Metabarcoding of fungal communities associated with bark beetles

| Journal:            | Molecular Ecology Resources |
|---------------------|-----------------------------|
| Manuscript ID:      | MER-15-0323                 |
| Manuscript Type:    | Resource Article            |
| Date Submitted by the Author: | 06-Sep-2015             |
| Complete List of Authors: | Miller, Kirsten; Natural History Museum, Life Sciences |
|                      | Hopkins, Kevin; Natural History Museum, Life Sciences |
|                      | Inward, Daegan; Forest Research, |
|                      | Vogler, Alfried; Imperial College/Natural History Museum of London, Natural Sciences |
| Keywords:           | Community Ecology, DNA Barcoding, Fungi, Insects |
Metabarcoding of fungal communities associated with bark beetles

Kirsten Miller, Kevin Hopkins, Daegan J. G. Inward, Alfried P. Vogler

1 Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD, UK.
2 Department of Life Sciences, Imperial College London, Silwood Park Campus, Ascot SL5 7PY, UK.
3 Centre for Forestry and Climate Change, Forest Research, Alice Holt Lodge, Farnham, Surrey GU10 4LH, UK.

* Corresponding author: Alfried P. Vogler, Fax: 0207 942 5229, E-mail: a.vogler@nhm.ac.uk

Short running title: Fungal communities on bark beetles

Word count of the main text including tables, figure captions and references: 6719

Word count of Summary: 249
Abstract

Many species of fungi are closely allied with bark beetles, including many tree pathogens, but their species richness and patterns of distribution remain largely unknown. We established a protocol metabarcoding of fungal communities directly from total genomic DNA extracted from individual beetles, showing that the ITS3/4 primer pair selectively amplifies the fungal ITS. Using three specimens of bark beetle, each representing different species, we test the diversity of fungal communities and the repeatability of PCRs conducted with different primer tags. We obtained 727 fungal Operational Taxonomic Units (OTUs) for the specimen of Hylaster ater, 435 OTUs in Tomicus piniperda and 294 OTUs in Trypodendron lineatum. Individual PCR reactions produced an average of only 229, 54 and 31 OTUs for the three specimens, yet pairwise similarity of communities from each PCR replicate was very high, in particular when using species abundance, and communities differed greatly from those of other beetle specimens. There was no bias resulting from the use of different primer tags or the inclusion of amplicons in separate libraries. The resulting ITS2 sequences correspond to diverse lineages of fungi, as identified with the Lowest Common Ancestor approach, and included known tree pathogens in Ophiostomaceae and Lestiomyceles. We conclude that Illumina MiSeq metabarcoding reliably captures fungal diversity associated with bark beetles, although numerous PCR replicates are recommended for an exhaustive sample. Direct PCR from beetle DNA extractions provides a rapid method for surveying the diversity of fungal species, their beetle ‘host’ species, associations with trees, biogeography and environmental variables.

Key words

Seolytinae, tree pathogens, Ophiostomaceae, ITS2, Illumina, primer tags
Introduction

Bark beetles (Scolytinae) are among the most destructive of forest pests, but much of the influence they exert is only made possible by the associations they form with fungi (Paine et al. 1997). Numerous fungal taxa have been isolated from these beetles and their galleries, including phytopathogenic species such as members of the genus *Ophiostoma*, and there is strong evidence to suggest that many of these associations represent symbiotic relationships (Six 2004). Fungal associates often provide important nutritional benefits to the beetle, and ambrosia beetles, which transport and cultivate symbiotic fungi as their primary food source, notably form a high proportion of the invasive scolytine species to have successfully established in the USA (Haack 2006) and Europe (Kirkendall & Faccoli 2010). Phytopathogenic fungi may assist in overwhelming the defenses of a host tree and facilitating a successful attack by the beetle (Paine et al. 1997), but pathogenicity may also provide the fungal associates with their own competitive advantage within the host tree (Six & Wingfield 2011). Current knowledge of this system is biased towards a relatively small subset of economically important species. However, reports are accruing of unforeseen levels of damage caused by both native and invasive bark beetle species across the globe (Hulcr & Dunn 2011; Humble & Allen 2006). The causes of these occurrences are not well understood but there is evidence to suggest they result from a combination of environmental change, active and passive movement of beetle hosts and fungal associates, and changes in pathogenicity (Hulcr & Dunn 2011; Ploetz et al. 2013; Six 2013). It is clear that we need to take a more holistic approach towards understanding these interactions in order to better predict how environmental change might affect a
native community and also how an invasive beetle might integrate into a new
ecosystem.

