

1 *De novo* Genome Assembly of *Geosmithia morbida*, the Causal Agent of Thousand Cankers Disease

2

3 Taruna Aggarwal,<sup>1\*</sup> Anthony Westbrook,<sup>2</sup> Kirk Broders,<sup>3</sup> Keith Woeste,<sup>4</sup> Matthew D MacManes<sup>1</sup>

4

5 <sup>1</sup>Department of Molecular, Cellular, & Biomedical Sciences, University of New Hampshire, 105 Main  
6 Street, Durham, NH 03824

7 <sup>2</sup>Department of Computer Science, University of New Hampshire, 105 Main Street, Durham, NH  
8 03824

9 <sup>3</sup>Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort  
10 Collins, CO 80523

11 <sup>4</sup>USDA Forest Service Hardwood Tree Improvement and Regeneration Center, Department of  
12 Forestry and Natural Resources, Purdue University, West Lafayette, IN 47907

13 \* Corresponding author

14

15 Email addresses:

16 TA: [ta2007@wildcats.unh.edu](mailto:ta2007@wildcats.unh.edu)

17 AW: [anthonyw@wildcats.unh.edu](mailto:anthonyw@wildcats.unh.edu)

18 KB: [Kirk.Broders@colostate.edu](mailto:Kirk.Broders@colostate.edu)

19 KW: [woeste@purdue.edu](mailto:woeste@purdue.edu)

20 MDM: [Matthew.MacManes@unh.edu](mailto:Matthew.MacManes@unh.edu)

21

22

23

24

25

26

27

28

29 **Abstract**

30 **Background:** *Geosmithia morbida* is a filamentous ascomycete that causes Thousand Cankers  
31 Disease in the eastern black walnut tree. This pathogen is commonly found in the western U.S.;  
32 however, recently the disease was also detected in several eastern states where the black walnut  
33 lumber industry is concentrated. *G. morbida* is one of two known phytopathogens within the genus  
34 *Geosmithia*, and it is vectored into the host tree via the walnut twig beetle.

35 **Results:** We present the first *de novo* draft genome of *G. morbida*. It is 26.5 Mbp in length and  
36 contains less than 1% repetitive elements. The genome possesses an estimated 6,273 genes, 277 of  
37 which are predicted to encode proteins with unknown functions. Approximately 31.5% of the proteins  
38 in *G. morbida* are homologous to proteins involved in pathogenicity, and 5.6% of the proteins contain  
39 signal peptides that indicate these proteins are secreted.

40 **Conclusions:** Several studies have investigated the evolution of pathogenicity in pathogens of  
41 agricultural crops; forest fungal pathogens are often neglected because research efforts are focused  
42 on food crops. *G. morbida* is one of the few tree phytopathogens to be sequenced, assembled and  
43 annotated. The first draft genome of *G. morbida* serves as a valuable tool for comprehending the  
44 underlying molecular and evolutionary mechanisms behind pathogenesis within the *Geosmithia*  
45 genus.

46 **Keywords:** *de novo* genome assembly, pathogenesis, forest pathogen, black walnut, walnut twig  
47 beetle.

48

49 **Introduction**

50 Studying molecular evolution of any phenotype is now made possible by the analysis of large  
51 amounts of sequence data generated by next-generation sequencing platforms. This is particularly  
52 beneficial in the case of emerging fungal pathogens, which are progressively recognized as a threat to  
53 global biodiversity and food security. Furthermore, in many cases their expansion is a result of  
54 anthropogenic activities and an increase in trade of fungal-infected goods [1]. Fungal pathogens  
55 evolve in order to overcome host resistance, fungicides, and to adapt to new hosts and environments.  
56 Whole genome sequence data have been used to identify the mechanisms of adaptive evolution

