



Evaluation of the authenticity of a highly novel environmental sequence from boreal forest soil using ribosomal RNA secondary structure modeling

Daniel J. Glass^a, Naoki Takebayashi^a, Link E. Olson^b, D. Lee Taylor^{a,*}

^a Institute of Arctic Biology, 311 Irving I Building, 902 N. Koyukuk Drive, University of Alaska Fairbanks, Fairbanks, AK 99775, United States

^b University of Alaska Museum, 907 Yukon Drive, Fairbanks, AK 99775, United States

ARTICLE INFO

Article history:

Received 28 August 2012

Revised 9 January 2013

Accepted 29 January 2013

Available online 9 February 2013

Keywords:

Novel lineage

Ribosomal RNA secondary structure

Pseudogene

rDNA

Fungi

Metagenomic DNA

ABSTRACT

The number of sequences from both formally described taxa and uncultured environmental DNA deposited in the International Nucleotide Sequence Databases has increased substantially over the last two decades. Although the majority of these sequences represent authentic gene copies, there is evidence of DNA artifacts in these databases as well. These include lab artifacts, such as PCR chimeras, and biological artifacts such as pseudogenes or other paralogous sequences. Sequences that fall in basal positions in phylogenetic trees and appear distant from known sequences are particularly suspect. Phylogenetic analyses suggest that a novel sequence type (NS1) found in two boreal forest soil clone libraries belongs to the fungal kingdom but does not fall unambiguously within any known phylum. We have evaluated this sequence type using an array of secondary-structure analyses. To our knowledge, such analyses have never been used on environmental ribosomal sequences. Ribosomal secondary structure was modeled for four rRNA loci (ITS1, 5.8S, ITS2, 5' LSU). These models were analyzed for the presence of conserved domains, conserved nucleotide motifs, and compensatory base changes. Minimal free energy (MFE) foldings and GC contents of sequences representing the major fungal clades, as well as NS1, were also compared. NS1 displays secondary rRNA structures consistent with other fungi and many, but not all, conserved nucleotide motifs found across eukaryotes. However, our analyses show that many other authentic sequences from basal fungi lack more of these conserved motifs than does NS1. Together our findings suggest that NS1 represents an authentic gene copy. The methods described here can be used on any rRNA-coding sequence, not just environmental fungal sequences. As new-generation sequencing methods that yield shorter sequences become more widely implemented, methods that evaluate sequence authenticity should also be more widely implemented. For fungi, the adjacent 5.8S and ITS2 loci should be prioritized. This region is not only suited to distinguishing between closely related species, but it is also more informative in terms of expected secondary structure.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

The number of uncultured environmental fungal DNA sequences in public databases has grown exponentially during the last two decades. For example, the number of fungal internal transcribed spacer (ITS) sequences deposited per year in the INSD rose from roughly 2000 sequences in the year 2000 to roughly 22,000 in 2007 (Ryberg et al., 2009). Sequences in the INSD are frequently

Abbreviations: NS1, Novel Sequence Type One; BFL, basal fungal lineages; rRNA, ribosomal RNA; rDNA, ribosomal DNA; ITSs, internal transcribed spacers; SSU, small subunit; LSU, large subunit; MFE, minimal free energy; CBC, compensatory base change; NUMT, mitochondrion-derived nuclear pseudogene; BNZ-LTER, Bonanza Creek Long Term Ecological Research Station; AFTOL, Assembling the Fungal Tree of Life; INSD, International Nucleotide Sequence Databases.

* Corresponding author. Fax: +1 907 474 6967.

E-mail addresses: ltaylor@iab.alaska.edu, dltaylor@alaska.edu (D.L. Taylor).

used as reference sequences in phylogenetic reconstructions (Bridge et al., 2003) or in BLAST (Altschul et al., 1999) searches to determine the identities of new sequences. Accurate phylogenetic reconstructions require comparisons of orthologous sequences, and if unidentified lab or biological artifacts such as pseudogenes are present, phylogenetic inferences may be misleading (Bensasson et al., 2001; Olson and Yoder, 2002). Therefore, it is imperative that sequences accessioned into the INSD do not include unacknowledged biological or lab artifacts. Several types of artifactual rDNA have been encountered, particularly chimeras (Osborne et al., 2005) and pseudogenes (including nuclear mitochondrial-derived DNAs, or NUMTs; Anthony et al., 2007). Chimeras are common artifacts that can confound phylogenetic inference and species discrimination. Chimeras can result from one of two separate phenomena. During PCR, when two different sequences are used as template DNA, the resulting product may

include a combination of these original sequences (Jumpponen, 2007). A study by Jumpponen (2007) found an alarmingly high proportion (>30%) of sequences in two fungal amplicon clone libraries to be chimeric. Many of these chimeric sequences were located at more basal positions in the phylograms than the DNA sequences from which they were derived. If undetected, this could have led to the spurious inference of novel higher-level taxa (Hugenholtz and Huber, 2003). Post-PCR chimeric DNA sequences can also be generated when multiple short contiguous or minimally-overlapping PCR products from different template sequences are inadvertently joined during sequence assembly to produce longer continuous sequences, as can happen in studies of ancient or degraded DNA (Olson and Hassanin, 2002).

Pseudogenes are copies of a gene that originated from a functional gene copy but are no longer functional. Therefore, they are no longer constrained by the same selective pressures as their functional counterparts (Perna and Kocher, 1996). In the case of eukaryotic tandemly-repeated nuclear ribosomal genes, this usually involves transposition to a novel chromosomal location, after which concerted evolution fails to maintain sequence homogeneity (Balakirev and Ayala, 2003). A particularly worrisome type of pseudogene for studies based on rDNA is the NUMT. NUMTs, or mitochondrial-derived nuclear pseudogenes, are the complementary genes in the mitochondrial genome that have been transposed into the nuclear genome (Richly and Leister, 2004). These are particularly difficult to detect, especially if recently derived, and are known to mislead phylogenetic inferences (Bensasson et al., 2001; Olson and Yoder, 2002; Perna and Kocher, 1996). How much impact the inclusion of a pseudogene will have on phylogenetic inferences depends on both the age of the pseudogene and the degree of differentiation needed to adequately address the research questions being asked. For example, the use of an anciently-derived pseudogene arising from a familiar, formally described taxon would most likely differ greatly in functionally-conserved regions of the sequence, and if detected, should be excluded from phylogenetic reconstructions. Otherwise, it could lead to long branches in phylograms that do not represent the true amount of evolutionary distance between taxa. Alternatively, a recently-derived pseudogene representing a very novel and potentially anciently-derived lineage would show little variation between paralogs compared to the amount of variation represented by the evolutionary history of the taxa being compared. If only deep phylogenetic affinities are being pursued and species-level resolution is not needed, the inclusion of a recently-derived pseudogene instead of the targeted orthologous sequences may have relatively little impact on placement of the sequence in a tree.

The majority of studies of fungal phylogenetics, including ours, have targeted ribosomal RNA (rRNA)-coding sequences encompassing one or all of the following nuclear rDNA loci: the small subunit (SSU), the large subunit (LSU), or the internal transcribed spacers (ITS) including the 5.8S region (Fig. 1). Fortunately, rRNA-coding sequences need to preserve core rRNA secondary structures for proper RNA processing. This has led to the evolutionary

conservation of certain domains and nucleotide motifs across the eukaryotic kingdom (Coleman, 2007). By analyzing the predicted secondary structure of an rRNA sequence and verifying that conserved domains and motifs are present, we can estimate whether the sequence is likely to code for functional rRNA (Harpke and Peterson, 2008; but see Olson and Yoder, 2002). To our knowledge, analyses of ribosomal secondary structure have never been undertaken on environmental sequences. This is surprising, given that secondary structure analysis is one of the more widely-implemented tools for validating the authenticity of rRNA gene copies (Harpke and Peterson, 2008; Olson and Yoder, 2002; Xiao et al., 2010; Zheng et al., 2008) and that community analyses of environmental microbial systems depend heavily on comparisons of only orthologous (versus paralogous) sequences.

A fungal amplicon clone library spanning partial ribosomal large subunit (LSU) and the internal transcriber spacers (ITS) constructed by Taylor et al. (2007) from boreal forest soils at the Bonanza Creek LTER near Fairbanks, AK, yielded several DNA sequences (~1200 bp) that were highly divergent from any publicly available sequences. Here we focus on one such sequence type, encountered multiple times, the level of divergence of which led us to question whether it belongs to the fungal kingdom (Eumycota) and whether it might be an artifact. We attempted to determine the phylogenetic affinities of this sequence type, which we label NS1. We then employed various methods to analyze its predicted secondary structure. We also evaluated which rRNA loci were most informative in this endeavor. Although the methods outlined have been employed on what appears to be a fungal DNA sequence, they are potentially applicable to any novel rDNA sequence.

