Genomic, transcriptomic, and metagenomic approaches for detecting fungal plant pathogens and investigating the molecular basis of fungal ice nucleation activity

Shu Yang

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Plant Pathology, Physiology, and Weed Science

Boris Vinatzer, Chair
Brian Badgley
Song Li
David Schmale III

December 15, 2021
Blacksburg, VA

Keywords: fungi, disease diagnosis, metagenomics, ice nucleation activity, comparative genomics, comparative transcriptomics

Genomic, transcriptomic, and metagenomic approaches for detecting fungal plant pathogens and investigating the molecular basis of fungal ice nucleation activity

Shu Yang

ABSTRACT

Fungi play important roles in various environments. Some of them infect plants and cause economically important diseases. However, many fungal pathogens cause similar symptoms or are even spread asymptomatically, making it difficult to identify them morphologically. Therefore, culture-independent, sequence-based diagnostic methods that can detect and identify fungi independently of the symptoms that they cause are desirable. Whole genome metagenomic sequencing has the potential to enable rapid diagnosis of plant diseases without culturing pathogens and designing pathogen-specific probes. In my study, the MinION nanopore sequencer, a portable single-molecule sequencing platform developed by Oxford Nanopore Technologies, was employed to detect the fungus *Calonectria pseudonaviculata* (*Cps*), the causal agent of the devastating boxwood blight disease of the popular ornamental boxwood (*Buxus* spp.). Various DNA extraction methods and computational tools were compared. Detection was sensitive with an extremely low false positive rate for most methods. Therefore, metagenomic sequencing is a promising technology that could be implemented in routine diagnostics of fungal diseases.

Other fungi may play important roles in the atmosphere because of their ice nucleation activity (INA). INA is the capacity of some particles to induce ice formation above the temperature that pure water freezes (-38°C). Importantly, INPs affect the ratio of ice crystals to liquid droplets in clouds, which in turn affects Earth's radiation balance and the intensity and frequency of precipitation. A few fungal species can produce ice nucleating particles (INPs) that cause ice formation at temperatures ≥ –10°C and they may be present in clouds. Two such fungal genera are *Fusarium* and *Mortierella* but little is known about their INPs and the genetic basis of their INA. In my study, *F. avenaceum* and *M. alpina* were examined in detail. INPs of both species were characterized and it was found that strains within both species varied in regards to the strength of INA. Whole genome sequencing and comparative genomic studies were then performed to identify putative INA genes. Differential expression analyses at different growth temperatures were also performed. INP properties of the two species shared similarities, both appearing to consist of secreted aggregates larger than 30 kDa. Low temperatures induced INA in both species. Lists of candidate INA genes were identified based on their presence in the strains with the strongest INA and/or induction of their expression at low temperatures and because they either encode secreted proteins or enzymes that produce other molecules known to have INA in other organisms. These genes can now be characterized further to help identify the fungal INA genes in both species. This can be expected to help increase our understanding of the role of fungal INA in the atmosphere.

Genomic, transcriptomic, and metagenomic approaches for detecting fungal plant pathogens and investigating the molecular basis of fungal ice nucleation activity

Shu Yang

GENERAL AUDIENCE ABSTRACT

Fungi are important to life on Earth and play roles in the environments that surround us. On the one hand, fungi can make plants sick and some plant diseases may even cause economic losses to farmers. If the cause of a disease can be identified accurately in an early stage before symptoms develop, disease transmission may be prevented and plants may be protected from disease. However, it is a challenge to find out which fungus causes which disease since symptoms of different fungal diseases look very similar. Typically, we have to wait for plants to become very sick or we have to isolate the fungus that causes a disease to identity it, which may be time-consuming and not lead to precise identification. DNA sequencing technologies have the potential to lead to more sensitive, faster, and more accurate disease diagnosis and, therefore, may help prevent disease outbreaks. In my study, the MinION nanopore sequencer, a small portable device, was used to detect the fungus causing boxwood blight on boxwood. By loading the DNA of unhealthy boxwood on the device, the boxwood blight pathogen was identified within a very short time. Thus, this method is a promising diagnostic method that may be applied to detect other plant fungal diseases as well.

On the other hand, fungi may affect Earth's climate by affecting how many water droplets in clouds are frozen, which in turn affects Earth's temperature and how often and how much it rains and snows. Fungi may affect the freezing of water droplets in clouds since some of them have ice nucleation activity (INA), which is the capacity to catalyze ice formation at a higher temperature than the temperature at which pure water freezes (-38°C), and they may be present in clouds. So far, INA has only been found in a few fungi, including the species *Fusarium avenaceum* and *Mortierella alpina*, but the mechanism of their INA is poorly understood. In my study, multiple *F. avenaceum* and *M. alpina* strains were examined in detail. Two approaches were used. First, strains in each species were compared with each other to find out how strong their INA is. Once it was found that they differed in their strength of INA, their genomes were sequenced and compared to find genes present in the most active strains and missing from the least active strains since it is these genes that may contribute to INA. It was also found that both fungal species had stronger INA when they were grown at lower temperatures. Therefore, the expression of their genes between higher and lower temperatures was compared to find the genes that were more highly expressed at lower temperatures since it is these genes that may cause INA. Based on previous studies, fungal INPs may either consist of secreted proteins or be the products of biosynthetic gene clusters. Therefore, the list of potential genes was reduced by looking for genes encoding either secreted proteins or biosynthetic gene clusters. The list of these potential INA genes will make it easier to identify the INA genes in *F. avenaceum* and *M. alpina* and determine the role of fungi in affecting the weather and climate on Earth.

Dedication

I would like to dedicate this thesis to my mother Prof. Yiping Shao, and my father Prof. Yingbiao Yang.

Acknowledgements

My journey to earning this degree has been long and challenging. I would not have been able to complete this journey without the support and encouragement of many people.

First and foremost, I would like to thank my advisor Dr. Boris Vinatzer for being so supportive and accommodating. I am grateful to have such a great mentor, who always supported me and encouraged me when I was struggling. I appreciate all the help he has given that allowed me to complete this degree during the pandemic. I would also like to thank all the members of my committee, Dr. David Schmale III, Dr. Song Li and Dr. Brian Badgley, for their advice and support.

I would like to thank members of the Vinatzer lab, both past and present: Haijie Liu, Dr. Marco Mechan-Llontop, Marcela A. Johnson, Parul Sharma, Mariah Rojas. They have been great colleagues and friends. I would also like to thank the Translational Plant Sciences Center and the Plant Pathology, Physiology, and Weed Science communities for allowing me to finish my studies. Additionally, I would like to thank Dr. Tim Smalley, my Master's advisor, for guiding me and supporting me since I was a student at the University of Georgia. Without his guidance and support, I would not have been able to pursue a doctoral degree.

Finally, I would like to thank my family and friends in China and other parts of the world for their love and support, especially during the pandemic. Last but not least, I would like to thank myself for not giving up and not losing the ability to love when encountering obstacles.

Table of Contents

List of Tables

List of Figures

Chapter 1. Metagenomic sequencing for detection and identification of the boxwood blight pathogen *Calonectria pseudonaviculata*

Shu Yang[1], Marcela A. Johnson[1,2], Mary Ann Hansen[1], Elizabeth Bush[1], Song Li[1], Boris A. Vinatzer[1*]

1 School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

2 Graduate Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, United States

*Corresponding author; email: vinatzer@vt.edu

Keywords: boxwood blight, disease diagnosis, metagenomics, MinION, nanopore sequencing

** Submitted to the journal Scientific Reports in July 2021

Attribution

B.A.V. and S.L. developed the project and supervised its execution. S.Y. performed the experiments and contributed Figure 1. S.Y. and M.A.J. performed the sequence analysis with S.Y. contributing Figures 3 and 4, Supplementary Figures 1, 2, and 3, Tables 3 and 4, and Supplementary Tables 1, 2 and 3 and M.A.J. contributing Figures 2 and 5 and Table 2. M.A.H. and E.B. provided plant samples and contributed advice on boxwood blight diseases and pathogen biology. S.Y. and B.A.V. wrote the manuscript with contributions from M.A.J. and all authors reviewed and revised the manuscript.

Abstract

Pathogen detection and identification are key elements in outbreak control of human, animal, and plant diseases. Since many fungal plant pathogens cause similar symptoms, are difficult to distinguish morphologically, and grow slowly in culture, culture-independent, sequence-based diagnostic methods are desirable. Whole genome metagenomic sequencing has emerged as a promising technique because it can potentially detect any pathogen without culturing and without the need for pathogen-specific probes. However, efficient DNA extraction protocols, computational tools, and sequence databases are required. Here we applied metagenomic sequencing with the Oxford Nanopore Technologies MinION to the detection of the fungus *Calonectria pseudonaviculata*, the causal agent of boxwood (*Buxus* spp.) blight disease. Two DNA extraction protocols, several DNA purification kits, and various computational tools were tested. All DNA extraction methods and purification kits provided sufficient quantity and quality of DNA. Several bioinformatics tools for taxonomic identification were found suitable to assign sequencing reads to the pathogen with an extremely low false positive rate. As few as 200 sequencing reads were required for pathogen detection in severely diseased samples and identification at strain-level resolution was approached as the number of sequencing reads was increased. We discuss how metagenomic sequencing could be implemented in routine plant disease diagnostics.

Introduction

The sooner a disease outbreak is detected and the causative agent is identified, the faster the outbreak can be controlled by implementing testing, quarantine, and isolation. This applies to human, animal, and plant diseases [1]. Boxwood blight is a devastating fungal plant disease of ornamentals in the Buxaceae family including boxwood (*Buxus* spp.), sweet box (*Sarcococca*

spp.), and pachysandra (*Pachysandra* spp.). Because boxwood is one of the most popular horticultural crops in the U.S. with annual sales of $126 million [2], boxwood blight has caused significant economic losses and is of great concern to the landscape and nursery industry and home growers. The disease is caused by two closely related fungal species, *Calonectria pseudonaviculata* (*Cps*) and *Calonectria henricotiae* (*Che*). While *Cps* is widely distributed in North America, western Asia and Europe, *Che* has only been observed in Europe so far [3]. *Cps* was first detected in the U.S. in 2011 and has since been reported in at least 30 states [4]. Since *Cps* mainly spreads through infected plant material, contaminated tools, and other surfaces, early and rapid pathogen detection to avoid the distribution of infected plant material to home growers, nurseries, and public parks is critical to managing this disease.

Several diagnostic methods have been used for the detection of boxwood blight. Traditional morphology-based methods use observation of spores under the microscope. This requires expertise and a relatively long incubation period of the collected plant material because sporulation may need to be induced first [4]. In some cases, it is even necessary to isolate and culture the pathogen before spores can be observed. Moreover, spores of *Cps* and *Che* are so similar that their differentiation is challenging [5] and there is even the risk that other fungi are mistaken for *Cps* [4].

Molecular detection methods have been developed for faster and more sensitive detection of *Cps*. Polymerase chain reaction (PCR)-based assays are commonly used for direct detection of *Cps* and have been validated using environmental samples. However, in the early stages of assay development, these tests had a risk of false-positive signals [6], and a trade-off between specificity and sensitivity in PCR-based assays has been found [7]. A set of new PCR-based protocols were developed to differentiate between *Cps* and *Che* but have only been validated on artificially inoculated plants [8]. Other molecular methods are based on Loop-mediated isothermal

amplification (LAMP) and have been shown to exhibit high specificity for pure cultures. These assays can discriminate between the target pathogen and closely related species that may be present in the rhizosphere with no false-positive results. However, validation of *Cps* in rhizosphere samples gave negative results [9]. Finally, Next-generation sequencing (NGS) using Illumina technology has also been used to identify *Cps* as the pathogen causing Sarcococca blight. This method was able to identify *Calonectria* at the species-rank, but only after DNA was obtained from pure fungal cultures [10].

Whole genome metagenomic sequencing is a promising new approach for pathogen detection and identification for disease diagnosis [11,12]. This culture-independent method consists in sequencing all DNA or RNA present in a sample, for example from a symptomatic host, and has been shown to provide accurate diagnosis. Since metagenomic sequencing does not rely on pathogen-specific probes or primers, little to no previous knowledge of the putative identity of the pathogen is required. Metagenomics approaches utilizing NGS have been used in clinical research and are gradually being adopted in diagnosing plant diseases as well [13,14]. To achieve a rapid diagnosis, the MinION nanopore sequencer, a single-molecule long-read sequencing platform developed by Oxford Nanopore Technologies Inc. (ONT) is particularly promising. It has several advantages over other NGS sequencing platforms: longer reads improve genome assembly and increase the precision of detection, first results are available minutes after a sequencing run is initiated, and it can be used almost anywhere, even in Space [15]. This portable sequencer has thus been used for metagenomic sequencing in medical research to successfully detect and sequence pathogens like Ebolavirus [16] and SARS-CoV-2 [17].

However, the MinION has limitations regarding sensitivity and accuracy. Read accuracy is around 90%, which is lower than that of the short read technology Illumina. Although accuracy

has recently been improved by increasing the accuracy with which raw signals obtained by the MinION are translated into base-pairs, a process called "base-calling" [18]. A more general challenge with metagenomics is that host genome sequences in the extracted DNA may represent the majority of the data [19] and non-pathogenic microorganisms associated with the host plant may reduce the percentage of pathogen sequences further [20], making it difficult to detect the causative agent.

With regard to plant disease diagnostics, metagenomic sequencing with the MinION using DNA or RNA extracted directly from plants enables rapid pathogen detection and identification in almost any laboratory or even in the field [20]. However, so far, the MinION has mainly been used to identify plant pathogenic viruses [21,22] and bacteria [23,24]. Few studies have reported using the MinION for detection of plant pathogenic fungi [19,25], which is challenging because of the poor representation of fungal genomes in reference databases and the technical difficulties in isolating high quality fungal DNA directly from plant tissue.

Here we applied metagenomic sequencing to the detection of *Cps* in naturally infected boxwood. The main objectives were to (i) find a DNA extraction method that yields a high concentration of pathogen DNA of sufficient quality for sequencing and (ii) develop a bioinformatics workflow that optimizes detection sensitivity and specificity of the pathogen. While we focused on *Cps* and boxwood, the developed approach should be adaptable to most fungal pathogens of most plants and thus contribute to the improvement of plant disease diagnostics for outbreak control in general.

Materials and Methods

Plant material

Naturally infected boxwood samples from various locations in Southwest Virginia were obtained from the Virginia Tech Plant Disease Clinic. Collection of plant material was done complying with institutional, national, and international guidelines and legislation. Samples were either moderately diseased or severely diseased (Supplementary Figure 1). Healthy boxwood collected in the towns of Blacksburg and Floyd, Virginia, where no boxwood blight had been recorded at the time, served as negative controls. Plant material was stored at 4°C for immediate use, otherwise at -80°C until DNA extraction.

Extraction methods used to prepare DNA for MinION sequencing

To determine the most efficient DNA extraction method, both moderately and severely diseased samples were either sonicated (without disrupting plant cells) or homogenized in liquid nitrogen (disrupting plant cells) (Figure 1). DNA was measured using a Thermo Scientific NanoDrop spectrophotometer.

For sonication, 4.5 g of plant tissue composed of twigs of moderately diseased or severely diseased plants were placed in a Ziploc bag containing nuclease-free water. Next, the bag was sonicated for 15 minutes to dislodge as many microorganisms as possible from the plant into the liquid and disrupt their cells. The liquid went through a vacuum filter flask to concentrate DNA on the filter membrane. DNA was extracted from the membrane using kits designed for water and soil samples, as shown in Table 1 (sample IDs starting with the letter S).

For homogenization, plant tissue composed of leaves and stems randomly picked from moderately diseased or severely diseased plants was ground in liquid nitrogen. 0.1 g of ground tissue was used for DNA extraction using kits as shown in Table 1 (sample IDs starting with the letter G). For extraction from severely diseased plant batch 1, 0.1 g of severely diseased boxwood

was ground and processed individually for each DNA extraction. However, to make plant samples more similar to each other and results obtained with different kits more comparable, this was changed for the later batches: several grams of tissue were ground together and then 0.1 g aliquots were used for individual DNA extractions. For the negative control, DNA was extracted with the ZymoBIOMICS DNA Miniprep Kit from a 0.1 g aliquot of ground, healthy plant tissue (sample ID: NC).

MinION library preparation and sequencing

MinION Library preparation was performed according to the native barcoding genomic DNA protocol (EXP-NBD104, EXPNBD114, and SQK-LSK109) [26] with minor modifications. The library was prepared using the Ligation Sequencing Kit (ONT; SQK-LSK109). For each run, first, DNA for each sample was repaired and end-prepped for each sample using the NEBNext Ultra II End Repair/dA-Tailing Module (New England Biolabs, Inc.; Catalog # E7546S). 90 µL AMPure XP beads were used for cleaning up repaired DNA. Then repaired DNA was washed on a magnetic rack using freshly made 70% ethanol and eluted with 25 µL nuclease-free water. Second, native barcode ligation was performed by mixing 22.5 µL of the elute with the Blunt/TA Ligase Master Mix (New England Biolabs, Inc.; Catalog # M0367S) and Native Barcode (ONT; Native Barcoding Expansion Kit EXP-NBD104). Barcoded DNA was cleaned up by another wash step using 90 µL AMPure XP beads, and DNA was eluted in 26 µL nuclease-free water. Then equimolar amounts of each barcoded DNA were pooled into a 1.5 mL microcentrifuge tube. Last, adapter ligation was performed by mixing the pooled barcoded sample with Adapter Mix (ONT; SQK-LSK109), NEBNext Quick Ligation Reaction Buffer (New England Biolabs, Inc.; Catalog # B6058S) and Quick T4 DNA Ligase (New England Biolabs, Inc.; Catalog # M2200S). Ligated

DNA was cleaned up with 60 μL AMPure XP beads, washed on a magnetic rack using Long Fragment Buffer (ONT; SQK-LSK109), and eluted with 15 μL Elution Buffer (ONT; SQK-LSK109).

Sequencing reactions were performed independently for each run on a MinION flow cell (ONT; FLO-MIN106 R9.4.1 Version) connected to a Mk1B device (ONT; MIN-101B) operated by the MinKNOW software (ONT, Inc. v19.12.2). Each flow cell was primed with the priming buffer prepared by mixing 30 μL Flush Tether (ONT; EXP-FLP002) with a tube of Flush Buffer (ONT; EXP-FLP002). 12 μL of the final library mixed with Sequencing Buffer (ONT; SQK-LSK109) and Library Loading Beads (ONT; SQK-LSK109) were loaded onto the SpotON sample port of the flow cell in a dropwise fashion. The sequencing run was stopped when all pores lost activity, usually after 48-72 hours. A new flow cell was used for each run. Sample IDs and descriptions are shown in Table 1. After sequencing, the raw files in FAST5 format, containing the electrical signals, were translated (base-called) with the ONT tool Guppy GPU (v3.2.2) into sequences with a minimum q-score of 7 and saved as FASTQ files for further analysis. The FASTQ files were then converted to FASTA files with an in-house shell script.

DNA extraction and Illumina sequencing

Healthy plant tissue (100 mg) and severely diseased plant tissue (100 mg) were homogenized in liquid nitrogen for DNA extraction for Illumina sequencing to serve as controls for MinION sequencing. DNA of healthy boxwood was extracted using Invisorb Spin Plant Mini Kit, and DNA of severely diseased boxwood was extracted using ZymoBIOMICS DNA Miniprep Kit.

Whole-genome sequencing of healthy boxwood was performed on an Illumina Nova Seq 6000 Platform (2 × 150 bp) at Novogene Corporation Inc. (Sacramento, CA). Low-quality reads

and adapters were removed by the company. Illumina sequencing of severely diseased plant tissue was performed on an Illumina HiSeq 3000 Platform (2 × 100 bp) at the Iowa State University DNA Facility, and the quality of reads was checked using FastQC v0.11.9 [27]. Reads were trimmed using Trimmomatic v0.39 [28] to remove adapters.

Metagenomic analysis

Two custom fungal genome databases were constructed for taxonomic assignment of fungal reads. First, to determine the DNA extraction method that yields the highest percentage of *Cps*, a small database containing only four fungal genomes of the family Nectriaceae was constructed: *Cps* CBS 139395, *Che* CBS 138102, *Fusarium graminearum* PH-1, and *Pseudonectria foliicola* AR2711 (downloaded from NCBI). The *Cps* genome was used to identify *Cps* reads and the Volutella blight pathogen *Pseudonectria foliicola* was included since it frequently co-infects boxwood with *Cps*. The *Che* genome was added as the negative control since it is closely related to *Cps* but is not present in the USA and the *F. graminearum* genome served as the second negative control since it is another member of the family Nectriaceae but does not cause disease on boxwood. A more extensive database (referred to as large database from here on) was used for a more in-depth characterization of the obtained metagenomes: all assembled genomes of *Cps*, *Che*, *F. graminearum*, *P foliicola* and *Pseudonectria buxi* (another Volutella blight pathogen) available at NCBI in April 2021 (Supplementary Table 1). Reads were trimmed with Porechop v0.2.4 [29] to remove adapters before using them with this database.

Three bioinformatics tools for taxonomic assignment of MinION readswere used: 1. BLASTN v2.10.0+ [30], 2. MetaMaps v0.1 [31], and 3. Kraken 2 v2.1.1 [32]. BLASTN was chosen because it is a commonly used tool to identify fungi [33]. The -evalue parameter was set to less than

0.001, and results were filtered for alignments longer than 1000 bp. For each read, the hit with the lowest E-value was used for taxonomic assignment. MetaMaps was specifically developed for taxonomic assignment of long metagenomic reads [31]. The parameter --perc_identity was set to 85, and hits were further filtered to an identity greater than 85% since hits with lower percentage identity were still reported even using the --perc_identity 85 parameter. Since MetaMaps provides a single taxonomic assignment for each read, ranking was not necessary. Kraken 2 is a popular tool for taxonomic read assignment that provides high accuracy and has faster speeds and lower memory requirements than the original Kraken [32,34]. It has been shown to work well for MinION reads [35] but was originally designed for short reads and was thus used for both MinION and Illumina reads. The default parameters were used for MinION reads, and the parameter --paired was used for Illumina reads.

Since contigs derived from assembled reads have a lower error rate than raw reads, *Cps* genomes were assembled to attempt identification of the *Cps* lineage present in our sample. *Cps* reads that had been pre-identified by BLASTN in samples G10, G11 and G12 using the extensive database were used as input. Canu v2.1.1 [36] was used for assembly and QUAST v5.0.2 [37] and BUSCO v5.0.0 [38] were used to assess the quality of the assembled *Cps* genome. CBS139395 served as the reference genome for QUAST. BUSCO was based on the lineage-specific profile library hypocreales_odb10. To explore strain-level identification, BLASTN and sourmash v4.0.0 [39] were then used in parallel to determine the similarity between the genome assemblies and the reference *Cps* genomes. For sourmash, the parameters -p, scaled=1000, and k=21 were used for generating signatures of the assembly and the reference genomes with the sketch dna command. The search command was then used to identify which *Cps* genome in the database was most similar

to the assemblies (measured as Jaccard similarity). For BLASTN, the same parameters as in the previous sections were used.

To determine the minimal number of MinION reads required to consistently detect *Cps*, reads were randomly sub-sampled 10 times at each of the following sub-sample sizes: 200, 300, 500, 700, and 1000. For each sub-sample, BLASTN hits for *Cps* were retrieved using the read IDs and counted.

All programs were run on Virginia Tech's high performance computer network ARC. For data visualization, R was used to generate the bubble plot. KronaTools v2.7.1 [40] was used to generate graphical interactive html taxonomy abundance piecharts.

Results

Experimental design overview

To determine the feasibility of culture-independent metagenomics for detection of the boxwood pathogen *Cps*, several DNA extraction methods, two DNA sequencing technologies, and several bioinformatics metagenomics analysis tools were compared. Because results of DNA extraction methods can only be compared with each other after sequencing and bioinformatics analyses have been completed and because it was not feasible to compare all combinations of protocols and tools, experiments and respective results were grouped as follows: 1. Identification of DNA extraction methods that provide DNA of sufficient quantity and quality for ONT MinION sequencing and a high percentage of *Cps* sequencing reads based on the analysis of all samples sequenced with the ONT MinION using two metagenomics tools and a small fungal reference database; 2. *Cps* identification using additional bioinformatics tools in combination with a large fungal genome database; 3. Comparison of results obtained with the ONT MinION to results obtained with the

Illumina sequencing platforms using a bioinformatics tool that can be used for both platforms; 4. Attempt at lineage-specific *Cps* identification after assembling sequencing reads; 5. Determination of the smallest number of MinION reads necessary to detect *Cps* in severely diseased samples.

DNA extraction from both, ground boxwood tissue and wash water of sonicated tissue, provides DNA of sufficient quality and quantity for detection of *Cps*

Two fundamentally different DNA extraction methods were tested: extraction of DNA from wash water of relatively large sonicated plant samples (4.5 g) and DNA extraction from a relatively small amount of plant tissue (0.1g) that was ground in liquid nitrogen (Figure 1). The rationale was that sonication can be expected to maximize the DNA of microorganisms that are easily separated from the host plant and should thus minimize contaminating plant DNA, whereas homogenization in liquid nitrogen efficiently frees DNA from all cells (plant, prokaryotic, and fungal) and can thus be expected to increase fungal DNA yield while also increasing plant DNA contamination.

Both extraction methods and all kits resulted in more than 1 µg per sample, which is the required minimum for use with the ONT MinION native barcoding genomic DNA protocol. DNA concentrations ranged widely from 76 ng/µL to over 1,133 ng/µL, but the majority of DNA extractions using either grinding or sonication yielded DNA concentrations in the range from 100 to 500 ng/µl and were similarly effective for both moderately and severely diseased samples (Table 1).

With regard to quality, we determined the A260/A280 (DNA/protein) and A260/A230 (DNA/other impurities) ratios, which for pure DNA are expected to be around 1.8 and 2.0-2.2, respectively. A260/A280 ratios were close to 1.8 for most samples independent of extraction

method and severity of disease (with the exception of one DNA sample extracted from a ground severely diseased sample, which had a ratio of only 0.89), suggesting low protein contamination in most samples. The A260/A230 ratio instead varied widely from almost 0 to 2.2, and DNA extracted from ground samples had generally lower ratios than DNA extracted from wash water after sonication, suggesting that more impurities were present in DNA extracted from ground samples. Severity of disease did not appear to affect the A260/A230 ratio.

Next, we analyzed the overall DNA sequencing output focusing on the total length of reads and the number of reads obtained per sample (Table 2). To make sequencing results comparable between samples (which were sequenced by combining different numbers of barcoded samples per flow cell) we normalized the total read length and number of reads per sample by multiplying them by the number of barcodes used in the respective flow cell. In other words, we computed the total read length and number of reads that we would have obtained if we had used an entire flow cell for each sample. Normalized read length varied between 5.4 to 26.2 gigabases (Gb) for DNA extracted from wash water of sonicated samples and between 2.9 and 22.9 Gb for DNA extracted from ground samples. The normalized number of reads varied similarly widely between 1.4 to 11.4 million (M) for DNA extracted from wash water of sonicated samples and between 1.4 and 19.2 M for DNA extracted from ground samples. Also, average read length and the length of the longest read varied widely for both extraction methods. As with DNA concentration and quality, severity of disease did not affect overall sequencing results. In summary, all extraction methods and kits were comparable in regard to overall DNA sequencing metrics and, unexpectedly, sequencing results did not correlate with either DNA concentration or DNA quality.

Finally, sequencing results were analyzed for the presence of *Cps* sequences. To do this, reads were classified taxonomically using two independent tools in parallel, BLASTN and

MetaMaps, and a small fungal reference library containing one *Cps* genome and one genome each of three additional species in the Nectriaceae family. While BLASTN generally identified twice as many reads as *Cps* compared to MetaMaps (Table 2 and Figure 2), the relative number of *Cps* reads between individual samples was the same for both tools, giving confidence that either tool could be used to compare samples with each other. Since BLASTN is the more widely used tool out of the two, only BLASTN results are reported in the next paragraphs.

Since samples differed from each other in the number of reads and total read length, we determined (1) the percentage of reads assigned to *Cps* out of all reads per sample (Table 2 and Figure 2) and (2) the percentage of the total length of reads identified as *Cps* out of the total length of reads per sample (Table 2). With regard to read number, DNA extracted from ground samples recovered a higher percentage of *Cps* reads (up to 9.44%) compared to DNA extracted from sonicated samples (only up to 0.15%). With regard to the percentage of the total length of *Cps* sequences out of the total sequencing length, DNA extracted from ground samples gave percentages of up to 12.52% compared to only 0.35% for sonicated samples. However, two samples obtained from ground tissue (G7 and G8) of the severely diseased batch 1 also had low percentages of *Cps* with regard to read number and length.

We cannot make any conclusions on individual DNA purification kits because most kits were only used once with moderately diseased boxwood samples and once with severely diseased boxwood samples. Additionally, DNA was sequenced on four separate flow cells (which quality is known to be inconsistent, in particular, with regard to the number of active pores). Importantly though, all kits performed sufficiently well to allow for downstream *Cps* detection.

As expected, a higher percentage of *Cps* reads was obtained from severely diseased samples (up to 9.44%) than from moderately diseased samples (up to 0.93%). Importantly, not a

single *Cps* read was found in the negative control DNA extracted from a healthy boxwood plant. With regard to the other fungal species included in the reference library, only a very small number of reads of *Che* and *Fusarium graminearum* were recovered. When the reads identified as *Che* using our small reference library were compared by BLASTN against the entire nt database at NCBI [41], these reads were more similar to other fungi or bacteria than to *Che* and were thus false positives. The ubiquitous boxwood pathogen *Pseudonectria foliicola* was found in all diseased samples in percentages similar or even higher than *Cps* but not in the healthy boxwood sample.

Robust *Cps* identification using BLASTN and Kraken 2 in combination with an expanded Nectriaceae genome database

For a more in-depth characterization of *Cps* and the other Nectriaceae family members in the metagenomic sequences, a large database containing all public genome assemblies of *Cps*, *Che*, *P. foliicola*, *P. buxi*, and *F. graminearum* was used. Although we had used BLASTN and MetaMaps to identify the best DNA extraction methods above, we replaced MetaMaps with Kraken 2 [32] here. Compared to MetaMaps, Kraken 2 has been used more widely in published metagenomic studies, is user-friendly, and has been shown to have high accuracy, low memory usage, and high speeds [32,34].

First, species-level taxonomic classification results obtained with Kraken 2 were compared with those obtained with BLASTN and showed that Kraken 2 also identified *Cps* in all diseased samples (Supplementary Table 2). Kraken 2 classified an even higher number of reads as one of the five fungal species present in the reference database than BLASTN. For example, Kraken 2 classified 26.62% of total reads in G10 as belonging to the five fungal species while BLASTN

only 20.75%. For the moderately diseased samples from which DNA was extracted after sonication, Kraken 2 identified 0.05 to ~ 0.11% of total reads as *Cps* (Supplementary Figure 2).

When looking specifically at *Cps*, 36.53 % of all reads assigned to one of the five Nectriaceae species in sample G10 were identified as *Cps* by Kraken 2, whereas 44.19% were identified as *Cps* by BLASTN (Figure 3). For G12, 37.83 % of fungal reads were identified as *Cps* by Kraken 2, whereas 45.76% were identified as *Cps* by BLASTN. This difference is due to the fact that Kraken 2 classified a subset of *Calonectria* reads at the *Calonectria* species complex rank without assigning them to an individual species, but our BLASTN pipeline assigned all fungal reads at the species rank.

Also when using the large reference database, a small subset of reads was identified as *Che* by both BLASTN and Kraken 2. However, these reads matched bacterial, yeast, or plant sequences when compared against NCBI's nt database (Supplementary Table 3 shows the results for sample S1 as example). The most remarkable new result using the large fungal database was the identification of the Volutella pathogen species *P. buxi* at an abundance similar to *P. foliicola*. The *P. buxi* reads were probably identified as *P. foliicola* when using the small database since *P. buxi* was not included in the smaller database. As with *Calonectria*, Kraken 2 classified some reads as *Pseudovaniculata* without species designation, while our BLASTN pipeline assigned all *Pseudovaniculata* reads to either *P. foliicola* or *P. buxi*. Approximately 0.5% of fungal reads in G10 and G12 were identified as *F. graminearum* but may belong to related *Fusarium* species since only *F. graminearum* genomes were included in the database, and it was thus not possible to distinguish between individual *Fusarium* species.

Unexpectedly, a small number of reads were identified as *Cps* by both Kraken 2 and BLASTN in the healthy negative control sample. Still, as the *Che* reads above, they were identified as false positives when comparing them to NCBI's nt database [41].

MinION and Illumina sequencing provide similar results in regard to *Cps* identification

To compare the results of ONT MinION long-read sequencing with the Illumina short-read platform, sample G10 and a negative control sample were sequenced using Illumina technology. Since Kraken 2 can be used for both short- and long-reads [35,42], we used Kraken 2 in combination with our large fungal database to compare the results from the two sequencing platforms. Illumina sequencing yielded 17,033,700 paired-end reads with a total length of 1.50 Gb compared to the 541,576 long reads with a total length of 1.96 Gb obtained by MinION sequencing (Supplementary Table 4). 9.73% of MinION reads and 6.14% of Illumina reads were identified as *Cps*, respectively (Figure 4). The lower percentage of Illumina reads identified as *Cps* was compensated by the higher percentage of Illumina reads that were assigned to the *Calonectria naviculata* species complex without species identification.

Since we had no DNA of the healthy boxwood left that we had used as the negative control for MinION sequencing, a different DNA sample of a healthy boxwood was sequenced with Illumina. Illumina sequencing yielded 271,857,762 paired-end reads with a total length of 40,778,664,300 bp for sample (Supplementary Table 4). As for the healthy negative control sample used with MinION sequencing, a very small number of reads of this sample were assigned to *Cps* (Figure 4). However, when these reads were compared with the entire nt database at NCBI using BLASTN, they were again found to be false positives.

Cps in diseased plants can be identified to a within-species cluster using sourmash and BLASTN

In a recent study, investigating the emergence of boxwood blight using population genomics, several clusters/lineages within the *Cps* species were identified [43]. Therefore, we wanted to determine if *Cps* reads in our samples could be assigned to one of the identified clusters. Since the program sourmash can identify bacterial genomes in metagenomes independently of taxonomy and without the need for NCBI taxonomic identifiers, we first attempted to use sourmash using all reads of samples G10, G11, and G12 as query and the same extended fungal database we had used with Kraken 2, but sourmash did not identify any fungal genome in any of the samples. However, Table 3 shows that when using only the reads that had been identified as *Cps* by BLASTN as query, sourmash did find them to have similarity to *Cps* genomes. The highest similarity was to the genomes of *Cps* isolates CBS139394 and CBS139395 (both isolated from sweet box in Maryland, USA [10]) followed by genome sequences of isolate CB002 (isolated from boxwood in Belgium [5]). Similarity was unexpectedly low (14-19%). Since the low similarity could have been due to sequencing errors present in individual reads, we then assembled all *Cps* reads from G10, G11, and G12 with the expectation that the assembled reads would have fewer errors and be more similar to the reference genomes. The *Cps* genome we obtained was 49,048,547 bp long and consisted of 1055 contigs. 48,291,239 bp of the assembly aligned with 88.746% of the chosen reference genome CBS139395 (Table 4). Although this revealed that our assembly covered most of the *Cps* genome, only 50.3% of genes were complete and 23.2% were fragmented compared to 96.6% of genes that were complete in the reference genome CBS139395 based on BUSCO [38] assessment (Table 4). When the assembled genome was used as query with sourmash against our fungal database, the genomes CBS139394, CBS139395, and CBS002 were again found to be most

similar, but now with a similarity value close to 73% (Table 3). When using BLASTN, the assembled *Cps* genome had a significantly higher number of best hits to CBS139395 than to all other genomes (Table 3).

Cps can be detected with as few as 200 MinION sequencing reads in severely diseased tissue

After showing that *Cps* can be identified with high specificity from naturally infected boxwood tissue using metagenomic sequencing with the ONT MinION, we wanted to investigate the minimal number of reads needed to detect *Cps*. We thus computationally sub-sampled samples G10, G11, and G12 to different read numbers generating 10 random subsamples for each size shown in Figure 5. Importantly, even for the sub-samples consisting of only 200 total reads, there was not a single sub-sample in either G10, G11, or G12 without *Cps* reads (Figure 5).

