MYCOLOGICAL RESEARCH XXX (2009) 1–18

British Mycological Society promoting fungal science

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

Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples

Jana M. U'REN[a], *James W. DALLING*[b], *Rachel E. GALLERY*[b,2], *David R. MADDISON*[c], *E. Christine DAVIS*[d,1], *Cara M. GIBSON*[c], *A. Elizabeth ARNOLD*[a,*]

[a]*Division of Plant Pathology and Microbiology, Department of Plant Sciences, 303 Forbes Building, University of Arizona, Tucson, AZ 85721 USA*
[b]*Department of Plant Biology, University of Illinois Urbana-Champaign, Urbana, IL 61801 USA*
[c]*Department of Entomology, University of Arizona, Tucson, AZ 85721 USA*
[d]*Department of Biology, Duke University, Durham, NC 27708 USA*

ARTICLE INFO

Article history:
Received 28 March 2008
Received in revised form
27 October 2008
Accepted 13 November 2008
Corresponding Editor: John Dighton

Keywords:
Alignment
Ancestral state reconstruction
Ascomycota
Barro Colorado Island
Endophytes
Phylogeny

ABSTRACT

Fungi associated with seeds of tropical trees pervasively affect seed survival and germination, and thus are an important, but understudied, component of forest ecology. Here, we examine the diversity and evolutionary origins of fungi isolated from seeds of an important pioneer tree (*Cecropia insignis*, *Cecropiaceae*) following burial in soil for five months in a tropical moist forest in Panama. Our approach, which relied on molecular sequence data because most isolates did not sporulate in culture, provides an opportunity to evaluate several methods currently used to analyse environmental samples of fungi. First, intra- and interspecific divergence were estimated for the nu-rITS and 5.8S gene for four genera of *Ascomycota* that are commonly recovered from seeds. Using these values we estimated species boundaries for 527 isolates, showing that seed-associated fungi are highly diverse, horizontally transmitted, and genotypically congruent with some foliar endophytes from the same site. We then examined methods for inferring the taxonomic placement and phylogenetic relationships of these fungi, evaluating the effects of manual *versus* automated alignment, model selection, and inference methods, as well as the quality of BLAST-based identification using GenBank. We found that common methods such as neighbor-joining and Bayesian inference differ in their sensitivity to alignment methods; analyses of particular fungal genera differ in their sensitivity to alignments; and numerous and sometimes intricate disparities exist between BLAST-based *versus* phylogeny-based identification methods. Lastly, we used our most robust methods to infer phylogenetic relationships of seed-associated fungi in four focal genera, and reconstructed ancestral states to generate preliminary hypotheses regarding the evolutionary origins of this guild. Our results illustrate the dynamic evolutionary relationships among endophytic fungi, pathogens, and seed-associated fungi, and the apparent evolutionary distinctiveness of saprotrophs. Our study also elucidates the diversity, taxonomy, and ecology of an important group of plant-associated fungi and highlights some of the advantages and challenges inherent in the use of ITS data for environmental sampling of fungi.

* Corresponding author. Tel.: +520 621 7212.
E-mail address: Arnold@ag.arizona.edu
[1] Current address: Department of Biology and Health Sciences, Meredith College, Raleigh, NC 27607 USA.
[2] Current address: Department of Zoology, University of Oxford, Oxford OX1 3PS UK.

0953-7562/$ – see front matter © 2008 The British Mycological Society. Published by Elsevier Ltd. All rights reserved.
doi:10.1016/j.mycres.2008.11.015

Introduction

Fungi and oomycetous pathogens are the major cause of seed mortality in soil for a variety of tropical pioneer trees, which depend on recruitment from seed banks to colonize gaps and other disturbances in intact forests (Hall & Swaine 1980; Alvarez-Buylla & Martínez-Ramos 1990; Dalling *et al.* 1997, 1998; Murray & Garcia 2002; O'Hanlon-Manners & Kotanen 2004; Gallery *et al.* 2007a, 2007b). Yet some fungi recovered from seeds are consistently associated with higher-than-average germination success, suggesting a positive role in seed survival or the process of germination itself (Gallery *et al.* 2007a). In the soil of lowland forests in Panama and Costa Rica, >80 % of intact seeds of the neotropical pioneer tree *Cecropia* (*Cecropiaceae*) are infected by cultivable *Ascomycota* (Gallery *et al.* 2007a, 2007b). These fungi are distributed heterogeneously and appear to be host-generalists with regard to infecting various pioneer species (Kluger *et al.* 2008), but they have differential effects on the species they infect (Gallery *et al.* 2007a, 2007b). The scale of diversity of seed-associated fungi, their transmission patterns, and their evolutionary relationships with other fungal guilds are unclear.

The aim of this study was to examine the diversity, taxonomic composition, and evolutionary origins of fungi associated with seeds of a representative neotropical pioneer tree. Our approach was based on isolating fungi from surface-sterilized seeds of *Cecropia insignis* following incubation of seeds for five months in the forest soil at Barro Colorado Island, Panama. Because fungi recovered in culture almost never sporulated, we sequenced the fast-evolving nu-rITS and the intervening 5.8S gene for all isolates, and used standard approaches for evaluating the identity, diversity, and evolutionary relationships of these fungi (e.g. BLAST searches to assign putative identity, designation of operational taxonomic units based on sequence similarity measures, and phylogenetic inference). However, like many researchers examining unknown fungi, we found that current methods have limitations that, to our knowledge, have not been addressed empirically in the literature.

To address our biological questions carefully, we used our dataset of 527 seed-associated fungi and their molecular sequence data to examine a series of methodological questions in the study of fungal environmental samples. We focused on ITS data, which often are informative at the species level for fungi, and have been proposed as the locus of choice for 'barcoding' environmental fungal samples (see http://www.allfungi.com/its-barcode.php). ITS data are appealing for environmental sampling of fungi because of the ease with which they can be amplified and their prevalence in large public databases (e.g. Lutzoni *et al.* 2004). However, these data are notoriously difficult to align across diverse taxa, precluding robust phylogenetic species concepts for diverse samples or distinctive fungi with few known close relatives (e.g. Vandenkoornhuyse *et al.* 2002; Schadt *et al.* 2003; Arnold *et al.* 2007). Despite this difficulty, many studies use ITS data in three primary ways.

First, some use percent similarity measures to delimit ITS genotype groups, and then use those groups as operational taxonomic units (OTU) for estimating diversity and other ecological parameters (e.g. Arnold & Lutzoni 2007). OTU are a proximate, but limited, solution for categorizing environmental samples, such as clones or *mycelia sterilia*, in the absence of reliable phylogenetic species concepts (e.g. O'Brien *et al.* 2005; Arnold & Lutzoni 2007; Hoffman & Arnold 2008). In most studies, authors do not justify their use of a given degree of sequence divergence for estimating species boundaries and instead rely upon an established level of ITS similarity (e.g. 95 %) (but see Arnold *et al.* 2007). A superior approach is the use of robust, multi-locus, phylogenetic species concepts (*sensu* Taylor *et al.* 2000), but this is largely untenable in environmental studies, which typically consist of single-locus datasets, geographically restricted sampling, and a single-guild focus (e.g. seed-associated fungi, which may include only a few members of a given fungal lineage, and may exclude closely related species or strains with different ecological modes). Because the utility of ITS for species-level diagnoses varies among clades (Seifert *et al.* 2007), empirical solutions, such as lineage-specific estimation of percent sequence divergence within and among species, are needed for rapidly estimating the number of distinctive biological units. To our knowledge no study has compared current software applications for assigning genotype groups, such as DOTUR (Schloss & Handelsman 2005) and Sequencher (as used by Arnold *et al.* 2007), prompting our empirical assessments of these tools.

Second, many studies estimate, with variable precision and accuracy, the taxonomic placement of unknown fungi using rapid distance-based and non-phylogenetic algorithms (Davis *et al.* 2003; Hogberg & Land 2004; Arenz *et al.* 2006). For example, taxonomic identification of sterile and/or uncultivable fungi from the environment often relies on BLAST searches of the NCBI GenBank database (Altschul *et al.* 1990) and/or comparisons with other databases [e.g. the Assembling the Fungal Tree of Life (AFToL) database: http://aftol.org] using BLAST or FASTA (Pearson 1998, 2000; Geml *et al.* 2005: http://www.borealfungi.uaf.edu). Although limitations of BLAST-based identification are widely acknowledged, most published accounts focus on the problems posed by the undersampling of fungal diversity and the prevalence of unidentified and misidentified sequences in public databases (e.g. Bridge *et al.* 2003; Harris 2003; Vilgalys 2003; Nilsson *et al.* 2005; but see Arnold *et al.* 2007), rather than on the quality of identification should all sequences in the database be reliably identified and representative of the diversity in a group. Complete representation of fungi in public databases remains a distant and perhaps unrealistic goal, but databases populated by reliably identified fungi, although geographically, ecologically, or taxonomically limited, are becoming a reality (e.g. the AFToL and UNITE databases; Kõljalg *et al.* 2005). An important step forward in environmental studies of fungi is to understand the quality of BLAST-based identification when matches are made to named taxa, leading us to explicitly address this issue in a phylogenetic framework.

Third, a growing number of studies use ITS data to infer the phylogenetic relationships of unknown fungi within particular genera, often noting the superiority of phylogenetic analyses over non-phylogenetic matching algorithms (Henry *et al.*

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

2000; Davis *et al*. 2003; Denman *et al*. 2003; Promputtha *et al*. 2007). However, these studies differ markedly in the methods used to align sequence data and the inference methods used to reconstruct phylogenetic relationships. Even when only closely related fungi are considered in ITS-based environmental samples, automated multiple-sequence alignments are often unsatisfactory, leading to variation among researchers in manual editing of alignments and the exclusion of ambiguous regions (Morrison & Ellis 1997; Landan & Graur 2007). Different alignments can produce different tree topologies, but the sensitivity of different inference methods to particular alignment characteristics is not clear (see Morrison & Ellis 1997; Lambert *et al*. 2003). In particular, it would be helpful to examine the effects of manual editing, the exclusion of ambiguous regions, model selection, and inference approaches on the resulting phylogenetic trees for environmental sampling data, and perhaps most importantly, to understand the sensitivity of particular fungal genera to particular alignment and inference methods.

To our knowledge, no study has critically evaluated this suite of methods – from empirical estimation of taxonomic boundaries using ITS data to the impact of different phylogenetic inference approaches – for a focal group of fungi. Here, we use a large data set of ITS sequences from cultivable fungi associated with seeds of *Cecropia insignis* to address the diversity, taxonomic composition, and evolutionary origins of fungi associated with this important neotropical pioneer tree. We first estimate percent ITS sequence divergence for conspecifics, sister taxa, and non-sister taxa in four representative genera of Ascomycota that are common among seed-associated fungi and for which published phylogenies exist (*Botryosphaeria, Colletotrichum, Mycosphaerella,* and *Xylaria*). We then apply these sequence divergence values to estimate species boundaries for seed-associated fungi representing the same classes of Pezizomycotina (Dothideomycetes and Sordariomycetes), which together represent >90% of cultivable seed-associated fungi at our study site (Gallery *et al*. 2007a). We use our results to estimate richness, diversity, transmission patterns, and congruence with another well-sampled guild at the same site – fungal endophytes from foliage, which previous studies have suggested are similar to seed-associated fungi (Gallery *et al*. 2007a). We evaluate methods for inferring the phylogenetic relationships of these unknown fungi and exemplar taxa from GenBank, examining the importance of manual vs. automated alignment methods, model selection, and inference methods. We then use our most robust methods to examine the quality of BLAST-based identification, considering the taxonomic information provided by NCBI's module for inferring fast minimum evolution trees and our own Bayesian analyses. Finally, we use ancestral state reconstructions to generate preliminary hypotheses regarding the evolutionary origins of seed-associated fungi in three focal genera (*Botryosphaeria, Mycosphaerella,* and *Xylaria*), considering relationships of seed-associated, endophytic, pathogenic, and saprotrophic fungi. Together, our analyses not only shed light on a distinctive, understudied, and ecologically important group of tropical plant-associated fungi, but also demonstrate several advantages and potential pitfalls inherent in the use of ITS data in the study of unknown fungi.