With the increasing numbers of novel fungal lineages that were first or only recorded in environmental DNA studies (Schadt et al., 2003; Lara et al., 2010), it is becoming increasingly valuable to include environmental sequences in fungal systematics (Hibbett et al., 2011). At the same time, molecular systematics of the Eumycota is undergoing radical revision in the light of intensive multi-taxon, multi-locus studies (James et al., 2006; see the special issue of Mycologia, 2006, volume 98, issue 6). There is strong molecular support for monophyly of the Eumycota (Baldauf and Palmer, 1993; Bruns et al., 1992; Steenkamp et al., 2006), a kingdom defined by heterotrophy, absorptive nutrition, chitinous walls, apical growth, haploid thalli, spindle-pole bodies, and trehalose as a major storage sugar, among other features. Some of the most primitive fungi have flagellated dispersal stages, while all of the more derived lineages lack flagella. About 20 years ago, a relatively stable system of five fungal phyla achieved universal recognition based, in part, on initial rDNA phylogenies (Bruns et al., 1992). These phyla included the simple and ancient flagellated water-molds assigned to the Chytridiomycota, which include the epidemic amphibian pathogen *Batrachochytrium dendrobatidis*, the ancient non-flagellated filamentous bread-molds of the Zygomycota, the transitional arbuscular mycorrhizal fungi now placed within the Glomeromycota (Schüßler et al., 2001), and the

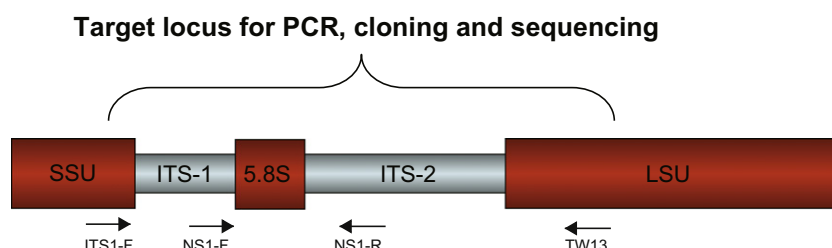


Fig. 1. Map of nuclear ribosomal gene-regions including primer binding sites.

two crown phyla. The two crown phyla are the Ascomycota, named for their flask-shaped reproductive cells, which include iconic taxa such as fission yeast, bakers yeast, *Penicillium*, *Neurospora* and most lichen-forming fungi, and the Basidiomycota, named for their club-shaped reproductive cells, which include the majority of ‘mushrooms’ as well as plant-pathogenic rusts and smuts.

Later multilocus molecular analyses have suggested that the two basal phyla are not monophyletic (O'Donnell et al., 2001) and that other deeply divergent lineages such as the recently named Cryptomycota (formerly the *Rozella*-clade; Jones et al., 2011) are interspersed with the former Zygomycota and Chytridiomycota. Several new phyla and subphyla have been proposed (Hibbett et al., 2007), including the Blastocladiomycota and Neocallimastigomycota (formerly members of Chytridiomycota), and the Entomophthoromycotina, Kickxellomycotina, Mucoromycotina, and Zoopagomycotina (formerly members of Zygomycota). However, because the relationships of the most ancient fungal lineages are poorly resolved and their systematics are in flux, these lineages are often simply referred to as the ‘basal fungal lineages’ (BFL). We use this colloquial identifier here for convenience, recognizing that it does not represent a natural grouping, and that one extant taxon cannot be basal to another extant taxon (see Krell and Cranston, 2004). Although this use of basal and basal fungal lineages is inconsistent with phylogenetic principles, we aimed for consistency with other mycological literature on the topic (see Hibbett et al., 2007; James et al., 2006; Tanabe et al., 2005; Voigt and Kirk, 2011).

2. Methods

2.1. Initial DNA extraction, amplification, cloning, and sequencing

The molecular methods used to isolate and identify fungal DNA sequences from soils within both the riparian black spruce site (FP5C: 64.71361404N, 148.1472629W) and the upland site (UP2A: 64.69546067N, 148.35568W) have been described in detail previously (Geml et al., 2009, 2010; Taylor et al., 2010). In brief, the segment (~1200 bp) encompassing the ribosomal internal transcribed spacers (ITS) and a portion (~700 bp) of the ribosomal large subunit (LSU; Fig. 1) was amplified from soil extracts using the fungal-specific PCR primers ITS1-F (Gardes and Bruns, 1993) and TW13 (Taylor and Bruns, 1999) or the slightly modified (longer, tagged) variants of these primers described in Taylor et al. (2008). Amplicons were cloned into pCR®4-TOPO vectors using a TOPO-TA kit (Invitrogen, Carlsbad, CA, USA) and sent to the Broad Institute of MIT and Harvard where transformations, automated clone-picking, and sequencing of clone libraries took place. Additional details are provided in the [Supplementary Materials](#).

2.2. Identification of novel sequences within clone libraries

A BLAST (Altschul et al., 1999) search against known fungi revealed a novel DNA sequence within the black spruce clone dataset that was divergent from known fungal phyla and, whose relationship to known fungal lineages was uncertain. Initial chimera checking was conducted by running separate BLAST sequence similarity searches for the ITS1 and ITS2 regions. For a more detailed description of these methods and stringency cutoffs, see Geml et al. (2009). This sequence was subsequently re-examined for chimerism using the program Uchime within the Uclust package (Edgar et al., 2011).

2.3. Primer design and reamplification of novel taxa

To evaluate the authenticity of this sequence, we used Primer3 (Koressaar and Remm, 2007) to design primers (NS1-F and NS1-R;

Fig. 1) to specifically target a 290 bp diagnostic region encompassing a portion of ITS1, the 5.8S, and a portion of ITS2 (for additional details see [Supplementary Information](#)). Using these primers, we performed PCR on an array of soil DNA extracts from Interior Alaska.

2.4. Phylogenetic analyses

In order to estimate the phylogenetic affinities of this novel sequence, we added the ~700 bp 5' LSU region of NS1 sequence to an alignment constructed by Tim Y. James (University of Michigan), which was similar to that of White et al. (2006). The LSU is widely used in fungal phylogenetics because it is sufficiently conserved to construct multiple sequence alignments containing distantly-related taxa, but variable enough to provide support for monophyletic genera and families. We also constructed a separate alignment of the highly conserved 5.8S region, combining taxa from the matrices published in Cullings and Vogler (1998) and James et al. (2006). The ITS spacer regions bracketing the 5.8S are hypervariable and consequently nearly impossible to use in multiple sequence alignments beyond a single genus. Both LSU and 5.8S alignments were optimized using T-Coffee (Notre-dame et al., 2000). The D1 and D2 loops of the LSU alignment were highly variable across the Eumycota, with suspect positional homology. We therefore pruned the resulting alignment to relatively conserved columns of bases using lenient settings in GBLOCKS (Castresana, 2000). The same 5.8S alignment was used for secondary structure analyses (see below), except that we removed the outgroups because our goal was to estimate a consensus fungal folding; we also removed the two uncultured environmental fungi sequences, since their authenticity cannot be verified. These alignments and the resulting tree files have been deposited in TreeBASE (Sanderson et al., 1994) under accession 13807.

We constructed trees using maximum likelihood and Bayesian methods. The Akaike Information Criterion in Modeltest 3.06 (Posada and Crandall, 1998) and MrModeltest (Nylander, 2004) was used to determine the best-fit evolutionary model to be implemented in RAXML (Stamatakis, 2006) and MrBayes (Ronquist and Huelsenbeck, 2003), respectively. For tree reconstructions, outgroups representing basal opisthokonts (LSU) or Viridiplantae (5.8S), but distinct from the kingdom Eumycota, were used. Tree-Graph2 (Stöver and Müller, 2010) was used to annotate trees and export graphics. Alignments and best trees were combined in Mesquite (Maddison and Maddison, 2008) for upload to TreeBASE.

To evaluate whether sequence divergence for NS1 is accelerated relative to the most closely-related taxa available, we carried out Tajima's relative rates tests using the 5.8S of NS1 and six representative basal fungi using MEGA (v. 5; Tamura et al., 2011).

2.5. Tests to evaluate the authenticity of NS1

We used a number of methods (described below) to evaluate functional attributes of the novel environmental sequence type, NS1. Comparative analyses were carried out using sequences spanning the kingdom Eumycota from the Assembling the Fungal Tree of Life (AFTOL) project (<http://aftol.org/>). We chose these sequences for comparison because they were derived from vouchered and taxonomically identified herbarium specimens and isolates (Lutzoni et al., 2004). Furthermore, the published sequences have been vetted using a series of automated and manual checks (Lutzoni et al., 2004). Hence, of available sequences, these are highly likely to constitute authentic orthologs rather than pseudogenes or other artifacts. The majority of the sequences used in our phylogenetic reconstructions were also used in the analyses described below. Those that were excluded were omitted for one of

two reasons: because we were unable to identify and separate loci, particularly the SSU (which was not used in these analyses), or because a sequence comprising all four loci was not available. GenBank sequences were used in place of AFTOL sequences if they exactly matched the corresponding AFTOL LSU sequence and, when possible, also matched the corresponding sequence in the AFTOL 5.8S alignment (James et al., 2006). GenBank sequences representing the species (*Sphaeronaemella fimicola*), believed to have the shortest ITS of any fungus, were also modeled for comparison (Fujita et al., 2001). Accession numbers for all sequences downloaded from GenBank, and for the two full-length NS1 sequences are provided in Table S1.

2.5.1. Modeling ribosomal secondary structure

We used a variety of programs from the Vienna RNA Websuite (<http://rna.tbi.univie.ac.at/>; Gruber et al., 2008; Hofacker, 2003) to search for anomalous features within NS1 that might suggest that it is an artifact. The secondary structure of NS1 was modeled using default settings in RNAfold. We modeled a folding of the entire original sequence, as well as independent foldings for each of the four constituent loci (ITS1, 5.8S, ITS2, and partial LSU). Minimal free energy structures were compared to previously modeled fungal secondary structures to determine if core conserved regions, including pan-eukaryotic homologies, were present.

RNAalifold (Bernhart et al., 2008; Hofacker et al., 2002) is a program designed to produce consensus secondary structures from multiple sequence alignments. Roughly 180 representative sequences of the major fungal lineages were extracted from the 5.8S and LSU AFTOL alignments of James et al. (2006) and used as a template against which to align our novel sequence. The LSU alignment predominately included the same taxa used in the phylogenetic reconstructions. These multiple sequence alignments were used to create consensus Eumycota secondary structures of the 5.8S and 5' LSU regions using default settings in RNAalifold. This program also produces a multiple sequence alignment with the conserved stem regions from the consensus secondary structure and their positional probabilities (i.e., the chance that a given sequence has that base present at that position) superimposed onto the alignment. We were thereby able to evaluate whether conserved regions in the primary sequence are also conserved in the secondary structure. We also compared output from RNAalifold to the RNAfold folding of the NS1 5.8S to determine how many compensatory base changes (wherein both paired nucleotides have changed while bond has been preserved) were present in NS1 relative to other fungi.