Discussion

Sensitive, specific, and fast pathogen detection is instrumental in plant disease control and management. Here we explored metagenomic sequencing using the ONT MinION and Illumina for detection and identification of the boxwood blight pathogen *Cps*.

To effectively use metagenomics for *Cps* detection, we first needed to identify a suitable DNA extraction method. We tested two protocols. One protocol aimed at minimizing host DNA by not disrupting host cells and assuming *Cps* could be separated from host tissue by washing and sonication. The other protocol was designed to obtain as much total DNA as possible by disrupting both host cells and fungal cells by grinding in liquid nitrogen. For most samples, disrupting host cells yielded more *Cps* sequencing reads than not disrupting host cells. This indicates that most *Cps* is likely to be embedded in host tissues upon infection, while only a small amount of *Cps*

exists on the host surface. However, for all samples, *Cps* reads were identified even in DNA extracted from wash water of sonicated tissue revealing that both protocols can be used to prepare DNA for metagenomic sequencing.

Compared to results using metagenomic sequencing for the identification of bacterial plant pathogens, the recovery of fungal pathogen reads in this study was relatively low. In fact, up to 60% of reads were identified as the bacterial pathogen *Xanthomonas perforans* in tomato plants naturally infected with bacterial spot [23]. However, for fungal plant pathogens, other studies reported recovery of very few pathogen reads. For example, DNA of wheat inoculated with fungal pathogens was extracted by homogenization using a protocol designed for fungi for long-read sequencing [44], and at most 5.7% of the total sequence length was identified as the target fungal pathogen by BLASTN [19]. Therefore, the DNA extraction methods used here for *Cps* and boxwood may have the potential to be successful with other fungal plant pathogens as well.

Compared to the detection of bacterial plant pathogens by metagenomic sequencing, fungal plant pathogens present another challenge. Prokaryotic genome databases include dozens, or even hundreds, of genome sequences for most bacterial plant pathogen species, while genome sequences of fungal plant pathogens are still relatively rare in genome databases. This could contribute to the relatively low number of sequencing reads identified as being of fungal origin compared to bacterial origin in some metagenomic studies [25]. In our study, we were unable to use the ONT-provided WIMP taxonomic classification tool for metagenomic analysis when starting this project because *Cps* genomes were not included in the WIMP database. We thus had to build our own custom databases for use with the bioinformatics tools employed here. Fortunately, several genome sequences of *Cps* and *Che* became publicly available by the end of this project and could be included in our large database. Although BLASTN, MetaMaps and Kraken 2 were

all adequate in identifying the target plant pathogen using our databases, sensitivity varied. For example, a larger number of *Cps* reads was identified by Kraken 2 compared to BLASTN for most diseased samples, and fewer false-positive reads were identified by Kraken 2 in the negative control. On the other hand, a significant number of reads was assigned to non-specific species complexes or genera in the family Nectriaceae.

It is worth noting that *Che*, which is not present in the USA, was identified in diseased samples at very low abundance of 0.001-0.807% by BLASTN (0.000-0.399% by MetaMaps, 0.012-0.312% by Kraken 2). This indicates that all three tools were mostly able to differentiate *Cps* from the closely related species *Che*. Moreover, besides these reads misidentified as *Che*, a small number of reads were identified as *Cps* in the negative healthy control sample. In both cases, when performing BLASTN on these potential false *Che*- and *Cps*-positive reads against the entire NCBI nt database, the best matches for these reads were plants, bacteria, and other fungi. For reads shorter than 100 nt, sometimes *Che* or *Cps* were the best hits but percent identity and bit-score were very low (data not shown). Therefore, the wrongly identified reads were mostly a result of using relatively small custom fungal databases lacking plant, bacteria, and other fungal genomes. We chose to use these relatively small custom databases to accelerate read identification but the resulting false positives are clearly a weakness resulting from this decision. Larger, more comprehensive databases and filtering out short reads can be expected to avoid false positives almost completely. However, it may be impossible to avoid all misidentifications since some reads may get misidentified because they align to genes highly conserved within the genus or family of interest.

It was expected that reads of the Volutella pathogens *P. foliicola* and *P. buxi* would be identified in all diseased samples since they are ubiquitous boxwood pathogens. However, it was

interesting that not a single read of either pathogen was identified in the two healthy negative control samples, suggesting that these pathogens only thrive in co-infection with *Cps*. It was also expected that very few reads of *F. graminearum* would be recovered because this species does not cause disease on boxwood. Also, prokaryotes were identified in all samples as described in Supplementary Results 1.

Besides distinguishing between species, metagenomics was shown to almost reach strain/lineage-level precision for plant pathogenic bacteria [23]. *Cps* has diversified into multiple lineages with several of them being present in the US [43,45]. Neither MetaMaps nor Kraken 2 can easily distinguish between lineages since they rely on NCBI taxIDs and only a single taxID is associated with each fungal species. Also, MinION reads have a relatively high error rate and Illumina reads are short, further complicating precise identification. However, we have shown here that assembling MinION reads made it possible to determine which public *Cps* genome sequences were most similar to the *Cps* sequences in some of our samples using either BLASTN or sourmash. Both tools identified the same three strains as best hits, including the strains CBS139395 and CBS139394, both isolated from sweet box (*Sarcococca* spp.) in the same location in Maryland, USA [10], and both members of clade B [43]. While this result is not sufficient to conclude that the *Cps* strain from our Virginia samples belongs to the same clade, it shows the potential of metagenomic sequencing to reach strain/lineage-level resolution not only for bacteria but also for fungi. Using the obtained *Cps* genome assembly as input into a single nucleotide polymorphism (SNP) pipeline for phylogenetic tree construction will be necessary to confidently assign it to clade B. Also, sequencing a sample on an entire flow cell should provide a higher number of *Cps* reads to obtain a better genome assembly compared to the one we were able to obtain, which had a limited number of complete genes.

Compared to Illumina sequencing, the MinION revealed several strengths. First, the requirements of DNA quantity and quality were lower. Second, with long reads, initial identification using the MinION can be made without assembling metagenomes. Also, its portability and ability to report results in real-time can't be matched by Illumina. Although the relatively high error rate of the MinION is often considered a weakness, it was not a limitation in our study. The increased length of reads compared to Illumina provided high confidence read identification and easily compensated for the higher error rate.

With regard to detection, 200 MinION reads would have been sufficient to consistently detect *Cps* in the samples with the highest percentage of *Cps* reads. The MinION was also able to detect *Cps* in moderately diseased boxwood, although the percentage of reads identified was lower than 1% and, therefore, a much higher number of reads would be required to confidently detect *Cps*. We did not have the opportunity to determine the detection limit for asymptomatic boxwood. Moreover, infection severity may vary significantly between different asymptomatic samples and it may thus be challenging to determine how many reads would be required to confidently conclude that *Cps* is absent. On the other hand, the very low false positive rate provides confidence in identifying an infection even when a very small number of *Cps* reads were detected. Since we had no access to *Cps*-specific molecular PCR or LAMP assays, we cannot compare detection sensitivity of metagenomic sequencing using the MinION with these assays and can only generally state that the sensitivity of metagenomic sequencing increases with the number of total sequencing reads that are generated. Therefore, if high sensitivity of detection is required, one can increase the total number of reads by using an entire flow cell per sample or even using more than one flow cell.

A current challenge with metagenomic sequencing for pathogen identification is that knowledge of bioinformatics is required when using many of the open-source tools designed for this purpose. Although the BLAST program can be performed locally, for higher speed and efficiency, it had to be installed on Virginia Tech's high performance computer network, ARC. To automate the comparison of every individual sequencing read to our databases and to summarize the obtained results, custom scripts needed to be written. Also, MetaMaps, Kraken 2, and sourmash were run on ARC because the amount of sequence data obtained in metagenomics is too much to handle for a standard laptop or desktop computer. This is an obvious challenge when trying to implement metagenomics into routine disease diagnostics. A user-friendly program interface and automated pipelines running at the back-end on a high-performance computing network will both be required. If these become available, a diagnostic clinic could extract DNA from a sample, prepare a sequencing library, and start a sequencing run within hours and obtain first results on the same day. This would represent a significant acceleration compared to any culture-dependent diagnostic technique and even applicable to the detection of emerging pathogens for which no specific qPCR test may be available.

In conclusion, we have shown here that using appropriate DNA extraction techniques and bioinformatics tools and genome databases, metagenomic sequencing using the ONT MinION can easily distinguish the boxwood blight pathogens *Cps* and *Che* from each other and from other fungal species. With some improvements to databases and parameters used in the classification pipeline, it should be possible to eliminate false positives to practically zero. Using a high enough number of reads, metagenomic sequencing with the ONT Minion can also reach very high sensitivity of detection and specificity can approach strain-level resolution. The main challenge to implementing metagenomic sequencing for plant pathogen identification in routine diagnostics

will be in providing access to high performance computing networks and user-friendly interfaces from which to run the necessary computational pipelines.

Acknowledgements

Funding for this project was provided by the Virginia Agricultural Council (PFJK4I7B). Funding to S.Y. was also provided in part by the School of Plant and Environmental Sciences at Virginia Tech. Funding to M.A.J. was also provided in part by the Virginia Tech graduate program in Genetics, Bioinformatics and Computational Biology at Virginia Tech. Funding to B.A.V. and S.L. was also provided in part by the Virginia Agricultural Experiment Station and the Hatch Program of the National Institute of Food and Agriculture, United States Department of Agriculture.

Additional information

The authors declare no competing interests.

Data availability

Sequencing data have been submitted to the NCBI SRA database under BioProject PRJNA750039, BioSamples SAMN20428190 to SAMN20428209 and SRA Accession numbers SRR15275531 to SRR15275520.

References

1 Rajapaksha, P. *et al.* A review of methods for the detection of pathogenic microorganisms. *Analyst* **144**, 396-411, doi:10.1039/c8an01488d (2019).

2 Calabro, J. M. in *Nursery Management* (2018).

3 Daughtrey, M. L. Boxwood blight: threat to ornamentals. *Annual Review of Phytopathology* **57**, 189-209, doi:10.1146/annurev-phyto-082718-100156 (2019).

4 Castroagudín, V. L. *et al.* Boxwood blight disease: a diagnostic guide. *Plant Health Progress* **21**, 291-300, doi:10.1094/php-06-20-0053-dg (2020).

5 Gehesquiere, B. *et al.* Characterization and taxonomic reassessment of the box blight pathogen *Calonectria pseudonaviculata*, introducing *Calonectria henricotiae* sp. nov. *Plant Pathology* **doi:10.1111/ppa.12401** (2015).

6 Gehesquiere, B. *et al.* qPCR assays for the detection of *Cylindrocladium buxicola* in plant, water, and air samples. *Plant Disease* **97**, 1082-1090, doi:10.1094/pdis-10-12-0964-re (2013).

7 Healy, S. E. *Biology and management of box blight caused by Cylindrocladium buxicola*, The University of Guelph, (2014).

8 Guo, Y. & Pooler, M. Real-time and conventional PCR tools for detection and discrimination of *Calonectria pseudonaviculata* and *C. henricotiae* causing boxwood blight. *Plant Disease* **105**, 164-168, doi:10.1094/pdis-09-19-2053-re (2021).

9 Malapi-Wight, M., Demers, J. E., Veltri, D., Marra, R. E. & Crouch, J. A. LAMP detection assays for boxwood blight pathogens: a comparative genomics approach. *Scientific Reports* **doi:10.1038/srep26140** (2016).

10 Malapi-Wight, M. *et al.* Sarcococca blight: use of whole-genome sequencing for fungal plant disease diagnosis. *Plant Disease* **100**, 1093-1100 (2016).

11 Adams, I. P. *et al.* Next-generation sequencing and metagenomic analysis: a universal diagnostic tool in plant virology. *Molecular Plant Pathology* **10**, 537-545, doi: 10.1111/j.1364-3703.2009.00545.x (2009).

12 Miller, R. R., Montoya, V., Gardy, J. L., Patrick, D. M. & Tang, P. Metagenomics for pathogen detection in public health. *Genome Medicine* **5**, 81, doi:10.1186/gm485 (2013).

13 Gu, W., Miller, S. & Chiu, C. Y. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu Rev Pathol* **14**, 319-338, doi:10.1146/annurev-pathmechdis-012418-012751 (2019).

14 Piombo, E. *et al.* Metagenomics approaches for the detection and surveillance of emerging and recurrent plant pathogens. *Microorganisms* **9**, 188 (2021).

15 Schadt, E. E., Turner, S. & Kasarskis, A. A window into third-generation sequencing. *Human Molecular Genetics* **19**, R227-R240, doi:10.1093/hmg/ddq416 (2010).

16 Deng, X. *et al.* Metagenomic sequencing with spiked primer enrichment for viral diagnostics and genomic surveillance. *Nature Microbiology* **5**, 443-454, doi:10.1038/s41564-019-0637-9 (2020).

17 Mostafa, H. H. *et al.* Metagenomic next-generation sequencing of nasopharyngeal specimens collected from confirmed and suspect COVID-19 patients. *mBio* **11**, e01969-01920, doi:10.1128/mBio.01969-20 (2020).

18 Leggett, R. M. & Clark, M. D. A world of opportunities with nanopore sequencing. *Journal of Experimental Botany* **68**, 5419-5429, doi:10.1093/jxb/erx289 (2017).

19 Hu, Y. *et al.* Pathogen detection and microbiome analysis of infected wheat using a portable DNA sequencer. *Phytobiomes Journal* **3**, 92-101, doi:10.1094/pbiomes-01-19-0004-r (2019).

20 Chalupowicz, L. *et al.* Diagnosis of plant diseases using the Nanopore sequencing platform. *Plant Pathology* **68**, 229-238, doi:10.1111/ppa.12957 (2019).

21 Bronzato Badial, A. *et al.* Nanopore sequencing as a surveillance tool for plant pathogens in plant and insect tissues. *Plant Disease* **102**, 1648-1652, doi:10.1094/pdis-04-17-0488-re (2018).

22 Filloux, D. *et al.* Nanopore-based detection and characterization of yam viruses. *Scientific Reports* **8**, 17879, doi:10.1038/s41598-018-36042-7 (2018).

23 Mechan Llontop, M. E. *et al.* Strain-level identification of bacterial tomato pathogens directly from metagenomic sequences. *Phytopathology* **110**, 768-779, doi:10.1094/PHYTO-09-19-0351-R (2020).

24 Xu, R. *et al.* MinION Nanopore-based detection of *Clavibacter nebraskensis*, the corn Goss's wilt pathogen, and bacteriomic profiling of necrotic lesions of naturally-infected leaf samples. *PLoS One* **16**, e0245333, doi:10.1371/journal.pone.0245333 (2021).

25 Loit, K. *et al.* Relative performance of MinION (Oxford Nanopore Technologies) versus Sequel (Pacific Biosciences) third-generation sequencing instruments in identification of agricultural and forest fungal pathogens. *Appl Environ Microbiol* **85**, e01368-01319, doi:10.1128/AEM.01368-19 (2019).

26 *Native barcoding genomic DNA (with EXP-NBD104, EXP-NBD114, and SQK-LSK109) Protocol*, <https://community.nanoporetech.com/protocols/native-barcoding-genomic-dna/> (2019).

27 Andrews, S. *et al. FastQC: a quality control tool for high throughput sequence data*, <http://www.bioinformatics.babraham.ac.uk/projects/fastqc> (2010).

28 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

29 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Completing bacterial genome assemblies with multiplex MinION sequencing. *Microbial Genomics* **3**, doi: 10.1099/mgen.0.000132 (2017).

30 Camacho, C. *et al.* BLAST+: architecture and applications. *BMC bioinformatics* **10**, 421, doi:10.1186/1471-2105-10-421 (2009).

31 Dilthey, A. T., Jain, C., Koren, S. & Phillippy, A. M. Strain-level metagenomic assignment and compositional estimation for long reads with MetaMaps. *Nature Communications* **10**, 3066, doi:10.1038/s41467-019-10934-2 (2019).

32 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).

33 Raja, H. A., Miller, A. N., Pearce, C. J. & Oberlies, N. H. Fungal identification using molecular tools: a primer for the natural products research community. *J Nat Prod* **80**, 756-770, doi:10.1021/acs.jnatprod.6b01085 (2017).

34 Wood, D. E. & Salzberg, S. L. Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biol* **15**, R46, doi:10.1186/gb-2014-15-3-r46 (2014).

35 Leidenfrost, R. M., Pöther, D.-C., Jäckel, U. & Wünschiers, R. Benchmarking the MinION: Evaluating long reads for microbial profiling. *Scientific Reports* **10**, 5125, doi:10.1038/s41598-020-61989-x (2020).

36 Koren, S. *et al.* Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res*, doi:10.1101/gr.215087.116 (2017).

37 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075, doi:10.1093/bioinformatics/btt086 (2013).

38 Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness. *Methods in Molecular Biology* **1962**, 227-245, doi:10.1007/978-1-4939-9173-0_14 (2019).

39 Brown, C. T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *The Journal of Open Source Software* **1**, doi:10.21105/joss.00027 (2016).

40 Ondov, B. D., Bergman, N. H. & Phillippy, A. M. Interactive metagenomic visualization in a Web browser. *BMC bioinformatics* **12**, 385, doi:10.1186/1471-2105-12-385 (2011).

41 NCBI. *BLAST nt database*, <https://ftp.ncbi.nlm.nih.gov/blast/db/> .

42 Nicholls, S. M., Quick, J. C., Tang, S. & Loman, N. J. Ultra-deep, long-read nanopore sequencing of mock microbial community standards. *GigaScience* **8**, doi:10.1093/gigascience/giz043 (2019).

43 LeBlanc, N., Cubeta, M. A. & Crouch, J. A. Population genomics trace clonal diversification and intercontinental migration of an emerging fungal pathogen of boxwood. *Phytopathology* **111**, 184-193, doi:10.1094/PHYTO-06-20-0219-FI (2021).

44 Hu, Y. High quality DNA extraction from Fungi_small scale. *protocols.io*, doi:dx.doi.org/10.17504/protocols.io.exmbfk6 (2016).

45 Castroagudín, V. L. *et al.* One clonal lineage of *Calonectria pseudonaviculata* is primarily responsible for the boxwood blight epidemic in the United States. *Phytopathology* **110**, 1845-1853, doi:10.1094/PHYTO-04-20-0130-R (2020).

Tables

Table 1. Metadata and DNA quantity and quality of samples used in this study

Sample description	Sample ID	Sequencing Date	Extraction methods	Kit	Flowcell ID	Number of barcodes used per flowcell	DNA quantity and quality		
							DNA concentration (ng/µL)	A260/ A280	A260/ A230
Moderately diseased plant	S1	Nov22	Sonication	DNeasy® PowerWater®	FAK95928	4	317.0	1.90	1.98
	S2	Nov22	Sonication	DNeasy® PowerSoil® Pro			479.5	1.93	1.74
	S3	Nov22	Sonication	ZymoBIOMICS™ DNA Miniprep			403.5	1.92	1.96
	G1	Nov22	Grinding	ZymoBIOMICS™ DNA Miniprep			76.0	1.89	1.72
	G2	Nov24	Grinding	DNeasy® PowerPlant® Pro	FAK96453	5	163.1	1.66	0.74
	G3	Nov24	Grinding	Invisorb® Spin Plant Mini			103.0	1.71	0.57
	G4	Nov24	Grinding	OmniPrep™			203.1	1.73	0.69
	G5	Nov24	Grinding	OmniPrep™ with RNAse			277.8	1.81	1.95
	G6	Nov24	Grinding	Gentra® Puregene®			314.6	1.45	0.45
Severely diseased plant (Batch 1)	S4	Dec12	Sonication	DNeasy® PowerWater®	FAN08223	4	151.0	1.89	2.02
	S5	Dec12	Sonication	DNeasy® PowerSoil® Pro			98.0	1.90	0.83
	G7	Dec12	Grinding	DNeasy® PowerPlant® Pro			76.8	0.89	0.20

	G8	Dec12	Grinding	Invisorb® Spin Plant Mini			135.1	1.77	1.49
	G9	Dec17	Grinding	OmniPrep™	FAN08200	5	170.3	2.10	2.23
Severely diseased plant (Batch 2)	G10	Dec17	Grinding	OmniPrep™			1,132.5	2.10	1.88
	G11	Dec17	Grinding	OmniPrep™ diluted 10 times and treated with Rnase			110.1	1.96	0.49
	G12	Dec17	Grinding	ZymoBIOMICS™ DNA Miniprep			77.4	2.19	0.08
	G13	Dec17	Grinding	Gentra® Puregene®			349.0	1.57	0.55
Healthy plant	NC	Negative control	Grinding	ZymoBIOMICS™ DNA Miniprep	FAO99127	2	82.9	1.87	1.53

Table 2. Summary of ONT MinION sequencing data obtained in this study (see Table 1 for sample metadata)

ID	Total read length (Gbp)	Normalized read length per flow cell (Gbp)	Total number of reads	Normalized number of reads per flow cell	Average read length (bp)	Longest read length (bp)	Number of *Cps* hits >= 1000 bp (based on BLASTN)	Number of *Cps* hits >= 85% id (based on MetaMaps)	Total read length of *Cps* hits >= 1000 bp (Mbp; based on BLASTN)	*Cps* reads (based on BLASTN) out of total reads (%)	*Cps* (based on BLASTN) read length out of total read length (%)	*Cps* genome coverage (×)
S1	1.35	5.41	429,098	1,716,392	3,152	89,153	174	86	0.64	0.04	0.05	0.012
S2	1.92	7.68	475,383	1,901,532	4,040	91,256	166	87	0.75	0.03	0.04	0.014
S3	1.37	5.49	354,893	1,419,572	3,864	65,510	349	192	1.16	0.10	0.08	0.021
G1	1.30	5.21	711,491	2,845,964	1,830	54,580	6,382	3,548	17.33	0.90	1.33	0.315
G2	2.55	12.73	2,020,441	10,102,205	1,260	54,580	18,797	9,269	46.68	0.93	1.83	0.849
G3	1.97	9.84	1,965,416	9,827,080	1,001	88,418	8,528	4,056	19.82	0.43	1.01	0.360
G4	2.91	14.57	2,724,170	13,620,850	1,069	64,110	4,859	1,841	9.67	0.18	0.33	0.176
G5	4.56	22.88	3,843,496	19,217,480	1,190	72,917	9,977	4,271	20.98	0.26	0.46	0.381
G6	0.60	2.98	468,312	2,341,560	1,274	50,579	3,027	1,190	6.82	0.65	1.14	0.124
S4	4.90	19.59	2,430,505	9,722,020	2,014	64,015	3,681	1,700	9.52	0.15	0.19	0.173
S5	6.55	26.19	2,846,336	11,385,344	2,300	189,652	9,379	3,763	22.78	0.13	0.35	0.414
G7	1.14	4.58	1,399,778	5,599,112	817	42,603	11,987	8,255	36.71	0.86	3.21	0.668
G8	2.31	9.24	2,839,930	11,359,720	813	369,167	13,146	7,697	38.09	0.46	1.65	0.693
G9	1.14	5.69	298,982	1,494,910	3,804	77,811	14,484	10,343	46.02	4.84	4.05	0.837
G10	2.06	10.30	549,134	2,745,670	3,749	73,800	46,460	40,409	257.82	8.46	12.52	4.670
G11	0.92	4.61	292,084	1,460,420	3,154	40,849	27,566	22,881	110.97	9.44	12.04	2.019
G12	3.46	17.31	894,828	4,474,140	3,868	53,435	65,192	59,892	386.05	7.29	11.15	7.022
G13	0.74	3.70	280,861	1,404,305	2,633	91,251	11,526	9,677	67.72	4.10	9.15	1.232
NC	2.10	4.19	846,387	1,692,774	2,476	28,907	0	0	0	0.00	0.00	0

[a] Gbp, giga base pairs; Mbp, mega base pairs.

Table 3. Percentage of *Cps* based on Jaccard similarity obtained with sourmash and *Cps* hits obtained with BLASTN

Reference genome	Accession number	G10 *Cps* reads	G11 *Cps* reads	G12 *Cps* reads	Assembled *Cps* genome	
		Similarity (%) by sourmash	Similarity (%) by sourmash	Similarity (%) by sourmash	Similarity (%) by sourmash	Number of hits by BLASTN
CBS139395	GCA_004380915.1	17.47	19.11	14.21	72.27	621
CBS139394	GCA_001696505.1	17.30	18.93	14.01	72.65	125
CB002	GCA_006505905.1	17.11	18.72	13.83	72.53	3
	GCA_004141935.1	17.11	18.72	13.83	72.53	
CT13	GCA_004380985.1	16.65	18.26	13.36	72.12	217
CBS114417	GCA_004381005.1	16.51	18.10	13.23	72.02	43
ODA1	GCA_004382225.1	15.69	17.36	12.53	68.81	31
NC-BB1	GCA_004381035.1	13.85	15.54	11.02	61.20	5
ICMP14368	GCA_004382245.1	10.73	12.56	8.48	48.09	5

Table 4. Assembly summary of assembled *Cps* reads that were pre-identified by BLASTN in samples G10, G11 and G12, and of reference genome CBS139395

	Assembled *Cps*	CBS139395
Assembly size (bp)	49,048,547	54,975,240
Number of contigs	1,055	27
Maximum contig length (bp)	419,837	5,578,780
N50 contig length (bp)	88,131	3,534,399
GC content (%)	48.12	46.36
Total aligned length (bp)	48,291,239	NA
Genome fraction (%)	88.746	NA
[a]Assembly BUSCO coverage (%)	C:50.3; F:23.2; M:26.5	C:96.6; F:0.2; M:3.2

[a] For BUSCO coverage, C stands for complete BUSCOs, F stands for fragmented BUSCOs, and M stands for missing BUSCOs.

Figures

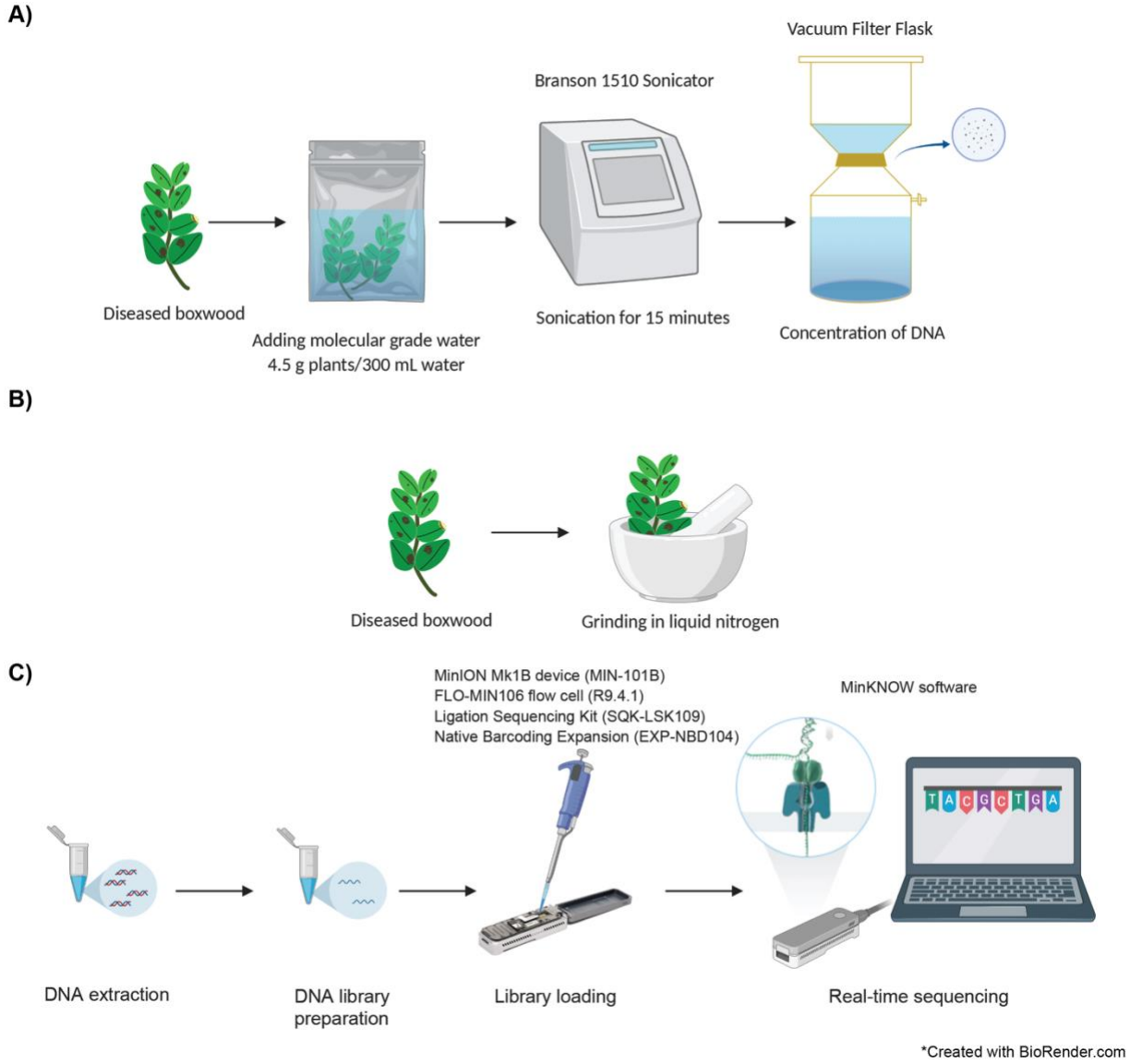


Figure 1. Pipelines for detection and identification of *Calonectria pseudonaviculata* (*Cps*). A) DNA extraction approach based on sonication without disrupting plant cells. B) DNA extraction approach based on homogenization in liquid nitrogen with disrupting plant cells. C) The MinION sequencing pipeline. Created with BioRender.com.

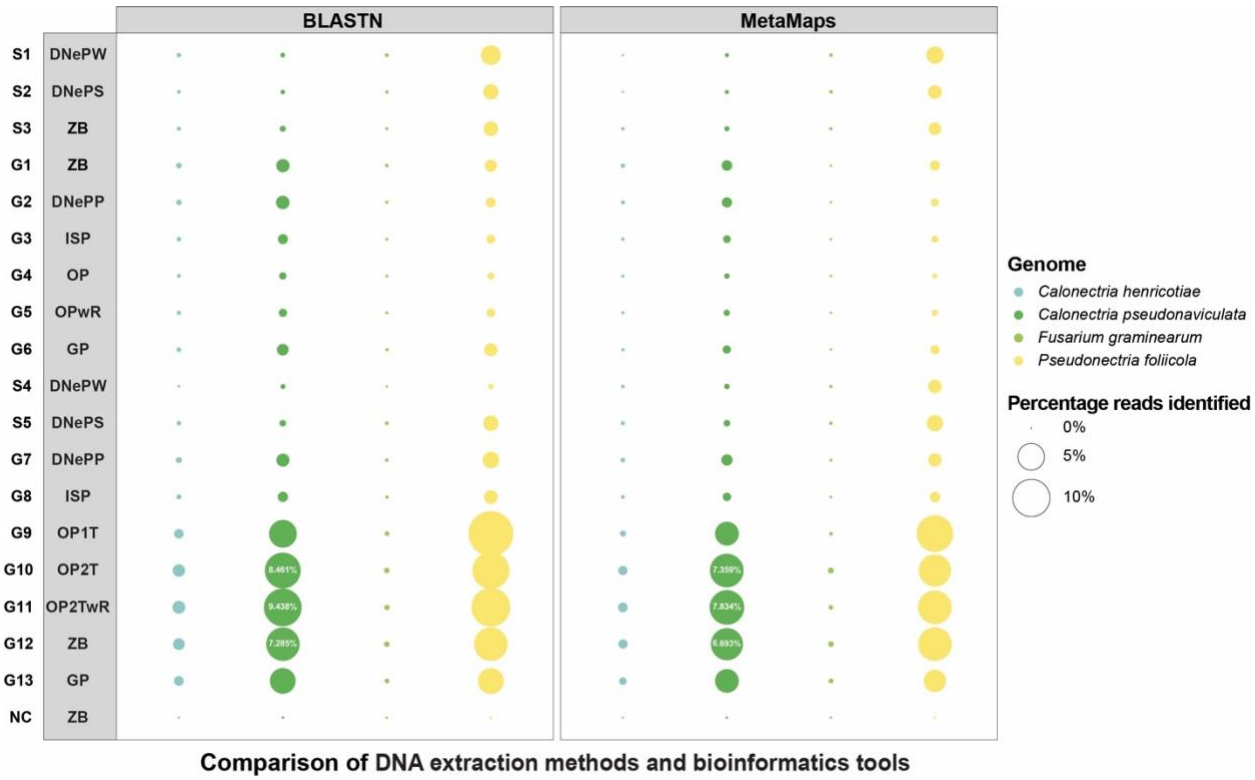


Figure 2. Bubble plot showing the percentage of sequencing reads assigned to four fungal species in each sequenced sample. The column on the left displays the sample IDs and the column to its right displays the abbreviations of DNA extraction kits (see Table 1). Bubble size is proportional to the percentage of reads assigned to the four species listed on the right based on the tools BLASTN and MetaMaps using a small fungal database containing one genome per fungal species.

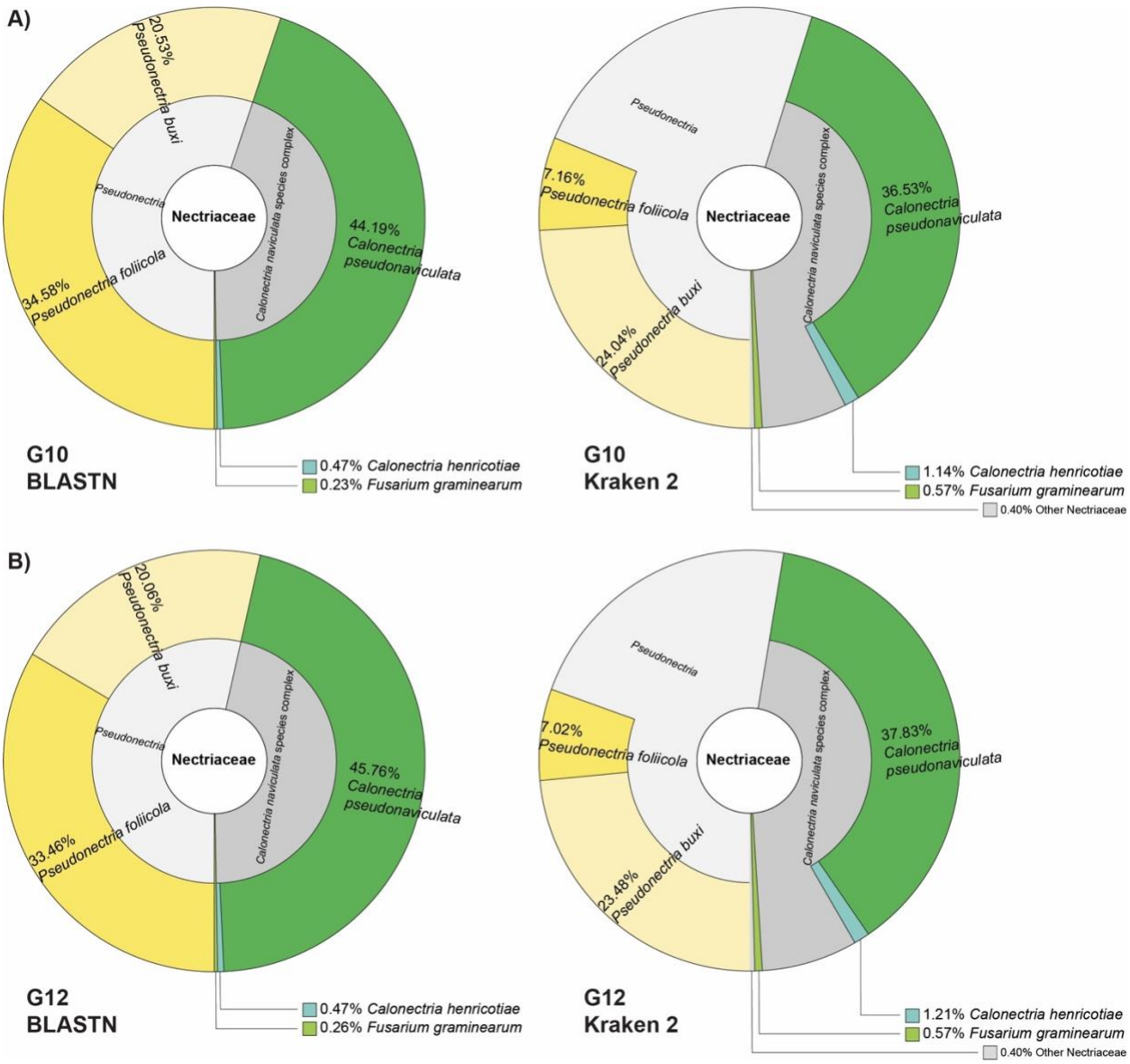


Figure 3. Krona plots showing the fraction of reads identified at the species, species complex, or genus rank as a percentage of the sequencing reads assigned to the family Nectriaceae using the tool Kraken 2 and a database of 29 genomes. The plots on the left display BLASTN results and the ones on the right Kraken 2 results. Each color represents a species, species complex, or genus. A) Results of G10, the sample processed by OmniPrep after homogenization in liquid nitrogen. B) Results of G12, the sample processed by ZymoBIOMICS DNA Miniprep Kit after homogenization in liquid nitrogen (See Table 1).