Methods

Mature infructescences of five *Cecropia insignis* individuals/site were collected from an array of 2 m^2 mesh seed traps 1 m above the forest floor (mesh size = 1 mm) and directly from tree canopies in lowland tropical forest at La Selva, Costa Rica (10°26′ N, 84°00′ W; 37–150 m above sea level) and Barro Colorado Island, Panama (BCI; 9°9′ N, 79°51′ W; 120–160 m above sea level) during the fruiting season in 2005. Immediately after collection, seeds were removed from infructescences, rinsed in 0.5 % sodium hypochlorite for 2 min, and surface-dried under sterile conditions in a darkroom. Seeds were sorted by maternal source into lots of 30, mixed with 10 g of autoclave-sterilized forest soil (115 °C for 2 h), and enclosed in nylon mesh bags (mesh size = 0.5 mm) that effectively exclude most arthropods while allowing seeds to be exposed to viruses, bacteria, fungi, and nematodes (Gallery *et al*. 2007a). Bags were buried 3 cm beneath the soil surface in 5 × 5 m plots below the crowns of four mature female *C. insignis* individuals at BCI (>50 m apart) and recovered after five months (December 2005).

At harvest, bags recovered beneath three crowns were emptied onto sterile filter paper in Petri dishes, moistened as needed, and allowed to germinate for two months following the methods of Gallery *et al*. (2007a). From each bag, four seeds that did not germinate were selected haphazardly, surface-sterilized by sequential immersion in 95 % ethanol (10 s), 0.5 % sodium hypochlorite (2 min), and 70 % ethanol (2 min) (Gallery *et al*. 2007a, 2007b), allowed to surface-dry under sterile conditions, and plated on 2 % malt extract agar (MEA) slants. Slants were incubated for six months. Cultivable filamentous fungi were observed in 81 % of seeds. Living vouchers were deposited in the Robert L. Gilbertson Mycological Herbarium at the University of Arizona (ARIZ).

Total genomic DNA was extracted from all isolates following the methods of Arnold & Lutzoni (2007), and the ITS region was amplified and sequenced following the methods of Gallery *et al*. (2007a). Bidirectional sequences were assembled automatically and quality scores assigned using *phred* and *phrap* (Ewing & Green 1998; Ewing *et al*. 1998; Gordon *et al*. 1998) with automation provided by the ChromaSeq package implemented in Mesquite version 1.91 (Maddison & Maddison: http://mesquiteproject.org). All assembled sequences were submitted to GenBank under accession numbers *FJ612603–FJ613109*.

Taxonomic placement for each fungus was estimated at the subphylum and class levels by BLAST searches in GenBank and by comparison against a phylogenetically referenced database of over 6 K ITS sequences for plant-symbiotic fungi (Arnold unpubl. data). The majority of seed-associated fungi were identified as *Pezizomycotina*, with a particular concentration in the *Sordariomycetes* (438 isolates), and *Dothideomycetes* (89 isolates), both of which are rich in saprotrophs, pathogens, and endophytes (e.g. James *et al*. 2006; Arnold *et al*. 2007; Arnold & Lutzoni 2007). Our sample included 507 isolates recovered in the current study, and 20 additional seed-associated fungi obtained in similar studies at BCI (Gallery *et al*. 2007a; Kluger *et al*. 2008).

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

Estimating intra- and interspecific divergence of ITS sequence data

Using previously published phylogenies, we examined the percent similarity of ITS sequences within and among currently recognized species in four genera that are common among plant-associated fungi in the neotropics: *Botryosphaeria* and *Mycosphaerella* (*Dothideomycetes*) and *Colletotrichum* and *Xylaria* (*Sordariomycetes*) (Lee *et al.* 2000; Denman *et al.* 2003; Du *et al.* 2005; Feau *et al.* 2006). We downloaded all available, named sequences from GenBank for each genus, and randomly chose a maximum of five sequences per species for subsequent analyses (Supplementary Material Table S1). Sequences were aligned in MacClade (Maddison & Maddison 2003) and genetic distances (percent sequence divergence over the entire ITS region, including the 5.8S gene) were calculated in PAUP* 4b10 (Swofford 2002) at three levels: intraspecific, sister taxa, and non-sister taxa (Table 1). Sister taxa included *Botryosphaeria ribis* and *B. protearum*, *Colletotrichum gloeosporioides* and *C. fragariae*, *C. graminicola* and *C. sublineolum*, *Mycosphaerella rubi* and *M. ribis*, *M. brassicae* and *M. ulmi*, *M. populorum* and *M. populicola*, *Xylaria arbuscula* and *X. mali*, *X. longipes* and *X. acuta*, and *X. polymorpha* and *X. hypoxylon* (following Lee *et al.* 2000; Denman *et al.* 2003; Du *et al.* 2005; Feau *et al.* 2006). All other interspecific comparisons were between non-sister taxa. Mean percent divergence and standard deviation, standard error, and 95 % confidence intervals were determined using JMP IN version 4.0.4 (SAS Institute, Cary, NC).

Richness and diversity of seed-associated fungi

Based on our empirical estimate of percent divergence between sister species, ITS genotype groups based on 95 % sequence similarity were used as operational taxonomic units (OTU; see below). Genotype groups were assembled using Sequencher v. 4.2.2 (Gene Codes Corporation, Ann Arbor, MI) following the methods of Arnold *et al.* (2007), and DOTUR (Schloss & Handelsman 2005). For Sequencher, consensus sequences were assembled into groups under the expectation of 40 % overlap following the methods of Arnold *et al.* (2007). For DOTUR, sequences were first aligned in MUSCLE (Edgar 2004) using default parameters, and the alignment was used to generate an uncorrected distance matrix in PAUP* 4b10 (Swofford 2002). The furthest neighbour algorithm was used in DOTUR (a conservative approach which assigns a sequence to a group only if it is X % similar to all sequences in that group), with 10 K iterations and jumbled sequence input. Although groups based on 95 % ITS similarity were used in subsequent analyses, sequences were grouped at several different levels of similarity (95–99 % sequence similarity; Table 2) to explore the congruence of each program's output and the effects on richness given different stringencies for defining OTU.

Table 1 – Divergence of fungal ITS sequences based on intraspecific and interspecific (sister, non-sister) comparisons using published phylogenies

Genus	Total no. of sequences	Total no. of species	No. of sister-species pairs[a]	Sister	Non-sister	Intra-specific
Xylaria[b]	15	6	3			
No. of pairwise sequence comparisons				11	48	7
Mean				4.18	7.29	1.43
S.D.				2.18	2.17	2.94
S.E.M.				0.66	0.31	1.11
Colletotrichum/Glomerella[c]	21	5	2			
No. of pairwise sequence comparisons				23	118	31
Mean				4.61	8.10	3.00
S.D.				3.55	1.99	3.21
S.E.M.				0.74	0.18	0.58
Botryosphaeria[d]	30	6	1			
No. of pairwise sequence comparisons				25	350	60
Mean				3.92	13.11	0.97
S.D.				0.49	4.31	2.19
S.E.M.				0.10	0.23	0.28
Mycosphaerella[e]	21	6	3			
No. of pairwise sequence comparisons				36	146	28
Mean				4.64	6.90	1.68
S.D.				1.50	1.76	2.13
S.E.M.				0.25	0.15	0.40
Mean divergence for all genera				4.39	10.43	1.65
S.E.M. divergence				0.22	0.17	0.23
Total	87	23	9			

a This is the number of pairs of species that are sister taxa. Each species is represented by multiple individuals.
b Lee *et al.* 2000.
c Du *et al.* 2005.
d Denman *et al.* 2003.
e Feau *et al.* 2006.

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

Table 2 – Comparison of the number of operational taxonomic units generated for 527 seed-associated fungi

Sequence divergence (% similarity)	DOTUR[a] (all sequences)	DOTUR[a] (S + D)[b]	Sequencher (all sequences)	Sequencher (S + D)[b]
1 (99 %)	272	278	202	158
2 (98 %)	250	253	177	138
3 (97 %)	235	235	170	126
4 (96 %)	221	226	162	117
5 (95 %)	207	212	155	108

The same datasets yielded markedly different estimates of genotypic richness at all levels of comparison, and differed in estimates of richness depending on the input of a single master file of all sequences, or independent analysis of sequences in each of two classes.
a DOTUR's furthest-neighbour algorithm was used.
b The result of 438 sordariomycete (S) and 89 dothideomycete (D) sequences analysed separately.

Species-accumulation curves and bootstrap (BS) estimates of total richness were inferred using 50 randomizations of sample order in EstimateS version 7.5 using both DOTUR and Sequencher OTU groups (Colwell 2005: http://viceroy.eeb.uconn.edu/EstimateS) (Fig 1). Diversity was calculated as Fisher's alpha, Shannon index, and Simpson's index to maximize comparability with other studies. To examine the similarity of fungi associated with seeds from different origins (La Selva and BCI), similarity indices based on presence/absence data (Jaccard's index) and isolation frequency (Morisita–Horn index) for each non-singleton genotype were calculated in EstimateS (Colwell 2005).

To examine congruence of seed-associated fungi with known foliar endophytes at the levels of genotype (99 % ITS similarity) and putative species (95 % ITS similarity; see below), seed-associated fungi were compared against a database of 6 K ITS sequences of foliar endophytes (Arnold unpubl. data) using Sequencher 4.2.2. That database includes representative endophytes from asymptomatic foliage of 32 host species in arctic, boreal, temperate, and tropical plant communities (Arnold unpubl. data; Arnold *et al.* 2007; Arnold & Lutzoni 2007), including 1150 sequences of endophytic fungi from BCI.

Examining BLAST-based identification methods

We assessed the quality of the BLAST-based taxonomic assignment of 41 isolates representing two focal genera (*Xylaria* and *Mycosphaerella*) using two phylogenetic approaches. First, GenBank was queried using the BLASTn algorithm for each sequence, and the top hits recorded. From the same search output, a fast minimum evolution tree then was generated automatically using NCBI's automated pairwise alignment, the Jukes–Cantor distance measure, and NCBI's online 'view tree' option (see http://www.ncbi.nlm.nih.gov/CBBResearch/Desper/FastME.html for details). For each resulting tree, we examined the placement of our query sequence relative to its top BLAST match.

Second, an alignment was generated for each genus using our query sequences; selected top BLASTn hits; the sequence of the first taxonomically named BLASTn hit, if the top hits were unidentified; and a selection of named species (see below). These alignments were assembled automatically in MUSCLE with manual editing (corresponding to category 2; see below). Model selection and Bayesian phylogenetic analyses were conducted as described below (i.e. GTR + I + G for 5 M generations). The data matrix for *Xylaria* contained 25 query sequences and 19 of their BLASTn hits, as well as 12 additional sequences representing the diversity of *Xylaria* in GenBank (Supplementary Material Table S1), for a total of 56 sequences. The *Mycosphaerella* matrix contained 51 sequences, 16 of which were unknown and 35 of which were NCBI sequences (14 top BLASTn hits and 21 additional sequences representing the diversity of *Mycosphaerella* in GenBank) (Supplementary Material Table S1). Redundant BLASTn hits from NCBI were removed prior to each analysis.

