis platani*. The species in red are pathogens, while the names in black are nonpathogens. Proteases were identified using BLASTp searches against the MEROPs database v10. S: serine, M: metallo, C: cysteine, A: aspartic, T:threonine, I: inhibitors, P: mixed, G: glutamic. ……………………………..………………………………………………………………………54

Figure 3.3
The Bayesian Markov Chain Monte Carlo (BMCMC) phylogeny was estimated using the mixed amino acid model in MrBayes (Ronquist et al. 2012) on a dataset containing 89,999 positions. This topology is identical to partitioned analyses conducted in RAxML (Stamatakis 2014). All nodes in BMCMC and ML analyses receive maximum support. The black circles symbolize classes. The color-shaded boxes at the right of the figure denote the orders within each class. The first and second numbers in parentheses represent the genome sizes in Mbp and the number of predicted protein models, respectively. Black and red branches correspond to non-pathogens and pathogens, which span multiple orders…………………………..………………55

Figure 3.4
The distance matrix illustrated as a heatmap. A value close to 0 indicates high similarity; whereas 1 represents divergent species. The largest distance between two species in the matrix was 0.38. Black and red branches correspond to non-pathogens and pathogens. The distance matrix was generated with Mash (Ondov et al. 2016)……………………………………..…..56

Figure 3.5
Number of shared PHI-base proteins based on Phibase accession IDs among *Geosmithia morbida*, *Geosmithia flava*, *Geosmithia putterillii*, and *Acremonium chrysogenum*……………………………..………………………………………………….………………....57

ABSTRACT

UNDERSTANDING THE EVOLUTION OF PATHOGENICITY WITHIN *GEOSMITHIA*

By

Taruna Aggarwal

University of New Hampshire, September 2016

Geosmithia morbida is a filamentous ascomycete that causes thousand cankers disease in the eastern black walnut tree. This pathogen is commonly found in the western US; however, recently the disease was also detected in several eastern states where the black walnut lumber industry is concentrated. *G. morbida* is one of two known phytopathogens within the genus *Geosmithia*, and it is vectored into the host tree via the walnut twig beetle. We present the first *de novo* draft genome of *G. morbida* (Chapter 2). It is 26.5 Mbp in length and contains less than 1% repetitive elements. The genome possesses an estimated 6,273 genes, 277 of which are predicted to encode proteins with unknown functions. Approximately 31.5% of the proteins in *G. morbida* are homologous to proteins involved in pathogenicity, and 5.6% of the proteins contain signal peptides that indicate these proteins are secreted.

Additionally, the genomes of *Geosmithia flava* and *Geosmithia putterillii* were assembled and compared with *G. morbida* (Chapter 3). The *G. flava* assembly composed of 1,819 scaffolds totaling in 29.47 Mbp in length, and *G. putterillii* genome contained 320 scaffolds consisting of 29.99 Mbp. Our results showed that all three *Geosmithia* species possess similar number of carbohydrate binding enzymes and proteases. We also constructed a Bayesian phylogeny that illustrates the evolutionary relationships between *Geosmithia* and other fungal species. Our phylogeny is consistent with topologies from previous studies.

Lastly, we identified genes under positive selection in *G. morbida* that could potentially contribute to pathogenicity. Our results showed 38 genes under selection in *G. morbida*; none of which were under selection in *G. clavigera*. These findings indicate that species-specific mechanisms might be the driving force behind the evolution of pathogenicity in both of these beetle-vectored fungal pathogens.

Chapter 1
Introduction

1.1 Brief summary of fungal evolution

Fungal species occupy diverse ecological niches that include both mutualistic and pathogenic relationships with their hosts. The latter niche is of particular interest because fungal pathogens cause severe economic and ecological damage that may be irreversible (Pennisi 2010, Fisher et al. 2012). For instance, Oerke (2006) ranked fungi high among the major pests and pathogens that collectively are responsible for 37% of rice and 27% of wheat crop losses worldwide. Fungal species are capable of infecting animals and plants alike, and they not only threaten wildlife and forest health, but also our global food supply (Fisher et al. 2012, Gurr et al. 2011).

Fungal pathogens have evolved a variety of mechanisms for adapting to their host and to dynamic environmental conditions. These adaptations evolve as consequences of random mutations, which can sweep through populations if they render a fitness advantage to the pathogen. It is important to note that several factors drive the evolution of pathogenicity in fungi and no single mechanism is species specific. Two means that propel the appearance of novel traits pertaining to fungal pathogenicity are briefly discussed below and include mobile genetic elements and horizontal gene transfer.

Mobile genetic elements are influential drivers of adaptive evolution (Stukenbrock & Croll 2014, Casacuberta & Gonzalez 2013). For example, de Jonge and colleagues (2013) recorded that specific strains of *Verticillium dahliae* possess lineage specific regions containing effector genes that are flanked by repetitive elements, such as retrotransposons. Additionally, the authors suggested that these repeat rich sequences also contribute to expression regulation because several genes located near these sequences displayed high expression levels during induced infection (de Jonge et al. 2013). Because *V. dahliae* is an asexual pathogen, chromosomal rearrangements allow advantageous mutations to occur that may alternatively result from sexual recombination in sexually reproducing pathogens. Similar findings were uncovered by another study that determined lineage specific regions in *Fusarium oxysporum* represented four additional chromosomes compared to closely related species (Ma et al. 2010). These chromosomes were rich in transposons and genes such as putative effectors, necrosis and ethylene-inducing proteins and carbohydrate binding enzymes (Ma et al. 2010).

In addition to evolution driven by mobile elements, adaptions can also arise from horizontal gene transfer (HGT)—a phenomenon well established in prokaryotes (Koonin et al. 2001). Though speculated to be uncommon, HGT does occur in eukaryotes including fungal species (Mehrabi et al. 2011, Fitzpatrick 2011). For example, Friesen and colleagues (2006) demonstrated that a gene called *ToxA*—encoding for a host-specific toxin—was horizontally transferred from one wheat fungal pathogen (*Stagonospora nodorum*) to another (*Pyrenophora tritici-repentis*). The predicted gene in both species was 99.7% similar at the nucleotide level, and a large 11 kb segment flanking *ToxA* in both fungi was also highly conserved (Friesen et al. 2006).

Another study used phylogenomic analyses to reveal HGT of 20 gene families from fungi to oomycetes and a single HGT from an oomycete to fungal lineage (Richards et al. 2011). These genes have putative functions in plant cell wall degradation, nutrient uptake, and suppression of plant immune response molecules (Richards et al. 2011). Numerous studies have illustrated the occurrence of HGT events in other fungal pathogens, such as *Alternaria alternata* (Hatta et al. 2002, Akagi et al. 2009), *Fusarium oxysporum* (Ma et al. 2010), and *Fusarium solani* (Temporini & VanEtten 2004, Coleman et al. 2009). The precise processes that aid in HGT are not fully understood; however, anastomosis is hypothesized to be involved in HGT events (Mehrabi et al. 2011, Fitzpatrick 2011, Xie et al 2008). Anastomosis is the fusion of vegetative hyphae that can result in exchange of genetic material between two different mycelia (Webster & Weber, 2007).

The mechanisms driving evolution in fungi are not limited to the two aforementioned means (mobile genetic elements and HGT) rather, they also include random mutations, sexual recombination, and epigenetics (Raffaele & Kamoun 2012). Furthermore, evolutionary changes are a product of multiple concomitant mechanisms. This thesis aims to provide insight into the evolution of fungal pathogenicity in relation to these existing proposed mechanisms by using genus *Geosmithia* as a model system. The study species is *Geosmithia morbida*, a bark beetle vectored fungal pathogen that infects eastern black walnut trees.

1.2 Coevolution of bark beetles and their fungal symbionts

Bark beetles belong to the subfamily Scolytinae and are vital forest insects that frequently associate with one or more fungal species. Beetles and fungi play ecologically significant roles in nutrient cycling; however they also can become pathogenic and cause extensive damage to conifers as well as hardwoods (Goheen & Hansen 1993, Paine et al. 1997). While both ambrosia and bark beetles associate with symbiotic fungal partners, mycophagy (fungi feeding) is common among ambrosia beetles and is rare among bark beetles that prefer to reproduce and feed on the nutrient-rich phloem. Nevertheless, four bark beetle genera, including *Dendroctonus*, are known to be mycophagous (Harrington 2005). The *Dendroctonus* genus harbors many of the most destructive and economically important conifer pests (Goheen & Hansen 1993, Harrington 2005). For instance, *D. frontalis* (southern pine beetle), *D. jeffreyi* (Jeffrey pine beetle), *D. brevicomis* (western pine beetle), and *D. ponderosae* (mountain pine beetle) are major forest insect pests in North America that associate with mutualistic fungi (Coyle et al. 2015, Otrosina et al. 1997, Six & Paine 1997, Owen et al. 1987).

The mountain pine beetle (MPB) historically has been found from central British Columbia to northern Mexico and from the pacific coast to southwestern regions of South Dakota (Safranyik et al. 2010). MPB primarily attacks lodgepole pine; however, they are capable of invading ponderosa, sugar, and western white pines. As warmer temperatures increase due to climate change, this beetle is beginning to migrate into eastern parts of both Canada and the

United States, including the Rocky Mountains in Colorado (Carroll et al. 2003, Meddens et al. 2012). MPB are estimated to have caused pine tree mortality in areas greater than 5.1 and 3.4 million hectares in British Columbia (2001-2010) and the western US (1997-2010), respectively (Meddens et al. 2012).

MPB interacts symbiotically with various ophiostomatoid fungi; however, the most threatening among these species is an ascomycete fungus—*Grosmannia clavigera*—that greatly hastens host death (Plattner et al. 2008, Tsui et al. 2012). Much is known about the ecology, population structure, genomics and detoxification methods of *G. clavigera* (Tsui et al. 2012, DiGuistini et al. 2011, Wang et al. 2012). Although important, *G. clavigera* is not the only bark beetle-associated fungus that is phytopathogenic. Another emerging bark beetle pest in the United States is *Pityophthorus juglandis* (walnut twig beetle) that carries the pathogenic fungus, *Geosmithia morbida* to *Juglans* species (Montecchio & Faccoli 2014). *Geosmithia morbida* is the causal agent of Thousand Cankers Disease that was originally detected in *J. nigra* (eastern black walnut) (Kolarik et al. 2011) and the focal species of this research.

1.3 Thousand Cankers Disease

Thousand Cankers Disease (TCD) is caused by the aggressive feeding of the walnut twig beetle (WTB) and its fungal partner, *Geosmithia morbida* (Tisserat et al. 2009). As the disease progresses, large necrotic cankers form in great numbers on branches and tree trunks; hence the name Thousand Cankers Disease (Tisserat et al. 2009). TCD was first documented in Colorado in 2001 when several eastern black walnuts were experiencing elevated levels of tree mortality (Tisserat et al. 2009). The black walnut is native to the eastern US, but it is planted throughout the western part of the country as a decorative tree.

To date, the disease has been detected in 16 states in the US (Figure 1.1) and parts of Europe (Tisserat et al. 2009, Montecchio & Faccoli 2014, Hadziabdic et al. 2014, Juzwik et al. 2016, Zerillo et al. 2014, RugmanJones et al. 2015, ThousandCankersDisease.com). Furthermore, *G. morbida* has been isolated from three wingnut species (*Pterocarya fraxinifolia*, *P. rhoifolia* and *P. stenoptera*), English walnuts (*J. regia*) in California, and butternut (*J. cinerea*) in Oregon (Serdani et al. 2013, Yaghmour et al. 2014, Hishinuma et al. 2016). Currently, infected tree removal is the only proposed method of mitigating the dispersal of TCD. The urgency to develop more effective regulatory mechanisms warrants a better understanding of this disease and its vector complex, which comprises the walnut twig beetle and the pathogenic fungus, *Geosmithia morbida* (ThousandCankersDisease.com).

1.4 Population distribution of *Geosmithia morbida*

G. morbida is a filamentous fungus (Ascomycota: Hypocreales) that was first described by Kolarik and colleagues (2011). Although *G. morbida* primarily causes tree mortality in *J. nigra*, various other *Juglans* species, such as *J. californica*, *J. cinerea*, *J. regia*, and *J. major*, are also susceptible to *G. morbida* based on greenhouse inoculation studies (Utley et al. 2013). The *G. morbida* population distribution in the US is best described as four highly diverse genetic clusters spanning three geographic regions (Zerillo et al. 2014, Figure 1.2). The source of *G. morbida* is unknown; however it is clear that the fungus is native to North America and some hypotheses have been proposed about its origin. Firstly, it has been postulated that WTB and *G. morbida* might be native associates of *J. major* (Arizona walnut tree indigenous to southwestern US), and a host shift from Arizona walnut to a more naïve eastern black walnut took place (Zerillo, et al. 2014). The beetle was recovered from *J. major* in 1896; whereas, WTB or *G. morbida* were not detected in *J. nigra* stands until 1959 in the western US (Cranshaw 2011). Though plausible, this hypothesis regarding a host shift was discounted by Zerillo and colleagues (2014) due to the lack of most common haplotypes in central Arizona and New Mexico regions (Figure 1.2). This suggested that *G. morbida* haplotypes from *J. major* are not ancestral to *G. morbida* populations found throughout other parts of the US. Another hypothesized origin of *G. morbida* is *J. californica*, which may be the native host of WTB and *G. morbida*. Two of the most common haplotypes, namely H02 and H03, were identified in

California and also found in other regions (Figure 1.2). This implied that California populations of *J. californica* might be the source populations hosting WTB and *G. morbida* (Zerillo et al. 2014).

Despite the lack of strong evidence explaining the origin of TCD epidemics in the US, it is likely that the disease spread is a consequence of anthropogenic activities and movement of infested wood (Zerillo et al. 2014). Black walnut trees are highly prized for their lumber quality, with the approximate market value of black walnuts in the US being over $5 billion (USDA-APHIS 2009). In addition, California alone provides 99% of the English walnuts consumed in the US. Therefore, limiting the expansion of TCD into the central and eastern walnut plantations is critical both to the maintenance of the walnut industry and to the central hardwood forest ecosystem health. To this end, our work aims to understand the evolution of pathogenicity within *Geosmithia morbida*, which will provide insight into TCD dispersal and the development of more effective control methods.

1.5 Thesis objectives

The central aim of this research is to determine the molecular mechanisms contributing to the evolution of pathogenicity within *Geosmithia morbida*. In order to meet this aim, we first sequenced, assembled and annotated a reference genome of *G. morbida*. We also characterized the *G. morbida* genome relative to two other fungal pathogens, *Fusarium solani* and *Grosmannia clavigera* (Chapter 1). Next, we identified genes under positive selection in *G. morbida*, and we compared the predicted protein models in *G. morbida* with two other *Geosmithia* species and their closest sister taxa within the order of Hypocreales. Lastly, we performed phylogenetic analyses to identify the evolutionary relationship between *Geosmithia* species and other species in the order Hypocreales (Chapter 2).