2.5.2. Minimal free energy regression

In order to determine the minimal free energies (MFEs) of several sequences at once, we used the Quikfold program on the mfold web server (<http://mfold.rna.albany.edu/?q=DINAMelt/Quikfold>). Settings were changed to RNA, and the RNA 3.0 model for folding was used; all other settings were left at the defaults. One structure per sequence was produced. We performed linear regression in R v 2.8.1 (R Development Core Team, 2008) comparing locus length and MFE. The ITS1 and ITS2 regions of NS1 and reference sequences were modeled separately. This allowed us to estimate a typical MFE for a given sequence length and determine if NS1 fell within the 95% prediction interval for each locus.

2.5.3. Comparison of ITS1 and ITS2 length and GC contents

Prior studies have suggested that although length may differ between ITS1 and ITS2 for a given species, the two loci will usually have very similar GC contents (Harpke and Peterson, 2006; Mullineux and Hausner, 2009; Torres et al., 1990; Zheng et al., 2008). In general, pseudogenes frequently have higher AT content (Harpke and Peterson, 2006). To test this for NS1, the InfoSeq tool on the

EMBOSS web server (Rice et al., 2000) was used to calculate GC content of each locus. This same program was used to calculate the length of each locus.

In order to determine if NS1 fell within the range of natural variation of GC content for these loci, we compared values for NS1 to reference sequences obtained from the AFTOL website and GenBank. We determined the mean GC content and standard deviation for each locus separately, plotted the distribution of values, and determined where NS1 fell within this distribution. We then compared our results for ITS1 and ITS2 to determine if NS1 varied in GC content between loci.

3. Results

3.1. PCR and phylogenetic results

Using our taxon-specific primers we were able to amplify the 290 bp fragment of NS1 from several soil extracts in addition to the sample from which the original clone sequence was obtained. Sequences from these amplicons closely matched the original singleton clone (Supplementary Information). In addition, a ~1200 bp sequence very similar to the original NS1 sequence, differing at only eight sites, was found in a clone library originating from BNZ-LTER site UP2A, an upland mixed white spruce (*Picea glauca*) and birch (*Betula neoalaskana*) stand (Supplementary Information). This sequence not only originated from a different clone library that was constructed from a different DNA extract, it was amplified using primers slightly modified from those used to originally amplify NS1. We also reamplified the 290 bp fragment for NS1 from soil extracts from this site. It is therefore nearly impossible that this sequence arose from lab contamination by the original amplicon.

We were incapable of successfully amplifying and sequencing any portion of the LSU or SSU for NS1 using our original taxon-specific primers and any of the LSU or SSU primers we tested. We tested multiple new primers designed to be paired with a specific LSU or SSU primer and selectively amplify NS1. Even with these primers we could not reproduce portions of our original NS1 sequence longer than the 290 bp fragment or amplify regions of the LSU beyond the length of our original NS1 amplicon.

A nucleotide discontinuous MEGABLAST search found only one match for the ITS1 segment of NS1. This uncultured soil sequence overlapped the entire sequence and shared 83% sequence identity (GenBank GQ921811.1). When only the ITS2 segment was screened, no significant matches were found. The ITS1–5.8S–ITS2 segment shared the same top BLAST match with the ITS1 segment but produced many more results corresponding to partial matches for the 5.8S. The entire NS1 sequence had top hits corresponding to environmental sequences with as high as 53% sequence overlap (corresponding predominately to 5.8S and LSU) and maximum identities below 90% for these short-overlapping segments.

The 5.8S alignment consisted of 174 taxa and 169 characters, of which 138 were variable. The LSU alignment consisted of 67 taxa and 506 characters (after trimming with GBLOCKS), of which 338 were variable. For the 5.8S alignment, the TVM + G and GTR + G models were selected according to the AIC using ModelTest and MrModelTest, respectively. We chose GTRCAT for RAxML searches, since TVM is not available, and used GTR + G in MrBayes. GTR + G + Γ was selected as the best-fit model for the LSU alignment by both ModelTest and MrModelTest. GTRCAT was used in RAxML, due to the author's cautions against including an invariant sites parameter, while GTR + G + Γ was used in MrBayes.

Placement of NS1 was poorly resolved in both 5.8S and LSU trees. However, there was statistical support for placement of NS1 within the Eumycota for both loci under both inference

methods. The LSU trees, based upon many more characters, were better resolved than the 5.8S trees, and retrieved most deep groupings of Eumycota that are now considered well-supported (Hibbett et al., 2007; see clade demarcations in Fig. 2). In the LSU, both Bayesian and likelihood methods grouped NS1 with the Entomophthoromycotina, with 0.98 posterior probability (PP) but only 46% bootstrap support (BS). However, the NS1-containing clade was placed within the Eumycota with 84% BS and 1.0 PP (Fig. 2). In contrast, the poorly resolved 5.8S trees placed NS1 in a clade with the Kickxellomycotina plus *Piptocephus* (Zoopagomycotina), but without bootstrap support and only 0.91 posterior probability (Fig. S1). The 5.8S trees placed the NS1-containing clade within

the Eumycota plus Metazoa (Opisthokonta) with 98% BS and 1.0 PP using the Viridiplantae as outgroup, but did not recover a monophyletic Eumycota. Interestingly, two environmental DNA sequences recovered through nBLAST searches labeled ‘uncultured fungus’ were grouped with NS1 at 1.0 and 0.95 PP (86% and 57% BS) in the 5.8S trees (Fig. S2). However, since the provenances of these sequences cannot be verified, we do not consider these relationships to provide additional support for placing NS1 within the Eumycota.

Although NS1 is likely fungal and appears to fall among the BFL, the lack of support for more precise placements suggests the trees be treated with caution. The lack of support may be due to the

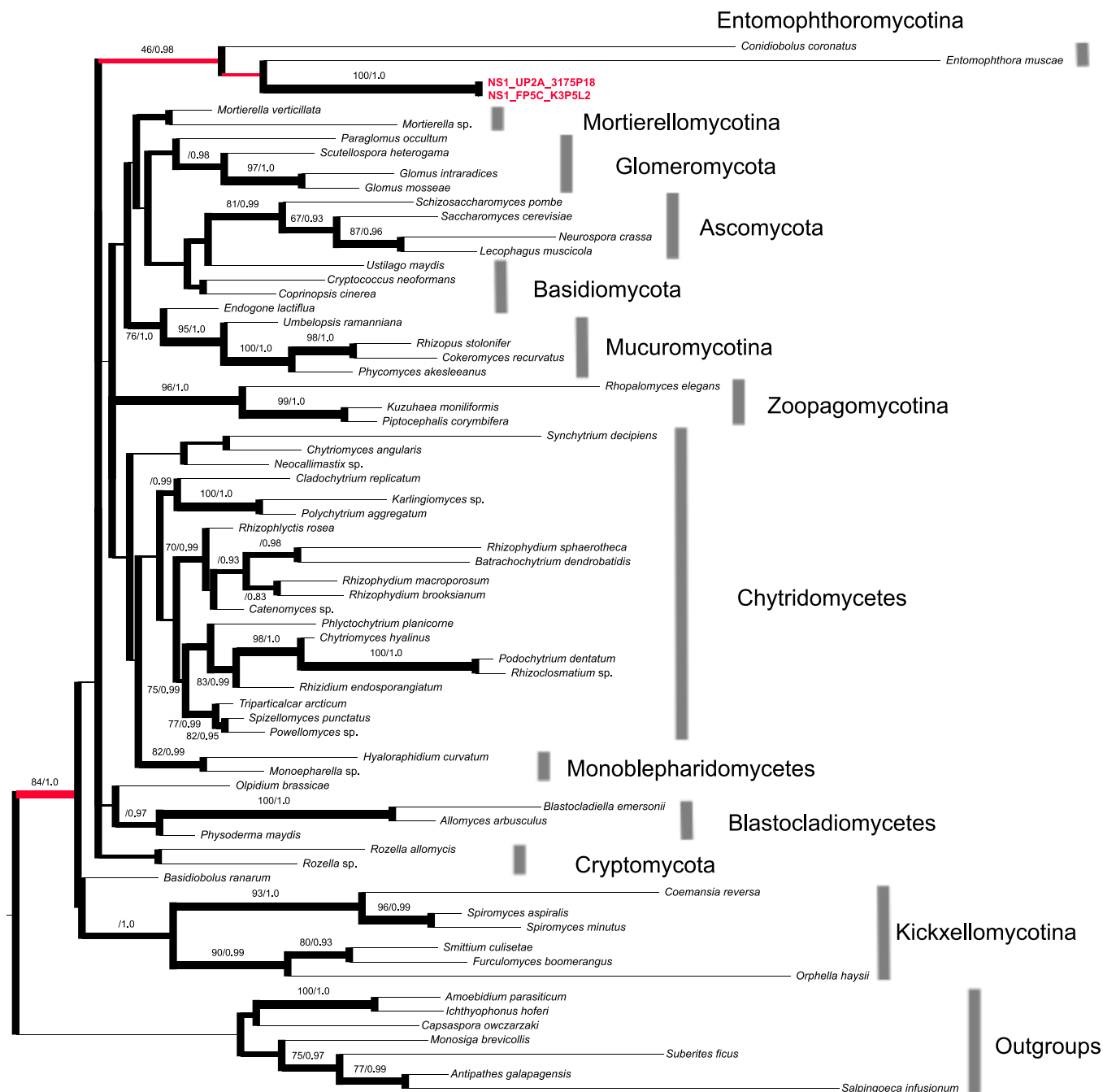


Fig. 2. Maximum likelihood phylogenetic reconstruction of the kingdom Eumycota based on partial LSU rDNA gene region. Branch thickness is proportional to bootstrap support. The first number above selected nodes is the maximum likelihood support from 1000 fast bootstrap replicates in RAxML; the number following the “/” is the posterior probability from 10,000,000 generations in MrBayes. Only bootstrap values about 70% and posterior probabilities above 80% are shown. When only a number after the / is given, the bootstrap support was below 70% while the posterior probability (shown) was above 80%.

considerable divergence between NS1 and other known sequences, which suggests that NS1 may represent a novel lineage.