For each minimum evolution and Bayesian analysis, we asked three questions: (1) how frequently was a query sequence reconstructed as sister to its top BLASTn hit; (2) of the query sequences that were sister to their top BLASTn hit, how frequently could they be identified to the genus or species level (i.e. how frequently were they sister to a named taxon); (3) what proportion of sequences failed to be placed with confidence (i.e. were placed in polytomies, were basal to clades containing multiple named taxa, or were sister to unidentified fungi, either from this study or from GenBank; Table 3).

Alignment, model selection, and phylogenetic analyses

Seed-associated and foliar endophytic *Botryosphaeria, Colletotrichum, Mycosphaerella,* and *Xylaria* were selected for phylogenetic analyses from the isolates recovered in this study and from a living library of 12 145 vouchers maintained at ARIZ (Arnold & Lutzoni 2007; Gallery *et al.* 2007a; Kluger *et al.* 2008). For seed-associated fungi in the same genotype group (based on 95 % sequence similarity), a single representative sequence was chosen for phylogenetic analyses. In sum, sequences from 38 seed-associated fungi (representing 116 isolates) and 80 foliar endophytes were included in our alignments: 15 seed-associated isolates and 21 endophytic isolates of *Botryosphaeria;* two and 39, respectively, of *Colletotrichum*; four and 12 *Mycosphaerella*; 17 and eight *Xylaria*. These seed-associated and endophyte sequences were aligned with named sequences of each genus (*Botryosphaeria*, 30 named sequences; *Colletotrichum*, 17; *Mycosphaerella*, 21; *Xylaria*, 12) obtained from GenBank (Supplementary Material Table S1).

We first assessed the effects of three different alignment methods on resulting phylogenetic trees. For each genus, we generated (1) an automatic, unedited alignment using MUSCLE (Edgar 2004) with default parameters (category 1); (2)

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

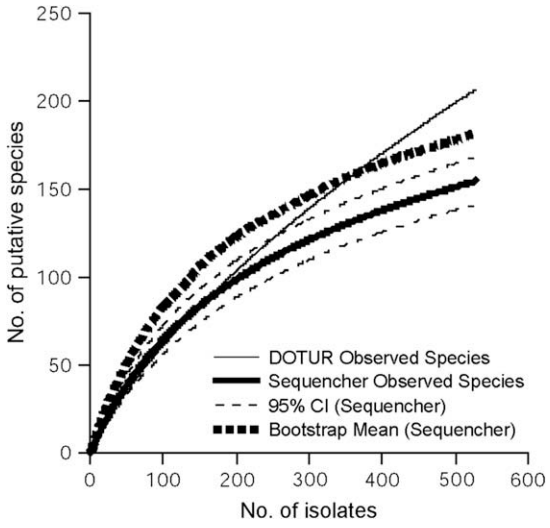

Fig 1 – Species accumulation curve of 155 ITS genotype groups for 527 seed-associated fungal isolates, 95 % confidence interval, and bootstrap estimate of total species richness. ITS genotype groups were delineated by sequence comparisons in Sequencher and DOTUR and reflect 95 % sequence similarity based on empirical estimates of sequence divergence in focal genera.

a default MUSCLE alignment edited manually in MacClade (Maddison & Maddison 2003) (category 2); and (3) a fully manual alignment generated in MacClade with ambiguous regions excluded (category 3). For category 3 alignments, 27 of 548 characters were excluded from the *Botryosphaeria* alignment (101–111, 401–406, 425, and 514–525), 89 of 514 from the *Colletotrichum* alignment (38–113 and 394–398), 75 of 518 from the *Mycosphaerella* alignment (16–40, 125–128, 142–145, 374–385 442–445, 456–459, 463–479, and 504–516), and 210 of 694 from the *Xylaria* alignment (20–29, 76–214, 246–265, 497–519, 565–572, 588–589, and 631–644). All alignments have been submitted to TreeBASE and can be downloaded from the website of the corresponding author (www.endophytes.org/alignments). Topologies and support values resulting from Bayesian analyses of each alignment under the same model of evolution (see below) were then compared within each genus.

The appropriate model of evolution was determined for each alignment using the Akaike Information Criterion (AIC) in ModelTest (Posada & Crandall 1998) (Supplementary Material Table S2). All alignments for *Botryosphaeria*, *Colletotrichum*, *Mycosphaerella*, and *Xylaria* were fitted to the GTR + I+ G model. However, *Mycosphaerella* and *Xylaria* category 1 (MUSCLE, default) and category 2 (MUSCLE, manually adjusted) alignments also were fitted to SYM + I+G. The *Xylaria* category 3 (fully manual) alignment was fitted to GTR + I+ G, SYM + I+ G and HKY + I+ G.

Phylogenetic estimations were conducted for each alignment and each selected model of evolution using MrBayes 3.1.1 (Huelsenbeck & Ronquist 2001) for 5 M generations each, using two independent runs (each with four chains) and sampling every 100th generation. To determine the burn-in for each analysis, the average s.d. of the split frequencies was evaluated, as well as plots of –lnL values. In general, the average s.d. of the split frequencies was <0.01 after a maximum of 3.69 M generations for all analyses. The convergence statistic for *Colletotrichum* category 1 and 2 alignments (0.0145 and 0.0264), the *Mycosphaerella* category 1 alignment using both GTR + I+ G and SYM + I+ G (0.0153 and 0.0102), and the *Xylaria* category 2 alignment using SYM + I+ G (0.0158) never dropped below 0.01, but plots of –lnL and other parameters demonstrated a stable plateau after 4 M generations. Therefore, a burn-in of 4 M generations was discarded from each run and the majority rule consensus tree and posterior probabilities (PP) were computed using the remaining 20,002 trees in PAUP* version 4b10 (Swofford 2002).

Because topologies inferred using Bayesian methods were largely consistent both among alignments and with regard to different models of evolution, a representative alignment (category 2: MUSCLE, with manual adjustment) and a single model (GTR + I+ G) was selected for subsequent phylogenetic estimations (Figs 2–5).

Neighbor-joining (NJ) analyses also were conducted for each alignment (categories 1, 2, 3) and each focal genus (*Botryosphaeria*, *Colletotrichum*, *Mycosphaerella*, and *Xylaria*) to assess the sensitivity of NJ to different alignments, and to compare topologies with Bayesian results (category 2 only) (Supplementary Figs S1–S4). NJ analyses were conducted using the default settings in PAUP* version 4b10 [uncorrected distance measure (p), equal rates, no invariant sites] (Swofford

Table 3 – Phylogenetic analysis of BLASTn identification of unknown fungi affiliated with *Mycosphaerella* and *Xylaria* using Bayesian analyses and minimum evolution analyses (implemented in GenBank)

No. of sequences	*Mycosphaerella*		*Xylaria*	
	Bayesian	Min. evolution	Bayesian	Min. evolution
Sister to top BLASTn hit	3	5	10	19
Identified to species	1	1	1	1
Identified to genus	2	4	5	14
Unidentified	0	0	4	4
Sister to other sequence	4	2	7	3
In polytomy	9	9	8	3
Total	16	16	25	25

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

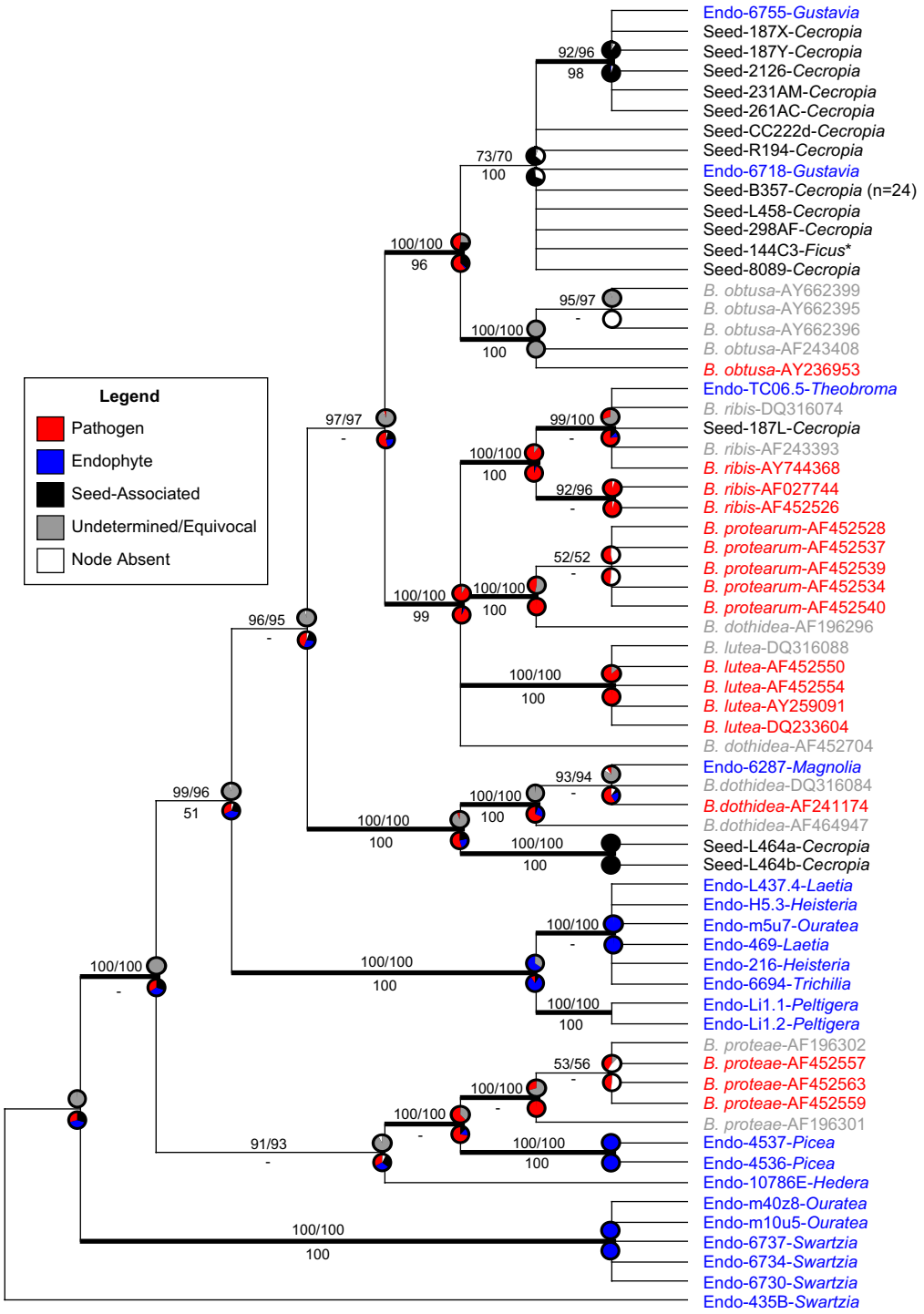

Fig 2 – Majority rule consensus tree of 20,002 trees generated from alignment method 2 (MUSCLE, with manual adjustment) of *Botryosphaeria* spp. using Bayesian analysis of ITS rDNA data, with reconstruction of ancestral states estimating the origins of fungal guilds. Taxon names of endophytes and seed-associated fungi indicate ecological role-numerical identifier-host genus. The number of seed-associated isolates with identical genotypes at 95% similarity is indicated in parentheses. Support values are PPs generated in MrBayes using model GTR + I + G for three different alignment methods: (1) MUSCLE unedited (before slash; category 1 alignment) (2) MUSCLE with manual editing (after slash; category 2 alignment) and (3) manual alignment in MacClade (below branch; category 3 alignment). A dash (-) indicates the node was not present in the majority rule consensus for that alignment. Thickened lines represent neighbor-joining bootstrap support ≥70% generated from the category 2 alignment. Pie charts indicate two methods for ancestral state reconstruction: (1) the proportion of trees in which a given state was significantly more likely than alternative states, or in which a node was equivocal or absent (top); and (2) among the trees in which a node was present, the average probability of any given state (bottom). The * indicates this sequence was obtained from *Ficus insipida* instead of *C. insignis*.