57 within fungi [2-4]. For instance, Stukenbrock et al. (2011) investigated the patterns of evolution in  
58 fungal pathogens during the process of domestication in wheat using all aligned genes within the  
59 genomes of wheat pathogens. They found that *Mycosphaerella graminicola*, a domesticated wheat  
60 pathogen (now known as *Zymoseptoria tritici*), underwent adaptive evolution at a higher rate than its  
61 wild relatives, *Mycosphaerella* S1 and *Mycosphaerella* S2. The study also revealed that many of the  
62 pathogen's 802 secreted proteins were under positive selection. A study by Gardiner et al. (2012),  
63 identified genes encoding aminotransferases, hydrolases, and kinases that were shared between  
64 *Fusarium pseudograminearum* and other cereal pathogens. Using genomic and phylogenetic  
65 analyses, the researchers demonstrated that these genes had bacterial origins. These studies  
66 highlight the various evolutionary means that fungal species employ in order to adapt to specific  
67 hosts, as well as the important role genomics and bioinformatics play in elucidating evolutionary  
68 mechanisms within the fungal kingdom.

69 Many tree fungal pathogens associate with bark beetles, which belong to the Scolytinae family  
70 [5]. As climate patterns change, both the beetles and their fungal symbionts are able to invade new  
71 territory and become major invasive forest pests on a global scale [6, 7]. A well-known example of an  
72 invasive pest is the mountain pine beetle and its symbiont, *Grosmannia clavigera* that has affected  
73 approximately 3.4 million of acres of lodgepole, ponderosa, and five-needle pine trees in Colorado  
74 alone since the outbreak began in 1996 [8,9]. Another beetle pest in the western U.S., *Pityophthorous*  
75 *judlandis* (walnut twig beetle), associates with several fungal species, including the emergent fungal  
76 pathogen *Geosmithia morbida* [10].

77 Reports of tree mortality triggered by *G. morbida* infections first surfaced in 2009 [12], and the  
78 fungus was described as a new species in 2011 [10]. This fungus is vectored into the host via *P.*  
79 *juglandis* and is the causal agent of thousand cankers disease (TCD) in *Julgans nigra* (eastern black  
80 walnut) [12]. This walnut species is valued for its wood, which is used for furniture, cabinetry, and  
81 veneer. Although *J. nigra* trees are planted throughout western U.S. as a decorative species, they are  
82 indigenous to eastern North America where the walnut industry is worth hundreds of millions of dollars  
83 [13]. In addition to being a major threat to the eastern populations of *J. nigra*, TCD is of great concern

84 because certain western walnut species including *J. regia* (the Persian walnut), *J. californica*, and *J.*  
85 *hindsii* are also susceptible to the fungus according to greenhouse inoculation studies [14].

86 The etiology of TCD is complex because it is a consequence of a fungal-beetle symbiosis. The  
87 walnut twig beetle, which is only known to attack members of genus *Juglans* and *Pterocarya*, is the  
88 most common vector of *G. morbida* [10]. Nevertheless, other beetles are able to disperse the fungus  
89 from infested trees [15, 16]. As vast numbers of beetles concentrate in the bark of infested trees,  
90 fungal cankers form and coalesce around beetle galleries and entrance holes. As the infection  
91 progresses, the phloem and cambium discolor and the leaves wilt and yellow. These symptoms are  
92 followed by branch dieback and eventual tree death, which can occur within three years of the initial  
93 infection [10]. Currently, 15 states in the U.S. have reported one or more incidences of TCD, reflecting  
94 the expansion of WTB's geographic range from its presumed native range in a few southwestern  
95 states [17].