Tajima's relative rates tests did not suggest that the 5.8S for NS1 was evolving at an accelerated rate relative to other basal fungi (Table S2).

3.2. Secondary structure and GC analyses of NS1

3.2.1. ITS1

ITS1 secondary structure has not been surveyed in as much detail as that of ITS2 (Mullineux and Hausner, 2009). Nonetheless, among the fungi studied, this locus is characterized by having a main central hairpin and may have smaller peripheral hairpins (Lalev and Nazar, 1998; Mullineux and Hausner, 2009). Even though this locus is relatively short in NS1 (90 bp), it possesses this main central hairpin (Fig. 3). When ITS1 length was plotted against MFE, we found NS1 to fall within the 95% prediction interval ($R^2 = 0.75$; Fig. S3). The GC content of the ITS1 for NS1 was above

the mean, but within one standard deviation of the mean as well ($n = 155$, mean = 0.46, s.d. = 0.12, NS1 = 0.51).

3.2.2. 5.8S

Similarities in 5.8S structure between the consensus folding and NS1 can be seen throughout the majority of the locus (Fig. 4). Areas that do differ between NS1 and the consensus folding appear to be more variable across fungi and are areas where NS1 displays similarities with sequences within the BFL.

Three regions within the 5.8S that have been proposed to be highly conserved across eukaryotes are motifs one, two, and three (Table 1; Harpke and Peterson, 2008). For NS1, motifs one and three are identical to both the sequence described in the literature and the consensus sequence. Motif two may be present in NS1, but the sequence does differ within this region (Table 1). However, this motif differs in fungi, particularly those belonging to the BFL, as well. Of 29 sites within the 5.8S where compensatory base changes (including GU pairs) appeared to occur in the majority of the fungal

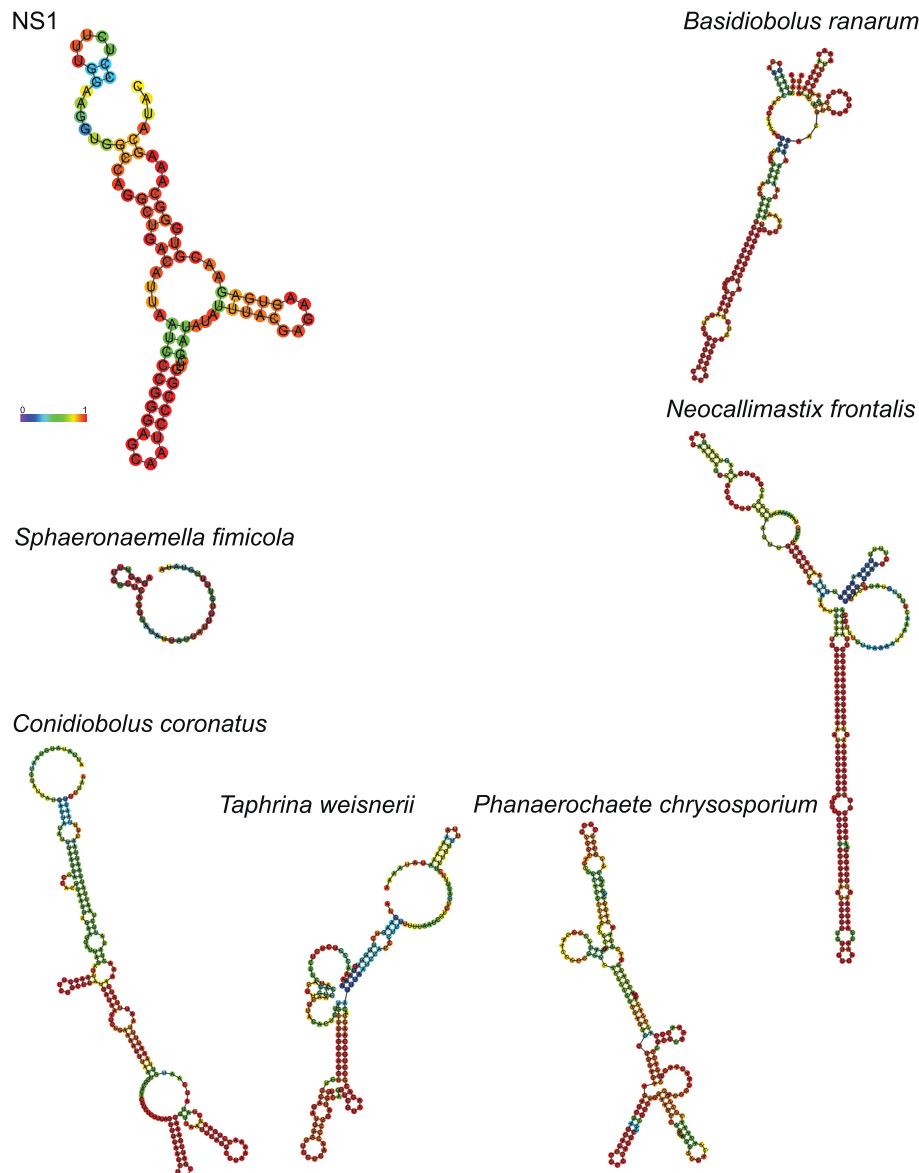


Fig. 3. ITS1 secondary structure for NS1 and representative taxa belonging to most of the described fungal phyla. Colors depict positional probabilities. The red end of the spectrum indicates that the majority of modeled foldings (not only the MFE folding shown) support such a pairing while the blue end of the spectrum indicates little support for the pairing shown.

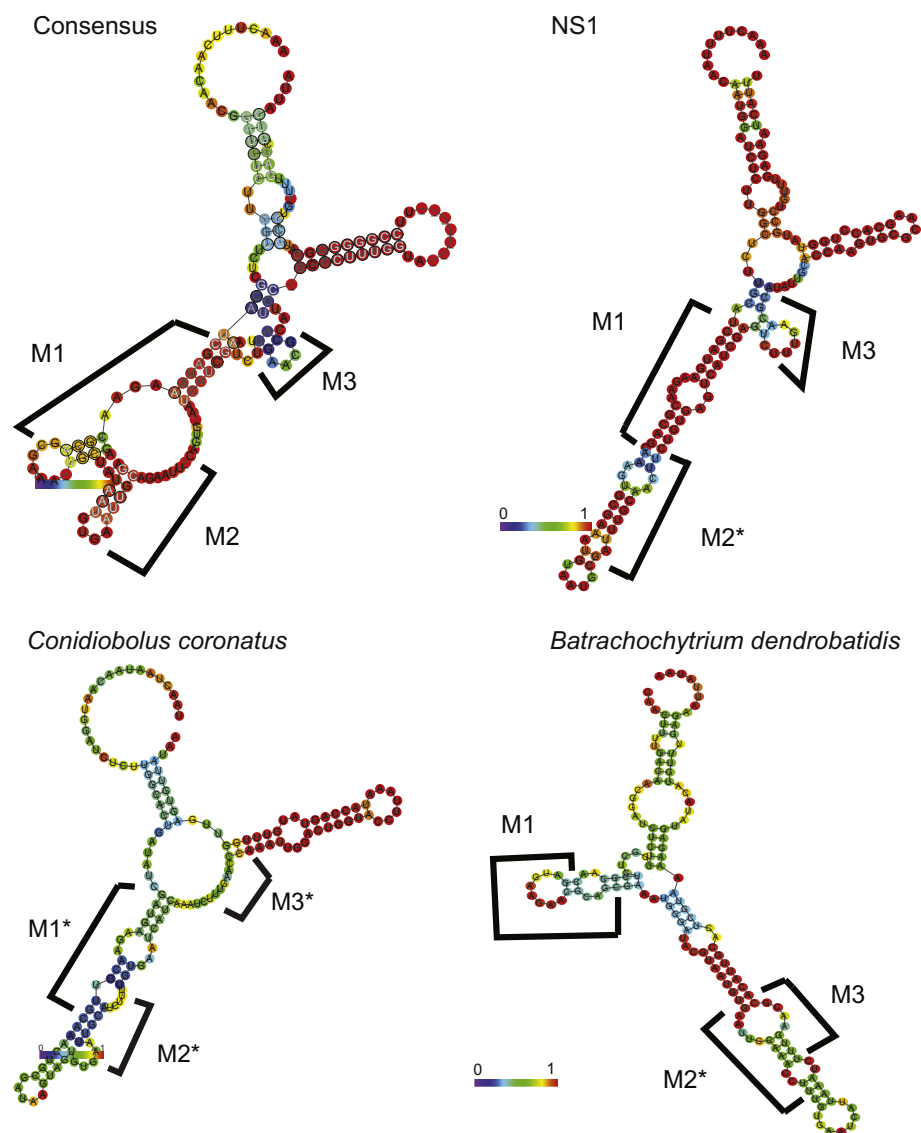


Fig. 4. 5.8S secondary structures for NS1, two comparative fungal taxa belonging to the BFL, and a consensus folding of 188 fungi. Motifs one, two, and three are abbreviated M1–M3. Asterisks indicate a variation in sequence from that described in literature.

	Motif 1	Motif 2	Motif 3
Consensus (literature)	CGATGAAGAACGCAGC	GAATTGCAGAAATTC	TTTGAACGCA
NS1	CGATGAAGAACGCAGC	GA TTT GC A -ACTTC	TTTGAACGCA
<i>Conidiobolus coronatus</i>	CGATGAAGAACG TT GC	GAATTGCAG TCTTT	TTTGAACCCA
<i>Batrachochytrium dendrobatidis</i>	CGATGAAGAACGCAGC	GAATTGCAGAA CCT	TTTGAACGCA

kingdom, we found NS1 to have 21 of these changes present (Table S3).