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

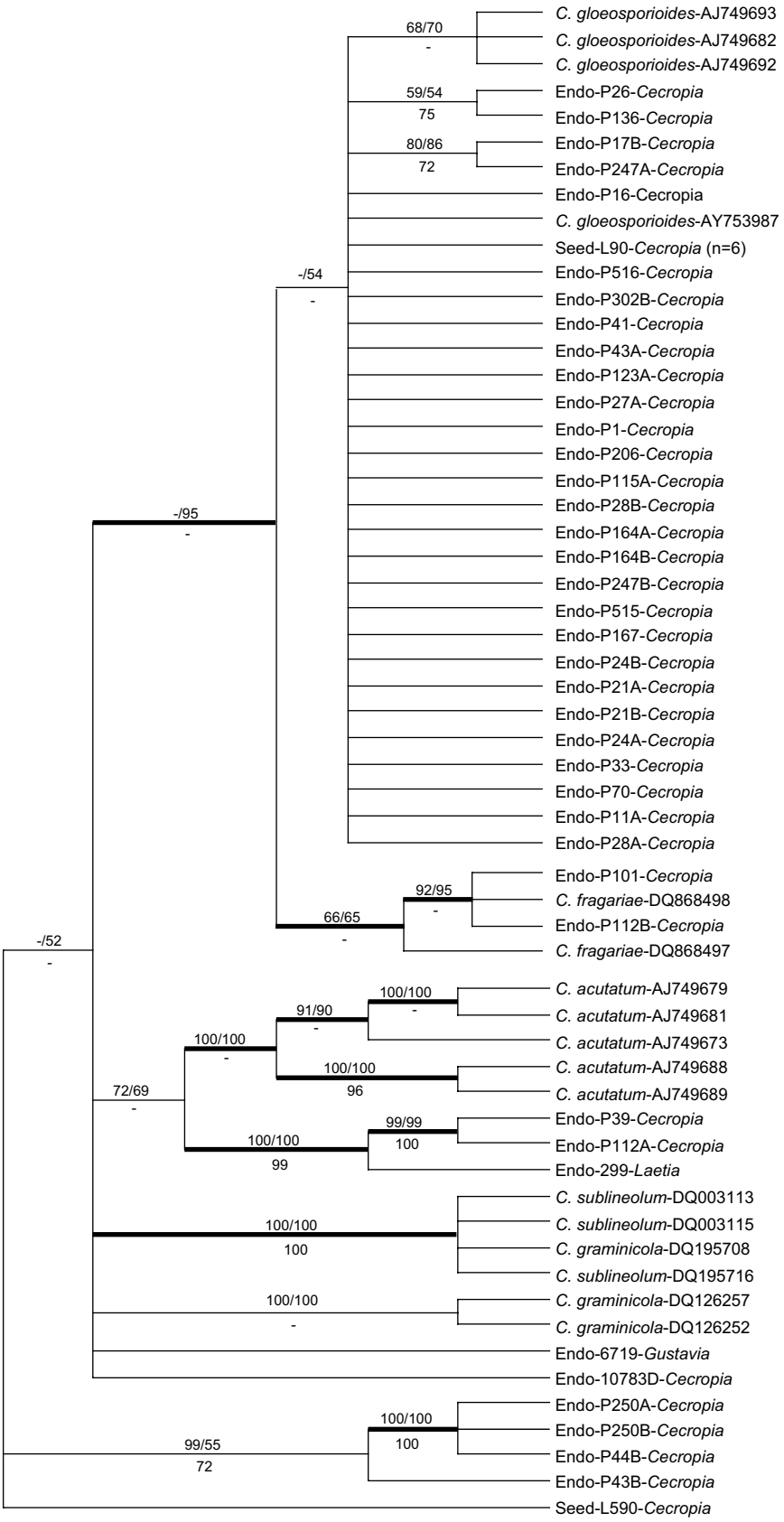

Fig 3 – Majority rule consensus tree of 20,002 trees generated from alignment method 2 (MUSCLE, with manual adjustment) of *Colletotrichum* spp. using Bayesian analysis of ITS rDNA data. Taxon names and support values are as described in Fig 2.

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

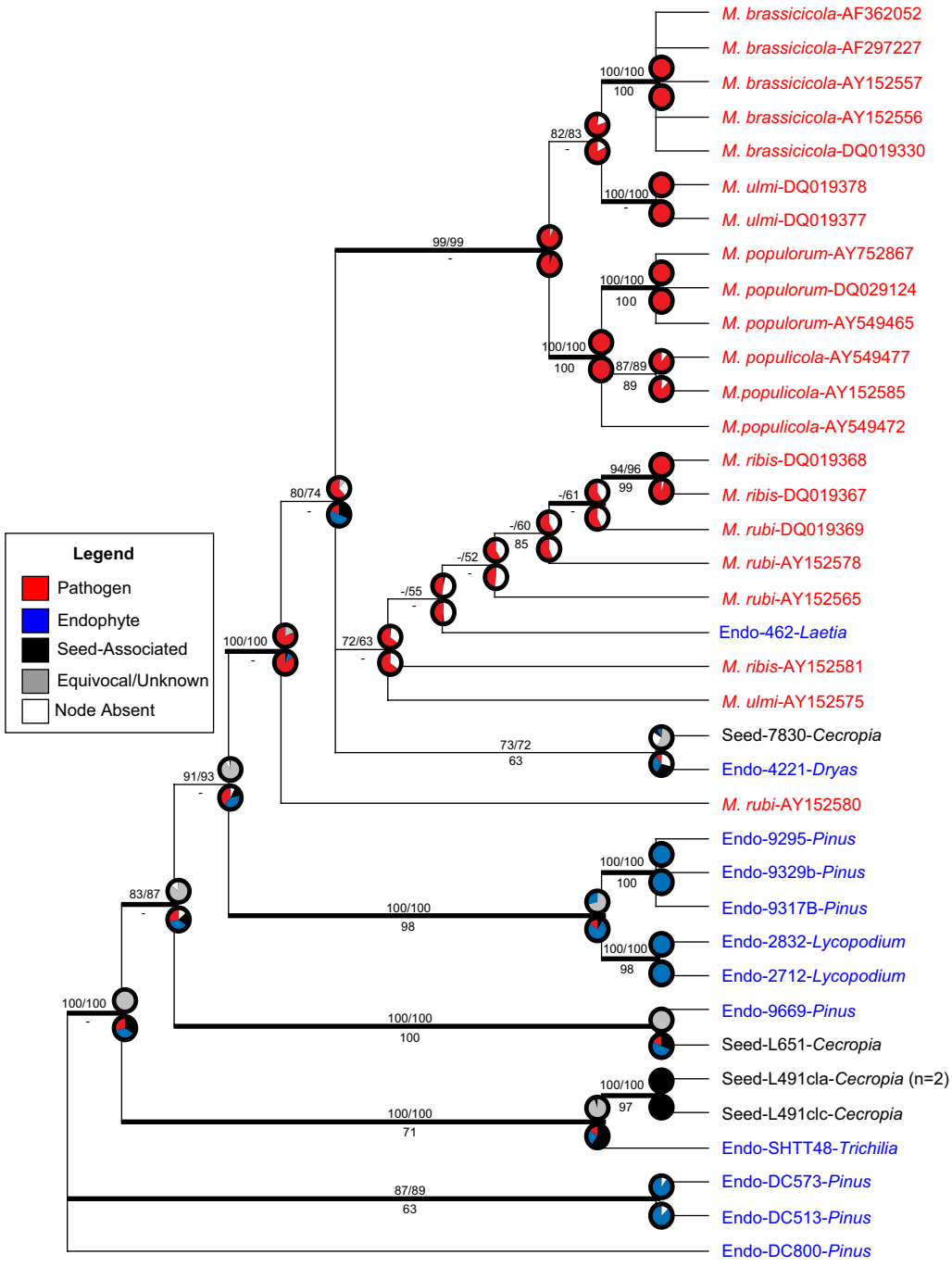

Fig 4 – Majority rule consensus tree of 20,002 trees generated from alignment method 2 (MUSCLE, with manual adjustment) of *Mycosphaerella* spp. using Bayesian analysis of ITS rDNA data, with reconstruction of ancestral states estimating the origins of fungal guilds. Taxon names, support values and pie charts are as described in Fig 2.

2002) with branch support determined through 1 K NJ BS replicates. Although models of evolution can be implemented using NJ, we chose default settings to represent the method most commonly found in the literature.

Differences between topologies, as a result of different alignments or inference methods, were defined as the sum of well-supported nodes present in the reference tree, but missing in the comparison tree, divided by the total number of nodes in the reference tree (Table 4), with well-supported nodes defined as those with ≥70 % NJ BS or ≥95 % Bayesian PP values. Resulting proportions were averaged by genus, inference method, or alignment category and analysed using ANOVA (if normally distributed) or Wilcoxon rank-sum tests (if distributions differed significantly from normal, inferred using the Shapiro–Wilk W statistic implemented in JMP IN). For convenience, we refer to these proportions as 'incongruency scores'. Higher incongruency scores indicated that a large proportion of well-supported

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

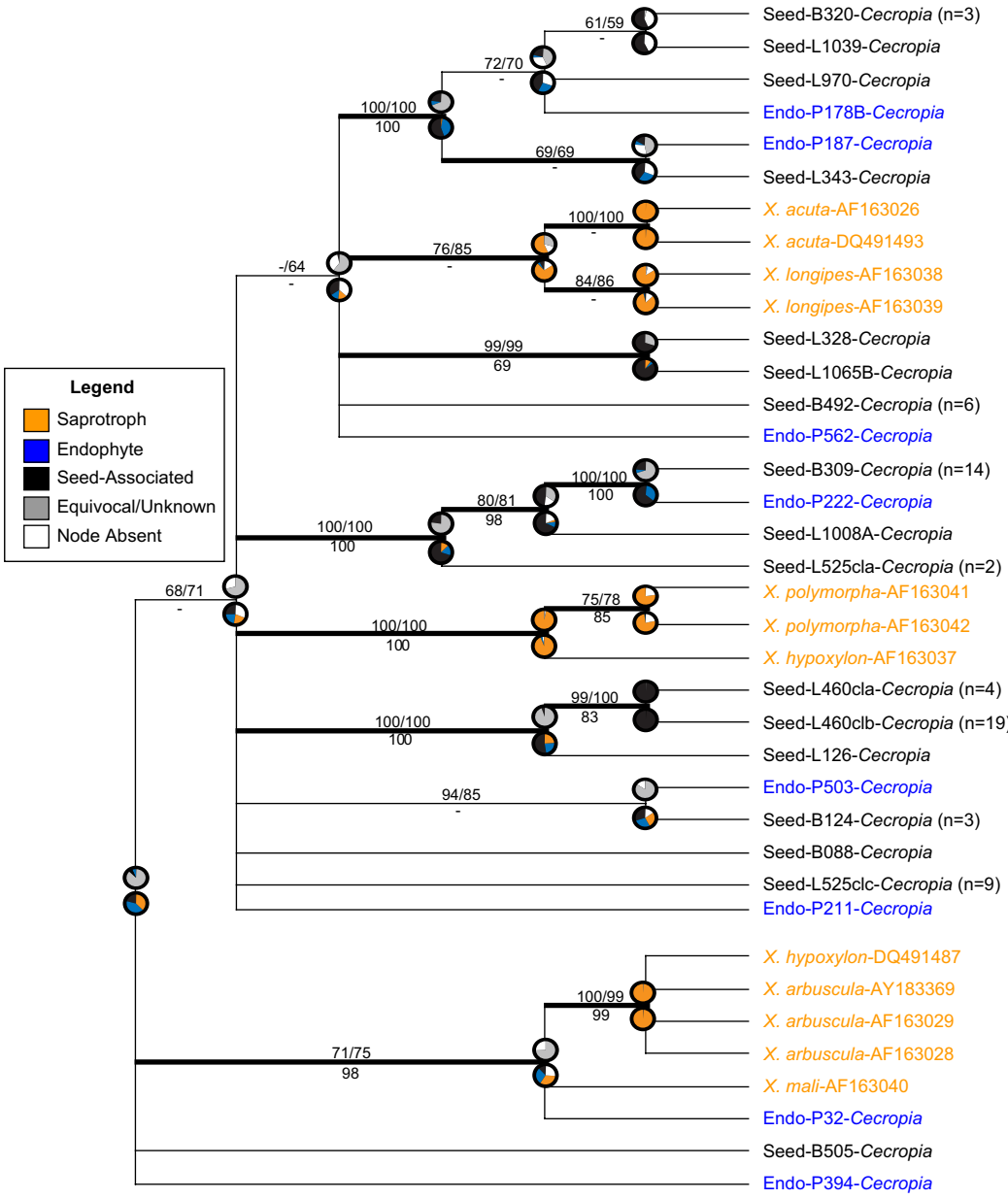

Fig 5 – Majority DNA rule consensus tree of 20,002 trees generated from alignment method 2 (MUSCLE, with manual adjustment) of *Xylaria* spp. using Bayesian analysis of ITS rDNA data, with reconstruction of ancestral states estimating the origins of fungal guilds. Taxon names, support values and pie charts are as described in Fig 2.

nodes present in a reference tree were absent from the comparison tree (i.e. trees were highly incongruent with one another). Lower scores indicated that a small proportion of well-supported nodes in a reference tree were absent from the comparison tree (i.e. trees were relatively congruent with one another; Table 3).