Recent developments in high throughput sequencing (HTS) have
revolutionized the study of fungal communities, revealing unexpectedly high
diversity that greatly surpasses counts from morphological studies and culturing
(Hawksworth 2012; Tedersoo et al. 2014). The great sequencing depth achievable
with recent HTS methods permits the characterization of entire microbial
communities comprising hundreds of species that are typically found in a single
environmental sample (Schmidt et al. 2013; Taylor et al. 2008; Tedersoo et al.
2014). A convenient genome region for studying this diversity is the Internal
Transcribed Spacer (ITS) of the rRNA locus that has emerged as a standard
‘barcode’ marker for fungal species detection and delimitation (Schoch et al. 2012),
and amplification from mixed samples (‘metabarcoding’) provides new possibilities
for studying the interactions of fungi and bark beetles. Only a single study
integrating HTS (using 454 pyrosequencing) and bark beetle associated fungal
communities currently exists. Kostovcik et al. (2015) investigated the mycobiota of
three ambrosia beetle species, with focus on the fungal communities isolated from
the mycangia, i.e. the specialized structures of the exoskeleton in many bark beetles
designed for the transport of fungal spores. Using this approach, they found several
dozen fungal OTUs associated with the mycangia of every specimen, which
represents many more species than could be isolated with traditional culturing
techniques. Additionally, the three beetle species included in this study displayed
divergent mycobiota, and diversity was further increased by geographical variation.
These discoveries suggest that the fungal diversity in this system is indeed very
large and that it cannot be sufficiently captured by conventional methods.
While Kostovcik et al.’s (2015) study focuses on the mycangia of ambrosia beetles, it is known that the spores and mycelia of pathogenic and symbiotic fungi can be transported affixed to beetle cuticles or within the insect gut (Harrington 2005). The full mycobiome of a beetle may be captured if DNA is extracted from the entire beetle specimen; fungal sequences can then be obtained by PCR amplification of ITS with fungus-specific primers that are believed to discriminate between beetle and fungal DNA templates (De Beeck et al. 2014). This approach also circumvents the need for tedious preparation of the mycangia and thus permits the rapid analysis of numerous samples. High sample numbers raise the prospect of establishing patterns that link fungal taxa with beetle (host) species or environmental conditions through frequency of association (Wright & Reeves 1992).

Processing of large numbers of specimens with HTS approaches also requires an efficient system for separation of these samples in the sequencing process. The addition of a sequence ‘tag’ at the amplification step is commonly used to attach a sample-specific nucleotide tag to each primer, so that resulting amplicons are identifiable in the post-sequencing bioinformatics analysis (Gloor et al. 2010). However, adding different tags to the primers may subtly chance the outcome of the PCR amplification and thus unique tags may cause tag-specific biases in community composition, as has been observed in bacterial communities (Berry et al. 2011; Faircloth & Glenn 2012).

The current study tests the feasibility of a HTS approach to the metabarcoding of fungal communities associated with bark beetles. First, we address the concern that fungus specific ITS primers are not sufficiently taxon specific to discriminate against the much greater amount of beetle template in the DNA extract. Secondly,
we test the reproducibility of individual PCRs in regard to species composition and abundance, in particular as differently tagged primers are used that may amplify slightly different sets of fungal template from the mixtures. The tests were based on 15 newly designed primer tags for the ITS3/4 primer pair widely used for amplification of the fungal ITS2 regions, followed by Illumia MiSeq sequencing of the resulting amplicons. We thus ask if metabarcoding of fungal communities can be achieved by amplicon sequencing directly from total DNA extracted from the insects, and how reproducible the recovery of complex fungal communities is when using differently tagged primers. The latter was addressed by repeated PCRs with differently tagged primer pairs on a test set of DNAs extracted from different beetle species and assessment of the resulting fungal communities. We find a huge diversity of fungi on each beetle specimen, and while separate PCRs result in somewhat different sets of fungal communities, these differences are not related to the variable primer tags or the subsequent step of library preparation in the DNA sequencing.

Materials and Methods

Samples used and experimental design

One specimen each of *Hylastes ater, Tomicus piniperda* and *Trypodendron lineatum* was obtained in ethanol-baited Lindgren Funnel traps (Table 1). For DNA extraction the thorax and abdomen were separated using sterilized forceps and both body parts were placed into a single well of a 96 well microtiter plate. A standard DNA extraction was carried out on each specimen following the BIOSPRINT 96
(Qiagen) tissue extraction protocol, by first adding 180 µl ATL buffer and 20 µl proteinase K to each specimen. Plates were incubated at 56°C for 10-12 hrs, centrifuged (500 g for 5 min), and 430 µl master mix (200 µl Buffer AL, 200 µl Isopropanol and 30 µl MagAttract Suspension G) was added to each well. Plates were then placed in the BIOSPRINT 96 workstation and the DNA was purified following the BIOSPRINT 96 DNA Tissue program. All extracted DNA was deposited in the Frozen Collection at the Natural History Museum, London.

**Tagged primer design and PCR**

We used the ITS3 and ITS4 primers (White *et al.* 1990) to amplify the internal transcribed spacer region 2 (ITS2) (Schoch *et al.*, 2012). The ITS2 region in fungi is 320 bp in length on average and lacks the insertions commonly found in ITS1, reducing length variation (Bazzicalupo *et al.* 2013; Martin & Rygiewicz 2005). These primers were modified by different 5’ identifier sequences of 5 to 9 base pairs (Supplementary Table 1) to generate 15 pairs of unique error-correcting tags, with an edit distance of ≥3 (i.e. the minimum number of indels and substitutions that might cause mis-identification of tags, (Degnan & Ochman 2012), as recommended by Qiu *et al.* (2003). Tags were length variable to avoid problems arising from low diversity sequences with Illumina sequencing protocols (Gloor *et al*. 2010). A 2-bp sequence was placed between the 5’ end of the primers and each tag to ensure a mismatch between the barcode and the template DNA (Zhou *et al*. 2011). Full length tagged primers were checked for hairpin formation, self-dimerization and cross-dimerization using EDITTAG (Faircloth & Glenn 2012).