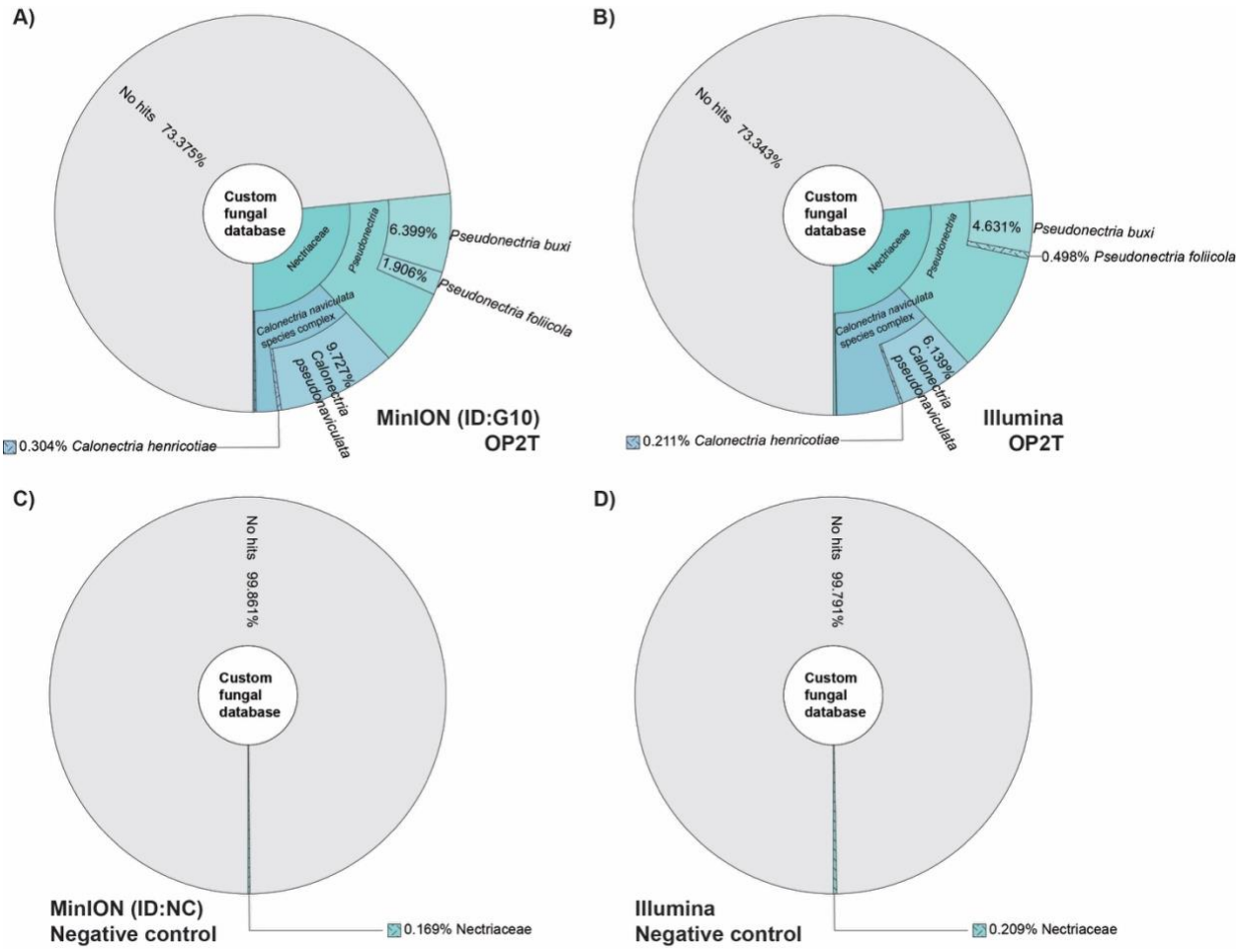


Figure 4. Krona plots showing the fraction of reads identified as members of the family Nectriaceae as a percentage of all sequencing reads using the tool Kraken 2 and a database of 29 genomes. Each color represents a clade. A) Results of G10 sequenced on the ONT MinION. B) Results of G10 sequenced on the Illumina HiSeq 3000 platform. C) Results of a healthy sample sequenced on the ONT MinION. D) Results of another healthy sample sequenced on the Illumina Nova Seq 6000 Platform.

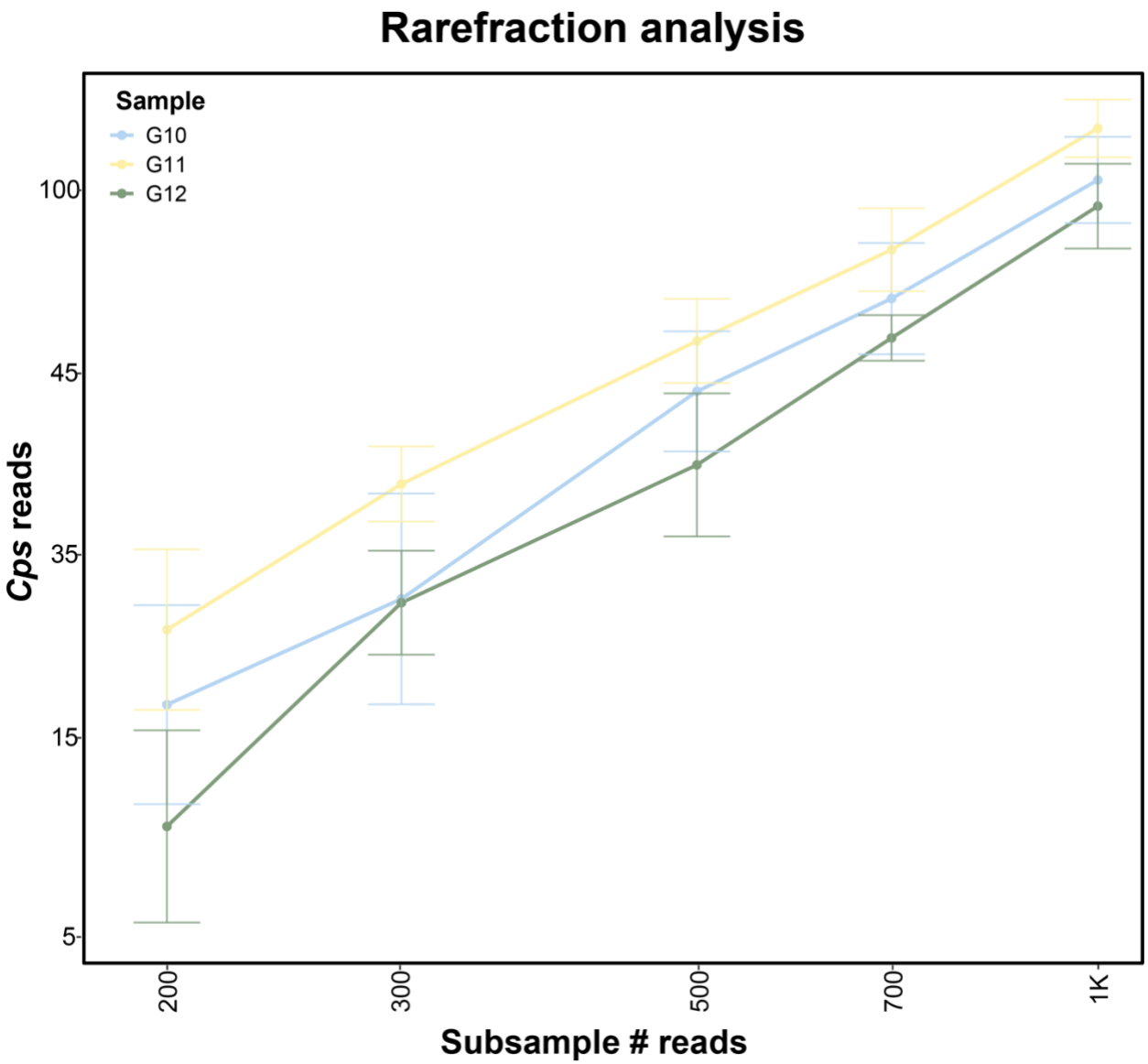


Figure 5. Detection limit analysis based on computational sub-sampling. Sub-samples were obtained by randomly extracting reads from original sequencing files. The X-axis shows the number of sub-sampled reads. The Y-axis shows the number of identified *Cps* reads. The circles represent the median value for each sub-sample size and error bars show the standard deviation among the 10 subsampling events.

Chapter 2. Literature review: fungal ice nucleation activity

Ice nucleation

Pure water does not freeze at the well-known melting point of water (0°C) but remains liquid as so-called “super-cooled” water until reaching -38°C (Koop et al. 2000). Below this temperature, ice crystals form spontaneously causing water to freeze. The process is called homogeneous ice nucleation. By contrast, in the process of heterogeneous ice nucleation, ice formation occurs at higher temperatures stimulated by so-called ice nuclei or ice nucleating particles (INPs). INPs are usually impurities present in water, which trigger the formation of ice crystals.

Ice nucleation activity (INA) is the capacity of INPs to catalyze ice formation at temperatures higher than the temperature at which pure water freezes. INPs directly impact atmospheric processes by affecting the ratio of frozen to liquid droplets in clouds, which in turn affects atmospheric radiative fluxes and, therefore, global warming (Matus and L'Ecuyer 2017) and the formation of precipitation (Murray et al. 2012). Since heterogeneous ice nucleation is the most common pathway for ice formation in the atmosphere (Niedermeier et al. 2011) and the number and type of INPs in the atmosphere affect cloud properties and precipitation (Yang H et al. 2019), it is important to study the most active INPs, which are of biological origin.

Ice-nucleating particles (INPs)

Ice-nucleating particles (INPs) come from various sources and are of different types. Primary atmospherically relevant INPs originate from natural sources like deserts, volcanic eruptions, and oceans. But INPs can also come from anthropogenic sources such as agricultural practices, air pollution caused by industrial processes and transportation, and deforestation. The INP types include mineral and desert dusts, metals and metal oxides, organics and glassy particles, biological

particles, soil dust, biomass and fossil fuel combustion aerosol particles, volcanic ash particles, and crystalline salts (Kanji et al. 2017). Non-biological INPs usually induce ice nucleation at temperatures < –15°C.

Biological INPs, which are usually associated with bacteria, fungi, pollen, and lichen, are very effective INPs and induce ice nucleation at temperatures ≥ –12°C (Lundheim and Zachariassen 1999) and can potentially affect cloud formation and precipitation (Morris et al. 2014; Joyce et al. 2019). Among them, bacterial INPs are the best characterized. The most active ones can induce INA at temperatures as high as –1°C (Morris et al. 2004).

Several bacterial species are known to have INA, including Gram-negative bacterium *Pseudomonas syringae* (Maki et al. 1974) and a few additional *Pseudomonas* species (Maki and Willoughby 1978; Anderson and Ashworth 1986), *Pantoea agglomerans* (synonym, *Erwinia herbicola*) (Lindow et al. 1978), *Pantoea ananatis* (synonyms, *Erwinia ananatis* and *Erwinia uredovora*)(Michigami et al. 1994) , and *Xanthomonas campestris* (Kim et al. 1987) as well as the Gram-positive bacterium *Lysinibacillus parviboronicapiens* (Failor et al. 2017).

Some of these ice nucleation-active (Ice^+) bacteria have been molecularly characterized. For Ice^+ bacteria that are Gram-negative, INA is usually conferred by a protein anchored to the outer membrane (Wolber et al. 1986; Morris et al. 2004). Some *Pantoea* and *Pseudomonas* strains secrete Ice^+ proteins that are associated with outer membrane vesicles (Phelps et al. 1986; Michigami et al. 1995; Šantl-Temkiv et al. 2015). In the Gram-positive bacterium *L. parviboronicapiens*, INPs are associated to a polyketide non-ribosomal peptide, and they are secreted, heat resistant, lysozyme, and proteinase resistant (Failor et al. 2017; Failor et al. 2021).

Less is known about INPs produced by other organisms. Pollen INPs are easily-suspendable, non-proteinaceous, and seem associated with polysaccharides (Pummer et al. 2012;

Dreischmeier et al. 2017). Fungal INPs have been poorly characterized so far. Studies have suggested that these particles may be proteinaceous (Lagzian et al. 2014; Pummer et al. 2015). On the other hand, urediniospores of rust fungi have also been found to have INA, but in this case INA may depend on a polysaccharide (Morris et al. 2013). INPs derived from viruses usually have a lower freezing temperature that can induce ice formation at about –20°C (Adams et al. 2021).

While INA is not as commonly associated with fungi as with bacteria, fungal INA can have the potential to impact the atmosphere. Fungi that have INA are widely present in soils over the globe. Previous studies have been proposed that soil can be a relevant source of atmospheric ice nucleating particles (INPs) (Schnell and Vali 1972; Schnell and Vali 1976; Conen et al. 2011). INPs could attach to soil particles and subsequently be released as aerosols (O'Sullivan et al. 2015). Also, spores of some fungal species that are found to have INA are disseminated by wind and are abundant in the atmosphere (Morris et al. 2013; Haga et al. 2014). Thus, it is worth investigating the characteristics of fungal INPs since such studies may provide useful information to better understand atmospheric processes.

Fungal ice nucleation activity (INA)

Fungal INA has been known only in a few ascomycotic and basidiomycotic fungal species (Pouleur et al. 1992; Hasegawa et al. 1994; Richard et al. 1996; Fröhlich-Nowoisky et al. 2015). It remains unknown why some fungi present INA. Possible reasons include that these fungi can cause frost damage in plants, thus acquiring nutrients from plants and/or initiating infection (Lindow 1983; Buttner and Amy 1989). Little is known about the characteristics of fungal INA molecules.

The genera *Fusarium* and the species *Mortierella alpina* are known to produce efficient ice nuclei, which induce ice nucleation at temperatures warmer than –12°C (Fröhlich-Nowoisky et al. 2015; Kunert et al. 2019). They have been the only fungi that have been reported repeatedly and consistently by different groups to have INA. Thus, *Fusarium* species and *M. alpina* are discussed and investigated in our study.

- ***Fusarium* species**

Fusarium species are filamentous fungi that are widely distributed in soils. Many of them are important plant pathogens and dispersed by wind. INA has been found mainly in *Fusarium acuminatum* and *Fusarium avenaceum* (Pouleur et al. 1992). Later, a few strains of *F. armeniacum*, *F. begoniae*, *F. concentricum*, *F. langsethiae*, *F. moniliforme*, *F. oxysporum*, and *F. tricinctum* have also been identified as Ice^+ (Tsumuki et al. 1995; Richard et al. 1996; Kunert et al. 2019). Although quite a few *Fusarium* species are known to have INA, the fungal INPs secreted by *Fusarium* species are still poorly characterized.

Some studies suggested that *Fusarium* INPs are proteinaceous compounds because of indirect evidence. For example, INA are lost after proteinase treatments (Tsumuki and Konno 1994), and a peak in the 280 nm UV absorbance has been observed for *Fusarium* INPs (O'Sullivan et al. 2015). Another study reported a predicted *Fusarium* gene to encode a protein that can confer INA in *E. coli* (Anastassopoulos 2001; Lagzian et al. 2014). However, it is not known if this gene is even expressed in *Fusarium* and this result has not been confirmed in any other studies. Most importantly, while this gene had been identified using the *ina* gene of *P. syringae* as a hybridization probe, this gene has no significant DNA sequence similarity to the *ina* gene in *P. syringae* and when the protein was tested for INA, a negative control was not included and detailed results were not reported. Thus, it cannot yet be concluded that *Fusarium* INPs are proteinaceous.

Physical and chemical characteristics of *Fusarium* INPs have been investigated as well. It was found that *Fusarium* INPs are smaller than 100 kDa in size, stable at pH levels from 2 to 12, tolerate temperature treatments up to 40-60°C, and have high stability under atmospherically relevant conditions (Pouleur et al. 1992; Hasegawa et al. 1994; Kunert et al. 2019). Also, *Fusarium* INPs appear to be aggregates of different size that can be separated into smaller sub-units with lower INA (Pouleur et al. 1992; Hasegawa et al. 1994; Kunert et al. 2019).

- ***Mortierella alpina***

Fungi in the Mortierellaceae family are also filamentous fungi that are widely distributed in soils. They are saprobes that are investigated in biotechnology as producers of polyunsaturated fatty acids (Shimizu and Yamada 1990; Shimizu et al. 1997; Certik et al. 1998; Certik and Shimizu 1999). *Mortierella alpina* is the only species in this family known to have INA so far in (Fröhlich-Nowoisky et al. 2015; Pummer et al. 2015). *M. alpina* INPs are even more poorly characterized than *Fusarium* INPs.

M. alpina INPs seem proteinaceous because INA is lost after proteinase treatments and after a chemical treatment that degrades proteins (Fröhlich-Nowoisky et al. 2015; Pummer et al. 2015). However, genes encoding *M. alpina* INPs have not been identified yet so it cannot be concluded yet if *M. alpina* INPs are proteinaceous or not. Since *M. alpina* is known for producing fatty acids, fatty acids may play a role in *M. alpina* INA, although long-chain fatty acids are less effective for inducing ice formation when warmer than –36°C (Qiu, Odendahl et al. 2017, DeMott, Mason et al. 2018).

Other known characteristics of *Mortierella* INPs include that they are smaller than 300 kDa in size and heat-sensitive (Fröhlich-Nowoisky et al. 2015; Vasebi et al. 2019). Concentrations of *M. alpina*-like INPs are higher in cold regions compared to warm regions (Conen and Yakutin

2018). A possible explanation is that *Mortierella* species are cold-adapted fungi and INA may be advantageous for them to overwinter and survive through protective extracellular freezing (Conen and Yakutin 2018). Thus, it is important to characterize INA of *M. alpina* and investigate if they play roles in the atmosphere and/or the soil.

Factors affecting biological ice nucleation activity (INA)

The expression of biological INA can be affected by a few factors, such as growth conditions and environmental conditions. Factors affecting bacterial INA are better understood, while only a few studies have investigated fungi.

Bacterial INA is dependent on the cell concentration. For example, *P. syringae* needs a certain number of cells to initiate INA at warmer temperatures, and the initial INA appears associated with intact cells (Maki et al. 1974). The efficiency of bacterial INA expression also differs in different growth phases (Deininger et al. 1988; Pooley and Brown 1991).

Growth media can affect the production of bacterial INPs as well. Nutrition starvation for nitrogen, phosphorus, sulfur, and iron have been found to enhance the expression of bacterial INPs (Nemecek-Marshall et al. 1993; Fall and Fall 1998). A much greater number of INPs is produced when bacteria are grown on agar media compared to growth in liquid media (Pooley and Brown 1991).

Environmental factors play an important role in affecting bacterial INA as well. Studying INA when bacteria are grown under environmental conditions that simulate the atmospheric environment can help understand the role of biological INPs in atmospheric processes. For example, when grown *in vitro*, low temperature exposure was found to induce INA in a few bacterial species (Rogers et al. 1987; Mueller et al. 1990; Nemecek-Marshall et al. 1993; Gurian-

Sherman and Lindow 1995; Fall and Fall 1998), but how temperature induces bacterial INA is not conclusive. On the contrary, UV radiation was found to inhibit bacterial INA. For example, *P. syringae* pv. *garcae* suffered a decline in INA after long exposure to UV radiation. Growth on plants may also affect INA. In fact, INA was found to be higher when *P. syringae* was grown on plants compared to when it was grown *in vitro* (O'Brien R and Lindow 1988).

Little is known about how environmental factors affect fungal INA. Only a few *Fusarium* strains have been investigated. Low temperature was found to enhance INA in some *Fusarium* strains but did not have any impact on others (Yanai et al. 1996; Humphreys et al. 2001). Cultivation may also affect *Fusarium* INA. An early study suggested that the threshold freezing temperature tends to increase with culture age (Richard et al. 1996).

There are some similarities and differences between factors affecting bacterial INA and fungal INA. Like bacteria, fungi grown on agar media have stronger INA than when they are grown in liquid media. However, INA was found to be lost in some *Fusarium* species after several subcultures when grown in liquid media (Tsumuki et al. 1995). This was never observed in bacteria. Also, unlike bacteria, nutrition starvation does not enhance fungal INA (Humphreys et al. 2001). Other Ice^+ fungi have not been investigated yet and factors affecting fungal INA have thus remained unclear. It is essential to examine how environmental factors affect INA in different fungal strains to better understand their possible roles in the atmosphere.

Genetics of biological ice nucleation activity (INA)

So far, genes encoding biological INA have only been identified in bacteria. Most of these genes encode a protein conferring the ice nucleation phenotype. These genes have been identified in species that belong to the genera *Pseudomonas* and *Pantoea* (formerly in the genus of *Erwinia*).

A single small region of DNA from *P. syringae*, named *inaZ*, was first identified to encode the ice nucleation phenotype (Green and Warren 1985; Orser et al. 1985). The *inaZ* gene can confer INA to *Escherichia coli* and restore this phenotype to *P. syringae* ice nucleation-inactive (Ice^-) strains. This gene sequence from *P. syringae* was then found to have homology to the INA gene *iceE* in *Pa. agglomerans* (synonym, *Erwinia herbicola*) (Warren Gareth and Corotto 1989). Another DNA fragment from another *Pseudomonas* species (*P. fluorescens*) can also confer INA to *E. coli* and was called *inaW* since it is related to the *inaZ* gene from *P. syringae* (Corotto et al. 1986; Warren G. et al. 1986). Then, *inaX* from *Xanthomonas campestris* pv. *translucens* was identified to encode an ice nucleation protein that has homology to *inaZ*, *inaW* and *iceE* (Zhao and Orser 1990). Later, *inaA* and *inaU* from *Pantoea ananatis* conferring INA was characterized, which also shares sequence similarity to the *ina* genes present in the *Pseudomonas* species (Abe et al. 1989; Michigami et al. 1994). Three more *ina* genes called *inaV*, *inaK* and *inaQ* from *P. syringae* were identified later on (Schmid et al. 1997; Li L et al. 2004; Li Q et al. 2012). All these genes are simply different alleles of the same gene and are all derived from the same ancestral sequence. They can thus be considered orthologues.

On the other hand, the Gram-positive bacterium *L. parviboronicapiens* employs a different mechanism to confer ice nucleation. The active molecule secreted that confers INA is heat resistant, lysozyme and proteinase resistant, suggesting it is not a protein (Failor et al. 2017). The gene conferring the ice nucleation phenotype was identified using a combination of comparative and functional genomics approaches comparing wide type strain and UV mutants. (Failor 2018). The gene is predicted to encode a polyketide non-ribosomal peptide (Failor et al. 2021).

While genes conferring fungal INA have not been identified yet, a predicted gene has been found in *F. acuminatum* (Lagzian et al. 2014). However, as mentioned above, it is not known if

this gene is even expressed in *Fusarium*. Also, besides that the predicted *ina* gene from *Fusarium* had no homology to the *P. syringae* ina gene, other sequences from *Fusarium* have not been found to have any homology to bacterial *ina* genes either (Hasegawa et al. 1994). These findings suggest that genes conferring fungal INA may not necessarily encode proteins but fungal INA may depend on a polyketide non-ribosomal peptide as found for *L. parviboronicapiens* (Failor et al. 2021). The objective of our study was to identify putative INA genes in *Fusarium* and *M. alpina*.

Comparative genomics

Comparative genomics is a discipline in computational biology that aims at identifying what is conserved and what is different among a group of genomes. (Nobrega and Pennacchio 2004). Differences may consist in the presence or absence of genes or in allelic differences of genes. Whole-genome sequencing typically uses next-generation sequencing platforms and third-generation sequencing platforms utilizing long read technologies are also available (Reuter et al. 2015). After sequencing, the core process in comparative genomics consists in alignment of DNA sequences by mapping nucleotides in one sequence onto another with the introduction of gaps (Hardison 2003). Thus, the matches between genomes can be described, which is important to study genome evolution and genome function. Comparative genomics approaches have been used in many studies of fungi to identify new genes and functional non-coding regions (Axelson-Fisk and Sunnerhagen 2006; Gasch 2007).

Comparative genomics approaches have been used in *Fusarium* to study pathogenicity. By comparing strains with narrow host range and strains with broad host range, mobile pathogenicity chromosomes have been identified (Ma et al. 2010). By comparing strains from the same species complex, genes contributing to the phenotypic variation and niche specialization

have been identified (Walkowiak et al. 2016). By comparing a set of strains from the same subspecies, a resistance gene accounting for plant disease resistance has been identified (Schmidt et al. 2016).

Comparative genomics approaches have been used for the Mortierellaceae family to study evolution (Massey and Garey 2007; Uehling et al. 2017; Vandepol et al. 2020). In these studies, strains from the Mortierellaceae family were compared with genomes at long phylogenetic distances. According to the Mortierellaceae phylogeny proposed by Vandepol et al. (2020), several former *Mortierella* species now belong to new genera.

Although genes encoding fungal INPs remain unknown, the ice nucleation phenotype varies among strains within the same *Fusarium* species and within the same *Mortierella* species. Thus, it should be possible to identify genes conferring INA using comparative genomics approaches by sequencing and comparing genomes of Ice^+ strains and Ice^- strains from the same fungal species.

Currently available *Fusarium* and *Mortierella* genomes

- ***Fusarium* species**

The sizes of assembled *Fusarium* genomes are between 34 Mb and 62 Mb according to the NCBI Assembly database (Kitts et al. 2016). By November 2021, 1,117 assembled *Fusarium* genomes are available. Most genomes are species of *F. graminearum* and *F. oxysporum*, and they are the most intensively studied *Fusarium* species. However, strains from these two species are not known to have INA (Kunert et al. 2019).

Other databases or sources are available online to study *Fusarium* species. For example, the Cyber-infrastructure for *Fusarium* (http://www.fusariumdb.org/) is publicly available (Geiser

et al. 2004). It consists of platforms supporting strain identification and phylogenetic analyses, archiving genomes from four species (*F. verticillioides*, *F. oxysporum* f. sp. *lycopersici*, *F. graminearum*, and *F. solani* f. sp. *batatas*), and an online community to share and preserve knowledge of *Fusarium* (Park et al. 2010). Still, these species are not known to have INA.

F. avenaceum is one of the very first *Fusarium* species reported to have INA and is better characterized than many other *Fusarium* species (Pouleur et al. 1992; Hasegawa et al. 1994; O'Sullivan et al. 2015; Kunert et al. 2019). Eleven *F. avenaceum* genomes available in the NCBI Assembly database (including the genome we sequenced and used as the reference for our study, F156N33 (Yang S et al. 2021)). Three genomes (Fa05001, FaLH03 and FaLH27) have been compared in one study, which suggested that *F. avenaceum* has the potential to produce various secondary metabolites including polyketides, non-ribosomal peptides, terpenes, alkaloids and indole-diterpenes (Lysøe et al. 2014). Since there is evidence that a polyketide non-ribosomal peptide is the molecule conferring INA in *L. parviboronicapiens*, this molecule may also play an important role in *Fusarium* INA. However, it is unknown if the other 10 *F. avenaceum* strains are Ice^+ or Ice^-. Thus, we could not use them to identify INA genes with comparative genomics approaches. Also, Fa05001 was the only genome that had been annotated, but not based on its transcriptomic data. Therefore, we sequenced, assembled, and annotated F156N33 with the genomic DNA and RNA data, which is an Ice^+ strain.

- ***Mortierella* species**

The sizes of assembled genomes in Mortierellaceae are also between 32 and 52 Mb according to the NCBI Assembly database (Kitts et al. 2016). By November 2021, 79 assembled genomes are available, and most of them used to belong to *Mortierella* but now in new genera. Currently, 28 *Mortierella* genomes are available.

M. alpina is the only species in Mortierellaceae known for INA (Fröhlich-Nowoisky et al. 2015; Pummer et al. 2015). Nine *M. alpina* genomes are available in the NCBI Assembly database (including the genome we sequenced and used as the reference for our study, LL118). The other 8 genomes were analyzed independently, but it is unknown if they are Ice^+ or Ice^- strains. The genome of ATCC#32222 was characterized to investigate the lipogenesis pathway and results suggested that this species can produce fatty acids and a complex mixture of glycerolipids, glycerophospholipids and sphingolipids (Wang et al. 2011). The genome of CCTCC M 207067 was sequenced to study arachidonic acid-rich oil production, but the genomic data was not used in this study (Nie et al. 2014). The draft genome of CDC-B6842 was obtained to increase our understanding of the molecular mechanisms at the basis of lipid production and metabolism (Etienne et al. 2014). The remaining five *M. alpina* genomes were used to resolve the Mortierellaceae phylogeny (Vandepol et al. 2020). Also, the five genomes used to construct phylogeny were the only genomes that had been annotated, but not based on their transcriptomic data. Therefore, we also sequenced, assembled, and annotated an Ice^+ strain (LL118) with genomic DNA and RNA data.

Transcriptomics analysis and RNA-seq

Transcriptome analysis is useful for characterizing the molecular basis of phenotypic variation in biology. Comparative transcriptomics has been used to identify changes in gene expression that correlate with phenotypic differences, thus unravelling the molecular pathways underlying complex traits (Rossouw et al. 2008; Cohen et al. 2010). Investigation of gene expression profiles in response to different conditions can provide an effective approach to identifying candidate genes that determine the phenotype of interest.

High-throughput RNA sequencing, RNA-seq, is one of the most common methods to analyze transcriptomes and to profile genome-wide gene expression, which can detect all transcripts in one sample, including messenger RNAs (mRNAs), long non-coding RNAs (lncRNAs) and small RNAs (sRNAs) (Hrdlickova et al. 2017). RNA-seq typically uses next-generation sequencing (NGS) platforms to sequence complementary DNA (cDNA) that is synthesized from an RNA sample. Sequencing reads are then mapped to a reference genome or transcriptome and assembled into a genomic feature of interest such as a gene, and the abundance of the feature is measured based on the number of reads (Oshlack et al. 2010). Differential expression analysis can be performed to identify genes that have changed significantly in abundance across experimental conditions. Gene regulatory network may be further analyzed and inferred.

So far, transcriptome analysis has not been performed to investigate the ice nucleation phenotype. Since it is known that fungal INA can either be induced or inhibited under certain conditions, it can be expected that genes encoding fungal INPs are expressed differently under the conditions either inducing or inhibiting INA. It should thus be possible to identify these genes using transcriptomics.

References

Abe K, Watabe S, Emori Y, Watanabe M, Arai S. 1989. An ice nucleation active gene of *Erwinia ananas*. Sequence similarity to those of *Pseudomonas* species and regions required for ice nucleation activity. FEBS letters. 258(2):297-300. eng.

Adams MP, Atanasova NS, Sofieva S, Ravantti J, Heikkinen A, Brasseur Z, Duplissy J, Bamford DH, Murray BJ. 2021. Ice nucleation by viruses and their potential for cloud glaciation. Biogeosciences. 18(14):4431-4444.

Anastassopoulos E. 2001. Συμβολή στη μελέτη της ευκαρυωτικής παγοπυρήνωσης: ανάπτυξη μεθόδου επιλογής φυτών καπνού (*Nicotiana tabacum* L.) ανθεκτικών στο πάγωμα και απομόνωση παγοπυρηνωτικού γονιδίου από τον μύκητα *Fusarium acumitatum*. University of Crete.

Anderson JA, Ashworth EN. 1986. The effects of streptomycin, desiccation, and UV radiation on ice nucleation by *Pseudomonas viridiflava*. Plant physiology. 80(4):956-960.

Axelson-Fisk M, Sunnerhagen P. 2006. Comparative genomics and gene finding in fungi. In: Sunnerhagen P, Piskur J, editors. Comparative Genomics: Using Fungi as Models. Berlin, Heidelberg: Springer Berlin Heidelberg; p. 1-28.

Buttner MP, Amy PS. 1989. Survival of ice nucleation-active and genetically engineered non-ice-nucleating *Pseudomonas syringae* strains after freezing. Appl Environ Microbiol. 55(7):1690-1694. eng.

Certik M, Sakuradani E, Shimizu S. 1998. Desaturase-defective fungal mutants: useful tools for the regulation and overproduction of polyunsaturated fatty acids. Trends in Biotechnology. 16:500-505.

Certik M, Shimizu S. 1999. Biosynthesis and regulation of microbial polyunsaturated fatty acid production. Journal of Bioscience and Bioengineering. 87(1):1-14.

Cohen D, Bogeat-Triboulot MB, Tisserant E, Balzergue S, Martin-Magniette ML, Lelandais G, Ningre N, Renou JP, Tamby JP, Le Thiec D et al. 2010. Comparative transcriptomics of drought responses in Populus: a meta-analysis of genome-wide expression profiling in mature leaves and root apices across two genotypes. BMC genomics. 11:630. eng.

Conen F, Morris CE, Leifeld J, Yakutin MV, Alewell C. 2011. Biological residues define the ice nucleation properties of soil dust. Atmos Chem Phys. 11(18):9643-9648.

Conen F, Yakutin MV. 2018. Soils rich in biological ice-nucleating particles abound in ice-nucleating macromolecules likely produced by fungi. Biogeosciences. 15(14):4381-4385.

Corotto LV, Wolber PK, Warren GJ. 1986. Ice nucleation activity of *Pseudomonas fluorescens*: mutagenesis, complementation analysis and identification of a gene product. EMBO J. 5(2):231-236. eng.

Deininger CA, Mueller GM, Wolber PK. 1988. Immunological characterization of ice nucleation proteins from *Pseudomonas syringae*, *Pseudomonas fluorescens*, and *Erwinia herbicola*. Journal of bacteriology. 170(2):669-675. eng.

Dreischmeier K, Budke C, Wiehemeier L, Kottke T, Koop T. 2017. Boreal pollen contain ice-nucleating as well as ice-binding 'antifreeze' polysaccharides. Scientific Reports. 7(1):41890.

Etienne KA, Chibucos MC, Su Q, Orvis J, Daugherty S, Ott S, Sengamalay NA, Fraser CM, Lockhart SR, Bruno VM. 2014. Draft genome sequence of *Mortierella alpina* isolate CDC-B6842. Genome Announc. 2(1):e01180-01113. eng.

Failor KC. 2018. Identification and characterization of ice nucleation active bacteria isolated from precipitation. Virginia Tech.

Failor KC, Liu H, Llontop MEM, LeBlanc S, Eckshtain-Levi N, Sharma P, Reed A, Yang S, Tian L, Lefevre C et al. 2021. Ice nucleation in a Gram-positive bacterium isolated from precipitation depends on a polyketide synthase and non-ribosomal peptide synthetase. The ISME Journal.

Failor KC, Schmale III DG, Vinatzer BA, Monteil CL. 2017. Ice nucleation active bacteria in precipitation are genetically diverse and nucleate ice by employing different mechanisms. The ISME Journal. 11(12):2740-2753.

Fall AL, Fall R. 1998. High-Level expression of ice nuclei in *Erwinia herbicola* is induced by phosphate starvation and low temperature. Current Microbiology. 36(6):370-376.

Fröhlich-Nowoisky J, Hill TCJ, Pummer BG, Yordanova P, Franc GD, Pöschl U. 2015. Ice nucleation activity in the widespread soil fungus *Mortierella alpina*. Biogeosciences. 12(4):1057-1071.