96 To date, *G. morbida* is one of only two known pathogens within the genus *Geosmithia*, which  
97 consists of mostly saprotrophic beetle-associated species (the other pathogen is *G. pallida*) [18]. The  
98 ecological complexity this vector-host-pathogen complex exhibits makes it an intriguing lens for  
99 studying the evolution of pathogenicity within the fungal kingdom. A well-assembled reference  
100 genome will enable us to identify genes unique to *G. morbida* that may be utilized to develop  
101 sequence-based tools for detecting and monitoring epidemics of TCD and for studying the evolution of  
102 pathogenesis within the *Geosmithia* genus. Here, we present a *de novo* genome assembly of  
103 *Geosmithia morbida*. The objectives of this study are to: 1) assemble the first, high-quality draft  
104 genome of this pathogen; 2) annotate the genome in order to better comprehend the evolution of  
105 pathogenicity in the *Geosmithia* genus; and 3) briefly compare the genome of *G. morbida* to two other  
106 fungal pathogens for which genomic data is available: *Fusarium solani*, a root pathogen that infects  
107 soybean, and *Grosmannia calvigera*, a pathogenic ascomycete that associates with the mountain pine  
108 beetle and kills lodgepole pines in North America.

109

## 110 **Methods**

### 111 **DNA extraction and Library Preparation**

112 DNA was extracted using the CTAB method as outlined by the Joint Genome Institute to extract DNA  
113 for Genome Sequencing from lyophilized mycelium of *G. morbida* (isolate 1262, host: *Juglans*  
114 *californica*) from southwestern California [19]. The total DNA concentration was measured using  
115 Nanodrop, and samples for sequencing were sent to Purdue University Genomics Core Facility in  
116 West Lafayette, Indiana. DNA libraries were prepared using the paired-end Illumina Truseq protocol  
117 and mate-pair Nextera DNA Sample Preparation kits with average insert sizes of 487bp and 1921bp  
118 respectively. These libraries were sequenced on the Illumina HiSeq 2500.

### 119 **Preprocessing Sequence Data**

120 We began by performing quality control checks on our raw sequence data generated by the Illumina  
121 platform. To assess the quality of our data, we ran FastQC (v0.11.2) (<https://goo.gl/xHM1zf>) [20] and  
122 SGA Preqc (v0.10.13) (<https://goo.gl/9y5bNy>) on our raw sequence reads [21]. Both tools aim to  
123 supply the user with information such as per base sequence quality score distribution (FastQC) and  
124 frequency of variant branches in *de Bruijn* graphs (Preqc) that aid in selecting appropriate assembly  
125 tools and parameters. The paired-end raw reads were corrected using a Bloom filter-based error  
126 correction tool called BLESS (v0.16) (<https://goo.gl/Kno6Xo>) [22]. Next, the error corrected reads were  
127 trimmed with Trimmomatic, version 0.32, using a Phred threshold of 2, following recommendations  
128 from MacManes (2014) (<https://goo.gl/FFoFjL>) [23]. NextClip, version 1.3.1, was leveraged to trim  
129 adapters in the mate-pair read set (<https://goo.gl/aZ9ucT>) [24]. The raw reads are available at  
130 <https://goo.gl/IMsMe5>.

### 131 **De novo genome assembly and evaluation**

132 The *de novo* genome assembly was constructed with ALLPaths-LG (v49414) (<https://goo.gl/03gU9Z>)  
133 [25]. The assembly was evaluated with BUSCO (v1.1b1) (<https://goo.gl/bMrXIM>), a tool that assesses  
134 genome completeness based on the presence of single-copy orthologs [26]. We also generated  
135 length-based statistics for our *de novo* genome with QUAST (v2.3) (<https://goo.gl/5KSa4M>) [27]. The  
136 raw reads were mapped back to the genome using BWA version 0.7.9a-r786 to further assess the  
137 quality of the assembly (<https://goo.gl/Scxgn4>) [28].