Evolutionary origins of seed-associated fungi

Using our most robust phylogenetic analyses, we reconstructed ancestral states to formulate preliminary hypotheses regarding the evolutionary origins of seed-associated fungi in *Botryosphaeria*, *Mycosphaerella*, and *Xylaria*, which contained multiple representatives of each ecological guild in well-supported clades. Terminal taxa from each genus were assigned a code for ecological mode (pathogens, saprotrophs, endophytes, seed-associated fungi, or missing data). Coding for *Mycosphaerella* and *Xylaria* taxa obtained from GenBank (Supplementary Material Table S1) was inferred from the known ecological mode of each species (e.g. *Xylaria arbuscula* and *X. polymorpha* as saprotrophs). In the case of *Botryosphaeria*, in which some pathogenic species have life cycles with an endophytic or latent phase (Smith *et al.* 1996; Swart *et al.* 2000; Denman *et al.* 2003), taxon coding was based on the original publication in which each

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

Table 4 – Comparison of Bayesian (MB) and NJ tree topologies

Type of alignment	Bayesian source tree						NJ source tree					
	Proportion of clades missing (missing/total)	*Botryosphaeria*	*Mycosphaerella*	*Colletotrichum*	*Xylaria*	Average for all genera	Proportion of clades missing (missing/total):	*Botryosphaeria*	*Mycosphaerella*	*Colletotrichum*	*Xylaria*	Average for all genera
Category 1	1A. MB tree compared with MB trees with different alignments	0.39	0.31	0.44	0.11	0.31 ± 0.15	1A. NJ tree compared with NJ trees with different alignments	0.40	0.24	0.38	0.35	0.34 ± 0.07
	1B. MB tree compared with category 1 NJ tree	0.09	0	0.11	0		1B. NJ tree compared with category 1 MB tree	0.16	0.24	0.23	0.25	
Category 2	2A. MB tree compared with MB trees with different alignments	0.38	0.29	0.50	0.11	0.32 ± 0.16	2A. NJ tree compared with NJ trees with different alignments	0.42	0.29	0.38	0.35	0.36 ± 0.05
	2B. MB tree compared with category 2 NJ tree	0.08	0	0.10	0		2B. NJ tree compared with category 2 MB tree	0.19	0.14	0.15	0.25	
Category 3	3A. MB tree compared with MB trees with different alignments	0.50	0.62	0.57	0.30	0.5 ± 0.14	3A. NJ tree compared with NJ trees with different alignments	0.90	0.57	0.75	0.69	0.73 ± 0.14
	3B. MB tree compared with category 3 NJ tree	0.05	0	0	0.10		3B. NJ tree compared with category 3 MB tree	0.27	0.24	0.58	0.31	
Average	A. All MB trees with different alignments	0.42 ± 0.07	0.41 ± 0.19	0.5 ± 0.07	0.17 ± 0.11		A. All NJ trees with different alignments	0.57 ± 0.28	0.37 ± 0.18	0.5 ± 0.21	0.46 ± 0.2	
	B. All MB trees and their corresponding NJ tree	0.07 ± 0.02	0	0.07 ± 0.06	0.03 ± 0.06		B. All NJ trees and their corresponding MB tree	0.21 ± 0.06	0.21 ± 0.06	0.19 ± 0.06	0.27 ± 0.03	

Values indicate incongruency scores as detailed in the text.

isolate and its ecological mode were described. Isolates from unknown substrates or those with unknown ecological modes were conservatively coded as missing data.

For each genus, ancestral states were reconstructed on each of 1 K randomly sampled trees from the posterior of 20,002 trees using the Markov k-state 1 parameter model in Mesquite version 2.0 (Maddison & Maddison: http://mesquiteproject.org). Results are shown on the majority rule consensus tree of all 20,002 trees for each genus. For each focal node, ancestral state reconstructions were summarized using pie charts that correspond to two analyses. First, we determined the proportion of trees in which a given ecological mode was significantly more likely than alternative ecological modes, or in which a node was equivocal or absent from the sample of trees (Arnold *et al.* in press). Second, we determined for each node the average likelihood of each ecological mode among the trees in which that node was present.

Results

Estimating intra- and interspecific divergence of ITS sequence data

Empirical estimates of within *versus* between-species ITS sequence divergence in four common genera of seed-associated fungi (*Botryosphaeria*, *Colletotrichum*, *Mycosphaerella*, and *Xylaria*) revealed a mean sequence divergence of 4.39 % (95 % C.I. = 3.96–4.82 %) for sister species and 10.43 % (95 % C.I. = 10.09–10.77 %) for non-sister taxa (Table 1). Mean intraspecific divergence was significantly lower (1.65 %; 95 % C.I. = 1.19–2.11 %). *Mycosphaerella* had the highest sister-group sequence divergence at 4.64 %, whereas *Botryosphaeria* had the lowest sister-group sequence divergence at 3.92 %.

The observation of *ca* 4.39 % mean sequence divergence between closely related (sister) species for four genera suggests that 95 % sequence similarity (5 % sequence divergence) is a reasonable and conservative estimate for delimiting species boundaries in our sample. Therefore, genotype groups based on 95 % sequence similarity were used to define OTU for the remainder of our study.

Diversity of seed-associated fungi

From 527 seed-associated fungi in the *Sordariomycetes* and *Dothideomycetes*, we recovered 155 putative species based on 95 % sequence similarity using Sequencher (Fisher's alpha = 74.01; Shannon index = 4.64; and Simpson's index = 84.10). Sixty-three putative species (40.7 %) occurred only once (i.e. were singletons; Table 2). When compared using 99 % sequence similarity in Sequencher, sequences clustered into 202 distinctive genotypes, of which 105 were singletons (52.0 %).

The furthest neighbour clustering method in DOTUR yielded a higher estimate of observed species- and genotypic richness than did Sequencher. At 95 % ITS similarity, DOTUR yielded 207 putative species (Fisher's alpha = 125.63; Shannon index = 4.53; Simpson's index = 35.47), and at 99 % sequence similarity, DOTUR distinguished 272 genotypes (Table 2). The

majority of these additional genotypes were singletons; for example, 69 % of genotypes inferred using 95 % similarity occurred only once.

Over the entire dataset, species-accumulation curves based on 95 % sequence similarity remained non-asymptotic regardless of the program used to infer genotype groups, and BS estimates of total richness consistently exceeded the 95 % confidence interval around observed species richness (Fig 1). The total richness estimated by the furthest neighbour algorithm in DOTUR significantly exceeded that estimated by BS analyses of groups inferred by Sequencher (Fig 1). For subsequent analyses, we used the more inclusive ITS genotype groups based on Sequencher to conservatively estimate species boundaries, recognizing that this method likely underestimates richness.

Evidence for horizontal transmission of seed-associated fungi

Following burial in the forest at BCI, seeds originally collected from Panama (BCI) and Costa Rica (La Selva) harboured similar fungal communities in terms of diversity and taxonomic composition. Among 262 isolates cultured from seeds originally harvested at BCI, 102 putative species were recovered (Fisher's alpha = 61.5). Among 265 isolates cultured from seeds originally harvested at La Selva, 106 putative species were recovered (Fisher's alpha = 66.3). Fungal communities associated with seeds of both origins were similar based on both presence/absence data for non-singletons (Jaccard's index = 0.50) and isolation frequency (Morisita–Horn index = 0.48). Among 92 putative species that were recovered more than once (non-singletons), 56 (60.9 %) were recovered from seeds of both geographic origins. The remaining 36 non-singleton species were found only in seeds originally harvested at BCI (16 putative species), or only in seeds originally harvested from La Selva (20 putative species). The observation that the majority of seed-associated fungi were shared between co-incubated seeds from different geographic origins, coupled with evidence from previous studies showing that *Cecropia* seeds rarely harbour cultivable fungi before soil exposure (Gallery *et al.* 2007a), is consistent with horizontal transmission.

Genotypic congruence of seed-associated and endophytic fungi

Analysis using the empirical estimate of 95 % ITS sequence similarity revealed that 40 of 155 putative species recovered from seeds in this study (25.8 %) were recovered previously as endophytic fungi from foliage at BCI. When more stringent analyses were conducted using 99 % sequence similarity, 32 of 202 genotypes (15.8 %) matched previously encountered foliar endophytic fungi, including representatives of the cosmopolitan genera *Alternaria, Cladosporium, Fusarium, Hypoxylon, Kabatiella, Nodulisporium, Pestalotiopsis, Trichoderma,* and *Xylaria* (genus-level taxonomy inferred using phylogenetic analyses published elsewhere; e.g. Arnold *et al.* in press). The remainder of our isolates did not match any fungi in our endophyte database at 95 or 99 % similarity, and were not perfect matches of any sequence data available through GenBank.

Assessment of BLASTn-based identification methods

Placement of our query sequences relative to their top BLASTn-hits was assessed using (1) analyses of our query sequences and their top hits under NCBI's automated pairwise alignment and minimum-evolution tree-generating tool, and (2) Bayesian analyses based on category 2 alignments of our query sequences, their top hits, and representative sequences from NCBI (Supplementary Material Table S1). Our goal was not to compare the outputs of these different inference methods per se, but instead to address the potential for the limited taxonomic sampling present in GenBank to overestimate certainty with regard to placement of unknown sequences.

Although BLASTn and the minimum-evolution module drew from exactly the same sequences in GenBank, incongruencies between BLASTn-based identification and identification based on the minimum-evolution module were common (Table 3). Only five of 16 *Mycosphaerella* sequences (31.3 %), and 19 of 25 *Xylaria* sequences (76.0 %), were reconstructed as sister to their top BLASTn hit using the automated minimum evolution analysis provided by NCBI (Table 3).

Overall, minimum evolution analyses were 1.7–3.2 times more likely than our Bayesian analyses to place query sequences as sister to their top BLASTn hit: only 18.8 % (*Mycosphaerella*) and 40.0 % (*Xylaria*) of our query sequences were reconstructed as sister to their top BLASTn hit in Bayesian analyses (Table 3). Instead, unknown endophytes and seed-associated fungi were often the closest relatives of our query sequences, disrupting the relationships between queries and GenBank sequences seen in minimum evolution trees. Accordingly, misidentification by BLASTn, indicated by sister relationships between query sequences and sequences other than their top BLASTn match, was highlighted roughly twice as frequently by our Bayesian inferences (Table 3). In general, we observed that the minimum evolution inference module in NCBI can over-inflate certainty regarding BLASTn-based identification because only sequences present in GenBank are considered.

