

2016

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

Taruna A. Schuelke
University of New Hampshire, Durham

Anthony Westbrook
University of New Hampshire, Durham

Kirk Broders
Colorado State University

Keith E. Woeste
Purdue University, woeste@purdue.edu

Matthew D. MacManes
University of New Hampshire, Durham

Follow this and additional works at: <https://docs.lib.purdue.edu/fnrpubs>

Recommended Citation

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 <https://doi.org/10.7717/peerj.1952>

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

Taruna A. Schuelke¹, Anthony Westbrook², Kirk Broders³, Keith Woeste⁴ and Matthew D. MacManes¹

¹ Department of Molecular, Cellular, & Biomedical Sciences, University of New Hampshire, Durham, New Hampshire, United States

² Department of Computer Science, University of New Hampshire, Durham, New Hampshire, United States

³ Department of Bioagricultural Sciences and Pest Management, Colorado State University, Fort Collins, Colorado, United States

⁴ Hardwood Tree Improvement and Regeneration Center, USDA Forest Service, West Lafayette, Indiana, United States

ABSTRACT

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 U.S.; 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*. 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. 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. *G. morbida* is one of the few tree phytopathogens to be sequenced, assembled and annotated. The first draft genome of *G. morbida* serves as a valuable tool for comprehending the underlying molecular and evolutionary mechanisms behind pathogenesis within the *Geosmithia* genus.

Submitted 21 January 2016

Accepted 29 March 2016

Published 2 May 2016

Corresponding author

Kirk Broders,

kirk.broders@colostate.edu

Academic editor

Abhishek Kumar

Additional Information and
Declarations can be found on
page 8

DOI 10.7717/peerj.1952

© Copyright

2016 Schuelke et al.

Distributed under

Creative Commons CC-BY 4.0

OPEN ACCESS

Subjects Genomics, Mycology

Keywords Pathogenesis, Black walnut, Forest pathogen, Walnut twig beetle, *De novo* genome assembly, *Geosmithia morbida*

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 U.S., *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 *Juglans 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 U.S. 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

(Kolařík, Kostovčík & Pažoutová, 2007; Kolařík, & 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 U.S. 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.

METHODS

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.

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, Lohse & Usadel, 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*).

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 (*Simão 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*).

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, Bradnam & Korf, 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>).

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, Hubley & Green, 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*).

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, Clements & Eddy, 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>).

RESULTS AND DISCUSSION

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 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.

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). 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.

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.

Table 1 Statistics for *Geosmithia morbida* sequence data. The values in bold are final number of reads used for assembly after quality check.

	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	

Table 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	98%

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 3).

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 (Table S1). For *F. solani* and *G. clavigera*, there were 4,855 and 2,387 genes with homologous PHI-base proteins (Tables S2 and S3).

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 (Table S4). 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 (Tables S5 and S6). Secreted proteins are essential for host-fungal interactions and are indicative of adaptation within fungal pathogens that require an array of mechanisms

Table 3 Repetitive elements profile for *Geosmithia morbida*, *Grosmannia clavigera*, and *Fusarium solani*. RepeatMasker (v4.0.5) was used to generate these values. Genomic data for *F. solani* and *G. clavigera* were downloaded from Ensembl Fungi.

	<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%

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*).

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). The HMMER *G. morbida* and effector proteins output files can be found in [Tables S7](#) and [S8](#).

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.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

Partial funding was provided by the New Hampshire Agricultural Experiment Station. Funding was also provided by the USDA Forest Service, Forest Health and Protection. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests

The authors declare no competing interests. Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

Author Contributions

- Taruna A. Schuelke conceived and designed the experiments, performed the experiments, analyzed the data, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.
- Anthony Westbrook analyzed the data, reviewed drafts of the paper.
- Kirk Broders conceived and designed the experiments, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper, conceived funding.
- Keith Woeste conceived and designed the experiments, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper, conceived funding.
- Matthew D. MacManes conceived and designed the experiments, analyzed the data, reviewed drafts of the paper.

DNA Deposition

The following information was supplied regarding the deposition of DNA sequences:

The pep files for *F. solani* (v2.0.29) and *G. clavigera* (kw1407.GCA_000143105.2.30) were downloaded from FungalEnsembl.

