

TECHNICAL ADVANCES

Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach

D. LEE TAYLOR,* MICHAEL G. BOOTH,* JACK W. MCFARLAND,* IAN C. HERRIOTT,* NIALL J. LENNON,† CHAD NUSBAUM† and THOMAS G. MARR*‡

*Institute of Arctic Biology, 311 Irving I Building, University of Alaska, Fairbanks, AK 99775, USA, †Broad Institute, Massachusetts Institute of Technology, 320 Charles Street, Cambridge, MA 02141, USA

Abstract

High throughput sequencing methods are widely used in analyses of microbial diversity, but are generally applied to small numbers of samples, which precludes characterization of patterns of microbial diversity across space and time. We have designed a primer-tagging approach that allows pooling and subsequent sorting of numerous samples, which is directed to amplification of a region spanning the nuclear ribosomal internal transcribed spacers and partial large subunit from fungi in environmental samples. To test the method for phylogenetic biases, we constructed a controlled mixture of four taxa representing the Chytridiomycota, Zygomycota, Ascomycota and Basidiomycota. Following cloning and colony restriction fragment length polymorphism analysis, we found no significant difference in representation in 19 of the 23 tested primers. We also generated a clone library from two soil DNA extracts using two primers for each extract and compared 456 clone sequences. Community diversity statistics and contingency table tests applied to counts of operational taxonomic units revealed that the two DNA extracts differed significantly, while the pairs of tagged primers from each extract were indistinguishable. Similar results were obtained using UniFrac phylogenetic comparisons. Together, these results suggest that the pig-tagged primers can be used to increase ecological inference in high throughput sequencing projects on fungi.

Keywords: community genetics, fungi, microbial communities, microbial ecology, new tools/technological developments

Received 27 August 2007; revision accepted 3 December 2007

Direct characterization of DNA from environmental samples has revolutionized our understanding of microbial evolution and ecology (Stahl *et al.* 1984; DeLong 1992). While many culture-independent methods for characterizing microbial communities have been developed, including dot-blots, microarrays, and fluorescent *in situ* hybridization, the most widely used methods involve polymerase chain reaction (PCR) amplification of diagnostic gene-regions, often the ribosomal small subunit. Most natural microbial communities

are diverse, and these environmental PCR approaches produce highly heterogeneous pools of amplicons. There are several categories of strategies to sort these sequence mixtures into meaningful snapshots of the community including fingerprinting methods, microarray hybridization and clone library sequencing methods. No arrays are currently available for characterization of fungal diversity. Fingerprinting methods seek to resolve different phylotypes through separation of fragments based on sequence differences including differences in restriction sites [amplified ribosomal DNA restriction site analysis, terminal restriction fragment length polymorphism (T-RFLP); Heyndrickx *et al.* 1996; Liu *et al.* 1997; Moeseneder *et al.* 1999] and differences in fragment migration through various acrylamide gel

Correspondence: D. Lee Taylor, Fax: 907-474-6967;
E-mail: lee.taylor@iab.alaska.edu

‡Present address: University of Colorado Health Sciences Center, Denver, CO 80262, USA

formulations [single-strand conformation polymorphism and denaturing gradient gel electrophoresis (DGGE); Muyzer & Smalla 1998; Rosado *et al.* 1998; Lowell & Klein 2001; Stach *et al.* 2001]. These methods offer two major advantages over clone library sequencing in that fingerprinting methods are relatively high throughput and can be applied to many samples at a modest cost. However, these methods have the striking disadvantage that they do not provide phylogenetically explicit information. For this reason, fingerprinting methods are often combined with clone library sequencing strategies, which can convert anonymous bands on T-RFLP profiles or DGGE gels into clearly identified taxa through standard sequence matching (e.g. BLAST searches) and phylogenetic analyses. Clone library sequencing is a more direct route to phylogenetically explicit data, but is more expensive and cumbersome to apply to more than a few samples. In fact, the vast majority of published microbial studies that have used clone library sequencing analyse only one to four samples (e.g. Borneman *et al.* 1996; Felske *et al.* 1999; Widmer *et al.* 1999; Acinas *et al.* 2004; Cottrell *et al.* 2005; Hansgate *et al.* 2005), although several studies have characterized larger numbers of samples (Zhou *et al.* 2002a; Bik *et al.* 2006; Flanagan *et al.* 2007). By 'sample', we mean the constellation of amplicons that end up in a single clone library. Such a sample may or may not result from pooling at various stages, e.g. pooling of DNA extracts, PCR products, or ligations. In any case, sequences from a sample constitute a single snapshot of the community of interest, or a single population from a statistical point of view.

The limited numbers of samples that can be routinely subject to high throughput sequencing approaches (e.g. genome centre Sanger sequencing of clone libraries or pyrosequencing of PCR products) is an impediment to answering a wide range of important ecological questions. Given that 'Ecology is concerned with patterns of distribution (where organisms occur) and with patterns of abundance (how many organisms occur) in **space** and **time**' (McGraw-Hill 1992), analysis of small numbers of samples is problematic. Ecologists often want to know how organisms are distributed across a landscape, and how their abundance varies through time or in response to particular natural or manipulated conditions. To address such basic questions requires analysis of discrete, replicated samples collected across time and space. These issues become daunting when one considers the extreme diversity, rapid turnover, and strong microscale spatial structure of many microbial communities, particularly in soil (Boerner *et al.* 1996; Nunan *et al.* 2001; Zhou *et al.* 2002b; Mummey & Stahl 2003). The difficulty lies not in the numbers of clones or PCR products that can be sequenced, but in the numbers of libraries that can be separately analysed. In a high throughput context, it has generally not been feasible to divide sequencing effort across many libraries due to the time, money and human intervention involved in handling

separate libraries or carrying out numerous pyrosequencing runs. Of course, if one is sequencing 'in house' without the use of extensive robotics, the numbers of libraries that can be handled is less a limitation than is sequencing throughput.

The goal of handling large numbers of discrete samples while carrying out high throughput sequencing also applies to certain genomics problems, for example, the construction of tissue-specific transcription profiles through construction and sequencing of (expressed sequence tag) EST libraries and pyrosequencing of defined loci. Recently, the limitations to processing of numerous tissue samples for EST analysis has been largely overcome through the inclusion of short sequence tags during the initial reverse transcription steps prior to library sequencing (Gavin *et al.* 2002; Scheetz *et al.* 2004). A similar approach has been used for pyrosequencing of PCR products (Binladen *et al.* 2007). These 'tagging' methods allow cDNA pools from many different tissues or PCR products from different samples to be combined *in vitro* prior to sequencing, then separated *in silico* by identification of the source tags after sequencing (Gavin *et al.* 2002; Scheetz *et al.* 2003; Binladen *et al.* 2007). To our knowledge, tagging approaches have not been applied to the characterization of microbial community structure, where representative abundances of sequences from each source taxon are critical.