1.6 Tables and figures

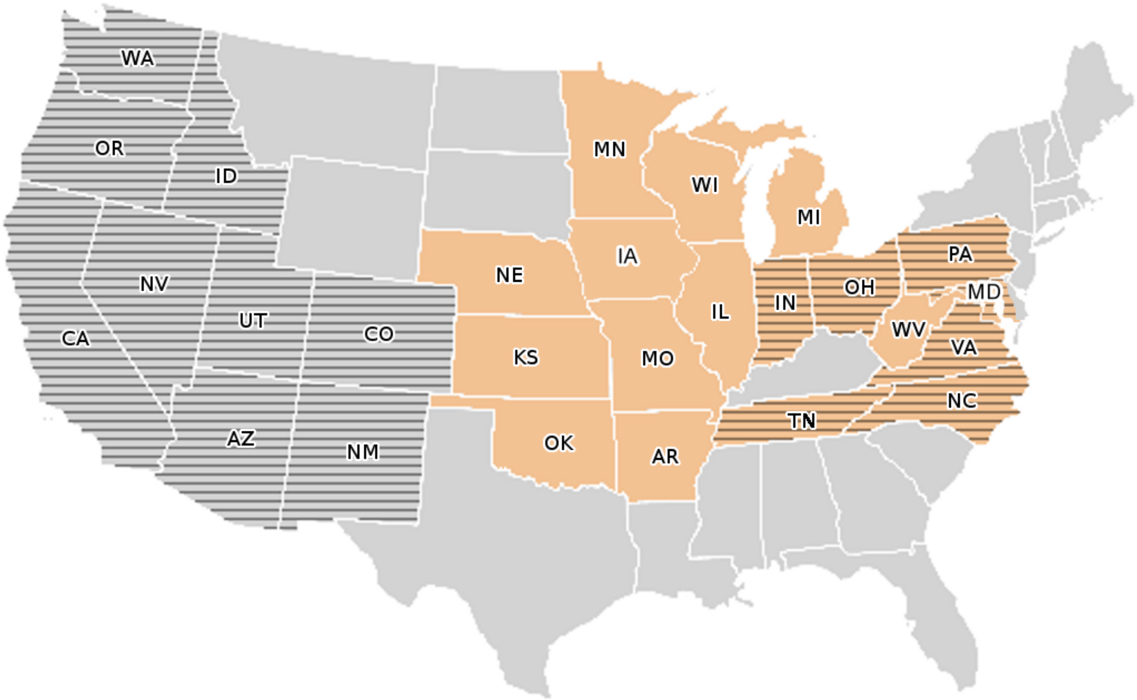

Figure 1.1. Map illustrating TCD distribution in the United States as of April 2015. The lined states depict regions where TCD has been confirmed and quarantine has been issue in the tan shaded areas. The states that are lined and tan shaded represent areas where disease was detected and quarantine was issued (www.thousandcankers.com).

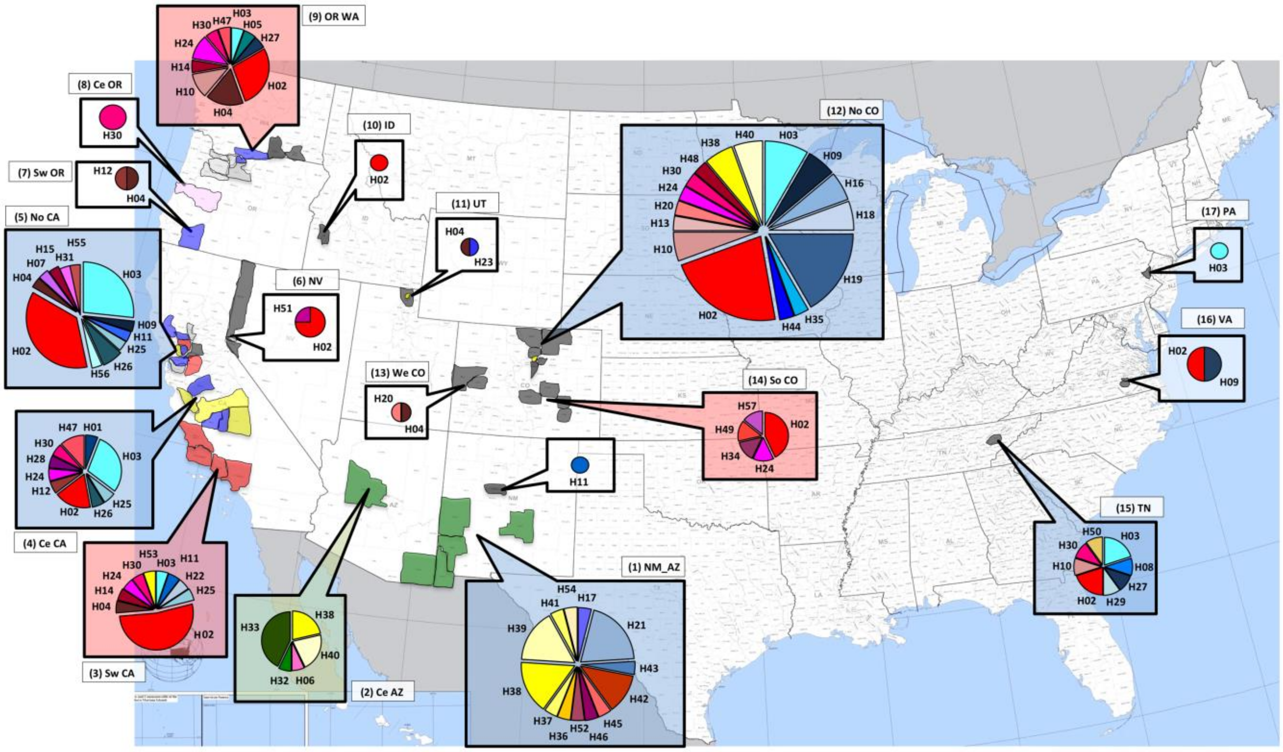

Figure 1.2. Map illustrating *Geosmithia morbida* haplotype distribution in the United States. The callouts correspond to three grouped geographic regions (blue=NW_AZ, central CA, northern CA and CO, TN; green=central AZ; red=southwestern CA, OR_WA, southern CO). The shaded wedges in each pie chart represent four genetic clusters (1=blue; 2=red/brown; 3=yellow; 4=green) (Zerillo et al. 2014).

1.7 References

1. Akagi Y, Akamatsu H, Otani H, Kodama M. 2009. Horizontal chromosome transfer, a mechanism for the evolution and differentiation of a plant-pathogenic fungus. Eukaryot Cell. 8:1732-1738.
2. Carroll AL, Taylor SW, Régnière J, Safranyik L. 2003. Effects of climate change on range expansion by the mountain pine beetle in British Columbia. In: Shore TL, Brooks JE, Stone JE, editors. Mountain Pine Beetle Symposium: Challenges and Solutions. Victoria, BC: Report BC-X-399, Canadian Forest Service, Pacific Forestry Centre. 223–232.
3. Casacuberta E, Gonzalez J. 2013. The impact of transposable elements in environmental adaptation. Mol Ecol. 22:1503-1517.
4. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, et al. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 5:e1000618.
5. Coyle DR, Klepzig KD, Koch FH, Morris LA, Nowak JT, et al. 2015. A review of southern pine decline in North America. Forest Ecol Manag. 349:134-148.
6. Cranshaw W. 2011. Recently recognized range extensions of the walnut twig beetle, *Pityophthorus juglandis* Blackman (Coleoptera: Curculionidae: Scolytinae), in the western United States. Coleopts Bull. 65:48-49.
7. de Jonge R, Bolton MD, Kombrink A, van den Berg GC, Yadeta KA, et al. 2013. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Res. 23:1271-1282.
8. DiGuistini S, Wang Y, Liao NY, Taylor G, Tanguay P, et al. 2011. Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. PNAS. 108:2504-2509.

9. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, et al. 2012. Emerging fungal threats to animal, plant and ecosystem health. Nature. 484:186-194.

10. Fitzpatrick DA. 2011. Horizontal gene transfer in fungi. FEMS Microbiol Lett. 329:1-8.

11. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, et al. 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet. 38: 953-956.

12. Goheen D and Hansen E. 1993. Effects of pathogens and bark beetles on forests in Beetle-Pathogen Interactions in Conifer Forests. Academic Press, London, UK. 175-196.

13. Gurr S, Samalova M, Fisher M. 2011. The rise and rise of emerging infectious fungi challenges food security and ecosystem health. Fungal Biol Rev. 25:181-188.

14. Hadziabdic D, Windham M, Baird R, Vito L, Cheng Q, et al. 2014. First report of *Geosmithia morbida* in North Carolina: the pathogen involved in thousand cankers disease of black walnut. Plant Dis. 98:992.

15. Harrington TC. 2005. Ecology and Evolution of Mycophagous Bark Beetles and Their Fungal Partners. Oxford University Press, Oxford, UK. 257-291.

16. Hatta R, Ito K, Hosaki Y, Tanaka T, Tanaka A, et al. 2002. A conditionally dispensable chromosome controls host-specific pathogenicity in the fungal plant pathogen *Alternaria alternata*. Genetics. 161:59-70.

17. Hishinuma SM, Dallara PL, Yaghmour MA, Zerillo MM, Parker CM, et al. 2016. Wingnut (Juglandaceae) as a new generic host for *Pityophthorus juglandis* (Coleoptera: Curculionidae) and the thousand cankers disease pathogen, *Geosmithia morbida* (Ascomycota: Hypocreales). Can Entomol. 148:83-91.

18. Juzwik J, McDermott-Kubeczko M, Stewart TJ, Ginzel MD. 2016. First Report of *Geosmithia morbida* on Ambrosia Beetles emerged from Thousand Cankers Disease *Juglans* in Ohio. Plant Dis. 100:1238.

19. Kolarik M, Freeland E, Utley C, Tisserat N. 2011. *Geosmithia morbida* sp. nov., a new phytopathogenic species living in symbiosis with the walnut twig beetle (*Pityophthorus juglandis*) on *Juglans* in USA. Mycologia. 103:325–332.

20. Koonin EV, Makarova KS, Aravind L. 2001. Horizontal Gene Transfer in Prokaryotes: Quantification and Classification. Annu Rev Microbiol. 55:709-742.

21. Ma LJ, Van Der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 464:367-373.

22. Meddens AJH, Hicke JA, Ferguson CA. 2012. Spatiotemporal patterns of observed bark beetle-caused tree mortality in British Columbia and the western United States. Ecol Appl. 22:1876–1891.

23. Mehrabi R, Bahkali AH, Abd-Elsalam KA, Moslem M, M'Barek SB, et al. 2011. Horizontal gene and chromosome transfer in plant pathogenic fungi affecting host range. FEMS Microbiol Rev. 35:542-554.

24. Montecchio L, Faccoli M. 2014. First record of thousand cankers disease *Geosmithia morbida* and walnut twig beetle *Pityophthorus juglandis* on *Juglans nigra* in Europe. Plant Dis. 98:696.

25. Oerke EC. 2006. Crop losses to pests. J Agric Sci. 144:31-43.

26. Otrosina WJ, Hess NJ, Zarnoch SJ, Perry TJ, Jones JP. 1997. Blue-stain fungi associated with roots of southern pine trees attacked by the southern pine beetle, *Dendroctonus frontalis*. Plant Dis. 81:942-945.

27. Owen DR, Lindahl KQ, Wood DL, Parmeter JR. 1987. Pathogenicity of fungi isolated from *Dendroctonus valens*, *D. brevicomis*, and *D. ponderosae* to ponderosa pine seedlings. Phytopathology. 77:631-636.

28. Paine TD, Raffa KF, Harrington TC. 1997. Interactions among Scolytid bark beetles, their associated fungi, and live host conifers. Annu Rev Entomol. 42:179-206.

29. Pennisi E. 2010. Armed and Dangerous. Science. 327:804-805.

30. Plattner A, Kim J, DiGuistini S, Breuil C. 2008. Variation in pathogenicity of a mountain pine beetle–associated blue-stain fungus, *Grosmannia clavigera*, on young lodgepole pine in British Columbia. Can J Plant Pathol. 30:457-466.

31. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol. 10:417-430.

32. Richards TA, Soanes DM, Jones MDM, Vasieva O, Leonard G, et al. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. PNAS. 108:15258-15263.

33. Rugman-Jones PF, Seybold SJ, Graves AD, Stouthamer R. 2015. Phylogeography of the Walnut Twig Beetle, *Pityophthorus juglandis*, the vector of thousand cankers disease in North American walnut trees. PLoS ONE. 10:e0118264.

34. Safranyik L, Carroll AL, Regniere J, Langor DW, Reil WG, et al. 2010. Potential for range expansion of mountain pine beetle into the boreal forest of North America. Can Entomol. 142:415–441.

35. Serdani M, Vlach JJ, Wallis KL, Zerillo M, McCleary T, et al. 2013. First Report of *Geosmithia morbida* and *Pityophthorus juglandis* Causing Thousand Cankers Disease in Butternut. Plant Health Progress. doi:10.1094/PHP-2013-1018-01-BR.

36. Six DL, Paine TD. 1997. *Ophiostoma clavigerum* is the mycangial fungus of the Jeffrey pine beetle, *Dendroctonus jeffreyi*. Mycologia. 89:858-866.

37. Stukenbrock EH, Croll D. 2014. The evolving fungal genome. Fungal Biol Rev. 28: 1–12.

38. Temporini ED, VanEtten HD. 2004. An analysis of the phylogenetic distribution of the pea pathogenicity genes of *Nectria haematococca* MPVI supports the hypothesis of their origin by horizontal transfer and uncovers a potentially new pathogen of garden pea: *Neocosmospora boniensis*. Curr Genet. 46: 29-36.

39. Thousand Cankers Disease. Available at http://www.thousandcankers.com/tcd-locations.php. Accessed 6 May 2016.

40. Tisserat N, Cranshaw W, Leatherman W, Utley C, Alexander K. 2009. Black walnut mortality in Colorado caused by the walnut twig beetle and thousand cankers disease. Plant Health Progress. doi:10.1094/PHP-2009-0811-01-RS.

41. Tsui CK, Roe AD, El-Kassaby YA, Rice AV, Alamouti SM, et al. 2012. Population structure and migration pattern of a conifer pathogen, *Grosmannia clavigera*, as influenced by its symbiont, the mountain pine beetle. Mol Ecol. 21:71-86.

42. USDA-APHIS. 2009. Pathway Assessment: *Geosmithia* sp. and *Pityophthorus juglandis* Blackman movement from the western into the eastern United States. Available at http://www.thousandcankers.com/media/docs/APHIS_Geosmithia_10_2009.pdf.

43. Utley C, Nguyen T, Roubtsova T, Coggeshall M, Ford TM, et al. 2013. Susceptibility of walnut and hickory species to *Geosmithia morbida*. Plant Dis. 97:601–607.

44. Wang Y, Lim L, DiGuistini S, Robertson G, Bohlmann J, et al. 2012. A specialized ABC efflux transporter GcABC-G1 confers monoterpene resistance to *Grosmannia clavigera*, a bark beetle-associated fungal pathogen of pine trees. New Phytol. 197:886-898.

45. Webster J, Weber R. 2007. Introduction to Fungi. Cambridge University Press, Cambridge, UK. 227.

46. Xie J, Fu Y, Jiang D, Li G, Huang J, et al. 2008. Intergeneric transfer of ribosomal genes between two fungi. BMC Evol. Biol. 8:1-7.

47. Yaghmour MA, Nguyen TL, Roubtsova TV, Hasey JK, Fichtner EJ, et al. 2014. First Report of *Geosmithia morbida* on English walnut and its paradox rootstock in California. Plant Dis. 98: 1441.

48. Zerillo MM, Caballero JI, Woeste K, Graves AD, Hartel C, et al. 2014. Population Structure of *Geosmithia morbida*, the Causal Agent of Thousand Cankers Disease of Walnut Trees in the United States. PLoS ONE. 9:e112847.

Chapter 2
De novo genome assembly of *Geosmithia morbida*, the causal agent of thousand cankers disease

2.1 Introduction

Studying molecular evolution of any phenotype is now made possible by the analysis of large amounts of sequence data generated by next-generation sequencing platforms. This is particularly beneficial for the study of emerging fungal pathogens, which are progressively recognized as a threat to global biodiversity and food security. Furthermore, in many cases their expansion is a result of anthropogenic activities and an increase in trade of fungal-infected goods (Fisher et al. 2012). Fungal pathogens are capable of evolving rapidly in order to overcome host resistance, fungicides, and to adapt to new hosts and environments. Whole genome sequence data are useful in identifying the mechanisms of adaptive evolution within fungi (Stukenbrock et al. 2011, Gardiner et al. 2012, Condon et al. 2013). For instance, Stukenbrock et al. (2011) investigated the patterns of evolution in fungal pathogens during the process of domestication in wheat using all aligned genes within the genomes of wheat pathogens. They found that *Zymoseptoria tritici*, a domesticated wheat pathogen (formerly known as *Mycosphaerella graminicola*), underwent adaptive evolution at a higher rate than its wild relatives, *Z. pseudotritici* and *Z. ardabiliae* (Stukenbrock et al. 2012). The study also revealed that many of the pathogen’s 802 secreted proteins were under positive selection. A study by Gardiner et al. (2012), identified genes encoding aminotransferases, hydrolases, and kinases that were shared between *Fusarium pseudograminearum* and other cereal pathogens. Using phylogenomic analyses, the researchers demonstrated that these genes had bacterial origins. These studies highlight the various evolutionary means that fungal species employ in order to adapt to specific hosts, as well as the importance of genomics and bioinformatics in elucidating evolutionary mechanisms within the fungal kingdom.

