

available at www.sciencedirect.comjournal homepage: www.elsevier.com/locate/funeco

Methodological Advances

Identifying wood-inhabiting fungi with 454 sequencing – what is the probability that BLAST gives the correct species?

Otso OVASKAINEN^{a,*}, Jussi NOKSO-KOIVISTO^a, Jenni HOTTOLA^a, Tiina RAJALA^b,
Taina PENNANEN^b, Heini ALI-KOVERO^a, Otto MIETTINEN^c, Petri OINONEN^c,
Petri AUVINEN^d, Lars PAULIN^d, Karl-Henrik LARSSON^e, Raisa MÄKIPÄÄ^b

^aDepartment of Biosciences, University of Helsinki, P.O. Box 65 (Viikinkaari 1), 00014 University of Helsinki, Finland

^bVantaa Research Unit, Finnish Forest Research Institute, Finland

^cFinnish Museum of Natural History, University of Helsinki, Finland

^dInstitute of Biotechnology, University of Helsinki, Finland

^eNatural History Museum, University of Oslo, P.O. Box 1172, Blindern, 0318 Oslo, Norway

ARTICLE INFO

Article history:

Received 28 August 2009

Revision received 5 January 2010

Accepted 5 January 2010

Available online 17 March 2010

Corresponding editor: Lynne Boddy

Keywords:

BLAST

Dead wood

DGGE

Fruit-body

Fungi

Logistic regression

Molecular identification

454 Sequencing

Species diversity

ITS region

ABSTRACT

When comparing environmental sequences with fully identified reference sequences, a common practice has been to rely on threshold values for sequence similarity. We develop a modelling approach that utilizes the self-consistency of the reference database to transfer sequence similarity to the probability of correct identification to a given taxonomic level. We model separately the probability of the focal species being in the reference database, and the probability that the best BLAST hit is correct, conditional on the species being in the reference database. We illustrate our approach in the context of 454 sequencing data on dead wood-inhabiting fungi, with a reference database containing 2262 ITS-sequences of 1145 species. We compare the species communities observed by 454 pyrosequencing, DGGE fingerprinting and fruit-body inventory. High-throughput sequencing calls for automated species identification with adequate assessment of identification error. Our results highlight that this is possible if a high-quality reference database with broad coverage is available.

© 2010 Elsevier Ltd and The British Mycological Society. All rights reserved.

Introduction

Dead wood-inhabiting aphyllorhous fungi (Basidiomycota) form a diverse, ecologically important and taxonomically well

known group of species (e.g., Hjortstam *et al.* 1987; Niemelä 2005; Kotiranta *et al.* 2009). In Finland and Sweden, one fourth of aphyllorhous species (ca. 1100 species, most of which are corticioids), and one third of polypore species

* Corresponding author. Tel.: +358 9 19157924.

E-mail address: otso.ovaskainen@helsinki.fi (O. Ovaskainen).

1754-5048/\$ – see front matter © 2010 Elsevier Ltd and The British Mycological Society. All rights reserved.

doi:10.1016/j.funeco.2010.01.001

(poroid Aphyllophorales; ca. 260 species) have been classified as threatened or near threatened (Rassi *et al.* 2001, Gärdenfors 2005), mainly because intensive forest management has reduced the availability of dead wood. So far, almost all field studies on wood-inhabiting fungi have been based on time-consuming and costly fruit body inventories. Due to temporal variation in occurrence of fruit bodies, a substantial proportion of the fungal community is not visible at a given point in time (Berglund *et al.* 2005; Porter *et al.* 2008), especially for species with annual fruit bodies which are not necessarily formed every year. The more readily observable perennial fruit bodies offer only a partial proxy for the diversity of the species with annual fruit bodies (Halme *et al.* 2009). Gaining a complete picture of species' distribution and prevalence would thus require studying also the mycelial stage inside dead wood. The aim of this paper is to test the feasibility of high-throughput sequencing and molecular identification to census species from environmental saw-dust samples.