We hypothesized that the addition of sequence tags at the 5' ends of a primer used in the PCR would allow a similar approach to pooling and subsequent separation of samples to be used in analyses of multiple microbial communities. However, PCR amplification from environmental samples is known to suffer from several potential biases in the representation of microbial communities. For example, kinetic bias occurs as PCR products accumulate and primers decline in abundance with increasing numbers of cycles and as the most abundant sequence types re-anneal and outcompete the primers for binding more often than the rare sequence types (Suzuki & Giovannoni 1996). The result is that taxon abundances approach equality, regardless of their starting ratios. Very low numbers of PCR cycles seem to help minimize this bias (Suzuki *et al.* 1996; Polz & Cavanaugh 1998). Because sequences with a high per cent G + C have higher melting temperatures and stronger secondary structures, they are at a competitive disadvantage in the PCR, leading to another form of bias. Finally, sequences which perfectly match the primers are at a competitive advantage over sequences with mismatches to the primers.

Bias may also occur at the cloning step, for example due to selection against particular insert sequences. In the case of adding variable tags to the ends of standard primers, particular tags could be more or less prone to terminal A addition (Brownstein *et al.* 1996), resulting in differential representation across tags when using TA cloning methods (Taylor *et al.* 2007). We therefore included an additional 'pig-tail' when constructing our 'pig-tagged' primers to

maximize nontemplate-dependent A addition (Brownstein *et al.* 1996). Due to concerns about potential PCR and cloning biases, we performed a detailed analysis to determine whether the addition of selected 10 base tags and a seven-base pig-tail to a standard primer would result in phylogenetic biases in the representation of a controlled mixture of four known fungal taxa. We also applied four of the primers to analyses of two soil DNA extracts from a natural environment containing numerous, unknown fungal taxa by pooling to create a single library. We find that 19 of 23 primers tested give statistically equivalent estimates of the simple, constructed fungal community. Even more encouragingly, the two tags applied to each of the soil samples yielded statistically indistinguishable patterns of fungal abundance, while the two soil samples yielded quite distinct sets of fungi. Hence, we suggest that the addition of carefully chosen and tested tags may dramatically expand the ecological inferences possible via high-throughput clone library sequencing.

Materials and methods

Primer design

Our primary concern is the characterization of fungal diversity in soil. We chose the fungal-selective primer ITS1-F (Gardes *et al.* 1991) and the universal eukaryote primer TW13 (White *et al.* 1990) as the starting points for primer design. These primers span the variable internal transcribed spacer (ITS) 1 and 2 regions of about 600 bp as well as about 700 bp of the more conserved nuclear ribosomal large subunit (LSU) structural region, allowing for species-level as well as deeper phylogenetic placement of unknown fungi, respectively. The primer ITS1-F has been widely used in studies of fungal diversity on mycorrhizal roots (Gardes *et al.* 1991; Gardes & Bruns 1996; Taylor & Bruns 1997) and in soil (Dickie *et al.* 2002; Rosling *et al.* 2003), and the ITS1-F/TW13 pair was efficient and reliable in prior testing on soil DNA extracts (Taylor *et al.* 2007). The first modification for our application was to balance and slightly raise the predicted annealing temperature of the ITS1-F/TW13 pair to 60 °C by adding four nucleotides to the 5' end of ITS1-F and two nucleotides to the 5' end of TW13. These base positions are highly (though not completely) conserved among the true fungi. The modified version of ITS1-F is designated ITS1-FL ('long'). Because ITS1-F provides the fungal selectivity in the PCR, we chose to then add the experimental tags to TW13. We downloaded a list of 288 previously designed 10-base tags from <http://genome.uiowa.edu/pubsoft/software.html>. These tags were designed to maximize their identification from resulting clone sequences in the face of sequencing errors by ensuring that each tag had an edit distance of 3 and a Hamming distance of 5 from every other tag (Scheetz *et al.* 2003). The practical implication of these rules is that tags suffering from up to one indel

and/or up to two substitutions/incorrect base calls can be identified correctly. This list of tags was then filtered to meet the following additional criteria. First, the 3' most base, which would link to the TW13 sequence, could not match the invariant fungal base C. Second, over the 10 bases of the tag, no more than three bases could match fungal bases, which are highly conserved in this region.

TW13 was modified further to minimize possible biases in the ligation step. A seven-base 'pig-tail' motif, GTTTCTT, shown to maximize nontemplate dependent A addition by *Taq* polymerase, was added to the 5' end of the primer (Brownstein *et al.* 1996). One base out of seven matches a conserved fungal base, and a second base matches some fungal sequences (see Fig. 1). The pig-tail is also beneficial because it moves the tag further from the vector-sequencing primer site, increasing the likelihood of high quality base calls for the tag, and provides an additional landmark of invariant sequence which can be used in the bioinformatic steps to help locate and identify the tag. The resulting 37-base oligos were then screened for the formation of significant hairpins, self-dimers and cross-dimers with ITS1-FL using the program NETPRIMER (Premier Biosoft, <http://www.premierbiosoft.com/netprimer/>). Oligos predicted to have unwanted hybridization features with free energies more negative than -6.5 kcal/mol were discarded. Of the 50 pig-tagged versions of TW13 passing this *in silico* screening step (see Table 1), the first 23 were then tested empirically.

Organisms and DNA isolation

The following four species were selected to represent the diversity of the kingdom Fungi to create a controlled mixed community: *Spizellomyces punctatus* ATCC 48900 (Chytridiomycota), *Mortierella alpina* ATCC 42430 (Zygomycota), *Peziza cf. varia* GAL 18629 (Ascomycota) and *Lichenomphalia umbellifera* GAL 18639 (Basidiomycota). The chytrid and zygomycete were obtained from the American Type Culture Collection and grown in the mycelial phase in Koch's K-L broth and potato dextrose broth, respectively, for 7 days, collected by centrifugation in 50 mL Falcon tubes, and then freeze-dried. The ascomycete and basidiomycete were collected as wild sporocarps in Interior Alaska, and dried over low heat, then deposited in the ALA Fungal Herbarium, housed at the University of Alaska Fairbanks. After grinding in a mortar and pestle under liquid nitrogen, genomic DNA was extracted from each of the four dried samples using the QIAGEN Plant DNeasy Maxi kit, following the manufacturer's instructions. The genomic DNA extracts were quantified using Pico Green (Molecular Probes) on an Analyst AD fluorometer, then normalized to 2.5 ng/μL.

To further test our tagging approach and illustrate an ecological application of the method, we analysed two composite soil samples that are part of a larger study of the