PCR was conducted for the three specimens with each of the 15 newly designed primer pairs to test for variability in amplification success. Each specimen-
tag combination was performed in three replicates, bringing the total to 45 distinct samples per specimen. A negative control was created by performing the DNA extraction procedure on double distilled water, after which the extract was subjected to PCR with each of the 15 primer pairs. This brings the total number of amplifications to 150. For each of these, three PCRs were conducted at a different annealing temperature (50, 53, 56°C), before they were combined for sequencing. Combining multiple PCR products is a commonly used approach in microbial studies as using a range of annealing temperatures reduces primer binding bias and counteracts the stochasticity of individual PCRs (Schmidt et al. 2013).

The following cycling conditions were used for PCRs: initial denaturation at 94°C for 300 s, followed by 30 cycles of denaturation at 94 °C for 60 s, annealing at 50/53/56°C for 60 s, extension at 72 °C for 30 s, and a final extension at 72 °C for 120 s. The PCR mix contained 1 µl template DNA, 0.8 mM dNTPs, 5 U/µM BIOTAQ™ DNA polymerase, 4 mM MgCl2, and 0.5 µM of forward and reverse primers in a 40 µL reaction volume. PCR products were visualized using GelRed™ (Biotium) a 1% agarose gel. Purification of PCR products was done using DNA Clean & Concentrator™ (Zymo Research). To determine the concentration of double-stranded DNA present, cleaned PCR products were then quantified using a Qubit 2.0 Fluorometer (Invitrogen) with the Qubit dsDNA HS Assay Kit (Invitrogen).

After amplification samples were pooled into ten separate libraries, following a scheme by which each replicate (set of 3) PCR is placed in a different pool (Supplementary table 2). KAPA library kits were used to prepare samples for Illumina sequencing (SKU: KK8201; Kapa Biosystems). The library preparation followed the manufacturer’s instruction except that a step to isolate the large DNA fragments was omitted. To increase the diversity of the sequences and prevent
problems arising from the low diversity of amplicons, samples were split across two separate runs and 50% of each run contained genomic DNA from an unrelated study, but the total sequencing effort equated to one full run. Paired-end sequencing was conducted on an Illumina MiSeq sequencer with v.3 chemistry and 2 x 300 cycles.

**Bioinformatic Processing**

Reads from each library were quality checked with Fastqc (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/, v0.11.2), and PRINSEQ-lite 0.20.4 (Schmieder & Edwards 2011) was used to remove reads with an average Q score of <25 and to trim bases below Q25 from 3’ ends, no N’s were allowed, and only read lengths of over >150 bp were retained. Fastq-join (Aronesty 2011) was then used to merge forward and reverse reads if they matched over >50 bp at a minimum similarity of 99%.

Merged reads were renamed to mark the library from which they were derived and all libraries were then pooled together into a single Fastq file. A custom Python script (remove_multiprimer.py) (Balint et al. 2014) was used to remove primer artifacts from all reads. Sequences were demultiplexed based on the unique tags and library labels using Qiime’s split_libraries.py script allowing for one-base mismatches (Caporaso et al. 2010). Paired-end sequences with implausible forward and reverse primer combinations were discarded. Finally, sequences were passed through the ITSx software (Bengtsson-Palme et al. 2013), which matches input sequences to profile hidden Markov models (HMMs) (Eddy 1998) that represent kingdom-wide sequence alignments of fungal ITS. The program identifies and removes the ITS flanking regions, and also removes non-fungal sequences.
Operational taxonomic unit (OTU) clustering of ITS2 sequences was done with UPARSE (Edgar et al. 2011) which uses maximum parsimony models to compare each sequence to an existing OTU database. This method has been used successfully by a number of studies to study fungal communities (Adams et al. 2013; Balint et al. 2014; Liu et al. 2015). Briefly, the UPARSE pipeline entails the removal of singletons, i.e. completely unique sequences (Brown et al. 2015), followed by dereplication to retain a single representative of each ITS2 sequence, followed by clustering. The program outputs a fasta file containing a single representative for each OTU. These were then checked for chimeras with UCHIME (Edgar et al. 2011). OTUs were defined by a similarity cut-off of 97%. Finally, to check that the OTUs were truly fungal in origin, a taxonomic approach was utilized. BLASTn searches (BLAST 2.2.28+; (Altschul et al. 1990) of OTU representative sequences were conducted against the NCBI nucleotide database, and identities of each OTU representative were established with the lowest common ancestor (LCA) method in MEGAN (Huson et al. 2011). OTUs identified as non-fungal with the approach were excluded from the analysis.

**Data Preparation and Statistical Analysis**

An OTU abundance table (beetle individual x fungal OTUs) was produced with UCHIME, whereby all ITS2 sequences are mapped back against the OTU representatives at a similarity of 99% and sequence counts are used as a measure of OTU abundance. Only OTUs supported by >10 sequences were retained. OTU accumulation curves were generated for replicates of each specimen to assess whether the sequencing effort was sufficient to capture the full fungal community. This was done using the ‘rarecurve’ function in the R package ‘vegan’ (Oksanen et
The saturation of curves was assessed using the ‘estimateR’ function (Supplementary Table 2).