The internal transcribed spacer (ITS) region of rDNA has been widely used for molecular identification of fungi (Köljalg *et al.* 2005; Peay *et al.* 2008; Ryberg *et al.* 2008), largely because of its relatively low level of intraspecific variation compared to high level of interspecific variation (Nilsson *et al.* 2008). Methods utilizing the ITS for fungal community identification from environmental samples include RFLP (restriction fragment length polymorphism, Karen *et al.* 1997; Johannesson & Stenlid 1999; Jasalavich *et al.* 2000), T-RFLP (terminal restriction fragment length polymorphism, Allmér *et al.* 2006), DGGE (denaturing gradient gel electrophoresis, Vainio & Hantula 2000), TGGE (temperature gradient gel electrophoresis, Kulhankova *et al.* 2006) and cloning combined with sequencing (Menkis *et al.* 2005; Kwasna *et al.* 2008). These methods have been successfully applied for describing patterns of fungal diversity and community turnover, but identifying all the individual species present in a given sample remains a challenge.

Recent advances in next generation sequencing platforms are making it increasingly feasible to obtain a high number of sequences with reasonable cost and workload, with first applications of 454 pyrosequencing to fungal communities starting to emerge (Buée *et al.* 2009; Jumpponen & Jones 2009). The amount of data generated by high-throughput sequencing, however, brings forth challenges for the analysis phase, and automated methods for data management and analysis become a necessity (Markowitz *et al.* 2006; Raes *et al.* 2007). Software that has been developed to aid in species identification include MEGAN (Huson *et al.* 2007) which classifies DNA fragments based on a lowest common ancestor algorithm, and CARMA (Krause *et al.* 2008) which utilizes the Pfam database (Finn *et al.* 2008) for placing the sequences in phylogenetic trees.

One of the most fundamental questions in algorithmic species identification concerns the reliability of the results. In the context of sequence similarity comparison, a common practice has been to set a threshold level (e.g., 97% similarity, Kwasna *et al.* 2008) below which the identifications are considered unreliable. Intraspecific variability, however, differs substantially among different groups of fungi (Nilsson *et al.* 2008; Hughes *et al.* 2009), and consequently taxon-specific threshold levels have been applied (Menkis *et al.* 2005). It is clear that even if sequence similarity is above the threshold, the identification is not 100% reliable. Conversely, if sequence

similarity is below the threshold, it is still possible that the query sequence and the reference sequence represent the same taxonomical unit. We thus argue that a less context-dependent approach would be to estimate the probability that the proposed identification is correct.

Probabilistic models have been developed for classification and identification problems in bacterial taxonomy (for review, see Gyllenberg & Koski 2001), where traditional approaches based on morphology are not feasible. However, probabilistic methods have seldom been applied to the molecular identification of higher taxa. In this paper, we take a step towards systematic and automated assessment of the reliability of fungal identifications from sequence data. Our method is tailored for environmental samples that represent a well defined species community for which a high coverage reference database is available, and it thus complements the general-purpose tools implemented in e.g., CARMA and MEGAN. We utilize the internal consistency of the reference database to model the probability of correct identification to a given taxonomical level, such as species or genus. The application of our approach requires the availability of multiple reference sequences for some of the species included in the database. We develop our approach in the context of the sequence similarity software suite NCBI-BLAST (Altschul *et al.* 1997), but it could be applied to any other algorithm that scores the similarity of the query sequence against sequences included in a reference database.

We illustrate our approach using spruce-associated wood-inhabiting fungi in Fennoscandia as a case study. We compiled a reference database for this target group with 2262 ITS-sequences of 1145 distinct species. We apply the statistical methods developed here to small-scale pilot data obtained by a combination of 454 pyrosequencing, DGGE analysis, and fruit-body inventory. While our main focus is in the reliability of species identification, we also examine what kind of sampling design would adequately represent the species diversity in dead wood.

