Fungal Ecology 35 (2018) 1–9

Contents lists available at ScienceDirect

Fungal Ecology

journal homepage: www.elsevier.com/locate/funeco

Methodological Advances

Data mining for discovery of endophytic and epiphytic fungal diversity in short-read genomic data from deciduous trees

Nicholas R. LaBonte [a, *], James Jacobs [b], Aziz Ebrahimi [c], Shaneka Lawson [c], Keith Woeste [c]

[a] University of Illinois Department of Crop Sciences, 1201 W. Gregory Dr, Urbana, IL 61801, USA
[b] USDA Forest Service Southwestern Region Forest Health Protection, 333 Broadway SE, Albuquerque, NM 87102, USA
[c] USDA Forest Service Hardwood Tree Improvement and Regeneration Center, 715 W. State Street, West Lafayette, IN 47907, USA

ARTICLE INFO

Article history:
Received 5 September 2017
Received in revised form
24 March 2018
Accepted 10 April 2018

Corresponding Editor: James White Jr.

Keywords:
Endophytes
Epiphytes
Microbiome
Illumina sequencing
Data mining
Metagenomics

ABSTRACT

High-throughput sequencing of DNA barcodes, such as the internal transcribed spacer (ITS) of the 16s rRNA sequence, has expanded the ability of researchers to investigate the endophytic fungal communities of living plants. With a large and growing database of complete fungal genomes, it may be possible to utilize portions of fungal symbiont genomes outside conventional marker sequences for community analysis of short-read data. We designed a bioinformatics pipeline to identify putative fungal coding sequences from 100 bp Illumina reads of DNA extracted from several angiosperm species (*Castanea, Juglans,* and *Ulmus*). Reads remaining after a two-step filtering process made up a small fraction of total reads (2–100 putative fungal reads per 10,000 plant reads) and were assigned to fungal genera and orders based on similarity to proteins from complete fungal genomes. Some of the taxa identified are known to be ubiquitous class 2 endophytes. We detected some differences in endophyte community composition based on ITS sequence data versus results from the short-read pipeline, particularly among *Ulmus.* ITS results in *Juglans* and *Castanea,* however, closely reflected results from the short-read pipeline, and both methods portrayed similar intergeneric differences in endophyte community composition.

1. Introduction

Endophytes and epiphytes belong to a wide range of fungal taxa that colonize above- and belowground portions of their hosts (Carroll, 1988; Rodriguez et al., 2009). Some endophytes, particularly among those symbiotic with grasses (Clay, 1998), confer enhanced tolerance of biotic and abiotic stresses, although phenotypic effects of endophytes are highly variable (Saikkonen et al., 1998). Among forest trees and their symbiotic fungi, few well-documented cases of situational mutualism exist (Faeth and Fagan, 2002). The majority of endophytic fungi in woody plants have no discernible positive or negative effect on host fitness, but the potential for fungal symbionts to confer or enhance essential plant phenotypes like disease resistance or drought tolerance sustains interest in studying them (Arnold, 2007; Rodriguez et al., 2009). Endophytes are difficult to observe *in vivo*, thus studies of endophyte and epiphyte fungal communities have traditionally relied on isolation of fungal symbionts in culture before identification or amplification and sequencing of DNA barcode sequences to assay the fungal community of a plant. Since culturing favors fast-growing fungi and underrepresents endophytes that do not grow in culture, DNA-based methods have increased in popularity (Arnold, 2007; Yahr et al., 2016), in particular, pyrosequencing of barcode amplicons (e.g., Tedersoo et al., 2010).

When genomic DNA is extracted from plant tissue, non-plant DNA from the cells of fungal pathogens (Hsiang and Goodwin, 2003), as well as endophytes - and epiphytes, if the sample is not surface-sterilized - will be present in the plant DNA sample. When whole-genome sequencing is performed on these samples using the Illumina platform, DNA is randomly sheared into fragments and sequenced as short reads. A (small) fraction of those short reads will belong to symbiotic fungi rather than host plants. Aligning to sequence databases to determine whether a given short read is of plant or fungal origin would be impossible if the read originates from repetitive and low-complexity sequence in intergenic regions. For reads that originate from the more conserved coding sequences of genes, however, 100 bases may be sufficient to identify a read's origin using a BLAST-like algorithm to align reads to a database of plant and fungal protein sequences. Identifying fungal reads to

* Corresponding author.
E-mail addresses: nrlabonte@gmail.com, nlabonte@illinois.edu (N.R. LaBonte).

https://doi.org/10.1016/j.funeco.2018.04.004
1754-5048/

order, genus, or species would be most useful for endophyte community profiling. Taxonomic assignment of reads is possible with barcode sequences such as internal transcribed spacers (ITS) which have been sequenced for numerous fungal species. Using predicted protein sequences from genome assemblies, however, has been dismissed as a method for studying fungal symbiont communities since the number of fungal taxa with whole-genome assemblies is too small to accurately represent fungal endophyte diversity. Recent studies however, successfully identified fungal transcripts within Norway spruce (*Picea abies*) and mangrove (*Avicennia marina*) transcriptomic data (Huang et al., 2014; Delhomme et al., 2015), and as more fungal genomes become available, taxonomic limitations may no longer be critical.

The number of sequenced fungal genomes is rapidly increasing. Over 1000 taxa (per GenBank) are represented across the major fungal lineages. Fungal genomes are, in general, much smaller (less than 100 Mb) and more tractable than plant and animal genomes, so sequenced species numbers are likely to increase continuously (Grigoriev et al., 2011). At the same time, whole-genome sequencing is becoming a more routine method for genotyping plants. Leveraging existing bioinformatics tools to identify fungal sequences in raw Illumina sequence data could unlock a trove of unused fungal sequence data from the DNA short reads generated by plant genome sequencing and resequencing projects.

Using a high-throughput sequence aligner, we attempted to profile endophyte communities from several hardwood tree taxa using whole-genome short-read Illumina sequence data from 43 individual trees. The questions we sought to answer in this project were: (1) What percent of reads in tree DNA sequence libraries are likely to come from fungi? (2) What is the taxonomic composition of the fungal reads in tree DNA samples A? and (3) do samples of putative fungal sequences parallel those from previous research on fungal endophytes in woody plants? If the answer to the third question is positive, we would expect to observe known endophyte-rich taxa to be well-represented, differences in putative fungal community composition among individual trees, taxa, and sites, and similar results between DNA-barcode (ITS) methods and the bioinformatics analysis described in this work.

2. Materials and methods

2.1. High-throughput sequencing

Genomic DNA was isolated from fresh leaves (chestnut, *Castanea* spp.), lyophilized leaves (*Castanea mollissima*), and dormant twigs (*Castanea* spp., walnuts (*Juglans* spp.), and American elm, *Ulmus americana*) by grinding tissues in liquid nitrogen with a mortar and pestle, followed by extraction using a CTAB buffer and phenol-chloroform (Table 1). Following quantification using 1.5% agarose gels and a NanoDrop spectrophotometer (Thermo Scientific), shotgun short-read libraries (paired-end 100 bp) were prepared and sequenced using an Illumina Hi-Seq 2500 by the Purdue Genomics Core Facility (https://www.purdue.edu/hla/sites/genomics/). Read trimming, de-multiplexing and initial filtering were also performed by the Genomics Core Facility.