Gasch AP. 2007. Comparative genomics of the environmental stress response in ascomycete fungi. Yeast. 24(11):961-976.

Geiser DM, del Mar Jiménez-Gasco M, Kang S, Makalowska I, Veeraraghavan N, Ward TJ, Zhang N, Kuldau GA, O'Donnell K. 2004. FUSARIUM-ID v. 1.0: a DNA sequence database for identifying *Fusarium*. European Journal of Plant Pathology. 110(5):473-479.

Green RL, Warren GJ. 1985. Physical and functional repetition in a bacterial ice nucleation gene. Nature. 317(6038):645-648.

Gurian-Sherman D, Lindow SE. 1995. Differential effects of growth temperature on ice nuclei active at different temperatures that are produced by cells of *Pseudomonas syringae*. Cryobiology. 32(2):129-138.

Haga DI, Burrows SM, Iannone R, Wheeler MJ, Mason RH, Chen J, Polishchuk EA, Pöschl U, Bertram AK. 2014. Ice nucleation by fungal spores from the classes *Agaricomycetes*, *Ustilaginomycetes*, and *Eurotiomycetes*, and the effect on the atmospheric transport of these spores. Atmos Chem Phys. 14(16):8611-8630.

Hardison RC. 2003. Comparative genomics. PLOS Biology. 1(2):e58.

Hasegawa Y, Ishihara Y, Tokuyama T. 1994. Characteristics of ice-nucleation activity in *Fusarium avenaceum* IFO 7158. Bioscience, Biotechnology, and Biochemistry. 58(12):2273-2274.

Hrdlickova R, Toloue M, Tian B. 2017. RNA-Seq methods for transcriptome analysis. Wiley Interdiscip Rev RNA. 8(1):10.1002/wrna.1364. eng.

Humphreys TL, Castrillo LA, Lee MR. 2001. Sensitivity of partially purified ice nucleation activity of *Fusarium acuminatum* SRSF 616. Current Microbiology. 42(5):330-338.

Joyce RE, Lavender H, Farrar J, Werth JT, Weber CF, D'Andrilli J, Vaitilingom M, Christner BC. 2019. Biological ice-nucleating particles deposited year-round in subtropical precipitation. Appl Environ Microbiol. 85(23):e01567-01519. eng.

Kanji ZA, Ladino LA, Wex H, Boose Y, Burkert-Kohn M, Cziczo DJ, Krämer M. 2017. Overview of ice nucleating particles. Meteorological Monographs. 58:1.1-1.33.

Kim HK, Orser C, Lindow SE, Sands DC. 1987. *Xanthomonas campestris* pv. *translucens* strains active in ice nucleation. Plant disease. 71(11):994-997. eng.

Kitts PA, Church DM, Thibaud-Nissen F, Choi J, Hem V, Sapojnikov V, Smith RG, Tatusova T, Xiang C, Zherikov A et al. 2016. Assembly: a resource for assembled genomes at NCBI. Nucleic acids research. 44(D1):D73-D80. eng.

Koop T, Luo B, Tsias A, Peter T. 2000. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. Nature. 406(6796):611-614.

Kunert AT, Pöhlker ML, Tang K, Krevert CS, Wieder C, Speth KR, Hanson LE, Morris CE, Schmale III DG, Pöschl U et al. 2019. Macromolecular fungal ice nuclei in *Fusarium*: effects of physical and chemical processing. Biogeosciences. 16(23):4647-4659.

Lagzian M, Latifi AM, Bassami MR, Mirzaei M. 2014. An ice nucleation protein from *Fusarium acuminatum*: cloning, expression, biochemical characterization and computational modeling. Biotechnology Letters. 36(10):2043-2051.

Li L, Kang DG, Cha HJ. 2004. Functional display of foreign protein on surface of *Escherichia coli* using N-terminal domain of ice nucleation protein. Biotechnology and bioengineering. 85(2):214-221. eng.

Li Q, Yan Q, Chen J, He Y, Wang J, Zhang H, Yu Z, Li L. 2012. Molecular characterization of an ice nucleation protein variant (*InaQ*) from Pseudomonas syringae and the analysis of its transmembrane transport activity in *Escherichia coli*. Int J Biol Sci. 8(8):1097-1108. eng.

Lindow SE. 1983. The role of bacterial ice nucleation in frost injury to plants. Annual Review of Phytopathology. 21(1):363-384.

Lindow SE, Arny D, Upper C. 1978. *Erwinia herbicola*: a bacterial ice nucleus active in increasing frost injury to corn. Phytopathology. 68(3):523-527.

Lundheim R, Zachariassen K. 1999. Applications of biological ice nucleators. Biotechnological Applications of Cold-Adapted Organisms. Springer; p. 309-317.

Lysøe E, Harris LJ, Walkowiak S, Subramaniam R, Divon HH, Riiser ES, Llorens C, Gabaldón T, Kistler HC, Jonkers W et al. 2014. The genome of the generalist plant pathogen *Fusarium avenaceum* is enriched with genes involved in redox, signaling and secondary metabolism. PLoS One. 9(11):e112703-e112703. eng.

Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature. 464(7287):367-373.

Maki LR, Galyan EL, Chang-Chien MM, Caldwell DR. 1974. Ice nucleation induced by *Pseudomonas syringae*. Appl Microbiol. 28(3):456-459. eng.

Maki LR, Willoughby KJ. 1978. Bacteria as biogenic sources of freezing nuclei. Journal of Applied Meteorology. 17(7):1049-1053.

Massey SE, Garey JR. 2007. A comparative genomics analysis of codon reassignments reveals a link with mitochondrial proteome size and a mechanism of genetic code change via suppressor tRNAs. Journal of Molecular Evolution. 64(4):399-410.

Matus AV, L'Ecuyer TS. 2017. The role of cloud phase in Earth's radiation budget. Journal of Geophysical Research: Atmospheres. 122(5):2559-2578.

Michigami Y, Abe K, Iwabuchi K, Obata H, Arai S. 1995. Formation of ice nucleation-active vesicles in *Erwinia uredovora* at low temperature and transport of *inaU* molecules into shed vesicles. Bioscience, Biotechnology, and Biochemistry. 59(10):1996-1998.

Michigami Y, Watabe S, Abe K, Obata H, Arai S. 1994. Cloning and sequencing of an ice nucleation active gene of *Erwinia uredovora*. Bioscience, Biotechnology, and Biochemistry. 58(4):762-764.

Morris CE, Conen F, Alex Huffman J, Phillips V, Pöschl U, Sands DC. 2014. Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. Global Change Biology. 20(2):341-351.

Morris CE, Georgakopoulos DG, Sands DC. 2004. Ice nucleation active bacteria and their potential role in precipitation. J Phys IV France. 121:87-103.

Morris CE, Sands DC, Glaux C, Samsatly J, Asaad S, Moukahel AR, Gonçalves FLT, Bigg EK. 2013. Urediospores of rust fungi are ice nucleation active at > −10 °C and harbor ice nucleation active bacteria. Atmos Chem Phys. 13(8):4223-4233.

Mueller GM, Wolber PK, Warren GJ. 1990. Clustering of ice nucleation protein correlates with ice nucleation activity. Cryobiology. 27(4):416-422.

Murray BJ, O'Sullivan D, Atkinson JD, Webb ME. 2012. Ice nucleation by particles immersed in supercooled cloud droplets [10.1039/C2CS35200A]. Chemical Society Reviews. 41(19):6519-6554.

Nemecek-Marshall M, LaDuca R, Fall R. 1993. High-level expression of ice nuclei in a *Pseudomonas syringae* strain is induced by nutrient limitation and low temperature. Journal of bacteriology. 175(13):4062-4070.

Nie Z, Deng Z, Zhang A, Ji X, Huang H. 2014. Efficient arachidonic acid-rich oil production by *Mortierella alpina* through a three-stage fermentation strategy. Bioprocess and Biosystems Engineering. 37(3):505-511.

Niedermeier D, Shaw RA, Hartmann S, Wex H, Clauss T, Voigtländer J, Stratmann F. 2011. Heterogeneous ice nucleation: exploring the transition from stochastic to singular freezing behavior. Atmos Chem Phys. 11(16):8767-8775.

Nobrega MA, Pennacchio LA. 2004. Comparative genomic analysis as a tool for biological discovery. The Journal of physiology. 554(Pt 1):31-39. eng.

O'Brien R D, Lindow SE. 1988. Effect of plant species and environmental conditions on ice nucleation activity of *Pseudomonas syringae* on leaves. Applied and environmental microbiology. 54(9):2281-2286. eng.

O'Sullivan D, Murray BJ, Ross JF, Whale TF, Price HC, Atkinson JD, Umo NS, Webb ME. 2015. The relevance of nanoscale biological fragments for ice nucleation in clouds. Sci Rep. 5:8082. eng.

Orser C, Staskawicz BJ, Panopoulos NJ, Dahlbeck D, Lindow SE. 1985. Cloning and expression of bacterial ice nucleation genes in *Escherichia coli*. Journal of bacteriology. 164(1):359-366. eng.

Oshlack A, Robinson MD, Young MD. 2010. From RNA-seq reads to differential expression results. Genome Biol. 11(12):220-220. eng.

Park B, Park J, Cheong K-C, Choi J, Jung K, Kim D, Lee Y-H, Ward TJ, O'Donnell K, Geiser DM et al. 2010. Cyber infrastructure for *Fusarium* : three integrated platforms supporting strain identification, phylogenetics, comparative genomics and knowledge sharing. Nucleic Acids Research. 39(suppl_1):D640-D646.

Phelps P, Giddings TH, Prochoda M, Fall R. 1986. Release of cell-free ice nuclei by *Erwinia herbicola*. Journal of bacteriology. 167(2):496-502. eng.

Pooley L, Brown TA. 1991. Effects of culture conditions on expression of the ice nucleation phenotype of *Pseudomonas syringae*. FEMS Microbiology Letters. 77(2-3):229-232.

Pouleur S, Richard C, Martin JG, Antoun H. 1992. Ice nucleation activity in *Fusarium acuminatum* and *Fusarium avenaceum*. Appl Environ Microbiol. 58(9):2960-2964. eng.

Pummer BG, Bauer H, Bernardi J, Bleicher S, Grothe H. 2012. Suspendable macromolecules are responsible for ice nucleation activity of birch and conifer pollen. Atmos Chem Phys. 12(5):2541-2550.

Pummer BG, Budke C, Augustin-Bauditz S, Niedermeier D, Felgitsch L, Kampf CJ, Huber RG, Liedl KR, Loerting T, Moschen T et al. 2015. Ice nucleation by water-soluble macromolecules. Atmos Chem Phys. 15(8):4077-4091.

Reuter JA, Spacek DV, Snyder MP. 2015. High-throughput sequencing technologies. Mol Cell. 58(4):586-597. eng.

Richard C, Martin JG, Pouleur S. 1996. Ice nucleation activity identified in some phytopathogenic *Fusarium species*. Phytoprotection. 77(2):83-92. En.

Rogers JS, Stall RE, Burke MJ. 1987. Low-temperature conditioning of the ice nucleation active bacterium, *Erwinia herbicola*. Cryobiology. 24(3):270-279.

Rossouw D, Næs T, Bauer FF. 2008. Linking gene regulation and the exo-metabolome: A comparative transcriptomics approach to identify genes that impact on the production of volatile aroma compounds in yeast. BMC genomics. 9(1):530.

Šantl-Temkiv T, Sahyoun M, Finster K, Hartmann S, Augustin-Bauditz S, Stratmann F, Wex H, Clauss T, Nielsen NW, Sørensen JH et al. 2015. Characterization of airborne ice-nucleation-active bacteria and bacterial fragments. Atmospheric Environment. 109:105-117.

Schmid D, Pridmore D, Capitani G, Battistutta R, Neeser JR, Jann A. 1997. Molecular organisation of the ice nucleation protein *InaV* from *Pseudomonas syringae*. FEBS letters. 414(3):590-594. eng.

Schmidt SM, Lukasiewicz J, Farrer R, van Dam P, Bertoldo C, Rep M. 2016. Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the *Fom-2* resistance gene in melon. New Phytologist. 209(1):307-318.

Schnell RC, Vali G. 1972. Atmospheric ice nuclei from decomposing vegetation. Nature. 236(5343):163-165.

Schnell RC, Vali G. 1976. Biogenic ice nuclei: Part I. terrestrial and marine Sources. Journal of Atmospheric Sciences. 33(8):1554-1564. English.

Shimizu S, Ogawa J, Kataoka M, Kobayashi M. 1997. Screening of novel microbial enzymes for the production of biologically and chemically useful compounds. Advances in Biochemical Engineering/Biotechnology. 58:45-87. eng.

Shimizu S, Yamada H. 1990. Production of dietary and pharmacologically important polyunsaturated fatty acids by microbiological processes. Comments on Agricultural and Food Chemistry. 2(3):211-235.

Tsumuki H, Konno H. 1994. Ice nuclei produced by *Fusarium* sp. isolated from the gut of the rice stem borer, *Chilo suppressalis* Walker (Lepidoptera: Pyralidae). Bioscience, Biotechnology, and Biochemistry. 58(3):578-579.

Tsumuki H, Yanai H, Aoki T. 1995. Identification of ice-nucleating active fungus isolated from the gut of the rice stem borer, *Chilo suppressalis* Walker (Lepidoptera: Pyralidae) and a search for ice-nucleating active *Fusarium* species. Japanese Journal of Phytopathology. 61(4):334-339.

Uehling J, Gryganskyi A, Hameed K, Tschaplinski T, Misztal PK, Wu S, Desirò A, Vande Pol N, Du Z, Zienkiewicz A et al. 2017. Comparative genomics of *Mortierella elongata* and its

bacterial endosymbiont *Mycoavidus cysteinexigens*. Environmental Microbiology. 19(8):2964-2983.

Vandepol N, Liber J, Desirò A, Na H, Kennedy M, Barry K, Grigoriev IV, Miller AN, O'Donnell K, Stajich JE et al. 2020. Resolving the Mortierellaceae phylogeny through synthesis of multi-gene phylogenetics and phylogenomics. Fungal Diversity. 104(1):267-289.

Vasebi Y, Mechan Llontop ME, Hanlon R, Schmale III DG, Schnell R, Vinatzer BA. 2019. Comprehensive characterization of an aspen (*Populus tremuloides*) leaf litter sample that maintained ice nucleation activity for 48 years. Biogeosciences. 16(8):1675-1683.

Walkowiak S, Rowland O, Rodrigue N, Subramaniam R. 2016. Whole genome sequencing and comparative genomics of closely related Fusarium Head Blight fungi: *Fusarium graminearum*, *F. meridionale* and *F. asiaticum*. BMC genomics. 17(1):1014.

Wang L, Chen W, Feng Y, Ren Y, Gu Z, Chen H, Wang H, Thomas MJ, Zhang B, Berquin IM et al. 2011. Genome characterization of the oleaginous fungus *Mortierella alpina*. PLoS One. 6(12):e28319-e28319. eng.

Warren G, Corotto L. 1989. The consensus sequence of ice nucleation proteins from *Erwinia herbicola*, *Pseudomonas fluorescens* and *Pseudomonas syringae*. Gene. 85(1):239-242.

Warren G, Corotto L, Wolber P. 1986. Conserved repeats in diverged ice nucleation structural genes from two species of *Pseudomonas*. Nucleic acids research. 14(20):8047-8060. eng.

Wolber PK, Deininger CA, Southworth MW, Vandekerckhove J, van Montagu M, Warren GJ. 1986. Identification and purification of a bacterial ice-nucleation protein. Proc Natl Acad Sci U S A. 83(19):7256-7260. eng.

Yanai H, Tsumuki H, Konno H, Maeda T. 1996. Ice nucleus production of *Fusarium moniliforme* var. *subglutinans* in relation to its growth characteristics. Bioscience, Biotechnology, and Biochemistry. 60(9):1516-1518.

Yang H, Xiao H, Guo C. 2019. Effects of aerosols as ice nuclei on the dynamics, microphysics and precipitation of severe storm clouds. Atmosphere. 10(12).

Yang S, Coleman JJ, Vinatzer BA. 2021. Genome Resource: Draft genome of *Fusarium avenaceum*, strain F156N33, isolated from the atmosphere above Virginia and annotated based on RNA sequencing data. Plant Disease.

Zhao J, Orser CS. 1990. Conserved repetition in the ice nucleation gene *inaX* from *Xanthomonas campestris* pv. *translucens*. Molecular and General Genetics MGG. 223(1):163-166.

Chapter 3. Exploring the genetic basis of ice nucleation activity in *Fusarium avenaceum*

Shu Yang[1], Jeffrey J. Coleman[2], Boris A. Vinatzer[1]

[1] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

[2] Department of Entomology and Plant Pathology, Auburn University, Auburn, AL, USA

Corresponding author: B. A. Vinatzer, vinatzer@vt.edu

Keywords: fungi, ice nucleation activity, biological ice nucleating particles, *Fusarium avenaceum*, comparative genomics, comparative transcriptomics

Abstract

Ice nucleation activity (INA) is the capacity of some particles to catalyze ice formation at a higher temperature than the temperature at which pure water freezes (–38°C). INA has profound impacts on atmospheric processes occurring in clouds, such as the formation of precipitation and changes in radiative fluxes. Biological ice nucleating particles (INPs) are among the most efficient INPs and include bacteria, pollen, lichen, and fungi. However, little is known about most biological INPs and their genetic basis. So far, INA has only been found in a few fungi, including *Fusarium avenaceum*. INA was first discovered in *F. avenaceum* in 1992, but the INPs produced by *F. avenaceum* are still poorly characterized. To characterize INA in *Fusarium* and to investigate its genetic basis, *F. avenaceum* was examined in detail. Fourteen *F. avenaceum* strains were screened for INA. Three strains were found to be less ice nucleation-active, and *Fusarium* INPs appeared to consist of secreted aggregates. Therefore, whole genome sequencing and comparative genomic

studies were performed to identify putative genes at the basis of *Fusarium* INA. Since growth temperature was also found to affect *Fusarium* INA, gene expression at different growth temperatures was also compared to further pinpoint the *Fusarium* INA genes. Overall, a list of candidate genes has been selected, which helps to identify the *Fusarium* INA genes. After the *Fusarium* INA genes have been identified, it will be easier to determine the role of fungal INA in atmospheric processes that affect weather and climate on Earth.

Introduction

Homogeneous ice nucleation is the process by which pure water freezes at temperatures below –38°C (Koop et al. 2000). By contrast, in the process of heterogeneous ice nucleation, ice formation occurs at warmer temperatures catalyzed by ice nucleating particles (INPs) serving as ice nuclei. Ice nucleation activity (INA) describes the capacity of INPs to catalyze ice formation at temperatures higher than the temperature at which pure water freezes. INPs play an important role in atmospheric processes by affecting the ratio of frozen to liquid droplets in clouds, which in turn affects atmospheric radiative fluxes (Matus and L'Ecuyer 2017) and the formation of precipitation (Murray et al. 2012). Biological particles are very effective INPs and can usually induce ice formation at temperatures ≥ –12°C (Lundheim and Zachariassen 1999). However, little is known about how biological INPs affect atmospheric processes and most of their characteristics.

Biological INPs are known to be associated with bacteria, fungi, viruses, pollen, lichen, and marine organics amongst others (Lundheim and Zachariassen 1999; Adams et al. 2021). However, only molecules responsible for bacterial INA have been identified. To date, the bacterial INPs are either proteins associated with Gram-negative bacterial genera *Pseudomonas*, *Pantoea*

and *Xanthomonas* (Maki et al. 1974; Lindow et al. 1978; Maki and Willoughby 1978; Anderson and Ashworth 1986; Phelps et al. 1986; Wolber et al. 1986; Kim et al. 1987; Michigami et al. 1994; Michigami et al. 1995; Morris et al. 2004; Šantl-Temkiv et al. 2015), or a polyketide non-ribosomal peptide produced by a type I iterative polyketide synthase non-ribosomal peptide synthetase (PKS-NRPS) in the Gram-positive bacterium *Lysinibacillus parviboronicapiens* (Failor et al. 2017; Failor 2018; Failor et al. 2021). Fungal INA was first found in *Fusarium* species (Pouleur et al. 1992), and only a few other genera have been found to be ice nucleation-active (Ice^+), such as *Mortierella* and *Puccinia* (Morris et al. 2013; Fröhlich-Nowoisky et al. 2015). So far, fungal INPs are still poorly characterized.

Fusarium species are filamentous fungi, many of which are important pathogens of plants and animals (Nelson et al. 1994). They are widely distributed in soils, and some *Fusarium* species can also be dispersed by wind and found in atmospheric samples (Palmero et al. 2011; Schmale III et al. 2012; O'Sullivan et al. 2015). *F. avenaceum* was one of the very first *Fusarium* species reported to have INA and has been better characterized than other fungi (Pouleur et al. 1992; Hasegawa et al. 1994; O'Sullivan et al. 2015; Kunert et al. 2019). Although a few studies investigated the physical and chemical characteristics of *Fusarium* INPs, knowledge of their properties is still limited. The freezing temperature of *Fusarium* INPs can be as high as –2°C. These INPs are smaller than 100 kDa in size, consist of aggregates composed of smaller subunits, are stable at pH ranging from 2 to 12, tolerate heat treatments up to 40-60°C, and maintain their INA under atmospherically relevant conditions and after long-term storage (Pouleur et al. 1992; Hasegawa et al. 1994; Kunert et al. 2019). The identity of the molecules(s) produced by *Fusarium* causing this activity remains unknown.

Some studies suggested that *Fusarium* INPs are proteinaceous compounds since INA was lost after proteinase treatments (Tsumuki and Konno 1994) and a peak UV absorbance was observed at 280 nm (O'Sullivan et al. 2015). An INA gene was reported in *F. acuminatum* that was identified by DNA hybridization using the INA gene *inaZ* from *Pseudomonas syringae* pv. *syringae*, and it was successfully expressed in *E. coli* (Anastassopoulos 2001; Lagzian et al. 2014). However, no evidence indicates that it is expressed in *Fusarium*, and these results have not been confirmed in any other studies. Therefore, at this point, it cannot be excluded that *Fusarium* INA may depend on a PKS-NRPS gene cluster as in *L. parviboronicapiens* instead of being dependent on a protein.

A recent study investigated the distribution of INA within the genus *Fusarium*. About 16% of 112 strains showed INA above –12°C, and at least seven *Fusarium* species included strains with INA (Kunert et al. 2019). These results indicated that INA varies even within the same species, which should make it possible to identify putative INA genes in *Fusarium* using comparative genomics approaches as was done to identify putative INA genes in *L. parviboronicapiens* (Failor et al. 2017; Failor 2018; Failor et al. 2021).

If the INA genes in *Fusarium* were to be identified, it would contribute to our basic understanding of the process of ice nucleation and the potential role of *Fusarium* INPs in atmospheric processes. The products encoded by these genes may also be a candidate for industrial applications (Gurian-Sherman and Lindow 1992). Therefore, 14 *F. avenaceum* strains were tested for ice nucleation, and comparative genomics approaches were used to identify genes associated with INA. Since we observed that the strength of INA in *F. avenaceum* depends on the temperature at which *F. avenaceaum* is grown, differential expression analysis was also used to identify

putative INA genes. Finally, since INPs are secreted, it was determined if any of the identified genes either encoded secreted proteins or were part of PKS-NRPS gene clusters.

Material and Methods

Fungal strains

Twelve *Fusarium avenaceum* were obtained from the Agricultural Research Service (ARS) culture collection, two strain was provided by courtesy of David G. Schmale III (Virginia Tech), one of which was originally from Kansas State University. See Table 1 for details. All strains were grown on potato dextrose agar (PDA) prior to being processed.

INA measurements

To obtain cumulative ice nucleation spectra of each strain, 0.5 mg of mycelium was collected from the center of PDA plates and suspended in 1 mL of nuclease-free water. These primary suspensions were used to make dilution series in nuclease-free water from 10^{-1} to 10^{-5}. Droplet-freezing assays were performed using thirty drops of 20 µL volume of the primary suspension and each dilution. Drops were deposited on parafilm boats floating on the glycerol bath in a cooling thermostat (LAUDA Alpha Cooling Thermostat RA24). The water used to make dilutions served as negative control. INA was tested at –6°C, –7°C, –8°C, –9°C, –10°C, –11°C, and –12°C. Drops were incubated for 10 minutes at each temperature, and the number of drops that froze in each group was recorded. The entire assay starting from preparation of suspensions was repeated three times. The number cumulative ice nuclei (IN) per gram of fungus at each temperature was inferred using the method developed by Vali (1971) and described by Failor et al. (2017). Analysis of variance

(ANOVA) was performed using R (v4.0.4) at certain temperature to determine if there were statistically significant differences in INA among strains/treatments.

Characterization of INPs

F. avenaceum strain F156N33 was used to investigate the properties of INPs. After the strain was grown on PDA for 7 days at room temperature, a primary suspension was made by suspending 5 mg of mycelium from the center of each plate in 50 mL of nuclease-free water. Next, 49 mL of the primary suspension was passed through a 0.22-μm-pore-size filter (Millex-GP Syringe Filter, 0.22 μm) to obtain the 0.22 μm filtrate. Then, 20 mL of the 0.22 μm filtrate were passed through a 30-kDa-pore-size filter (Macrosep Advance Centrifugal Devices with Omega Membrane 30K) for 10 min at 5,000 rpm. This step separated INPs below approximately 5 nm in size since these INPs could pass through the 30kDa filter and ended up in the filtrate. INPs above approximately 5 nm in diameter were retained by the 30kDa filter, and could be resuspended from the filter and constituted the 30 kDa retentate. The retentate was resuspended from the filter membrane with 500 μL of nuclease-free water. In parallel, another 20 mL of the 0.22 μm filtrate were passed through another 30 kDa filter with the same centrifugation setting but 10 mL of nuclease-free water were added to the filter and the filters were centrifuged again. This washing step was performed a total of ten times. The final 30 kDa filter retentate was obtained by resuspending the retentate in 500 μL of nuclease-free water. The primary suspension, the 0.22 μm filtrate, the 30 kDa filtrate, the original 30 kDa retentate, the final 30 kDa retentate, the tenth filtrate, and 10^{-1} to 10^{-5} dilutions of each fraction were used to infer cumulative ice nucleation spectra.

The primary suspension and the 0.22 μm filtrate were also used for investigating the effect of storage at extreme low temperature. They were stored at –80°C for 30 days, 60 days, and 90

days, respectively, prior to INA tests. These suspensions and their 10^{-1} to 10^{-5} dilutions were used to infer cumulative ice nucleation spectra as described above.

Investigating the effect of growth conditions on INA

F. avenaceum strain F156N33 was also studied to investigate the effect of growth temperature. The strain was grown for about 30 days at 6°C, room temperature, or 28°C, respectively. For the effect of culture age, the strain was grown for 7 days, 14 days, 21 days, 28 days, and 35 days, respectively, always at room temperature. The primary suspension and 10^{-1} to 10^{-5} dilutions were made as described above for each of the treatments and were used to infer cumulative ice nucleation spectra as described above.

Genome and transcriptome sequencing and assembly

Genomic DNA of all 14 *F. avenaceum* strains was extracted from mycelium grown on PDA using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research). Total RNA of *F. avenaceum* strain F156N33 was extracted using the RNeasy® Plant Mini Kit (QIAGEN) after INA was confirmed to ensure INA genes were expressed. DNA and RNA were sequenced on an Illumina Nova Seq 6000 Platform at Novogene Corporation Inc. (Sacramento, CA). Low-quality reads and adapters were removed by the company. The quality of reads was checked using FastQC v0.11.9 (Andrews et al. 2010). The methods used for genome and transcriptome assembly have been described in detail by Yang et al. (2021).

Phylogenetic analyses

Assemblies of 14 *F. avenaceum* strains were used, and *F. tricinctum* strain INRA 104 (GenBank accession: OVTS00000000) served as the outgroup. Sequences of translation elongation factor 1-alpha (TEF-1α), RNA polymerase II largest subunit (RPB1), and RNA polymerase II second largest subunit (RPB2) were identified by BLASTN v2.10.0+ (Camacho et al. 2009) using a custom database containing sequences of these genes in other *Fusarium* species obtained from GenBank (TEF-1α: FFUJ_05795 from *F. fujikuroi*; RPB1: FFUJ_00736 from *F. fujikuroi*; RPB2: FFUJ_07996 from *F. fujikuroi*). Multiple sequence alignments for each of the three genes were performed using Clustal W (Thompson et al. 1994) in MEGA 7 (Kumar et al. 2016) with the default setting, and the resulting alignments were manually edited.

Phylogenetic analyses were performed using the Maximum Likelihood method in MEGA 7 (Kumar et al. 2016) for each of the three genes as well as a concatenated gene dataset. The best nucleotide substitution model was determined for each of the single genes and the combined dataset using MEGA 7 (Kumar et al. 2016). Maximum Likelihood trees for each of the single genes and the combined dataset were generated using the best nucleotide substitution model accordingly with 1000 bootstrap replications in MEGA 7 (Kumar et al. 2016).

Gene prediction and genome annotation

The assembled 14 genomes were annotated using the MAKER annotation pipeline (v3.01.03) (Campbell et al. 2014) with a combination of evidence-based methods and *ab initio* gene prediction as previously described (Yang et al. 2021). In brief, for each genome assembly, the previously assembled transcriptome of F156N33 by both Trinity and StringTie served as EST evidence, and the proteome of *Fusarium graminearum* (UniProt Proteomes accession: UP000070720), served as

protein homology evidence. *Ab initio* gene annotation were performed by SNAP v2013-02-16 (Korf 2004) and AUGUSTUS v3.4.0 (Stanke et al. 2008) afterwards.

Functional annotation, prediction of the signal peptide and prediction of secondary metabolite genes

Functional annotation was performed using InterProScan v5.46-81.0 for the presence of Pfam domains with terms from the Gene Ontology (Jones et al. 2014). BLASTP from BLAST v2.10.0+ (Camacho et al. 2009) was also used to find regions of local similarity against the February 2021 release of the Swiss-Prot database (Consortium 2020). Prediction of the signal peptide was performed by SignalP v5.0b (Almagro Armenteros et al. 2019). Phyre2 was used to predict protein function based on the homology of predicted protein structures with the structure of proteins with known function (Kelley et al. 2015). Prediction of metabolic gene clusters was performed by the fungal version of antiSMASH 6 (Blin et al. 2021).

Pan-genome analysis and prediction of orthologues

Pan-genome analysis and searching for orthologous genes in the genome of 14 *F. avenaceum* strains was performed by GET_HOMOLOGUES-EST v3.4.2 (Contreras-Moreira and Vinuesa 2013), which clustered homologous gene families using the OrthoMCL v1.4 (Li et al. 2003) clustering algorithm. Genes present in Ice^+ strains and Ice^- strains were identified using the script (parse_pangenome_matrix.pl) included in this program.

Gene expression analysis

As described above, RNA-seq reads from *F. avenaceum* strain F156N33 grown at 6°C (3 replicates) and room temperature (2 replicates) were obtained. Each replicate was aligned to the genome assembly of F156N33 using STAR v2.7.8a (Dobin et al. 2013), generating five alignment files in BAM format. These five BAM files were subjected to featureCounts v2.0.1 (Liao et al. 2014) with parameters -p -B -C (multi-mapped reads excluded) as well as -p -B -O -M (multi-mapped reads included) to determine the number of reads mapped to each gene. The read counts were normalized by DESeq2 package in R (Love et al. 2014). Thus, differential expressed genes (DEGs) were identified with the following parameters: "padj (adjusted P value) < 0.05 and log2FoldChange > 1 using DESeq2.

DNAseq and variant calling and analyses

F. avenaceum strain F156N33 served as the reference genome for DNAseq, which allowed us to determine presence and absence of each gene of strain F156N33 in each of the 13 other genomes based on read alignment to independently confirm the GET_HOMOLOGUES-EST v3.4.2 (Contreras-Moreira and Vinuesa 2013) results. DNA reads were mapped on the reference genome using BWA-MEM2 v2.2.1 (Vasimuddin et al. 2019). The mapping quality was assessed by Qualimap v2.2.2 (Okonechnikov et al. 2016) and the number of reads mapped to each gene was determined by featureCounts v2.0.1 (Liao et al. 2014) with parameters -p -B -O -M. The alignment files were processed by Genome Analysis Toolkit v4.0 (GATK) (McKenna et al. 2010). GATK 'SortSam' and 'MarkDuplicates (Picard)' were used to remove duplicated mapping reads.

Variants were called by GATK 'HaplotypeCaller' (Poplin et al. 2018), followed by extracting SNPs using GATK 'SelectVariants'and filtering SNPs using GATK4'VariantFiltration'. Variants with a quality depth less than 2 ("QD < 2.0"), a quality score

less than 30 ("QUAL < 30"), a strand odds ratio larger than 3.0 ("SOR > 3.0"), a Fisher score larger than 60 ("FS > 60.0") and a root mean square of the mapping quality a quality less than 40.0 ("MQ < 40.0") were filtered out (GATK 2021). Base quality score recalibration was applied to filtered variants using GATK 'ApplyBQSR'. The resulting variants were called and filtered for the second time using the same programs and parameters described above. The filtered SNPs were annotated by SnpEff v5.0e (Cingolani et al. 2012).

Results and Discussion

INA varies within *F. avenaceum*

To determine the distribution of INA within the species *F. avenaceum*, cumulative ice nucleation spectra were obtained for 14 strains. All strains presented INA to some extent. Thirteen strains started to freeze around –6°C to –7°C, while freezing of *F. avenaceum* NRRL 54396 was only observed at –9°C and below (Fig. 1). Among the 13 strains that started to induce freezing at –6°C to –7°C, two of them (NRRL 13826 and NRRL 36457), showed significantly lower activity than the other 11 strains that started freezing at –6°C to –7°C (p-value = 3.73E-05 at –6°C). However, the 11 most active strains still varied significantly based on the inferred cumulative number of IN per gram of mycelium with 4 strains (NRRL 66272, NRRL 13316, NRRL 54754, and F156N33) having very similar and higher INA at the lowest tested temperatures compared to the other 7 strains. In regards to the cumulative number of IN produced per gram of mycelium, the 11 most active strains produced 10^7 to 10^{10} IN/g at –8°C and below with *F. avenaceum* NRRL 66272 showing the absolute highest number of IN/g.

Because INA was detected in all strains with the number of IN/g of mycelium varying gradually between strains, we conclude that INA in *F. avenaceum* is rather a quantitative trait than

a qualitative trait. The strength of INA may depend on the presence or absence of several INA genes and/or be affected by allelic differences in one or several INA genes.

F. avenaceum INPs appear to be secreted aggregates that are prone to be separated by centrifugatin and washing, and they are stable at –80°C

To investigate the properties of *F. avenaceum* INPs, filtration was performed for the primary suspension of strain F156N33 using filters with two different pore sizes (0.22 μm and 5 nm, the approximate pore size of a 30kDa filter). The 0.22 μm filtrate showed high INA with approximately 10^6 IN per gram of mycelium at –7°C and reaching 10^8 IN/g at –12°C. This is only a tenfold reduction compared with the primary mycelial suspension of strain F156N33.

The number of IN/g was reduced approximately 10,000 to 100,000-fold at –7°C (p-value = 0.0381) and –8°C (p-value = 0.0392) and still 1,000-fold at –9°C to –12°C (for example, p-value = 0.0066 at –9°C) compared with the primary suspension after passing through a 30 kDa filter (Fig. 2A). The retentate collected from the 30 kDa filter was tenfold reduced at –7°C (p-value = 0.0152) and –8°C (p-value = 0.0921) compared with the 0.22 μm filtrate but presented INA similar to the 0.22 μm filtrate at lower temperatures (for example, p-value = 0.3500 at –9°C). This suggests that *F. avenaceum* INPs are secreted from mycelial cells and that most *F. avenaceum* INPs have a diameter of at least 5nm corresponding to a mass above 30 kDa. This is in agreement with earlier results by Kunert et al. (2019) who found *Fusarium* INPs to have a mass of over 100kDa.