Many tree fungal pathogens associate with bark beetles in the family Scolytinae (Six & Wingfield, 2011). With climate change, beetles and their fungal symbionts can invade new territory and become major invasive forest pests on a global scale (Kurz et al. 2008, Sambaraju et al. 2012). A well-known example of an invasive pest is the mountain pine beetle and its symbiont, *Grosmannia clavigera* that has affected approximately 3.4 million of acres of lodgepole, ponderosa, and five-needle pine trees in Colorado alone since the outbreak began in 1996 (Massoumi Alamouti et al. 2014, Colorado State Forest Service 2015). Another beetle pest in the western US, *Pityophthorus juglandis* (walnut twig beetle), associates with several fungal species, including the emergent fungal pathogen *Geosmithia morbida* (Tisserat et al. 2009, Kolarik et al. 2011).

Reports of tree mortality triggered by *G. morbida* infections first surfaced in 2009 (Kolarik et al. 2011), while the fungus was described as a new species in 2011 (Tisserat et al. 2009). This fungus is vectored into the host via *P. juglandis* and is the causal agent of thousand cankers disease (TCD) in *Julgans nigra* (eastern black walnut) (Zerillo et al. 2014). This walnut species is valued for its wood, which is used for furniture, cabinetry, and veneer. Although *J. nigra* trees are planted throughout western US as a decorative species, they are indigenous to eastern North America where the walnut industry is worth hundreds of millions of dollars (Rugman-Jones et al. 2015, Zerillo et al. 2014). In addition to being a major threat to the eastern populations of *J. nigra*, TCD is of great concern because certain western walnut species including *J. regia* (the Persian walnut), *J. californica*, and *J. hindsii* are also susceptible to the fungus according to greenhouse inoculation studies (Utley et al. 2013).

The etiology of TCD is complex because it is a consequence of a fungal-beetle symbiosis. The walnut twig beetle, which is only known to attack members of genera *Juglans* and *Pterocarya*, is the most common vector of *G. morbida* (Kolarik et al. 2011). Nevertheless, other beetles are able to disperse the fungus from infested trees (Kolarik et al. 2007, Kolarik & Jankowiak, 2013). As vast numbers of beetles concentrate in the bark of infested trees, fungal cankers form and coalesce around beetle galleries and entrance holes. As the infection progresses, the phloem and cambium discolor and the leaves wilt and yellow. These symptoms are followed by branch dieback and eventual tree death, which can occur within three years of the initial infection (Kolarik et al. 2011). Currently, 15 states in the US have reported one or more incidences of TCD, reflecting the expansion of WTB's geographic range from its presumed native range in a few southwestern states (Rugman-Jones et al. 2015). Additionally, TCD has also been found in Europe where walnut species are planted for timber (Montecchio & Faccoli 2014).

To date, *G. morbida* is one of only two known pathogens within the genus *Geosmithia*, which consists of mostly saprotrophic beetle-associated species (the other pathogen is *G. pallida*) (Lynch et al. 2014). The ecological complexity this vector-host-pathogen system exhibits makes it an intriguing lens for studying the evolution of pathogenicity. A well-assembled reference genome will enable us to identify genes unique to *G. morbida* that may be utilized to develop sequence-based tools for detecting and monitoring epidemics of TCD and for exploring the genomic features of *Geosmithia* species, which may help explain the evolution of pathogenicity. Here, we present a *de novo* genome assembly of *Geosmithia morbida*. The objectives of this study are to: 1) assemble the first, high-quality draft genome of this pathogen; 2) annotate the genome to better understand the genomic composition of *Geosmithia* species; and 3) briefly compare the genome of *G. morbida* to two other fungal pathogens for which genomic data are available: *Fusarium solani*, a root pathogen that infects soybean, and *Grosmannia calvigera*, a pathogenic ascomycete that associates with the mountain pine beetle and kills lodgepole pines in North America.

2.2 Methods

2.2.1 DNA extraction and library preparation

DNA was extracted using the CTAB method as outlined by the Joint Genome Institute for Genome Sequencing from lyophilized mycelium of *G. morbida* (isolate 1262, host: *Juglans californica*) from southwestern California (Kohler & Francis 2015). The total DNA concentration was measured using Nanodrop, and samples for sequencing were sent to Purdue University Genomics Core Facility in West Lafayette, Indiana. DNA libraries were prepared using the paired-end Illumina Truseq protocol and mate-pair Nextera DNA Sample Preparation kits with average insert sizes of 487 and 1921 bp, respectively. These libraries were sequenced on the Illumina HiSeq 2500 using a single lane with a maximum read length of 101 bp.

2.2.2 Preprocessing sequence data

To assess the quality of our data, we ran FastQC (v0.11.2) (https://goo.gl/xHM1zf) (Andrews 2015) and SGA Preqc (v0.10.13) (https://goo.gl/9y5bNy) on our raw sequence reads (Simpson 2013). Both tools aim to supply the user with information such as per base sequence quality score distribution (FastQC) and frequency of variant branches in de Bruijn graphs (Preqc) that aid in selecting appropriate assembly tools and parameters. The paired-end raw reads were corrected using a Bloom filter-based error correction tool called BLESS (v0.16) (https://goo.gl/Kno6Xo) (Heo et al. 2014). Next, the error corrected reads were trimmed with Trimmomatic, version 0.32, using a Phred threshold of 2, following recommendations from MacManes (2014) (https://goo.gl/ FFoFjL) (Bolger et al. 2014). NextClip, version 1.3.1, was leveraged to trim adapters in the mate-pair read set (https://goo.gl/aZ9ucT) (Leggett et al. 2014).

2.2.3 *De novo* genome assembly and evaluation

The de novo genome assembly was constructed with ALLPaths-LG (v49414) (https://goo.gl/03gU9Z) (Gnerre et al. 2011). The assembly was evaluated with BUSCO (v1.1b1) (https://goo.gl/bMrXIM), a tool that assesses genome completeness based on the presence of single-copy orthologs (Simao et al. 2015). We also generated length-based statistics for our de novo genome with QUAST (v2.3) (https://goo.gl/ 5KSa4M) (Gurevich et al.

2013). The raw reads were mapped back to the genome using BWA version 0.7.9a-r786 to further assess the quality of the assembly (https://goo.gl/ Scxgn4) (Li & Durbin 2009).

2.2.4 Structural and functional annotation of *G. morbida* genome

We used the automated genome annotation software Maker version 2.31.8 (Cantarel et al. 2008). Maker identifies repetitive elements, aligns ESTs, and uses protein homology evidence to generate ab initio gene predictions (https://goo.gl/JiLA3H). We used two of the three gene prediction tools available within the pipeline, SNAP and Augustus. SNAP was trained using gff files generated by CEGMA v2.5 (a program similar to BUSCO) (Parra et al. 2007). Augustus was trained with *Fusarium solani* protein models (v2.0.26) downloaded from Ensembl Fungi (EnsemblFungi 2015). In order to functionally annotate the genome, the protein sequences produced by the structural annotation were blasted against the Swiss-Prot database, and target sequences were filtered for the best hits (Swiss-Prot 2015). A small subset of the resulting annotations was visualized and manually curated in WebApollo v2.0.1 (Lee et al. 2013). The final annotations were also evaluated with BUSCO (v1.1b1) (https://goo.gl/thTGzH).

2.2.5 Assessing repetitive elements profile

To assess the repetitive elements profile of *G. morbida*, we masked only the interspersed repeats within the assembled scaffolds with RepeatMasker (v4.0.5) (https://goo.gl/ TXrbr3) (Smit et al. 1996) using the sensitive mode and default values as arguments. In order to compare the repetitive element profile of *G. morbida* with *F. solani* (v2.0.29) and *G. clavigera* (kw1407.GCA_000143105.2.30), the interspersed repeats of these two fungal pathogens were also masked with RepeatMasker. The genome and protein data of these fungi were downloaded from Ensembl Fungi (EnsemblFungi 2015).

2.2.6 Identifying putative proteins contributing to pathogenicity

To identify putative genes contributing to pathogenicity in *G. morbida*, a BLASTp search was conducted for single best hits at an e-value threshold of 1e-6 or less against the PHI-base database (v3.8) (https://goo.gl/CEEVY0) that contains experimentally confirmed genes from fungal, oomycete and bacterial pathogens (PHI-base 2015). The search was performed using the same parameters for *F. solani* and *G. clavigera*. To identify the proteins that contain signal peptides, we used SignalP (v4.1) (https:// goo.gl/JOe5Dh), and compared results from *G. morbida* with those from *F. solani* and *G. clavigera* (Peterson et al. 2011). Lastly, to find putative protein domains involved in pathogenicity in *G. morbida*, we performed a HMMER (version 3.1b2) (Finn et al. 2011) search against the Pfam database (v28.0) (Finn et al. 2014) using the protein sequences as query. We conducted the same search for sequences of 17 known effector proteins, then extracted and analyzed domains common between the effector sequences and *G. morbida* (https://goo.gl/Y9IPZs).

2.3 Results and discussion

2.3.1 Data processing

A total of 28,027,726 paired-end (PE) and 41,348,578 mate-pair (MP) reads were generated with approximately 109x and 160x coverage, respectively (Table 2.1). Of the MP reads, 67.7% contained adapters that were trimmed using NextClip (v1.3.1). We corrected errors within the PE reads using BLESS (v0.16) at a kmer length of 21. After correction, low-quality reads (phred score < 2) were trimmed with Trimmomatic (v0.32) resulting in 99.75% reads passing. In total, 16,336,158 MP and 27,957,268 PE reads were used to construct the *de novo* genome assembly.

2.3.2 Assembly features

The *G. morbida* de novo assembly was constructed with AllPaths-LG (v49414). The assembled genome consisted of 73 contigs totaling 26,549,069 bp, which is comparable to certain other Ascomycetes such as *Acremonium chrysogenum* and *Ustilaginoidea virens* with genome sizes of 28.6 and 30.2 Mbp, respectively. The largest contig length was 2,597,956 bp, and the NG50 was 1,305,468 bp. The completeness of the genome assembly was assessed using BUSCO, a tool that scans the genome for the presence of single-copy orthologous groups present in more than 90% of fungal species. Of 1,438 single-copy orthologs specific to fungi, 98% were complete in our assembly, and 4.3% were duplicated BUSCOs. Only one ortholog was missing from the genome (Table 2.2). We used BWA to map the unprocessed, raw MP and PE reads back to the genome to further evaluate the assembly, and 87% of the MP and 90% of the PE reads mapped to our reference genome.

2.3.3 Gene annotation

The automated genome annotation software Maker v2.31.8 was used to identify structural elements in the *G. morbida* assembly generated by AllPaths-LG. Of the total 6,273 proteins that were predicted, 5,996 had protein-homology evidence in the Swiss-Prot database and only 277 (4.41%) of the total genes encoded for proteins of unknown function. Even though the total of 6,273 proteins is lower than the average number of 11,129 genes in Ascomycota, this number is within the range of the 4,657 and 27,529 coding genes within the phylum (Mohanta & Bae 2015). The completeness of the functional annotations was evaluated using BUSCO, and 95% of the single copy orthologs were present in this protein set and only 7% were duplicated BUSCOs.

2.3.4 Repetitive elements

Repetitive elements represented 0.81% of the total bases in *G. morbida*. The genome contained 152 retroelements (class I) that were mostly composed of long terminal repeats (n = 146) and 60 DNA transposons (class II). In comparison, the genomes of *G. clavigera* and *F. solani* contained 1.14 and 1.47%, respectively. *G. clavigera* possesses 541 retroelements (0.79%) and 66 DNA transposons (0.04%), whereas the genome of *F. solani* is comprised of 499 (0.54%) and 515 (0.81%) retroelements and transposons, respectively. The larger number of repeat elements in *F. solani* may explain its relatively large genome size—51.3 Mbp versus *G. clavigera's* 29.8 Mbp and *G. morbida's* 26.5 Mbp (Table 2.3).

2.3.5 Identifying putative pathogenicity genes

We blasted the entire predicted protein set against the PHI-base database (v3.8) to identify a list of putative genes that may contribute to pathogenicity within *G. morbida*, *F. solani*, and *G. clavigera*. We determined that 1,974 genes in *G. morbida* (31.47% of the total 6,273 genes) were homologous to protein sequences in the database. For *F. solani* and *G. clavigera*, there were 4,855 and 2,387 genes with homologous PHI-base proteins.

2.3.6 Identifying putative secreted proteins

A search for the presence of putative secreted peptides within the protein sequences of *G. morbida*, *F. solani* and *G. clavigera* showed that approximately 5.6% (349) of the *G. morbida* sequences contained signal peptides. Of the 349 sequences containing putative signal peptides, only 27 encoded proteins of unknown function. Roughly 8.8 and 6.9% of the proteins of *F. solani* and *G. clavigera* possess signal peptides. Secreted proteins are essential for host-fungal interactions and are indicative of adaptation within fungal pathogens that require an array of mechanisms to overcome plant host defenses. Even though the precise means by which fungal proteins are trafficked into the host are unclear, secreted proteins are known to be essential for the translocation of fungal proteins into the host cells (Petre & Kamoun 2014). For instance, race 1 strains of *Verticillium dahliae*, a common cause of vascular wilt disease in plants, secretes a protein called Ave1 that induces host immunity response suggesting this protein is crucial for virulence (de Jonge et al. 2012). Another example of a secreted protein is Ecp6 in fungal pathogen *Cladosporium fulvum* that prevents chitin-activated detection by the host plant (de Jonge et al. 2010).

2.3.7 Identifying protein domains

We conducted a HMMER search against the pfam database (v28.0) using amino acid sequences for *G. morbida* and 17 effector proteins from various fungal species. For *G. morbida*, there were 6,023 unique protein domains out of a total of 43,823 Pfam hits. A total of 17 domains, which comprised 1,000 hits, were shared between *G. morbida* and known effector proteins. The three most common protein domains in *G. morbida* with a putative effector function belonged to short-chain dehydrogenases (n = 111), polyketide synthases (n = 94) and

NADH dehydrogenases (n = 86).

2.4 Conclusion

This work introduces the first genome assembly and analysis of *Geosmithia morbida*, a fungal pathogen of the black walnut tree that is vectored into the host via the walnut twig beetle. The *de novo* assembly is composed of 73 scaffolds totaling in 26.5 Mbp. There are 6,273 predicted proteins, and 4.41% of these are unknown. In comparison, 68.27% of *F. solani* and 26.70% of *G. clavigera* predicted proteins are unknown. We assessed the quality of our genome assembly and the predicted protein set using BUSCO, and found that 98 and 95% of the single copy orthologs specific to the fungal lineage were present in both, respectively. These data are indicative of our assembly's high quality and completeness. Our BLASTp search against the PHI-base database revealed that *G. morbida* possesses 1,974 genes that are homologous to proteins involved in pathogenicity. Furthermore, *G. morbida* shares several domains with known effector proteins that are key for fungal pathogens during the infection process.

Geosmithia morbida is one of only two known fungal pathogens within the *Geosmithia* genus (Lynch et al. 2014). The genome assembly introduced in this study can be leveraged to explore the molecular mechanisms behind pathogenesis within this genus. The putative list of pathogenicity genes provided in this study can be used for future comparative genomic analyses, knock-out, and inoculation experiments. Moreover, genes unique to *G. morbida* may be utilized to develop DNA sequence-based tools for detecting and monitoring ongoing and future TCD epidemics.

2.5 Tables and figures

Table 2.1. Statistics for *Geosmithia morbida* sequence data.