Materials and methods

Reference database

We constructed a database of fully identified ITS-sequences of wood-inhabiting fungi, in particular the species associated with Norway spruce (*Picea abies*). The database, referred to as SAF (spruce-associated fungi), includes 443 species of polypores, corticioids, agarics, and hydroid and stereoid fungi occurring in Northern Europe (Knudsen & Vesterholt 2008). The species (see [Supplementary material](#) for a list of the species included) are represented by 615 sequences, multiple sequences per species being present especially for the common species. We combined the SAF database with the UNITE database (Köljalg *et al.* 2005) containing 1647 ITS-sequences of 711 species of ectomycorrhizal asco- and basidiomycetes. The two databases share nine species, thus the total number of species is 1145. The original number of sequences was larger than the 2262 sequences included here, but we excluded all sequences with potential quality problems. The taxonomic and nomenclatural aspects of the

database were streamlined following Hansen & Knudsen (1997), Knudsen & Vesterholt (2008) and Niemelä (2005), except for the genera *Cortinarius*, *Laccaria* and *Amanita*, for which the species-level nomenclature and taxonomy remain in a state of flux (e.g., Froslev et al. 2005; Geml et al. 2006).

Modelling the probability of correct identification

We applied the NCBI-BLAST algorithm to match environmental ITS1-sequences (termed query sequences) to the reference sequences contained in the SAF-UNITE database. Our main aim was to identify the species represented by the query sequences, and to assess the reliability of the identifications either to the species level or to the genus level. However, the BLAST algorithm simply provides the sequence similarity between the query sequence and the reference sequences, and thus does not directly indicate how reliably the best matching reference sequence belongs to the same species (or genus) than the query sequence. We develop a simple statistical model that translates sequence similarity to the probability of correct identification, and parameterize the model by testing how frequently the BLAST search correctly identifies query sequences of known species.

Assume that a query sequence s representing an unknown species is compared for similarity using BLAST against a reference database R , returning as the best BLAST result a reference sequence $i \in R$ and an associated quality index q . By “correct identification” we mean that the query sequence s and the best matching reference sequence i represent the same taxonomic unit. We may consider if the identification is correct to any taxonomic level, such as species or genus. The quality index q is the explanatory variable that we use in logistic regression to model the probability of correct identification. The quality index q can include any information that is useful in assessing the likelihood of correct identification, such as the high-scoring segment pairs (HSP) score or the length of the query sequence. To transform the quality index q into a probability of correct identification, we decompose the probability that the BLAST result is correct as

$$P(s = i|q) = P(s = i|q, s \in R)P(s \in R|q). \quad (1)$$

Here $P(s = i|q, s \in R)$ is the probability that the BLAST result is correct, conditional on the quality index q and the assumption that the unknown species s is in the database, whereas the second component $P(s \in R|q)$ represents the probability that the unknown species s is in the database, conditional on the quality index q .

Both of the above probabilities can be estimated if multiple sequences are available for the species included in the reference database. We first simplified the combined SAF-UNITE reference database so that it contained only one sequence representing each of the 1145 species, and then compared the remaining 1117 additional sequences against the simplified database using BLAST. To examine the effect of sequence length, we used query sequences of length 150, 200, ..., 550, 600 bp. As our environmental data come from the ITS1-region, in the main analysis we cut these query sequences from the beginning of the ITS1-region (starting from the end of the primer ITS1F). To examine the effect of the genetic region, we repeated the analyses with cutting the sequences from the

beginning of the ITS2-region (starting from the end of the primer 58SF).

We included the following explanatory variables in the quality index q :

The overlapping length of the query sequence and the reference sequence (termed *length*); the HSP score divided by the overlapping length of the sample and reference sequences (termed *identity*); the difference between the best BLAST hit and the second best BLAST hit, calculated as the difference in the HSP scores divided by the *length* of the best BLAST hit (termed *gap*); a class variable indicating if the best BLAST hit was from the SAF database (*database* = 1) or from the UNITE database (*database* = 0). For the case of species-level analyses, a class variable indicating if the best BLAST hit belonged to a genus within which the classification or the nomenclature of species were considered unclear (*problematic genus* = 1 for the genera *Cortinarius*, *Laccaria* and *Amanita*, otherwise *problematic genus* = 0).