Intriguingly, the combined 30 kDa retentate and the 30 kDa filtrate had higher INA than that of the primary suspension, in particular at –12°C, at which temperature we found 10^{14} IN/g. This suggests that during filtration larger *Fusarium* INPs may have separated into smaller INPs. To follow up on this hypothesis, we washed the 30kDa filter 10 times. In other words, after the

first centrifugation, water was added to the filter and centrifugation was repeated. This process was then performed another 9 times. Each filtrate was tested for INA and the final retentate was tested for INA as well. Even after 10 washes, the cumulative number of IN per gram of original mycelium for the 30 kDa retentate was still 10^7/g at –12°C, only approximately 10-fold lower compared to the first, unwashed 30 kDa retentate (p-value = 0.2410) (Fig. 2A). Moreover, the filtrate from the tenth wash still showed similar INA to the final, washed 30 kDa retentate (p-value = 0.0835). This suggests that INPs even smaller than 5nm are generated from the original INPs during each washing and centrifugation step. Therefore, INPs produced by *F. avenaceum* appear to consist of aggregates that can be separated into smaller units and *F. avenaceum* INPs thus rather consist of relatively small proteins, non-ribosomal peptides, or polyketides that interact with each other forming the relatively large INPs that initially do not pass through the 30kDa filter. While Kunert et al. (2019) also concluded that *Fusarium* INPs consist of aggregates, here we show that the smallest aggregates that retain INP may be even smaller than what they suggested based on their experiments using a 100kDa filter.

INA stability after storage at extreme low temperature was also investigated. A temperature of –80°C did not affect INA significantly (for example, p-value = 0.9880 at –8°C), and *Fusarium* INPs retained high INA even after 90 days of storage (Fig. 2B). This result shows that *Fusarium* INPs can persist for a long time without losing INA at extreme low temperature, making it possible for them to have a lasting effect in the atmosphere as suggested previously by Kunert et al. (2019).

Growth temperature affects *Fusarium* INA, while the length of growth time does not have a significant impact

To investigate how growth temperature affects INA in *F. avenaceum*, strain F156N33 was grown at different temperatures for about 30 days prior to determining cumulative IN spectra. Observing the mycelia by eye, growth temperature affected their morphology (Fig. S1). Although growth was slow at 6°C, the mycelium covered the entre plates after 30 days of growth at 6°C. When instead grown at 28°C, the mycelium never covered more than half the plates.

In terms of INA, INA was induced when grown at 6°C compared to growth at room temperature while it was reduced after growth at 28°C (Fig. 3A). More precisely, the cumulative number of IN/g of mycelium at –12°C was around 10^{13} when grown at 6°C, 10^{9} when grown at room temperature, and only 10^{6} when grown at 28°C (p-value = 0.0015). Previous studies on bacterial INA also reported that low temperature (15°C) induce INA in *Pseudomonas syringae* although even lower temperature (9°C) inhibited INA. How temperature induces bacterial INA was not shown conclusively (Mueller et al. 1990; Gurian-Sherman and Lindow 1995). We hypothesize that INA in *F. avenaceum* may be higher at lower temperatures because the expression of INA genes may be induced at lower temperatures and be repressed at higher temperatures. The higher expression of INA genes may then lead to a higher production of INPs. On the other hand, it is also possible that the structure or size of INPs that forms at lower growth temperatures is different from the structure or size of INPs formed at higher temperatures. Finally, post-translational factors could be involved that alter the surface of INPs at lower growth temperatures.

We also noted that INA varied among replicates when grown at 28°C but was relatively stable among replicates grown at 6°C and room temperature (Fig. 3A). This result suggests that expression of the underlying INA gene(s) may also be unstable at 28°C.

The impact of culture age was also investigated using *F. avenaceum* F156N33 grown for 35 days at room temperature. While no significant change in INA was observed between the 7-

day, 14-day, and 21-day time points, the number of cumulative IN/g dropped approximately tenfold after 28 and 35 days of growth, in particular at temperatures from –8°C to –12°C, although not significantly (for example, p-value = 0.3190 at –8°C) (Fig. 2B). At –6°C, INA was inconsistent among replicates making it challenging to compare INA between cultures of different length. This result suggests that *F. avenaceum* continues to produce INPs as long as the mycelium grows and that INPs may start degrading when growth stops sometimes after 21 days of culture.

Phylogenetic analyses reveal that *F. avenaceum* strains form several within-species clusters and that the strength of INA does not correlate with phylogeny

Whole genome sequencing and genome assembly were performed for all 14 *F. avenaceum* strains. The genome coverage of the assemblies ranged from 49× to 61×. Assembly sizes ranged from 36.8 Mb to 49.7 Mb, and the G+C content ranged from 48.20 % to 48.50% with one exception (50.72% for strain NRRL 54396). The BUSCO quality assessment was based on the lineage-specific profile library hypocreales_odb10 (4,494 genes) and revealed that more than 97.6% of genes were present in all 14 assemblies, indicating a high quality of genome assembly for all strains. The assembly statistics are shown in Table 2.

All 14 strains were identified as *F. avenaceum* based on BLASTN. Sequences of translation elongation factor 1-alpha (TEF-1α), RNA polymerase II largest subunit (RPB1), and RNA polymerase II second largest subunit (RPB2) were extracted from each of the 14 assemblies to perform phylogenetic analyses. Since *F. tricinctum* is a closely related species, the genome of reference strain *F. tricinctum*, INRA104, was chosen as the outgroup.

All *F. avenaceum* strains formed one large clade separated from the outgroup *F. tricinctum*. Three to four major sub-clades formed, depending on the genes that were used (Fig. 4). The TEF-

1α gene sequences had the lowest rate of variability and the lowest phylogenetic resolution with six strains having an identical sequence and forming a single clade with one additional strain (Fig. 4A) while the RPB1-based ML tree and the multilocus sequence tree had the highest rate of variability (Fig. 4B and 4C).

When comparing the trees with the geographic origin of strains and their substrates of isolation (see Table 1 and Fig. 4), strains did not cluster together based on where and what they were isolated from. Also, although the three strains, NRRL 13826, NRRL 36457, and NRRL 54396, with the relatively lowest INA were phylogenetically distinct from our reference strain F156N33 according all four ML trees, they clustered together with other strains with high INA, for example, strain NRRL 66272.

In summary, based on the four genes used for phylogenetic analysis, the strength of INA in *F. avenaceum* does correlate with phylogeny. This could be either due to convergent evolution with multiple independent gene acquisitions (of the same or different INA genes) or independent mutations (in the same or different INA genes) that increase INA, or due to multiple gene loss events or multiple mutations that lead to a decrease in INA.

Putative INA genes in strain *F. avenaceum* F156N33

F. avenaceum F156N33 was annotated using its RNA-seq data. Thus, it served as the reference genome in our study. 11,233 genes were predicted in F156N33. Since we hypothesize that *F. avenaceum* INPs are either secreted protein aggregates or consist in products of PKS–NRPS gene clusters, a list of putative INA genes was obtained: 1,155 genes were predicted to encode proteins with signal peptides and 59 genes were predicted to belong to PKS-NRPS gene clusters. Therefore, a total of 1,214 genes are putative INA genes in F156N33 (Supplementary Table 1).

Comparative genomics approach to reduce the number of putative INA genes in *F. avenaceum*

To reduce the number of putative INA genes in *F. avenaceum*, we performed a pan-genome analysis. To do this, orthologous groups were identified for all 14 strains and genes were clustered into orthologous groups. One predicted transcript (Gene ID: KAF25_002500) was skipped by the program because it was longer than 25,000 bp. 554 genes were identified as redundant isoforms. Among 10,678 orthologue groups, 8,210 orthologues were found to be present in all 14 strains.

We first hypothesized that one or more INA genes may be present in strain F156N33 and the other 3 most active strains (NRRL 13316, NRRL 54754, and NRRL 66272) and absent from the least active strain NRRL 54396, ignoring all strains with intermediate INA. 82 genes were so identified (Supplementary Table 2).

Based on phenotypic results, *Fusarium* INPs were likely to be secreted molecules. Therefore, we assumed that *Fusarium* INPs were secreted proteins. Ten genes among these 82 genes were predicted to encode proteins with signal peptides. Four of these genes had no annotation or were annotated as proteins of unknown function based on InterProScan or BLASTP. Five genes were found to encode enzymes, and most of them belonged to glycosyl hydrolase families. One gene was found to be a hydrophobic surface binding protein A, annotated as cell wall mannoprotein 1 by BLASTP. Accodring to Phyre2, the 4 unknow proteins were predicted to encode enzymes, 2 of which were hydrolases (Supplementary Table 3).

Since polyketide non-ribosomal peptides are a kind of biological INP (Failor et al. 2021) and PKS-NRPS gene clusters are involved in the biosynthesis of mycotoxins in *Fusarium* (Desjardins and Proctor 2007), we assumed that PKS-NRPS gene clusters could be also involved

in producing INPs in *Fusarium*. One gene among these 82 genes was predicted to be a PKS-NRPS gene, and another 3 genes fell into one PKS-NRPS cluster. The PKS-NRPS gene (Gene ID: KAF25_005439) was predicted to encode an alcohol dehydrogenase, and the cluster this gene belonged to was most similar to the fusaridione A biosynthetic gene cluster from *Fusarium heterosporum*. The cluster the other 3 genes belonged to was most similar to the bikaverin biosynthetic gene cluster from *Fusarium fujikuroi* IMI 58289, and these 3 genes were predicted to encode either enzymes or a protein of unknown function.

Next, we tested a slightly different hypothesis. We assumed again that the strain with the lowest INA, strain NRRL 54396, was truly INA negative and had no INA gene but that INA genes were present in all 11 strains with either high or medium activity. Only 23 genes were so identified (Supplementary Table 4), and they were a subset of the above 82 genes in Supplementary Table 2. Only 2 genes were predicted to encode proteins with signal peptides; one (Gene ID: KAF25_008711) was a protein of unknown function by InterProScan or BLASTP and was predicted to encode a hydrolase (cellulose-binding protein) by Phyre2. Another gene (Gene ID: gene-KAF25_007828) encoded the hydrophobic surface binding protein A mentioned above.

Finally, we tested the hypothesis that some INA genes may be present in all but the three strains with the lowest INA. Only one gene (Gene ID: KAF25_004243) met this criterium based on the pan-genome analysis. However, the probability for this gene encoding a secreted peptide was only 0.0854% and it was not predicted to be a biosynthetic gene. Therefore, this gene is unlikely to be a candidate INA gene.

It is also important to point out that the described results of the orthologue-based pangenome analysis are mostly in agreement with the results from a read-based DNAseq analysis whereby reads of each strain were aligned against the annotated genome of F156N33 and

presence/absence was determined based on the number of reads that aligned to each F156N33 gene (Supplementary Table 6). While some the genes found to be absent based on the orthologue analysis had some aligned reads, this number was generally much lower than the number of reads that aligned to genes found to be present in the orthologue-based analysis.

In conclusion, most genes present in the most active strains but absent from the least active strain(s) encoded enzymes with predicted function. We think that such enzymes are unlikely to form aggregates that have INA. However, one or more of the genes encoding proteins that have signal peptides but are of unknown function remain promising INA gene candidates. It will be interesting to predict their structure and determine if any of them may possess characteristics that allow them to form aggregates and/or that have other characteristics predicted to make surfaces ice nucleation-active (Lin et al. 2018). Also, the genes that are part of biosynthetic gene clusters need to be pursued further as putative *F. avenaceum* INA genes and the products of these biosynthetic gene clusters need to be predicted and compared to the predicted polyketide non-ribosomal peptide of *L. parviboronicapiens* (Failor et al. 2021).

It is also possible that one or more of the 8,210 genes present in all 14 strains can be responsible for INA and that the difference in INA is due to allelic differences in some of these genes. Or, as we will investigate next, INA genes may be present in all strains but be expressed to a significantly higher level at low temperatures when INA was found to be highest compared to higher temperatures, when INA was found to be lower.

Transcriptomics approach to reduce the number of putative INA genes in *F. avenaceum*

Based on our phenotyping results, INA in *F. avenaceum* F156N33 was higher at 6°C than at room temperature and compared to 28°C. Therefore, we hypothesized that among the putative genes in

strain F156N33, the genes that were more highly expressed at 6°C than at room temperature, were more likely to be INA genes than the genes that were expressed higher at room temperature than at 6°C or the genes that were equally expressed at both temperatures. We could not include gene expression data at 28°C since the slow growth of *F. avenaceum* at 28°C did not allow us to extract enough RNA for sequencing.

Differential expression (DE) analyses were performed for uniquely mapped reads only as well as by including multi-mapped reads. Multi-mapped reads did not make much of a difference. In fact, 1,451 genes were found upregulated at 6°C for uniquely mapped reads only, and 1,536 genes were found when multi-mapped reads were included. In total, 1,630 genes were upregulated at 6°C (Supplementary Table 1).

Similarly to the comparative genomics analysis above, we looked next at which of the 1,630 differentially regulated genes were either predicted to encode proteins that contained signal peptides or that fell within biosynthetic gene clusters.

318 genes among the 1,630 genes were predicted to encode proteins with signal peptides (Fig. 5A). 98 of these genes had no annotation or are uncharacterized based on InterProScan or BLASTP. 178 genes were predicted to encode enzymes, so they were unlikely to be candidate INA genes. Most of these enzymes were predicted hydrolases. The remaining 42 genes included genes encoding site-specific binding proteins, genes encoding specific domains, virulence factors, and genes involved in production of molecules such as antifungal proteins, mycotoxins and fungal hydrophobins.

One gene coding for a secreted hydrophobin, annotated by BLASTP as Rodlet protein, is particularly interesting. Phyre2 found this protein to align with high confidence over most of its length with the Roda hydrophobin of *Aspergillus fumigatus*, the Hyd1 hydrophobin from

Schizophyllum commune, and the hydrophobin Mpg1 from the rice blast fungus *Magnaporthe oryzae*. These hydrophobins have in common that they self-assemble into large aggregates at hydrophobic:hydrophilic interfaces. These aggregates represent amyloid-like structures and are called "rodlets" (Kwan et al. 2006). However, the structure of the formed aggregate appears difficult o predict. Therefore, we hypothesize that an aggregate of secreted *F. avenaceum* hydrophobin molecules could potentially form a particle that induces ice formation on its surface.

According to Phyre2, 89 out of the 98 unknown genes mentioned above had predictions (Supplementary Table 5). 26 genes were predicted to encode enzymes, 13 of which encoded hydrolases. Interestingly, the putative INA gene of *F. acuminatum* reported in previous studies to encode a protein with INA when expressed in *E. coli* (Anastassopoulos 2001; Lagzian et al. 2014) belongs to the glycoside hydrolase family 16 based on InterProScan. However, this gene was originally identified using the *P. syringae* INA gene as a hybridization probe although the gene has no actual homology to the *P. syringae* INA gene. Moreover, the reported INA test results for this protein did not include a negative control, a cumulative IN spectrum was not reported, nor were results repeated independently. Therefore, it is highly unlikely that this protein has INA.

The remaining 63 genes found here were involved in cell transport, transcription, binding, or production of certain proteins such as mycotoxins, hormones, and antimicrobial proteins. However, some predictions had very low confidence, so functions of these genes are uncertain. On the other hand, 103 genes among 1,630 genes upregulated at 6°C were predicted to fall into 31 PKS-NRPS clusters (Fig. 5B). 21 out of these 103 genes were predicted to be PKS-NRPS genes distributed among 14 clusters. Many of these 21 genes were predicted to encode AMP-binding enzymes, and there were a few predicted to encode dehydrogenases. 7 genes (Gene IDs: KAF25_000207, KAF25_001383, KAF25_001960, KAF25_002505, KAF25_003892,

KAF25_008718, KAF25_008732) were associated with polyketide synthases or polyketide synthase dehydratases.

In summary, many genes found to be upregulated at 6°C in *F. avenaceum* are predicted to encode metabolic enzymes. Therefore, they are unlikely candidate INA genes. However, other genes were found to be either part of biosynthetic gene clusters that are predicted to produce polyketide non-ribosomal peptides, that encode secreted proteins of yet unknown function, or in one case, encode a predicted secreted hydrophobin, a class of proteins known to self-assemble into larger aggregates. All of these genes thus represent putative INA genes and will deserve further investigation.

Combining comparative genomics and transcriptomics results to reduce the list of putative INA genes in *F. avenaceum*

We finally compared the list of putative INA genes based on the comparative genomics and the transcriptomics results. If we assume that INA genes are absent from the least active strain NRRL 54396 and present in the 4 strains with the highest INA (F156N33, NRRL 13316, NRRL 54754, and NRRL 66272), only one gene is predicted to encode a protein with a signal peptide upregulated at 6°C (Gene ID: KAF25_006087) and no gene was found to be located within PKS-NRPS clusters and upregulated at 6°C. Since the gene KAF25_006087 is predicted to encode a peptidase/proteinase, it is unlikely to be a promising INA gene candidate.

If assuming INA genes are absent from the least active strain NRRL 54396 and present in the 11 active strains, none of the genes is predicted to encode proteins with a signal peptide or to be located within PKS-NRPS clusters and found to be upregulated at 6°C.

Conclusions

Here we found that although all *F. avenaceum* strains available to us had INA, the strength of INA varied among strains. This suggests that either the presence or absence of several genes contribute to the strength of INA in *F. avenaceum* or that allelic differences in one or more INA genes affect INA. Moreover, INA in *F. avenaceum* is associated with secreted aggregates that appear to consist of subunits as small as 5nm in diameter and that are stable at –80°C. INA in *F. avenaceum* is higher at lower temperatures suggesting that expression of INA genes may be induced at lower temperatures and be repressed at higher temperatures. Comparing the gene content of strains with different strengths of INA and the expression of genes at different temperatures, we have obtained a list of putative INA genes that either encode secreted proteins or that are located in biosynthetic gene clusters. However, genes outside of this list cannot be ruled out because allelic differences in genes present in all strains may be the main determinants of INA in *F. avenaceum*. Unfortunately, fourteen strains are not sufficient to test this hypothesis by performing a genome-wide association study (GWAS) (Tam et al. 2019) and we do not have access to any additional strains at this point. Therefore, to ultimately identify the INA genes in *F. avenaceum*, either more strains will need to be analyzed to conduct GWAS and/or candidate INA genes need to be confirmed experimentally through either mutational analysis in *F. avenaceum* or gain-of-function experiments expressing them in *Fusarium* species that do not have INA.

Once the INA genes in *F. avenaceum* have been identified, INA genes in other *Fusarium* species may be identified by homology. The identified INA genes can then be searched in metagenomic sequences from various environments to determine the ecological role of INA in *Fusarium*, in particular, it will be possible to investigate if *Fusarium* INA genes are present in the atmosphere and contribute to atmospheric processes. Finally, identifying the *Fusarium* INA genes

and their products will contribute to our basic understanding of the process of INA itself, which is still poorly understood (Coluzza et al. 2017).

References

Adams MP, Atanasova NS, Sofieva S, Ravantti J, Heikkinen A, Brasseur Z, Duplissy J, Bamford DH, Murray BJ. 2021. Ice nucleation by viruses and their potential for cloud glaciation. Biogeosciences. 18(14):4431-4444.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 37(4):420-423.

Anastassopoulos E. 2001. Συμβολή στη μελέτη της ευκαρυωτικής παγοπυρήνωσης: ανάπτυξη μεθόδου επιλογής φυτών καπνού (*Nicotiana tabacum* L.) ανθεκτικών στο πάγωμα και απομόνωση παγοπυρηνωτικού γονιδίου από τον μύκητα *Fusarium acumitatum*. University of Crete.

Anderson JA, Ashworth EN. 1986. The effects of streptomycin, desiccation, and UV radiation on ice nucleation by *Pseudomonas viridiflava*. Plant physiology. 80(4):956-960.

Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Institute, Babraham, UK. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 49(W1):W29-W35.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics. 10(1):421.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Current Protocols in Bioinformatics. 48:4.11.11-14.11.39. eng.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin). 6(2):80-92. eng.

Coluzza I, Creamean J, Rossi MJ, Wex H, Alpert PA, Bianco V, Boose Y, Dellago C, Felgitsch L, Fröhlich-Nowoisky J et al. 2017. Perspectives on the future of Ice nucleation research: research needs and unanswered questions identified from two international workshops. Atmosphere. 8(8):138.

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79(24):7696-7701. eng.

Desjardins AE, Proctor RH. 2007. Molecular biology of *Fusarium* mycotoxins. International Journal of Food Microbiology. 119(1):47-50.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21. eng.

Failor KC. 2018. Identification and characterization of ice nucleation active bacteria isolated from precipitation. Virginia Tech.

Failor KC, Liu H, Llontop MEM, LeBlanc S, Eckshtain-Levi N, Sharma P, Reed A, Yang S, Tian L, Lefevre C et al. 2021. Ice nucleation in a Gram-positive bacterium isolated from precipitation depends on a polyketide synthase and non-ribosomal peptide synthetase. The ISME Journal.

Failor KC, Schmale III DG, Vinatzer BA, Monteil CL. 2017. Ice nucleation active bacteria in precipitation are genetically diverse and nucleate ice by employing different mechanisms. The ISME Journal. 11(12):2740-2753.

Fröhlich-Nowoisky J, Hill TCJ, Pummer BG, Yordanova P, Franc GD, Pöschl U. 2015. Ice nucleation activity in the widespread soil fungus *Mortierella alpina*. Biogeosciences. 12(4):1057-1071.

GATK. 2021. (How to) Filter variants either with VQSR or by hard-filtering. https://gatk.broadinstitute.org/hc/en-us/articles/360035531112--How-to-Filter-variants-either-with-VQSR-or-by-hard-filtering.

Gurian-Sherman D, Lindow SE. 1992. Ice nucleation and its application. Current Opinion in Biotechnology. 3(3):303-306.

Gurian-Sherman D, Lindow SE. 1995. Differential effects of growth temperature on ice nuclei active at different temperatures that are produced by cells of *Pseudomonas syringae*. Cryobiology. 32(2):129-138.

Hasegawa Y, Ishihara Y, Tokuyama T. 1994. Characteristics of ice-nucleation activity in *Fusarium avenaceum* IFO 7158. Bioscience, Biotechnology, and Biochemistry. 58(12):2273-2274.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236-1240. eng.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 10(6):845-858.

Kim HK, Orser C, Lindow SE, Sands DC. 1987. *Xanthomonas campestris* pv. *translucens* strains active in ice nucleation. Plant disease. 71(11):994-997. eng.

Koop T, Luo B, Tsias A, Peter T. 2000. Water activity as the determinant for homogeneous ice nucleation in aqueous solutions. Nature. 406(6796):611-614.

Korf I. 2004. Gene finding in novel genomes. BMC bioinformatics. 5(1):59.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 33(7):1870-1874. eng.

Kunert AT, Pöhlker ML, Tang K, Krevert CS, Wieder C, Speth KR, Hanson LE, Morris CE, Schmale III DG, Pöschl U et al. 2019. Macromolecular fungal ice nuclei in *Fusarium*: effects of physical and chemical processing. Biogeosciences. 16(23):4647-4659.

Kwan AHY, Winefield RD, Sunde M, Matthews JM, Haverkamp RG, Templeton MD, Mackay JP. 2006. Structural basis for rodlet assembly in fungal hydrophobins. Proc Natl Acad Sci U S A. 103(10):3621.

Lagzian M, Latifi AM, Bassami MR, Mirzaei M. 2014. An ice nucleation protein from *Fusarium acuminatum*: cloning, expression, biochemical characterization and computational modeling. Biotechnology Letters. 36(10):2043-2051.

Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13(9):2178-2189. eng.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30(7):923-930.

Lin C, Corem G, Godsi O, Alexandrowicz G, Darling GR, Hodgson A. 2018. Ice nucleation on a corrugated surface. Journal of the American Chemical Society. 140(46):15804-15811.

Lindow SE, Arny D, Upper C. 1978. *Erwinia herbicola*: a bacterial ice nucleus active in increasing frost injury to corn. Phytopathology. 68(3):523-527.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550.

Lundheim R, Zachariassen K. 1999. Applications of biological ice nucleators. Biotechnological Applications of Cold-Adapted Organisms. Springer; p. 309-317.

Maki LR, Galyan EL, Chang-Chien MM, Caldwell DR. 1974. Ice nucleation induced by *Pseudomonas syringae*. Appl Microbiol. 28(3):456-459. eng.

Maki LR, Willoughby KJ. 1978. Bacteria as biogenic sources of freezing nuclei. Journal of Applied Meteorology. 17(7):1049-1053.

Matus AV, L'Ecuyer TS. 2017. The role of cloud phase in Earth's radiation budget. Journal of Geophysical Research: Atmospheres. 122(5):2559-2578.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20(9):1297-1303. eng.

Michigami Y, Abe K, Iwabuchi K, Obata H, Arai S. 1995. Formation of ice nucleation-active vesicles in *Erwinia uredovora* at low temperature and transport of *inaU* molecules into shed vesicles. Bioscience, Biotechnology, and Biochemistry. 59(10):1996-1998.

Michigami Y, Watabe S, Abe K, Obata H, Arai S. 1994. Cloning and sequencing of an ice nucleation active gene of *Erwinia uredovora*. Bioscience, Biotechnology, and Biochemistry. 58(4):762-764.

Morris CE, Georgakopoulos DG, Sands DC. 2004. Ice nucleation active bacteria and their potential role in precipitation. J Phys IV France. 121:87-103.

Morris CE, Sands DC, Glaux C, Samsatly J, Asaad S, Moukahel AR, Gonçalves FLT, Bigg EK. 2013. Urediospores of rust fungi are ice nucleation active at > −10 °C and harbor ice nucleation active bacteria. Atmos Chem Phys. 13(8):4223-4233.

Mueller GM, Wolber PK, Warren GJ. 1990. Clustering of ice nucleation protein correlates with ice nucleation activity. Cryobiology. 27(4):416-422.

Murray BJ, O'Sullivan D, Atkinson JD, Webb ME. 2012. Ice nucleation by particles immersed in supercooled cloud droplets [10.1039/C2CS35200A]. Chemical Society Reviews. 41(19):6519-6554.

Nelson PE, Dignani MC, Anaissie EJ. 1994. Taxonomy, biology, and clinical aspects of Fusarium species. Clinical Microbiology Reviews. 7(4):479-504.

O'Sullivan D, Murray BJ, Ross JF, Whale TF, Price HC, Atkinson JD, Umo NS, Webb ME. 2015. The relevance of nanoscale biological fragments for ice nucleation in clouds. Sci Rep. 5:8082. eng.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. Bioinformatics. 32(2):292-294. eng.

Palmero D, Rodríguez JM, de Cara M, Camacho F, Iglesias C, Tello JC. 2011. Fungal microbiota from rain water and pathogenicity of *Fusarium* species isolated from atmospheric dust and rainfall dust. Journal of Industrial Microbiology and Biotechnology. 38(1):13-20.

Phelps P, Giddings TH, Prochoda M, Fall R. 1986. Release of cell-free ice nuclei by *Erwinia herbicola*. Journal of bacteriology. 167(2):496-502. eng.

Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, Kling DE, Gauthier LD, Levy-Moonshine A, Roazen D et al. 2018. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv.201178.

Pouleur S, Richard C, Martin JG, Antoun H. 1992. Ice nucleation activity in *Fusarium acuminatum* and *Fusarium avenaceum*. Appl Environ Microbiol. 58(9):2960-2964. eng.

Šantl-Temkiv T, Sahyoun M, Finster K, Hartmann S, Augustin-Bauditz S, Stratmann F, Wex H, Clauss T, Nielsen NW, Sørensen JH et al. 2015. Characterization of airborne ice-nucleation-active bacteria and bacterial fragments. Atmospheric Environment. 109:105-117.

Schmale III DG, Ross SD, Fetters TL, Tallapragada P, Wood-Jones AK, Dingus B. 2012. Isolates of *Fusarium graminearum* collected 40–320 meters above ground level cause Fusarium head blight in wheat and produce trichothecene mycotoxins. Aerobiologia. 28(1):1-11.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 24(5):637-644.

Tam V, Patel N, Turcotte M, Bossé Y, Paré G, Meyre D. 2019. Benefits and limitations of genome-wide association studies. Nature Reviews Genetics. 20(8):467-484.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic acids research. 22(22):4673-4680. eng.

Tsumuki H, Konno H. 1994. Ice nuclei produced by *Fusarium* sp. isolated from the gut of the rice stem borer, *Chilo suppressalis* Walker (Lepidoptera: Pyralidae). Bioscience, Biotechnology, and Biochemistry. 58(3):578-579.

UniProt Consortium. 2020. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 49(D1):D480-D489.

Vali G. 1971. Quantitative evaluation of experimental results on the heterogeneous freezing nucleation of supercooled liquids. Journal of the Atmospheric Sciences. 28(3):402-409.

Efficient architecture-aware acceleration of BWA-MEM for multicore systems. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS); 20-24 May 2019 2019.

Wolber PK, Deininger CA, Southworth MW, Vandekerckhove J, van Montagu M, Warren GJ. 1986. Identification and purification of a bacterial ice-nucleation protein. Proc Natl Acad Sci U S A. 83(19):7256-7260. eng.

Yang S, Coleman JJ, Vinatzer BA. 2021. Genome Resource: Draft genome of *Fusarium avenaceum*, strain F156N33, isolated from the atmosphere above Virginia and annotated based on RNA sequencing data. Plant Disease.

Tables

Table 1. List of *F. avenaceum* strains tested for INA

Source	Accession	Substrate	Sampling location	Accession(s) in other collections
USDA ARS Culture Collection (NRRL)	13316	Turf soil	Pennsylvania, USA	NA
	13826	Carnation	California, USA	NA
	36457	Barley kernel	USA	CBS 409.86 /FRC R-8509/IMI 309353
	54396	Soil	Easter Lilly Research Borrkings, Oregon, USA	F49
	54754	Corn	Pennsylvania, USA	A-28077
	66272	Wheat	Washington, USA	A-28073
	66944	Seedling of spruce	Pennsylvania, USA	A-28042
	66946	Plant roots, Douglas fir tree	Oregon, USA	A-28020
	66947	Seedling of douglas fir	Pennsylvania, USA	A-28035
	66948	Sugar pine tree seedling	Oregon, USA	A-28040
	66949	Seedling of spruce	Pennsylvania, USA	A-28041
	66950	Seedling of spruce	Pennsylvania, USA	A-28043
Kansas State University	11440	NA	NA	
Virginia Tech	F156N33	Atmosphere	Virginia, USA	

Table 2. Assembly summary of 14 *F. avenaceum* strains

Strain	Coverage (×)	Assembly size (bp)	Number of contigs	Maximum contig length (bp)	Minimum contig length (bp)	Average contig length (bp)	Median contig length (bp)	N50 contig length (bp)	GC content (%)	[a]Assembly BUSCO coverage (%)
F156N33	51	41,175,306	214	3,233,628	210	192,487	1,075	1,472,944	48.44	C:97.8; F:0.5; M:1.7
11440	52	42,933,485	897	1,984,632	200	48,011	538	1,024,532	48.33	C:97.7; F:0.5; M:1.8
NRRL 13316	61	41,704,585	964	2,164,424	200	43,399	579	843,661	48.26	C:97.7; F:0.5; M:1.8
NRRL 13826	51	44,694,304	1,465	1,780,814	200	30,660	493	779,627	48.35	C:97.6; F:0.5; M:1.9
NRRL 36457	54	38,761,238	308	2,047,741	234	125,928	1,068	918,031	48.38	C:97.7; F:0.6; M:1.7
NRRL 54396	51	49,692,405	626	1,709,761	200	79,506	659	691,401	50.72	C:97.8; F:0.5; M:1.7
NRRL 54754	50	38,889,550	234	4,244,225	229	166,303	875	1,329,635	48.48	C:97.8; F:0.5; M:1.7
NRRL 66272	49	39,140,173	251	2,646,538	226	156,028	1,000	1,190,042	48.47	C:97.6; F:0.5; M:1.9
NRRL 66944	59	36,826,384	216	3,227,342	231	170,590	845	1,246,826	48.48	C:97.6; F:0.4; M:2.0
NRRL 66946	54	40,603,425	397	2,016,945	200	102,387	945	972,920	48.42	C:97.7; F:0.5; M:1.8
NRRL 66947	61	40,212,181	320	2,482,218	228	125,771	996	1,127,118	48.34	C:97.9; F:0.5; M:1.6
NRRL 66948	57	38,139,861	290	2,379,904	202	131,584	1,103	1,046,193	48.31	C:97.9; F:0.5; M:1.6
NRRL 66949	52	37,127,996	260	3,604,089	234	142,881	1,021	1,215,522	48.41	C:97.7; F:0.4; M:1.9
NRRL 66950	51	40,854,987	317	4,247,833	214	128,969	1,117	1,141,783	48.38	C:97.8; F:0.5; M:1.7

[a] For BUSCO coverage, C stands for complete BUSCOs, F stands for fragmented BUSCOs, and M stands for missing BUSCOs.