	Paired-end		Mate-pair	
Number of reads	28,027,726	**27,957,268**	41,348,578	**16,336,158**
Average insert size (bp)	487		1921	
Average coverage	109x		160x	

The values in bold are number of trimmed, error corrected and filtered reads that were used for the assembly.

Table 2.2. *Geosmithia morbida* reference genome assembly statistics generated using QUAST (v2.3)

Number of sequences	73
Largest scaffold length	2,597,956
N50	1,305,468
L50	7
Total assembly length	26,549,069
GC%	54.31
BUSCOs completeness	95%

Table 2.3. Repetitive elements profile for *Geosmithia morbida*, *Grosmannia clavigera* and *Fusarium solani*.

	G. morbida	*G. clavigera*	*F. solani*
Genome size	26.5 Mbp	29.8 Mbp	51.3 Mbp
% Repetitive element	0.81%	1.14%	1.47%
% Retroelements	0.10%	0.79%	0.54%
% DNA transposons	0.02%	0.04%	0.81%

RepeatMasker (v4.0.5) was used to generate the above values. Genomic data for *F. solani* and *G. clavigera* were downloaded from Ensembl Fungi.

2.6 References

1. Andrews S. 2015. FastQC. Cambridge: Babaraham Institute. Available at http://www. bioinformatics.babraham.ac.uk/projects/fastqc/. Accessed 12 Dec. 2015.
2. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114–2120.
3. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, et al. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188–196.
4. Colorado State Forest Service. 2015. Mountain Pine Beetle. Fort Collins: Colorado State Univeristy. Available at http://csfs.colostate.edu/forest-management/common-forest-insects- diseases/mountain-pine-beetle/. Accessed 15 April 2015.
5. Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, et al. 2013. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* Pathogens. PLoS Genet. 9:e1003233.
6. de Jonge R, van Esse HP, Kombrink A, Shinya T, Desaki Y, et al. 2010. Conserved fungal lysm effector Ecp6 prevents chitin-triggered immunity in plants. Science 329:953–955.
7. de Jonge R, van Esse HP, Maruthachalam K, Bolton MD, Santhanam P, et al. 2012. Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome RNA sequencing. PNAS. 109: 5110–5115.
8. EnsemblFungi. 2015. Available at http://fungi.ensembl.org/index.html. Accessed 14 Nov. 2015.
9. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.
10. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39:W29–W37.

11. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, et al. 2012. Emerging fungal threats to animal, plant and ecosystem health. Nature. 484:186–194.

12. Gardiner DM, McDonald MC, Covarelli L, Solomon PS, Rusu AG, et al. 2012. Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. PLoS Pathog. 8:e1002952.

13. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, et al. 2011. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. PNAS. 108:1513–1518.

14. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075.

15. Heo Y, Wu X-L, Chen D, Ma J, Hwu W-M. 2014. BLESS: bloom filter-based error correction solution for high-throughput sequencing reads. Bioinformatics. 30:1354–1362.

16. Kohler A, Francis M. 2015. Genomic DNA Extraction. Available at http://1000.fungalgenomes.org/ home/wp-content/uploads/2013/02/genomicDNAProtocol-AK0511.pdf. Accessed 12 Dec. 2015.

17. Kolarik M, Freeland E, Utley C, Tisserat N. 2011. *Geosmithia morbida* sp. nov., a new phytopathogenic species living in symbiosis with the walnut twig beetle (*Pityophthorus juglandis*) on *Juglans* in USA. Mycologia. 103:325–332.

18. Kolarik M, Jankowiak R. 2013. Vector affinity and diversity of *Geosmithia* fungi living on subcortical insects inhabiting Pinaceae species in Central and Northeastern Europe. Micro Ecol. 66:682–700.

19. Kolarik M, Kostovcik M, Pazoutova S. 2007. Host range and diversity of the genus *Geosmithia* (Ascomycota: Hypocreales) living in association with bark beetles in the Mediterranean area. Mycological Res. 111:1298–1310.

20. Kurz WA, Dymond CC, Stinson G, Rampley GJ, Neilson ET, et al. 2008. Mountain pine beetle and forest carbon feedback to climate change. Nature. 452:987–990.

21. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, et al. 2013. Web Apollo: a web-based genomic annotation editing platform. Genome Biol.14:R93.

22. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. 2014. NextClip: an analysis and read preparation tool for Nextera long mate pair libraries. Bioinformatics. 30:566–568.

23. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.

24. Lynch SC, Wang DH, Mayorquin JS, Rugman-Jones PF, Stouthamer R, Eskalen E. 2014. First Report of *Geosmithia pallida* Causing Foamy Bark Canker, a new disease on coast live oak (*Quercus agrifolia*), in association with *Pseudopityophthorus pubipennis* in California. Plant Dis. 98:1276.

25. MacManes MD. 2014. On the optimal trimming of high-throughput mRNA sequence data. Available at http://dx.doi.org/10.1101/000422.

26. Massoumi Alamouti S, Haridas S, Feau N, Robertson G, Bohlmann J, Breuil C. 2014. Comparative genomics of the pine pathogens and beetle symbionts in the genus *Grosmannia*. Mol Biol Evol. 31:1454–1474.

27. Mohanta TK, Bae H. 2015. The diversity of fungal genome. Biol Proced Online. 17:1–9.

28. Montecchio L, Faccoli M. 2014. First record of thousand cankers disease *Geosmithia morbida* and walnut twig beetle *Pityophthorus juglandis* on *Juglans nigra* in Europe. Plant Dis. 98:696.

29. Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23:1061–1067.

30. Peterson TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.

31. Petre B, Kamoun S. 2014. How do filamentous pathogens deliver effector proteins into plant cells? PLoS Biol. 12:1–7.

32. PHI-base. 2015. The pathogen–host interaction database. Available at http://www.phi-base.org/. Accessed 22 Nov. 2015.

33. Rugman-Jones PF, Seybold SJ, Graves AD, Stouthamer R. 2015. Phylogeography of the walnut twig beetle, *Pityophthorus juglandis*, the vector of thousand cankers disease in North American walnut trees. PLoS ONE. 10:e118264.

34. Sambaraju KR, Carroll AL, Zhu J, Stahl K, Moore RD, Aukema BH. 2012. Climate change could alter the distribution of mountain pine beetle outbreaks in western Canada. Ecography. 35:211–223.

35. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 31:3210–3212.

36. Simpson JT. 2013. Exploring genome characteristics and sequence quality without a reference. Available at http://arxiv.org/abs/1307.8026.

37. Six DL, Wingfield MJ. 2011. The role of phytopathogenicity in bark beetle-fungus symbioses: a challenge to the classic paradigm. Ann Rev Entomol. 56:255–272.

38. Smit AFA, Hubley R, Green P. 1996. RepeatMasker. Available at http://www.repeatmasker.org.

39. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, 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.

40. Stukenbrock EH, Quaedvlieg W, Javan-Nikhah M, Zala M, Crous PW, McDonald BA. 2012. *Zymoseptoria ardabiliae* and *Z. pseudotritici*, two progenitor species of the septoria tritici leaf blotch fungus *Z. tritici*. Mycologia. 104:1397–1407.

41. Swiss-Prot. 2015. Available at http://www.uniprot.org/. Downloaded 6 May 2015.

42. Tisserat N, Cranshaw W, Leatherman D, Utley C, Alexander K. 2009. Black walnut

mortality in colorado caused by the walnut twig beetle and thousand cankers disease. Plant Health Progress. 1–10. DOI 10.1094/PHP-2009-0811-01-RS.

43. Utley C, Nguyen T, Roubtsova T, Coggeshall M, Ford TM, et al. 2013. Susceptibility of walnut and hickory species to *Geosmithia morbida*. Plant Dis. 97:601–607.

44. Zerillo MM, Caballero JI, Woeste K, Graves AD, Hartel C, et al. 2014. Population structure of *Geosmithia morbida*, the causal agent of thousand cankers disease of walnut trees in the United States. PLoS ONE. 9:e112847.

Chapter 3
Understanding the Evolution of Pathogenicity within *Geosmithia*

3.1 Introduction

Geosmithia (Ascomycota: Hypocreales) is a newly described genus that largely contains saprotrophic beetle-associated fungal species (Kolarik et al. 2005, Kolarik et al. 2011). The genus was first proposed in 1979 for fungi that were formerly placed in genus *Penicillium* (Pitt 1979). *Geosmithia* species are filamentous fungi that commonly associate with phloeophagous bark beetles. However, some *Geosmithia* fungi, such as *G. eupagioceri* and *G. microcorthyli*, are known to affiliate with ambrosia beetles (Kolarik & Jankowiak 2013). *Geosmithia* species and their beetle associates occupy a variety of hosts including pines, oaks, junipers, angiosperms and walnut trees (Kolarik & Kirkendall 2010, Kolarik & Jankowiak 2013, Kolarik et al. 2007). An understanding of the ecology and diversity of symbiotic relationships between these fungi and their beetle associates is limited, but has recently started to be explored (Kolarik & Jankowiak 2013, Kolarik et al. 2007). While most species in *Geosmithia* are saprotrophic, two species are known to be pathogenic—*G. pallida* (Lynch et al. 2014) and *G. morbida* (Tisserat et al. 2009), the latter of which is the focal species of this study.

Geosmithia morbida causes thousand cankers disease (TCD) in *Juglans nigra* (eastern black walnut) and is vectored into the host by *Pityophthorus juglandis*, commonly known as the walnut twig beetle (WTB) (Kolarik et al. 2011). The earliest mortality incidences of black walnut trees were noted in Colorado, US in 2001. Since then, nine western states (CO, WA, OR, ID, NV, UT, CA, NM, AZ) and seven eastern states (PA, OH, IN, MD, VA, TN, NC) have reported one or more incidences of TCD (Zerillo et al. 2014). This increase in TCD is likely a consequence of the expansion of WTB's geographic range, which was present in only four counties of California, Arizona and New Mexico in the 1960s. However, as of 2014, the beetle has been detected in over 115 counties in the western and eastern US (Rugman-Jones et al. 2014).

The origin of this pathogen is not clear, however, it has been hypothesized that *G. morbida* may have undergone a host shift from *J. major* (Arizona black walnut) to a more naïve host, *J. nigra*, because the fungus does not cause disease in the Arizona black walnut, and neither WTB nor *G. morbida* were observed in the native range of *J. nigra* until 2010. It is also essential to note *J. nigra* is not indigenous to western US and has been planted through the region as an ornamental species. An alternative prediction based on *G. morbida* population genetic data suggests that the origin of *G. morbida* and WTB are the walnut populations of southern California, where the pathogen has been isolated from both healthy and diseased *J. californica* trees (Zerillo et al. 2014).

Some early symptoms of infection by *G. morbida* include yellowing, wilting and thinning of the foliage followed by branch dieback and tree death within 2-3 years after the initial infestation (Kolarik et al. 2011, Tisserat et al. 2009). Little is known about the specific means *G. morbida* employs for initiating and maintaining the infection, or what benefits, if any, the fungus imparts to the WTB vector. However, previous studies have demonstrated that fungal pathogens that occupy similar ecological niches as *G. morbida* must be capable of enduring and combating toxic host environments used by plants to resist infection. For instance, *Grosmannia clavigera* is a fungal symbiont of the mountain pine beetle, and this fungus can detoxify metabolites such as terpenoids and phenolics produced by the host as defense mechanisms (DiGuistini et al. 2011).

In a recent study, we developed a reference genome of *Geosmithia morbida*, which consisted of 73 scaffolds totaling 26.5 Mbp in length (Chapter 2, Schuelke et al. 2016). The fungus possesses 6,273 predicted proteins; over 30% of these peptides are homologous to proteins implicated in pathogenicity. In this work, we compare the reference genome of pathogenic and host specific species *G. morbida* with two closely related non-pathogenic and generalist species, *G. flava* and *G. putterillii*. To do so, we identify putative genes under positive selection that may be involved in the specialization of a pathogenic life strategy dependent on a single beetle vector and narrow, but potentially expanding, host range. We also present a species phylogeny, estimated using single-copy orthologs, which confirms the placement of *Geosmithia* species in the order Hypocreales and that their closest fungal relative is *Acremonium chrysogenum*. The primary goal of this study was to gain insight into the evolution of pathogenicity within *G. morbida*. To do so, we identified potential genes that are experiencing adaptive selection in *G. morbida* and *G. clavigera* because these two tree pathogens are vectored into the host via beetle associates. We predicted these fungal pathogens to have common effector proteins that are under positive selection.

3.2 Methods

3.2.1 DNA extraction and sequencing

The CTAB method delineated by the Joint Genome Institute was used to extract DNA for genome sequencing from lyophilized mycelium of *Geosmithia flava* and *Geosmithia putterillii* (Kohler et al. 2011). Total DNA concentration was measured with Nanodrop, and DNA sequencing was conducted at Purdue University Genomics Core Facility in West Lafayette, Indiana. DNA libraries were prepared using the paired-end Illumina Truseq protocol and sequenced on an Illumina HiSeq 2500 using a single lane. Mean insert sizes for *G. flava* and *G. putterillii* were 477bp and 513bp, correspondingly. Table 3.1 lists genetic, geographic, and host information for each *Geosmithia* species used in this study.

3.2.2 Preprocessing sequence data

The raw paired-end reads for *G. flava* and *G. putterillii* were corrected using BFC (version r181) (Li 2015). BFC utilizes a combination of hash table and bloom-filter to count *k*-mers for a given read and correct errors in that read based on the *k*-mer support. Because BFC requires interleaved reads as input, khmer was leveraged to interleave as well as split the paired-end reads before and after the error correction stage, respectively (Crusoe et al. 2015). Next, low quality bases and adapters in error corrected reads were trimmed with Trimmomatic, version 0.32, using a Phred threshold of 4 (Bolger et al. 2014).

3.2.3 Assembly construction

Genome assemblies were constructed with ABySS 1.9.0 using four k-mer sizes of 61, 71, 81, and 91 (Simpson et al. 2009). The resulting assemblies were evaluated using BUSCO (v1.1b1) (Simão et al. 2015), which assess completeness based on the presence of universal single-copy orthologs within fungi. Length-based statistics were generated with QUAST v2.3 (Gurevich et al. 2013). Final assemblies were manually chosen based on length-based and genome completeness statistics. Furthermore, the raw reads of *G. flava* and *G. putterillii* were mapped back to their corresponding genomes using BWA version 0.7.9a-r786 to assess the quality of the chosen assemblies (Li & Durbin 2009).

3.2.4 Structural and Functional Annotation

We utilized the automated annotation software Maker version 2.31.8 to structurally annotate the genomes of *G. flava* and *G. putterillii* (Cantarel et al. 2008). We used two of the three gene prediction tools available within the pipeline, SNAP and Augustus. SNAP was trained using gff files generated by CEGMA v2.5 (a program similar to BUSCO) (Parra et al. 2007). Augustus was trained with *Fusarium solani* protein models (v2.0.26) downloaded from Ensembl Fungi (EnsemblFungi 2015). The protein sequences generated by the structural annotation were blasted against the Swiss-Prot database to functionally annotate the genomes of *G. flava* and *G. putterillii* (Swiss-Prot 2015).

3.2.5 Assessing repetitive elements profile

To evaluate the repetitive elements profile of *G. flava* and *G. putterillii*, we masked the interspersed repeats within the assembled genomes with RepeatMasker 4.0.5 using the sensitive mode and default values as arguments (Smit et al. 1996).

3.2.6 Identifying putative genes involved in host-pathogen interactions

To search for putative genes contributing to pathogenicity, we conducted a BLASTp search with an e-value threshold of 1e-6 against the PHI-base database that includes known genes implicated in pathogenicity (PHI-base 2015). Additionally, we identified proteins that contain signal peptides and lack transmembrane domains in each *Geosmithia* species as well as their close relative, *Acremonium chrysogenum*, with SignalP 4.1 and TMHMM 2.0 (Peterson et al. 2011, Krogh et al. 2001).

3.2.7 Identifying carbohydrate-active proteins and peptidases

To identify enzymes capable of degrading carbohydrate molecules in species belonging to Hypocreales and *G. clavigera*, we performed a HMMER search against the CAZy database released July 2015 (Lombard et al. 2014) and filtered the results following the developer's recommendations. Lastly, we profiled the proteolytic enzymes present in species under examination using the *MEROPS* database 10.0 (Rawlings et al. 2016).