### 138 **Structural and Functional Annotation of *G. morbida* genome**

139 We used the automated genome annotation software Maker version 2.31.8. Maker identifies repetitive  
140 elements, aligns ESTs, and uses protein homology evidence to generate *ab initio* gene predictions  
141 (<https://goo.gl/JiLA3H>) [29]. We used two of the three gene prediction tools available within the  
142 pipeline, SNAP and Augustus. SNAP was trained using gff files generated by CEGMA v2.5 (a  
143 program similar to BUSCO). Augustus was trained with *Fusarium solani* protein models (v2.0.26)  
144 downloaded from Ensembl Fungi [30, 34]. In order to functionally annotate the genome, the protein  
145 sequences produced by the structural annotation were blasted against the Swiss-Prot database, and  
146 target sequences were filtered for the best hits [31]. A small subset of the resulting annotations was  
147 visualized and manually curated in WebApollo v2.0.1 [32]. The final annotations were also evaluated  
148 with BUSCO (v1.1b1) (<https://goo.gl/thTGzH>).

#### 149 **Assessing Repetitive Elements Profile**

150 To assess the repetitive elements profile of *G. morbida*, we masked only the interspersed repeats  
151 within the assembled scaffolds with RepeatMasker (v4.0.5) (<https://goo.gl/TXrbr3>) [33] using the  
152 sensitive mode and default values as arguments. In order to compare the repetitive element profile of  
153 *G. morbida* with *F. solani* (v2.0.29) and *G. clavigera* (kw1407.GCA\_000143105.2.30), the  
154 interspersed repeats of these two fungal pathogens were also masked with RepeatMasker. The  
155 genome and protein data of these fungi were downloaded from Ensembl Fungi [34].

#### 156 **Identifying putative proteins contributing to pathogenicity**

157 To identify putative genes contributing to pathogenicity in *G. morbida*, a BLASTp search was  
158 conducted for single best hits at an e-value threshold of 1e-6 or less against the PHI-base database  
159 (v3.8) (<https://goo.gl/CEEVY0>) that contains experimentally confirmed genes from fungal, oomycete  
160 and bacterial pathogens [35]. The search was performed using the same parameters for *F. solani* and  
161 *G. clavigera*. To identify the proteins that contain signal peptides, we used SignalP (v4.1)  
162 (<https://goo.gl/JOe5Dh>), and compared results from *G. morbida* with those from *F. solani* and *G.*  
163 *clavigera* [36]. Lastly, to find putative protein domains involved in pathogenicity in *G. morbida*, we  
164 performed a HMMER (version 3.1b2) [37] search against the Pfam database (v28.0) [38] using the  
165 protein sequences as query. We conducted the same search for sequences of 17 known effector

166 proteins, then extracted and analyzed domains common between the effector sequences and *G.*

167 *morbida* (<https://goo.gl/Y9IPZs>).

168

## 169 **Results and Discussion**

### 170 **Data Processing**

171 A total of 28,027,726 PE and 41,348,578 MP forward and reverse reads were generated with  
172 approximately 56x and 83x coverage respectively (Table 1). Of the MP reads, 67.7% contained  
173 adapters that were trimmed using NextClip (v1.3.1). We corrected errors within the PE reads using  
174 BLESS (v0.16) at a kmer length of 21. After correction, low-quality reads (phred score < 2) were  
175 trimmed with Trimmomatic (v0.32) resulting in 99.75% reads passing. In total, 16,336,158 MP and  
176 27,957,268 PE reads were used to construct the *de novo* genome assembly.

177

178 Table 1. Statistics for *Geosmithia morbida* sequence data.

	Paired-end		Mate-pair	
Number of reads	28,027,726	<b>27,957,268</b>	41,348,578	<b>16,336,158</b>
Average insert size (bp)		487		1921
Average coverage		56x		83x

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

181

### 182 **Assembly Features**

183 The *G. morbida de novo* assembly (available at <https://goo.gl/6P8zmY>) was constructed with  
184 AllPaths-LG (v49414). The assembled genome consisted of 73 contigs totaling 26,549,069 bp. The  
185 largest contig length was 2,597,956 bp, and the NG50 was 1,305,468 bp. The completeness of the  
186 genome assembly was assessed using BUSCO, a tool that scans the genome for the presence of  
187 single-copy orthologous groups present in more than 90% of fungal species. Of 1,438 single-copy  
188 orthologs specific to fungi, 95% were complete in our assembly, and 3.6% were fragmented BUSCOs.  
189 Only 0.8% of the orthologs were missing from the genome (Table 2). We used BWA to map the

190 unprocessed, raw MP and PE reads back to the genome to further evaluate the assembly, and 87%  
191 of the MP and 90% of the PE reads mapped to our reference genome.

