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.