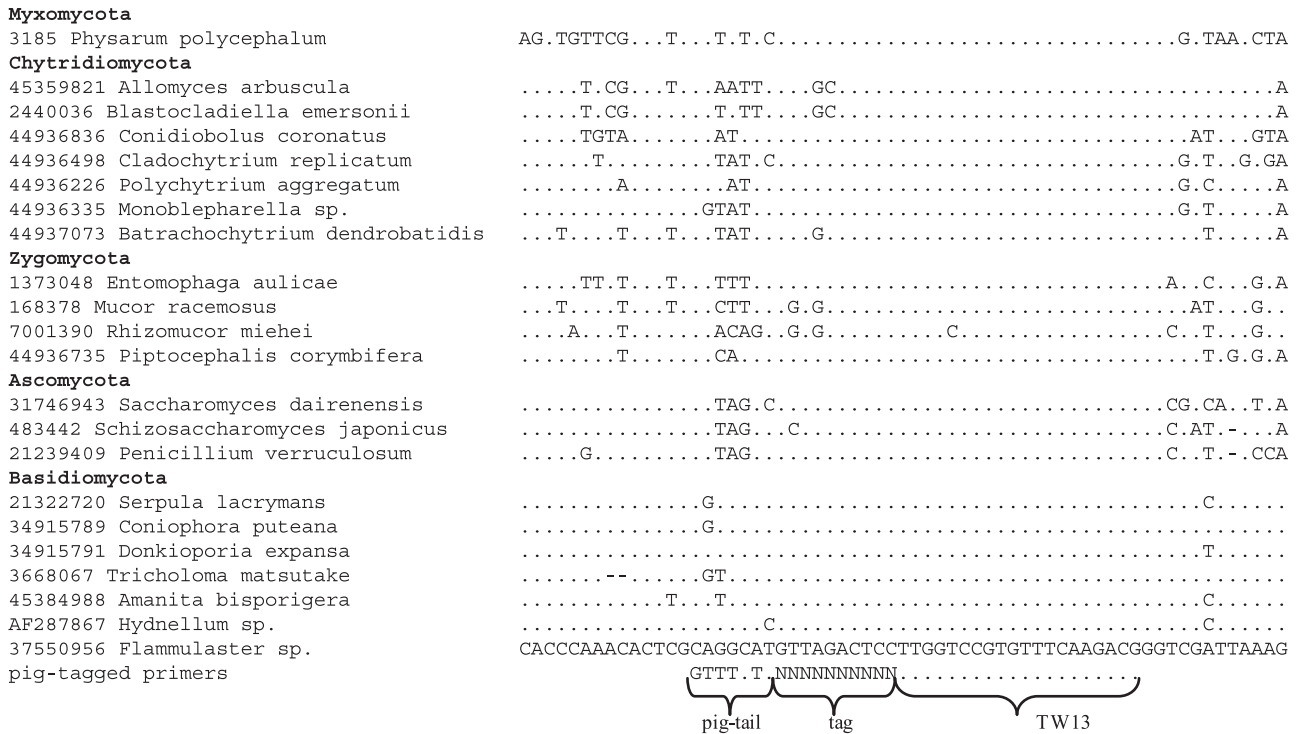


Fig. 1 Alignment of pig-tagged primer motif to fungal and slime-mould ribosomal large subunit sequences. Bases identical to those of *Flammulaster* sp. (Basidiomycota) are indicated with a dot. Sequences are reverse complement to the nuclear large subunit coding strand. The 3' end of TW13 directs DNA synthesis in the direction of the ITS spacers and the nuclear small subunit. See Table 1 for the nucleotides of the 10-base tags indicated with Ns here.

seasonal dynamics of the soil fungal community within a spruce (*Picea glauca*, *P. mariana*) forest located at the University of Alaska Fairbanks campus, under the auspices of the Bonanza Creek Long-term Ecology Research (LTER) programme (<http://www.lter.uaf.edu/>). Organic portions from soil cores were deposited in 50 mL Falcon tubes, and stored at -80 °C until freeze-drying. The soils were then ground on a ball mill at -20 °C and two sets of 10 replicate cores from different arrays of plots (all within a single 150 × 150 m site) were composited to create two soil pools. Genomic DNA was extracted from 1 g of soil from each of the two composite samples using the Mo Bio Powersoil kit following the manufacturer's instructions. The soil DNA extracts were normalized to 2.5 ng/μL after Pico Green quantification.

PCR

Each of the four known taxa was amplified separately using ITS1-FL and TW13-68P (one of the 23 pig-tagged primers) then cloned in order to generate reference RFLP patterns. Subsequently, aliquots of the four genomic DNA extracts were pooled in equimolar proportions and amplified using ITS1-FL in combination with each of the 23 pig-tagged TW13 primers. For comparative purposes, a parallel library

was constructed by amplifying the mixed template with the original, unmodified versions of ITS1-F and TW13. Seven replicate PCRs were carried out and subsequently pooled for each primer pair in order to average stochastic events within individual PCRs. PCRs utilized 25 μL Amersham Ready-To-Go beads, 0.5 μM primers, and the following cycling conditions: initial denaturation at 96 °C for 2 min followed by 25 cycles of denaturation at 94 °C for 30 s, annealing at 57 °C for 40 s and extension at 72 °C for 3 min, with a final extension at 72 °C for 10 min.

One of the composited soil DNA extracts was amplified, separately, with ITS1-FL and the pig-tagged primers TW13-064P and TW13-102P, while the other sample was amplified with pig-tagged primers TW13-067P and TW13-126P. PCR conditions were the same as above, except that only three replicate PCRs were carried out for each of the four primer/sample combinations.

Clone library construction and RFLP analysis

To minimize the cloning of primer-dimers and other short inserts, 100 μL of the pooled fragments from the controlled mix as well as the complex soil amplicons were size-fractionated over Chroma Spin 400 columns (BD Biosciences), then concentrated through DNA Clean and Concentrator-5

Table 1 Primer sequences. The primer ITS1-FL was modified from the primer ITS1-F published by Gardes and Bruns (1993). The pig-tagged primers were generated through the addition of two conserved fungal bases, a variable 10 base tag and an invariant seven base pig-tail to the TW13 primer (T.J. White unpublished, see <http://plantbio.berkeley.edu/~bruns/primers.html>; this primer has been used in previous publications, including Taylor & Bruns (1999). Asterisks (*) denote the three empirically tested primers which produced taxonomic ratios different from those of the remaining 19 primers