192

193 Table 2. *Geosmithia morbida* reference genome assembly statistics generated using QUAST (v2.3)

	Scaffolds
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%

194

## 195 Gene annotation

196 The automated genome annotation software Maker v2.31.8 was used to identify structural elements in  
197 the *G. morbida* assembly generated by AllPaths-LG. Of the total 6,273 proteins that were predicted,  
198 5,996 returned with hits against the Swiss-Prot database—only 277 (4.41%) of the total genes  
199 encoded for proteins of unknown function. The completeness of the functional annotations was  
200 evaluated using BUSCO, and 94% of the single copy orthologs were present in this protein set. The  
201 transcript and protein files are available at <https://goo.gl/svTmKp> and <https://goo.gl/pB9y5l>.

## 202 Repetitive Elements

203 Repetitive elements represented 0.81% of the total bases in the *G. morbida* genome (available at  
204 <https://goo.gl/wDq2xP>). The genome contained 152 retroelements (class I) that were mostly  
205 composed of long terminal repeats (n=146) and 60 DNA transposons (class II). In comparison, the  
206 genomes of *G. clavigera* and *F. solani* contained 1.14% and 1.47% respectively (available at  
207 <https://goo.gl/8zXAIH> and <https://goo.gl/YQAM2N>). *G. clavigera* possesses 541 retroelements  
208 (0.79%) and 66 DNA transposons (0.04%), whereas the genome of *F. solani* is comprised of 499  
209 (0.54%) and 515 (0.81%) retroelements and transposons respectively. The larger number of repeat



210 elements in *F. solani* may explain its relatively large genome size —51.3 Mbp versus *G. clavigera*'s  
211 29.8 Mbp and *G. morbida*'s 26.5 Mbp (Table 3).

212

213 Table 3. Repetitive elements profile for *Geosmithia morbida*, *Grosmannia clavigera* and *Fusarium*  
214 *solani*.

	<i>G. morbida</i>	<i>G. clavigera</i>	<i>F. solani</i>
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%

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

217

### 218 **Identifying and classifying putative pathogenicity genes**

219 We blasted the entire predicted protein set against the PHI-base database (v3.8) to identify a list of  
220 putative genes that may contribute to pathogenicity within *G. morbida*, *F. solani*, and *G. clavigera*. We  
221 determined that 1,974 genes in *G. morbida* (31.47% of the total 6,273 genes) were homologous to  
222 protein sequences in the database (available at <https://goo.gl/SZA4Kd>). For *F. solani* and *G.*  
223 *clavigera*, there were 4,855 and 2,387 genes with homologous PHI-base proteins (available at  
224 <https://goo.gl/Rm8Zx7> and <https://goo.gl/fjrrvm>).

### 225 **Identifying putative secreted proteins**

226 A search for the presence of secreted peptides within the protein sequences of *G. morbida*, *F. solani*  
227 and *G. clavigera* showed that approximately 5.6% (349) of the *G. morbida* protein sequences  
228 contained putative signal peptides (available at <https://goo.gl/Qz8gUr>). Of the 349 sequences  
229 containing putative signal peptides, only 27 encoded proteins of unknown function. Roughly 8.8% and  
230 6.9% of the proteins of *F. solani* and *G. clavigera* possess signal peptides (available at  
231 <https://goo.gl/mTu7Ok> and <https://goo.gl/PZdSNc>). Secreted proteins are essential for host-fungal

232 interactions and are indicative of adaptation within fungal pathogens that require an array of  
233 mechanisms to overcome plant host defenses.