3.2.8 Phylogenetic analysis

3.2.8a Taxon Sampling

In order to determine phylogenetic position of *Geosmithia*, we combined the predicted peptide sequences from three *Geosmithia* species described here with the predicted peptide sequences of an additional 17 fungal genomes that represent the breadth of pathogens and non-pathogens within Ascomycota. Our dataset contained 11 pathogens and 9 non-pathogens. Table 3.2 lists the species used in this study and additional information regarding their taxonomy, ecological roles, and source databases.

3.2.8b Inferring Orthology

Orthologous peptide sequences among the 20 fungal genomes were determined using OrthoFinder version 0.3.0 (Emms & Kelly 2015). OrthoFinder performs an all-versus-all BLASTp (Altschul et al. 1990) search among a set of protein coding genes to infer orthogroups and aligns them using MAFFT (Katoh & Standley 2013). These orthogroups may contain paralogs as well as orthologs; because datasets rich in paralogs can confound phylogenomic analysis, the alignment files produced by OrthoFinder were parsed to recover only those orthogroups that contained single-copy orthologs from each of the 20 species. This resulted in 1,916 total orthogroups with 100% taxon occupancy.

3.2.8b Trimming Alignments

For each alignment, regions that contained gap rich sites were removed using –*gappout* option in trimAl v1.4.rev15 (Capella-Gutiérrez et al. 2009). Next, all files containing orthogroups were renamed so the respective headers among these files were identical and individual alignments were concatenated. Concatenation resulted in a single fasta file containing all 1,916 partitions with 1,054,662 sites at 100% taxon occupancy. This initial alignment was further filtered using MARE (v.0.1.2) (Misof et al. 2013), which reduced the character matrix to 247,627 sites. This reduced fasta alignment was converted into a partitioned phylip formatted file. Next, the best-fit substitution models for each partition and a global partitioning scheme were determined with PartitionFinder (v1.1.1) using hcluster clustering algorithm and default parameters (Lanfear et al. 2014).

3.2.8d Constructing Phylogeny

Maximum likelihood (ML) analysis was conducted in RaxML v 8.1.20 (Stamatakis 2014) leveraging the partitioning scheme determined by PartitionFinder. The ML tree and 200 bootstrap replicates were conducted in a single analysis using the –*f a* option. In addition, we conducted Bayesian Markov Chain Monte Carlo (BMCMC) analysis in MrBayes 3.2.6 (Ronquist et al. 2012). For MrBayes analysis, we specified the mixed amino acid model prior and ran the fully partitioned tree search for 215,000 generations. A consensus tree was then generated after discarding 50% of the run as burnin. The nexus file, including MrBayes block, provides other details of the MrBayes analysis (Supplementary File A).

3.2.9 Measuring genomic distances using MinHash

In addition to assessing the phylogenetic relationships among the fungal species in this study, we calculated approximate pair-wise distances based on whole genome sequences with a tool called Mash that utilizes the MinHash technique (Ondov et al. 2016). Mash reduces large clusters of sequences into MinHash sketches with a user-defined size and estimates a Mash distance and a corresponding *P*-value. We constructed a distance matrix with data from 100,000 sketches and default k-mer value of 21.

3.2.10 Detecting genes under positive selection

To identify genes under positive selection in *G. morbida* lineage, we compared *G. morbida* with all non-pathogens from the aforementioned 20 fungi used to estimate the species tree in Figure 3.1. Among this batch of 10 fungal species, we detected 22,908 protein orthogroups using OrthoFinder that contained paralogs as well as orthologs. Of these, only 9,560 orthogroups were alignable with MAFFT because many groups consisted of only one sequence from a single species (Katoh & Standley 2013). A total of 3,327 orthogroups, composed of single-copy orthologs, were sieved and corresponding coding DNA sequences for each peptide in these partitions were extracted using custom scripts (available at our Github repository – see link below).

The coding DNA sequences were then aligned with MACSE v1.01.b (Ranwez et al. 2011). This Java-based utility accounts for frameshifts and premature stop codons in coding sequences during the alignment process and outputs aligned protein and nucleotide sequences. In order to filter out alignments with frameshifts and internal stop codons, we utilized a program called PAL2NAL v14 (Suyama et al. 2006). This software searches for complementary regions between multiple protein alignments and the corresponding coding DNA sequences, and omits any problematic codons from the output file. This cleaning step reduced the number of 3,327 orthogroups to 2,798 that were used for detecting genes under selective pressures.

We used the branch-site model (BSM) in the CodeML program of package PAML v4.8 for selection analysis (Yang 2007). BSM permits ω (dN/dS) to vary among sites and branches and thus, allowing the identification of specific branches and sites subjected to selection. We computed two models in order to calculate and compare the likelihood values: a null model with a fixed ω value of 1 and an alternative model that estimates ω in the foreground branch, which is *G. morbida* in our case. In the effort to reduce false positives, we implemented the Benjamini-

Hochberg correction method when comparing likelihood ratios for null and alternative models using a *P*-value threshold of 0.05. We performed similar BLAST searches as mentioned previously to characterize the functions of these proteins, identify proteins with signal peptides and transmembrane domains, and assess which genes encode for putative pathogenic proteins.

We repeated the above procedures for detecting genes under selection in *Grosmannia clavigera* because this fungal pathogen plays an ecological role similar to *G. morbida*. By performing these analyses, we sought to uncover genes under adaptive evolution in both beetle-vectored tree pathogens.

3.3 Code Availability

All commands and scripts used in our analyses are available at GitHub repository associated with this paper (https://github.com/macmanes-lab/GeosmithiaComparativeGenomics).

3.4 Results and discussion

3.4.1 Assembly features

We recently assembled a reference genome for a *G. morbida* strain isolated from *J. californica* in Southern California. The reference contained 73 scaffolds with an estimated size of 26.5 Mbp. We predicted 6,273 protein models in this reference that were generated in-silico using the Maker annotation pipeline (Cantarel et al. 2008). In this work, we sequenced strains of *G. flava* and *G. putterillii* at approximately 102x and 131x coverage, respectively. The *G. flava* assembly composed of 1,819 scaffolds totaling in 29.47 Mbp in length, and *G. putterillii* genome contained 320 scaffolds with 29.99 Mbp. Both genomes possess 98% of the single-copy orthologs present in more than 90% of the fungal species. Additionally, 97% and 98% of the raw reads mapped back to *G. flava* and *G. putterillii* genome assemblies, respectively (Table 3.4). These length-based, completeness, and mapping statistics attest to the high quality of our genome assemblies. We estimated *G. flava* and *G. putterillii* possess 6,976 and 7,086 peptides, respectively. This increase in number of predict proteins may provide a larger toolkit enabling *G.*

flava and *G. putterillii* to interact with multiple bark beetle species, however experimental work is necessary to support this speculation.

An estimated 0.80% of *G. morbida* reference genome sequence represented repeats; however, 0.63% and 0.64% of the sequences in *G. flava* and *G. putterillii* consisted of repetitive elements. There are 60, 42, and 15 DNA transposons in *G. morbida*, *G. flava*, and *G. putterillii*, respectively. Furthermore, *G. morbida* possesses only 152 retroelements, whereas *G. flava* and *G. putterillii* have 401 and 214 of such elements, correspondingly.

3.4.2 Identifying putative genes involved in pathogenicity

The full BLASTp search results against the PHI-base database (v4.0) for *G. morbida*, *G. flava*, *G. putterillii* and *Acremonium chrysogenum* are available in the supplementary material (Table S1). Approximately 32%, 34%, and 34% of the total proteins in *G. morbida*, *G. flava* and *G. putterillii* respectively share homology with protein sequences in the database. The number of unknown proteins with hits in the PHI-base database is similar for *G. morbida* (26) and *G. flava* (28) in comparison to *G. putterillii* (37).

In comparison to *A. chrygosgenum* in Table 3.6, *G. morbida* contains four percent more proteins that putatively play a role during or after the infection process. The three *Geosmithia* species share 961 of the PHI-base proteins with their closest relative, *Acremonium chrysogenum*. *G. morbida* possesses only 14 unique PHI-base proteins when compared to *G. flava*, *G. putterillii* and *A. chrysogenum* (Figure 3.3). However, nine of the 14 genes encoded products that were proven to disrupt pathogenicity; whereas the remaining predicted proteins did not phenotypically affect pathogenicity. One of these unique genes important for pathogenicity encodes for polyketide synthase, which is involved in microbial secondary metabolism that confers virulence (Tsai et al. 1998, Gaffoor et al. 2005). Other genes that may contribute to pathogenicity in *G. morbida* are involved in environmental stress response and conidium and appresorium formation (Cervantes-Chávez et al. 2011, Chen et al. 2008, Ryder & Talbot 2015). *G. morbida* contains Calcineurin regulatory subunit B, a gene that controls stress response. Studies show that transformants lacking this gene were unable to form proper cell-wall in *Ustilago hordei* (Cervantes-Chávez et al. 2011, Kraus & Heitman 2003).

Moreover, the protein product encoded by a gene known as Rac1, is a GTPase that plays a role in directed cellular growth and the development of appressorium that allows the pathogens to penetrate the host surface (Chen et al. 2008, Rolke & Tudzynski 2008, Harris 2011). Another gene present only in *G. morbida* that affects appressorial construction encodes for adenylate cyclase that is located within the cell membrane and catalyzes cAMP from ATP (Steer 1975). cAMP monitors signaling pathways that control morphogenesis, hyphal development, and virulence in various fungal pathogens (Choi & Dean 1997, Adachi & Hamer 1998, D'Souza & Heitman 2001, Barhoom & Sharon 2004). Although the specific penetration mechanisms are not known in *G. morbida*, the presence of Rac1 and adenylate cyclase homologs suggests that this fungus employs similar invasion strategies as other known plant pathogens such as *Magnaporthe grisea* (rice blast fungus), *Claviceps purpurea* (ergot fungus), and *Ustilago maydis* (corn smut fungus).

3.4.3 Identifying putative secreted proteins

A total of 349, 403, and 395 proteins in *G. morbida*, *G. flava*, and *G. putterillii* contain signal peptides respectively. Of these putative signal peptides, *G. morbida* encodes 27 proteins (7.7%) with unknown function, whereas *G. flava* and *G. putterillii* contain 29 (7.2%) and 30 (7.6%) unknown proteins. The difference in percent of unknown proteins with signal peptides is minimal among the three genomes. For each species, proteins containing signal peptides were subjected to a membrane protein topology search using TMHMM v2.0. There were 237, 281, and 283 proteins in *G. morbida*, *G. flava*, and *G. putterillii* that lacked any transmembrane protein domains. Again, these numbers of proteins are very similar.

3.4.4 Profiling carbohydrate active enzymes and peptidases

CAZymes are carbohydrate active enzymes that break down plant structural components enabling initiation and establishment of infection. We assessed the CAZymatic profile of all species in the order Hypocreales, *Geosmithia* species, and *Grosmannia clavigera* (Figure 3.1). The glycoside hydrolase (GH) family members dominated all protein models followed by glycosyltransferase (GT) family. The two most prominent families among all fungal species were GH3 and GH16 (supplementary material Table S2). GH3 hydrolases are involved in cell wall degradation and defense against the host immune system, and GH16 enzymes fulfill a wide range of cellular functions including transporting amino acids. The third most representative family was GH18; however *G. morbida* only contains four of these enzymes. In contrast, this number for other species ranges from 9 to 31 enzymes. Along with acetylglucosaminidases, family GH18 harbors chitinases that assist in the production of carbon and nitrogen. In terms of other CAZyme families, all fungi express a similar overall distribution, with *F. solani* being the only exception, which contains more CAZymes than any other pathogen and non-pathogen. This *Fusarium* species is a necrotrophic pathogen that is hypothesized to possess more CAZymes than biotrophic and hemibiotrophic fungi. This discrepancy may be due to the fact that necrotrophic pathogens require an extensive toolkit to promote host cell death as quickly as possible; whereas biotrophs need to keep the host alive and dispensing large number of degradation enzymes can be detrimental to that aim (Zhao et al. 2013).

In addition to profiling CAZymes, we also performed a BLAST search against the peptidase database, Merops v10.0 (Rawlings et al. 2016), for each Hypocreales, *C. platani*, and *G. clavigera* genome. Among the pathogens, *G. morbida* has the third highest percent of predicted proteases after *Cordyceps militaris* (insect pathogen) and *G. clavigera* (Figure 3.2). Moreover, *G. flava* and *G. putterillii* have largest percent of peptidases among the nonpathogenic fungi. All three *Geosmithia* species illustrate similar proteolytic profiles and contain no glumatic and mixed peptidases.

3.4.5 Inferring phylogeny

Even though the *Geosmithia* genus was first established in 1979, it has only recently been described in depth. One of the main objectives in this study was to uncover the phylogenetic relationship between *Geosmithia* species and other fungal pathogens using protein coding DNA sequence data. In order to determine the broader evolutionary history of *Geosmithia* species, we constructed a maximum likelihood (ML) and Bayesian Markov Chain Monte Carlo (BMCMC) phylogenies using 1,916 single-copy orthologs from *G. morbida*, *G. putterillii*, *G. flava*, and 17 additional fungal taxa (Table 3.2). Our final dataset consisted of 11 pathogens and 9 non-pathogens.

After trimming and filtering, our 1,916 orthogroups contained approximately $1e10^6$ amino acid sites in total. The topologies of trees generated under ML and BMCMC were identical and all branches in all analyses received bootstrap support of 100% (ML) and posterior probabilities of 1.0 (BMCMC) (Figure 3.3). Both analyses resulted in identical tree topology that is consistent with prior work (Fitzpatrick et al. 2006, Wang et al. 2009). Our phylogenetic analysis places *Geosmithia* species in the order of Hypocreales and confirms that the closest relative to this genus is *Acremonium chrysogenum* (Kolarik et al. 2011).

3.4.6 Genomic distances

The genomic distance dendrogram does not align with the phylgeny built using single copy orthologs (Figure 3.4). The *Geosmithia* clade is consistent with the predicted species tree in Figure 3.3, however five of the 13 Hypocreales group separately from the others. *Trichoderma virens* is the most distantly related fungus. The disparity between the phylogeny and the distance based dendrogram can simply be explained by their divergent construction approaches. The species phylogeny (Figure 3.3) depends solely on protein coding sequences, but the heatmap (Figure 3.4) incorporates both coding and noncoding regions of the genomes.

3.4.7 Genes under positive selection

In order to detect genes under positive selection in the *G. morbida* lineage, we first searched for all single-copy orthologs shared among the 9 non-pathogens and *G. morbida* using OrthoFinder (v0.3.0). Briefly, this program conducts all-versus-all BLASTp searches to find orthologous peptides. Using a custom Python script, we extracted the corresponding coding sequences for each protein in the 3,327 orthogroups containing 1:1 orthologs. These orthogroups were aligned using MACSE v1.01b and cleaned with PAL2NAL v14. This step resulted in 2,798 multiple sequence alignments that were used for selection analysis.

To identify coding sequences and sites experiencing selection, we leveraged the branch-site model in PAML's codeml program (v4.8). *Geosmithia morbida* was selected as the foreground branch. Our results showed 38 genes to be under positive selection at an adjusted *P*-value < 0.05. Next, we performed a functional search for each protein by blasting the peptide sequences against the NCBI non-redundant and pfam databases. We determined that several were involved in catabolic activity, gene regulation, and cellular transport.

For instance, a cullin3-like protein belonging to a group of structurally similar molecules involved in protein degradation, such as the Skp-Cullin-F-box (SCF) ubiquitin ligase complex, was predicted to be under positive selection (Pintard et al. 2004, Cardozo & Pagano 2004). Furthermore, a ubiquitin-conjugating enzyme (E2) that interacts with cullin3 to prepare substrate for degradation, also had a dn/ds > 1, indicating that both genes are under positive selection within *G. morbida*. Although little is known regarding the precise functional abilities of these complexes, it is possible these proteins are involved in pathogenicity of *G. morbida*. Previous studies have also implicated ubiquitin ligase complexes in infection and disease development (Duyvesteijn et al. 2005, Han et al. 2007).