Forward primer	
ITS1-FL	CAAACTTGGTCATTTAGAGGAAGTAA
Empirically tested, pig-tagged reverse primers	
TW13-2P	GTTTCTT-AGAATGTCCT-TTGGTCCGTGTTTCAAGACG
TW13-6P*	GTTTCTT-CTGCTAGTAG-TTGGTCCGTGTTTCAAGACG
TW13-23P	GTTTCTT-TGTACCGATG-TTGGTCCGTGTTTCAAGACG
TW13-32P	GTTTCTT-TGCCAGTAAT-TTGGTCCGTGTTTCAAGACG
TW13-38P	GTTTCTT-CTCACTTAGG-TTGGTCCGTGTTTCAAGACG
TW13-43P*	GTTTCTT-AGACCTTCTG-TTGGTCCGTGTTTCAAGACG
TW13-59P	GTTTCTT-CGCCGTTATA-TTGGTCCGTGTTTCAAGACG
TW13-64P	GTTTCTT-TGTGGTTCAT-TTGGTCCGTGTTTCAAGACG
TW13-65P	GTTTCTT-CATCTCTACT-TTGGTCCGTGTTTCAAGACG
TW13-67P	GTTTCTT-AGCCGCCGAT-TTGGTCCGTGTTTCAAGACG
TW13-68P	GTTTCTT-CGTAGTGA-TTGGTCCGTGTTTCAAGACG
TW13-69P*	GTTTCTT-CTAATGGACG-TTGGTCCGTGTTTCAAGACG
TW13-87P	GTTTCTT-ATAGCAAGAT-TTGGTCCGTGTTTCAAGACG
TW13-91P	GTTTCTT-AGGCATCACA-TTGGTCCGTGTTTCAAGACG
TW13-99P	GTTTCTT-ATCTAATATG-TTGGTCCGTGTTTCAAGACG
TW13-102P	GTTTCTT-TAGTAATAT-TTGGTCCGTGTTTCAAGACG
TW13-103P	GTTTCTT-TACGATGATG-TTGGTCCGTGTTTCAAGACG
TW13-106P	GTTTCTT-TGGCTTAGAG-TTGGTCCGTGTTTCAAGACG
TW13-112P	GTTTCTT-AGAGGAGACG-TTGGTCCGTGTTTCAAGACG
TW13-121P	GTTTCTT-TTACAGGTGA-TTGGTCCGTGTTTCAAGACG
TW13-123P	GTTTCTT-AACAGCCATA-TTGGTCCGTGTTTCAAGACG
TW13-126P	GTTTCTT-CATAGCACTG-TTGGTCCGTGTTTCAAGACG
TW13-127P	GTTTCTT-CTATGTATA-TTGGTCCGTGTTTCAAGACG
Pig-tagged reverse primers, not empirically tested	
TW13-136P	GTTTCTT-AGAAGGAGTG-TTGGTCCGTGTTTCAAGACG
TW13-146P	GTTTCTT-CGTGTCGTAA-TTGGTCCGTGTTTCAAGACG
TW13-147P	GTTTCTT-CTCATCTGAA-TTGGTCCGTGTTTCAAGACG
TW13-148P	GTTTCTT-AACTTAATAG-TTGGTCCGTGTTTCAAGACG
TW13-153P	GTTTCTT-TAGTGATTA-TTGGTCCGTGTTTCAAGACG
TW13-154P	GTTTCTT-ATGTGATCCT-TTGGTCCGTGTTTCAAGACG
TW13-156P	GTTTCTT-ACAACAGTCA-TTGGTCCGTGTTTCAAGACG
TW13-162P	GTTTCTT-ATTGACGACA-TTGGTCCGTGTTTCAAGACG
TW13-179P	GTTTCTT-TAGGTCATCA-TTGGTCCGTGTTTCAAGACG
TW13-182P	GTTTCTT-ACTCGCTCCG-TTGGTCCGTGTTTCAAGACG
TW13-190P	GTTTCTT-ATACTCCTAA-TTGGTCCGTGTTTCAAGACG
TW13-191P	GTTTCTT-TTGGTTGGCT-TTGGTCCGTGTTTCAAGACG
TW13-198P	GTTTCTT-CTCTAGCCAT-TTGGTCCGTGTTTCAAGACG
TW13-206P	GTTTCTT-TATCTGGCTG-TTGGTCCGTGTTTCAAGACG
TW13-209P	GTTTCTT-AATTGGATCA-TTGGTCCGTGTTTCAAGACG
TW13-215P	GTTTCTT-CTATGGTCTG-TTGGTCCGTGTTTCAAGACG
TW13-217P	GTTTCTT-ATAGTGTGG-TTGGTCCGTGTTTCAAGACG
TW13-219P	GTTTCTT-TGCTGGCGTA-TTGGTCCGTGTTTCAAGACG
TW13-226P	GTTTCTT-TGCTGTGACT-TTGGTCCGTGTTTCAAGACG
TW13-229P	GTTTCTT-ACTAAGGTTG-TTGGTCCGTGTTTCAAGACG
TW13-242P	GTTTCTT-AGTTTCACATA-TTGGTCCGTGTTTCAAGACG
TW13-248P	GTTTCTT-TCAGCAATGG-TTGGTCCGTGTTTCAAGACG
TW13-254P	GTTTCTT-ACTGGTCATG-TTGGTCCGTGTTTCAAGACG
TW13-266P	GTTTCTT-CTACATGAG-TTGGTCCGTGTTTCAAGACG
TW13-275P	GTTTCTT-AACATAGACT-TTGGTCCGTGTTTCAAGACG
TW13-279P	GTTTCTT-CATCATAAT-TTGGTCCGTGTTTCAAGACG
TW13-284P	GTTTCTT-TAATAGATTG-TTGGTCCGTGTTTCAAGACG
TW13-285P	GTTTCTT-TCTCCTGTAG-TTGGTCCGTGTTTCAAGACG

columns (Zymo Research), as in Taylor *et al.* (2007). The templates were quantified on a Nanodrop spectrophotometer and normalized to 25 ng/μL prior to ligation. The Invitrogen TOPO TA for sequencing kit with the pCR4.0 vector was used for cloning. Ligation reactions utilized 1.0 μL of vector, 1.0 μL of salt solution and 4.0 μL of template and were incubated on the bench for 30 min, then frozen at -80 °C until used for transformation. For the controlled mixture ligations, chemical transformation with Invitrogen TOP10 cells was carried out following the manufacturer's instructions, followed by spreads on LB plus kanamycin (50 mg/mL) plates and manual colony picking. For the two complex soil libraries, electroporation of GC10 Thunderbolt cells was carried out, followed by robotic colony picking.

Colonies chosen at random from each controlled mixture ligation were subject to direct PCR amplification using the vector primers M13F and M13R at 0.5 μM final concentration with Sigma Red *Taq* Ready-Mix increased to 2.15 mM MgCl in 25-μL volumes. Colony PCR was carried out with cell lysis and initial denaturation at 96 °C for 8 min, followed by cycles of denaturation at 92 °C for 30 s, annealing at 50 °C for 45 s, extension at 72 °C for 2 min for 35 cycles. The enzymes *AluI*, *HinfI*, *MboI* and *HaeIII* (New England Biolabs) were tested for their ability to distinguish clones produced by the four organisms in the controlled mixture. All four enzymes produced one or two unique patterns for each taxon; *AluI* was chosen for testing of each tagged primer amplification from the mixed templates. To test whether the four template taxa from the artificial mixed community were similarly represented in the libraries created with each of the 23 pig-tagged primers, 95 colonies were screened by *AluI* RFLPs for each of the 23 pig-tagged primers. Restriction digests were performed in 20 μL volumes containing 12 μL of PCR product, 1× BSA, 1× buffer NEB-2 and 2.5 U of *AluI* with incubation overnight at 37 °C. Digested products were separated in gels containing 1% Seakem agarose and 2% Nusieve agarose at 170 V for approximately 2 h. If all bands were close to the expected size, a clone was assigned to one of the four taxa. Clones producing unexpected fragment sizes were excluded from subsequent analyses.

To confirm the identities of the four fungi, we sequenced two colonies from each of the reference clone libraries representing the four fungal phylotypes. Colonies were selectively amplified using vector primers and reaction conditions as above. PCR products were cleaned using QIAGEN QIAquick columns and cycle-sequenced using 2.0 μL ABI BigDye 3.1 Terminator Mix, 3.0 μL of 5× buffer, and 0.25 μM sequencing primer and 40 ng of PCR product in a total volume of 12 μL per reaction. We used two primer sets: (i) vector primers M13F and M13R, and (ii) internal primers ITS4 and Ctb6 (White *et al.* 1990). Products from cycle sequencing were cleaned over Sephadex and run

through an ABI 3100 capillary sequencer. Sequences have been deposited in GenBank under Accessions EU292220–EU292675 (www.ncbi.nlm.nih.gov).

Sequencing of clones from the complex soil libraries was carried out as described previously (Taylor *et al.* 2007). Plasmids were amplified using the TempliPhi kit (Nelson *et al.* 2002), followed by cycle sequencing with ABI BigDye 3.0 and electrophoresis on an ABI 9700.