Permutational multivariate ANOVA (PERMANOVA) was conducted to test the effect of tag identity, specimen identity and library membership on the fungal community obtained from each replicate. PERMANOVA is a non-parametric, resemblance-based permutation method for analyzing compositional differences in communities belonging to different groups (Anderson 2006). The PERMANOVAs used pairwise dissimilarity matrices between all communities as an input, to test whether the dissimilarities between pairs of communities from all groups are significantly different from the pairwise dissimilarities of communities from the same group.

Two different pairwise dissimilarity matrices were analyzed by PERMANOVA, one based upon sequence read abundance (used here as a proxy for OTU abundance per replicate), and one based upon presence-absence (incidence). The fungal OTU abundance table was therefore used to compute two separate compositional dissimilarity matrices, each based on a different pairwise community dissimilarity index. We used the Sørensen index, which uses the incidence data, and the Bray-Curtis index, which uses abundance data.

Results

Read processing and OTU delimitation

Over 27 million paired end reads were obtained from one full Illumina MiSeq run. A total of 16.6 million (60.8%) reads were removed at various stages during the bioinformatics processing, leaving 10.7 million for further analysis (Fig. 1). The
greatest proportion of reads (25.9% across all libraries) was lost during quality filtering and again in the final filtering step against the fungal ITS database using ITSx (18.9%) (Fig. 1, Supplementary Fig. S1). Primer artefacts and complications with merging of paired reads were minimal, while unexpected combinations of forward and reverse primer tags also resulted in substantial losses of reads. The high proportion of non-fungal ITS sequences without match to the reference Hidden Markov model was evident also from the Lowest Common Ancestor (LCA) analysis (Huson et al. 2011) that included numerous sequences not assignable to any taxon, or assigned to plants and beetles (specifically Cucujiformia, consistent with the bark beetle ‘hosts’), which presumably were co-amplified by the fungus-specific primers. These sequences were no longer seen after the ITSx filtering step, except for a single OTU that was classified as non-fungal in the LCA analysis and removed.

The number of post-filtering sequences differed greatly among specimens, with H. ater (specimen 1) showing the highest numbers, followed by T. lineatum (specimen 3) and T. piniperda (specimen 2) (Table 1). Sequence numbers for each of the three specimens were largely uniform within a sample (ANOVA: $F_{(2,131)} = 137.5; p < 0.0001$; Supplementary Table 2), indicating high reproducibility of the PCR within replicates of the same beetle DNA extract.

The assembled sequences were grouped into a total of 1760 OTUs using the Uparse software, of which 612 OTUs were removed because their associations were supported by fewer than 10 ITS2 sequences. 19 OTUs were also present in the negative control and were removed from all other samples where they occurred, and one OTU was removed because its closest Blast match was to an angiosperm (Theobroma cacao). This left 1128 OTUs for statistical analysis. The number of
OTUs derived from these sequences for the three specimens differed greatly, ranging from 229 OTUs on average (range=149-323; s.d.=40) for *H. ater*, 54 OTUs (range=24-87; SD=16) for *T. piniperda*, and 31 (range=9-55; SD=9) OTUs for *T. lineatum*. Thus, variation in OTU richness was highly correlated with the beetle specimen (ANOVA of log OTU: $F_{(2,131)} = 589.6; p < 0.0001$), although not to the same degree as the number of ITS2 sequences obtained for these specimens (Table 1). The OTUs could be assigned taxonomically to various groups of fungi including major lineages of Ascomycota and Basidiomycota using the LCA method (Fig. 2). In total, in all PCR replicates combined there were 727 OTUs found in *H. ater* (specimen one), 435 OTUs in *T. piniperda* (specimen two) and 294 OTUs in *T. lineatum* (specimen three). OTUs differed greatly in their abundance (number of sequences associated with an OTU), following a typical ‘hollow curve’ distribution of species abundance, with a few highly abundant entities and a long tail of rare ones. This abundance profile was highly reproducible among the 15 PCR replicates based on the SD on read counts for the individual OTUs (Fig. 3).

**Reproducibility of PCR**

A PERMANOVA analysis was conducted on the measures of dissimilarity of the communities based on incidence (presence/absence) and relative abundance of OTUs. In both tests, the set of OTUs derived from the PCR replicates carried out on a single DNA extraction was highly similar to each other, indicating high reproducibility of the PCR in the replicates, but differed greatly from the OTU set obtained from another specimen. Specimen identity explained a high proportion of the data for OTU incidence (PERMANOVA: Pseudo$F = 84.21$, $R^2 = 0.56$, $p=0.001$), and this proportion is even higher when taking into account the OTU
abundance (PseudoF = 1521.50, $R^2 = 0.96$, $p=0.001$) (Fig. 4). The latter showed that the OTU set obtained from a given specimen is virtually identical for each PCR replicate (Fig. 4b).

PERMANOVAs testing the effect of primer identity on the fungal communities revealed that the primer tag used for each sample had no significant effect (PseudoF = 0.49, $R^2 = 0.05$, $p = 1$; Fig. 3c). The effect of library was also found to be non-significant (PseudoF = 1.27, $R^2 = 0.01$, $p=0.235$). These results indicate that the choice of tag has only minimal influence on the recovered OTU set, while the library preparation was highly reproducible.