Figures

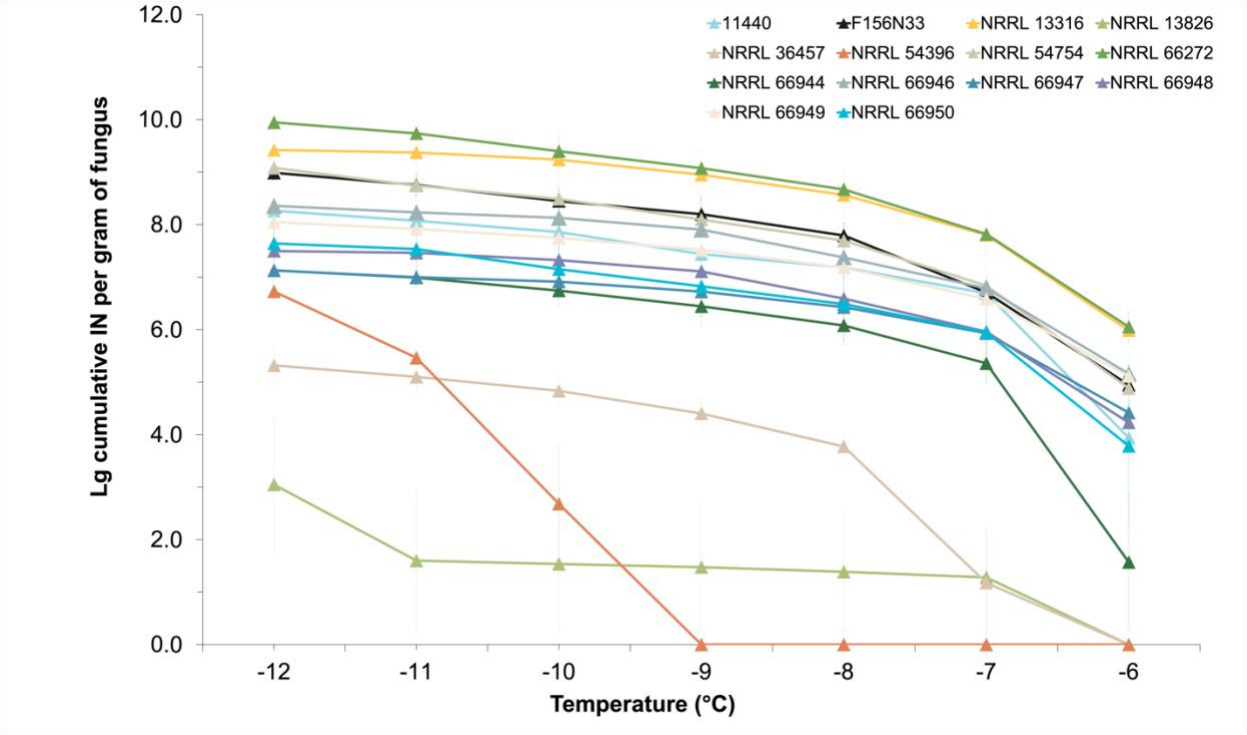
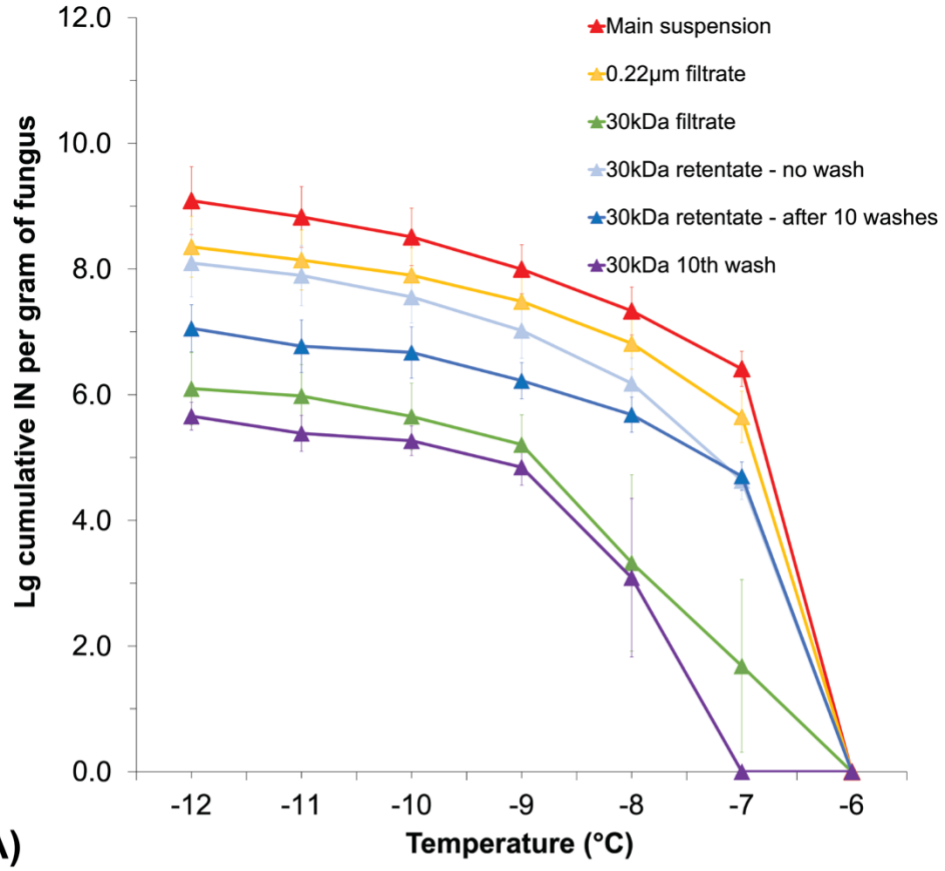
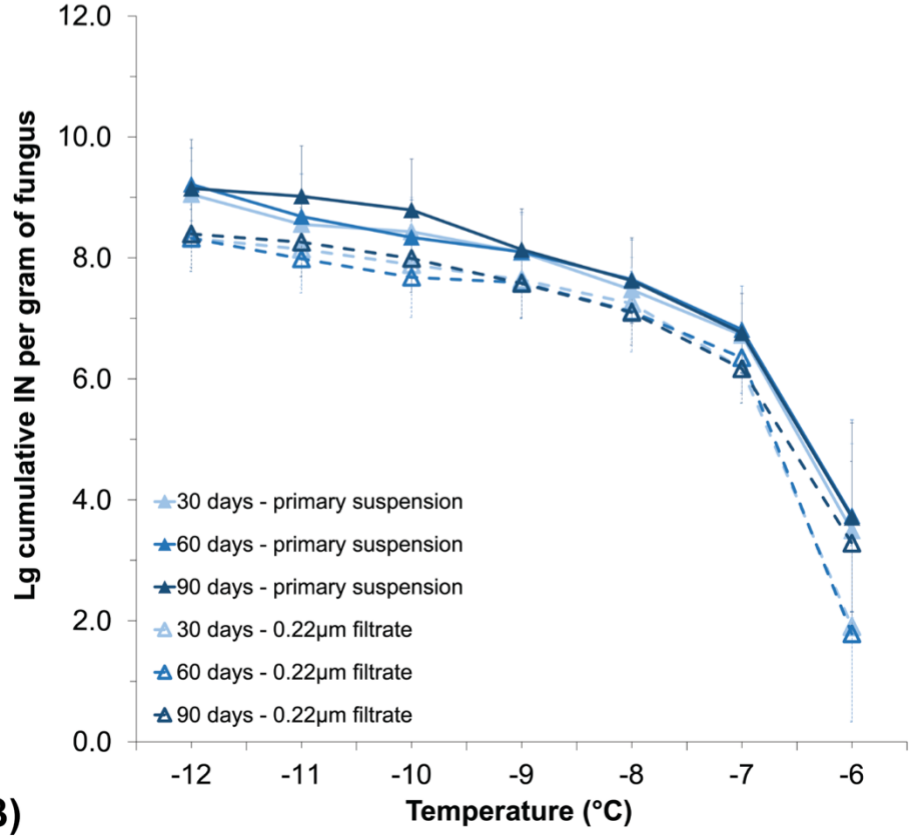


Figure 1. Cumulative ice nucleation spectra of 14 *F. avenaceum* strains. All cultures were grown at room temperature for 7 days. Results are primary suspensions based on droplet freezing assays at -6, -7, -8, -9, -10, -11, and -12°C. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.

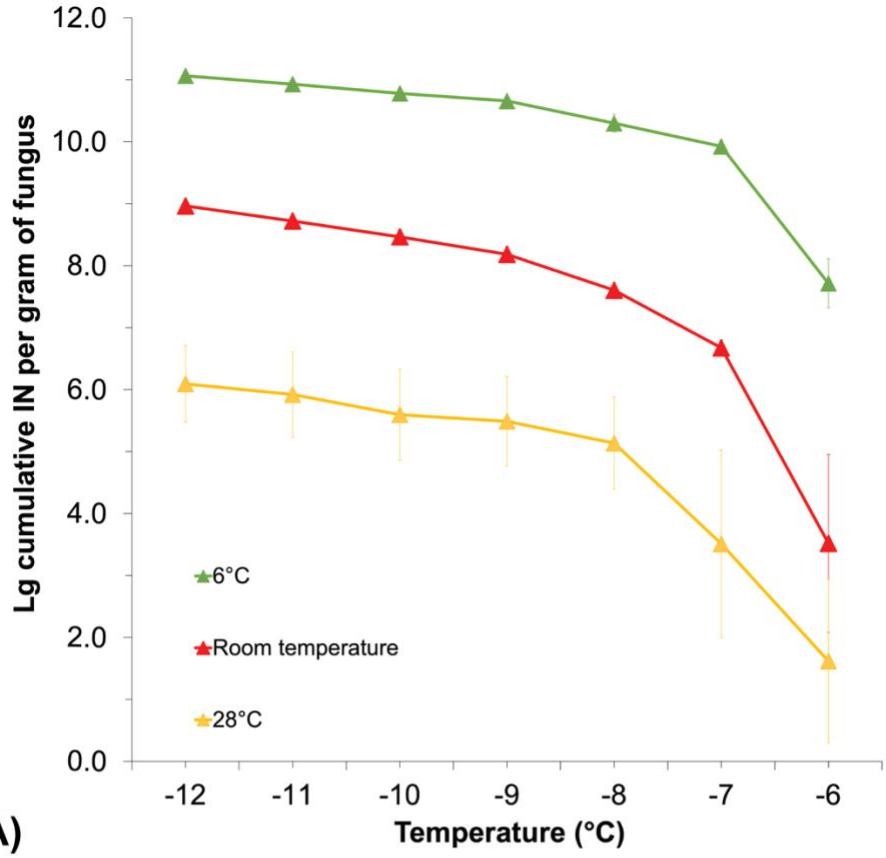

Main suspension
0.22µm filtrate
30kDa filtrate
30kDa retentate - no wash
30kDa retentate - after 10 washes
30kDa 10th wash
Lg cumulative IN per gram of fungus
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

A)

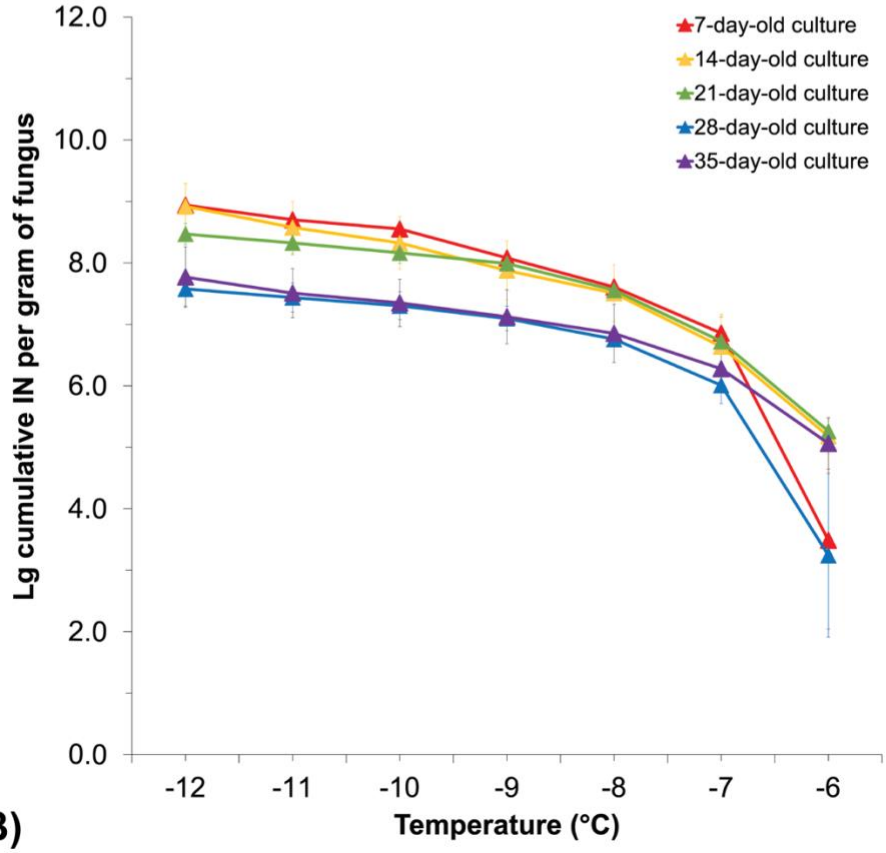

30 days - primary suspension
60 days - primary suspension
90 days - primary suspension
30 days - 0.22µm filtrate
60 days - 0.22µm filtrate
90 days - 0.22µm filtrate
Lg cumulative IN per gram of fungus
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

B)

Figure 2. Cumulative ice nucleation spectra of *F. avenaceum* F156N33 grown at room temperature for 7 days. A) Results are primary suspensions, 0.22 μm filtrates, 30 kDa filtrates, original 30 kDa retentates, washed 30 kDa retentates, and last washes based on droplet freezing assays. B) Results are primary suspensions and 0.22 μm filtrates stored at -80°C based on droplet freezing assays. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Lg cumulative IN per gram of fungus
6°C
Room temperature
28°C
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

A)


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Lg cumulative IN per gram of fungus
7-day-old culture
14-day-old culture
21-day-old culture
28-day-old culture
35-day-old culture
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

B)

Figure 3. Cumulative ice nucleation spectra of *F. avenaceum* F156N33 A) grown at room 6°C, temperature, and 28°C, respectively, for about 30 days; B) grown at room temperature for 7 days, 14 days, 21 days, 28 days, and 35 days, respectively. Results are primary suspensions based on droplet freezing assays. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.

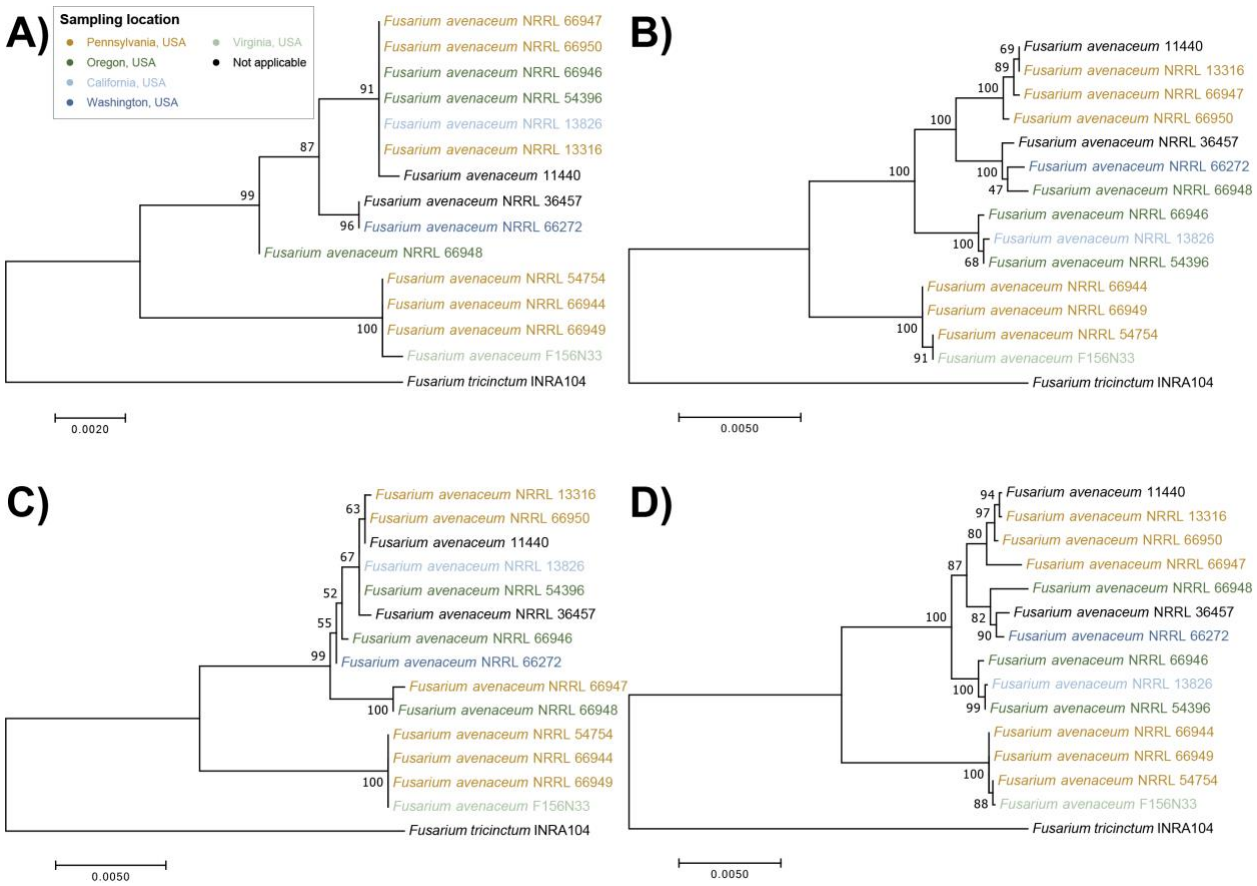


Figure 4. Maximum Likelihood (ML) trees constructed based on sequences of A) translation elongation factor 1-alpha (TEF-1α), B) RNA polymerase II largest subunit (RPB1), C) RNA polymerase II second largest subunit (RPB2), and D) combined four-locus data set using the best nucleotide substitution model with 1000 bootstrap replications. Each color represents a state where the strain was isolated.

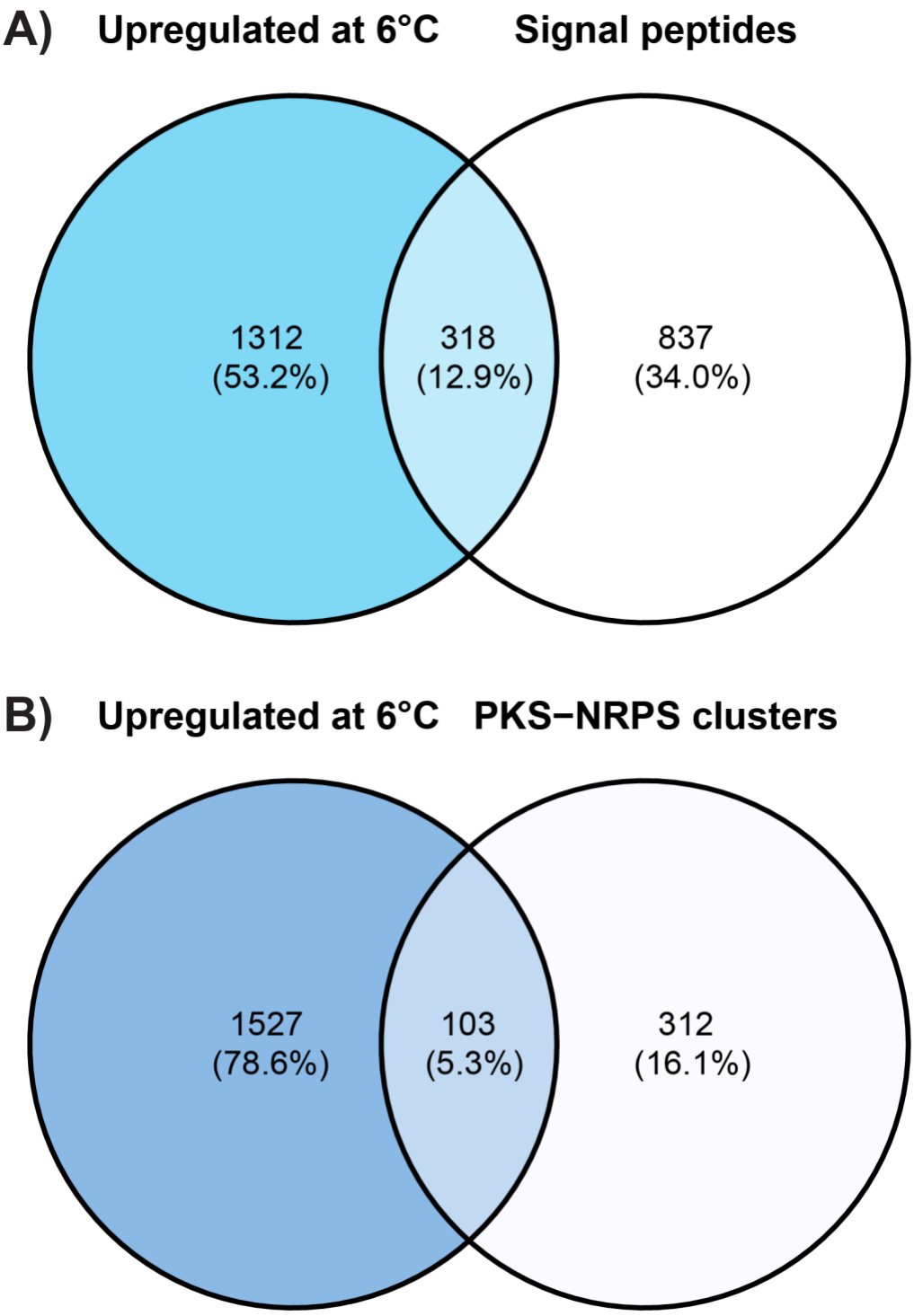


Figure 5. Venn diagrams. A) Overlap of genes that were upregulated at 6°C and predicted to encode signal peptides. B) Overlap of genes that were upregulated at 6°C and predicted to fall within PKS-NRPS clusters.

Supplementary tables

Supplementary table 1. List of genes in F156N33

Supplementary table 2. List of genes that were present in F156N33, NRRL 13316, NRRL 54754 and NRRL 66272 but absent from NRRL 54396

Supplementary table 3. Phyre2 predictions of genes in Supplementary table 2 that had no annotations or were uncharacterized by InterProScan or BLASTP

Supplementary table 4. List of genes that were present in 11 ice nucleation active strains but absent from NRRL 54396

Supplementary table 5. Phyre2 predictions of genes that were upregulated at 6°C and predicted to encode signal peptides with no annotations or were uncharacterized by InterProScan or BLASTP

Supplementary table 6. Number of reads of each strain that aligned to each F156N33 gene

Supplementary figures

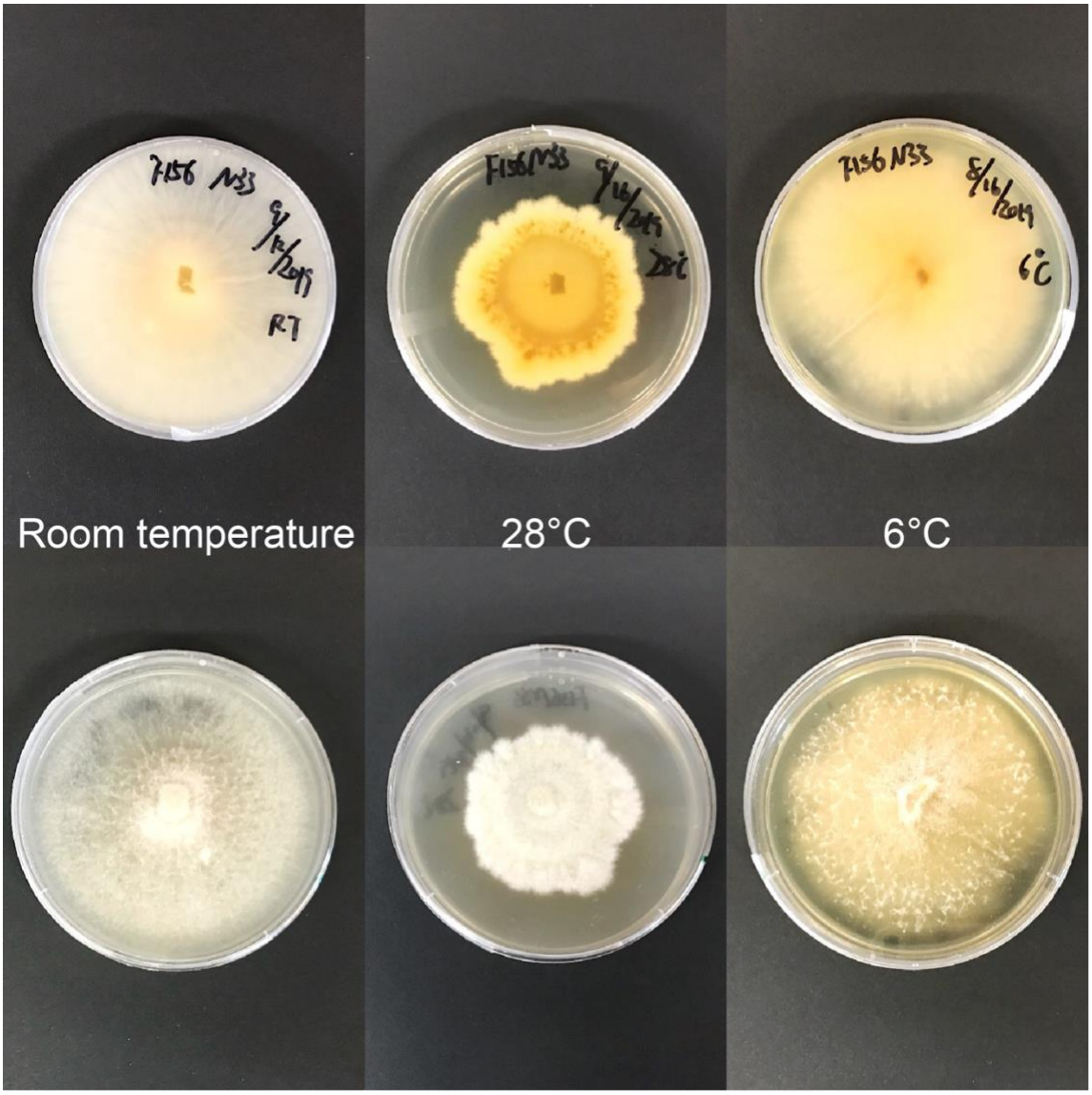


Figure S1. Morphology of *F. avenaceum* F156N33 grown at room temperature, 28°C, and 6°C, respectively, for approximately 30 days. The upper photos show a view at the bottom of the plates while the lower photos show a view through the Petri dish cover.

Chapter 4. Exploring the genetic basis of ice nucleation activity in the common soil fungus *Mortierella alpina*

Shu Yang[1], Boris A. Vinatzer[1]

[1] School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, USA

Corresponding author: B. A. Vinatzer, vinatzer@vt.edu

Keywords: fungi, ice nucleation activity, biological ice nucleating particles, *Mortierella alpina*, comparative transcriptomics

Abstract

Ice Nucleation activity (INA) is the ability of some particles to induce ice formation above the freezing temperature of pure water. Very few bacterial species and fungi display INA at temperatures above –10°C. The common soil fungus *Mortierella alpina* is one of them. INA in *M. alpina* has been shown to be heat and proteinase sensitive suggesting that INA depends on a folded protein. However, it cannot be excluded that INA depends instead on the product of a polyketide synthase non-ribosomal peptide synthetase as found for INA in the bacterial species *Lysinibacillus parviboronicapiens*. It has also been shown that INA in *M. alpina* is associated with secreted ice nucleation particles (INPs) smaller than 300 kDa in size. However, the genes necessary for INA in *M. alpina* are unknown. Here we report that differential centrifugation suggests that *M. alpina* INPs consist of aggregates that can be separated into INPs smaller than 30kDa and that *M. alpina* INPs are stable when stored at –80°C for up to three months. Also, the strength of INA in *M. alpina*

increases as the growth temperature drops from 28°C to room temperature and to 6°C. Hypothesizing that the strength of INA in *M. alpina* correlates with the expression level of the genes necessary for INA, we compared gene expression in one *M. alpina* strain between the three tested growth temperatures. 560 genes were more highly expressed at either 6°C or room temperature or both compared to 28°C and 106 of these genes encoded secreted proteins. We discuss which of these genes may encode putative INA molecules and should be further investigated.

Introduction

Soil has been proposed to be a relevant source of atmospheric ice nucleating particles (INPs), which are known to affect the ratio of frozen to liquid cloud droplets and thus earth's radiation balance and the amount and intensity of precipitation (Schnell and Vali 1972; Schnell and Vali 1976; Conen et al. 2011). INPs that induce ice formation at temperatures warmer than –10°C have been found to be mostly of biological origin (Christner et al. 2008) and a wealth of circumstantial evidence suggests that INPs of biological origin contribute to atmospheric INPs (Pratt et al. 2009; Conen et al. 2011; Creamean et al. 2013; O'Sullivan et al. 2014; O'Sullivan et al. 2018).

Biological INPs include bacteria, fungi, viruses, pollen, lichen, and marine organics amongst others (Lundheim and Zachariassen 1999; Adams et al. 2021). Fungal INA was first found, and has been best characterized, in the genus *Fusarium* (Pouleur et al. 1992; Kunert et al. 2019). Very few additional fungal genera have been found to be ice nucleation-active (Ice^+) since then with one of them being the widespread and abundant soil fungus *Mortierella alpina*, which can initiate freezing at −5°C to −6°C (Fröhlich-Nowoisky et al. 2015; Pummer et al. 2015; Hill et

al. 2016). To date, *M. alpina* is the only species in the Mortierellaceae family known to include strains with ice nucleation activity (INA).

The Mortierellaceae family consists in filamentous fungi that belong to the subphylum Mortierellomycotina (Spatafora et al. 2016). Species in the Mortierellaceae family are distributed in a diversity of environments but are most commonly associated with soils, the rhizosphere, and plant roots (Summerbell 2005; Nagy et al. 2011). These species are of industrial, agricultural, and clinical relevance (Carter et al. 1973; Kikukawa et al. 2018; Ozimek and Hanaka 2021). *M. alpina*, a saprobic member in this family, is considered an important oleaginous fungus that can be used for the production of the polyunsaturated fatty acid (PUFA) arachidonic acid (ARA) on an industrial scale (Shimizu and Yamada 1990; Shimizu et al. 1997; Certik et al. 1998; Certik and Shimizu 1999).

M. alpina strains with INA have been isolated from soils (Fröhlich-Nowoisky et al. 2015; Hill et al. 2016) and leaf litter (Vasebi et al. 2019) and have been implied as source of INA in fresh water (Knackstedt et al. 2018). Some studies suggested that *M. alpina* INPs are proteinaceous compounds since INA was lost after proteinase and chemical treatments (Fröhlich-Nowoisky et al. 2015). Other known characteristics of *M. alpina* INPs include that they are smaller than 300 kDa in size and are heat-sensitive (Fröhlich-Nowoisky et al. 2015; Pummer et al. 2015; Vasebi et al. 2019). However, there is no direct evidence for *M. alpina* INPs being proteins and genes encoding *M. alpina* INPs have not been identified. So far, the molecular basis of *M. alpina* INPs is thus poorly understood.

Since very little is currently known about any fungal INPs, if *Mortierella* INA genes were to be identified, this knowledge would significantly increase our understanding of fungal INA in the soil and the atmosphere. Therefore, in this study, multiple species within the Mortierellaceae

family were screened for INA. After it was found that INA was limited to *M. alpina*, one the most active strains, LL118, was chosen for genome sequencing (Yang and Vinatzer 2021). Since the strength of INA in this strain increased as growth temperature decreased, we hypothesized that genes necessary for INA would also be expressed more highly at lower temperatures compared to higher temperatures. Based on this hypothesis, putative INA genes were identified.

Material and Methods

Mortierellaceae strains

Seventeen strains belonging to 13 Mortierellaceae species were investigated (Table 1). Except for *M. alpina* LL118, which was isolated from an Aspen leaf litter sample, all other strains were provided by Dr. Gregory Bonito (Michigan State University). All strains were grown on potato dextrose agar (PDA) prior to being processed unless otherwise specified.

INA testing

After Mortierellaceae strains were grown for 7 days at room temperature, 1 cm^2 of mycelium was taken from the center of Petri dishes and suspended in 1 mL of nuclease-free water. INA testing was performed using a droplet freezing assay with a glycerol bath in a cooling thermostat (LAUDA Alpha Cooling Thermostat RA24) as previously described by Failor et al. (2017). The water used to make the suspensions served as negative control. INA was tested on 30 drops per strains at –6°C, –7°C, –8°C, –9°C, –10°C, –11°C, and –12°C. Drops were incubated for 10 minutes at each temperature, and the number of frozen drops was recorded. The assay was repeated two times. For strains for which a cumulative ice nucleation spectrum was obtained, 0.5 mg of mycelia from the center of each plate were suspended in 1 mL of nuclease-free water to make a primary suspension,

which was then used to make a dilution series from 10^{-1} to 10^{-5}. Thirty drops per dilution were tested and the number of ice nuclei (IN) per gram of fungus at each temperature was calculated based on the method developed by Vali (1971). To determine if any differences in INA among treatments were statistically significant, analysis of variance (ANOVA) was performed at certain temperature using R (v4.0.4).

Investigation of INP properties

To investigate the effect of filtration on INA, *M. alpina* strain LL118 was grown on PDA for 14 days at room temperature. A primary suspension was made by suspending 5 mg of mycelia collected from the center of Petri dishes in 50 mL of nuclease-free water. 49 mL of the primary suspensions were passed through a 0.22-µm-pore-size filter (Millex-GP Syringe Filter, 0.22 µm). 20ml of the so obtained 0.22 µm filtrate were passed through a 30-kDa-pore-size filter (Macrosep Advance Centrifugal Devices with Omega Membrane 30K) for 10 min at 5,000 rpm. This step separated INPs smaller than approximately 5 nm, which passed through the 30kDa filter, from those larger than 5 nm, which were retained in in the 30kDa retentate. The latter was resuspended from the filter membrane with 500 µL of nuclease-free water. In parallel, another 20 mL of the 0.22 µm filtrate were passed through a separate 30 kDa filter but 10 mL of nuclease-free water were added to the filter and centrifugation was repeated several times. After each centrifugation, the filtrate was stored separately and another 10 mL of nuclease-free water were added to the filter. Centrifugation was repeated until INA in the filtrate became almost undetectable. After the last centrifugation, the 30 kDa retentate was resuspended from the filter membrane with 500 µL of nuclease-free water. The primary suspension, the 0.22 µm filtrate, the 30 kDa filtrate, the original

30 kDa retentate, the washed 30 kDa retentate, the final filtrate, and 10^{-1} to 10^{-5} dilutions of each fraction were used to characterize INA in droplet freezing assays.

To investigate the effect of storage at extreme low temperature, the primary suspension and the 0.22 μm filtrate were stored at –80°C for 30 days, 60 days, and 90 days, respectively, prior to INA testing.

To investigate the effect of growth temperature on INA in *M. alpina*, strain LL118 was grown on PDA for about 30 days at 6°C, room temperature, or 28°C. For the effect of culture age, the strain was grown for 7 days, 14 days, 21 days, 28 days, and 35 days at room temperature. The primary suspension and 10^{-1} to 10^{-5} dilutions were made as described above for each of the treatments and used to characterize INA in droplet freezing assays.

Sequencing and assembly

The methods of genome and transcriptome sequencing has been described by Yang and Vinatzer (2021). In brief, genomic DNA of 5 *M. alpina* strains was extracted from mycelium grown in potato dextrose broth (PDB) using the ZymoBIOMICS DNA Miniprep Kit (Zymo Research) and was sequenced on an Illumina HiSeq 3000 Platform (2 × 100 bp) at the Iowa State University DNA Facility. Total RNA of *M. alpina* LL118, an Ice^{+} strain, was extracted from mycelium grown on PDA at 6°C (3 replicates), room temperature (3 replicates), and 28 °C (3 replicates) using the RNeasy® Plant Mini Kit (QIAGEN) and was sequenced an Illumina Nova Seq 6000 Platform at Novogene Corporation Inc. (Sacramento, CA). DNA reads were trimmed using Trimmomatic v0.39 (Bolger et al. 2014) to remove adapters. For RNA reads, low-quality reads and adapters were removed by the company. The quality of reads was checked and confirmed prior to genome assembly and annotation using FastQC v0.11.9 (Andrews et al. 2010).

Reads were aligned against the respective genome using STAR v2.7.8a (Dobin et al. 2013). The resulting alignment in BAM format and the genome assembly in FASTA format were used for transcriptome assembly. Genome-guided *de novo* assembly was performed by Trinity v2.12.0 (Haas et al. 2013). The completeness of the assembled transcriptome was assessed by the tool Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.0.0 with the hypocreales_odb10 dataset in transcriptome mode (Seppey et al. 2019). Transcripts were also assembled by StringTie v2.1.5 (Pertea M et al. 2015) to obtain the GTF file, which was later converted to a GFF file by GffRead v0.12.1 (Pertea G and Pertea 2020).

Gene prediction and genome annotation

The MAKER annotation pipeline (v3.01.03) (Campbell et al. 2014) was used for genome annotation combining evidence-based methods and *ab initio* gene predictions. In the first round, the annotation pipeline incorporated transcript evidence from the assembled transcriptome described above and protein evidence from the most closely related species available *Linnemannia elongata* (Uehling et al. 2017) (ID: UP000078512) ("est2genome" and "protein2genome" options in MAKER). Next, two rounds of MAKER were performed for *ab initio* gene prediction ("snaphmm" option in MAKER) after training with SNAP v2013-02-16 (Korf 2004) in each round. The resulting GFF file from the third round was used to train the AUGUSTUS model for *M. alpina* strain LL118 (Stanke et al. 2008). In the final round, the pipeline incorporated *ab initio* gene prediction from AUGUSTUS v3.4.0 using the trained AUGUSTUS model ("augustus_species" option in MAKER), which generated the final GFF file containing annotation data. The final GFF file was functionally annotated with InterProScan v5.46-81.0 (Jones et al. 2014) for the presence of Pfam domains and BLASTP from BLAST v2.10.0+ (Camacho et al.

2009) for searching against the Swiss-Prot database released in February 2021 (Consortium 2020) (E-value, $< 1 \times 10^{-6}$). The quality of the genome annotation was assessed by BUSCO v5.0.0 (Seppey et al. 2019) in the protein mode that evaluates the completeness of annotated gene sets. The BUSCO assessment was performed against the lineage-specific profile library mucoromycota_odb10 (1,614 genes).

Functional annotation, prediction of the signal peptide and prediction of secondary metabolite genes

Functional annotation was performed using InterProScan v5.46-81.0 for the presence of Pfam domains with terms from the Gene Ontology database (Jones et al. 2014). BLASTP of the BLAST v2.10.0+ suite (Camacho et al. 2009) was used to find regions of local similarity against the February 2021 release of the Swiss-Prot database (Consortium 2020). Prediction of the signal peptide was performed by SignalP v5.0b (Almagro Armenteros et al. 2019). Phyre2 was also used to predict and analyze proteins (Kelley et al. 2015). Prediction of secondary metabolite genes was performed by the fungal version of antiSMASH 6 (Blin et al. 2021).