Sequence processing

A random subset of the clones obtained was chosen for analysis. Reads from the same clone were assembled using ALIGNER (CodonCode). A script was written to scan sequences by first identifying a perfect match to either the core TW13 sequence or the pig-tail sequence, then searching for tags with an edit distance of 2 or less from the predicted tag sequences. All consensus base positions with phred scores below 15 were converted to Ns. All sequences were aligned with CLUSTAL W and imported into SE-AL (Rambaut 1996), then the primer plus vector sequences were removed. Separate ITS and LSU portions of the alignment were created, gaps were removed and sequences exported. Any remaining dirty ends internal to the primer positions (or at the edges of the ITS or LSU) were removed using a window size of 20 and an ambiguity limit of 5% in the TRIMSEQ program of the EMBOSS package. Sequences with greater than 2% Ns overall were discarded, as were sequences inferred to be of chimeric origin. Chimera detection was attempted using three strategies: (i) the native chimera detection component of CAP 3 was applied during assembly of complete ITS plus LSU sequences into contigs at several per cent identity ($-p$) settings, (ii) the LSU alignment was submitted to the program BELLEROPHON, available on the internet at foo.maths.uq.edu.au/~huber/bellerophon.pl (Huber *et al.* 2004), and (iii) ITS1 and LSU regions from each clone were submitted to separate FASTA comparisons with fungal sequences from GenBank using our web server (<http://www.borealfungi.uaf.edu/>). The GenBank organism field is part of the output from our site, allowing rapid comparison of the taxonomic affinities of the ITS1 vs. LSU regions. A final, balanced set of 114 clones with ITS and LSU portions from each of the four TW13 tags was carried through all subsequent analyses.

OTU discrimination, phylogenetic and statistical analysis

We tested whether the counts of the four known taxa in the controlled mixture differed between primers by contingency table analyses to obtain Pearson χ^2 statistics for homogeneity in JMP 6.0 (SAS Institute). Primers with the largest χ^2 values were removed from the data set sequentially until a homogeneous set of primers was identified.

In the wider analysis of unknown fungi from the soil libraries, operational taxonomic units (OTUs) were distinguished by assembling ITS sequences using the CAP 3 program (Huang & Madan 1999) as described previously in Taylor *et al.* (2007). Our target was 97% sequence identity which is a reasonable approximation to species limits in many fungi. Using fabricated test data sets, we found that the following parameter settings for CAP 3 result in assembly of sequences with 3% or less divergence: maximum overhang per cent length = 60, match score factor = 6, overlap per cent identity cutoff = 96, clipping range = 6. Counts of clones belonging to each OTU from each tag were tabulated in Excel and exported to ESTIMATES 8.0 (Colwell & Coddington 1994). A variety of community diversity and similarity statistics were then calculated for the populations of OTUs recovered from each of the four tags. The community similarity indices range from zero to 1.0, with higher scores indicating greater similarity. In addition, counts of OTUs were compared across the four tags using the Pearson χ^2 contingency table test in JMP 4.0. For every contingency test, the rarest OTUs were removed until less than 20% of the expected cell values were below 5, in order to avoid severe violations of the χ^2 distribution. In most cases, OTUs with total counts below 10 were not included in these tests.

Because the ITS region is unalignable across diverse fungi, only the LSU portions of each clone sequence were utilized for phylogenetic comparisons of tags and soil samples. Sequences from three chytrids from GenBank (DQ273798, DQ273779 and DQ273825) were added to the CLUSTAL W LSU alignment to serve as outgroup, followed by manual improvement of the alignment in SE-AL. A search for the maximum-likelihood tree was carried out using GARLI 0.95 (Zwickl 2006). The GTR + G + I model of sequence evolution was used, along with default parameters for numbers of populations, search termination criteria, etc. The best tree found was then imported to the web portal for UniFrac (Lozupone *et al.* 2006), along with a file listing the tag with which each clone sequence was associated. To infer whether the four tags recovered phylogenetically distinct arrays of fungi, we carried out the *P* test described by Martin (2002), as well as a closely related test using the UniFrac sequence divergence metric, which takes branch lengths into account rather than solely counting branches under a parsimony reconstruction approach. We first tested for any difference among the four tags, followed by pairwise tests for differences between tags. For both tests, the *P* values represent the probability that the two communities are the same, and have been Bonferroni-corrected for the numbers of comparisons. We also carried out UPGMA clustering of the communities recovered from each tag using the UniFrac metric with significance testing by jackknifing via 100 resamplings of 70 sequences per tag (Lozupone *et al.* 2006).

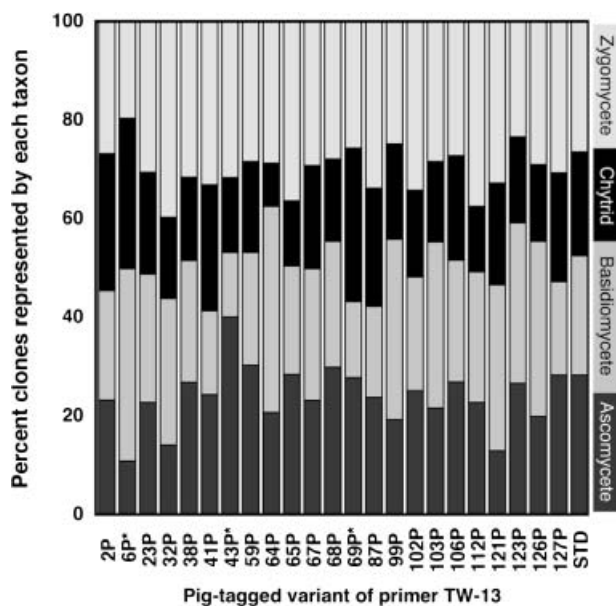


Fig. 2 Taxonomic distributions of fungal ITS clones derived from amplifications of a DNA mixture with pig-tagged primers (variants of the universal primer, TW13) and a fungus-specific primer, ITS1-FL. The three primers marked with asterisks (*), 6P, 43P and 69P, produced distributions significantly different from those produced by the other primers. Removal of those distributions from the larger set resulted in a statistically homogenous group of distributions produced by 19 TW13 variants plus the untagged, standard primers.

Results

Four-taxon mixture

As a first test for phylogenetic bias with the pig-tagged primers, we constructed a community using equal genomic DNA ratios of four distantly related fungi (Chytridiomycete, Zygomycete, Ascomycete, Basidiomycete). Percentages of sampled clones representing the four taxa recovered for each of the 23 pig-tagged primers are shown in Fig. 2. In general, the results suggest that the addition of a variable tag along with a conserved pig-tail to the primer TW13 does not influence the representation of the four tested taxa. We utilized equal ratios of genomic DNA extracts from the four taxa to create our artificial mixed community, but obtained unequal average ratios of clones representing the four taxa in the libraries. There are many possible explanations for the imbalance in ratios, including differences in genome sizes and copy numbers of the ribosomal repeats, which are well known in both prokaryotes and eukaryotes (Rogers & Bendich *aj* 1987; Klappenbach *et al.* 2001). However, genome sizes and repeat numbers are not known for any of our test fungi.

All four taxa were recovered from each of the 24 libraries analysed (including the library from the standard, untagged

primers) and proportions of the taxa recovered were similar across the libraries. However, contingency table analysis rejected homogeneity of count distributions across the 24 primers ($P < 0.0001$). After sequentially removing the three primers with the largest χ^2 values, TW13–6P, TW13–43P and TW13–69P, the null hypothesis of homogeneity among taxon ratios was accepted ($P = 0.09$). A fourth primer, TW13–64P, was a borderline outlier; its removal further reduces the χ^2 significance to 0.28. It is notable that the unmodified ITS1-F and TW13 fell within the homogeneous core group, further supporting the conclusion that the modified primers do not introduce strong biases in either the PCR, ligation or transformation steps. The major advantage of this approach is that a large number of pig-tagged primers could be tested empirically using a simple RFLP method.