NS1 also fell within less than one-half a standard deviation of the mean MFE value of 5.8S foldings ($n = 184$, mean = -44.8 , s.d. = 4.3 , NS1 = -46.9) and has a GC content within one standard deviation ($n = 188$, mean = 0.456 , s.d. = 0.026 , NS1 = 0.43).

3.2.3. ITS2

An MFE folding of the NS1 ITS2 locus revealed the presence of structures conserved in most eukaryotes (Fig. 5). The core structure of this region is a central bulge with four hairpin loops radiating

from it. These are designated helices I–IV. Of these, helices II and III are the most conserved (Coleman, 2007). The ITS2 MFE folding for NS1 included helices II and III, but also had base pairings with low support within the central bulge region, and lacked the other two helices. Helix II is usually composed of fewer than 12 pairings, is never branched, and contains a U–U bulge at the base of the helix. NS1 did not have a U–U bulge at the base of helix II but did have this pair of bases present (Fig. 5). Like many other fungi, NS1 had a bulge caused by another mispairing, in this case the presence of an unpaired guanine (Table 2). Helix III is the longest and contains a distinctive motif on the 5' end just before the apex (Schultz et al.,

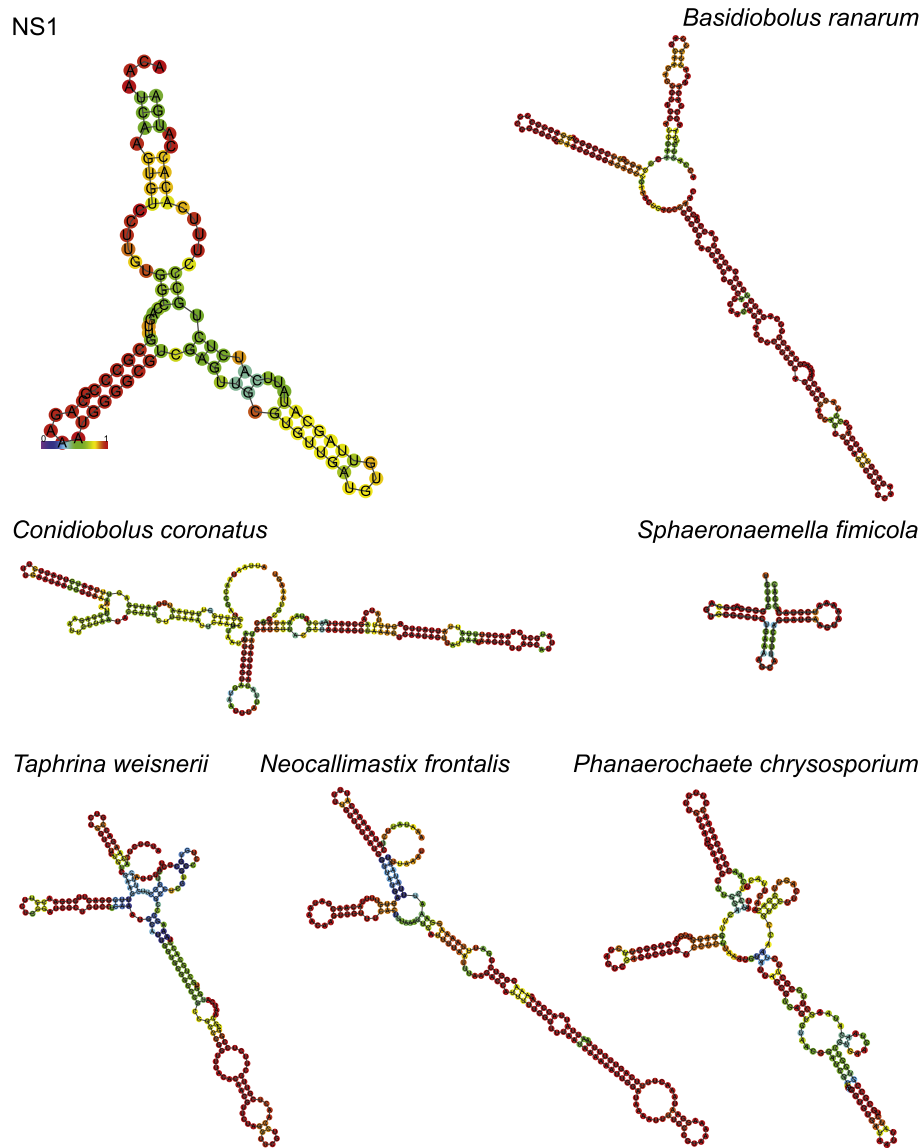


Fig. 5. ITS2 secondary structure for NS1 and representative taxa belonging to most of the described fungal phyla. Colors depict positional probabilities. The red end of the spectrum indicates that the majority of modeled foldings (not only the MFE folding shown) support such a pairing while the blue end of the spectrum indicates little support for the pairing shown.

Table 2

Comparison of highly conserved motifs within the ITS2 gene region. Fungi representing most of the major clades are compared to the consensus motif described in previous literature. Asterisks denote that the group given is undergoing taxonomic revision and is not a true phylum.

Taxon	Phylum	Hairpin II Bulge	Hairpin III 5' Motif
Consensus (literature)	NA	UU Mismatch	UGGU
NS1	NA	Unpaired G	UGAU
<i>Basidiobolus ranarum</i>	BFL*	Unpaired G	UGGU
<i>Conidiobolus coronatus</i>	BFL*	Unpaired U	AGUU
<i>Neocallimastix frontalis</i>	BFL*	UU Mismatch	UGGU
<i>Sphaeronaemella fimicola</i>	Ascomycota	Unpaired A	AGU
<i>Taphrina weisnerii</i>	Ascomycota	UU Mismatch	AGGU
<i>Phanaerochaete chrysosporium</i>	Basidiomycota	UU Mismatch	UGGU

2005; Wolf et al., 2005; Coleman, 2007). This is most often UGGU but can differ slightly; variations such as UGG, GGU, and UGGGU have been observed (Schultz et al., 2005). Helix III was present in NS1 and contained a motif similar, but not identical, to UGGU (Table 2, Fig. 5).

When ITS2 length was plotted against MFE, we found NS1 to fall within the 95% prediction interval ($R^2 = 0.51$; Fig. S4). The ITS2 GC content of NS1 was within one-half a standard deviation of the mean ($n = 155$, mean = 0.49, s.d. = 0.11, NS1 = 0.52).

NS1 had identical GC contents for ITS1 and ITS2 (51%). It also had very similar locus lengths (ITS1 = 90 bp; ITS2 = 97 bp).

3.2.4. LSU

The consensus folding for the LSU of all the reference fungi and NS1 had low probability support for many regions and is not shown. Foldings for two fungi belonging to the BFL are shown with NS1 for visual comparison (Fig. S2). Similarities in structure, especially near the 5' end, are noticeable. Many of these structures appear similar to those described for the B domain of the LSU (Ben Ali et al., 1999; De Rijk et al., 1999). We also note that conserved primary sequences among fungal LSU are present and easily aligned in NS1 (see TreeBASE 13807).

When compared to other fungal sequences for the conserved portion of the LSU we analyzed, NS1 had a predicted minimal free energy value for the folding that fell within one standard deviation

of the mean ($n = 182$, mean = 93.2, s.d. = 9.2, NS1 = 101.6). The GC content for this region of rRNA in NS1 fell within one standard deviation of the mean as well ($n = 175$, mean = 0.465, s.d. = 0.035, NS1 = 0.49).

4. Discussion

4.1. Amplification of longer segments

Phylogenetic and BLAST results suggested that NS1 is highly divergent from any previously described taxon and falls with weak support in the kingdom Eumycota. If NS1 represents a highly divergent lineage of fungi, the low level of sequence similarity to known fungi across the 5' LSU suggests that this sequence might represent a novel class or phylum. However, we were unable to amplify longer portions of the LSU or amplify the SSU for NS1. These additional characters might have allowed better supported phylogenetic placement of NS1.

Several factors may have contributed to our inability to reamplify DNA fragments for NS1 greater than 290 bp. The quality of DNA in the extracts could have diminished over time due to multiple freeze–thaw cycles. Due to NS1 being low in abundance (1 out of 384 clones in 2003 FP5C organic horizon clone library; 1 out of 931 clones in 2004 UP2A organic horizon clone library), there would be a high chance that PCR would be unsuccessful.

Additionally, when the primers we designed were paired with widely used SSU and LSU primers they appeared to be prone to dimerization. Since primer-dimer formation is a competitive reaction with template DNA amplification in PCR, this issue is more problematic when targeting templates present at low concentrations, such as NS1. Although PCR conditions were optimized to reduce this issue, it may have contributed to our inability to amplify NS1 adequately to successfully clone or directly sequence these amplicons. Only a limited number of potential primers could be designed that were NS1-specific. Due to the small size of the more variable regions (ITS1 and ITS2) available as potential priming sites, an ideal primer could not be found.

4.2. Evidence suggesting that NS1 is an authentic rRNA gene copy

Although the evidence remains somewhat ambiguous, our results nonetheless support the view that NS1 is a functional gene copy and thus truly represents a novel fungal lineage rather than a biological or lab artifact.

If NS1 is of chimeric origin, our secondary structure analyses would likely have provided an indication. Due to the nature of chimera formation and the necessity that certain structures be maintained for proper RNA processing and function, had NS1 been a chimera it would most likely have differed from other fungi in secondary structure far more than we detected. Furthermore, none of the chimera tests we conducted suggest that NS1 is chimeric. We also found two full-length (1200 bp) NS1 sequences that were nearly identical in separate clone libraries that were constructed using different source DNAs and were amplified with different PCR primers. Finally, we were able to amplify a 290 bp region of NS1 from multiple different soil extracts. These results collectively suggest that NS1 is not simply a PCR artifact. The odds of creating the same chimera multiple times from different samples are decidedly low. Furthermore, NS1 appears to be particularly divergent from other fungi throughout the entire 1200 bp sequence, particularly the ITS regions. This also suggests that it is not a chimera. A chimera would be expected to be highly similar to other fungi throughout the portions of its sequence from which the templates originated.