2.2. Database generation and alignment

Our protocol was partly based on the work of Hsiang and Goodwin (2003), who developed a smaller-scale pipeline to identify fungal expressed sequence tags (ESTs) from plant EST data. A database of predicted protein sequences from 225 fungal genomes (Supplementary File 1) was created using the Joint Genome Initiative MycoCosm fungal genomics resource (http://genome.jgi.doe.gov/programs/fungi/index.jsf) (Grigoriev et al., 2011, 2014). The rapid sequence alignment program DIAMOND (Buchfink et al., 2015) was used to align filtered and trimmed DNA reads (.FASTQ format) to the fungal protein database using a BLASTX algorithm with default settings (E-value cutoff for reporting = 0.001). Protein sequences were used instead of nucleotide to avoid spurious alignments of low-complexity DNA in short reads of plant origin to similar fungal-origin DNA. Subsequently, reads with alignments to the fungal database were extracted from the FASTQ reads file, and submitted to a second alignment, also using DIAMOND under the same default settings, to the NCBI nr protein database to verify fungal origin. A read was designated “fungal” if greater than half of the sequences aligned to it (maximum = 25) were of fungal origin. Similarly, a sequence was assigned to Basidiomycota if a majority of aligned sequences belonged to Basidiomycota. Genus and species were assigned to each fungal read based on the best aligned database sequence, based on E-value. The numbers of reads aligning to fungal genera, and to Ascomycota and Basidiomycota in general, were tallied for all 43 samples. Simpson's diversity index was calculated for ascomycete and basidiomycete sequences in each sample, using number of reads as a proxy for the abundance of fungi assigned to a genus. Genus was used as an identifier not because the method can actually identify short reads to genus, but rather to provide a way of grouping the most taxonomically similar reads together, analagous to the OTUs (operational taxonomic units) derived from DNA barcode sequencing. The 40 most strongly-represented genera (representing over 95% of the putative fungal reads identified in all samples) were grouped by order to provide a more realistic taxonomic assignment of putative fungal sequences.

2.3. ITS amplicon sequencing

To validate the results of our short-reads alignment method, we analyzed the ITS fungal barcoding region in (1) an independent set of *Juglans cinerea* and *J. ailantifolia* samples and (2) 15 of the same DNA samples we submitted for whole-genome sequencing and analyzed using the random genomic short reads pipeline. Sequences of amplified ITS regions for six independent samples of walnut (2 each *J. cinerea, Juglans* × *bixbyi,* and *J. ailantifolia;* Table 1) were obtained to compare with the results of the shotgun-sequence analysis. Genomic DNA was extracted from surface-sterilized twigs. Primers ITS-1 (TCCGTAGGTGAACCTGCGG) forward and ITS-4 (TCCTCCGCTTATTGATATGC) were used to amplify the nuclear internal transcribed spacer barcode sequence. A single-step 30-cycle PCR was used for amplification. Amplicons were multiplexed, purified using Agencourt AMPure beads (Agencourt Bioscience Corporation, MA, USA) and sequenced using Roche 454 FLX Titanium equipment and reagents. PCR and sequencing were conducted by MR DNA, Shallowater, TX (www.mrdnalab.com). 454 reads were demultiplexed and processed using the UPARSE OTU identification pipeline of USEARCH (Edgar, 2010, 2013) to quality-filter reads, remove singletons, assemble identical sequences, and derive OTUs.

The same ITS-1 and ITS-4 primers and PCR protocols were used to amplify fungal ITS sequences from the same plant DNA samples that were used to generate whole-genome sequences (Table 1). PCR products were pooled by taxon (2 samples of *U. americana*; 4 samples of *C. mollissima*; *J. cinerea* and *J. bixbyi*; *J. ailantifolia* and *Juglans mandshurica*; 3 samples of *Juglans regia*; 2 hybrids of *J. regia* with other *Juglans*) prior to sequencing using WideSeq, a service provided by the Purdue Genomics Core Facility that uses an Illumina MiSeq to generate short reads. The same UPARSE OTU pipeline (Edgar, 2010, 2013) was used to identify OTUs. Taxonomic classification of OTUs was conducted by submitting to BLASTn using the nr nucleotide database (https://blast.ncbi.nlm.nih.gov/Blast.

Table 1
Short reads aligning to fungal predicted protein sequences from whole-genome short-read datasets of *Castanea*, *Juglans*, and *Ulmus*.

Sample	Species	Tissue	Site[a]	Reads[b]	Fungal/1000[c]
Cd01	*Castanea dentata*	twig	Indiana-1	174966	1.0792
Cd02	*C. dentata*	twig	Indiana-2	242544	0.6950
Cx01	*C. dentata* × (*C. dentata* × *C. mollissima*)	twig	Virginia-1	220927	1.6083
Cx02	*C. dentata* × *C. sativa*	twig	Ohio-1	84512	0.4160
Cx03	*C. mollissima* × (*C. mollissima* × *C. dentata*)	twig	Ohio-1	221726	0.9941
Cx04	*C. mollissima* × (*C. mollissima* × *C. dentata*)	twig	Ohio-1	216963	1.1383
Cx05	*C. mollissima* × (*C. dentata* × *C. sativa*)	twig	Ohio-1	97868	0.4698
Cx06	*C. mollissima* × (*C. dentata* × *C. sativa*)	leaf	Ohio-1	172815	1.6289
Cx07	*C. mollissima* × (*C. dentata* × *C. sativa*)	twig	Ohio-1	145642	1.3971
Cm01	*C. mollissima*	leaf	Ohio-1	36523	1.0270
Cm02[d]	*C. mollissima*	leaf	Guizhou	69016	0.7934
Cm03[d]	*C. mollissima*	leaf	Hebei	21264	0.3400
Cm04[d]	*C. mollissima*	leaf	Shaanxi	100222	1.3473
Cm05[d]	*C. mollissima*	leaf	Yunnan	126268	1.6878
Cm06	*C. mollissima*	twig	Ohio-1	113634	0.7274
Cm07	*C. mollissima*	twig	Ohio-1	92692	0.7528
Cm08	*C. mollissima*	twig	Ohio-1	105482	0.7083
Cm09	*C. mollissima*	leaf	Ohio-1	61039	0.6538
Cm10	*C. mollissima*	leaf	Ohio-1	23730	0.2861
Cm11	*C. mollissima*	twig	Ohio-1	145785	0.6536
Cm12	*C. mollissima*	leaf	Ohio-1	30158	0.2875
Cm13	*C. mollissima*	twig	Ohio-1	26335	0.1286
Cm14	*C. mollissima*	twig	Ohio-1	153296	0.7082
Cm15	*C. mollissima*	twig	Ohio-1	184418	0.9123
Cm16	*C. mollissima*	twig	Ohio-1	134528	0.6064
Cm17	*C. mollissima*	leaf	Ohio-1	13945	0.2487
Cm18	*C. mollissima*	leaf	Ohio-1	47182	0.2265
Cm19	*C. mollissima*	twig	Ohio-1	193441	1.5873
Cm20	*C. mollissima*	twig	Ohio-1	140663	0.6764
Ja[d]	*Juglans ailantifolia*	twig	Indiana-2	126715	2.0513
Jc[d]	*J. cinerea*	twig	Indiana-2	223556	3.0264
Jmaj	*J. major*	twig	Indiana-2	143344	2.9514
Jman[d]	*J. mandshurica*	twig	Indiana-2	384224	3.7778
Jn	*J. nigra*	twig	Indiana-2	1512551	11.4725
Jr01[d]	*J. regia*	twig	Indiana-3	36345	0.8519
Jr02[d]	*J. regia*	twig	Indiana-2	251560	3.0502
Jr03[d]	*J. regia*	twig	Indiana-2	135074	1.4709
Jx01[d]	*J. ailantifolia* × *J. cinerea*	twig	Indiana-2	74346	1.9097
Jx02[d]	*J. regia* × (*J. nigra* × *J. regia*)	twig	Indiana-2	43067	1.0592
Jx03[d]	*J. regia* × *J. cinerea*	twig	Indiana-2	189166	2.8016
Ua01[d]	*Ulmus americana*	twig	Ohio-2	89144	0.2197
Ua02[d]	*Ulmus americana*	twig	Ohio-2	133245	0.2887