We examined the sequences of the failing primers, and we were unable to detect any patterns that might explain their observed biases. For example, the numbers of tag bases matching fungal bases in the different clades did not differ among the passing and failing primers.

Communities of unknown soil fungi

As a second test of the pig-tagging approach, we carried out amplification and clone library construction using two natural soil samples. In contrast to our RFLP analyses of constructed communities, analyses of soil communities required intensive sequencing efforts to even partially characterize fungal diversity. For this reason, we only analysed four of the 23 pig-tagged primers. However, these soil community comparisons offer the distinct advantage of high species diversity and ecological realism. After removing clones with greater than 2% of base calls with phred scores below 15 and one clone which appeared to be of nonfungal origin, we analysed 464 clones in detail. When the per cent identity ($-p$) threshold was set at 75%, CAP 3 did not detect any chimeras in the complete ITS + LSU clone sequences. At 90%, one putative chimera was detected, which was judged as equivocal when inspected manually because it was a low match to the putative parental sequences throughout (FASTA match of 69% to a member of the Helotiaceae in the ITS region vs. a match of 87% to a member of the Lecanoraceae in the LSU region). This sequence was retained in the final data set. Using a $-p$ setting of 97%, two additional putative chimeras were listed, but our FASTA comparisons of ITS1 and the LSU indicated nonchimeric origins of these sequences. Both regions had top hits to *Mortierella* for one clone, while both regions had top hits to *Lachnum* for the other clone. Because these two taxa were rare in these samples, a chimeric origin seems unlikely. BELLEROPHON reported numerous putative chimeras in the LSU data set. However, when these clones were inspected manually, none were found to be definite chimeras. In most

Table 2 OTU and phylogenetic comparisons of tags. The dominant OTU χ^2 is the probability (P value) of homogeneity from the Pearson test computed in JMP 6.0

Metric	Within soil sample		Across soil samples			
	TW13-064 vs. TW13-102	TW13-067 vs. TW13-126	TW13-064 vs. TW13-067	TW13-064 vs. TW13-126	TW13-067 vs. TW13-102	TW13-102 vs. TW13-126
Dominant OTU χ^2	0.994	0.478	< 0.0001	< 0.0001	0.0002	< 0.0001
Classic Sorenson	0.458	0.471	0.302	0.382	0.196	0.253
Chao–Sorenson–EST	0.911	0.933	0.568	0.636	0.490	0.491
Morisita–Horn	0.888	0.947	0.551	0.542	0.568	0.570
P test	0.360	1.0	≤ 0.06	≤ 0.06	≤ 0.06	≤ 0.06
UniFrac test	1.0	1.0	1.0	0.06	0.06	≤ 0.06

cases, the putative chimeric sequence and the two putative parents were each unique throughout the LSU fragment. In other cases, sets of putative chimeras and parents fell within a single OTU (hence, even if PCR recombination had occurred, it would have no effect on the OTU-based comparisons and little if any effect on the phylogenetic comparisons). In contrast, the FASTA-matching and comparison pinpointed 12 potential chimeras, of which six were deemed likely chimeras after manual inspection. These six chimeras were removed, along with three randomly selected additional clones to create the balanced data set of 114 clones from each tag.

The communities recovered from the two soil samples were highly diverse and undersampled: 121 OTUs, 69 of which were singletons, were distinguished across the 456 sequences. As shown in Table 2, community similarity indices that compare only species richness gave low similarity scores across all pairs of tags due to the stochastic nature of the recovery of rare taxa in undersampled libraries (Chao *et al.* 2005; Taylor *et al.* 2007). However, similarities between tags from the same soil sample were higher than similarities across soil samples for all indices and all comparisons. Furthermore, indices such as Morisita–Horn that consider abundances yielded high similarity estimates for the pairs of tags from the sample soil composite (0.888 and 0.947) and considerably lower similarities for comparisons of tags from different soil samples (from 0.542 to 0.570). The Chao–Sorenson–EST index is a modified version of Sorenson similarity which takes in account both species abundances in each sample, as well as estimates of the undetected shared species derived from the numbers of singletons and doubletons in each sample (Chao *et al.* 2005); it gave similar results as the Morisita–Horn index. Contingency table analyses of counts of the most abundant OTUs provided a similar perspective on the data. Counts from the pairs of the tags from the same soil sample were homogenous ($P = 0.478$ and 0.994), while counts from tags from different soil samples were heterogeneous ($P = 0.0002$ to < 0.0001).

In contrast to community diversity statistics which weigh all distinct species equally in comparing samples, phylogenetic methods take levels of sequence divergence into account in comparing communities, while placing less emphasis on species per se. The P test described by Maddison & Slatkin (1991) and applied to microbial communities by Martin (2002) is one such method, which has become quite popular among microbial ecologists. Lozupone & Knight (2005) have expanded upon this framework by considering genetic distance in the form of branch length unique to each community in their UniFrac metric, as opposed to the original P test which considers only branching topology. Analyses of the LSU sequences recovered from the four tags using UniFrac revealed similar trends as seen when comparing OTU composition. The Bonferroni-corrected P test showed no difference between the pairs of tags from the same soil sample (both $P = 0.36$ – 1.0), while tags from different soils were different (both $P \leq 0.06$; Table 2). Results using the UniFrac metric were similar (Table 2), except that only one pair of tags from different soil samples was significantly different (TW13–102P vs. TW13–126P, $P \leq 0.06$), while two other differences in two of the other three pairs from different samples were marginally different ($P = 0.06$). The two comparisons between tags from the same sample were nonsignificant ($P = 1.0$). Cluster analysis using the UniFrac metric grouped each pair of tags from the same sample with high jackknife support (Fig. 3).

Discussion

The statistically unbiased results we obtained from 19 of the tags when applied to a controlled mixture of four fungal taxa is a promising indication of the utility of our approach. Because these four taxa represent the four major Phyla within the true Fungi, these primers could be meaningfully applied to environmental samples expected to contain extremely diverse fungi. We did observe variation in the ratios of the four taxa across tagged primers. The contribution of various types of sampling error to this

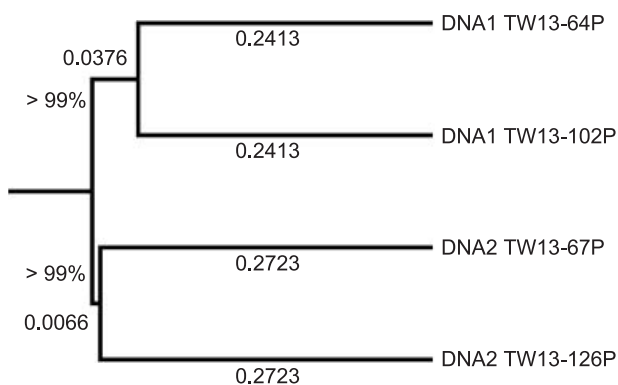


Fig. 3 UPGMA cluster analysis of UniFrac distances based upon the maximum-likelihood LSU tree obtained for the 456 clones. The taxon labels indicate the source soil DNA extract (1 vs. 2) and the pig-tagged primer used. Branch lengths from the UniFrac distance metric are given below or near branches, and jackknife support for the two nodes from 100 resampling runs are also shown. Note that the arrays of sequences obtained from the same soil sample with the different pig-tagged primers cluster together.

variation is unknown but likely high. A small subset of the initial population of molecules is sampled at each step from PCR, through ligation and, finally, colony picking. We are unaware of studies that have carefully quantified both sampling error and systematic biases at each of these steps in an environmental microbial diversity context. Such an undertaking would be laborious but worthwhile.