### Completeness of OTU sequencing

Species accumulation curves conducted on OTUs with >10 reads were fully saturated at the sequencing depth used, even for the most diverse specimen one (Fig. 5), indicating that diversity estimates would not increase with deeper sequencing. However, the replicates differed substantially in the total number of OTUs produced, indicating that PCR success of each replicate was highly variable. To account for the great variation in read numbers among replicates for comparisons of species richness in the replicates, we resampled the reads according the lowest number of sequences of any replicate of 1832 reads. Accumulation curves of the resampled reads failed to reach the full set of OTUs in two of three cases (Supplementary Fig. S2). Rarefaction based on that read number found that OTU richness was significantly reduced for specimen one ($t = 26.40$, df = 58.84, $p < 0.0001$), but not for specimen two or three ($t = 0.16$, df = 85.88, $p < 0.87$; $t = 1.31$, df = 87.12, $p < 0.19$, respectively) (Supplementary Fig. S3).
**Discussion**

**Reproducibility of community composition in metabarcoding**

Metabarcoding is changing the way ecological communities are studied, as it is becoming possible to assess the full extent of species diversity in complex assemblages due to the great sequencing depth and high read accuracy of the Illumina technology. The current study shows that metabarcoding of fungal communities associated with bark beetles can be performed reliably straight from DNA extracted from the beetle specimen without prior dissection of fungus-bearing structures. Metabarcoding conducted in this way detected a much greater species richness than recognised previously with culturing (Giordano *et al.* 2012; Kolarik & Jankowiak 2013) and DNA sequencing from mycangia (Kostovcik *et al.* 2015). However, the methodology is dependent on the accuracy of the PCR that potentially introduces biases in the representation of species and their abundances in a natural community. Different primer tags may add to this bias, which raises the question about the reproducibility and comparability of samples using this approach.

Variability of PCR amplifications is not usually tested, and it is often assumed that pooling of a few PCR replicates at different annealing temperatures is sufficient to overcome the stochastic variability of the amplification process (e.g. (Schmidt *et al.* 2013). We find that the 45 PCRs from a given sample specimen produce fungal communities that based on pairwise community similarity differ very little in species composition and in particular abundance (Fig. 3). There is also no effect of the choice of tag sequence on the outcome of the inferred community composition. In addition, we find that the fungal OTUs are recovered in similar abundance in the replicates (Fig. 3), further demonstrating the reproducibility of the PCRs.
Despite the apparent consistency among replicates in the NMDS analysis and rank abundance curves, there was a surprising level of variation in the total number of OTUs detected among the PCR replicates (Supplementary Table X), which is also evident in the species accumulation curves that consistently reach an asymptote for each PCR replicate, but at different levels (Fig. 5, Supplementary Table 3). As a proportion of the average number of species detected, the effect was strongest in the beetle specimen with the lowest number of fungal species (specimen three, *T. lineatum*). This type of variation is apparently not captured by the NMDS on pairwise community similarity, which would lead to the expectation that species numbers are also largely uniform. However, as this is not the case, the variation is likely affecting the rare species (tail of the hollow-curve distribution). Detailed investigation revealed that the number of sequences supporting high abundance OTUs is consistent among replicates, indicating that the method is able to reliably find a core of high abundance OTUs, but that low abundance OTUs tend to be patchily distributed and inconsistent between PCR replicates (data not shown). It appears that these low abundance OTUs are driving the variation in OTU incidence but that their influence is downweighted by the relative abundance analysis. Thus the mycobiome of specimen 3 (*T. domesticum*) displayed the lowest diversity and the greatest variability in community composition. Conversely, specimen 1 (*H. ater*) had the highest diversity and showed the lowest variability in OTU numbers. The observation of higher variation in OTU recovery for relatively small communities and rare species may indicate stochastic effects of the PCR amplification from mixed species communities and the chance amplification of different fungal templates if they are present in low abundance. Similar evidence is accumulating for the idea that the concentration of the original DNA extract is
primarily responsible for apparently stochastic variability in microbial communities (Kennedy et al. 2014; Salter et al. 2014). Our findings are consistent with this suggestion as low fungal diversity is likely to translate to lower DNA concentrations. Results also suggest that the fungal OTUs shared between different specimens display significantly different read abundance between the specimens, further supporting the power of these methods for reproducible and accurate determination of fungal community composition, at least for the common OTUs. A caveat is that this type of reproducibility may only apply to the PCR step, but does not necessarily reflect the true diversity of the natural community whose amplification may be skewed in a systematic way (Elbrecht & Leese 2015). However, these PCR induced biases are likely to be smoothed out by conducting multiple replicates that each contribute various components of the community (Fig. 5). Hence, library construction and sequencing in metabarcoding should be conducted on a large number of PCR replicates.