We estimated the probability $P(s = i|q, s \in R)$ with Bayesian logistic regression by letting the response variable be one if the species was identified correctly (BLAST returned the same species) and zero if the species was misidentified. To estimate the probability $P(s \in R|q)$, we generated BLAST results for cases in which the species represented by the query sequence was or was not present in the reference database. To obtain BLAST results of the latter type, we removed the focal species (or in case of genus-level analyses, all species representing the focal genus) from the reference database. Our aim was to use this method to identify dead wood associated fungi (polypores, corticioids, and ectomycorrhizal asco- and basidiomycetes) from environmental samples originating from spruce logs in Southern Finland. We assumed that within this target group, a species represented by a typical environmental sequence was included in the SAF-UNITE database with 80% probability, and we thus weighted in the logistic regression for $P(s \in R|q)$ the two types of samples by 0.8 and 0.2. For the genus-level analyses, we used the weights 0.9 and 0.1.

Environmental samples

In Mar. 2008, we sampled saw-dust from four spruce logs, using a 10 mm electric drill, in a protected semi-natural spruce-dominated forest 30 km north of Helsinki (Rörstrand, Sipoo) in Finland. The four logs, referred to as A, B, C and D, were 30–36 cm in diameter, and covered the range of decay classes from 1b to 4 (cf. Hottola & Siitonen 2008). Tree bark and epiphytes were removed from the drilling points prior to the sampling. The logs A, B and C were sampled at 0.3 m and 1.3 m from the base and then at 3 m intervals, whereas log D was sampled throughout at 1 m intervals. The number of samples depended on the length of the log, leading to 10 drilling points for logs A and B, eight drilling points for log C, and 23 drilling points for log D. The resulting 51 samples were kept separate for the DGGE analysis but combined into eight pooled samples for the 454 sequencing. For the latter case, samples 1 and 2 consisted of the material from one drilling point at breast height (1.3 m) for the logs A and B, respectively. Samples 3–6 were bulk samples originating from 0.3 m, 1.3 m, and then at 3 m intervals for the logs A–D, respectively. Sample 7 included material from all 1 m interval drilling points of log D. Sample 8

was a composite sample including material from all 51 drilling points.

All logs were inspected for fruit bodies of polypore species. The fruit body inventory was conducted twice (Mar. and Sep. 2008) to account for species-specific variation in timing and duration of fruiting. Fruit bodies that could not be reliably identified in the field were collected for microscopical species identification.

DNA extraction and PCR

Saw-dust from each drilling point was stored at -20°C . DNA extraction from 150 to 250 mg (fresh weight) samples was carried out using Power Soil DNA isolation kit (MoBio Laboratories, USA). Release of DNA was performed using FastPrep cell disrupter (Qbiogene, France) for 3×20 s at 4 m s^{-1} and incubation for 30 min at 65°C . Samples were amplified with 25 cycles using Phusion enzyme (Funnymen Ltd, Finland) prior to 454 sequencing and 35 prior to DGGE with the primers ITS1F (Gardes & Bruns 1993) and ITS2 (White *et al.* 1990).

454 sequencing

Each sample was marked with a specific tag sequence (Tag 1 = AGCAGC, Tag 2 = TACAGC, Tag 3 = TAGCTA, Tag 4 = TCT GTA, Tag 5 = ACAGCT, Tag 6 = CTA CTG, Tag 7 = CAGCTC, Tag 8 = TGTACG) incorporated in the primers. The PCR reactions were purified separately with Ampure (Agencourt Bioscience Corporation, USA) and the products were quantified with Nanodrop (Thermo Scientific, USA) and Bioanalyzer 2100 using DNA 1000 chips (Agilent Technologies, USA). One unit of amplicon DNA was applied to the emulsion (EM) PCR based on the Nanodrop result. One EM PCR tube for each sample was used. The samples were mixed in the emulsion breaking and the obtained 218 000 beads were sequenced using GC FLX. Stringent analysis software settings were used for the obtained sequences. Out of the total of 72 807 sequences, we ignored sequences shorter than 150 bp, resulting in 62 214 sequences.