The results we obtained from natural, complex soil communities are complementary and even more encouraging than those from the controlled mixture. All analyses show the fungi recovered from pairs of tags from the same soil sample to be highly similar (Morisita–Horn similarity) and statistically indistinguishable (χ^2 test of abundant OTUs, P test and UniFrac tests). In contrast, analyses show the fungi recovered from the two soil samples to be more divergent and statistically different in three of the four tests. The biological significance of these results is supported by the fact that both OTU-based community statistics and phylogenetic comparisons paint the same picture. The repeatability of the characterization of fungal communities from a highly diverse, natural substrate using different primers provides robust validation of this approach. Given the sampling error associated with every step of the process, the degree of coincidence between the tags is remarkable. It is also interesting that the OTU analyses detected stronger differences between the two soil samples than did the phylogenetic analyses. Based upon BLAST results and inspection of the phylogenetic tree (data not shown), we attribute this result to the presence of a similar diversity of deep fungal lineages in all tag/sample combinations, while certain closely related species are present in one soil sample or the other.

Using these primers, 19 samples can be pooled then subjected to the same colony picking and sequencing approach that would formerly have been used to analyse a single sample. Tag identification methods have been developed and used in a number of EST studies (Gavin *et al.* 2002; Scheetz *et al.* 2003; Scheetz *et al.* 2004). We have modified the tag-finding algorithm from ESTprep (Scheetz *et al.* 2003) in order to identify tags in fungal clone library sequences, and we will make these tools available on our fungal identification web site, <http://www.borealfungi.uaf.edu/> (Geml *et al.* 2005). An approximately twentyfold increase in sampling offers the potential for considerably increased insight into the spatial and temporal dynamics of complex fungal communities in soil. Furthermore, the steps we have carried out in design and testing of tagged primers should be applicable to other eukaryotes, as well as Bacteria and Archaea, for which high throughput sequencing approaches are widely used to characterize environmental diversity.

Our pig-tagged approach is most likely to be useful in very high throughput, automated sequencing projects, where the manual handling of many separate clone libraries becomes problematic. While next-generation sequencing such as with the GS20 pyrosequencing platform or the Solexa platform are promising, they are not yet informative approaches for the identification of fungal species, since the smallest variable unit book-ended by conserved primer sites would be either ITS1 or ITS2, requiring dual end read lengths of at least 300 bp. However, when adequate read lengths are obtained, these methods will be especially well suited to a tagging approach, due to the highly parallel processes in which large numbers of templates are randomly distributed across a single sequencing plate (Margulies *et al.* 2005). Binladen *et al.* (2007) utilized all possible two-base tags and a few four-base tags at the 5' end of their primers when testing the utility of tagging for multiplexing PCR samples in GS20 platform pyrosequencing. While the vast majority of sequence reads could be related to their source via the tag sequence, very strong biases were found in the abundances of sequences from different samples that were statistically correlated with the tag bases in both the first and second positions. Hence, it appears that a two-base tag GS20 pyrosequencing approach is not currently viable for microbial community analyses where *abundances* of reads from different samples and taxa are critical data points. In contrast, we find statistically equivalent abundances of taxa amplified using different tags, meaning that our method lacks the strong biases found by Binladen *et al.* Our approach also has the advantage of the capacity to correctly identify tags even in the presence of limited sequencing errors. In our genome-centre sequencing of fungal communities in the Bonanza Creek LTER site in Interior Alaska, we have found that our tagging approach has greatly increased our opportunities for drawing ecological inferences due to the division of sequencing effort across many more samples.

We expect that our approach, or carefully validated variants thereof, will be of increasing utility to a wide spectrum of microbial ecologists using either next generation platforms or Sanger sequencing.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant no. 0317144. We thank Mitali Patil for assistance in the laboratory, Dan Cardin for writing the script to identify tags, Shawn Huston and James Long for general computational support, and Gary A. Laursen for the *Peziza* and *Lichenomphalia* specimens.