**Data filtering and effect on species composition**

Prior to the community analysis extensive data filtering is needed that potentially causes a skew in species composition and specifically may affect rare species. The initial bioinformatics data treatment resulted in removal of totally ~61% of reads. Following widely used practice, as a precaution we also removed all OTUs represented by <10 sequences, thus removing several hundred OTUs, and inevitably bias against the detection of rare species. For the bioinformatics protocols we essentially followed the steps of Bálint et al.’s (2014) study of soil fungi, albeit using different software. This approach removed a similar proportion overall, but the sequence reads were lost at slightly different stages of the filtering.
process (Supplementary Fig. 1). Specifically, in the current study more data were removed in the first step of quality filtering eliminating reads with ambiguous base calls based on Phred quality scores. This error rate depends on the specific Illumina run and possibly the quality of the template DNA (in this case obtained from specimens in ethanol-filled traps emptied only once per week). Our procedure was more resilient at the stage of pairing of reads and primer base calls. This step was carried out with the Fastq-join software that uses error probabilities to accept the merged paired-end reads, and produced hardly any incorrect combinations.

However, the subsequent step of demultiplexing removed a large portion of reads, as implausible combinations of forward and reverse tags are removed, but the loss is lower than in the Bálint et al.’s (2014) study. This high proportion of mismatched amplicons demonstrates the need for tagging of primers on both ends of the amplicon, as used in our approach, rather than reliance on the paired-end combination, to establish any incidence of tag switching and chimera formation that would confound the subsequent OTU formation.

Finally, many sequences were removed at the ITSx stage that eliminates non-fungal sequences, which was much higher when compared to Balint et al. (2014) (Supplementary Fig. 1). To a small extent this was due to co-amplification from the beetle DNA templates, present in high concentrations in the DNA extracts. The LCA analysis confirmed the co-amplification of Coleoptera (Fig. 2), but showed this to be a minor component besides sequences assigned to angiosperms and other unidentified sequences, and the ITSx procedure cleanly removed these. We conclude that co-amplification of the beetle ‘host’ with the ITS3/4 primers is a manageable problem, while the primers pick up a great phylogenetic diversity of fungi, including the main lineages of Ascomycota and Basidiomycota. All current
universal primer pairs for the amplification of fungal ITS exhibit some form of bias (Schoch et al. 2012) and those with lower levels of bias are prone to the amplification of non-fungal ITS as those from plants (Martin & Rygiewicz 2005). Standard ITS primers have been shown to exhibit taxonomic preferences (Elbrecht & Leese 2015; Tedersoo et al. 2014). Thus, a trade-off between primer-specificity and primer-bias towards particular groups of fungi seems unavoidable and, if used for amplification of insect-associated fungi, any alternative to the ITS3/4 or amplification with multiple primer sets also would need to be tested for their specificity to insects.

The power of ITS2 for determining community composition

The ITS2 locus used here comprises only around 300 bp, which limits the power of this locus for species delimitation and phylogenetic placement, together with ITS length variation among lineages and intra-individual variation due to the tandem-repeat genomic structure of the rRNA locus (Hillis & Davis 1988). In addition, concerns have been raised over the biological significance of entities defined by a single marker and simplistic cut-off values for defining OTUs (Yamamoto & Bibby 2014). Ultimately, better-validated reference sequences are required for a more secure assignment to taxonomic groups, using macroscopic morphological characters and culturing, where possible, to link these sequences to species and clades and to their characters. Yet, many validated taxonomic records are now available for the ITS region and the marker has high potential for an unbiased characterization of phylogenetically diverse mixtures because of the conserved primer recognition sequences for amplification bracketing variable regions for species-level diagnostics.
The LCA analysis provided identifications generally consistent with fungal communities expected to be associated with bark beetles and woody plants, including some important tree pathogens such as Ophiostomaceae and Lestiomyceles including *Chalara*, the agent of ash dieback. Thus, from a pragmatic perspective the ITS region is an appropriate species marker and was chosen as the ‘formal barcode’ for the fungi (Schoch *et al.* 2012), which then provides the framework for metabarcoding of communities. In addition, the link to an external taxonomic identification is not strictly necessary if applied to a given set of samples for comparisons within a given study seeking to establish internal pattern within a sample set. The ITS2 region has been the marker of choice for analogous studies of complex fungal communities of the soil and has been successful to discover a huge uncharacterized diversity of fungi and to infer macroecological patterns (Tedersoo *et al.* 2015; Tedersoo *et al.* 2014). Sequencing straight from insect samples provides a rapid way of assessing large-scale patterns of diversity for fungi associated with wood-boring beetles. The current study demonstrates a huge diversity of putative species of fungi. Their correlations with ecological parameters, host species and biogeographic distributions can be assessed readily by metabarcoding of existing and newly collected DNA extracts from bark beetles.

**Acknowledgements**

This work was funded by a NERC PhD studentship and the NHM Biodiversity Initiative.