DGGE analyses

DGGE analyses were applied individually to all 51 samples in addition to the eight pooled samples. The PCR amplification and DGGE analysis were performed as in Korkama-Rajala *et al.*

(2008). DGGE gels were analyzed using GelCompar II software (Applied Maths, Sint-Martens-Latem, Belgium). Representative DGGE-bands of each mobility group were excised, re-amplified with a reduced number of cycles (25) and re-run in DGGE. Single-banded samples were amplified with a primer pair ITS1F–ITS2, purified (High Pure PCR Purification Kit, Roche, Mannheim, Germany) and sequenced with SEQ 8000 DNA analysis system using Quick Start kit (Beckman Coulter Inc., USA). The sequences obtained from the DGGE-bands were manually checked and edited using the BioEdit Sequence Alignment Editor (Hall 1999).

Identifying the environmental sequences

We applied the NCBI-BLAST algorithm to compare the environmental sequences (with primers removed) against the SAF-UNITE database and the GenBank database (Benson *et al.* 2009). To avoid the possible homopolymer bias generated by 454 sequencing, we truncated all homopolymers to at most 4 bp for both the environmental sequences and for the SAF-UNITE reference database, and set the low complexity filtering on for the BLAST search. We estimated for each query sequence the probability of correct identification to the species level and to the genus level using the statistical model described above. The pipeline was implemented using Biopython modules with a standalone BLAST-server. The source code is available (see [Supplementary material](#)).

Results

Probability of correct identification

Based on the ITS1-region, the conditional probability of correct identification $P(s = i|q, s \in R)$ increased with identity and gap (Table 1, Fig 1) for both the species-level and the genus-level analyses. Thus, when using a known sequence (a multiple sequence extracted from the SAF database) as a query sequence, the best BLAST hit went to the correct species in cases for which the sequence similarity of the best match was high (high identity) and the sequence similarity of the second best hit was low (high gap). The length of the sequence had only minor predictive power. As expected, the probability $P(s = i|q, s \in R)$ was lower for the genera that were identified as problematic. Excluding these three genera (present only in

Table 1 – Modelling the probability of correct identification to species or to genus level. $P(s = i|q, s \in R)$ is the probability by which the BLAST result is correct conditional on the target species being in the reference database and $P(s = R|q)$ is the probability that the target species is in the reference database (see Material and methods). The values in the table show the regression coefficients (median estimate and the 95% central highest posterior density interval) of the logistic regression model. The estimates for length have been multiplied by 1 000. Results based on the ITS1-region (for similar results for ITS2-region, see Supplementary material)

Model	Level	Database	Identity	Length	Gap	Problematic genus
$P(s = i q, s \in R)$	Species	-0.8 (-1.0, -0.6)	2.8 (2.3, 3.2)	0.5 (0.07, 0.9)	19.1 (18.2, 20.2)	-0.88 (-1.0, -0.7)
$P(s = i q, s \in R)$	Genus	-1.4 (-1.8, -1.0)	2.9 (2.1, 3.8)	0.6 (-0.8, 2.0)	6.9 (5.3, 9.2)	
$P(s \in R q)$	Species	0.19 (0.10, 0.28)	6.6 (6.4, 6.8)	-0.4 (-0.6, -0.2)	4.0 (3.7, 4.2)	-0.4 (-0.5, -0.3)
$P(s \in R q)$	Genus	0.05 (-0.05, 0.14)	12.0 (11.8, 12.2)	-2.9 (-3.3, -2.6)	3.5 (2.9, 3.9)	