Additionally, our analysis showed a regulatory protein homologous to the basic leucine zipper (bZIP) transcription factor under selection. The bZIP proteins are similar to AP-1 transcription factors and monitor several developmental and physiological processes including oxidative stress responses in eukaryotes (Corrêa et al. 2008). Fungal pathogens such as the rice blast fungus *Magnaporthe oryzae* express AP1-like transcription factor called MoAP1 that contains bZIP domain. MoAP1 is highly active during infection and is translocated from the cytoplasm to the nucleus in response to oxidative stress induced by H_2O_2 (Guo et al. 2011). Furthermore, the researchers showed that MoAP1 regulates enzymes such as laccase and glutamate decarboxylase that are involved in lignin breakdown and metabolism of γ-aminobutyric acid, respectively (Janusz et al. 2013, Baldrian 2005, Solomon & Oliver 2002). Some of the other positively selected genes include ABC1 transporter, proteases, proteins involved in apoptosis and DNA replication and repair. Lastly, only five of the 38 genes encoded proteins with unknown functions. A complete list of genes and their functions is provided in the supplementary material (Table S3).

3.4.8 Transmembrane protein and effector genes

Our analysis of the 38 proteins under positive selection showed that 11 of these possess at least one or more transmembrane domains. Given nearly 30% of the positively selected genes are membrane bound suggests that interactions with the host surface are drivers of evolution within *G. morbida*. Transmembrane proteins are important mediators between a host and its pathogens during microbial invasion. Fungal pathogens either penetrate a surface or enter the host through a wound or opening such as stomata in order to gain access to the nutrients in the plant (Chisholm et al. 2006). Once the infiltration process is completed, pathogens are exposed to host plasma membrane receptors that detect pathogen-associated molecular patterns (PAMP) and induce PAMP-triggered immunity (PTI) to prevent further proliferation of the microbe. Transmembrane proteins a fungal pathogen expresses within its membrane are crucial during PTI because they are responsible for suppressing PTI directly or by secreting effector molecules, which contain signal peptides necessary for proper targeting and transport (Boller & He 2009, Chisholm et al. 2006). However, we found no protein that contained a signal peptide indicating none of these proteins are secretory in nature. This finding is significant because it demonstrates that the secretome of *G. morbida* is not under positive selective pressures, and that this pathogen may be utilizing conserved effector proteins.

We also performed a BLASTp search against the phibase database (version 4.0) and found that seven out of 38 proteins shared homology with experimentally confirmed genes involved in pathogenicity. These seven putatively pathogenic genes do not contain any transmembrane domains or signal peptides.

3.4.9 Genes under adaptive evolution in beetle-vectored fungal pathogens

In addition to detecting genes under selective pressures in *G. morbida*, we performed the same selection analysis for *Grosmannia clavigera* to identify candidate proteins that may help explain adaptations these beetle-vectored species have evolved in light of their specific ecological roles. We found that *G. clavigera* possesses 42 positively selected genes that share protein domains with only two of the 38 genes predicted to be under selection in *G. morbida* (Supplementary table S4). The two overlapping motifs are methyltransferase and protein kinase domains. Our KEGG analysis exhibited no common pathways between *G. morbida* and *G. clavigera*. These results confirm that organisms can evolve vastly different mechanisms that give rise to similar phenotypic traits. Although both *G. morbida* and *G. clavigera* have similar niches, their host ranges are unrelated and the beetle vectors are also distinct. Given that many factors govern the evolution of each player in a vector-host-pathogen complex, it makes sense that the suite of pathogenic tools does not overlap between *G. morbida* and *G. clavigera*, but rather represents independent evolutionary trajectories for the development of pathogenicity in each fungal species.

3.5 Conclusion

This study aims to provide insight into the evolution of pathogenicity within *Geosmithia morbida*, a beetle vectored pathogen that is the causal agent of Thousand Cankers Disease in *Julgans nigra* (eastern black walnut). Here, we present *de novo* genome assemblies of two nonpathogenic *Geosmithia* species, *G. flava* and *G.* putterillii, and employ comparative genomics approach to uncover the molecular factors contributing to pathogenicity in *G. morbida*.

G. flava and *G. putterillii* have estimated genome sizes of 29.6 Mbp and 30.0 Mbp, correspondingly. These assemblies are larger than the genome of *G. morbida*, which measures 26.5 Mbp in length. Furthermore, in contrast to other species in the phylogeny (Figure 3.3), tree fungal associates, namely *Geosmithia* species, *G. clavigera*, and *C. platani* have reduced genomes and gene content. We predict this genome and gene content reduction is a result of evolving specialized lifestyles with a narrow host range. For instance, all three *Geosmithia* species and *G. clavigera* are vectored into their respective hosts via bark beetles, which may restrict the evolutionary processes because these fungi must adapt to their vectors and hosts simultaneously. Moreover, possessing genes that are not essential for this specialized lifestyle may impose a fitness disadvantage on the pathogen. A recent study characterizing the genome of mycoparasite *Escovopsis weberi*, exhibited that specialized pathogens tend to have smaller genomes and predicted protein sets because they lack genes that are not required beyond their restricted niche when compared to close generalist relatives (de Man et al. 2016).

Our results also illustrate that three *Geosmithia* species are highly similar based on genomic distances estimated with Mash. Furthermore, although one might expect that *G. morbida* harbors more carbohydrate binding enzymes and peptidases conferring pathogenicity only in *G. morbida*, our results indicate that all three species have similar enzymatic profiles (Figures 3.1 and 3.2). Despite these congruencies, our PAML analysis showed the presence of 38 genes under positive selection in *G. morbida* when compared to other nonpathgens within the order Hypocreales. These genes encode for proteins that have been implicated in pathogenicity in other fungal pathogens such as *Magnaporthe oryzae*. Additionally, we found peptides with protein kinase and methyltransferase domains that are under positive selection in both *G. morbida* and *G. clavigera.* Proteins kinases were previously shown to be under strong positive selection in *G. clavigera* (Alamouti et al. 2014). This result suggests the key contributions that protein kinases make in initiating signal transduction pathways during pathogen host interactions. Our study identified a small set of significant genes that are potentially involved in the evolution of pathogenicity in the genus *Geosmithia*. Functional experiments will be needed in order to valid our predictions.

3.6 Tables and Figures

Table 3.1. Species, geographic regions, *Juglans* host information for *Geosmithia morbida*, *Geosmithia flava*, and *Geosmithia putterillii*.

Species	Isolate	Cluster	Haplotype	Geographic region	County	Host
*G. morbida**	1262	1	H03	California	Ventura	*J. californica*
G. flava	CCF3333	-	-	Czech Republic	-	*Castenea sativa*
G. putterillii	CCF4204	-	-	California	-	*J. californica*

Table 3.2. Fungal species used for phylogenetic analysis

Species	Class	Order	Ecological role	Download source	References
Geosmithia morbida	Sordariomycetes	Hypocreales	Pathogen	-	Schuelke et al. 2016
Geosmithia flava	Sordariomycetes	Hypocreales	Non-pathogen	-	-
Geosmithia putterillii	Sordariomycetes	Hypocreales	Non-pathogen	-	-
Acremonium chrysogenum	Sordariomycetes	Hypocreales	Beneficial	FungalEnsembl	Terfehr et al. 2014
Stanjemonium griseum	Sordariomycetes	Hypocreales	Saprotrophic	JGI	Used with permission
Trichoderma virens	Sordariomycetes	Hypocreales	Beneficial	JGI	Kubicek et al. 2011
Trichoderma reesei	Sordariomycetes	Hypocreales	Saprotrophic	FungalEnsembl	Martinez et al. 2008
Ustilaginoidea virens	Sordariomycetes	Hypocreales	Biotrophic pathogen	FungalEnsembl	Zhang et al. 2014
Cordyceps militaris	Sordariomycetes	Hypocreales	Insect pathogen	FungalEnsembl	Zheng et al. 2011
Myrothecium inundatum	Sordariomycetes	Hypocreales	Saprotrophic	JGI	Used with permission
Fusarium solani	Sordariomycetes	Hypocreales	Necrotrophic pathogen	FungalEnsembl	Coleman et al. 2009
Fusarium graminearum	Sordariomycetes	Hypocreales	Necrotrophic pathogen	FungalEnsembl	Trail et al. 2003, Cuomo et al. 2007, Ma et al. 2010
Ceratocystis platani	Sordariomycetes	Microascales	Pathogen	FungalEnsembl	Belbahir 2015
Neurospora crassa	Sordariomycetes	Sordariales	Saprotrophic	FungalEnsembl	Galagan et al. 2003
Chaetomium globosum	Sordariomycetes	Sordariales	Saprotrophic	JGI	Berka et al. 2011
Grosmannia clavigera	Sordariomycetes	Ophiostomatales	Pathogen	FungalEnsembl	DiGuistini et al. 2011

Eutypa lata	Sordariomycetes	Xylariales	Pathogen	JGI	Blanco-Ulate et al. 2013
Botrytis cinerea	Leotiomycetes	Helotiales	Necrotrophic pathogen	FungalEnsembl	Amselem et al. 2011, Staats & van Kan, 2012

Table 3.3. Statistics for *Geosmithia morbida* isolates, *Geosmithia flava* and *Geosmithia putterillii* sequence data.

Species	Total read pairs		Est. coverage	
G. morbida	14,013,863*	20,674,289*	109*	160*
G. flava	16,183,281		102	
G. putterillii	19,711,745		131	

*These values are for paired-end read data for *G. morbida* from Schuelke et al. 2016.

Table 3.4. Length-based statistics for *Geosmithia morbida* isolates, *Geosmithia flava*, and *Geosmithia putterillii* generated with QUAST v2.3.

Species	Est. genome size (Mbp)	*k*-mer for ABySS assembly	Scaffold count	Largest scaffold	NG50*	LG50*	Genome completeness
G. morbida	26.5	NA[1]	73	2,597,956	1,305,468	7	**98**
G. flava	29.6	91	1,819	1,534,325	460,430	22	**98**
G. putterillii	30.0	91	320	2,758,267	1,379,352	9	**98**

The average GC content for *G. morbida*, *G. flava*, and *G. putterillii* equals 54%, 52%, and 55.5% respectively. The estimated genome sizes of *G. morbida*, *G. flava*, and *G. putterillii* are 26.5 Mbp, 29.6 Mbp, and 30.0 Mbp, respectively. The genome completeness values were produced with BUSCO v1.1b1. These percentages represent genes that are complete and not duplicated or fragmented.

1Genome assembly for *G. morbida* was constructed using AllPaths-LG (v49414). See Chapter 2 for further details.

*NG50 is the scaffold length such that considering scaffolds of equal or longer length produce 50% of the bases of the reference genome. LG50 is the number of scaffolds with length NG50.

Table 3.5. Repetitive elements profile of *Geosmithia* and 9 additional fungal species generated with RepeatMasker v4.0.5.

	Genome size (Mbp)	% GC	% Bases masked	Num. of Retroelements	Num. of DNA transposons
G. morbida	26.5	54	0.81	152	60
G. flava	29.6	52	0.63	401	42
G. putterillii	30.0	55.5	0.64	214	15

Table 3.6. Gene count and PHI-base results for *Geosmithia morbida*, *Geosmithia flava*, *Geosmithia putterillii*, and *Acremonium chrysogenum*.

Species	Total number of genes	% of total genes homologous to pathogenic genes	% of unique genes homologous to pathogenic genes
G. morbida	6,273	36	32
G. putterillii	7,086	40	34
G. flava	6,976	38	34
A. chrysogenum	8,901	32	28

Table 3.7. Functions of 14 PHI-base proteins that are present only in *Geosmithia morbida*.

Accession ID	Function	Mutant phenotype	Pathogen
PHI:101	Polyketide synthase	Reduced virulence	*Aspergillus fumigatus*
PHI:1183	Protein kinase	Reduced virulence	*Fusarium graminearum*
PHI:138	Histidine kinase	Reduced virulence	*Candida albicans*
PHI:1509	Transcription factor	Unaffected pathogenicity	*Fusarium graminearum*
PHI:1635	Transcription factor	Unaffected pathogenicity	*Fusarium graminearum*
PHI:1780	Transcription factor	Unaffected pathogenicity	*Fusarium graminearum*
PHI:2054	Actin cytoskeleton organization, polarized cellular growth, conidiogenesis	Loss of pathogenicity	*Magnaporthe oryzae*
PHI:2339	Cell-wall integrity	Reduced virulence	*Ustilago hordei*
PHI:241	Adenylate cyclase	Loss of pathogenicity	*Cryptococcus neoformans*
PHI:2425	T-toxin production	Reduced virulence	*Cochliobolus heterostrophus*
PHI:2342	Necrosis and ethylene-inducing protein	Unaffacted pathogenicity	*Botrytis elliptica*
PHI:2924	Lipase	Unaffacted pathogenicity	*Fusarium oxysporum*
PHI:72	Aspartyl proteinase	Reduced virulence	*Candida albicans*
PHI:2504	Putative α-1,3-glucan synthase	Unaffacted pathogenicity	*Aspergillus fumigatus*

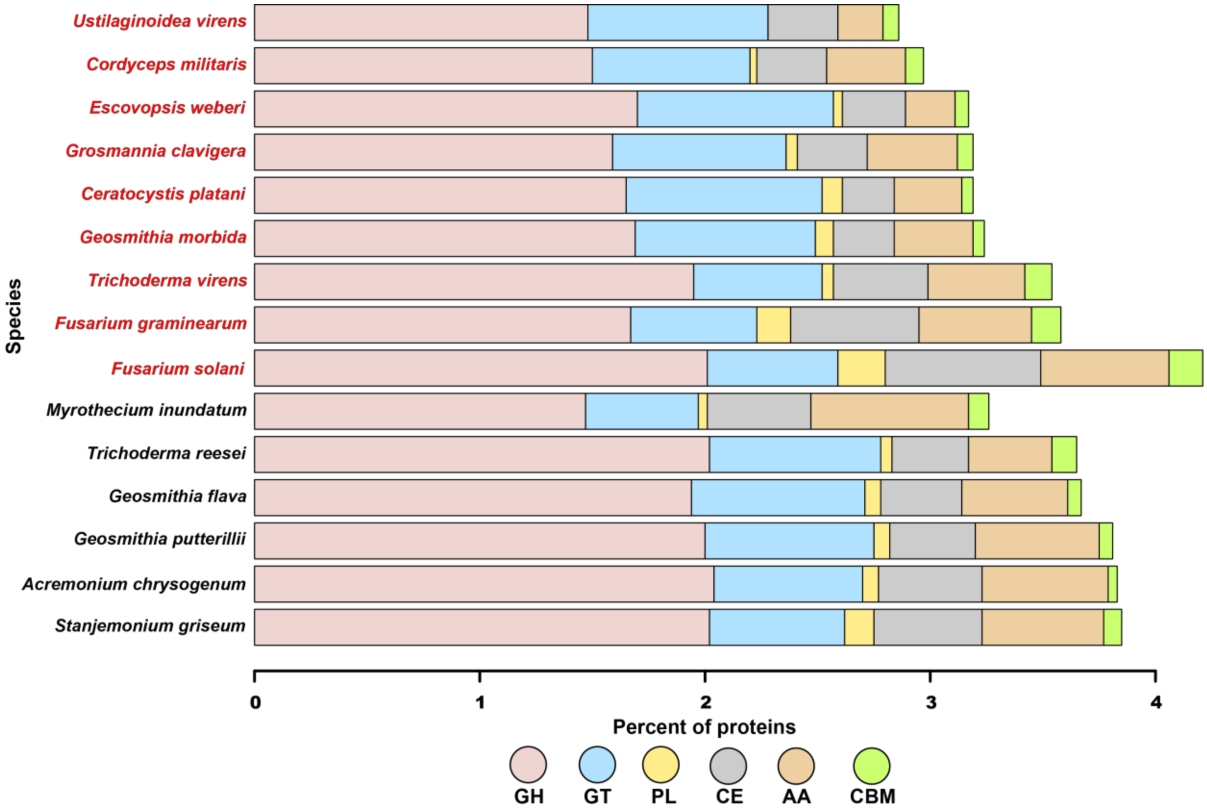

Figure 3.1. Carbohydrate active enzymes (CAZymes) distribution for *Geosmithia* species, other Hypocreales, and *Ceratocystis platani*. The species in red are pathogens, while the names in black are nonpathogens. CAZymes were identified with HMMer searches of dbCAN peptide models. GH: glycoside hydrolases, GT: glycosyltransferases, PL: polysaccharide lyases, CE: carbohydrate esterases, AA: auxiliary activities enzymes, and CBM: carbohydrate-binding molecules.