Gene expression analysis

RNA-seq reads from each the three replicates from each of the three growth temperatures were aligned to the genome assembly of the same strain (LL118) using STAR v2.7.8a (Dobin et al. 2013) generating nine alignment files in BAM format. These nine BAM files were subjected to featureCounts v2.0.1 (Liao et al. 2014) with parameters -p -B -C (multi-mapped reads excluded) as well as -p -B -O -M (multi-mapped reads included) to determine the number of reads mapped to each gene. The read counts were normalized by the DESeq2 package in R (Love et al. 2014).

Differential expressed genes (DEGs) were identified by pairwise comparisons with the following parameters: "padj (adjusted P value) < 0.05 and log2FoldChange > 1 using DESeq2.

Results and discussion

M. alpina is the only species in the Mortierellaceae family that has INA and INA varies within *M. alpina*

To determine the distribution of INA within the Mortierellaceae family, droplet-freezing assays were performed at temperatures between –6°C and –12°C for 17 strains in 13 Mortierellaceae species. Out of all tested strains, only the *M. alpina* strains presented INA. Two of the strains had strong INA, one had intermediate INA, and two strains were found to be negative (Table 1). These latter two strains did not induce ice formation at temperatures as low as –12°C. However, these results will need to be confirmed because some INA was detected even in these two strains when the concentration of mycelium was increased (data not shown).

The two most active Ice^+ strains, *M. alpina* strains LL118 and NVP153, were tested in detail computing cumulative ice nucleation spectra. Strain LL118 initiated ice nucleation at –8°C, and NVP153 initiated ice nucleation at –7°C (Fig. 1). LL118 produced close to 10^{10} IN per gram of mycelium and NVP153 produced $10^{8.5}$ g^{-1} at –9°C and below.

Since high INA was detected in two *M. alpina* strains, lower INA was observed in a third strain, and almost no INA was detected in another two strains, INA may be a quantitative trait in *M. alpina*. However, INA in these strains and in additional strains will need to be precisely quantified to make a firm conclusion. If INA were to be confirmed to be a quantitative trait in *M. alpina*, the strength of INA may depend on the presence or absence of several INA genes or it may be affected by allelic differences in one or several INA genes.

Repeated and differential centrifugation indicates that *M. alpina* INPs consist in aggregates that can be separated into smaller units

To investigate the properties of *M. alpina* INPs, filtration using filters with different pore sizes was performed. The primary suspension of strain LL118 produced approximately 10^{10} IN per gram of mycelium at –10°C, and the 0.22 µm filtrate had about 10^9 g^{-1} (Fig. 2A). These results indicate that *M. alpina* INPs are likely to be secreted. Compared with the primary suspension, INA was reduced approximately 1,000-fold after passing through a 30 kDa filter at –10°C (p-value = 0.0059). The retentate collected from 30 kDa filters without washing presented almost the same INA compared to the 0.22 µm filtrate (p-value = 0.1396), suggesting that most *M. alpina* INPs are larger than 30 kDa (approximately 5 nm). This is in agreement with earlier results by Fröhlich-Nowoisky et al. (2015) that *M. alpina* INPs consist in extracellular proteins that are smaller than 300 kDa.

However, the 30kDa filtrate still contained 10^7 g^{-1} of IN at –10°C, suggesting that some *M. alpina* INPs are even smaller than 30 kDa. This was confirmed when the 30 kDa retentate was washed. In fact, after the 30 kDa filter was washed ten times (by adding water followed by centrifugation ten times), it only contained 10^6 IN per gram of mycelium at –10°C, an approximately 1,000-fold reduction compared with the unwashed 30 kDa retentate (p-value = 0.0102) (Fig. 2A). Therefore, we conclude that *M. alpina* INPs represent aggregates composed of smaller sub-units that can be separated from each other and still maintain high INA. This property has been also found in *Fusarium* INPs (Kunert et al. 2019), suggesting fungal INPs may share this common feature.

***M. alpina* INPs maintain INA at –80°C**

INA of *M. alpina* was found to be stable after storage at an extreme low temperature of –80°C for 90 days (Fig. 2B) indicating that *M. alpina* INPs may be able to persist in the atmosphere over long periods of time, which makes it possible that they may in fact play a role in atmospheric processes.

High growth temperature inhibits *M. alpina* INA while the age of the culture does not

To investigate how growth temperature affects INA in *M. alpina*, strain LL118 was grown in parallel at 6°C, room temperature, and 28°C, for about 30 days prior to performing droplet-freezing assays. Growth temperature affected the morphology of cultures (Fig. S1). Also, growth was slow at 6°C but mycelium was very fluffy. Mycelium was not fluffy at 28°C. In terms of INA, INA was reduced dramatically after growth at 28°C compared to growth at 6°C and room temperature (p-value = 0.0004) (Fig. 3A). In fact, the cumulative number of IN per gram of mycelium was similar between cultures grown at 6°C and room temperature, around 10^{10} g^{-1} at –10°C (p-value = 0.5260), but was reduced 10^5 fold for cultures grown at 28°C comparing to cultures grown at room temperature (p-value = 0.0046).

To investigate how culture age affects *M. alpina* INA, strain LL118 was grown for 7 days, 14 days, 21 days, 28 days, and 35 days, respectively. The length of growth time did not significantly affect INA (for example, p-value = 0.4460 at –10°C), and no correlation was found between days and the intensity of INA (Fig. 3B).

Since INA in *M. alpina* was higher after growth at 6°C and room temperature compared to growth at 28°C, we hypothesize that the expression of INA genes may be induced when *M. alpine* grows at 6°C and room temperature compared to when it grows at 28°C. While INA was similar for mycelium grown between 6°C and room temperature when INA was assayed at –10°C and

lower, INA of *M. alpina* mycelium grown at 6°C was numerically stronger than INA of *M. alpina* mycelium grown at room temperature when INA was assayed between –6°C and –12°C. Thus, INA genes in *M. alpina* may be more highly induced when grown at 6°C compared to growth at room temperature. We had found the same trend for *F. avenaceum* INPs (See chapter 3). 28°C consistently repressed INA in both genera. However, the difference in INA between growth at 6°C and room temperatures was more pronounced in *F. avenaceum*. Instead of differences in gene expression at different temperatures, it is also possible that the structure and/or size of *M. alpina* and *F. avenaceum* INPs and post-translational modifications of the putative INA proteins are different at different temperatures and cause the observed differences in INA.

Transcriptomics identifies a series of putative INA genes

Based on our phenotyping results, INA in *M. alpina* is suppressed at high temperatures but is highly induced at room temperature and even more so at 6°C. Differential expression (DE) analyses were thus performed for a series of comparisons between growth temperatures. All analyses were performed both, using uniquely mapped reads only or including multi-mapped reads. In order not to exclude potential candidates, we report the results using uniquely mapped reads in combination with multi-mapped reads.

1,384 genes were upregulated in the pairwise comparison between 6°C and 28°C, and 946 genes were upregulated in the pairwise comparison between room temperature and 28°C. 560 of these genes were found to be upregulated in both of these pairwise comparisons. Considering INA was slightly higher at 6°C than at room temperature, genes that were more highly expressed at 6°C compared with room temperature were also identified (962 genes). Finally, among the 560 genes

upregulated both at 6°C compared to 28°C and at room temperature compared to 28°C, 304 genes were more highly expressed at 6°C compared with room temperature (Supplementary Table 1).

Since we confirmed previous results that *M. alpina* INPs are secreted and *M. alpina* INPs were previously hypothesized to be proteinaceous, we next determined which of the upregulated genes were predicted to encode secreted proteins. In total, 647 genes in the genome of strain LL118 were predicted to encode signal peptides, indicating that they are secreted. Of these, 106 genes were upregulated when *M. alpina* was growth at both 6°C and room temperature compared to 28°C (Supplementary Table 1). 57 genes out of these 106 genes had no annotations or were uncharacterized by InterProScan or BLASTP. 39 genes were predicted to encode enzymes of known function. The remaining 10 genes were most likely to be associated with the cell membrane. According to Phyre2, also the genes encoding proteins of unknown function were predicted to encode enzymes, structural proteins, and proteins associated with cell membrane (Supplementary Table 2). However, most predictions had low confidence or low sequence identify. Therefore, these predictions should be considered with caution.

On the other hand, of the 106 genes predicted to encode signal peptides and upregulated at both temperatures, 74 genes were also upregulated after growth at 6°C compared to growth at room temperature (Supplementary Table 1). These genes were thus considered the most likely candidate genes. Still, more than half of them had no annotations or were uncharacterized by InterProScan or BLASTP, and those with annotations are most likely to be predicted to encode enzymes.

Alternatively, *M. alpina* INPs could be the product of a polyketide synthase non-ribosomal peptide synthetase (PKS-NRPS) as was recently shown for the Gram positive bacterium *L parviboronicapiens* (Failor et al. 2021). However, only 13 genes in the entire *M. alpina* LL119

genome were found to be located within PKS-NRPS clusters, and none of them were upregulated at the lower temperatures.

Also, *M. alpina* is known for producing arachidonic acid. It remains unknown whether arachidonic acid can initiate ice nucleation at higher temperature. Previous studies suggested that long-chain fatty acids are not effective at nucleating ice at temperatures above –36°C (Qiu et al. 2017; DeMott et al. 2018). However, it may still interesting to look at genes involved in fatty acid synthesis since they may play a role in combination with associated proteins or polyketides or non-ribosomal peptides.

Finally, we performed a pan-genome analysis based on the hypothesis that INA genes could be present in the two strains with the highest INA (LL118 and NVP153) and absent from the strains with the lowest INA. 9 genes were so identified (Supplementary Table 3). However, none of them were predicted to encode signal peptides, and only one gene was upregulated at both 6°C and room temperature. This may be a limitation of having included only 5 strains in our study. More strains would need be to be included to conduct a genome-wide association study (GWAS), which would also be able to determine if allelic differences are at the basis of difference in INA between strains instead of complete presence/absence.

Conclusions

Here we investigated 17 strains in 13 Mortierellaceae species and only a subset of the *Mortierella alpina* strains had INA, suggesting it is a feature unique to *M. alpina* among the Mortierellaceae. *M. alpina* INPs are likely to be secreted aggregates. Therefore, INA genes may either encode proteins with signal peptides or fall within PKS-NRPS clusters. Since only 5 *M. alpina* strains were examined, it is impossible to conclude if the presence/absence of several genes or allelic

differences contribute to the strength of INA in *M. alpina.* Growth temperature strongly affects INA in *M. alpina*, suggesting that expression of INA genes may be higher at lower temperatures compared with high temperatures. Therefore, we have obtained a list of putative INA genes that are more highly expressed at lower temperatures. Based on the hypothesis that INPs are likely to be secreted proteins or polyketide non-ribosomal peptides, we have reduced the number of putative INA genes in *M. alpina*. However, to identify which gene(s) is (are) responsible for *M. alpina* INA, more strains will need to be analyzed. Either GWAS will need to be conducted and/or mutational analyses need to be performed through experiments.

References

Adams MP, Atanasova NS, Sofieva S, Ravantti J, Heikkinen A, Brasseur Z, Duplissy J, Bamford DH, Murray BJ. 2021. Ice nucleation by viruses and their potential for cloud glaciation. Biogeosciences. 18(14):4431-4444.

Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol. 37(4):420-423.

Andrews S, Krueger F, Segonds-Pichon A, Biggins L, Krueger C, Wingett S. 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Institute, Babraham, UK. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema Marnix H, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research. 49(W1):W29-W35.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 30(15):2114-2120. eng.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. 2009. BLAST+: architecture and applications. BMC bioinformatics. 10(1):421.

Campbell MS, Holt C, Moore B, Yandell M. 2014. Genome annotation and curation using MAKER and MAKER-P. Current Protocols in Bioinformatics. 48:4.11.11-14.11.39. eng.

Carter ME, Cordes DO, Di Menna ME, Hunter R. 1973. Fungi isolated from bovine mycotic abortion and pneumonia with special reference to *Mortierella wolfii*. Research in Veterinary Science. 14(2):201-206.

Certik M, Sakuradani E, Shimizu S. 1998. Desaturase-defective fungal mutants: useful tools for the regulation and overproduction of polyunsaturated fatty acids. Trends in Biotechnology. 16:500-505.

Certik M, Shimizu S. 1999. Biosynthesis and regulation of microbial polyunsaturated fatty acid production. Journal of Bioscience and Bioengineering. 87(1):1-14.

Christner BC, Morris CE, Foreman CM, Cai R, Sands DC. 2008. Ubiquity of biological ice nucleators in snowfall. Science. 319(5867):1214. eng.

Conen F, Morris CE, Leifeld J, Yakutin MV, Alewell C. 2011. Biological residues define the ice nucleation properties of soil dust. Atmos Chem Phys. 11(18):9643-9648.

Creamean JM, Suski KJ, Rosenfeld D, Cazorla A, DeMott PJ, Sullivan RC, White AB, Ralph FM, Minnis P, Comstock JM et al. 2013. Dust and biological aerosols from the Sahara and Asia influence precipitation in the western U.S. Science. 339(6127):1572-1578. eng.

DeMott PJ, Mason RH, McCluskey CS, Hill TCJ, Perkins RJ, Desyaterik Y, Bertram AK, Trueblood Jonathan V, Grassian VH, Qiu Y et al. 2018. Ice nucleation by particles containing long-chain fatty acids of relevance to freezing by sea spray aerosols [10.1039/C8EM00386F]. Environmental Science: Processes & Impacts. 20(11):1559-1569.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 29(1):15-21. eng.

Failor KC, Liu H, Llontop MEM, LeBlanc S, Eckshtain-Levi N, Sharma P, Reed A, Yang S, Tian L, Lefevre C et al. 2021. Ice nucleation in a Gram-positive bacterium isolated from precipitation depends on a polyketide synthase and non-ribosomal peptide synthetase. The ISME Journal.

Failor KC, Schmale III DG, Vinatzer BA, Monteil CL. 2017. Ice nucleation active bacteria in precipitation are genetically diverse and nucleate ice by employing different mechanisms. The ISME Journal. 11(12):2740-2753.

Fröhlich-Nowoisky J, Hill TCJ, Pummer BG, Yordanova P, Franc GD, Pöschl U. 2015. Ice nucleation activity in the widespread soil fungus *Mortierella alpina*. Biogeosciences. 12(4):1057-1071.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M et al. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols. 8(8):1494-1512.

Hill TCJ, DeMott PJ, Tobo Y, Fröhlich-Nowoisky J, Moffett BF, Franc GD, Kreidenweis SM. 2016. Sources of organic ice nucleating particles in soils. Atmos Chem Phys. 16(11):7195-7211.

Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G et al. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics. 30(9):1236-1240. eng.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. Nature Protocols. 10(6):845-858.

Kikukawa H, Sakuradani E, Ando A, Shimizu S, Ogawa J. 2018. Arachidonic acid production by the oleaginous fungus *Mortierella alpina* 1S-4: A review. Journal of Advanced Research. 11:15-22.

Knackstedt KA, Moffett BF, Hartmann S, Wex H, Hill TCJ, Glasgo ED, Reitz LA, Augustin-Bauditz S, Beall BFN, Bullerjahn GS et al. 2018. Terrestrial origin for abundant riverine

nanoscale ice-nucleating particles. Environmental Science & Technology. 52(21):12358-12367.

Korf I. 2004. Gene finding in novel genomes. BMC bioinformatics. 5(1):59.

Kunert AT, Pöhlker ML, Tang K, Krevert CS, Wieder C, Speth KR, Hanson LE, Morris CE, Schmale III DG, Pöschl U et al. 2019. Macromolecular fungal ice nuclei in *Fusarium*: effects of physical and chemical processing. Biogeosciences. 16(23):4647-4659.

Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 30(7):923-930.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15(12):550.

Lundheim R, Zachariassen K. 1999. Applications of biological ice nucleators. Biotechnological Applications of Cold-Adapted Organisms. Springer; p. 309-317.

Nagy LG, Petkovits T, Kovács GM, Voigt K, Vágvölgyi C, Papp T. 2011. Where is the unseen fungal diversity hidden? A study of *Mortierella* reveals a large contribution of reference collections to the identification of fungal environmental sequences. The New phytologist. 191(3):789-794. eng.

O'Sullivan D, Murray BJ, Malkin TL, Whale TF, Umo NS, Atkinson JD, Price HC, Baustian KJ, Browse J, Webb ME. 2014. Ice nucleation by fertile soil dusts: relative importance of mineral and biogenic components. Atmos Chem Phys. 14(4):1853-1867.

O’Sullivan D, Adams MP, Tarn MD, Harrison AD, Vergara-Temprado J, Porter GCE, Holden MA, Sanchez-Marroquin A, Carotenuto F, Whale TF et al. 2018. Contributions of biogenic material to the atmospheric ice-nucleating particle population in North Western Europe. Scientific Reports. 8(1):13821.

Ozimek E, Hanaka A. 2021. *Mortierella* species as the plant growth-promoting fungi present in the agricultural soils. Agriculture. 11(1):7.

Pertea G, Pertea M. 2020. GFF utilities: GffRead and GffCompare. F1000Research. 9(304).

Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 33(3):290-295.

Pouleur S, Richard C, Martin JG, Antoun H. 1992. Ice nucleation activity in *Fusarium acuminatum* and *Fusarium avenaceum*. Appl Environ Microbiol. 58(9):2960-2964. eng.

Pratt KA, DeMott PJ, French JR, Wang Z, Westphal DL, Heymsfield AJ, Twohy CH, Prenni AJ, Prather KA. 2009. *In situ* detection of biological particles in cloud ice-crystals. Nature Geoscience. 2(6):398-401.

Pummer BG, Budke C, Augustin-Bauditz S, Niedermeier D, Felgitsch L, Kampf CJ, Huber RG, Liedl KR, Loerting T, Moschen T et al. 2015. Ice nucleation by water-soluble macromolecules. Atmos Chem Phys. 15(8):4077-4091.

Qiu Y, Odendahl N, Hudait A, Mason R, Bertram AK, Paesani F, DeMott PJ, Molinero V. 2017. Ice nucleation efficiency of hydroxylated organic surfaces Is controlled by their structural fluctuations and mismatch to ice. Journal of the American Chemical Society. 139(8):3052-3064.

Schnell RC, Vali G. 1972. Atmospheric ice nuclei from decomposing vegetation. Nature. 236(5343):163-165.

Schnell RC, Vali G. 1976. Biogenic ice nuclei: Part I. terrestrial and marine Sources. Journal of Atmospheric Sciences. 33(8):1554-1564. English.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: assessing genome assembly and annotation completeness. Methods in Molecular Biology. 1962:227-245. eng.

Shimizu S, Ogawa J, Kataoka M, Kobayashi M. 1997. Screening of novel microbial enzymes for the production of biologically and chemically useful compounds. Advances in Biochemical Engineering/Biotechnology. 58:45-87. eng.

Shimizu S, Yamada H. 1990. Production of dietary and pharmacologically important polyunsaturated fatty acids by microbiological processes. Comments on Agricultural and Food Chemistry. 2(3):211-235.

Spatafora JW, Chang Y, Benny GL, Lazarus K, Smith ME, Berbee ML, Bonito G, Corradi N, Grigoriev I, Gryganskyi A et al. 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. Mycologia. 108(5):1028-1046.

Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve *de novo* gene finding. Bioinformatics. 24(5):637-644.

Summerbell RC. 2005. Root endophyte and mycorrhizosphere fungi of black spruce, *Picea mariana*, in a boreal forest habitat: influence of site factors on fungal distributions. Studies in Mycology. 53:121-145.

Uehling J, Gryganskyi A, Hameed K, Tschaplinski T, Misztal PK, Wu S, Desirò A, Vande Pol N, Du Z, Zienkiewicz A et al. 2017. Comparative genomics of *Mortierella elongata* and its bacterial endosymbiont *Mycoavidus cysteinexigens*. Environmental Microbiology. 19(8):2964-2983.

UniProt Consortium. 2020. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research. 49(D1):D480-D489.

Vali G. 1971. Quantitative evaluation of experimental results on the heterogeneous freezing nucleation of supercooled liquids. Journal of the Atmospheric Sciences. 28(3):402-409.

Vasebi Y, Mechan Llontop ME, Hanlon R, Schmale III DG, Schnell R, Vinatzer BA. 2019. Comprehensive characterization of an aspen (*Populus tremuloides*) leaf litter sample that maintained ice nucleation activity for 48 years. Biogeosciences. 16(8):1675-1683.

Yang S, Vinatzer B. 2021. Draft genome sequence of *Mortierella alpina* strain LL118, isolated from an aspen (*Populus tremuloides*) leaf litter sample. Microbiology Resource Announcements. 10(47):e00864-00821.

Tables

Table 1. List of Mortierellaceae strains tested for ice nucleation activity. Ice nucleation-active strains are marked with a plus (+), ice nucleation inactive strains are marked with a minus (-).

Species	Strain ID	Ice nucleation activity	Synonym
Dissophora globulifera	REB-010B	-	*Mortierella selenospora*
Dissophora ormata	1234	-	*Dissophora ornata*
Entomortierella echinosphaera	1233	-	*Mortierella echinosphaera*
Entomortierella lignicola	JL12	-	*Mortierella selenospora*
Gryganskiella cystojenkinii	1230	-	*Mortierella cystojenkinii*
Linnemannia elongata	NVP64-	-	*Mortierella elongata*
Linnemannia gamsii	AM1032	-	*Mortierella gamsii*
Linnemannia hyalina	AM1038	-	*Mortierella hyalina*
Lunasporangiospora selenospora	1228-	-	*Mortierella selenospora*
Mortierella alpina	AD071	-	*Mortierella alpina*
Mortierella alpina	LL118	+	*Mortierella alpina*
Mortierella alpina	NVP153	+	*Mortierella alpina*
Mortierella alpina	NVP17b	-	*Mortierella alpina*
Mortierella alpina	NVP47	+	*Mortierella alpina*
Mortierella zonata	1226	-	*Mortierella zonata*
Podila humilis	1414-	-	*Mortierella humilis*
Podila minutissima	NVP1	-	*Mortierella minutissima*

Figures

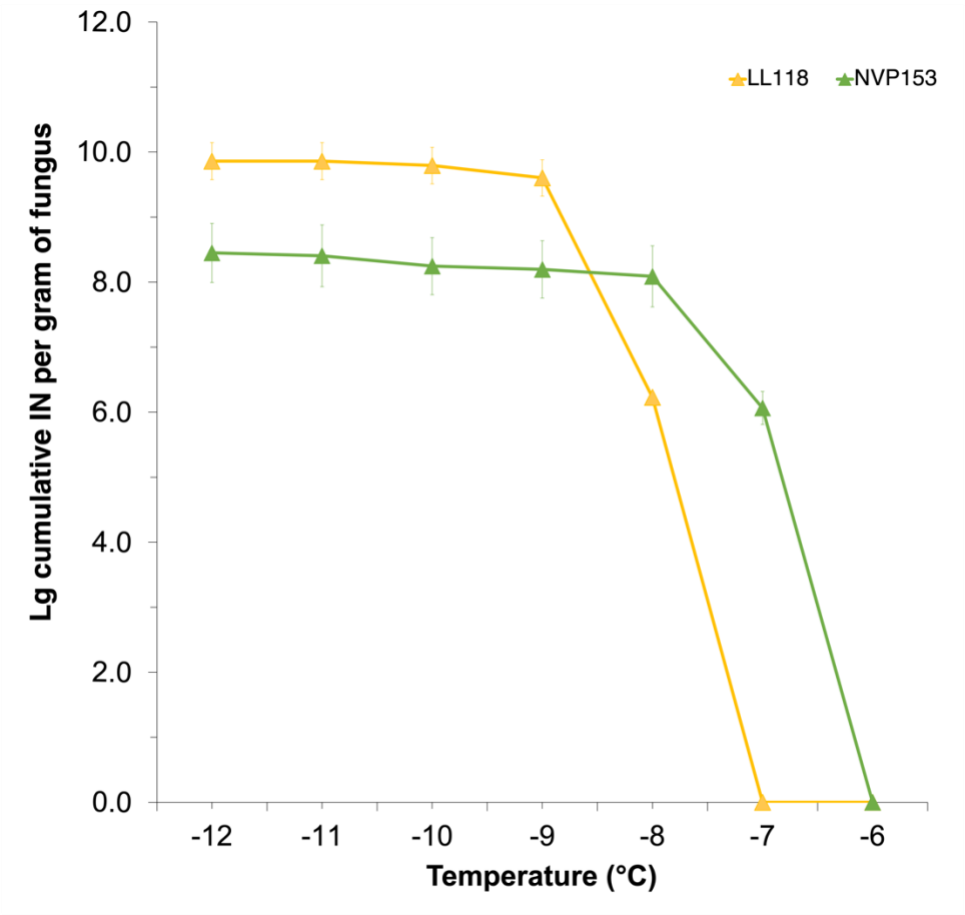
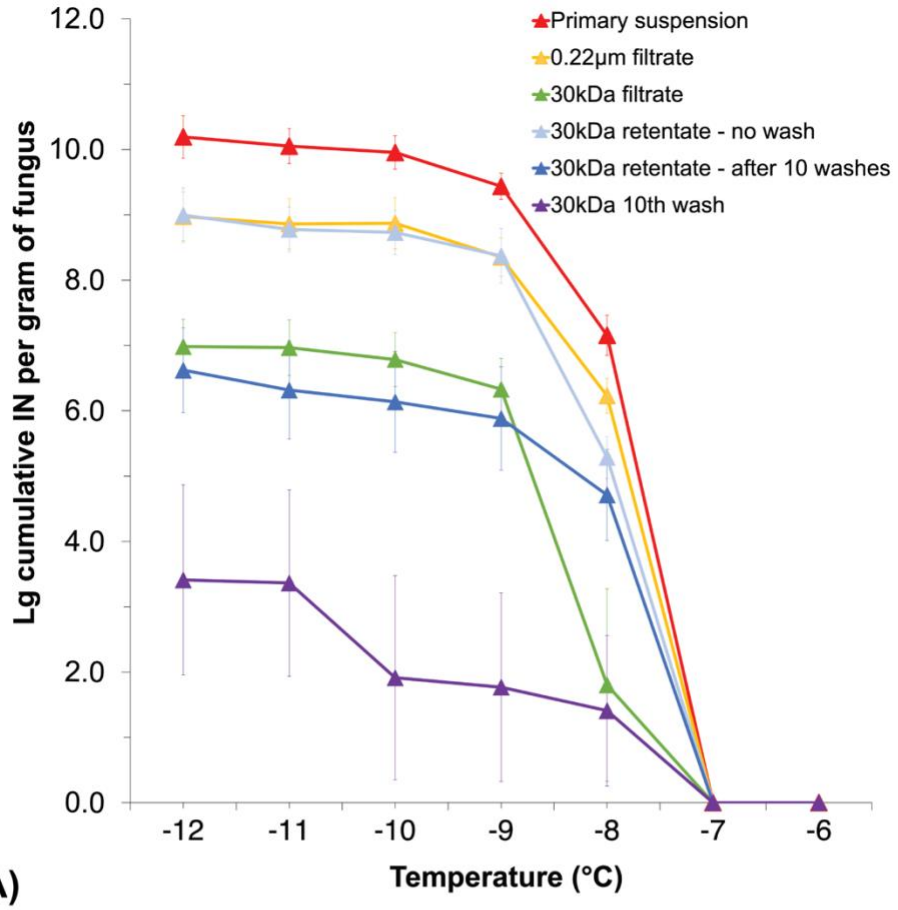
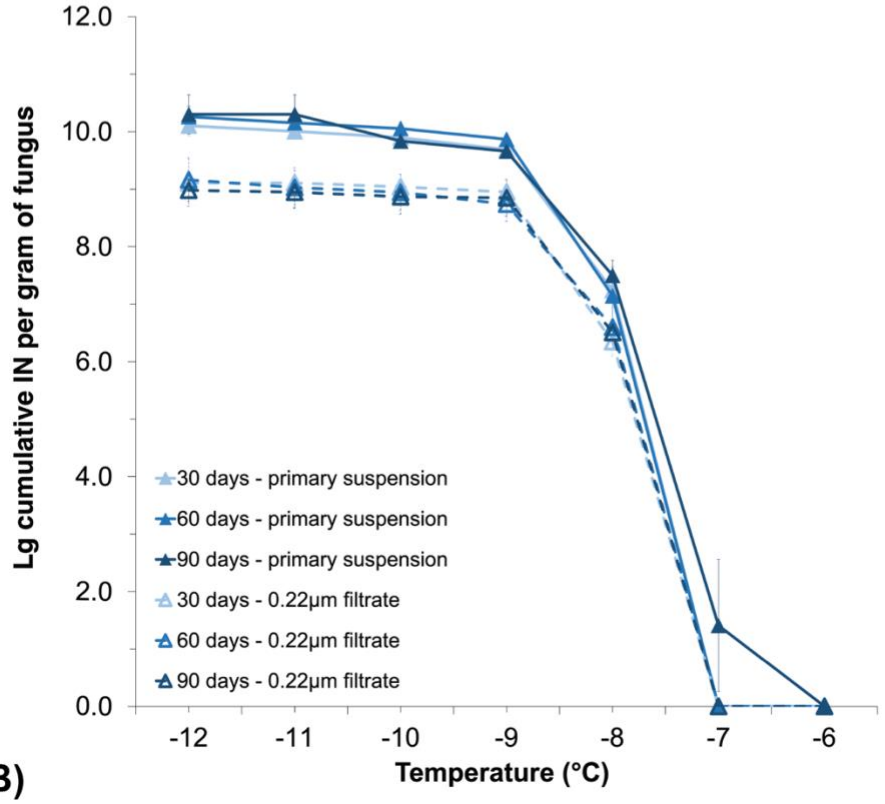


Figure 1. Cumulative ice nucleation spectra of two ice nucleation-active *Mortierella alpina* strains. All cultures were grown at room temperature for 14 days. Results are primary suspensions based on droplet freezing assays at -6, -7, -8, -9, -10, -11, and -12°C. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.

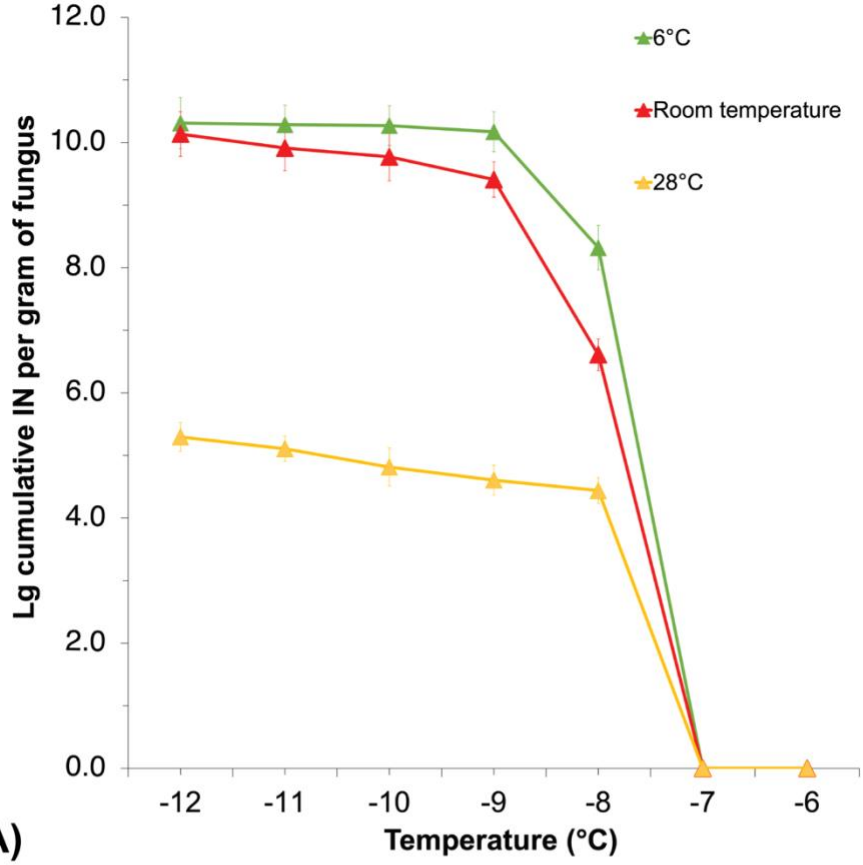

Primary suspension
0.22µm filtrate
30kDa filtrate
30kDa retentate - no wash
30kDa retentate - after 10 washes
30kDa 10th wash
Lg cumulative IN per gram of fungus
Temperature (°C)
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-12
-11
-10
-9
-8
-7
-6

A)

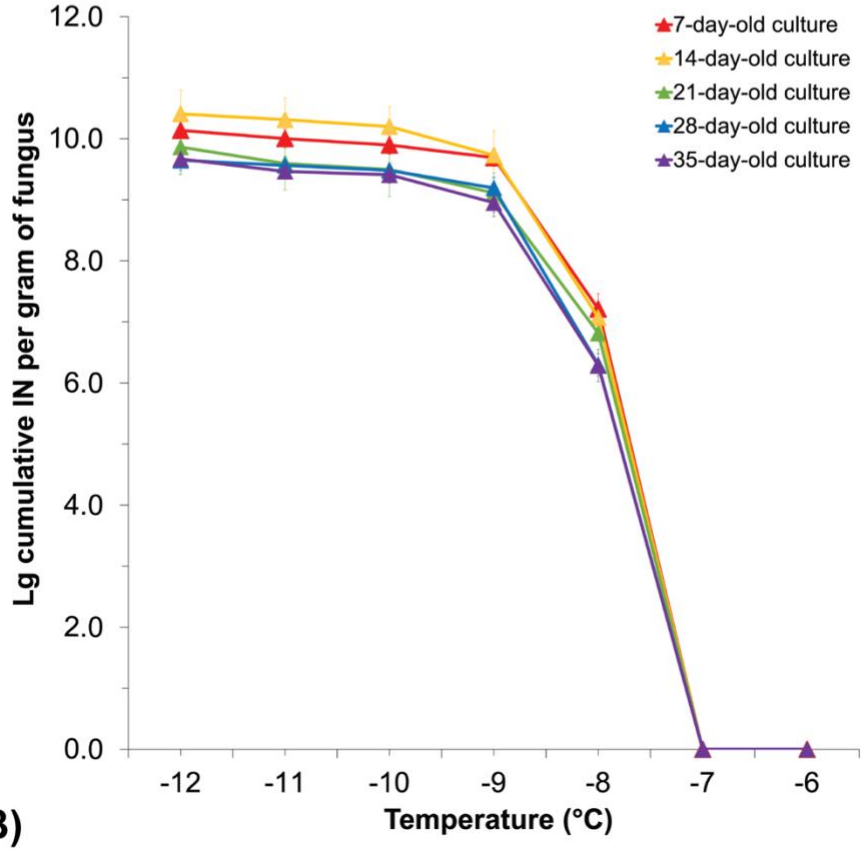

30 days - primary suspension
60 days - primary suspension
90 days - primary suspension
30 days - 0.22µm filtrate
60 days - 0.22µm filtrate
90 days - 0.22µm filtrate
Lg cumulative IN per gram of fungus
Temperature (°C)
12.0
10.0
8.0
6.0
4.0
2.0
0.0
-12
-11
-10
-9
-8
-7
-6

B)

Figure 2. Cumulative ice nucleation spectra of *Mortierella alpina* LL118 grown at room temperature for 14 days. A) Results are primary suspensions, 0.22 μm filtrates, 30 kDa filtrates, original 30 kDa retentates, washed 30 kDa retentates, and last washes based on droplet freezing assays. B) Results are primary suspensions and 0.22 μm filtrates stored at -80°C based on droplet freezing assays. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Lg cumulative IN per gram of fungus
6°C
Room temperature
28°C
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

A)


12.0
10.0
8.0
6.0
4.0
2.0
0.0
Lg cumulative IN per gram of fungus
7-day-old culture
14-day-old culture
21-day-old culture
28-day-old culture
35-day-old culture
-12
-11
-10
-9
-8
-7
-6
Temperature (°C)

B)

Figure 3. Cumulative ice nucleation spectra of *Mortierella alpina* LL118 A) grown at room 6°C, temperature, and 28°C, respectively, for about 30 days; B) grown at room temperature for 7 days, 14 days, 21 days, 28 days, and 35 days, respectively. Results are primary suspensions based on droplet freezing assays. Each data point represents a mean number (±SEM) obtained from three replicates. IN: ice nuclei.