The overall structure for all four loci was consistent with known (functional) sequences. Therefore, NS1 is not likely to be an ancient pseudogene. ITS1 secondary structure was similar to descriptions for *Schizosaccharomyces pombe* (Fig. 3; Lalev and Nazar, 1998) and several other ascomycetes (Bridge et al., 2008; Hausner and Wang, 2005; Mullineux and Hausner, 2009). It should be noted that all the fungal species that have had their ITS1 structure described are ascomycetes and that no basal fungi have had their ITS1 secondary structure modeled previously.

The overall structure of 5.8S for NS1 was similar to both that described in the literature and the 5.8S secondary structure we produced using the fungal consensus sequence (Fig. 4; Vaughn and Sperbeck, 1984). NS1 also contained the two most conserved hairpins within ITS2 (Fig. 5). It appears that these hairpins are also the only ones present in ascomycetes with particularly short ITS2 (Hausner and Wang, 2005). Overall, these results suggest that NS1 has probably retained the most important regions of ITS2 for proper RNA processing.

Along with the observation that the overall secondary structures for all four loci were similar to previously described secondary structures, we also found that two of the most conserved domains within the 5.8S, motifs I and III, were present in NS1. These domains play an important role in RNA processing and if absent or altered would strongly suggest a pseudogene (Table 1; Harpke and Peterson, 2008). Motif II in NS1 did differ from the sequence described by Harpke and Peterson (2008), but because it differed in other fungi as well, we believe that constraints on the primary sequence may be less strict than previously suggested (Table 1).

Additionally, 5.8S pseudogenes have been shown to have higher rates of sequence divergence than their functional counterparts (Harpke and Peterson, 2008). However, NS1 did not demonstrate accelerated sequence divergence rates relative to any of the potential sister taxa tested. This is consistent with secondary structure analyses suggesting that NS1 is unlikely to be a pseudogene. On the other hand, this finding must be viewed with caution, since no closely-related taxa are available for comparison, which may diminish the sensitivity of the relative rate test.

Comparing minimal free energies (MFE) can lend additional support to analyses of secondary structure (Harpke and Peterson, 2006; Zheng et al., 2008). It quickly demonstrates if the sequence in question falls within the known range of fungal rRNA MFE values. We plotted both ITS1 and ITS2 length against MFE to determine if MFEs scale with length and found a significant correlation consistent with patterns described in other taxa (Supplementary Information; Zheng et al., 2008). NS1 had MFEs consistent with the overall trend for structures of that length for both ITS1 and ITS2. This complements our findings from analyzing secondary structure and suggests that the most important structures were present, and the conformations for ITS1 and ITS2 were not due to chance.

Although ITS1 and ITS2 originated independently, they are believed to undergo concerted evolution (Lalev and Nazar, 1999; Hausner and Wang, 2005; Mullineux and Hausner, 2009) and the retention of similar GC content is a vestige of this process (Torres et al., 1990). Thus, pseudogenes or chimeras, neither of which are believed to undergo concerted evolution, might be expected to have divergent GC contents between ITS1 and ITS2, whereas functional genes should have similar GC contents because of selective constraints. Torres et al.'s (1990) study of ITS GC content only included 20 sequences representing the entire eukaryotic domain. Although the study did include fungi, only the Dikarya were represented. Our findings support the results of this study for most of the species tested, as have many other studies (Harpke and Peterson, 2006; Mullineux and Hausner, 2009; Xiao et al., 2010; Zheng et al., 2008). NS1 was particularly consistent with this rule, having identical GC content for both loci.

4.3. Evidence calling into question the authenticity of NS1

The secondary structures and nucleotide motifs that were inconsistent with universal patterns described in the literature appeared to be exceptions in other organisms as well (Schultz et al., 2005). As already mentioned, within the 5.8S region, the nucleotide sequence within motif II did differ slightly from the universal sequence for eukaryotes, but other basal fungal sequences used for comparison differed too (Table 1). The highly conserved motifs I and III were present in NS1, while the basal fungi used for comparison showed variation within these regions (Table 1). These three motifs, particularly motif II, may not be as universally diagnostic for pseudogenes as suggested by Harpke and Peterson (2008), at least for basal fungi.

NS1 displayed variation in conserved domains within ITS2 as well. The minimal free energy folding of ITS2 for NS1 did not include the UU bulge at the base of hairpin II that is widespread in eukaryotes, but did include the two uracil bases, suggesting that the minimal free energy folding model did not accurately depict the *in vivo* conformation of this portion of RNA. Alternatively, NS1 as well as many fungi appeared to have a bulge on helix II, but this was not always caused by unpaired uracils (Table 2). Perhaps the presence of a bulge is necessary for proper processing, but the origin of the bulge is under less selective constraint.

NS1 does appear to have hairpin III, but the highly conserved UGGU motif near the 5' apex of the loop differed slightly (UGAU). Schultz et al. (2005) state that this motif differs slightly in some eukaryotes, although a UGAU variant was not mentioned. Because two of the regions most conserved across the entire eukaryotic kingdom did vary in our sequence, the authenticity of NS1 is not certain. However, these regions are known to vary in other organisms as well (Coleman, 2007) and are frequently absent or altered in basal fungi. We also found variation within these regions in several of our reference sequences (Table 2). Therefore, they may not be as conserved in fungi as in other organisms.

Overall, the 5.8S and ITS2 regions were the most informative for comparing secondary structures and determining the presence of conserved regions. The secondary structure evidence that did not support NS1 representing functional rDNA also appeared to show variation in other fungal taxa. Based on the weight of the evidence, we conclude that NS1 is most likely a functional gene copy. NS1 may well be a recently derived pseudogene, but that would not seem to fully explain its divergence from all known fungal lineages.

4.4. Phylogenetic affinities

We set out to not only determine the secondary structure of NS1, but also to determine if it was fungal in origin. Phylogenetic analyses based on alignable components of the 5.8S and LSU agree in placing NS1 within the Eumycota with some statistical support, but disagree with respect to the containing clade, even at the level of phylum. Hopefully, longer segments representing the NS1 lineage will eventually be recovered, either from metagenomic DNA or from a known organism (e.g. a currently unsequenced member of the BFL). Longer rDNA sequences and/or additional loci will clearly be required in order to confidently place this lineage within a phylum or lower taxon.

4.5. Broader implications and future research

As microbial community ecology studies shift towards next-generation sequencing methods such as pyrosequencing, selecting a maximally informative region to amplify becomes more pressing, because less information is retained in shorter sequences (Nilsson et al., 2009). If a locus will be used in phylogenetic reconstructions spanning the fungal kingdom, it is critical that the region targeted meet two criteria: (1) It should be useful for distinguishing

between closely related species while also resolving deeper level relationships (Nilsson et al., 2009); and (2) it should be useful for determining whether the sequence represents an authentic, potentially functional gene copy (Mai and Coleman, 1997; Schultz et al., 2005). When either ITS1 or ITS2 is amplified with the flanking, highly-conserved 5.8S region, both ITS1 and ITS2 meet this first requirement, but ITS1 is less well suited to the latter.

ITS1 has widely been proposed as the ideal fungal marker because it varies more than ITS2 and is therefore better for distinguishing between closely related species (Chen et al., 2001; Hinrikson et al., 2005). One of the most often utilized forward primers in fungal ITS1 amplification is ITS1-F (Gardes and Bruns, 1993). A major issue with targeting ITS1 is the presence of an intron at the 3' end of the SSU in many fungi (Vralstad et al., 2002; Perotto et al., 2000). This intron is often amplified when the ITS1-F primer is used (Vralstad et al., 2002), which may push the amplicon length beyond the reach of next-generation sequencing methods. PCR length-biases towards species with shorter ITS segments have been shown when using the ITS1-F primer as well (Ihrmark et al., 2012). ITS2 shows moderate variation and can therefore still distinguish between species nearly as well as, and sometimes better than, ITS1 (Nilsson et al., 2008). However, ITS2 is better suited to evaluation of whether rRNA secondary structure is maintained (Schultz et al., 2005; Coleman, 2007). Not only are the conserved domains within ITS2 well described, there is a database containing these structures for thousands of organisms that can be used for comparison (Koetschan et al., 2010). It is therefore logical to target this region preferentially over ITS1, even though it shows slightly less variation.

All of these analyses benefit from having a broad representation of well-identified fungi to be used for comparison. The basal fungal lineages have much poorer representation within the INSD than do the Dikarya. At the time of writing, over 184,000 Dikarya sequences containing the 5.8S locus had been deposited, while the BFLs were represented by fewer than 5000. Of these roughly 5000 sequences, approximately 2000 belonged to the Mucoromycotina, while the Entomophthoromycotina and Kickxellomycotina were both represented by fewer than 100 nucleotide sequences (see <http://www.ncbi.nlm.nih.gov/taxonomy/txstat.cgi>).

If most known fungal species were represented and multiple sequences per species were available, it would be possible to determine nucleotide diversity (π) and identify nucleotide motifs conserved at the infraordinal level. These two methods, particularly the latter, have been shown to be the best way to identify pseudogenes in closely related species (Harpke and Peterson, 2006; Zheng et al., 2008). Because close relatives to be used as reference sequences are needed for these methods to perform well, they were not applicable to our study at this time.