### 234 **Identifying protein domains**

235 We conducted a HMMER search against the pfam database (v28.0) using amino acid sequences for  
236 *G. morbida* and 17 effector proteins from various fungal species. For *G. morbida*, there were 6,023  
237 unique protein domains out of a total of 43,823 Pfam hits. A total of 17 domains, which comprised  
238 1,000 hits, were shared between *G. morbida* and known effector proteins. The three most common  
239 protein domains in *G. morbida* with a putative effector function belonged to short-chain  
240 dehydrogenases (n=111), polyketide synthases (n=94) and NADH dehydrogenases (n=86). The  
241 HMMER *G. morbida* and effector proteins output files are located at <https://goo.gl/r8B7uk> and  
242 <https://goo.gl/mkn5aB> respectively.

243

### 244 **Conclusion**

245 This work introduces the first genome assembly and analysis of *Geosmithia morbida*, a fungal  
246 pathogen of the black walnut tree that is vectored into the host via the walnut twig beetle. The *de novo*  
247 assembly is composed of 73 scaffolds totaling in 26.5 Mbp. There are 6,273 predicted proteins, and  
248 4.41% of these are unknown. In comparison, 68.27% of *F. solani* and 26.70% of *G. clavigera*  
249 predicted proteins are unknown. We assessed the quality of our genome assembly and the predicted  
250 protein set using BUSCO, and found that 95% and 94% of the single copy orthologs specific to the  
251 fungal lineage were present in both respectively. These data are indicative of our assembly's high  
252 quality and completeness. Our BLASTp search against the PHI-base database revealed that *G.*  
253 *morbida* possesses 1,974 genes that are homologous to proteins involved in pathogenicity.  
254 Furthermore, *G. morbida* shares several domains with known effector proteins that are key for fungal  
255 pathogens during the infection process.

256 *Geosmithia morbida* is one of only two known fungal pathogens within the *Geosmithia* genus  
257 [18]. The genome assembly introduced in this study can be leveraged to explore the molecular  
258 mechanisms behind pathogenesis within this genus. The putative list of pathogenicity genes provided  
259 in this study can be used for future comparative genomic analyses, knock-out, and inoculation

260 experiments. Moreover, genes unique to *G. morbida* may be utilized to develop DNA sequence-based  
261 tools for detecting and monitoring ongoing and future TCD epidemics.

262

### 263 **Competing interests**

264 The authors declare no competing interests.

265

### 266 **Authors' contributions**

267 TA extracted DNA, prepared samples for sequencing, wrote the manuscript. TA and MDM assembled  
268 and evaluated the genome. AW annotated the genome. KB and KW helped conceive and fund the  
269 project and assisted in manuscript editing. MDM and KB contributed to the writing of the manuscript.

270

### 271 **Acknowledgements**

272 Partial funding was provided by the New Hampshire Agricultural Experiment Station. This is Scientific  
273 Contribution Number 2652. Funding was also provided by the USDA Forest Service, Forest Health  
274 and Protection. Mention of a trademark, proprietary product, or vendor does not constitute a  
275 guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its  
276 approval to the exclusion of other products or vendors that also may be suitable.