Supplementary tables

Supplementary table 1. List of genes that were upregulated at both 6°C and room temperature in LL118

Supplementary table 2. Phyre2 predictions of genes that were upregulated at both 6°C and room temperature in LL118 with had no annotations or were uncharacterized by InterProScan or BLASTP

Supplementary table 3. List of genes that were present in LL118 and NVP153 but absent from AD071 and NVP17b

Supplementary figures

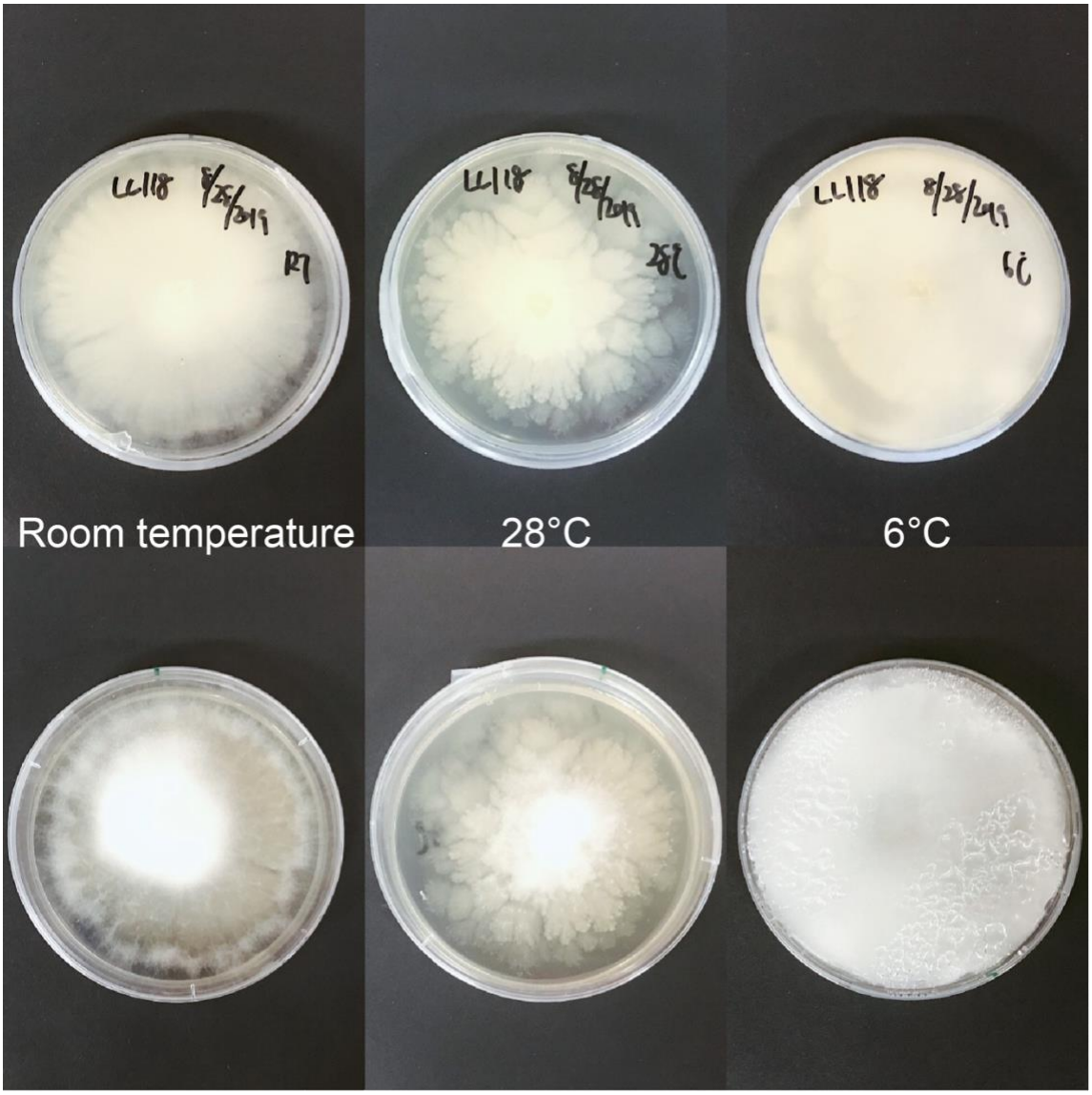


Figure S1. Morphology of ice nucleation spectra of *Mortierella alpina* LL118 grown at room temperature, 28°C, and 6°C, respectively, for about 30 days

Chapter 5. Conclusions and future directions

Detection of boxwood blight

This work suggests that metagenomic sequencing with the ONT MinION is a robust approach to detect *Calonectria pseudonaviculata* (*Cps*) from naturally infected boxwood at the species level with high sensitivity and accuracy, and it has the potential to detect it at the strain level. Therefore, the ONT MinION can be implemented in routine diagnostics of plant fungal pathogens. Suitable DNA extraction methods and bioinformatics tools with appropriate fungal genome databases are critical to detecting the target fungus and were identified here. However, to confirm which exaction method works the best, replicates using the same plants will be needed.

To determine the detection sensitivity, using artificially inoculated plants is desirable. Plants can be inoculated with inoculum ranging from low to high concentrations and subjected to DNA extraction using the same method after the disease progresses. In this way, the lowest inoculum concentration for the pathogen to be identified by metagenomic sequencing using the ONT MinION in combination with different bioinformatic analysess can be determined.

On the other hand, direct comparison with other diagnostics methods is needed to determine relative sensitivity. To obtain meaningful comparisons, more sets of DNA from the same sample and the same extraction method need to be sequenced with the ONT MinION and Illumina and compared to other methods, such as qRT-PCR.

In this study, detecting a fungal pathogen at the strain level was attempted. Fungal pathogens can diverge into numerous within-species clusters with different traits (e.g., virulence). Therefore, it is worth investigating of how to detect fungi at the strain level. So far, strain-level detection has not been fully explored for most fungal pathogens. Improvements in bioinformatics tools and fungal genome databases are needed to achieve within-species identification of fungi.

In conclusion, our study suggests that metagenomic sequencing, especially using the ONT MinION, is a promising technology that could be implemented in routine diagnostics of fungal diseases. There is still room for improvements in bioinformatics tools to increase detection sensitivity and accuracy.

Fungal ice nucleation

The work of investigating ice nucleation activity (INA) in *Fusarium* and *Mortierella* has expanded the knowledge of fungal ice nucleating particles (INPs). The main results of INA in *Fusarium* and *Mortierella* indicated: (1) their INPs appeared to be secreted aggregates smaller than 30 kDa in size and prone to be separated by external forces; (2) their INA was induced at low temperatures; (3) candidate INA genes were more likely to be those upregulated at low temperatures and predicted to be either signal peptides or polyketide non-ribosomal peptides; (4) hydrophobins could be good candidates for fungal INPs. Results were consistent with what Kunert et al. (2019) reported, *i.e.*, that *Fusarium* INPs consist of aggregates smaller than 100 kDa.

To further explore the INA gene(s) responsible for producing INPs in *Fusarium* and *Mortierella*, more experiments and analyses are needed. First, more strains of both species are needed to perform genome-wide association studies and find if any gene(s) or variants are associated with the strength of INA. Second, mutational analyses and gain-of-function experiments are needed by knock-out the candidate gene(s) in ice-nucleating active (Ice^+) strains to see if strains become ice-nucleating inactive (Ice^-), and knock-in of the candidate gene(s) in Ice^- to see if strains become Ice^+.

Another important aspect of investigating fungal INA is to determine its ecological roles. Little is known about the ecological roles of ice nucleation in general. To determine the ecological

role of fungi, experiments using both Ice$^+$ and Ice$^-$ strains can be done after INA genes are identified. Taking *Fusarium* species as the example, *F. avenaceum* and *F. graminearum* occupy similar ecological niches (Coluzza et al. 2017), and *F. graminearum* is generally Ice$^-$ (Pouleur et al. 1992; Kunert et al. 2019). After *Fusarium* INA gene(s) are determined, gene knockouts can be performed for a *F. avenaceum* Ice$^+$ strain so its Ice$^-$ mutate can be obtained. Co-inoculation of soil with (1) a *F. avenaceum* Ice$^+$ strain and a *F. graminearum* strain, (2) the *F. avenaceum* Ice$^-$ mutate and the same *F. graminearum* strain, can be performed at the same incubation condition to investigate the survival of these two species. Thus, whether ice nucleation benefits the survival can be determined.

On the other hand, while in the bioprecipitation theory it has been proposed that INPs can potentially induce precipitation (Sands et al. 1982; Morris et al. 2014), some INPs can be scrubbed from the atmosphere and may not be involved in precipitation (Hanlon et al. 2017). The knowledge of the link between sources and sinks of atmospheric INPs and precipitation is still limited. Since fungal spores can be dispersed and transported into the atmosphere, they may have direct impacts on precipitation. Therefore, investigating INA in fungal spores is needed. If desirable, spores of Ice$^+$ fungal strains need to be separated from mycelium. These spores could then be collected and tested for INA to determine their atmospheric importance.

In conclusion, our study has contributed to uncovering the properties of some fungal INPs and we proposed putative genes and molecules of these INPs. A better understanding of the ecological role of fungal INA and other biological INA will help answer the question of why INA exists in these organisms. Ultimately, the relative contribution of fungal INA to atmospheric processes that affect weather and climate on Earth could then be determined.

References

Coluzza I, Creamean J, Rossi MJ, Wex H, Alpert PA, Bianco V, Boose Y, Dellago C, Felgitsch L, Fröhlich-Nowoisky J et al. 2017. Perspectives on the future of Ice nucleation research: research needs and unanswered questions identified from two international workshops. Atmosphere. 8(8):138.

Hanlon R, Powers C, Failor KC, Monteil C, Vinatzer B, Schmale III D. 2017. Microbial ice nucleators scavenged from the atmosphere during simulated rain events. Atmospheric Environment. 163.

Kunert AT, Pöhlker ML, Tang K, Krevert CS, Wieder C, Speth KR, Hanson LE, Morris CE, Schmale III DG, Pöschl U et al. 2019. Macromolecular fungal ice nuclei in *Fusarium*: effects of physical and chemical processing. Biogeosciences. 16(23):4647-4659.

Morris CE, Conen F, Alex Huffman J, Phillips V, Pöschl U, Sands DC. 2014. Bioprecipitation: a feedback cycle linking Earth history, ecosystem dynamics and land use through biological ice nucleators in the atmosphere. Global Change Biology. 20(2):341-351.

Pouleur S, Richard C, Martin JG, Antoun H. 1992. Ice nucleation activity in *Fusarium acuminatum* and *Fusarium avenaceum*. Appl Environ Microbiol. 58(9):2960-2964. eng.

Sands D, Langhans VE, Scharen AL, Smet G. 1982. The association between bacteria and rain and possible resultant meteorological implications. Journal of the Hungarian Meteorological Service. 86:148-152.

Appendix A: Supplemental material for Chapter 1. Metagenomic sequencing for detection and identification of the boxwood blight pathogen *Calonectria pseudonaviculata*

Shu Yang[1], Marcela A. Johnson[1,2], Mary Ann Hansen[1], Elizabeth Bush[1], Song Li[1], Boris A. Vinatzer[1*]

1 School of Plant and Environmental Sciences, Virginia Tech, Blacksburg, VA, United States

2 Graduate Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, VA, United States

*Corresponding author; email: vinatzer@vt.edu

** Submitted to the journal Scientific Reports in July 2021

Supplementary Tables

Supplementary Table 1. List of genomes used in the extensive custom database withBLASTN, Kraken 2, and/or sourmash

Genome	Strain	Accession number
Calonectria henricotiae	CB077	GCA_004380935.1
	CBS138102	GCA_004380885.1
	NL009	GCA_004380965.1
	NL017	GCA_004382205.1
Calonectria pseudonaviculata	CB002	GCA_004141935.1
	CB002	GCA_006505905.1
	CBS139394	GCA_001696505.1
	CBS139395	GCA_004380915.1
	CBS14417	GCA_004381005.1
	CT13	GCA_004380985.1
	ICMP14368	GCA_004382245.1
	NC-BB1	GCA_004381035.1
	ODA1	GCA_004382225.1
Fusarium graminearum	233423	GCA_000966635.1
	241165	GCA_000966645.1
	CS3005	GCA_000599445.1
	DAOM180378	GCA_001717915.1
	FG078	GCA_006942295.1
	FN009	GCA_900476405.1
	ITEM124	GCA_002352725.1
	MDC_Fg1	GCA_900492705.1
	MDC_Fg13	GCA_901446245.1
	NRRL28336	GCA_001717905.1
	PH-1	GCA_900044135.1
	TaB10	GCA_012959185.1
Pseudonectria foliicola	AR2711	GCA_002911195.1
	JAC18-02	GCA_003693505.1
Pseudonectria buxi	AR2414	GCA_003693545.1
	JAC17-19	GCA_003693515.1

Supplementary Table 2. Fungal hits obtained with BLASTN and Kraken 2

Sample ID	# Total reads	BLASTN							Kraken 2							
		# Hits of *Calonectria henricotiae*	# Hits of *Calonectria pseudonaviculata* (*Cps*)	# Hits of *Fusarium graminearum*	# Hits of *Pseudonectria foliicola*	# Hits of *Pseudonectria buxi*	% Identified reads	% *Cps* hits	# Hits of *Calonectria henricotiae*	# Hits of *Calonectria pseudonaviculata* (*Cps*)	# Hits of *Fusarium graminearum*	# Hits of *Pseudonectria foliicola*	# Hits of *Pseudonectria buxi*	# Hits of non-specific Nectriaceae	% Identified reads	% *Cps* hits
S1	424,599	31	225	54	8,897	1,203	2.452	0.053	320	264	202	1843	1761	10022	3.394	0.062
S2	469,952	10	188	36	5,312	763	1.342	0.040	426	234	107	1284	1113	5314	1.804	0.050
S3	351,180	9	410	18	3,555	471	1.271	0.117	366	400	90	802	704	4170	1.860	0.114
G1	694,805	91	6,817	58	4,638	513	1.744	0.981	913	16,342	316	1,199	1,098	25,572	6.540	2.352
G2	1,965,786	219	20,010	69	8,005	598	1.470	1.018	3,041	53,424	631	2,411	1,871	81,733	7.280	2.718
G3	1,914,124	103	9,117	42	5,777	429	0.808	0.476	2,391	42,315	485	2,278	1,712	86,925	7.111	2.211
G4	2,644,721	60	5,281	53	4,025	280	0.367	0.200	1,561	24,633	480	1,551	1,228	56,420	3.247	0.931
G5	3,721,368	113	10,798	86	10,060	691	0.584	0.290	2,542	40,843	933	3,400	2,552	98,586	4.000	1.098
G6	454,974	38	3,174	21	3,808	269	1.607	0.698	450	8,944	231	1,040	719	25,519	8.111	1.966
S4	2,358,938	73	4,377	313	28,795	16,572	2.125	0.186	3,158	27,282	1,479	8,420	50,680	101,482	8.160	1.157
S5	2,755,500	180	11,002	335	51,566	27,900	3.302	0.399	4,256	33,325	1,650	13,050	51,947	92,271	7.131	1.209
G7	1,343,224	166	13,087	72	17,263	13,392	3.274	0.974	1,266	26,970	466	4,729	61,216	70,771	12.315	2.008
G8	2,736,033	173	14,276	105	21,025	20,780	2.060	0.522	1,903	36,942	889	5,994	95,452	99,711	8.804	1.350
G9	295,648	229	15,178	45	34,849	15,430	22.233	5.134	617	16,082	244	6,986	21,340	36,876	27.785	5.440
G10	541,576	531	49,667	254	38,866	23,074	20.753	9.171	1,644	52,677	824	10,321	34,658	44,071	26.625	9.727
G11	289,025	297	29,440	104	23,887	13,364	23.213	10.186	901	30,723	365	5,359	18,832	26,969	28.769	10.630
G12	885,884	719	69,724	396	50,980	30,563	17.201	7.871	2,438	76,053	1,153	14,107	47,211	60,102	22.696	8.585
G13	277,471	146	12,417	61	9,330	5,516	9.900	4.475	560	16,224	244	2,527	11,523	20,397	18.551	5.847
NC	831,673	5	137	60	207	21	0.052	0.016	103	18	438	44	132	672	0.169	0.002

Supplementary Table 3. The 31 reads of sample S1 that had been identified as *C. henricotiae* when using a custom library were compared against NCBI's entire database using BLASTN and the best hits for each read are shown below. None of the reads was identified as *C. henricotiae* showing that using a larger library would have eliminated the false positives for *C. henricotiae.*

Sequence ID	Accession	Species	Percent identity (%)	Length (bp)	E-value
0d8e0467-c80a-431b-b29c-d72f4d94fa8f	CP017483.1	*Stenotrophomonas* sp.	85.247	3240	0
0ed46597-f80f-402a-87b7-7be3d6bbb8ed	KU668563.1	*Clonostachys rosea*	84.512	1414	0
124c36d8-165a-4c82-9030-36113710ebfe	KF757229.1	*Acremonium chrysogenum*	83.256	2150	0
126aaebc-5f57-478d-8a1b-a30d020613ff	MT447058.1	*Orbiocrella petchii*	78.953	2789	0
1e2fc753-f7fd-4321-8855-6cd0c5ab4b40	NC_043850.1	*Paecilomyces penicillatus*	75.835	1407	7.98E-158
2d363f0a-7175-4b13-a85d-f907435e20f1	CP023323.1	*Cordyceps militaris*	74.481	964	5.44E-90
3262e0b6-f8bc-4093-adf9-d4be35b3d690	KT585676.1	*Lecanicillium saksenae*	79.4	2165	0
3aba7db5-201d-450d-b83c-7709695964dc	NC_043850.1	*Paecilomyces penicillatus*	78.261	2139	0
3e1f7f4a-54e9-463c-9b84-0db5b7956e92	MK213319.1	*Clonostachys rosea*	75.211	2013	0
42b156f6-9ef0-44a5-b9fb-8914fb642537	NC_043850.1	*Paecilomyces penicillatus*	77.275	2275	0
44777610-d88e-4828-8d1d-3ff9b36120e5	KT731105.1	*Nectria cinnabarina*	78.602	2790	0
5316e12b-191e-4e9d-9668-706ec8be082e	XM_024893633.1	*Trichoderma citrinoviride*	75.984	1778	0
6caeeee3-a331-45a9-b101-3c78b57141f9	CP014168.1	*Sphingomonas panacis*	74.589	913	3.93E-79
6efc4b38-515b-4b96-9284-1bebe77298af	LR792747.1	*Metarhizium brunneum*	81.251	4955	0
7568d60a-e985-47c4-ace7-f22e61d6439f	NC_043850.1	*Paecilomyces penicillatus*	81.072	2277	0
863f5f43-ccb0-4c40-ae4c-3cbc1c4915a9	CP052902.1	*Fusarium oxysporum* f. sp. *koae*	72.368	1748	5.78E-107
8b79c6be-8f97-4eb6-85e7-e62975ba0aff	NC_043850.1	*Paecilomyces penicillatus*	80.19	2110	0
950c6db0-18be-4135-9fae-67bc6c6ccabc	XM_025725124.1	*Fusarium venenatum*	88.06	67	7.89E-09
98190aab-298d-4fa4-abc5-18cbe8b6313b	LR792747.1	*Metarhizium brunneum*	81.07	5251	0
a27132f4-0f48-4e33-a1ef-a06746182cae	KU668563.1	*Clonostachys rosea*	80.963	1329	0

a6b89184-2497-48d3-a0ce-e007159b32ac	NC_043850.1	*Paecilomyces penicillatus*	85.686	1537	0
ca4e70c0-6a70-4949-82bc-3bcd08762d93	NC_043850.1	*Paecilomyces penicillatus*	80.244	2293	0
d2ff4fb3-c2b7-43a2-ad13-53389d4a38ef	CP020875.1	*Trichoderma reesei*	77.891	1963	0
d7436683-d95c-421b-8f63-620af96158d2	CP049930.1	*Ustilaginoidea virens*	75.232	1292	9.89E-151
db59b963-441a-4d8e-a100-988235f39aa8	NC_043850.1	*Paecilomyces penicillatus*	82.965	2260	0
de19d969-1acc-4388-8da8-f4519d096555	KP742838.1	*Fusarium mangiferae*	73.01	2249	8.37E-144
e660b8a2-a17b-4f02-abf1-825159431c8f	KU668563.1	*Clonostachys rosea*	84.768	1162	0
f43202f4-32dc-489a-938e-cb48c8f760f8	MT123351.1	*Calonectria ilicicola*	79.068	3153	0

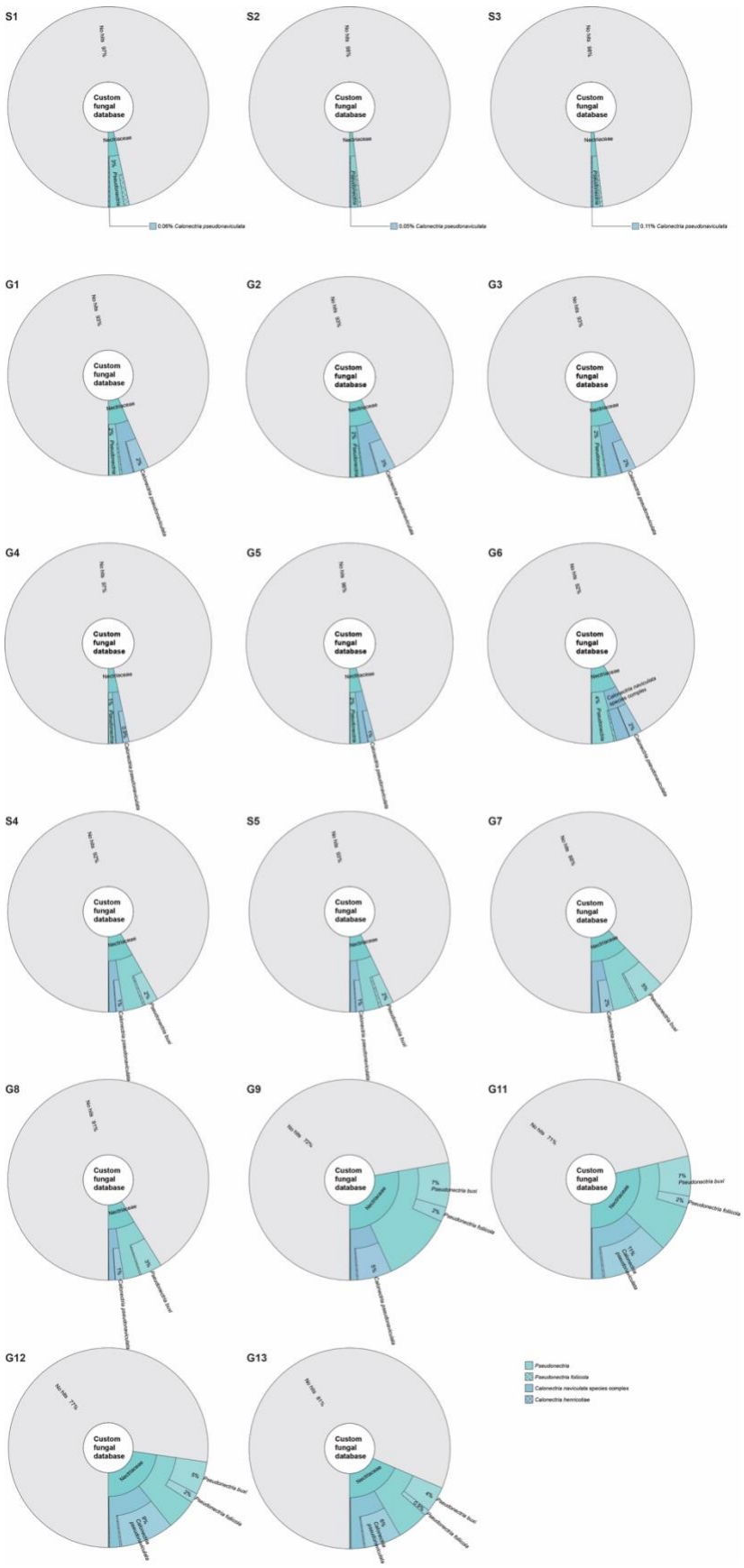
Note that two reads had no hits at all.

Supplementary Table 4. Sequencing statistics of G10 with ONT MinION and Illumina HiSeq

Sample	G10		Negative control	
Platform	Illumina	MinION	Illumina	MinION
# Toatl reads	17,033,700	541,576	271,857,762	831,673
Total read length (bp)	1,498,691,820	1,955,966,076	40,778,664,300	1,945,418,169
Max read length (bp)	100	65,962	150	28,837
Min read length (bp)	2	2	150	4
Avg read length (bp)	88	3,612	150	2,339

Supplementary Figures

Supplementary Figure 1. Diseased boxwood naturally infected with *C. pseudonaviculata*


S1
S2
S3
G1
G2
G3
G4
G5
G6
S4
S5
G7
G8
G9
G11
G12
G13
Custom fungal database
Nectriaceae
Pseudonectria
Calonectria naviculata species complex
Pseudonectria buxi
Pseudonectria foliicola
Calonectria hemricotiae
0.08% Calonectria pseudonaviculata
0.05% Calonectria pseudonaviculata
0.11% Calonectria pseudonaviculata

Supplementary Figure 2. Krona plots based on the percentage reads classified using the custom database with 29 genomes by Kraken 2

Supplementary Results 1

Boxwood infected with *Cps* harbored different prokaryotic communities compared to healthy boxwood

Methods

A database of all NCBI RefSeq genomes in April 2021 was built to profile boxwood-associated prokaryotic communities using Kraken 2 v2.1.1 [1]. Pavian v1.0 [2] was used to visualize the taxonomic classification of the prokaryotic sequences.

Results

Since metagenomics provides a complete picture of all microbes present in a sample, we decided to also take a look at the bacterial communities associated with boxwood. Therefore, Kraken 2 analysis was performed against a database containing all assembled bacterial genomes in RefSeq. MinION reads were combined based on disease severity and DNA extraction method, forming five composite samples: moderately diseased boxwood extracted after sonication, moderately diseased boxwood extracted by grinding, severely diseased boxwood extracted after sonication, severely diseased boxwood extracted by grinding, and healthy boxwood extracted by grinding. Supplementary Figure 3 shows the taxonomic profile of each group.

Because our DNA samples only came from a small number of boxwood samples, this analysis is necessarily descriptive and preliminary. However, some interesting trends were observed. Proteobacteria was the phylum with the highest relative abundance for all five composite samples (Supplementary Figure 3). Only for healthy boxwood, Firmicutes was the second most

abundant phylum (Supplementary Figure 3E). Another interesting feature of the healthy boxwood sample was the relatively high abundance of *Pasteurella multocida* followed by *Vibrio anguillarum* as the second most abundant species.

The only common feature among all symptomatic samples and absent from the healthy plant sample was the presence of *Pseudomonas putida* and *Pseudomonas fulva*. Bacterial communities of the severely diseased samples also included a relatively high abundance of *Pseudomonas monteilii* and *Pseudomonas rizosphereae*. While *Stenotrophomonas maltophilia* and other *Stenotrophomonas* species were relatively abundant in moderately diseased boxwood, the genera *Pantoea* and *Erwinia* were of higher relative abundance in severely diseased boxwood. Some bacterial species, such as *Buchnera aphidicola* and *Massilia oculi*, may have been present due to insects or due to human handling.

Discussion

Although these results are preliminary because of the small number of samples that were examined, the prokaryotic community composition appears to differ between healthy and diseased boxwood. The role of the detected bacterial species in regard to its effect on disease severity will need to be determined. Also, if some boxwood-associated community members may possibly have disease-suppressive ability that could be leveraged for disease management, similar to bacteria previously isolated from recycling irrigation systems [3], should be determined.

References

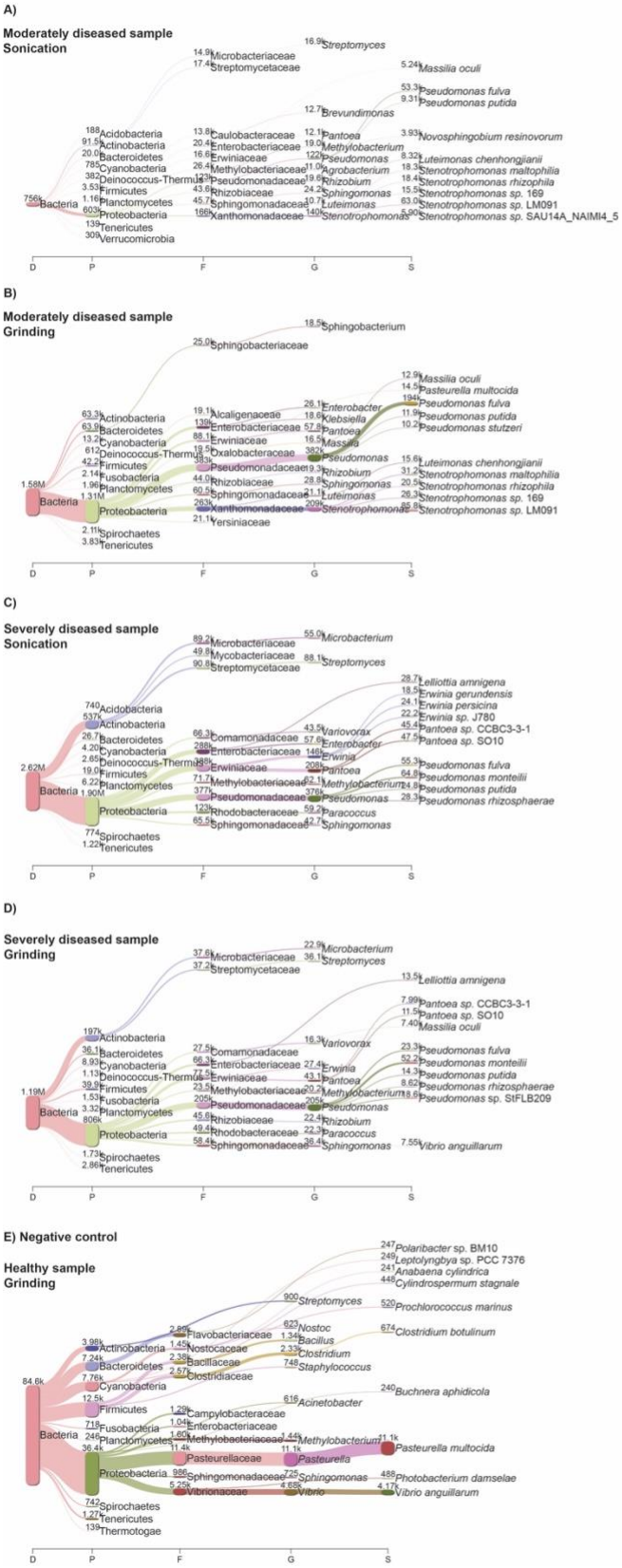
1 Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257, doi:10.1186/s13059-019-1891-0 (2019).

2 Breitwieser, F. P. & Salzberg, S. L. Pavian: interactive analysis of metagenomics data for microbiome studies and pathogen identification. *Bioinformatics* **36**, 1303-1304, doi:10.1093/bioinformatics/btz715 (2020).

3 Yang, X. & Hong, C. Biological control of boxwood blight by *Pseudomonas protegens* recovered from recycling irrigation systems. *Biological Control* **124**, 68-73, doi:https://doi.org/10.1016/j.biocontrol.2018.01.014 (2018).

Supplementary Figure 3. Sankey plots for taxonomic profiling of the metagenome based on reads mapped against RefSeq complete bacterial genomes. The heights of the rectangles indicate the number of reads assigned per taxa and rank, also indicated above/next to each taxa. A) Results of all moderately diseased samples (S1-S3) that were sonicated. B) Results of all moderately diseased samples (G1-G6) that were homogenized in liquid nitrogen. C) Results of all severely diseased samples (S4-S5) that were sonicated. D) Results of all severely diseased samples (G7-G13) that were homogenized in liquid nitrogen. E) Results of the negative control (NC).

See next page for Figure


A)
Moderately diseased sample
Sonication
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Planctomycetes
Proteobacteria
Tenericutes
Verrucomicrobia
Microbacteriaceae
Streptomycetaceae
Caulobacteraceae
Enterobacteriaceae
Erwiniaceae
Methylobacteriaceae
Pseudomonadaceae
Rhizobiaceae
Sphingomonadaceae
Xanthomonadaceae
Streptomyces
Massilia oculi
Pseudomonas fulva
Pseudomonas putida
Brevundimonas
Pantoea
Methylobacterium
Pseudomonas
Agrobacterium
Rhizobium
Sphingomonas
Luteimonas
Stenotrophomonas
Novosphingobium resinovorum
Luteimonas chenhongjianii
Stenotrophomonas maltophilia
Stenotrophomonas rhizophila
Stenotrophomonas sp. 169
Stenotrophomonas sp. LM091
Stenotrophomonas sp. SAU14A_NAIMI4_5
D
P
F
G
S
B)
Moderately diseased sample
Grinding
Bacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
Sphingobacteriaceae
Alcaligenaceae
Enterobacteriaceae
Erwiniaceae
Oxalobacteraceae
Pseudomonadaceae
Rhizobiaceae
Sphingomonadaceae
Xanthomonadaceae
Yersiniaceae
Sphingobacterium
Massilia oculi
Pasteurella multocida
Pseudomonas fulva
Pseudomonas putida
Pseudomonas stutzeri
Enterobacter
Klebsiella
Pantoea
Massilia
Pseudomonas
Rhizobium
Sphingomonas
Luteimonas
Stenotrophomonas
Luteimonas chenhongjianii
Stenotrophomonas maltophilia
Stenotrophomonas rhizophila
Stenotrophomonas sp. 169
Stenotrophomonas sp. LM091
D
P
F
G
S
C)
Severely diseased sample
Sonication
Bacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
Microbacteriaceae
Mycobacteriaceae
Streptomycetaceae
Comamonadaceae
Enterobacteriaceae
Erwiniaceae
Methylobacteriaceae
Pseudomonadaceae
Rhodobacteraceae
Sphingomonadaceae
Microbacterium
Streptomyces
Variovorax
Enterobacter
Erwinia
Pantoea
Methylobacterium
Pseudomonas
Paracoccus
Sphingomonas
Lelliottia amnigena
Erwinia gerundensis
Erwinia persicina
Erwinia sp. J780
Pantoea sp. CCBC3-3-1
Pantoea sp. SO10
Pseudomonas fulva
Pseudomonas monteilii
Pseudomonas putida
Pseudomonas rhizosphaerae
D
P
F
G
S
D)
Severely diseased sample
Grinding
Bacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
Microbacteriaceae
Streptomycetaceae
Comamonadaceae
Enterobacteriaceae
Erwiniaceae
Methylobacteriaceae
Pseudomonadaceae
Rhizobiaceae
Rhodobacteraceae
Sphingomonadaceae
Microbacterium
Streptomyces
Variovorax
Erwinia
Pantoea
Methylobacterium
Pseudomonas
Rhizobium
Paracoccus
Sphingomonas
Lelliottia amnigena
Pantoea sp. CCBC3-3-1
Pantoea sp. SO10
Massilia oculi
Pseudomonas fulva
Pseudomonas monteilii
Pseudomonas putida
Pseudomonas rhizosphaerae
Pseudomonas sp. StFLB209
Vibrio anguillarum
D
P
F
G
S
E) Negative control
Healthy sample
Grinding
Bacteria
Actinobacteria
Bacteroidetes
Cyanobacteria
Firmicutes
Fusobacteria
Planctomycetes
Proteobacteria
Spirochaetes
Tenericutes
Thermotogae
Flavobacteriaceae
Nostocaceae
Bacillaceae
Clostridiaceae
Campylobacteraceae
Enterobacteriaceae
Methylobacteriaceae
Pasteurellaceae
Sphingomonadaceae
Vibrionaceae
Streptomyces
Nostoc
Bacillus
Clostridium
Staphylococcus
Acinetobacter
Methylobacterium
Pasteurella
Sphingomonas
Vibrio
Polaribacter sp. BM10
Leptolyngbya sp. PCC 7376
Anabaena cylindrica
Cylindrospermum stagnale
Prochlorococcus marinus
Clostridium botulinum
Buchnera aphidicola
Pasteurella multocida
Photobacterium damselae
Vibrio anguillarum
D
P
F
G
S