[a] For detailed description of sample location, including coordinates, see Table S1.
[b] Short reads with alignments to fungal predicted protein sequences.
[c] Number of short reads aligned to fungal protein database per 1000 short reads of plant origin.
[d] For these samples, ITS barcodes were amplified using PCR in order to compare results with the short-read alignment pipeline.

cgi). The number of reads clustered to each OTU during the de-replication stage of the process was used to rank OTUs as a measure of relative abundance. For WideSeq-generated reads, all reads were submitted to a BLASTn alignment using HS-BLASTN (Chen et al., 2015) against a curated fungal ITS database (Schoch et al., 2014).

To compare taxonomic distribution of putative endophyte and epiphyte fungi identified by the shotgun-sequence based method and the more traditional ITS amplicon method, ITS sequences for the 20 representative species of the most commonly-identified genera from the short-read analysis were obtained from GenBank. Neighbor-joining trees (Saitou and Nei, 1987) were calculated using the Maximum Composite Likelihood method (Tamura et al., 2004) and visualized in MEGA 7.0.14 (Kumar et al., 2015).

3. Results

3.1. Identification of fungal reads from shotgun-sequence data

The number of reads identified as fungal using the two-step alignment process (first to a fungal protein database, and then to the nr protein database) was small, about 0.04–0.09% of reads in *Castanea*, 0.2–0.3% in *Juglans*, and 0.02% in *Ulmus* (Table 1). The second alignment step drastically reduced putative fungal read numbers by eliminating plant sequences from highly conserved genes, which aligned to the fungi-only database; these plant-origin sequences made up 75–90% of the initial putative fungal sequence sets. The highest percentage of fungal reads was found in a *Juglans nigra* sample, while the lowest percentage was found in *C. mollissima*.

3.2. Taxonomic assignment of reads

In most of the samples, many more reads were assigned to Ascomycota than Basidiomycota, *a fortiori* in *Juglans* (4.8 ascomycete reads per basidiomycete on average; range 1.6–12.5) rather than *Ulmus* and *Castanea* (about 2.5 ascomycete reads per basidiomycete on average). Two *Castanea* samples contained considerably more sequences identified to Basidiomycota than Ascomycota; both individuals were complex hybrids (*C. mollissima* × (*dentata* × *sativa*)).

Among ascomycetes, larger percentages of sequences in most samples segregated into the orders Pleosporales, Capnodiales,

Table 2
Most frequently observed orders identified from short reads and ITS sequence operational taxonomic units (OTUs) in *Juglans ailantifolia*, *Juglans cinerea*, and an interspecific hybrid.

Order	*Juglans ailantifolia*			*Juglans mandshurica*		*Juglans cinerea*			*Juglans* × *bixbyi*			ITS-Jc + Jxb
	Ja[a]	ITS-JA133[b]	ITS-JA1097[c]	Jmand	ITS-Ja + Jm	Jc	ITS-JC22	ITS-JC988	Jx01	ITS-1	ITS-2	
Pleosporales	0.41	0.60	0.57	0.35	0.40	0.21	0.49	0.01	0.26	0.58	0.16	0.26
Capnodiales	0.11	0.24	0.11	0.14	0.22	0.05	0.18	0.08	0.08	0.03	0.04	0.15
Chaetothyriales	0.04	0.01	0.02	0.05	0.00	0.18	0.30	0.31	0.04	0.10	0.02	0.10
Xylariales	0.01	0.01	0.13	0.01	0.00	0.01	0.00	0.49	0.01	0.00	0.22	0.00
Helotiales	0.03	0.07	0.00	0.03	0.00	0.01	0.00	0.00	0.02	0.03	0.00	0.00
Tremellales	0.04	0.00	0.00	0.04	0.10	0.05	0.00	0.00	0.02	0.00	0.01	0.13
Taphrinales	0.03	0.00	0.00	0.01	0.07	0.02	0.00	0.00	0.03	0.00	0.01	0.16
Unknown Ascomycete	0.08	0.03	0.14	0.10	0.03	0.04	0.03	0.02	0.11	0.00	0.26	0.01

[a] Percent of putative fungal-origin 100 bp Illumina reads from plant tissue sample aligning to a given fungal order.
[b] Percent of ITS short reads from Rosemont, MN accession aligning to OTUs identified to the given order.
[c] Percent of ITS short reads from West Lafayette, IN accession aligning to OTUs identified to the given order.

Chaetothyriales, Dothideales, Eurotiales, and "Ascomycota *incertae sedis*," which was mostly accounted for by sequences similar to *Oidiodendron* sp. (Leotiomycetes) and fungal sp. 11243, an unclassified fungus isolated from decaying leaf litter (Matsui et al., 2015). Taphrinales appeared to be an unusually well-represented component in *Ulmus*. Among basidiomycetes, Agaricales, Polyporales, Tremellales, Ustilaginales, and Uredinales accounted for the largest portion of putative fungal sequences (Table S2). Notably, Ustilaginales accounted for a larger proportion of putative fungal reads in *Castanea* samples that contained more basidiomycete than ascomycete sequences.

When classification was carried out at the genus level, *Bipolaris*, *Phaeosphaeria* (*Parastagonospora*), *Setosphaeria*, and *Pyrenophora* were the most abundant genera in samples dominated by Pleosporales (Table S3). In samples where Capnodiales accounted for more reads than Pleosporales, *Aureobasidium* and *Baudoinia* were among the most abundant genera. In samples with a strong Chaetothyriales component (particularly some *Juglans*), *Cyphellophora*, *Cladophialophora*, *Exophiala*, and *Capronia* accounted for relatively large (1000s-10000s) numbers of reads. *Colletotrichum* and *Fusarium* also accounted for large numbers of reads in most samples. Among basidiomycetes, there was less variation among samples and taxa in the most frequently observed genera. *Hebeloma*, *Melampsora*, *Dichomitus*, *Pisolithus*, and *Tilletiaria* tended to be the most abundant genera, except for the Ustilaginales-rich chestnut samples, which contained more reads from *Anthracocystis*, *Moesziomyces*, and *Ustilago*. *Cryptococcus* (Tremellales) was the most abundant basidiomycete genus in some samples, particularly walnuts. In one chestnut sample, *Gloeophyllum* (Polyporales) was by far the most frequently observed basidiomycete; and made up a diminutive number of reads in every other sample. In ITS sequencing results, Tremellales and Erythrobasidiales were frequently observed but Agaricales and Polyporales were almost entirely absent (Table S4).