277

278

279

280

281

282

283

284

285

286

287

288 **References**

- 289 1. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, et al. Emerging fungal  
290 threats to animal, plant and ecosystem health. *Nature*. 2012;484:186-194.
- 291 2. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, et al. The making of a new  
292 pathogen: Insights from comparative population genomics of the domesticated wheat pathogen  
293 *Mycosphaerella graminicola* and its wild sister species. *Genome Res*. 2011;21:2157–2166.
- 294 3. Gardiner DM, McDonald MC, Covarelli L, Solomon PS, Rusu AG, Marshall M, et al.  
295 Comparative Pathogenomics Reveals Horizontally Acquired Novel Virulence Genes in Fungi  
296 Infecting Cereal Hosts. *PLoS Pathog*. 2012;8:e1002952–22.
- 297 4. Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otililar R, et al. Comparative Genome  
298 Structure, Secondary Metabolite, and Effector Coding Capacity across *Cochliobolus*  
299 Pathogens. *PLoS Genet*. 2013;9:e1003233.
- 300 5. Six DL, Wingfield MJ. The role of phytopathogenicity in bark beetle-fungus symbioses: A  
301 challenge to the classic paradigm. *Annu Rev Entomol*. 2011;56:255-272.
- 302 6. Kurz WA, Dymond CC, Stinson G, Rampley GJ, Neilson ET, Carroll AL, et al. Mountain pine  
303 beetle and forest carbon feedback to climate change. *Nature*. 2008;452:987–990.
- 304 7. Sambaraju KR, Carroll AL, Zhu J, Stahl K, Moore RD, Aukema BH. Climate change could alter  
305 the distribution of mountain pine beetle outbreaks in western Canada. *Ecography*.  
306 2012;35:211–223.
- 307 8. Massoumi Alamouti S, Haridas S, Feau N, Robertson G, Bohlmann J, Breuil C. Comparative  
308 Genomics of the Pine Pathogens and Beetle Symbionts in the Genus *Grosmannia*. *Mol Biol*  
309 *Evol*. 2014;31:1454–1474.
- 310 9. Mountain Pine Beetle: [http://csfs.colostate.edu/forest-management/common-forest-insects-](http://csfs.colostate.edu/forest-management/common-forest-insects-diseases/mountain-pine-beetle/)  
311 [diseases/mountain-pine-beetle/](http://csfs.colostate.edu/forest-management/common-forest-insects-diseases/mountain-pine-beetle/). Accessed 15 April 2015.
- 312 10. Kolarik M, Freeland E, Utley C, Tisserat N. *Geosmithia morbida* sp. nov., a new  
313 phytopathogenic species living in symbiosis with the walnut twig beetle (*Pityophthorus*  
314 *juglandis*) on *Juglans* in USA. *Mycologia*. 2011;103:325–332.
- 315 11. Tisserat N, Cranshaw W, Leatherman D, Utley C, Alexander K. Black Walnut Mortality in