Overall, only 25 % (*Mycosphaerella*) and 32 % (*Xylaria*) of query sequences that were sister to their top BLASTn hit in minimum evolution analyses were also sister to their top hit in our Bayesian analyses. Polytomies, which may accurately reflect unresolved placement given the taxon sampling present in the analysis or the limitations of the phylogenetic signal in the data, occurred with similar frequency in Bayesian and minimum evolution trees for *Mycosphaerella*, but were twice as common in Bayesian trees than in minimum evolution analyses for *Xylaria*.

Bayesian analyses indicated that 81.3 % (*Mycosphaerella*) and 76 % (*Xylaria*) of our query sequences could not be assigned to a named taxon, as demonstrated by (1) sister relationships to unclassified NCBI sequences or, more frequently, unidentified sequences from our seed-associated or endophytic fungi, or (2) uncertain placement, either in polytomies with terminal taxa that might represent multiple species or placement basal to a clade containing multiple taxa (Supplementary Material Figs S5–S6; Table 3). Overall, Bayesian analyses only allowed one query sequence to be placed to the species level in *Mycosphaerella* (i.e. sister to an identified

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

species of *Mycosphaerella*), and two to genus (i.e. sister to sequences identified as *Mycosphaerella* sp.; Supplementary Material Fig S5). Bayesian analyses for *Xylaria* identified one sequence to the species level (i.e. sister to an identified species of *Xylaria*), and five to a higher taxonomic level (e.g. sister to sequences identified only as *Xylariaceae* sp. or Xylarialean fungi; Supplementary Material Fig S6).

Both minimum evolution and Bayesian analyses revealed that query sequences whose top BLASTn matches are to unidentified fungi should not be 'identified' on the basis of the first named species in the list of BLASTn results (Supplementary Material Figs S5–S6). In only one case was a query sequence reconstructed as sister to the first named hit when the top BLASTn hit was to an unidentified fungus: Seed-L126-*Cecropia*, for which the top BLASTn hit was an unnamed sequence, was reconstructed as sister to the first identified isolate among its BLASTn hits (DQ485958, *Botryosphaeria rhodina*; Supplementary Material Fig S6). One query sequence (Endo-P562-*Cecropia*) was placed in a polytomy with both its top hit (fungal endophyte DQ485962) and its first named hit (*Xylaria* sp. EF423534).

Importance of alignment, model selection, and inference method

We generated three alignments (categories 1, 2, and 3) for one dataset per genus, and examined resulting topologies following Bayesian and NJ analyses. Our goal was to evaluate the sensitivity of each inference method and fungal genus to different alignment methods.

We found that within each genus, Bayesian analyses were relatively robust to the different alignments and models of evolution used here: different alignments and evolutionary models yielded different topologies for each genus, but differences among those topologies were not supported by significant PPs (Figs 2–5). However, Bayesian analyses revealed that genera differed from one another in their sensitivity to alignment methods (one-way ANOVA based on incongruency scores, $F_{3,8} = 4.403$; $P = 0.042$). The highest incongruency score was observed in *Colletotrichum* (0.50 ± 0.07), which indicated that *ca* 50 % of nodes were lost when trees from the different alignments of the same *Colletotrichum* matrix were compared (Table 4). In contrast, *Xylaria* had a significantly lower incongruency score (i.e. more similar topologies) when Bayesian trees based on category 1, 2, and 3 alignments were compared with each other (0.17 ± 0.11) relative to the other three genera (post-hoc Student's *t*-test, $t_3 = 2.306$, $P < 0.05$; Table 4). In contrast, genera did not differ significantly from one another in their responses to different alignment methods under NJ (Table 4; one-way ANOVA, $F_{3,8} = 0.456$, $P = 0.720$).

Within each genus, NJ analyses were more sensitive to different alignment methods than were Bayesian analyses: average incongruency scores were higher for comparisons among alignments under NJ than under Bayesian methods (Table 4). NJ analyses were particularly sensitive to differences between category 3 alignments and the automated and semi-automated alignment methods used here (incongruency score for category 3 *versus* other alignments under NJ $= 0.73 \pm 0.14$). Topologies of NJ trees inferred from category 1 and category 2 alignments had more resolved terminal nodes than did category 3 trees, which in turn lacked resolution due to the exclusion of ambiguously aligned regions (data not shown).

Overall, NJ trees had higher average incongruency scores when compared against Bayesian trees (0.19–0.27) than did Bayesian trees when compared against NJ trees (0–0.07). These results highlight the greater resolution provided by NJ analyses, which was lost when Bayesian analyses were performed. All genera experienced a similar loss of resolution under Bayesian methods relative to NJ methods (one-way ANOVA, $F_{3,7} = 1.290$, $P = 0.35$), and a similar gain in resolution under NJ relative to Bayesian methods (Wilcoxon signed-rank test, chi-square $= 4.327$, $P = 0.228$, df $= 3$).

Phylogenetic relationships and evolutionary origins of seed-associated fungi

Our most robust phylogenetic analysis methods (Bayesian analyses based on category 2 alignments implementing GTR + I + G) revealed several well-supported clades containing both seed-associated fungi and endophytes for three of the four focal genera, including four clades in *Botryosphaeria* (Fig 2), one in *Mycosphaerella* (Fig 4), and three in *Xylaria* (Fig 5). The *Colletotrichum* tree contained one clade with pathogens, seed-associated fungi, and endophytes, but this clade was not well supported and lacked resolution (Fig 3), and thus was not used in ancestral state reconstructions.

Ancestral state reconstruction illustrated several preliminary hypotheses for the evolutionary history of seed-associated fungi, as well as their relationships to pathogens, endophytes, and saprotrophs in *Botryosphaeria*, *Mycosphaerella,* and *Xylaria*. In general, more frequent evolutionary associations were seen among endophytes, pathogens, and seed-associated fungi than between any of those guilds and saprotrophs (Figs 2, 4, 5). However, the evolutionary interplay of these ecological modes differed among genera.

Within *Botryosphaeria*, ancestral state reconstructions suggested that several endophytes arose from putatively seed-associated ancestors (e.g. Fig 2: Endo-6755-*Gustavia* and Endo-6718-*Gustavia*), and highlighted several transitions from pathogenicity to seed-association (e.g. the lineage containing *B. obtusa*). Transitions from endophytism to seed-association were not observed.

In *Mycosphaerella*, ancestral state reconstructions for three clades containing both endophytes and seed-associated fungi reveal that the greatest proportion of trees contain ancestors whose ecological mode reconstructions were equivocal (Fig 4). However, reconstructions that calculate the average probability of any given ecological mode among the trees in which a node was present indicate a seed-associated ancestor for one endophyte (Endo-SHTT48-*Trichilia*). Both methods of ancestral state reconstruction suggest that the clade containing all of the pathogens, as well as two endophytes and one seed-associated fungus, arose from a pathogenic ancestor.

In *Xylaria*, several endophytes may have arisen from seed-associated ancestors (e.g. Endo-178B-*Cecropia*, Endo-187-*Cecropia,* and Endo-P222-*Cecropia*; Fig 5). No transitions of seed-associated or endophytic fungi to or from saprotrophy were observed, suggesting that seed-associated fungi and endophytes may be evolutionarily distinct from saprotrophic lineages (Fig 5).

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

Discussion

Many seed-associated fungi remain sterile in culture, and therefore cannot be identified on the basis of reproductive morphology. In addition, phylogenetic species concepts are difficult to apply rigorously in survey data, which are geographically limited, often restricted to focal guilds, and typically are not exhaustive in terms of capturing the diversity of focal lineages of fungi. Therefore, designating operational taxonomic units on the basis of percent sequence similarity provides a useful proxy for estimating species boundaries (e.g. O'Brien *et al.* 2005; Arnold & Lutzoni 2007). However, the degree of sequence similarity used to designate fungal species boundaries often differs among studies because empirical estimates are rarely, if ever, generated. This is in contrast to bacteriology, where empirical studies have suggested that two bacteria belong to the same putative species if DNA–DNA cross-hybridization levels are >70 %, which in turn corresponds to a 16S rDNA sequence similarity of >97 % (Stackebrandt & Göbel 1994).

We empirically estimated the amount of ITS sequence divergence for four representative fungal genera. The ITS locus was chosen due to its extensive use for species-level diagnoses in fungal environmental studies and systematics (see Arnold *et al.* 2007). Although rates of inter- and intraspecific divergence differ among fungal lineages (Seifert *et al.* 2007), we observed relatively consistent patterns of sequence divergence within species, between sister species, and between non-sister species in four genera (Table 1). Focal *Dothideomycetes* had both the highest and the lowest sister-group sequence divergence estimates, whereas focal *Sordariomycetes* were intermediate.

Empirical estimation of percent divergence for estimating fungal species boundaries is important for environmental studies of fungi. However, we note the following caveats for future work. First, our estimates are based on published phylogenies, which represent only a minority of the potential geographic and taxonomic diversity in each genus. Were each genus sampled to completion, resulting estimates of intra- and interspecific sequence divergence might change markedly. We expect that tropical members of these genera are most likely to be undersampled relative to their temperate counterparts. Second, most published phylogenies do not contain explicit information regarding phylogenetic uncertainty, and in many cases support values for nodes are not given. Our estimates rely on robust phylogenies and would be changed if these topologies prove incorrect. Third, we assume a similar rate of change in the ITS region across different fungal lineages and among fungi from different ecological modes or geographic origins (e.g. pathogens *versus* endophytes; strongly seasonal temperate climates *versus* aseasonal tropical climates), but this may not hold true. Fourth, very few phylogenies based on loci other than ITS are available for these genera, limiting the robustness of our estimate. Finally, we acknowledge the limitations imposed by using only a single locus to infer species boundaries (see Taylor *et al.* 2000), which is at best a flawed, but still necessary, approach for categorizing fungal environmental samples.

Richness and diversity of seed-associated fungi

A limited number of computational methods are available for comparing sequences efficiently and accurately to designate operational taxonomic units. We found marked differences between two software packages in terms of the richness of our unknown fungi and the composition of OTU. DOTUR provided much higher richness estimates than did Sequencher: from the same datasets, DOTUR recovered as many as 120 more genotype groups (based on 99 % sequence similarity), and as many as 94 more putative species (based on 95 % similarity; Table 2). This striking disparity reveals a major challenge for biodiversity studies of unknown fungi.

We found that Sequencher can be sensitive to sample composition (see Table 2; differences between analyses with all sequences *versus Sordariomycetes* and *Dothideomycetes* only), and its sequence comparison algorithm is not publicly available. However, it is computationally straightforward and its output has been compared favourably with other tools for delimiting ITS genotype groups, including comparisons of ITS OTU with phylogenetic analyses of the nu-rLSU (Arnold *et al.* 2007).

DOTUR (Schloss & Handelsman 2005) has been used extensively for bacterial environmental studies based on 16S rDNA, and it is both fast and robust to sample composition (i.e. summing the number of OTU for the *Sordariomycetes* and *Dothideomycetes*, analysed separately, yields a value that is similar to analyses of the entire dataset; Table 2). However, when a corrected distance matrix is the input for DOTUR, the method is sensitive to sequence length. Moreover, DOTUR requires an alignment to generate a distance matrix. This may limit its utility for loci, such as ITS, for which reliable alignments can be difficult to generate in phylogenetically diverse datasets. Because DOTUR was designed for analysis of bacterial 16S data, it should be used with care when estimating fungal OTU based on ITS data.