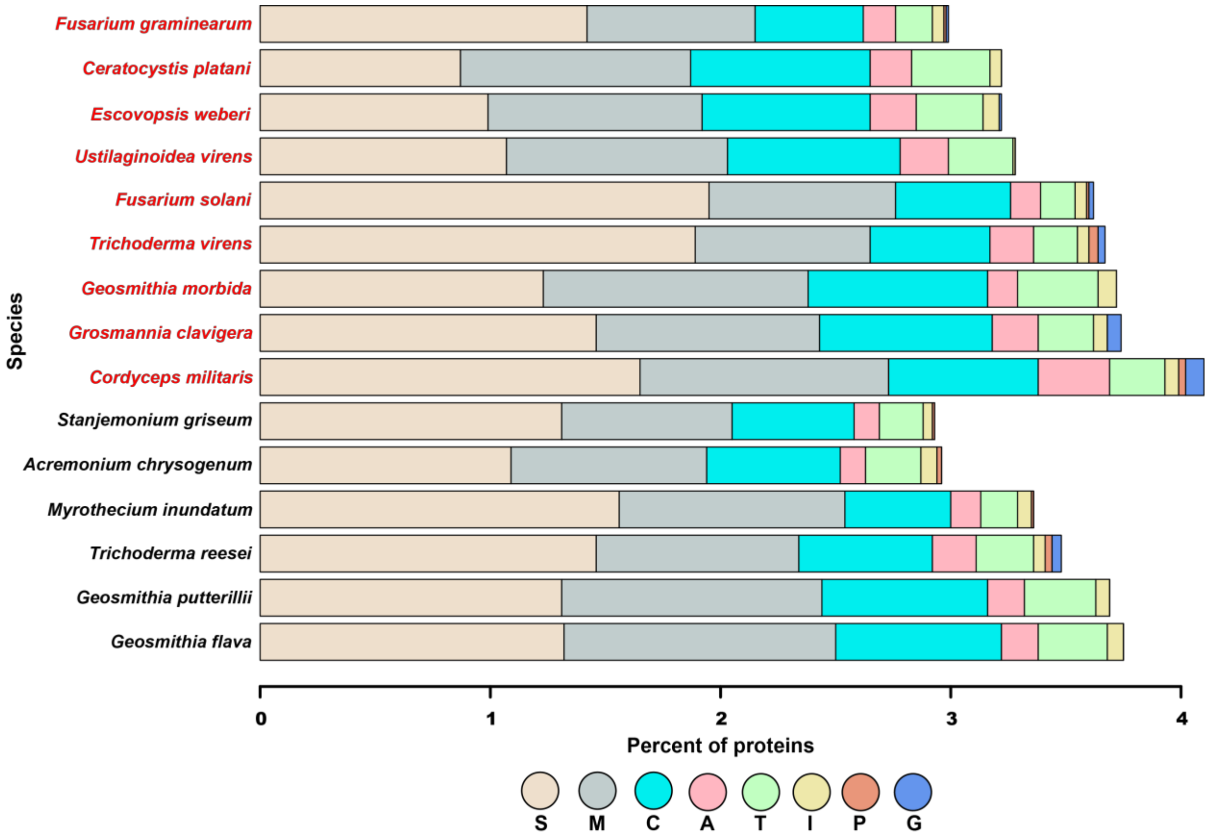

Figure 3.2. Proteolytic enzymes distribution for *Geosmithia* species, other Hypocreales, and *Ceratocystis platani*. The species in red are pathogens, while the names in black are nonpathogens. Proteases were identified using BLASTp searches against the MEROPs database v10. S: serine, M: metallo, C: cysteine, A: aspartic, T:threonine, I: inhibitors, P: mixed, G: glutamic.

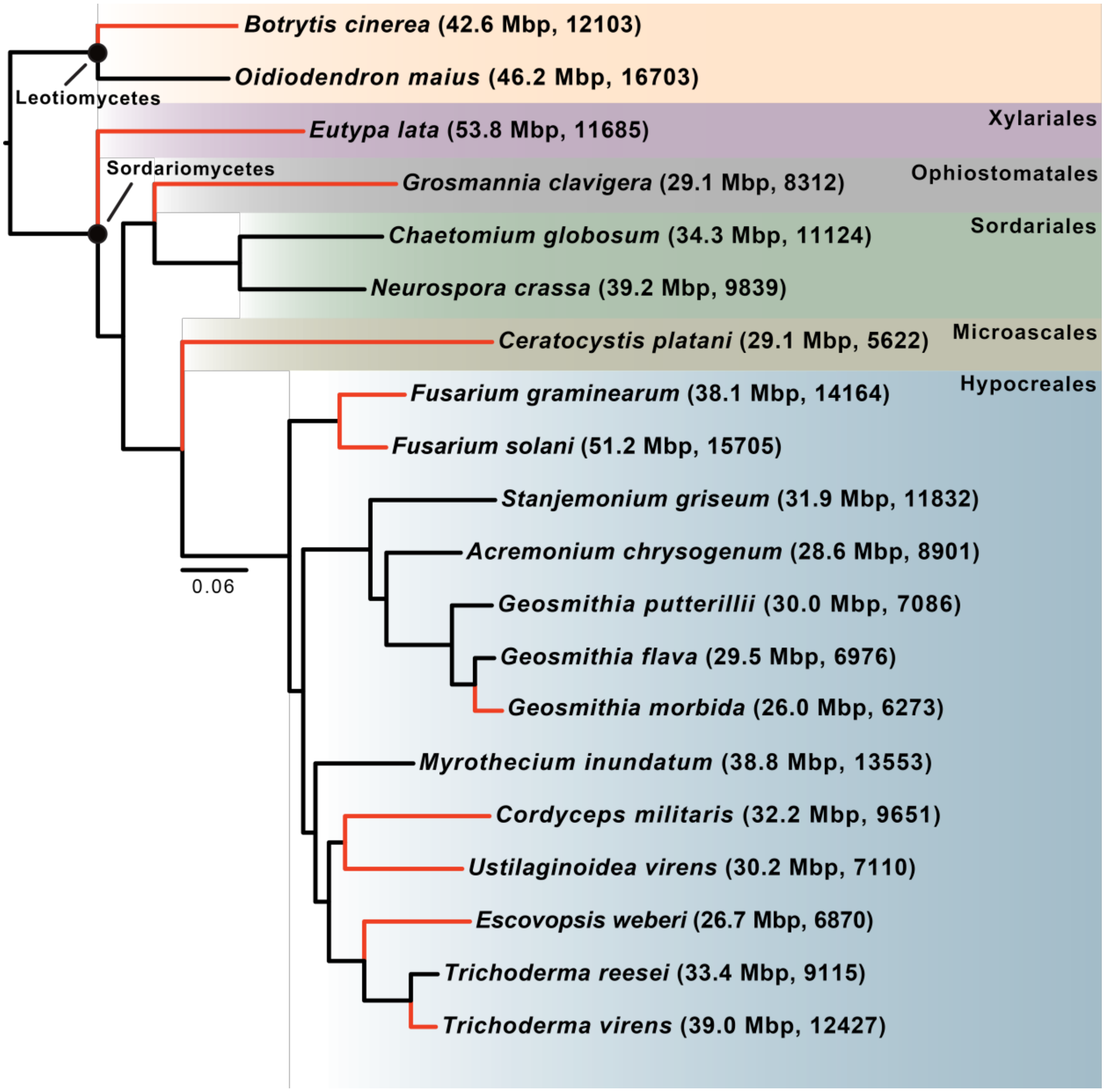
Botrytis cinerea (42.6 Mbp, 12103)
Oidiodendron maius (46.2 Mbp, 16703)
Leotiomycetes
Eutypa lata (53.8 Mbp, 11685)
Xylariales
Sordariomycetes
Grosmannia clavigera (29.1 Mbp, 8312)
Ophiostomatales
Chaetomium globosum (34.3 Mbp, 11124)
Sordariales
Neurospora crassa (39.2 Mbp, 9839)
Ceratocystis platani (29.1 Mbp, 5622)
Microascales
Fusarium graminearum (38.1 Mbp, 14164)
Hypocreales
Fusarium solani (51.2 Mbp, 15705)
Stanjemonium griseum (31.9 Mbp, 11832)
Acremonium chrysogenum (28.6 Mbp, 8901)
0.06
Geosmithia putterillii (30.0 Mbp, 7086)
Geosmithia flava (29.5 Mbp, 6976)
Geosmithia morbida (26.0 Mbp, 6273)
Myrothecium inundatum (38.8 Mbp, 13553)
Cordyceps militaris (32.2 Mbp, 9651)
Ustilaginoidea virens (30.2 Mbp, 7110)
Escovopsis weberi (26.7 Mbp, 6870)
Trichoderma reesei (33.4 Mbp, 9115)
Trichoderma virens (39.0 Mbp, 12427)

Figure 3.3. The Bayesian Markov Chain Monte Carlo (BMCMC) phylogeny was estimated using the mixed amino acid model in MrBayes (Ronquist et al. 2012) on a dataset containing 89,999 positions. This topology is identical to partitioned analyses conducted in RAxML (Stamatakis 2014). All nodes in BMCMC and ML analyses receive maximum support. The black circles symbolize classes. The color-shaded boxes at the right of the figure denote the orders within each class. The first and second numbers in parentheses represent the genome sizes in Mbp and the number of predicted protein models, respectively. Black and red branches correspond to non-pathogens and pathogens, which span multiple orders.

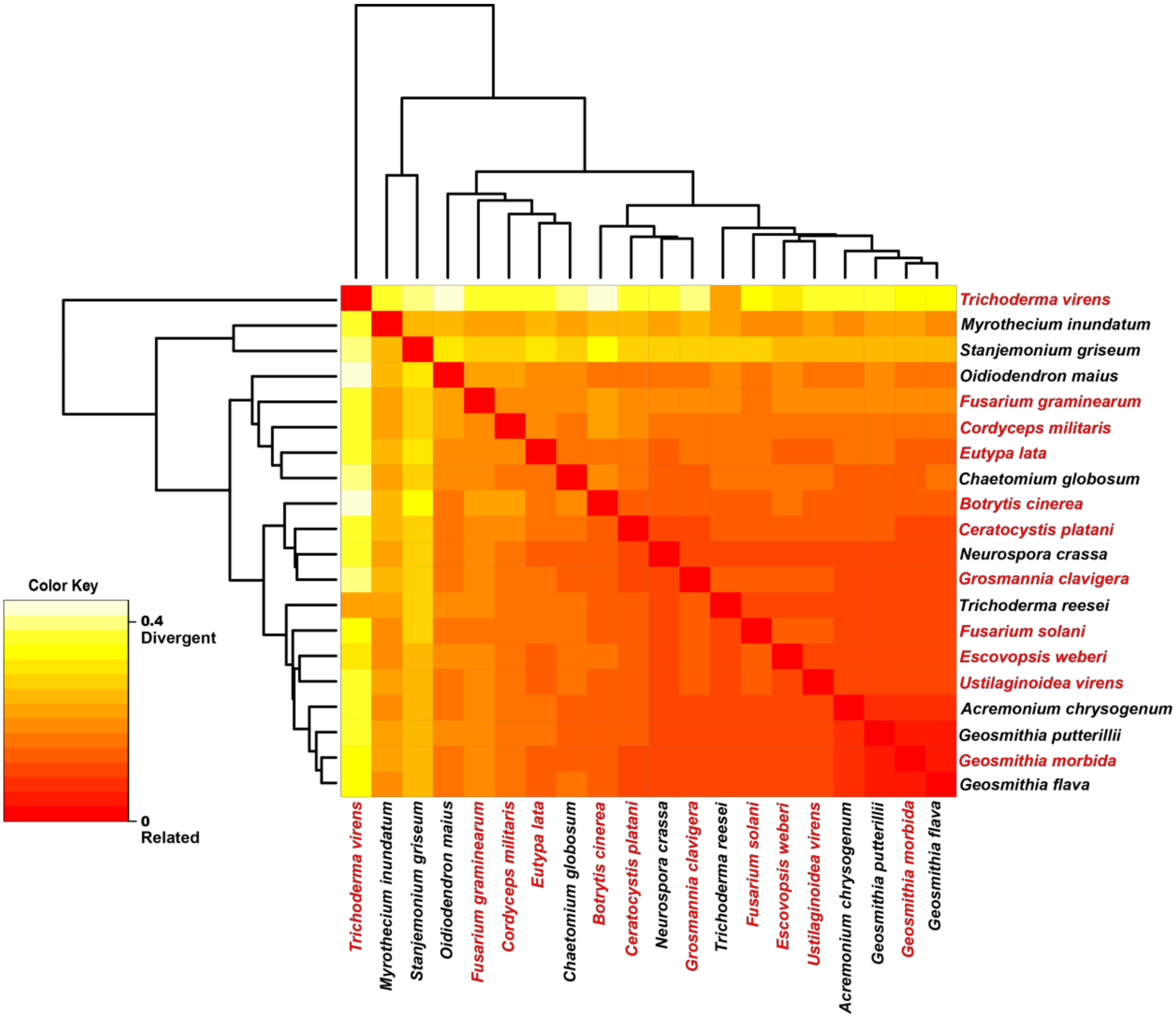

Figure 3.4. The distance matrix illustrated as a heatmap. A value close to 0 indicates high similarity; whereas 1 represents divergent species. The largest distance between two species in the matrix was 0.38. Black and red branches correspond to non-pathogens and pathogens. The distance matrix was generated with Mash (Ondov et al. 2016).

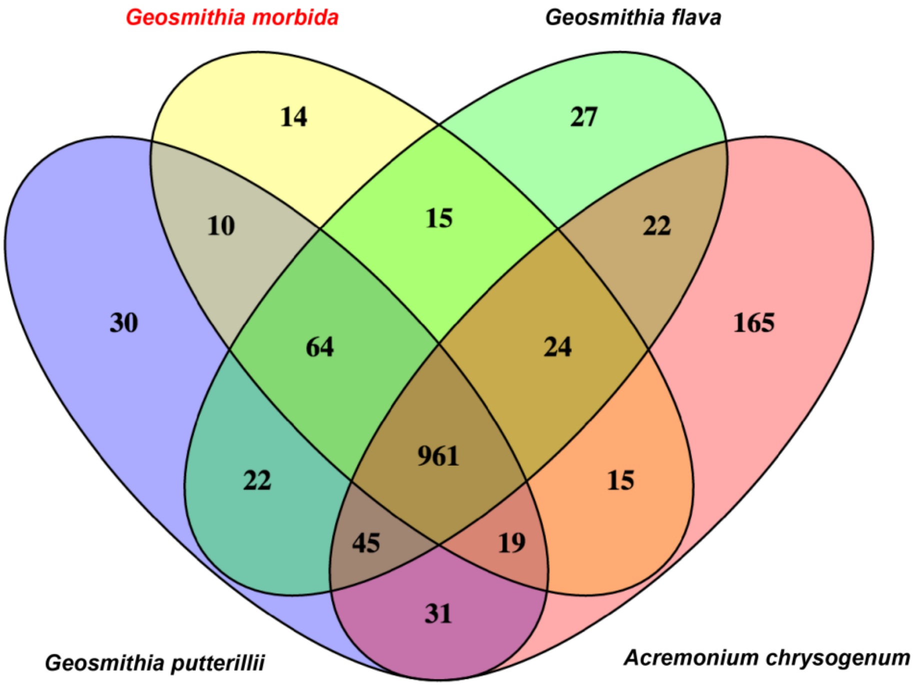

Figure 3.5. Number of shared PHI-base proteins based on Phibase accession IDs among *Geosmithia morbida*, *Geosmithia flava*, *Geosmithia putterillii*, and *Acremonium chrysogenum*.