3.3. ITS data from Juglans and comparison with WGS method

Using USEARCH, 55 OTUs were identified in the Minnesota *J. ailantifolia* (JA-ITS-1) sample, 34 from the Indiana *J. ailantifolia* (JA-ITS-2) sample, 30 from JC-ITS-1, and 23 from JC-ITS-2. The number of OTUs identified from contigs of 10 reads or more was 14 for JA-ITS-1, 9 for JA-ITS-2, 9 for JC-ITS-1, 9 for JC-ITS-2, 13 for JCxS-1, and 12 for JCxS-2. When OTU consensus sequences were aligned to the nucleotide nr NCBI database using BLASTn, the taxonomic profile of both *J. ailantifolia* samples was similar, dominated by Pleosporales (Table 2). The taxonomic makeup of OTUs from *J. cinerea* and *J.* × *bixbyi* varied more between IN and MN samples than between two samples of *J. ailantifolia* from the same pair of locations (Table 2). Taxonomic profiling of pooled ITS clones derived from the same *J. ailantifolia, J.* × *bixbyi,* and *J. cinerea* samples that were used for whole-genome short read profiling led to somewhat similar results: a majority of ITS reads from *J, ailantifolia* and the closely related *J. mandshurica* aligned to Pleosporales, but only 26% of ITS reads from a pooled sample of *J. cinerea* and *J.* × *bixbyi* aligned to Pleosporales (Table 2). While Chaetothyriales made up 10% of the ITS reads in this *J. cinerea* sample, Taphrinales (16%) and Tremellales

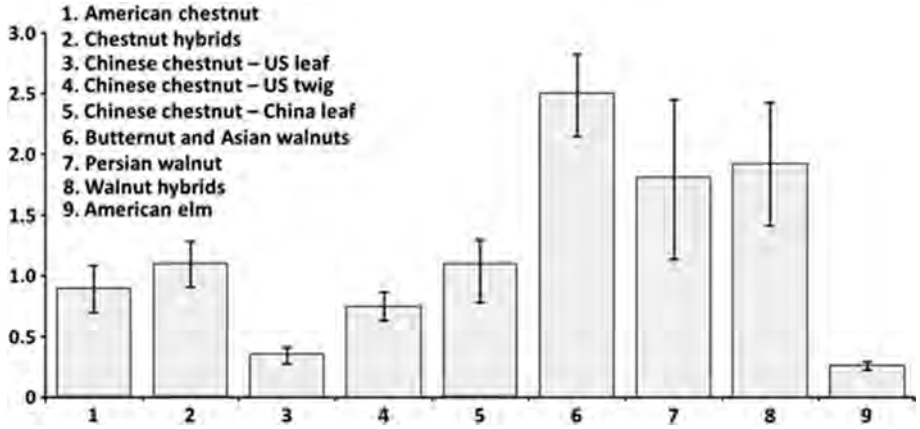

Fig. 1. Fungal reads identified per 1000 plant genome short reads averaged over samples for *Castanea*, *Juglans*, and *Ulmus*. Error bars display standard error of the mean. *Juglans* section Rhysocaryon (average 7.3 fungal reads/1000) is not displayed.

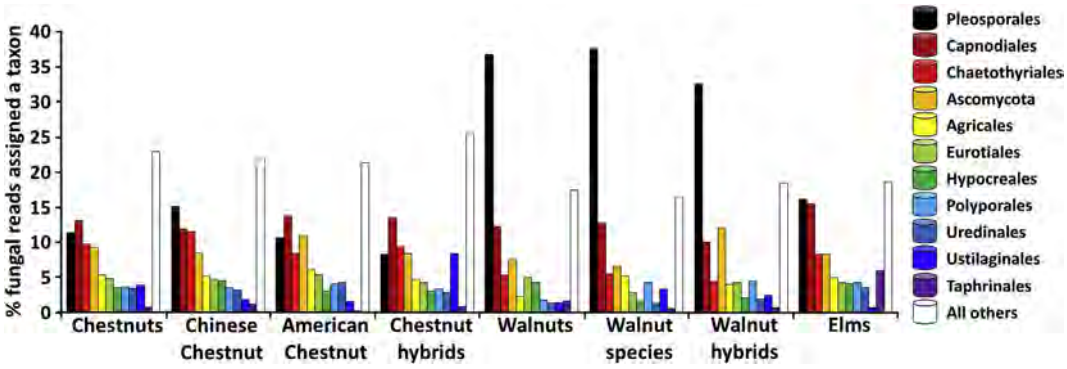

Fig. 2. Percent of fungal reads from whole-genome sequence data from three genera and interspecific hybrids (Table 1) assigned to the 11 most frequently observed taxonomic units. "Ascomycota" indicates reads identified to Ascomycota, but not to any order within Ascomycota.

(12%) were better-represented than in the short-reads analysis. Pooled ITS clones from two American elms were mostly (>50%) aligned to Capnodiales, quite different from the results of the whole genome short-read analysis (Figs.2 and 3) but results for Taphrinales in elm were notably similar (Table 3, Fig. 3). For *J. regia* and hybrids, ITS results mirrored results from the whole-genome short read analysis: Pleosporales dominated fungal sequences, followed by Capnodiales, with other orders rarely identified. When a neighbor-joining tree was built using OTU consensus sequences from *J. ailantifolia, J. cinerea*, hybrid, and representative ITS sequences from the most common genera assigned using the short-read pipeline, OTUs composed of the highest numbers of reads tended to group with the most common taxa identified in the short-read pipeline (Fig. 4). This result was true particularly for *J. ailantifolia* but also in *J. cinerea*, where a larger proportion of fungal reads were identified from the Chaetothyriales (Fig. 4).