Our DOTUR analyses used the furthest neighbour algorithm, whereby each sequence in an OTU must be X % similar to all other sequences, rather than X % similar to only the most similar sequence in that group. When analyses in DOTUR were repeated using that program's nearest neighbour algorithm (whereby each sequence within an OTU only must have X % similarity to the most similar sequence in the group), our estimate of the putative number of species was more consistent with Sequencher (data not shown). In future studies, it would be useful to validate both DOTUR and Sequencher output using data sets consisting of a known number of species, perhaps focusing on a particularly well-studied and widely sampled fungal genus.

When conservative methods are used to estimate species boundaries (95 % sequence similarity and Sequencher groupings, as used by Arnold & Lutzoni 2007), the diversity of fungi recovered among only the *Dothideomycetes* and *Sordariomycetes* in this study exceeds that found in similar seed-burial experiments at the same site: we observed a Fisher's alpha value of 74.01, whereas Gallery *et al.* (2007a) reported a Fisher's alpha value of 38.2 for a smaller data set representing four species of *Cecropia*. Despite having more than doubled the number of fungi recovered by Gallery *et al.* (2007a), our study did not yet achieve statistically complete sampling of novel species or genotypes: bootstrap estimates are 25 % higher than observed species richness (Fig 1). Further sampling is needed to capture the high diversity of seed-associated fungi at this site, underscoring the tremendous richness of fungi capable of infecting seeds of a single tree species in one tropical forest.

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

Assessing BLAST-based identification

Two phylogenetic methods, one automated, distance-based, and restricted to sequences in GenBank (minimum evolution, implemented in NCBI) and one model-based, using manually adjusted automated alignments and considering unpublished sequences (Bayesian), were compared to examine BLASTn-based identification of 41 sequences of unknown fungi. BLAST identification was considered accurate if a query sequence and its top BLAST hit were sister to each other in resulting topologies.

Although BLASTn and the minimum-evolution module relied on the same sequences in GenBank, incongruencies between these methods were observed for 17 of 41 sequences examined (41.5 %; Table 3). In turn, Bayesian inference using an alignment of unknown sequences, their top BLAST hit, and the next named BLAST hit (if the top hit was to an unidentified sample in GenBank), as well as identified sequences from NCBI, even more rarely placed query sequences with their top BLASTn matches. Although our focal genera are well-represented in GenBank, with ITS sequences of at least 92 species of *Mycosphaerella* and 13 species of *Xylaria* present in the database at the time of our study, the closest relatives of our query sequences are mostly uncharacterized and/or are not yet included in public databases. We conclude that BLASTn and the NCBI minimum-evolution module can falsely 'over-identify' unknown fungi: for both of the focal genera in this study, the majority of sequences were reconstructed in richer Bayesian analyses in polytomies with, or as sister to, other unknown fungi whose sequence data are not yet in public databases.

We realize the attractiveness of BLAST lies in its rapid and concise estimates of taxonomy, in contrast to phylogenetic methods that are more time-consuming and hard to interpret (e.g. see Little & Stevenson 2007 for discussion). However, our study and others that have highlighted the additional problems associated with misidentified sequences in GenBank (see Vilgalys 2003) suggest that BLASTn results should not be used to assign names to unknown sequences, but rather as a guide for estimating taxonomic placement. Caution is especially warranted when the top BLASTn matches are to unidentified fungi; using the first named fungus in the list of BLASTn hits almost always misidentified the query sequence in our analyses. Overall, our results support studies from other fields (e.g. genome annotation) which demonstrate that BLASTn comparisons do not always yield the closest relative (Koski & Golding 2001), highlight the importance of populating GenBank with additional sequence data to improve the quality of both BLASTn and minimum-evolution analyses, and indicate that analyses using rich datasets are likely to be most informative.

More generally, our study integrated unknown fungi into phylogenies with putatively closely related taxa and highlighted several cases in which seed-associated and endophytic fungi represent well-supported clades that are distinct from the placement of any closely related sequences in GenBank (e.g. Supplementary Material Figs S5–S6). Whether this outcome reflects a geographic artefact due to the paucity of published fungal sequences from Panama or instead embodies true phylogenetic distinctiveness of seed-associated fungi is not yet clear. Our results provide a reminder that both database-driven identification and phylogenetically assigned taxon names are sensitive to the availability of sequence data for closely related fungi.

Assessing alignment and model selection

Alignment methods strongly affect downstream phylogenetic analyses of molecular data (Ogden & Rosenberg 2006). Aligning ITS sequences can be especially difficult even within fungal genera, and criteria for alignment quality differ among researchers. Therefore, three different alignment methods were examined in the context of Bayesian and NJ phylogenetic inference methods.

The different alignments examined here resulted in no well-supported topological conflicts when relationships for each genus were inferred using Bayesian methods. However, trees generated from our category 1 (MUSCLE only) and category 2 (MUSCLE with manual adjustment) alignments were most similar to each other. This was an expected result given the low level of sequence editing that differentiated the category 2 alignments from category 1. Topologies generated from the Bayesian analyses of the category 3 (manual) alignments have less resolution than category 1 and category 2 alignments, reflecting manual exclusion of ambiguously aligned regions. Model selection did not have a large effect on the tree topologies generated in this study, possibly reflecting that the different models implemented here, based on ModelTest analyses, are similar to one another (e.g. SYM + I+ G is the same as GTR + I+ G except for fixed equal nucleotide frequencies).

Although NJ may perform poorly relative to other phylogenetic inference methods in many cases (see Hall 2005), many studies rely on NJ for estimating the relationships of unknown fungi because of the algorithm's speed with large datasets (e.g. Schadt *et al.* 2003). Our results suggest that NJ inferences are much more sensitive to ambiguous regions in alignments than are Bayesian methods (Table 4). NJ analyses also tended to give higher statistical support to terminal (species-level) clades than did Bayesian analyses (Supplementary Material Figs S3–S6), possibly overestimating confidence in those relationships. Overall, these results are not unexpected given results observed for simulated data sets, wherein NJ has been shown to be more sensitive to alignment error (Ogden & Rosenberg 2006). We conclude that caution in the use of NJ methods alone is warranted given the difficulty of aligning ITS data for many fungal taxa, the lack of phylogenetic uncertainty in NJ analyses, and the potential to overestimate certainty regarding terminal relationships.

Our Bayesian analyses demonstrate that genera differ significantly in their sensitivity to alignment methods. This reflects differences in the variability in ITS sequence data within particular genera, and highlights the need to consider the identity of the fungi prior to selecting an analysis approach. In contrast, genera did not differ in their sensitivity to different alignments when NJ analyses were performed. In general, BS values are generally in agreement with PPs for category 2 alignments under both inference methods (Figs 2–5), which supports

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

the validity of our subsequent analyses regarding the evolution of seed-associated fungi in these genera.

Evolutionary origins of seed-associated fungi

After incubation below *Cecropia* trees in lowland tropical forest, intact seeds of *C. insignis* are frequently infected by a diverse array of filamentous fungi, especially in the *Sordariomycetes* and *Dothideomycetes*. Given their diversity (this study; see also Gallery *et al.* 2007a), their spatial heterogeneity in the forest (see Kluger *et al.* 2008), and their variable evolutionary relationships with regard to pathogens, endophytes, and saprotrophs (this study), it is likely that these fungi play diverse ecological roles. Recent studies have suggested that the most common seed-associated fungi in tropical forests are host-generalists (e.g. Kluger *et al.* 2008), but that they still have profound effects on seed germination for particular tree species, as well as site-specific effects on seed survival (see Gallery *et al.* 2007a; Kluger *et al.* 2008). Our data complement these studies by not only providing a first hypothesis for the phylogenetic relationships of seed-associated fungi in several genera, but also an initial set of hypotheses regarding their evolutionary and ecological relationships with representative pathogenic, endophytic and saprotrophic fungi.

Consistent with previous suggestions of horizontal transmission (see Gallery *et al.* 2007a; Kluger *et al.* 2008), our data show that seeds from different geographic regions (Costa Rica, Panama) accumulate diverse but highly similar communities of fungi following burial together in Panama. However, the prevalence of known endophytic genotypes among seed-associated fungi, almost certainly an underestimate of the congruence of these guilds given that neither endophyte nor seed-associated fungi have been sampled to saturation in this forest, leaves open the possibility that some fungi in seeds are vertically transmitted endophytes.

Whereas sequence identity can reveal the overlap of fungi among different guilds, phylogenetic analyses provide insight into the evolutionary relationships of different fungal taxa and ecological modes. Our results indicate close, well-supported relationships of seed-associated fungi and endophytes in three of four focal genera (*Botryosphaeria*, *Mycosphaerella*, and *Xylaria*). Likelihood reconstructions were performed to find the character state of each node that maximizes the probability of getting the observed states (i.e. seed-associated, endophyte, pathogen or saprotroph) for terminal taxa. For all three genera, the second ancestral state reconstruction method used here, which considers the average probability of any given ecological mode among trees in which that node was present, reconstructed ecological states more often than the first method. However, the first method, which provides the proportion of posterior trees in which a given state was significantly more likely than alternative states, or in which a node was equivocal or absent, was able to unequivocally reconstruct some nodes closer to the tips of these trees.

Within these focal genera, the fungi we recovered from surface-sterilized seeds appear to share a closer evolutionary history with pathogens and endophytes than with saprotrophs. However, the observation of putatively saprotrophic fungi as endophytes in other studies (e.g. Fisher & Petrini 1992; Hyde *et al.* 2007; Promputtha *et al.* 2007), and the possibility that some seed-associated fungi are saprotrophs despite being isolated from the interior of surface-sterilized seeds, requires further attention. Given that endophytes of tropical trees gain entry to the leaf in order to colonize intercellularly, it follows that endophytes and seed-colonizing fungi could share a close relationship with pathogens, which also gain entry to living tissue to cause disease. Future work will involve more extensive sampling of pathogens, endophytes, and seed-associated fungi to more clearly reveal these relationships, and will rely in part on trustworthy analysis methods. The latter issue motivated and provided the context for the present study, and will continue to be improved by careful attention from mycologists, enhanced methods, and expansion of environmental samples beyond single-locus datasets.

Acknowledgements

We gratefully acknowledge the College of Agriculture and Life Sciences at the University of Arizona and the National Science Foundation for supporting this research (DEB-0200413, DEB-0342925, DEB-0516564, and DEB-0640996 to AEA, and DEB-0343953 to JWD). We further thank NSF for fostering discussion that informed this work through the Fungal Environmental Sampling and Informatics Network (FESIN; DEB-0639048 to Tom Bruns, Karen Hughes, and AEA). JMU was supported by an NSF-IGERT Fellowship in Genomics at the University of Arizona while completing this study. We thank François Lutzoni and A. Jon Shaw for logistical support at Duke University, and Malkanthi Gunatilaka, K. Lindsay Higgins, Michele Hoffman, Megan McGregor, Alyson Paulick, Evelyn Sanchez, Brett Wolfe and Mariana del Olmo Ruiz for technical assistance. The helpful suggestions of two anonymous reviewers improved this manuscript, for which we are most grateful.

Supplementary material

Supplementary data associated with this article can be found in the online version, at doi: 10.1016/j.mycres.2008.11.015

REFERENCES

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, 1990. Basic local alignment search tool. *Journal of Molecular Biology* **215**: 403–410.

Alvarez-Buylla ER, Martínez-Ramos M, 1990. Seed bank versus seed rain in the regeneration of a tropical pioneer tree. *Oecologia* **84**: 314–325.

Arenz BE, Held BW, Jurgens JA, Farrell RL, Blanchette RA, 2006. Fungal diversity in soils and historic wood from the Ross Sea region of Antarctica. *Soil Biology & Biochemistry* **38**: 3057–3064.