**References**

Adams RI, Miletto M, Taylor JW, Bruns TD (2013) The diversity and distribution of fungi on residential surfaces. *PLoS One* 8.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic Local Alignment Search Tool. *Journal of Molecular Biology* 215, 403-410.
Anderson MJ (2006) Distance-based tests for homogeneity of multivariate
dispersions. Biometrics 62, 245-253.
Aronesty E (2011) ea-utils: Command-line tools for processing biological sequencing
data http://code.google.com/p/ea-utils.
Balint M, Schmidt PA, Sharma R, Thines M, Schmitt I (2014) An Illumina
metabarcoding pipeline for fungi. Ecology and Evolution 4, 2642-2653.
Bazzicalupo AL, Balint M, Schmitt I (2013) Comparison of ITS1 and ITS2 rDNA in
454 sequencing of hyperdiverse fungal communities. Fungal Ecology 6, 102-
109.
Bengtsson-Palme J, Ryberg M, Hartmann M, et al. (2013) Improved software
detection and extraction of ITS1 and ITS2 from ribosomal ITS sequences of
fungi and other eukaryotes for analysis of environmental sequencing data.
Methods in Ecology and Evolution 4, 914-919.
Berry D, Ben Maftoudh K, Wagner M, Loy A (2011) Barcoded primers used in
multiplex amplicon pyrosequencing bias amplification. Applied and
Environmental Microbiology 77, 7846-7849.
Brown SP, Veach AM, Rigdon-Huss AR, et al. (2015) Scraping the bottom of the
barrel: are rare high throughput sequences artifacts? Fungal Ecology 13, 221-
225.
Caporaso JG, Kuczynski J, Stombaugh J, et al. (2010) QIIME allows analysis of high-
throughput community sequencing data. Nature Methods 7, 335-336.
De Beeck MO, Lievens B, Busschaert P, et al. (2014) Comparison and validation of
some ITS primer pairs useful for fungal metabarcoding studies. Plos One 9.
Degnan PH, Ochman H (2012) Illumina-based analysis of microbial community
diversity. ISME Journal 6, 183-194.
Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14, 755-763.
Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves
sensitivity and speed of chimera detection. Bioinformatics 27, 2194-2200.
Elbrecht V, Leese F (2015) Can DNA-based ecosystem assessments quantify species
qundance? Testing primer bias and biomass-sequence relationships with an
innovative metabarcoding protocol. PLoS One 10.
Faircloth BC, Glenn TC (2012) Not all sequence tags are created equal: designing and
validating sequence identification tags robust to indels. Plos One 7.
Giordano L, Garbelotto M, Nicolotti G, Gonthier P (2012) Characterization of fungal
communities associated with the bark beetle Ips typographus varies depending
on detection method, location, and beetle population levels. Mycological
Progress 12, 127–140.
Gloor GB, Hummelen R, Macklaim JM, et al. (2010) Microbiome profiling by
Illumina sequencing of combinatorial sequence-tagged PCR products. PLoS
One 5.
Haack RA (2006) Exotic bark- and wood-boring Coleoptera in the United States:
recent establishments and interceptions. Canadian Journal of Forest Research
36, 269-288.
Harrington T (2005) Ecology and evolution of mycophagous bark beetles and their
fungal partners. In: Ecological and evolutionary advances in insect-fungal
associations (eds. Vega FE, Blackwell M), pp. 257–291. Oxford University
Press.
Hawksworth DL (2012) Global species numbers of fungi: are tropical studies and
molecular approaches contributing to a more robust estimate? Biodiversity and
Conservation 21, 2425-2433.
Hillis DM, Davis SK (1988) Ribosomal DNA: Intraspecific polymorphisms, concerted evolution, and phylogeny reconstruction. *Systematic Zoology* **37**.

Hulcr J, Dunn RR (2011) The sudden emergence of pathogenicity in insect-fungus symbioses threatens naive forest ecosystems. *Proceedings of the Royal Society B-Biological Sciences* **278**, 2866-2873.

Humble LM, Allen EA (2006) Forest biosecurity: alien invasive species and vectored organisms. *Canadian Journal of Plant Pathology* **28**, S256-S269.

Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21**, 1552-1560.

Kennedy K, Hall MW, Lynch MDJ, Moreno-Hagelsieb G, Neufeld JD (2014) Evaluating bias of Illumina-based bacterial 16S rRNA gene profiles. *Applied and Environmental Microbiology* **80**, 5717-5722.

Kirkendall LR, Faccoli M (2010) Bark beetles and pinhole borers (Curculionidae, Scolytinae, Platypodinae) alien to Europe. *Zookeys*, 227-251.

Kolarik M, Jankowiak R (2013) Vector affinity and diversity of *Geosmithia* fungi living on subcortical insects inhabiting Pinaceae species in Central and Northeastern Europe. *Microbial Ecology* **66**, 682-700.

Kostovcik M, Bateman CC, Kolarik M, et al. (2015) The ambrosia symbiosis is specific in some species and promiscuous in others: evidence from community pyrosequencing. *ISME Journal* **9**, 126-138.

Liu T, Li CM, Han YL, et al. (2015) Highly diversified fungi are associated with the achorophyllous orchid *Gastrodia flavilabella*. *BMC Genomics* **16**.

Martin KJ, Rygiewicz PT (2005) Fungal-specific PCR primers developed for analysis of the ITS region of environmental DNA extracts. *BMC Microbiology* **5**.

Oksanen J, Blanchet G, Kindt R, et al. (2010) *vegan*: Community Ecology Package. *R* package version 1.17-2 available at [http://cran.r-project.org/](http://cran.r-project.org/).

Paine TD, Raffa KF, Harrington TC (1997) Interactions among scolytid bark beetles, their associated fungi, and live host conifers. *Annual Review of Entomology* **42**, 179-206.