4. Discussion

As fungal whole-genome and predicted protein sequences accumulate in public sequence databases, our ability to identify fungal endophytes from shotgun-sequenced plant genomes will improve. Our results show it is currently possible to obtain a coarse-scale snapshot of the fungal endophyte community of deciduous trees using a bioinformatics pipeline based on an initial alignment to a relatively small fungal protein database selected for taxonomic breadth and a second alignment to the NCBI nr protein database for more precise taxonomic identification and for filtering out plant reads that may have aligned to the fungal protein database. When short reads from Japanese walnut (*J. ailantifolia*) and butternut (*J. cinerea*) were identified to order using this alignment pipeline, orders comprising the most reads corresponded to the most commonly identified orders based on ITS pyrosequencing. In particular, fungi from the Chaetothyriales were found in ITS sequences of all three *J. cinerea* samples, but were nearly absent in all three *J. ailantifolia* ITS samples, including the same *Jc* and *Ja* genomic DNA samples used for whole-genome sequencing. Very few short reads were identified from Chaetothyriales in our *J. ailantifolia* whole-genome data, whereas 20% of short reads from *J. cinerea* were identified to Chaetothyriales. ITS results from a subsample of Chinese chestnut (*C. mollissima*) depicted an endophyte community dominated in nearly equal proportions by Pleosporales, Capnodiales, and Chaetothyriales, similar to the short-reads analysis, (Figs. 2 and 3). For *Ulmus* genomic DNA samples, the short-reads analysis derived a significantly lower estimate of the Capnodiales component of the endophyte community than ITS; Capnodiales dominated fungal ITS sequences from the same *Ulmus* DNA samples to an extent not seen in the short-reads results. Quantitative differences in representation of fungal taxa between the ITS and short-reads method may be the result of biases in the barcode amplification process (Bellemain et al., 2010). Both ITS and the short-reads analysis detected a small but notable *Aspergillus* (Eurotiales) component in Persian walnut (*J. regia*) samples (Table 3) and *Taphrina* (Taphrinales) in *Ulmus*. Based on branch lengths among known fungal genera and ITS OTUs in neighbor-joining phylogenetic trees (Figs. 3 and 4), as well as the genus

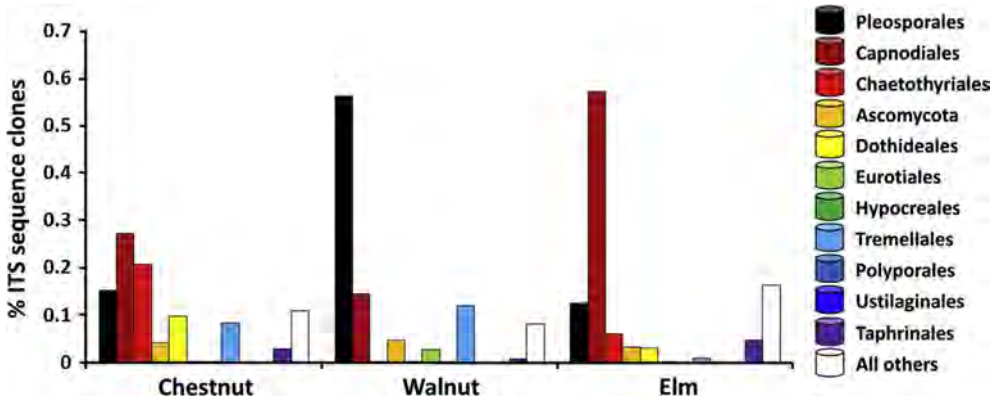

Fig. 3. Percent of ITS sequence clones from pooled samples of individual chestnut (*Castanea mollissima*) walnut (*Juglans regia*) and elm (*Ulmus americana*) (Table 1) assigned to frequently observed taxonomic units from the short-read analysis (Fig. 2). Agaricales and Uredinales were not significantly represented in ITS results. "Ascomycota" indicates reads identified to Ascomycota, but not to any order within Ascomycota.

Table 3
Fungal genera and orders most frequently assigned as the top hit to whole-genome sequence (WGS) short reads from *Juglans regia*, *Castanea mollissima,* and *Ulmus americana* with comparison to results of ITS sequencing. Order was determined using the NCBI Taxonomy browser (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser).

ITS genus[a]	order	%[b]	WGS reads genus[c]	order	%[d]
Juglans regia					
Alternaria	Pleosporales	0.49	*Bipolaris*	Pleosporales	0.11
Cladosporium	Capnodiales	0.20	*Parastagonospora*	Pleosporales	0.10
Vishniacozyma	Tremellales	0.17	*Pyrenophora*	Pleosporales	0.08
Didymella	Pleosporales	0.14	*Aspergillus*	Eurotiales	0.11
Stagonosporopsis	Pleosporales	0.06	*Colletotrichum*	Glomerellales	0.05
Plectosphaerella	Incertae sedis	0.06	*Setosphaeria*	Pleosporales	0.05
Aspergillus	Eurotiales	0.04	*Pestalotiopsis*	Xylariales	0.01
Ulmus americana					
Pseudotaeniolina	Capnodiales	0.16	*Zymoseptoria*	Capnodiales	0.05
Teratosphaericola	Capnodiales	0.14	*Aureobasidium*	Dothideales	0.07
Teratosphaeria	Capnodiales	0.14	*Taphrina*	Taphrinales	0.06
Cladosporium	Capnodiales	0.13	*Dothistroma*	Capnodiales	0.04
Alternaria	Pleosporales	0.04	*Bipolaris*	Pleosporales	0.10
Taphrina	Taphrinales	0.04	*Cladophialophora*	Chaetothyriales	0.02
Strelitziana	Chaetothyriales	0.03	*Pseudocercospora*	Capnodiales	0.03
Castanea mollissima					
Cladosporium	Capnodiales	0.21	*Aureobasidium*	Dothideales	0.10
Trichomerium	Chaetothyriales	0.11	*Sphaerulina*	Capnodiales	0.04
Aureobasidium	Dothideales	0.08	*Cladophialophora*	Chaetothyriales	0.05
Alternaria	Pleosporales	0.08	*Bipolaris*	Pleosporales	0.07
Filobasidium	Filobasidiales	0.06	*Parastagonospora*	Pleosporales	0.04
Camptophora	Chaetothyriales	0.05	*Baudoinia*	Capnodiales	0.03
Taphrina	Taphrinales	0.03	*Oidiodendron*	incertae sedis	0.02

[a] Genus assigned to ITS reads by BLASTn alignment with a curated fungal ITS database.
[b] Percent of ITS WideSeq reads assigned to genus.
[c] Genus assigned to random Illumina short reads by BLASTx alignment to fungal protein sequence databases.
[d] Percent of short reads identified to Ascomycota assigned to genus, averaged over individual samples (3 for *Juglans regia*, 2 for *Ulmus americana*,4 for *Castanea mollissima*).

identification of ITS reads from the sequenced plant samples, the short-read pipeline was generally not able to identify reads to genus, although some genera (*Cryptococcus, Taphrina* and *Aureobasidium*) were frequently identified using both methods. Branch lengths between OTUs and known fungal ITS sequences were generally equal to or longer than branch lengths between known genera. The best-represented genera in the short-reads analysis made up a much smaller percentage of the total number of sequences than the best-represented ITS sequences (Table 3) across all samples, which may indicate that the short-reads analysis assigned reads from a single fungal taxon to several closely-related genera.