- 316 Colorado Caused by the Walnut Twig Beetle and Thousand Cankers Disease. PHP. 2009:1–  
317 10. doi:10.1094/PHP-2009-0811-01-RS
- 318 12. Zerillo MM, Caballero JI, Woeste K, Graves AD, Hartel C, Pscheidt JW, et al. Population  
319 Structure of *Geosmithia morbida*, the Causal Agent of Thousand Cankers Disease of Walnut  
320 Trees in the United States. PLoS ONE. 2014;9(11):e112847.
- 321 13. Newton L, Fowler G. Pathway Assessment: *Geosmithia* sp. and *Pityophthorus juglandis*  
322 Blackman movement from the western into the eastern United States. US Dept. of Agriculture  
323 Animal and Plant Health Inspection Service. 2009. Accessed 22 Dec 2014.
- 324 14. Utley C, Nguyen T, Roubtsova T, Coggeshall M, Ford TM, Grauke LJ et al. Susceptibility of  
325 Walnut and Hickory Species to *Geosmithia morbida*. Plant Disease. 2013;97(5):601–607.
- 326 15. Kolařík M, Kostovčík M, Pažoutová S. Host range and diversity of the genus *Geosmithia*  
327 (Ascomycota: Hypocreales) living in association with bark beetles in the Mediterranean area.  
328 Mycological Res. 2007;111:1298–1310.
- 329 16. Kolařík M, Jankowiak R. Vector Affinity and Diversity of *Geosmithia* Fungi Living on Subcortical  
330 Insects Inhabiting *Pinaceae* Species in Central and Northeastern Europe. Microb Ecol.  
331 2013;66:682–700.
- 332 17. Rugman-Jones PF, Seybold SJ, Graves AD, Stouthamer R. Phylogeography of the Walnut  
333 Twig Beetle, *Pityophthorus juglandis*, the Vector of Thousand Cankers Disease in North  
334 American Walnut Trees. PLoS ONE. 2015;10:e0118264.
- 335 18. Lynch SC, Wang DH, Mayorquin JS, Rugman-Jones PF, Stouthamer R, Eskalen E. First  
336 Report of *Geosmithia pallida* Causing Foamy Bark Canker, a New Disease on Coast Live Oak  
337 (*Quercus agrifolia*), in Association with *Pseudopityophthorus pubipennis* in California. 2014;  
338 98:1276.
- 339 19. Kohler A, Francis M. Genomic DNA Extraction. [http://1000.fungalgenomes.org/home/wp-](http://1000.fungalgenomes.org/home/wp-content/uploads/2013/02/genomicDNAProtocol-AK0511.pdf)  
340 [content/uploads/2013/02/genomicDNAProtocol-AK0511.pdf](http://1000.fungalgenomes.org/home/wp-content/uploads/2013/02/genomicDNAProtocol-AK0511.pdf). Accessed 12 Dec 2015.
- 341 20. Andrews S. FastQC. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>. Accessed 12  
342 Dec 2015.
- 343 21. Simpson JT. Exploring genome characteristics and sequence quality without a reference.

- 344 arXiv. 2013; <http://arxiv.org/abs/1307.8026>.
- 345 22. Heo Y, Wu X-L, Chen D, Ma J, Hwu W-M. BLESS: bloom filter-based error correction solution  
346 for high-throughput sequencing reads. *Bioinformatics*. 2014;30:1354–1362.
- 347 23. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data.  
348 *Bioinformatics*. 2014;30:2114–2120.
- 349 24. Leggett RM, Clavijo BJ, Clissold L, Clark MD, Caccamo M. NextClip: an analysis and read  
350 preparation tool for Nextera long mate pair libraries. *Bioinformatics*. 2014;30(4):566-568.
- 351 25. Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, et al. High-quality draft  
352 assemblies of mammalian genomes from massively parallel sequence data. *PNAS*.  
353 2011;108:1513–1518.
- 354 26. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing  
355 genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*.  
356 2015:1–3 (2015).
- 357 27. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUASt: quality assessment tool for genome  
358 assemblies. *Bioinformatics*. 2013;29:1072–1075.
- 359 28. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.  
360 *Bioinformatics*. 2009;25:1754–1760.
- 361 29. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B. MAKER: An easy-to-use annotation  
362 pipeline designed for emerging model organism genomes. *Genome Res*. 2008;18:188-196.
- 363 30. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in  
364 eukaryotic genomes. *Bioinformatics*. 2007;23:1061-1067.
- 365 31. Swiss-Prot: (<http://www.uniprot.org/>). Downloaded 6 May 2015.
- 366 32. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a  
367 web-based genomic annotation editing platform. *Genome Biol*. 2013;4:R93.
- 368 33. Smit AFA, Hubley R, Green P. RepeatMasker: (<http://www.repeatmasker.org>).
- 369 34. EnsemblFungi: (<http://fungi.ensembl.org/index.html>). Accessed 14 Nov 2015.
- 370 35. PHI-base: The Pathogen - Host Interaction Database: (<http://www.phi-base.org/>). Accessed 22  
371 Nov 2015.

- 372 36. Peterson TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides  
373 from transmembrane regions. *Nature Methods*. 2011;8:785-786.
- 374 37. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching.  
375 *Nucl Acids Res*. 2011;1-9.
- 376 38. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein  
377 families database. *Nucl Acids Res*. 2014;42:D222-D230.