Data Availability

The following information was supplied regarding data availability:

The raw reads and assembled sequences reported in this manuscript are available at European Nucleotide Archive under Project Number PRJEB13066. The in silico generated transcript and protein files are located at Dryad ([doi:10.5061/dryad.d18mc](https://doi.org/10.5061/dryad.d18mc)). The code is available at https://github.com/macmanes-lab/Geosmithia_manuscript.

Supplemental Information

Supplemental information for this article can be found online at <http://dx.doi.org/10.7717/peerj.1952#supplemental-information>.

REFERENCES

- Andrews S. 2015. FastQC. Cambridge: Babraham Institute. Available at <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/> (accessed 12 December 2015).
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114–2120 DOI 10.1093/bioinformatics/btu170.
- 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 Research* 18(1):188–196 DOI 10.1101/gr.6743907.
- Colorado State Forest Service. 2015. Mountain Pine Beetle. Fort Collins: Colorado State University. Available at <http://csfs.colostate.edu/forest-management/common-forest-insect-diseases/mountain-pine-beetle/> (accessed 15 April 2015).
- Condon BJ, Leng Y, Wu D, Bushley KE, Ohm RA, Otililar R, Martin J, Schackwitz W, Grimwood J, MohdZainudin N, Xue C, Wang R, Manning VA, Dhillon B, Tu ZJ, Steffenson BJ, Salamov A, Sun H, Lowry S, LaButti K, Han J, Copeland A, Lindquist E, Barry K, Schmutz J, Baker SE, Ciuffetti LM, Grigoriev IV, Zhong S, Turgeon BG. 2013. Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* Pathogens. *PLoS Genetics* 9(1):e1003233 DOI 10.1371/journal.pgen.1003233.
- de Jonge R, van Esse HP, Kombrink A, Shinya T, Desaki Y, Bours R, van der Krol S, Shibuya N, Joosten MHAJ, Thomma BPHJ. 2010. Conserved fungal lysm effector Ecp6 prevents chitin-triggered immunity in plants. *Science* 329(5994):953–955 DOI 10.1126/science.1190859.
- de Jonge R, van Esse HP, Maruthachalam K, Bolton MD, Santhanam P, Saber MK, Zhang Z, Usami T, Lievens B, Subbarao KV, Thomma BPHJ. 2012. Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome RNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America* 109(13):5110–5115 DOI 10.1073/pnas.1119623109.
- EnsemblFungi. 2015. Available at <http://fungi.ensembl.org/index.html> (accessed 14 November 2015).
- Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. 2014. Pfam: the protein families database. *Nucleic Acids Research* 42(D1):D222–D230 DOI 10.1093/nar/gkt1223.
- Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* 39(Suppl. 2):W29–W37 DOI 10.1093/nar/gkr367.
- Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ. 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484(7393):186–194 DOI 10.1038/nature10947.
- Gardiner DM, McDonald MC, Covarelli L, Solomon PS, Rusu AG, Marshall M, Kazan K, Chakraborty S, McDonald BA, Manners JM. 2012. Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathogens* 8(9):e1002952. DOI 10.1371/journal.ppat.1002952.
- Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB. 2011. High-quality draft

- assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences of the United States of America* **108**(4):1513–1518 DOI [10.1073/pnas.1017351108](https://doi.org/10.1073/pnas.1017351108).
- Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013.** QUASt: quality assessment tool for genome assemblies. *Bioinformatics* **29**(8):1072–1075 DOI [10.1093/bioinformatics/btt086](https://doi.org/10.1093/bioinformatics/btt086).
- 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**(10):1354–1362 DOI [10.1093/bioinformatics/btu030](https://doi.org/10.1093/bioinformatics/btu030).
- 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 December 2015).
- 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**(2):325–332 DOI [10.3852/10-124](https://doi.org/10.3852/10-124).
- Kolařík M, Jankowiak R. 2013.** Vector affinity and diversity of *Geosmithia* fungi living on subcortical insects inhabiting *Pinaceae* species in Central and Northeastern Europe. *Microbial Ecology* **66**(3):682–700 DOI [10.1007/s00248-013-0228-x](https://doi.org/10.1007/s00248-013-0228-x).
- Kolařík M, Kostovčík M, Pažoutová S. 2007.** Host range and diversity of the genus *Geosmithia* (Ascomycota: Hypocreales) living in association with bark beetles in the Mediterranean area. *Mycological Research* **111**(11):1298–1310 DOI [10.1016/j.mycres.2007.06.010](https://doi.org/10.1016/j.mycres.2007.06.010).
- Kurz WA, Dymond CC, Stinson G, Rampley GJ, Neilson ET, Carroll AL, Ebata T, Safranyik L. 2008.** Mountain pine beetle and forest carbon feedback to climate change. *Nature* **452**(7190):987–990 DOI [10.1038/nature06777](https://doi.org/10.1038/nature06777).
- Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. 2013.** Web Apollo: a web-based genomic annotation editing platform. *Genome Biology* **14**:R93 DOI [10.1186/gb-2013-14-8-r93](https://doi.org/10.1186/gb-2013-14-8-r93).
- 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**(4):566–568 DOI [10.1093/bioinformatics/btt702](https://doi.org/10.1093/bioinformatics/btt702).
- Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14):1754–1760 DOI [10.1093/bioinformatics/btp324](https://doi.org/10.1093/bioinformatics/btp324).
- 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 Disease* **98**(9):1276 DOI [10.1094/PDIS-03-14-0273-PDN](https://doi.org/10.1094/PDIS-03-14-0273-PDN).
- MacManes MD. 2014.** On the optimal trimming of high-throughput mRNA sequence data. Available at <http://dx.doi.org/10.1101/000422>.
- 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*. *Molecular Biology and Evolution* **31**(6):1454–1474 DOI [10.1093/molbev/msu102](https://doi.org/10.1093/molbev/msu102).
- Mohanta TK, Bae H. 2015.** The diversity of fungal genome. *Biological Procedures Online* **17**(8):1–9 DOI [10.1186/s12575-015-0020-z](https://doi.org/10.1186/s12575-015-0020-z).
- 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 Disease* **98**(5):696 DOI [10.1094/PDIS-10-13-1027-PDN](https://doi.org/10.1094/PDIS-10-13-1027-PDN).
- Parra G, Bradnam K, Korf I. 2007.** CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23**(9):1061–1067 DOI [10.1093/bioinformatics/btm071](https://doi.org/10.1093/bioinformatics/btm071).

- Peterson TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8:785–786 DOI 10.1038/nmeth.1701.
- Petre B, Kamoun S. 2014. How do filamentous pathogens deliver effector proteins into plant cells? *PLoS Biology* 12(2):1–7 DOI 10.1371/journal.pbio.1001801.
- PHI-base. 2015. The pathogen–host interaction database. Available at <http://www.phi-base.org/> (accessed 22 November 2015).
- 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(2):e118264 DOI 10.1371/journal.pone.0118264.
- 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(3):211–223 DOI 10.1111/ecog.2012.35.issue-3.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19):3210–3212 DOI 10.1093/bioinformatics/btv351.
- Simpson JT. 2013. Exploring genome characteristics and sequence quality without a reference. Available at <http://arxiv.org/abs/1307.8026>.
- Six DL, Wingfield MJ. 2011. The role of phytopathogenicity in bark beetle–fungus symbioses: a challenge to the classic paradigm. *Annual Review of Entomology* 56(1):255–272 DOI 10.1146/annurev-ento-120709-144839.
- Smit AFA, Hubley R, Green P. 1996. RepeatMasker. Available at <http://www.repeatmasker.org>.
- Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, McDonald BA, Wang J, Schierup MH. 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 Research* 21(12):2157–2166 DOI 10.1101/gr.118851.110.
- 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(6):1397–1407 DOI 10.3852/11-374.
- Swiss-Prot. 2015. Available at <http://www.uniprot.org/>. Downloaded 6 May 2015.
- 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.
- Utley C, Nguyen T, Roubtsova T, Coggeshall M, Ford TM, Grauke LJ, Graves AD, Leslie CA, McKenna J, Woeste K, Yaghmour MA, Seybold SJ, Bostock RM, Tisserat N. 2013. Susceptibility of walnut and hickory species to *Geosmithia morbida*. *Plant Disease* 97(5):601–607 DOI 10.1094/PDIS-07-12-0636-RE.
- Zerillo MM, Caballero JI, Woeste K, Graves AD, Hartel C, Pscheidt JW, Tonos J, Broders K, Cranshaw W, Seybold SJ, Tisserat N. 2014. Population structure of *Geosmithia morbida*, the causal agent of thousand cankers disease of walnut trees in the United States. *PLoS ONE* 9(11): e112847 DOI 10.1371/journal.pone.0112847.