Ploetz RC, Hulcr J, Wingfield MJ, de Beer Zw (2013) Destructive tree diseases associated with ambrosia and bark beetles: black swan events in tree pathology? *Plant Disease* **97**, 856-872.

Salter SJ, Cox MJ, Turek EM, et al. (2014) Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biology* **12**.

Schmidt PA, Balint M, Greshake B, et al. (2013) Illumina metabarcoding of a soil fungal community. *Soil Biology and Biochemistry* **65**, 128-132.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**, 863-864.

Schoch CL, Seifert KA, Huhndorf S, et al. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 6241-6246.

Six DL (2004) Bark beetle-fungus symbioses. In: *Insect Symbiosis* (eds. Bourtzis K, Miller T), pp. 97-114. CRC Press, FL, USA.

Six DL (2013) The bark beetle holobiont: Why microbes matter. *Journal of Chemical Ecology* **39**, 989-1002.

Six DL, Wingfield MJ (2011) The role of phytopathogenicity in bark beetle-fungus symbioses: A challenge to the classic paradigm. In: *Annual Review of Entomology* (eds. Berenbaum MR, Cardé RT, Robinson GE), pp. 255-272.
Taylor DL, Booth MG, McFarland JW, et al. (2008) Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach. *Molecular Ecology Resources* 8, 742-752.

Tedersoo L, Anslan S, Bahram M, et al. (2015) Shotgun metagenomes and multiple primer pair-barcode combinations of amplicons reveal biases in metabarcoding analyses of fungi. *MycoKeys*, 1-43.

Tedersoo L, Bahram M, Polme S, et al. (2014) Global diversity and geography of soil fungi. *Science* 346, 1078-+.

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 J, White TJ), pp. 315-322. Academic Press, San Diego.

Wright DH, Reeves JH (1992) On the meaning and measurement of nestedness of species assemblages. *Oecologia* 92, 416-428.

Yamamoto N, Bibby K (2014) Clustering of fungal community internal transcribed spacer sequence data obscures taxonomic diversity. *Environmental Microbiology* 16, 2491-2500.

Zhou JZ, Wu LY, Deng Y, et al. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *ISME Journal* 5, 1303-1313.

**Data accessibility**

Final OTU sequence assembly: online Supporting Information

**Author contribution**

KM and APV: designed the project, wrote the paper; KM and DJGI: designed sampling and collected specimens; KM and KH: generated and analysed data
Table 1. Average number of sequences and OTUs for each beetle specimen (standard deviation in brackets) before and after rarefaction (at 1832 sequences).

| Specimen                  | BMNH number | Number sequences (average) | Number OTUs (SD) | Number OTUs rarefied (SD) |
|---------------------------|--------------|----------------------------|------------------|---------------------------|
| *Hylastes ater* (specimen 1) | 1047606      | 196,935 (101,767)          | 229 (40)         | 107 (8)                   |
| *Tomicus piniperda* (specimen 2) | 1047128      | 6,031 (3,059)              | 54 (16)          | 53 (15)                   |
| *Trypodendron lineatum* (specimen 3) | 1046457      | 28,539 (13,716)            | 31 (9)           | 28 (8)                    |
Legend to Figures

Fig. 1. The proportion of reads removed by the bioinformatics pipeline at each processing step for the current study and the study of Bálint et al. (2014).

Fig. 2. Lowest common ancestor analysis using MEGAN before ITSx treatment.

Fig. 3. Rank abundance curve for the number of fungal OTUs showing the mean log number of reads associated with each OUT, separately for each of the three specimens used in this study. Error bars represent standard error of the mean for each of 45 replicates (15 separate tags and three PCR replicates each).

Fig. 4. NMDS plots showing each point as the fungal community of an individual replicate based on (a and c) Sørensen (incidence) pairwise dissimilarities, (b) Bray-Curtis (abundance) pairwise dissimilarities. For (a) and (b) colours represent specimen of origin, grey polygons showing specimen groupings. For c colours represent the primer tag used and grey polygons show primer groupings.

Fig. 5. OTU accumulation curves for each of 45 PCR replicates in three specimens. Note the different scale of the horizontal axis in the three panels.
Supplementary Fig. S1. Percentage of initial reads remaining after each processing step for each library in this study (Library 1 – 10) and for pooled samples from Bálint et al. (2014).

Supplementary Fig. S2. Log fungal OTU richness differences for unrarified and rarified matrices for each specimen. Only for Specimen 1 there was a significant difference between rarefied and unrarefied treatments ($p < 0.0001$).

Supplementary Fig. S3. Rarified OTU accumulation curves for each sample demonstrating that rarefaction prevents certain samples from reaching an asymptote a) for beetle specimen 1, b) for beetle specimen 2 and c) for beetle specimen 3.
Supplementary figure. Megan generated classification of pre-ITSx merged reads based upon BLASTn searches against GenBank nucleotide (E-value ≤ 1 e-5).
Specimen 1: *Hylastes ater*

Specimen 2: *Tomicus piniperda*

Specimen 3: *Trypodendron domesticum*
(a) Specimens - incidence

(b) Specimens - relative abundance

(c) Primer - incidence
Specimen 1 - *Hylastes ater*

Specimen 2 - *Tomicus piniperda*

Specimen 3 - *Trypodendron domesticum*