Arnold AE, Henk DA, Eells RL, Lutzoni F, Vilgalys R, 2007. Diversity and phylogenetic affinities of foliar fungal endophytes in loblolly pine inferred by culturing and environmental PCR. *Mycologia* **99**: 185–206.

Arnold AE, Lutzoni F, 2007. Diversity and host range of foliar fungal endophytes: are tropical leaves biodiversity hotspots? *Ecology* **88**: 541–549.

Arnold A.E., Miadlikowska J., Higgins K.L., Sarvate S.D., Gugger P., Way A., Hofstetter V., Kauff F. & Lutzoni. Hyperdiverse fungal endophytes and endolichenic fungi elucidate the evolution of major ecological modes in the *Ascomycota*. *Systematic Biology* In press.

Bridge PD, Roberts PJ, Spooner BM, Panchal G, 2003. On the unreliability of published DNA sequences. *New Phytologist* **160**: 43–48.

Colwell RK, 2005. EstimateS: Statistical estimation of species richness and shared species from samples. Online, at <http://viceroy.eeb.uconn.edu/EstimateS>.

Dalling JW, Swaine MD, Garwood NC, 1997. Soil seed bank community dynamics in seasonally moist lowland tropical forest, Panama. *Journal of Tropical Ecology* **13**: 659–680.

Dalling JW, Swaine MD, Garwood NC, 1998. Dispersal patterns and seed bank dynamics of pioneer trees in moist tropical forest. *Ecology* **79**: 564–578.

Davis EC, Franklin JB, Shaw AJ, Vilgalys R, 2003. Endophytic *Xylaria (Xylariaceae)* among liverworts and angiosperms: phylogenetics, distribution, and symbiosis. *American Journal of Botany* **90**: 1661–1667.

Denman S, Crous PW, Groenwald JZ, Slippers B, Wingfield BD, Wingfield MJ, 2003. Circumscription of *Botryosphaeria* species associated with *Proteaceae* based on morphology and DNA sequence data. *Mycologia* **95**: 294–307.

Du M, Schardl CL, Nuckles EM, Vaillancourt LJ, 2005. Using mating-type gene sequences for improved phylogenetic resolution of *Colletotrichum* species complexes. *Mycologia* **97**: 641–658.

Edgar RC, 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792–1797.

Ewing B, Green P, 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**: 186–194.

Ewing B, Hillier L, Wendl MC, Green P, 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Research* **8**: 175–185.

Feau N, Hamelin RC, Bernier L, 2006. Attributes and congruence of three molecular datasets: inferring phylogenies among *Septoria*-related species from woody perennial plants. *Molecular Phylogenetics and Evolution* **40**: 808–829.

Fisher PJ, Petrini O, 1992. Fungal saprobes and pathogens as endophytes of rice (*Oryza sativa* L). *New Phytologist* **120**: 137–143.

Gallery RE, Dalling JW, Arnold AE, 2007a. Diversity, host affinity, and distribution of seed-infecting fungi: a case study with *Cecropia*. *Ecology* **88**: 582–588.

Gallery RE, Dalling JW, Wolfe B, Arnold AE, 2007b. Role of seed-infecting fungi in the recruitment limitation of neotropical pioneer species. In: Dennis A, Green R, Schupp E, Westcott D (eds), Chapter 23 in Seed Dispersal: Theory and its Application in a Changing World. CABI Press. pp. 479–498.

Geml J, Via Z, Long J, Huston S, Marr T, Taylor DL, 2005. The Fungal Metagenomics Project. <http://www.borealfungi.uaf.edu>.

Gordon D, Abajian C, Green P, 1998. Consed: a graphical tool for sequence finishing. *Genome Research* **8**: 195–202.

Hall BG, 2005. Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. *Molecular Biology and Evolution* **22**: 792–802.

Hall JB, Swaine MD, 1980. Seed stocks in Ghanaian forest soils. *Biotropica* **12**: 256–263.

Harris DJ, 2003. Can you bank on GenBank? *Trends in Ecology & Evolution* **18**: 317–319.

Henry T, Iwen PC, Hinrichs SH, 2000. Identification of *Aspergillus* species using internal transcribed spacer regions 1 and 2. *Journal of Clinical Microbiology* **38**: 1510–1515.

Hoffman M, Arnold AE, 2008. Geography and host identity interact to shape communities of endophytic fungi in cupressaceous trees. *Mycological Research* **112**: 331–344.

Hogberg N, Land CJ, 2004. Identification of *Serpula lacrymans* and other decay fungi in construction timber by sequencing of ribosomal DNA — a practical approach. *Holzforschung* **58**: 199–204.

Huelsenbeck JP, Ronquist F, 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**: 754–755.

Hyde KD, Bussaban B, Paulus B, Crous PW, Lee S, McKenzie EHC, Photita W, Lumyong S, 2007. Diversity of saprobic microfungi. *Biodiversity and Conservation* **16**: 7–35.

James TY, Kauff F, Schoch CL, Matheny PB, Hofstetter V, Cox CJ, Celio G, Gueidan C, Fraker E, Miadlikowska J, Lumbsch HT, Rauhut A, Reeb V, Arnold AE, Amtoft A, Stajich JE, Hosaka K, Sung GH, Johnson D, O'Rourke B, Crockett M, Binder M, Curtis JM, Slot JC, Wang Z, Wilson AW, Schüßler A, Longcore JE, O'Donnell K, Mozley-Standridge S, Porter D, Letcher PM, Powell MJ, Taylor JW, White MM, Griffith GW, Davies DR, Humber RA, Morton JB, Sugiyama J, Rossman AY, Rogers JD, Pfister DH, Hewitt D, Hansen K, Hambleton S, Shoemaker RA, Kohlmeyer J, Volkmann-Kohlmeyer B, Spotts RA, Serdani M, Crous PW, Hughes KW, Matsuura K, Langer E, Langer G, Untereiner WA, Lücking R, Büdel B, Geiser DM, Aptroot A, Diederich P, Schmitt I, Schultz M, Yahr R, Hibbett DS, Lutzoni F, McLaughlin DJ, Spatafora JW, Vilgalys R, 2006. Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature* **443**: 818–822.

Kluger CG, Dalling JW, Gallery RE, Sanchez E, Weeks-Galindo C, Arnold AE, 2008. Prevalent host-generalism among fungi associated with seeds of four neotropical pioneer species. *Journal of Tropical Ecology* **24**: 351–354.

Kõljalg U, Larsson KH, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM, 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* **166**: 1063–1068.

Koski LB, Golding GB, 2001. The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution* **52**: 540–542.

Lambert C, Campenhout J-MVan, DeBolle X, Depiereux E, 2003. Review of common sequence alignment methods: clues to enhance reliability. *Current Genomics* **4**: 131–146.

Landan G, Graur D, 2007. Heads or tails: a simple reliability check for multiple sequence alignments. *Molecular Biology and Evolution* **24**: 1380–1383.

Lee JS, Ko KS, Jung HS, 2000. Phylogenetic analysis of *Xylaria* based on nuclear ribosomal ITS1–5.8S–ITS2 sequences. *FEMS Microbiology Letters* **187**: 89–93.

Little DP, Stevenson DW, 2007. A comparison of algorithms for the identification of specimens using DNA barcodes: examples from gymnosperms. *Cladistics* **23**: 1–21.

Lutzoni F, Kauff F, Cox CJ, McLaughlin D, Celio G, Dentinger B, Padamsee M, Hibbett D, James TY, Baloch E, Grube M, Reeb V, Hofstetter V, Schoch C, Arnold AE, Miadlikowska J, Spatafora J, Johnson D, Hambleton S, Crockett M, Shoemaker R, Sung GH, Lücking R, Lumbsch T, O'Donnell K, Binder M, Diederich P, Ertz D, Gueidan C, Hansen K, Harris RC, Hosaka K, Lim YW, Matheny B, Nishida H, Pfister D, Rogers J, Rossman A, Schmitt I, Sipman H, Stone J, Sugiyama J, Yahr R, Vilgalys R, 2004. Assembling the fungal tree of life: progress, classification and evolution of subcellular traits. *American Journal of Botany* **91**: 1446–1480.

Maddison DR, Maddison WP, 2003. *MacClade 4.06: Analysis of Phylogeny and Character Evolution*. Sinauer Associates, Sunderland, MA.

Morrison DA, Ellis JT, 1997. Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of Apicomplexa. *Molecular Biology and Evolution* **14**: 428–441.

Murray KG, Garcia M, 2002. Contributions of seed dispersal to recruitment limitation in a Costa Rican cloud forest. In: Levey DJ, Silva WR, Galetti M (eds), Seed Dispersal and Frugivory:

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015

ecology, Evolution and Conservation. CAB International Press, Wallingford, UK, pp. 323–338.
Nilsson RH, Kristiansson E, Ryberg M, Larsson KH, 2005. Approaching the taxonomic affiliation of unidentified sequences in public databases — an example from the mycorrhizal fungi. *BMC Bioinformatics* **6**.
O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys R, 2005. Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology* **71**: 5544–5550.
O'Hanlon-Manners DL, Kotanen PM, 2004. Evidence that fungal pathogens inhibit recruitment of a shade-intolerant tree, white birch (*Betula papyrifera*), in understory habitats. *Oecologia* **140**: 650–653.
Ogden TH, Rosenberg MS, 2006. Multiple sequence alignment accuracy and phylogenetic inference. *Systematic Biology* **55**: 314–328.
Pearson WR, 1998. Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* **276**: 71–84.
Pearson WR, 2000. Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **132**: 185–219.
Posada D, Crandall KA, 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817–818.
Promputtha I, Lumyong S, Dhanasekaran V, McKenzie EHC, Hyde KD, Jeewon R, 2007. A phylogenetic evaluation of whether endophytes become saprotrophs at host senescence. *Microbial Ecology* **53**: 579–590.
Schadt CW, Martin AP, Lipson DA, Schmidt SK, 2003. Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science* **301**: 1359–1361.
Schloss PD, Handelsman J, 2005. Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71**: 1501–1506.
Seifert KA, Samson RA, Dewaard JR, Houbraken J, Levesque CA, Moncalvo JM, Louis-Seize G, Hebert PDN, 2007. Prospects for fungus identification using C01 DNA barcodes, with *Penicillium* as a test case. *Proceedings of the National Academy of Sciences USA* **104**: 3901–3906.
Smith H, Wingfield MJ, Crous PW, Coutinho TA, 1996. *Sphaeropsis sapinea* and *Botryosphaeria dothidea* endophytic in *Pinus* spp and *Eucalyptus* spp in South Africa. *South African Journal of Botany* **62**: 86–88.
Stackebrandt E, Göbel BM, 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology* **44**: 846–849.
Swart L, Crous PW, Petrini O, Taylor JE, 2000. Fungal endophytes of *Proteaceae*, with particular emphasis on *Botryosphaeria proteae*. *Mycoscience* **41**: 123–127.
Swofford DL, 2002. *PAUP*: Phylogenetic analysis using parsimony (*and other methods) Version 4.0b10*. Sinauer Associates, Sunderland, MA.
Taylor JW, Jacobson DJ, Kroken S, Kasuga T, Geiser DM, Hibbett DS, Fisher MC, 2000. Phylogenetic species recognition and species concepts in fungi. *Fungal Genetics and Biology* **31**: 21–32.
Vandenkoornhuyse P, Baldauf SL, Leyval C, Straczek J, Young JPW, 2002. Extensive fungal diversity in plant roots. *Science* **295**: 2051.
Vilgalys R, 2003. Taxonomic misidentification in public DNA databases. *New Phytologist* **160**: 4–5.

Please cite this article in press as: Jana M U'Ren et al., Diversity and evolutionary origins of fungi associated with seeds of a neotropical pioneer tree: a case study for analysing fungal environmental samples, Mycological Research (2009), doi:10.1016/j.mycres.2008.11.015