Some bias towards fungal taxa that are better-represented in public sequence repositories may have affected results from the short-read method. Xylariales was a major component of *J. cinerea* ITS sequences in the sample from Indiana (Table 2), but was very rarely the assigned order for fungal short reads. A survey of GenBank in January 2017 indicated that genome data from nine Xylariales accessions was available, while there were 35 Pleosporales genomes available. Taxa that were under-represented in databases may have been under-represented in taxonomic profiles based on random short reads; the most abundant taxa in databases might be somewhat over-identified in random short read results. The best-represented fungal orders in GenBank, Hypocreales (109 genomes) and Eurotiales (79 genomes), however, were only rarely identified in short-read data from the taxa in the study; while ITS results corroborated the rarity of Hypocreales, the number of reads identified to this order was considerably lower using ITS (Figs. 2 and 3). It is also possible that the near-absence of Xylariales was based on individual differences in the endophyte mycobiota of *J. cinerea*, as neither JC-ITS-1 nor JC-ITS-2 were the same tree as the *J. cinerea* whole-genome sample. While biases due to the taxonomic limitations of fungal whole-genome sequences may have been an issue for the short-read pipeline, the method also eliminated some potential sources of error that exist in more conventional barcode pyrosequencing methods for fungal community profiling. Since there is no PCR step, biases introduced by the PCR process, which can have a large effect on OTU identification in pyrosequencing results (Bellemain et al., 2010; Lee et al., 2012) may be reduced when fungi are identified from shotgun-sequenced short reads (Sharpton, 2014). Some combination of these differences may have led to the large discrepancy observed between ITS results and short-read results for *Ulmus*, where Capnodiales comprised 60% of ITS sequences but only 15% of short reads; Capnodiales (23 genomes as of 1/31/2017) is somewhat less well-represented in GenBank whole-genome sequences than Pleosporales (35 genomes).

There was considerable variability in fungal endophyte communities among individual species within *Juglans* and *Castanea,* as well as among individuals sampled from the same species, but differences among genera were nevertheless apparent. *Ulmus* tended to have a much lower proportion of fungal/plant short reads than the other genera. This is probably due to the autopolyploid (tetraploid) nature of the two *U. americana* that were sequenced. Tetraploid elms have approximately twice as much plant DNA per nucleus as the diploid walnuts and chestnuts, and the volume of fungal reads identified in elm was approximately half that of several chestnut samples (Table 1). *Juglans* samples on the other hand, contained more putative fungal reads than *Castanea* or *Ulmus* on average (Fig. 1), and the fungal communities of walnuts tended to be more thoroughly dominated by the most common taxa (usually Pleosporales). *Castanea* tended to have a larger number of rare taxa identified from short-read data, and the most common grouping made up a relatively small percentage of fungal reads. The same general pattern of diversity was observed by analysis of ITS clones and whole-genome short reads. While identification to genus is uncertain based on short-read data, several of the most commonly-observed fungal genera revealed in short reads from *Castanea* were known endophytes of woody plants. Genera identified from short reads containing known endophyte species

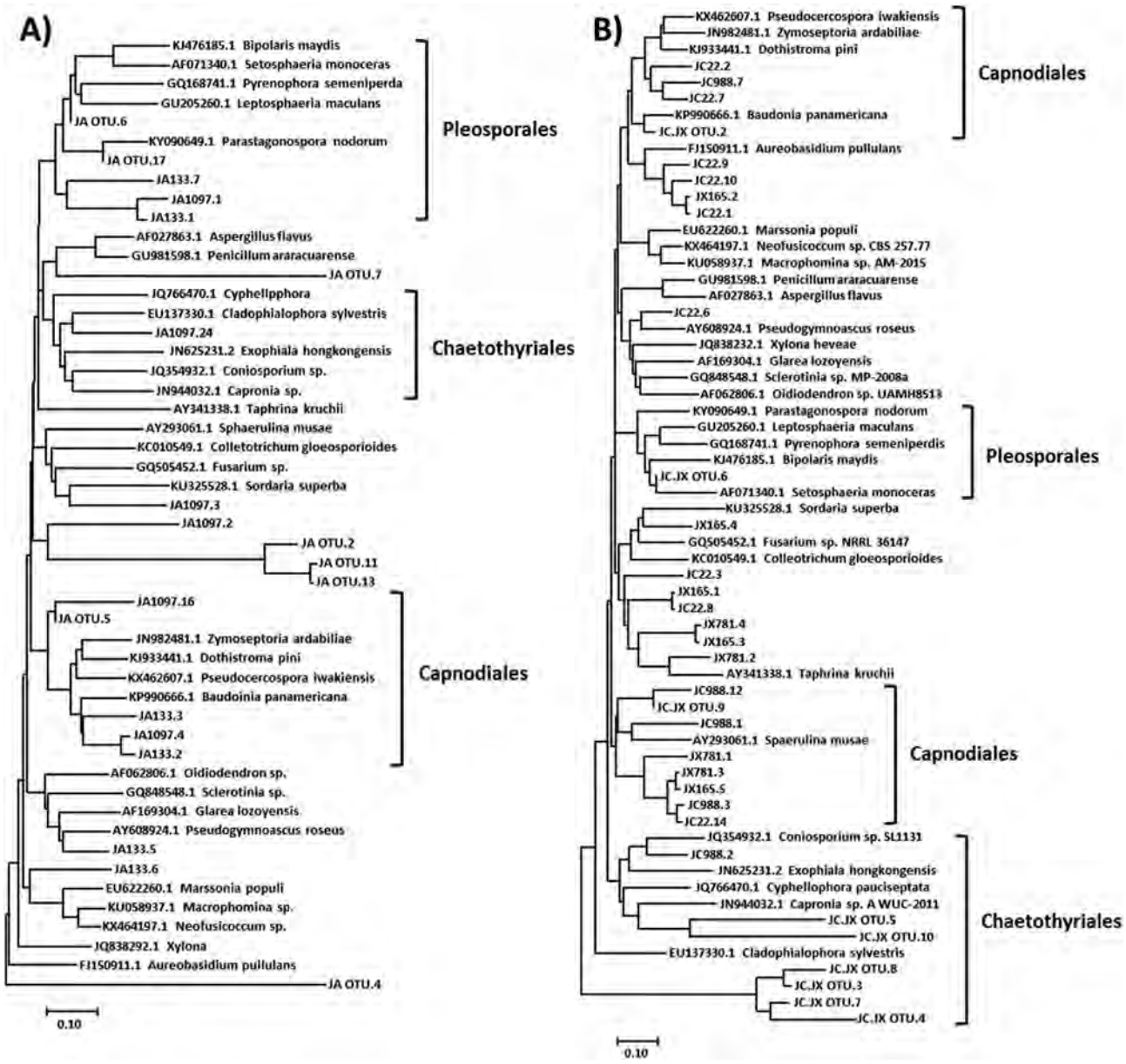

Fig. 4. **(A)** Neighbor-joining tree (computed in MEGA) showing evolutionary relationships of ITS sequence OTUs from three accessions of *Juglans ailantifolia* with database ITS sequences representing the most frequently observed fungal genera from short-read alignments to fungal predicted protein databases. The number next to each OTU represents in size rank (1 = largest number of reads aligned to OTU). "JA OTU" indicates ITS sequences from the same sample that was used in the short reads analysis. **(B)** Neighbor-joining tree (computed in MEGA) showing evolutionary relationships of ITS sequence OTUs from two accessions of *Juglans cinerea* with database ITS sequences representing the most frequently observed fungal genera from short-read alignments to fungal predicted protein databases. The number next to each OTU represents in size rank (1 = largest number of reads aligned to OTU). "JC.JX OTU" indicates pooled ITS sequences from the same DNA samples used in the short-reads analysis.

included *Bipolaris* (Orlandelli et al., 2012; Ali et al., 2016), *Oidiodendron* (Couture et al., 1983; Palmer et al., 2008), *Aureobasidium* (Pugh and Buckley, 1971; Khan et al. 2016), *Parastagonospora* (Sanzio Pimenta et al., 2012), *Cladophialophora* (Usui et al., 2016), and *Cryptococcus* (Santos Gai et al., 2009).