5. Conclusions

Our phylogenetic results suggest that NS1 is fungal in origin. More precise phylogenetic placements might be made if longer segments could be amplified, as in Porter et al. (2008). We modeled portions of the NS1 sequence to characterize ribosomal RNA secondary structure. Secondary structure analyses suggest that it represents an authentic gene copy. MFE and GC analyses further support the conclusion that it is not an artifact. Hence, NS1 may well be a pseudogene, but that would not seem to fully explain its divergence from all known fungal lineages. If shorter rRNA segments are going to be targeted in the future, as next-generation sequencing becomes more widespread, our findings suggest prioritizing the 5.8S and ITS2. These regions are not only suited to distinguishing between closely related species, but they are also more informative regarding the authenticity of a sequence.

Acknowledgments

We would like to thank Stefan Washietl for his valuable input on the applicability of different programs on the RNA Websuite to our research questions. We would also like to thank Tim James for supplying reference sequences and advice on constructing phylogenies of the basal fungal lineages. The entire Taylor lab provided input and assistance throughout this research project and deserves thanks as well. This material is based on work supported by the National Science Foundation through awards EF-0333308 and DEB-0620579, Alaska EPSCoR NSF award EPS-0701898 and the State of Alaska, and by the USDA Forest Service, Pacific Northwest Research Station through agreement number RJVA-PNW-01-JV-11261952-231.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ympev.2013.01.018>.

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., 1999. Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410.
- Anthony, N.M., Clifford, S.L., Bawe-Johnson, M., Abernethy, K.A., Bruford, M.W., Wickings, E.J., 2007. Distinguishing gorilla mitochondrial sequences from nuclear integrations and PCR recombinants: Guidelines for their diagnosis in complex sequence databases. *Mol. Phylogenet. Evol.* 43, 553–566.
- Balakirev, E.S., Ayala, F.J., 2003. Pseudogenes: Are they “junk” or functional DNA? *Annu. Rev. Genet.* 37, 123–151.
- Baldauf, S.L., Palmer, J.D., 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. USA* 90, 11558.
- Ben Ali, A., Wuyts, J., Wachter, R.D., Meyer, A., Van de Peer, Y., 1999. Construction of a variability map for eukaryotic large subunit ribosomal RNA. *Nucl. Acids Res.* 27, 2825–2831.
- Bensasson, D., Zhang, D., Hartl, D.L., Hewitt, G.M., 2001. Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol. Evol.* 16, 314–321.
- Bernhart, S.H., Hofacker, I.L., Will, S., Gruber, A.R., Stadler, P.F., 2008. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinform.* 9, 13.
- Bridge, P.D., Roberts, P.J., Spooner, B.M., Panchal, G., 2003. On the unreliability of published DNA sequences. *New Phytol.* 160, 43–48.
- Bridge, P.D., Schlitt, T., Cannon, P.F., Buddie, A.G., Baker, M., Borman, A.M., 2008. Domain II hairpin structure in ITS1 sequences as an aid in differentiating recently evolved animal and plant pathogenic fungi. *Mycopathologia* 166, 1–16.
- Bruns T.D., Vilgaly, R., Barns S.M., Gonzalez D.S., Lane D.J., Simon L., Stickle S., Szaro T.M., Weisburg W.G., et al., 1992. Evolutionary relationships within the fungi: analyses of nuclear small subunit rRNA sequences. *Mol. Phylogenet. Evol.* 1, 231–241.
- Castresana, J., 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.* 17, 540–552.
- Chen, Y.C., Eisner, J.D., Kattar, M.M., Rassoul-Barrett, S.L., Lefe, K., Bui, U., Limaye, A.P., Cookson, B.T., 2001. Polymorphic internal transcribed spacer region 1 DNA sequences identify medically important yeasts. *J. Clin. Microbiol.* 39, 4042–4051.
- Coleman, A.W., 2007. Pan-eukaryote ITS2 homologues revealed by RNA secondary structure. *Nucl. Acids Res.* 35, 3322–3329.
- Cullings, K.W., Vogler, D.R., 1998. A 5.8 S nuclear ribosomal RNA gene sequence database: applications to ecology and evolution. *Mol. Ecol.* 7, 919–923.
- De Rijk, P., Robbrecht, E., de Hoog, S., Caers, A., Van de Peer, Y., De Wachter, R., 1999. Database on the structure of large subunit ribosomal RNA. *Nucl. Acids Res.* 27, 174–178.
- Edgar, R.C., Haas, B.J., Clemente, J.C., Quince, C., Knight, R., 2011. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27, 2194–2200.
- Fujita, S.I., Senda, Y., Nakaguchi, S., Hashimoto, T., 2001. Multiplex PCR using internal transcribed spacer 1 and 2 regions for rapid detection and identification of yeast strains. *J. Clin. Microbiol.* 39, 3617–3622.
- Gardes, M., Bruns, T.D., 1993. ITS primers with enhanced specificity for basidiomycetes – application to the identification of mycorrhizae and rusts. *Mol. Ecol.* 2, 113–118.
- Geml, J., Laursen, G.A., Timling, I., McFarland, J.M., Booth, M.G., Lennon, N., Nusbaum, C., Taylor, D.L., 2009. Molecular phylogenetic biodiversity assessment of arctic and boreal ectomycorrhizal *Lactarius* Pers. (Russulales; Basidiomycota) in Alaska, based on soil and sporocarp DNA. *Mol. Ecol.* 18, 2213–2227.
- Geml, J., Laursen, G.A., Herriott, I.C., McFarland, J.M., Booth, M.G., Lennon, N., Nusbaum, H.C., Taylor, D.L., 2010. Phylogenetic and ecological analyses of soil and sporocarp DNA sequences reveal high diversity and strong habitat partitioning in the boreal ectomycorrhizal genus *Russula* (Russulales; Basidiomycota). *New Phytol.* 187, 494–507.
- Gruber, A.R., Lorenz, R., Bernhart, S.H., Neubock, R., Hofacker, I.L., 2008. The Vienna RNA Websuite. *Nucl. Acids Res.* 36, W70–W74.
- Harpke, D., Peterson, A., 2006. Non-concerted ITS evolution in *Mammillaria* (Cactaceae). *Mol. Phylogenet. Evol.* 41, 579–593.
- Harpke, D., Peterson, A., 2008. 5.8S motifs for the identification of pseudogenic ITS regions. *Botany* 86, 300–305.
- Hausner, G., Wang, X., 2005. Unusual compact rDNA gene arrangements within some members of the Ascomycota: evidence for molecular co-evolution between ITS1 and ITS2. *Genome* 48, 648–660.
- Hibbett, D.S., Binder, M., Bischoff, J.F., Blackwell, M., Cannon, P.F., Eriksson, O.E., Huhndorf, S., James, T., Kirk, P.M., Lücking, R., et al., 2007. A higher-level phylogenetic classification of the Fungi. *Mycol. Res.* 111, 509–547.
- Hibbett, D.S., Ohman, A., Grotzer, D., Nuhn, M., Kirk, P., Nilsson, R.H., 2011. Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biol. Rev.* 25, 38–47.
- Hinrikson, H.P., Hurst, S.F., Lott, T.J., Warnock, D.W., Morrisson, C.J., 2005. Assessment of ribosomal large-subunit D1–D2, internal transcribed spacer 1, and internal transcribed spacer 2 regions as targets for molecular identification of medically important *Aspergillus* species. *J. Clin. Microbiol.* 43, 2092–2103.
- Hofacker, I.L., Fekete, M., Stadler, P.F., 2002. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319, 1059–1066.
- Hofacker, I.L., 2003. Vienna RNA secondary structure server. *Nucl. Acids Res.* 31, 3429–3431.
- Hugenholtz, P., Huber, T., 2003. Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Syst. Evol. Microbiol.* 53, 289–293.
- Ihrmark, K., Bödeker, I.T.M., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., Strid, Y., Stenlid, J., Brandström-Durling, M., Clemmensen, K.E., Lindahl, B.D., 2012. New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol. Ecol.* 82, 666–677.
- James, T.Y., Kauff, F., Schoch, C.L., Matheny, P.B., Hofstetter, V., Cox, C.J., Celio, G., Gueidan, C., Fraker, E., Miadlikowska, J., Lumbsch, H.T., Rauhut, A., Reeb, V., Arnold, A.E., Amtoft, A., Stajich, J.E., Hosaka, K., Sung, G., Johnson, D., O'Rourke, B., Crockett, M., Binder, M., Curtis, J.M., Slot, J.C., Wang, Z., Wilson, A.W., Schußler, A., Longcore, J.E., O'Donnell, K., Mozley-Standridge, S., Porter, D., Letcher, P.M., Powell, M.J., Taylor, J.W., White, M.M., Griffith, G.W., Davies, D.R., Humber, R.A., Morton, J.B., Sugiyama, J., Rossman, A.Y., Rogers, J.D., Pfister, D.H., Hewitt, D., Hansen, K., Hambleton, S., Shoemaker, R.A., Kohlmeyer, J., Volkmann-Kohlmeyer, B., Spotts, R.A., Serdani, M., Crous, P.W., Hughes, K.W., Matsuura, K., Langer, E., Langer, G., Untereiner, W.A., Lücking, R., Budel, B., Geiser, D.M., Aptroot, A., Diederich, P., Schmitt, I., Schultze, M., Yahr, R., Hibbett, D.S., Lutzoni, F., McLaughlin, D.J., Spatafora, J.W., Vilgaly, R., 2006. Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* 443, 818–822.
- Jones, M.D., Forn, I., Gadelha, C., Egan, M.J., Bass, D., Massana, R., Richards, T.A., 2011. Discovery of novel intermediate forms redefines the fungal tree of life. *Nature* 474, 200–203.
- Jumpponen, A., 2007. Soil fungal communities underneath willow canopies on a primary successional glacier forefront: rDNA sequence results can be affected by primer selection and chimeric data. *Microb. Ecol.* 53, 233–246.
- Koetschan, C., Förster, F., Keller, A., Schleicher, T., Ruderisch, B., Schwarz, R., Müller, T., Wolf, M., Schultze, J., 2010. The ITS2 Database III-sequences and structures for phylogeny. *Nucl. Acids Res.* 38, D275–D279.
- Koressaar, T., Remm, M., 2007. Enhancements and modifications of primer design program Primer3. *Bioinformatics* 23, 1289–1291.
- Krell, F.T., Cranston, P.S., 2004. Which side of the tree is more basal? *Syst. Entom.* 29, 279–281.
- Lalev, A.I., Nazar, R.N., 1998. Conserved core structure in the internal transcribed spacer 1 of the *Schizosaccharomyces pombe* precursor ribosomal RNA. *J. Mol. Biol.* 284, 1341–1351.
- Lalev, A.I., Nazar, R.N., 1999. Structural equivalence in the transcribed spacers of pre-rRNA transcripts in *Schizosaccharomyces pombe*. *Nucl. Acids Res.* 27, 3071–3078.
- Lara, E., Moreira, D., López-García, P., 2010. The environmental clade LKM11 and *Rozella* form the deepest branching clade of Fungi. *Protist* 161, 116–121.
- Lutzoni, F., Kauff, F., Cox, C.J., McLaughlin, D., Celio, G., Dentinger, B., Padamsee, M., Hibbett, D., James, T.Y., Baloch, E., Grube, M., Reeb, V., Hofstetter, V., Schoch, C., Arnold, A.E., Miadlikowska, J., Spatafora, J., Johnson, D., Hambleton, S., Crockett, M., Shoemaker, R., Sung, G., Lücking, R., Lumbsch, T., O'Donnell, K., Binder, M., Diederich, P., Ertz, D., Gueidan, C., Hansen, K., Harris, R.C., Hosaka, K., Lim, Y., Matheny, B., Nishida, H., Pfister, D., Rogers, J., Rossman, A., Schmitt, I., Sipman, H., Stone, J., Sugiyama, J., Yahr, R., Vilgaly, R., 2004. Assembling the fungal tree of life: Progress, classification and evolution of subcellular traits. *Am. J. Bot.* 91, 1446–1480.
- Maddison, W.P., Maddison, D.R., 2008. Mesquite: a modular system for evolutionary analysis. *Evolution* 62, 1103–1118.
- Mai, J.C., Coleman, A.W., 1997. The internal transcribed spacer 2 exhibits a common secondary structure in green algae and flowering plants. *J. Mol. Evol.* 44, 258–271.
- Mullineux, T., Hausner, G., 2009. Evolution of rDNA ITS1 and ITS2 sequences and RNA secondary structures within members of the fungal genera *Grossmannia* and *Leptographium*. *Fungal Genet. Biol.* 46, 855–867.
- Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N., Larsson, K., 2008. Intraspecific ITS variability in the kingdom Fungi as expressed in the