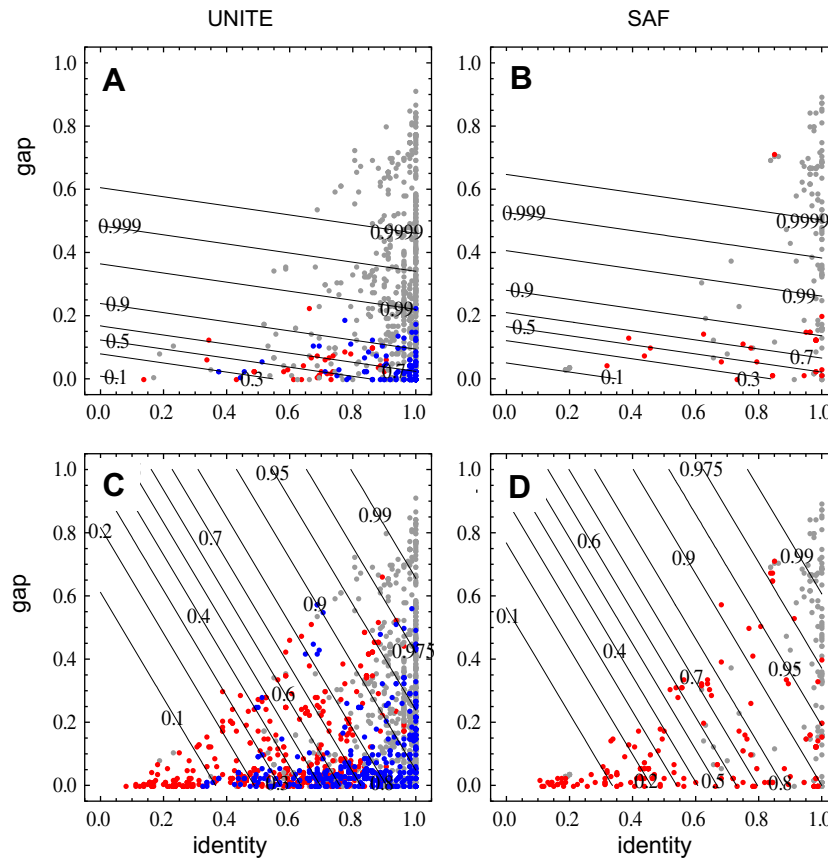


Fig 1 – Estimating the probability of correct identification by logistic regression. The upper panels (A and B) show the probability $P(s=i|q, s \in R)$ by which the best BLAST hit represents the correct species conditional on the species being in the reference database. The lower panels (C and D) show the probability $P(s=R|q)$ by which the species is in the reference database. The left-hand panels (A and C) show the results for the UNITE database and the right-hand panels (B and D) for the SAF database. The dots represent the data consisting of known sequences extracted from the database (shown only for sequence length 200 bp). Grey dots indicate correct identifications, and blue (for the problematic genera) and red (for the other genera) dots misidentifications. The contour lines show the predictions of the fitted logistic regression models, with explanatory variables *length* set to 200 and *problematic genera* set to 0. See [Supplementary material](#) for a corresponding figure based on the ITS2-region.

UNITE), the probability $P(s=i|q, s \in R)$ was higher for UNITE than for SAF (Table 1).

The probability $P(s=R|q)$ that the query species is in the database also increased with *identity* and *gap* (Table 1, Fig 1). Thus, when using a known sequence (a multiple sequence extracted from the SAF database) as a query sequence, the *identity* and *gap* values of the best BLAST hit were high if another sequence of the same species was kept in the database, but these values were low if the correct species was missing from the database. With given *identity* and *gap* values, the probability $P(s=R|q)$ was somewhat higher for the SAF database than for the UNITE database, and it was lower for the genera that were identified as problematic. The *length* of the sequence actually had a small negative effect (Table 1). This somewhat counterintuitive result is explained by the fact that short sequences are classified unreliable already for the reason that they tend to have a low *gap* value.

Combining the models $P(s=i|q, s \in R)$ and $P(s=R|q)$ by Eq. (1), we can use the outcome of the BLAST algorithm (the

values of *identity*, *gap*, *length*, *database* and *problematic genus* corresponding to the best hit) to estimate the probability by which the species behind an environmental query sequence is the same species as the best matching hit from the SAF–UNITE database. The combined probability $P(s=i|q)$ is illustrated by the contour lines in Fig 2, showing the estimated probability of the best BLAST hit being correct, either to the species level (upper panels) or to the genus level (lower panels). Comparing Figs 1 and 2 shows that for high values of *identity* and *gap*, the main issue was whether the focal species is included in the reference database, as in this part of the parameter space the conditional probability of correct identification (given that the species is included in the reference database) was very high.