Perhaps the most notable difference between ITS and whole-genome short read results was a shift in the best-represented basidiomycete taxa, from Agaricales, Polyporales, and Uredinales (short-read analysis) to Tremellales and Erythrobasidiales (ITS). Agaricales and Polyporales include many close fungal associates of angiosperm trees: *Hebeloma* (Agaricales), the main basidiomycete genus identified from chestnut samples using the short-read method, has been identified as an ectomycorrhizal symbiont of American chestnut (Bauman et al., 2008). Horizontal gene transfer (Richards et al., 2009; Yue et al., 2012) between trees and fungal associates (wood decay fungi and ectomycorrhizae) may explain why alignment of plant and fungal short reads to a fungal gene database identified Agaricales and Polyporales, but ITS amplicon sequencing of the same DNA samples did not. Although it is perhaps more likely that differences in results between ITS and the short-reads pipeline are due to biases associated with the PCR process (Bellemain et al., 2010), such as differences in primer specificity or amplicon size among taxa, future improvements to the short-reads pipeline could include a step where gene sequences with evidence of horizontal gene transfers are excluded from the initial alignment.

Interspecific hybridization is prevalent in many temperate forest tree taxa, so for this study we obtained whole-genome sequences for several interspecific hybrids. In some studied cases, such as the *Quercus grisea* × *Quercus gambelii* hybrid complex, the endophytic community of hybrids is intermediate between parent species and shows a linear relationship with parental species phenotype (Gaylord et al., 1996). In *Populus* hybrids, infection rates of a common endophyte were positively associated with genetic introgression from one of the hybrid parents (Bailey et al., 2005). In both of these cases, hybrid parents were naturally co-occurring, but the hybrids analyzed here were all the offspring of a native North American species (e.g., *J. cinerea, Castanea dentata*) and an introduced tree (e.g. *J. ailantifolia, C. mollissima*). Endophyte communities of introduced trees may differ markedly from native species growing in the same locations, potentially including a higher proportion of common, cosmopolitan species (Hoffman and Arnold, 2008) and may differ from the endophyte community observed in the species' native range (Fisher et al., 1993). The dominant endophyte orders identified for *C. mollissma* samples from China in our study differed from *C. mollissima* grown in the United States (Table S2). The fungal communities of hybrids did not appear to be a simple average or have a linear relationship with tree species genome content. Two multispecies hybrid chestnuts (*C. dentata* × (*sativa* × *mollissima*)) had an unusually large number of reads identified to Ustilaginales (Table S2), indicating a fungal association arising in hybrids that was not found in the surrounding individuals of the parental species. Yeasts from the Ustilaginales have been documented as endophytes (Abdel-Motaal et al., 2009). The ecological effects of novel endophyte associations in chestnut hybrids could affect the ecological success of chestnut blight-resistant backcrossed chestnuts (Burnham et al., 1986) if they are restored to the eastern North American deciduous forest.

5. Conclusions

The bioinformatics pipeline described here (available from N. LaBonte by request; github.com/nrlabonte) has the potential to identify DNA sequences of fungal symbionts from short-read transcriptome and whole-genome data of plants. Although only a small portion of any fungal genome is recovered, a picture of the fungal symbiont community of the sequenced plant can be obtained that corresponds closely with results from widely-used ITS amplicon sequencing. As more fungal genomes are sequenced, the power of this technique to identify fungal species present in a plant sample will increase. In cases where fungal endophyte communities are of interest, our method allows investigators to profile endophytes in plant sequence data without additional PCR or sequencing efforts.

Acknowledgements

The genome sequencing of chestnuts described here would not have been possible without the financial support of a Frederick M. Van Eck research Scholarship at Purdue University, along with small grants from the American Chestnut Foundation, Northern Nut Growers Association, and the Indiana Academy of Sciences; the authors thank these institutions for their support.

Mention of a trademark, proprietary product, or vendor does not constitute a guarantee or warranty of the product by the U.S. Department of Agriculture and does not imply its approval to the exclusion of other products or vendors that also may be suitable.

Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.funeco.2018.04.004.

References

Abdel-Motaal, F.F., El-Zayat, S.A., Kosaka, Y., El-Sayed, M.A., Nassar, M.S.M., Ito, S., 2009. Four novel ustilaginomycetous anamorphic yeast species isolated as endophytes from the medicinal plant *Hyoscyamus muticus*. Asian J. Plant Sci. 8 (8), 526–535.

Ali, L., Khan, A.L., Hussain, J., Al-Harrasi, A., Wagas, M., Kang, S.M., Al-Rawahi, A., Lee, I.J., 2016. Sorokiniol: a new enzymes inhibitory metabolite from a fungal endophyte Bipolaris sorokiniana LK12. BMC Microbiol. 16 (103).

Arnold, A.E., 2007. Understanding the diversity of foliar endophytic fungi: progress, challenges, and frontiers. Fungal Biol. Rev. 21 (2), 51–66.

Bailey, J.K., Deckert, R., Schweitzer, J.A., Rehill, B.J., Lindroth, R.L., Gehring, C., Whitham, T.G., 2005. Host plant genetics affect hidden ecological players: links among *Populus*, condensed tannins, and fungal endophyte infection. Can. J. Bot. 83, 356–361.

Bauman, J.M., Keiffer, C.H., McCarthy, B.C., Hiremath, S., 2008. Belowground ectomycorrhizal symbiosis of American chestnut (Castanea dentata) following Ohio coal mine reclamation. In: Proceedings 93rd ESA Annual Meeting. Milwaukee, WI, August 3-8 2008.

Bellemain, E., Carlsen, T., Brochmann, C., Coissac, E., Taberlet, P., Kauserud, H., 2010. ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. BMC Microbiol. 10, 189.

Buchfink, B., Xie, C., Huson, D., 2015. Fast and sensitive protein alignment using DIAMOND. Br. J. Pharmacol. 12, 59–60.

Burnham, C.R., Rutter, P.A., French, D.W., 1986. Breeding blight-resistant chestnuts. Plant Breed. Rev. 4, 347–397.

Carroll, G., 1988. Fungal endophytes in stems and leaves: from latent pathogen to mutualistic symbiont. Ecology 69 (1), 2–9.

Chen, Y., Ye, W., Zhang, Y., Xu, Y., 2015. High speed BLASTN: an accelerated MegaBLAST search tool. Nucleic Acids Res. 43 (16), 7762–7768.

Clay, K., 1998. Fungal endophytes of grasses: a defensive mutualism between plants and fungi. Ecology 69 (1), 10–16.

Couture, M., Fortin, J.A., Dalpe, Y., 1983. Oidiodendron griseum Robak: an endophyte of ericoid mycorrhiza in Vaccinium spp. New Phytol. 95 (3), 375–380.