- international sequence databases and its implications for molecular species identification. *Evol. Bioinform. Online* 4, 193–201.
- Nilsson, R.H., Ryberg, M., Abernethy, K., Sjökvist, E., Kristiansson, E., 2009. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *FEMS Microbiol. Lett.* 296, 97–101.
- Notredame C., Higgins D.G., Heringa J., others, 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* 302, 205–218.
- Nylander, J., 2004. MrModeltest, Version 2. Uppsala University, Uppsala, Sweden, Evolutionary Biology Centre.
- O'Donnell, K., Lutzoni, F.M., Ward, T.J., Benny, G.L., 2001. Evolutionary relationships among mucoralean fungi (Zygomycota): evidence for family polyphyly on a large scale. *Mycologia* 286–297.
- Olson, L.E., Hassanin, A., 2002. Contamination and chimerism are perpetuating the legend of the snake-eating cow with twisted horns (*Pseudonovibos spiralis*). A case study of the pitfalls of ancient DNA. *Mol. Phylogenet. Evol.* 27, 545–548.
- Olson, L.E., Yoder, A.D., 2002. Using secondary structure to identify ribosomal numts: Cautionary examples from the human genome. *Mol. Biol. Evol.* 19, 93–100.
- Osborne, C.A., Galic, M., Sangwan, P., Janssen, P.H., 2005. PCR-generated artefact from 16S rRNA gene-specific primers. *FEMS Microbiol. Lett.* 248, 183–187.
- Perna, N.T., Kocher, T.D., 1996. Mitochondrial DNA: Molecular fossils in the nucleus. *Curr. Biol.* 6, 128–129.
- Perotto, S., Nepote-Fus, P., Saletta, L., Bandi, C., Young, J.P.W., 2000. A diverse population of introns in the nuclear ribosomal genes of ericoid mycorrhizal fungi includes elements with sequence similarity to endonuclease-coding genes. *Mol. Biol. Evol.* 17, 44–59.
- Porter, T.M., Schadt, C.W., Rizvi, L., Martin, A.P., Schmidt, S.K., Scott-Denton, L., Vilgalys, R., Moncalvo, J.M., 2008. Widespread occurrence and phylogenetic placement of a soil clone group adds a prominent new branch to the fungal tree of life. *Mol. Phylogenet. Evol.* 46, 635–644.
- Posada, D., Crandall, K.A., 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- R Development Core Team, 2008. R: A language and environment for statistical computing. R Foundation Statistical Computing.
- Rice, P., Longden, I., Bleasby, A., others, 2000. EMBOSS: the European molecular biology open software suite. *Trends in genetics* 16, 276–277.
- Richly, E., Leister, D., 2004. NUMTs in sequenced eukaryotic genomes. *Mol. Biol. Evol.* 21, 1081–1084.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Ryberg, M., Kristiansson, E., Sjökvist, E., Nilsson, R.H., 2009. An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytol.* 181, 471–477.
- Sanderson, M., Donoghue, M., Piel, W., Eriksson, T., 1994. TreeBASE: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *Am. J. Bot.* 81, 183.
- Schadt, C.W., Martin, A.P., Lipson, D.A., Schmidt, S.K., 2003. Seasonal dynamics of previously unknown fungal lineages in tundra soils. *Science* 301, 1359–1361.
- Schultz, J., Maisel, S., Gerlach, D., Muller, T., Wolf, M., 2005. A common core of secondary structure of the internal transcribed spacer 2 (ITS2) throughout the Eukaryota. *RNA* 11, 361–364.
- Schüßler, A., Schwarzott, D., Walker, C., 2001. A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycol. Res.* 105, 1413–1421.
- Stamatakis, A., 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22, 2688–2690.
- Steenkamp, E.T., Wright, J., Baldauf, S.L., 2006. The protistan origins of animals and fungi. *Mol. Biol. Evol.* 23, 93–106.
- Stöver, B.C., Müller, K.F., 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinform.* 11, 7.
- Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., Kumar, S., 2011. MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol. Biol. Evol.* 28, 2731–2739.
- Tanabe, Y., Watanabe, M.M., Sugiyama, J., 2005. Evolutionary relationships among basal fungi (Chytridiomycota and Zygomycota): Insights from molecular phylogenetics. *J. Gen. Appl. Microbiol.* 51, 267–276.
- Taylor, D.L., Bruns, T.D., 1999. Community structure of ectomycorrhizal fungi in a *Pinus muricata* forest: minimal overlap between the mature forest and resistant propagule communities. *Mol. Ecol.* 8, 1837–1850.
- Taylor, D.L., Herriott, I.C., Long, J., O'Neill, K., 2007. TOPO TA is A-OK: a test of phylogenetic bias in fungal environmental clone library construction. *Environ. Microbiol.* 9, 1329–1334.
- Taylor, D.L., Booth, M.G., McFarland, J.W., Herriott, I.C., Lennon, N.J., Nusbaum, C., Marr, T.G., 2008. Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach. *Mol. Ecol. Resour.* 8, 742–752.
- Taylor, D.L., Herriott, I.C., Stone, K.E., McFarland, J.W., Booth, M.G., Leigh, M.B., 2010. Structure and resilience of fungal communities in Alaskan boreal forest soils. *Can. J. For. Res.* 40, 1288–1301.
- Torres, R.A., Ganai, M., Hemleben, V., 1990. GC balance in the internal transcribed spacers ITS-1 and ITS-2 of nuclear ribosomal-RNA genes. *J. Mol. Evol.* 30, 170–181.
- Vaughn, J.C., Sperbeck, S.J., 1984. A universal model for the secondary structure of 5.8S ribosomal-RNA molecules, their contact sites with 28S ribosomal-RNAs, and their prokaryotic equivalent. *Nucl. Acids Res.* 12, 7479–7502.
- Voigt, K., Kirk, P.M., 2011. Recent developments in the taxonomic affiliation and phylogenetic positioning of fungi: impact in applied microbiology and environmental biotechnology. *Applied microbiology and biotechnology* 90, 41–57.
- Vralstad, T., Myhre, E., Schumacher, T., 2002. Molecular diversity and phylogenetic affinities of symbiotic root-associated ascomycetes of the Helotiales in burnt and metal polluted habitats. *New Phytol.* 155, 131–148.
- White, M.M., James, T.Y., O'Donnell, K., Cafaro, M.J., Tanabe, Y., Sugiyama, J., 2006. Phylogeny of the Zygomycota based on nuclear ribosomal sequence data. *Mycologia* 98, 872–884.
- Wolf, M., Achtziger, M., Schultz, J., Dandeker, T., Muller, T., 2005. Homology modeling revealed more than 20,000 rRNA internal transcribed spacer 2 (ITS2) secondary structures. *RNA* 11, 1616–1623.
- Xiao, L.Q., Moller, M., Zhu, H., 2010. High nrDNA ITS polymorphism in the ancient extant seed plant *Cycas*: Incomplete concerted evolution and the origin of pseudogenes. *Mol. Phylogenet. Evol.* 55, 168–177.
- Zheng, X.Y., Cai, D.Y., Yao, L., Teng, Y., 2008. Non-concerted ITS evolution, early origin and phylogenetic utility of ITS pseudogenes in *Pyrus*. *Mol. Phylogenet. Evol.* 48, 892–903.