References

- Acinas SG, Klepac-Ceraj V, Hunt DE *et al.* (2004) Fine-scale phylogenetic architecture of a complex bacterial community. *Nature*, **430**, 551–554.
- Bik EM, Eckburg PB, Gill SR *et al.* (2006) Molecular analysis of the bacterial microbiota in the human stomach. *Proceedings of the National Academy of Sciences, USA*, **103**, 732–737.
- Binladen J, Gilbert MTP, Bollback JP *et al.* (2007) The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing. *PLoS One*, **2**, e197.
- Boerner REJ, Demars BG, Leicht PN (1996) Spatial patterns of mycorrhizal infectiveness of soils along a successional chronosequence. *Mycorrhiza*, **6**, 79–90.
- Borneman J, Skroch PWO, Sullivan KM *et al.* (1996) Molecular microbial diversity of an agricultural soil in Wisconsin. *Applied and Environmental Microbiology*, **62**, 1935–1943.
- Brownstein MJ, Carpten JD, Smith JR (1996) Modulation of non-templated nucleotide addition by tag DNA polymerase: primer modifications that facilitate genotyping. *BioTechniques*, **20**, 1004–1006.
- Chao a, Chazdon RL, Colwell RK, Shen TJ (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters*, **8**, 148–159.
- Colwell RK, Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **345**, 101–118.
- Cottrell MT, Waidner LA, Yu LY, Kirchman DL (2005) Bacterial diversity of metagenomic and PCR libraries from the Delaware River. *Environmental Microbiology*, **7**, 1883–1895.
- DeLong EF (1992) Archaea in coastal marine environments. *Proceedings of the National Academy of Sciences, USA*, **89**, 5685–5689.
- Dickie IA, Xu B, Koide RT (2002) Vertical niche differentiation of ectomycorrhizal hyphae in soil as shown by T-RFLP analysis. *New Phytologist*, **156**, 527–535.
- Felske A, Wolterink A, van Lis R, de Vos WM, Akkermans ADL (1999) Searching for predominant soil bacteria: 16S rDNA cloning versus strain cultivation. *FEMS Microbiology Ecology*, **30**, 137–145.
- Flanagan JL, Brodie EL, Weng L *et al.* (2007) Loss of bacterial diversity during antibiotic treatment of intubated patients colonized with *Pseudomonas aeruginosa*. *Journal of Clinical Microbiology*, **45**, 1954–1962.
- Gardes M, Bruns TD (1993) ITS primers with enhanced specificity for Basidiomycetes — application to the identification of mycorrhizae and rusts. *Molecular Ecology*, **2**, 113–118.
- Gardes M, Bruns TD (1996) Community structure of ectomycorrhizal fungi in a *Pinus muricata* forest: above- and below-ground views. *Canadian Journal of Botany*, **74**, 1572–1583.
- Gardes M, White TJ, Fortin JA, Bruns TD, Taylor JW (1991) Identification of indigenous and introduced symbiotic fungi in ectomycorrhizae by amplification of nuclear and mitochondrial ribosomal DNA. *Canadian Journal of Botany-Revue Canadienne de Botanique*, **69**, 180–190.
- Gavin AJ, Scheetz TE, Roberts CA *et al.* (2002) Pooled library tissue tags for EST-based gene discovery. *Bioinformatics*, **18**, 1162–1166.
- Geml J, Via Z, Long J *et al.* (2005) *The Fungal Metagenomics Project Website*, <http://www.borealfungi.uaf.edu/>.
- Hansgate AM, Schloss PD, Hay AG, Walker LP (2005) Molecular characterization of fungal, community dynamics in the initial stages of composting. *FEMS Microbiology Ecology*, **51**, 209–214.
- Heyndrickx M, Vauterin L, Vandamme P, Kersters K, De Vos P (1996) Applicability of combined amplified ribosomal DNA restriction site analysis (ARDRA) patterns in bacterial phylogeny and taxonomy. *Journal of Microbiological Methods*, **26**, 247–259.
- Huang XQ, Madan A (1999) CAP 3: a DNA sequence assembly program. *Genome Research*, **9**, 868–877.
- Huber T, Faulkner G, Hugenholtz P (2004) BELLEROPHON: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics*, **20**, 2317–2319.
- Klappenbach JA, Saxman PR, Cole JR, Schmidt TM (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Research*, **29**, 181–184.
- Liu W-T, Marsh TL, Cheng H, Forney LJ (1997) Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA. *Applied and Environmental Microbiology*, **63**, 4516–4522.
- Lowell JL, Klein DA (2001) Comparative single-strand conformation polymorphism (SSCP) and microscopy-based analysis of nitrogen cultivation interactive effects on the fungal community of a semiarid steppe soil. *FEMS Microbiology Ecology*, **36**, 85–92.
- Lozupone C, Hamady M, Knight R (2006) UniFrac — an online tool for comparing microbial community diversity in a phylogenetic context. *BMC Bioinformatics*, **7**, 371.
- Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.
- Maddison WP, Slatkin M (1991) Null models for the number of evolutionary steps in a character on a phylogenetic tree. *Evolution*, **45**, 1184–1197.
- Margulies M, Egholm M, Altman WE *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.
- Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology*, **68**, 3673–3682.
- McGraw-Hill I (1992) *McGraw-Hill Encyclopedia of Science and Technology*, 7th edn. McGraw-Hill, New York.
- Moeseneder MM, Arrieta JM, Muyzer G, Winter C, Herndl GJ (1999) Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. *Applied and Environmental Microbiology*, **65**, 3518–3525.
- Mummy DL, Stahl PD (2003) Spatial and temporal variability of bacterial 16S rDNA-based T-RFLP patterns derived from soil of two Wyoming grassland ecosystems. *FEMS Microbiology Ecology*, **46**, 113–120.

- Muyzer G, Smalla K (1998) Application of denaturing gradient gel electrophoresis (DGGE) and temperature gradient gel electrophoresis (TGGE) in microbial ecology. *Antonie Van Leeuwenhoek*, **73**, 127–141.
- Nelson JR, Cai YC, Giesler TL *et al.* (2002) TempliPhi, phi29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *BioTechniques*, **44**–47.
- Nunan N, Ritz K, Crabb D *et al.* (2001) Quantification of the *in situ* distribution of soil bacteria by large-scale imaging of thin sections of undisturbed soil. *FEMS Microbiology Ecology*, **37**, 67–77.
- Polz MF, Cavanaugh CM (1998) Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology*, **64**, 3724–3730.
- Rambaut A (1996) *SE-AL: Sequence Alignment Editor*. Oxford University, Oxford, UK.
- Rogers SO, Bendich AJ (1987) Ribosomal-RNA genes in plants – variability in copy number and in the intergenic spacer. *Plant Molecular Biology*, **9**, 509–520.
- Rosado AS, Duarte GF, Seldin L, Van Elsas JD (1998) Genetic diversity of *nifH* gene sequences in *Paenibacillus azotofixans* strains and soil samples analyzed by denaturing gradient gel electrophoresis of PCR-amplified gene fragments. *Applied and Environmental Microbiology*, **64**, 2770–2779.
- Rosling A, Landeweert R, Lindahl BD *et al.* (2003) Vertical distribution of ectomycorrhizal fungal taxa in a podzol soil profile. *New Phytologist*, **159**, 775–783.
- Scheetz TE, Laffin JJ, Berger B *et al.* (2004) High-throughput gene discovery in the rat. *Genome Research*, **14**, 733–741.
- Scheetz TE, Trivedi N, Roberts CA *et al.* (2003) ESTprep: preprocessing cDNA sequence reads. *Bioinformatics*, **19**, 1318–1324.
- Stach JEM, Bathe S, Clapp JP, Burns RG (2001) PCR-SSCP comparison of 16S rDNA sequence diversity in soil DNA obtained using different isolation and purification methods. *FEMS Microbiology Ecology*, **36**, 139–151.
- Stahl D, Lane D, Olsen G, Pace N (1984) Analysis of hydrothermal vent-associated symbionts by ribosomal RNA sequences. *Science*, **224**, 409–411.
- Suzuki MT, Giovannoni SJ (1996) Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Applied and Environmental Microbiology*, **62**, 625–630.
- Taylor DL, Bruns TD (1997) Independent, specialized invasions of ectomycorrhizal mutualism by two nonphotosynthetic orchids. *Proceedings of the National Academy of Sciences, USA*, **94**, 4510–4515.
- Taylor DL, Bruns TD (1999) Community structure of ectomycorrhizal fungi in a *Pinus muricata* forest: minimal overlap between the mature forest and resistant propagule communities. *Molecular Ecology*, **8**, 1837–1850.
- Taylor DL, Herriott IC, Long J, O'Neill K (2007) TOPO-TA is A-OK: a test of phylogenetic bias in fungal clone library construction. *Environmental Microbiology*, **9**, 1329–1334.
- White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: *PCR Protocols: A Guide to Methods and Applications* (eds Innis MA, Gelfand DH, Sninsky JJ, White TJ), pp. 315–322. Academic Press, San Diego, California.
- Widmer F, Shaffer BT, Porteous LA, Seidler RJ (1999) Analysis of *nifH* gene pool complexity in soil and litter at a Douglas fir forest site in the Oregon Cascade Mountain Range. *Applied and Environmental Microbiology*, **65**, 374–380.
- Zhou J, Xia B, Treves DS *et al.* (2002a) Spatial and resource factors influencing high microbial diversity in soil. *Applied and Environmental Microbiology*, **68**, 326–334.
- Zhou JZ, Xia BC, Treves DS *et al.* (2002b) Spatial and resource factors influencing high microbial diversity in soil. *Applied and Environmental Microbiology*, **68**, 326–334.
- Zwickl DJ (2006) *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD Dissertation, The University of Texas.

Supplementary material

The following supplementary material is available for this article:

Table S1 Counts of all OTUs across the four pig-tagged primers

This material is available as part of the online article from:

<http://www.blackwell-synergy.com/doi/abs/10.1111/j.1755-0998.2008.02094.x>

(This link will take you to the article abstract).

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.