Delhomme, N., Sundstrom, G., Zamani, N., Lantz, H., Lin, Y.-C., Hvidsten, T.R., et al., 2015. Serendipitous meta-transcriptomics : the fungal community of Norway spruce (*Picea abies*). PLoS One 10 (9) e0139080.

Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26 (19), 2460–2461.

Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. Br. J. Pharmacol. 10 (10), 996–998.

Faeth, S.H., Fagan, W.F., 2002. Fungal endophytes: common host plant symbionts but uncommon mutualists. Integr. Comp. Biol. 42 (2), 360–368.

Fisher, P.J., Petrini, O., Sutton, B.C., 1993. A comparative study of endophytes in leaves, zylem and bark of *Eucalyptus nitens* in Australia and England. Sydowia 45 (2), 338–345.

Gaylord, E.S., Preszler, R.W., Boecklen, W.J., 1996. Interactions between host plants, endophytic fungi, and a phytophagous insect in an oak (*Quercus grisea* x *Q. gambelii*) hybrid zone. Oecologia 105, 336–342.

Grigoriev, I.V., Cullen, D., Goodwin, S.B., Hibbett, D., Jeffries, T.W., Kubicek, C.P., Kuske, C., Magnuson, J.K., Martin, F., Spatafora, J.W., Tsang, A., Baker, S.E., 2011. Fuelling the future with fungal genomics. Mycology 2 (3), 192–209.

Grigoriev, I.V., Nikitin, R., Haridas, S., Kuo, A., Ohm, R., Otillar, R., Riley, R., Salamov, A., Zhao, X., Korzeniewski, F., Smirnova, T., Nordberg, H., Dubchak, I., Shabalov, I., 2014. MycoCosm portal: gearing up for 1000 fungal genomes. Nucleic Acids Res. 42 (1), D699–D704.

Hoffman, M.T., Arnold, E., 2008. Geographic locality and host identity shape fungal endophyte communities in cupressaceous trees. Mycol. Res. 112 (3), 331–334.

Hsiang, T., Goodwin, P.H., 2003. Distinguishing plant and fungal sequences in ESTs from infected plant tissues. J. Microbiol. Meth. 54, 339–351.

Huang, J., Lu, X., Zhang, W., Huang, R., Chen, S., Zheng, Y., 2014. Transcriptome sequencing and analysis of leaf tissue of *Avicennia marina* using the Illumina platform. PLoS One 9 (9), e108785.

Khan, A.L., Al-Harrasi, A., Al-Rawahi, A., Al-Farsi, Z., Al-Mamari, A., Waqas, A., Asaf, S., Elyassi, A., Mabood, F., Shin, J.-H., Lee, I.-J., 2016. Endophytic fungi from Frankincense Tree improves host growth and produces extracellular enzymes and indole acetic acid. PLoS One 11 (6), e0158207. https://doi.org/10.1371/journal.pone.0158207.

Kumar, S., Stecher, G., Tamura, K., 2015. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol. Biol. Evol. 33 (7), 1870–1874.

Lee, C.K., Herbold, C.W., Polson, S.W., Wommack, K.E., Williamson, S.J., McDonald, I.R., Cary, S.C., 2012. Groundtruthing next-gen sequencing for microbial ecology- biases and errors in community structure estimates from PCR amplicon pyrosequencing. PLoS One 7 (9) e44224. https://doi.org/10.1371/journal.pone.0044224.

Matsui, M., Yokoyama, T., Nemoto, K., Kumagai, T., Terai, G., Arita, M., Machida, M., Shibata, T., 2015. Genome sequence of Fungal Species No. 11243, which produces the antifungal antibiotic FR901469. Genome Announc. 3 (2) e00118–15.

Orlandelli, R.C., Alberto, R.N., Rubin Filho, C.J., Pamphile, J.A., 2012. Diversity of endophytic fungal community associated with Piper hispidum (Piperaceae) leaves. Genet. Mol. Res. 11 (2), 1575–1585.

Palmer, J.M., Lindner, D.L., Volk, T.J., 2008. Ectomycorrhizal characterization of an

American chestnut (Castanea dentata) dominated community in western Wisconsin. Mycorrhiza 19, 27–36.
Pugh, G.J.F., Buckley, N.G., 1971. Aureobasidium pullulans: an endophyte in sycamore and other trees. Trans. Br. Mycol. Soc. 57 (2), 227–231.
Richards, T.A., Soanes, D.M., Foster, P.G., Leonard, G., Thornton, C.R., Talbot, N.J., 2009. Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfers between plants and fungi. Plant Cell 21 (7), 1897–1911.
Rodriguez, R.J., White Jr., J.F., Arnold, A.E., Redman, R.S., 2009. Fungal endophytes: diversity and functional roles. New Phytol. 182 (2), 314–330.
Saikkonen, K., Faeth, S.H., Helander, M., Sullivan, T.J., 1998. Fungal endophytes: a continuum of interactions with host plants. Annu. Rev. Ecol. Systemat. 29, 319–343.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Santos Gai, C., Teixeira Lacava, P., Maccheroni Jr., W., Glienke, C., Luiz Araujo, W.L., Miller, T.A., Azevedo, J.L., 2009. Diversity of endophytic yeasts from sweet orange and their localization by scanning electron miscroscopy. J. Basic Microbiol. 49 (5), 441–451.
Sanzio Pimenta, R., Moreira da Silva, J.F., Buyer, J.S., Janisiewicz, W.J., 2012. Endophytic fungi from plums (*Prunus domestica*) and their antifungal activity against Monilinia fruticola. J. Food Protect. 75 (10), 1883–1889.
Schoch, C.L., Robbertse, B., Robert, V., et al., 2014. Finding Needles in Haystacks: Linking Scientific Names, Reference Specimens and Molecular Data for Fungi. Database (Oxford) 2014: bau061.
Sharpton, T.J., 2014. An introduction to the analysis of shotgun metagenomic data. Front. Plant Sci. 5, 209.
Tamura, K., Nei, M., Kumar, S., 2004. Prospects for inferring very large phylogenies using the neighbor-joining method. Proc. Natl. Acad. Sci. Unit. States Am. 101, 11030–11035.
Tedersoo, L., Nilsson, R.H., Jairus, T., Sadam, A., Saar, I., Bahram, M., Bechem, E., Chuyong, G., Koljag, U., 2010. 454 pyrosequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. New Phytol. 188 (1), 291–301.
Usui, E., Takashima, Y., Narisawa, K., 2016. *Cladophialophora inabaensis* sp. nov., a new species among the dark septate endophytes from a secondary forest in Tottori, Japan. Microb. Environ. 31 (3), 357–360.
Yahr, R., Schoch, C.L., Dentinger, T.M., 2016. Scaling up discovery of hidden diversity in fungi: impacts of barcoding approaches. Phil. Trans. Roy. Soc. Bot. 371, 201503356.
Yue, J., Hu, X., Sun, H., Yang, Y., Huang, J., 2012. Widespread impact of horizontal gene transfer on plant colonization of land. Nat. Commun. 3, 1152. https://doi.org/10.1038/ncomms2148.