3.7 References

1. Adachi K, Hamer JE. 1998. Divergent cAMP Signaling Pathways Regulate Growth and Pathogenesis in the Rice Blast Fungus *Magnaporthe grisea. Plant Cell. 10:1361-1373.*
2. Alamouti SM, Haridas S, Feau N, Robertson G, Bohlmann J, et al. 2014. Comparative Genomics of the Pine Pathogens and Beetle Symbionts in the Genus *Grosmannia*. Mol Biol Evol. 31:1454-1474.
3. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–10.
4. Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, et al. 2011. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. Plos Genet. 7:e1002230.
5. Baldrian P. 2005. Fungal laccases-occurrence and properties. FEMS Microbiol Rev. 30:215-242.
6. Barhoom S, Sharon A. 2004. cAMP regulation of “pathogenic” and “saprophytic” fungal spore germination. Fungal Genet Biol. 41:317-326.
7. Belbahri L. 2015. Genome sequence of *Ceratocystis platani*, a major pathogen of plane trees. Available at http://www.ncbi.nlm.nih.gov/nuccore/814603118.
8. Berka RM, Grigoriev IV, Otillar R, Salamov A, Grimwood J, et al. 2011. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. Nat Biotechnol. 29:922-927.
9. Blanco-Ulate B, Rolshausen PE, Cantu D. 2013. Draft Genome Sequence of the Grapevine Dieback Fungus *Eutypa lata* UCR-EL1. Genome Announc. 1:e00228-13.
10. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30:2114-2120.

11. Boller T, He SY. 2009. Innate immunity in plants: An arms race between pattern recognition receptors in plants and effectors in microbial pathogens. Science. 324:742-744.

12. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18:188-196

13. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 25:1972-1973.

14. Cardozo T, Pagano M. 2004. The SCF ubiquitin ligase: insights into a molecular machine. Nat Rev Mol Cell Biol. 5:739-751.

15. Cervantes-Chávez JA, Ali S, Bakkeren G. 2011. Response to environmental stresses, cell-wall integrity, and virulence are orchestrated through the calcineurin pathway in *Ustilago hordei*. Mol Plant Microbe Interact. 24:219-232.

16. Chen J, Zheng W, Zheng S, Zhang D, Sang W, et al. 2008. Rac1 is required for pathogenicity and Chm1-dependent conidiogenesis in rice fungal pathogen *Magnaporthe grisea*. PLoS Pathog. 4:e1000202.

17. Chisholm ST, Coaker G, Day B, Staskawicz BJ. Host-microbe interactions: shaping the evolution of the plant immune response. Cell. 124:803-814.

18. Choi W, Dean RA. 1997. The adenylate cyclase gene MAC1 of *Magnaporthe grisea* controls appressorium formation and other aspects of growth and development. Plant Cell. 9:1973-1983.

19. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, et al. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 5:e1000618.

20. Corrêa LG, Riano-Pachon DM, Schrago CG, dos Santos RV, Mueller-Roeber B, et al. 2008. The Role of bZIP Transcription Factors in Green Plant Evolution: Adaptive Features Emerging from Four Founder Genes. PLoS One. 3:e2944.

21. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, et al. 2015. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res. 4:900.

22. Cuomo CA, Gldener U, Xu JR, Trail F, Turgeon BG, et al. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science. 317:1400-1402.

23. D’Souza C, Heitman J. 2001. Conserved cAMP signaling cascades regulate fungal development and virulence. FEMS Microbiol Rev. 25:349-364.

24. de Man TJB, Stajich JE, Kubicek CP, Teiling C, Chenthamara K, et al. 2016. Small genome of the fungus *Escovopsis weberi*, a specialized disease agent of ant agriculture. PNAS. 113:3567-3572.

25. DiGuistini S, Wang Y, Liao NY, Taylor G, Tanguay P, et al. 2011. Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. PNAS. 108:2504-2509.

26. Duyvesteijn RG, van Wijk R, Boer Y, Rep M, Cornelissen BJ, et al. 2005. Frp1 is a *Fusarium oxysporum* F-box protein required for pathogenicity on tomato. Mol Microbiol. 57:1051-1063.

27. Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 16:157:2-14.

28. EnsemblFungi. 2015. Available at http://fungi.ensembl.org/index.html. Accessed 14 Nov. 2015.

29. Fitzpatrick DA, Logue ME, Stajich JE, Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evol Biol. 6:99.

30. Gaffoor I, Brown DW, Plattner R, Proctor RH, Qi W, et al. 2005. Functional analysis of the Polyketide Synthase Genes in the Filamentous Fungus *Gibberella zeae* (Anamorph *Fusarium graminearum*). Eukaryot Cell. 4:1926-1933.

31. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature. 422:859-868.

32. Guo M, Chen Y, Du Y, Dong Y, Guo W, et al. 2011. The bZIP Transcription Factor MoAP1 Mediates the Oxidative Stress Response and Is Critical for Pathogenicity of the Rice Blast Fungus *Magnaporthe oryzae*. PLoS Pathog. 7:e1001302.

33. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 29:1072–1075.

34. Han YK, Kim MD, Lee SH, Yun SH, Lee YW. 2007. A novel F-box protein involved in sexual development and pathogenesis in *Gibberella zeae*. Mol Microbiol. 63:768-779.

35. Harris SD. 2011. Cdc42/Rho GTPases in fungi: variations on a common theme. Mol Microbiol. 79:1123-1127.

36. Janusz G, Kucharzyk KH, Pawlik A, Staszczak M, Paszczynski AJ. 2013. Fungal laccase, manganese peroxidase and lignin peroxidase: Gene expression and regulation. Enzyme Microb Tech. 52:1-12.

37. Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. Mol Biol Evol. 30:772-780.

38. Kohler A, Francis M, Costa M. 2011. Genomics DNA Extraction. http://1000.fungalgenomes.org/home/wp-content/uploads/2013/02/genomicDNAProtocol-AK0511.pdf. Accessed 12 Dec. 2015.

39. Kolarik M, Freeland E, Utley C, Tisserat N. 2011. *Geosmithia morbida* sp. nov., a new phytopathogenic species living in symbiosis with the walnut twig beetle (*Pityophthorus juglandis*) on *Juglans* in USA. Mycologia.103:325–332.

40. Kolarik M, Jankowiak R. 2013. Vector Affinity and Diversity of *Geosmithia* Fungi Living on Subcortical Insects Inhabiting *Pinaceae* Species in Central and Northeastern Europe. Microb Ecol. 66:682–700.

41. Kolarik M, Kirkendall LR. 2010. Evidence for a new lineage of primary ambrosia fungi in *Geosmithia* Pitt (Ascomycota: *Hypocreales*). Fungal Biol. 114:676–689.

42. Kolarik M, Kostovcik M, Pazoutova S. 2007. Host range and diversity of the genus *Geosmithia* (Ascomycota: *Hypocreales*) living in association with bark beetles in the Mediterranean area. Mycological Res. 111:1298–1310.

43. Kolarik M, Kubatova A, van Cepicka I, Pazoutova S, Srutka P. 2005. A complex of three new white-spored, sympatric, and host range limited *Geosmithia* species. Mycological Res. 109:1323–1336.

44. Kraus PR, Heitman J. 2003. Coping with stress: calmodulin and calcineurin in model and pathogenic fungi. Biochem Bioph Res Co. 311:1151-1157.

45. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 305:567-580.

46. Kubicek CP, Herrera-Estrella A, Seidl-Seiboth V, Martinez DA, Druzhinina IS, et al. 2011. Comparative genome sequence analysis underscores mycoparasitism as the ancestral life style of *Trichoderma*. Genome Biol. 12:R40.

47. Lanfear R, Calcott B, Kainer D, Mayer C, Stamatakis A. 2014. Selecting optimal partitioning schemes for phylogenomic datasets. BMC Evol Biol. 14:82.

48. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 25:1754–1760.

49. Li H. 2015. BFC: correcting Illumina sequencing errors. Bioinformatics. 31:2885-2887.

50. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. 2014. The Carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 42:D490–D495.

51. Lynch SC, Wang DH, Mayorquin JS, Rugman-Jones PF, Stouthamer R, Eskalen E. 2014. First Report of *Geosmithia pallida* Causing Foamy Bark Canker, a new disease on coast live oak (*Quercus agrifolia*), in association with *Pseudopityophthorus pubipennis* in California. Plant Dis. 98:1276.

52. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 464:367-373.

53. Mabey JE, et al. 2004. CADRE: the Central Aspergillus Data REpository. Nucleic Acids Res. 32:D401-405.

54. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, et al. 2008. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat. Biotechnol. 26:553-560.

55. Misof B, Meyer B, von Reumont BM, Kuck P, Misof K, Meusemann K. 2013. Selecting informative subsets of sparse supermatrices increases the chance to find correct trees. BMC Bioinformatics. 14:348.

56. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, et al. 2016. Mash: fast genome and metagenome distance estimation using MinHash. BioRxiv. Available at http://dx.doi.org/10.1101/029827

57. Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 23:1061-1067.

58. Peterson TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785-786.

59. PHI-base. 2015. The pathogen–host interaction database. Available at http://www.phi-base.org/. Accessed 22 Nov. 2015.

60. Pintard L, Willems A, Peter M. 2004. Cullin-based ubiquitin ligases: Cul3-BTB complexes join the family. EMBO J. 23:1681-1687.

61. Pitt JI. 1979. *Geosmithia*, *gen. nov.* for *Penicillium lavendulum* and related species. Can J Botany. 57:2021-2030.

62. Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: **M**ultiple **A**lignment of **C**oding **SE**quences Accounting for Frameshifts and Stop Codons. PLoS One. 6:e22594.

63. Rawlings ND, Barrett AJ, Finn RD. 2016. Twenty years of the *MEROPS* database of proteolytic enzymes, their substrates and inhibitors. Nucleic Acids Res. 44:D343-D350.

64. Rolke Y, Tudzynski P. 2008. The small GTPase Rac and the p21-activated kinase Cla4 in *Claviceps purpurea*: interaction and impact on polarity, development and pathogenicity. Mol Microbiol. 68:405-423.

65. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 61:539–542.

66. Rugman-Jones PF, Seybold SJ, Graves AD, Stouthamer R. 2015. Phylogeography of the walnut twig beetle, *Pityophthorus juglandis*, the vector of thousand cankers disease in North American walnut trees. PLoS ONE. 10:e118264.

67. Ryder LS, Talbot NJ. 2015. Regulation of appressorium development in pathogenic fungi. Curr Opin Plant Biol. 26:8-13.

68. Schuelke TA, Westbrook A, Broders K, Woeste K, MacManes MD. 2016. *De novo* genome assembly of *Geosmithia morbida*, the causal agent of thousand cankers disease. PeerJ. 4:e1952.

69. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 1–3.

70. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: A parallel assembler for short read sequence data. Genome Res. 19:1117-1123.

71. Smit AFA, Hubley R, Green P. 1996. RepeatMasker. Available at http://www.repeatmasker.org.

72. Solomon PS, Oliver RP. 2002. Evidence that γ-aminobutyric acid is a major nitrogen source during *Cladosporium fulvum* infection of tomato. Planta. 214:414-420.

73. Staats M, van Kan JA. 2012. Genome update of *Botrytis cinerea* strains B05.10 and T4. Eukaryot Cell. 11:1413-1414.

74. Stamatakis A. 2014. RAxML Version 8: A tool for Phylogenetic Analysis and Post-analysis of Large Phylogenies. Bioinformatics. 30:1312–1313.

75. Steer ML. 1975. Adenyl cyclase. Ann Surg. 182:603-609.

76. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res. 34:W609-W612.

77. Swiss-Prot. 2015. Available at http://www.uniprot.org/. Downloaded 6 May 2015.

78. Terfehr D, Dahlmann TA, Specht T, Zadra I, Kurnsteiner H, Kuck U. 2014. Genome Sequence and Annotation of *Acremonium chrysogenum*, Producer of the β-Lactam Antibiotic Cephalosporin C. Genome Announc. 2: e00948-14.

79. Tisserat N, Cranshaw W, Leatherman D, Utley C, Alexander K. 2009. Black walnut mortality in colorado caused by the walnut twig beetle and thousand cankers disease. Plant Health Progress. 1–10. DOI 10.1094/PHP-2009-0811-01-RS.

80. Traeger S, et al. 2013. The genome and development-dependent transcriptomes of *Pyronema confluens*: a window into fungal evolution. PLoS Genet. 9:e1003820.

81. Trail F, Xu JR, San Miguel P, Halgren RG, Kistler HC. 2003. Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). Fungal Genet Biol. 38:187-197.

82. Tsai HF, Chang YC, Washburn RG, Wheeler MH, Kwon-Chung KJ. 1988. The developmentally regulated alb1 gene of *Aspergillus fumigatus*: its role in modulation of conidial morphology and virulence. J. Bacteriol. 180:3031-3038.

83. Wang H, Xu Z, Gao L, Hao B. 2009. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol Biol. 9:195.

84. Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 24:1586-1591.

85. Zerillo MM, Caballero JI, Woeste K, Graves AD, Hartel C, et al. 2014. Population structure of *Geosmithia morbida*, the causal agent of thousand cankers disease of walnut trees in the United States. PLoS ONE. 9:e112847.

86. Zhang Y, Zhang K, Fang A, Han Y, Yang J, et al. 2014. Specific adaptation of Ustilaginoidea virens in occupying host florets revealed by comparative and functional genomics. Nat Commun. 5:3849.

87. Zhao Z, Liu H, Wang C, Xu J. 2013. Erratum to: Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 15:6.

88. Zheng P, Xia Y, Xiao G, Xiong C, Hu X, et al. 2011. Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional Chinese medicine. Genome Biol. 12:R116.

Closing remarks

Several studies have investigated the evolution of pathogenicity in pathogens of agricultural crops; forest fungal pathogens are often neglected because research efforts are focused on food crops. Furthermore, studies that do investigate various aspects of a vector-host-pathogen complex concentrate their attention on the geographic and genetic composition and distribution of the fungal pathogen. This research is significant because it explores the processes contributing to the evolution of a novel trait of pathogenicity in *Geosmithia morbida*, which is a beetle-associated pythopathogen and the causal agent of thousand cankers disease (TCD).

G. morbida is one of two known pathogens within the genus *Geosmithia*. Little is known about the precise means this pathogen utilizes to infect its host *Juglans nigra*. This pathogen and its vector, the walnut twig beetle, are of great concern. As of 2015, one or more TCD events have been recorded in several western and eastern states in the US. TCD not only threatens *J. nigra* native stands in eastern US, but also certain western walnut populations such as *J. regia*, *J. californica*, and *J. hindsii*.

Our work here investigates the evolution of pathogenicity within the genus *Geosmithia*. First, we presented the first draft genomes of three *Geosmithia* species—*G. morbida*, *G. flava*, and *G. putterillii*. Our overall findings exhibited that these three species are highly similar and closely relate to *Acremonium chrysogenum* in order Hypocreales. In comparison to other plant pathogens and saprobes, *G. morbida* has a relatively small genome composed of 26.5 Mbp and 6,273 protein models. The ecological restrictions imposed on *G. morbida* could explain this reduction in genome size because this fungus must adapt to its vector and host at the same time.

Although *G. morbida* and *G. clavigera* have similar ecological niches, both fungal pathogens have different set of genes undergoing positive selection. Nonetheless, some of the genes under selection in *G. morbida* have been implicated in the infection process in other pathogens as such *Magnaporthe oryzae* (rice blast fungus). For instance, a positively selected gene in *G. morbida* encodes a bZIP-like transcription factor, which is involved in gene regulation during oxidative stress response in *M. oryzae*.

Acquiring insights into the evolutionary and molecular processes that give rise to novel traits in fungal pathogen is essential for the development of disease control and monitoring techniques. Our results will be instrumental for future studies that are necessary to corroborate our findings with experimental data.