For the ITS2-region, the results were qualitatively similar, though for this region *identity* got more weight than for the ITS1-region, and the *length* of the query sequence became an important predictor especially for the genus-level analyses (Supplementary material).

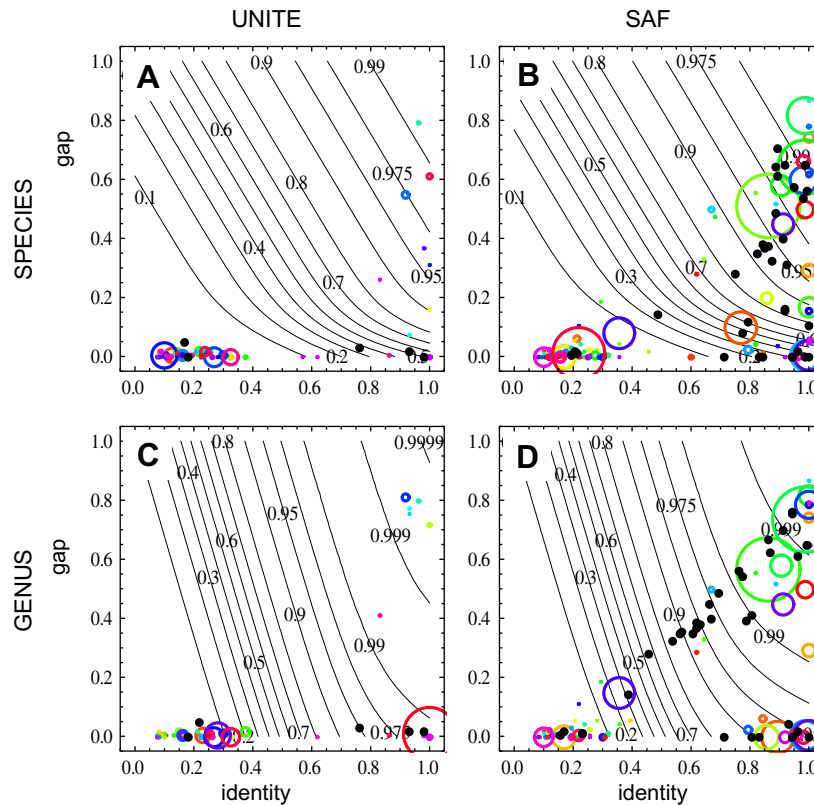


Fig 2 – Species diversity and probabilities of correct identification in the environmental samples. The contour lines show the combined probability $P(s=i|q)$ by which the best BLAST represents the correct result to species level (A and B) or genus level (C and D), assuming the sequence length of 200. The coloured circles show the empirical data based on 454 sequencing. Each colour represents a different candidate name, the area of the circle being proportional to the number of sequences assigned to that candidate name. The black dots show the corresponding data based on DGGE analysis. A and C (B and D) show species for which the best hit was found from the UNITE (SAF) database.

Species diversity in environmental samples

Out of the 62214 sequences obtained by 454 sequencing, 51% could be identified to the species level at least with 90% confidence, these sequences representing 30 different species (Fig 2). An additional 9% of the sequences could be identified with 90% confidence to the genus level, amounting to 14 genera that were not included in the species-level identifications. The remaining 40% of the sequences remained unidentified even to the genus level, if 90% probability of correct identification was used as the threshold. Using the same criteria, the DGGE data revealed nine different species and six additional genera (Figs 2 and 3). Comparing the results obtained by 454 sequencing with those from DGGE analysis confirms the expectation that DGGE is likely to reveal especially the most dominating species. For example, out of the sequences obtained for the pooled sample 8, the DGGE analysis revealed 42% of the species (identified to the species level with at least 90% confidence) and 70% of the sequences (see Supplementary material).

A total of 15 species were observed as fruit bodies, but only five of these were discovered by 454 sequencing with at least 90% confidence (Fig 3). If the 90% confidence threshold was relaxed, 454 sequencing found four additional species present

in the fruit-body data. Out of the six species not included even in the full list of candidate names obtained by 454 sequencing, five species were present as a genus-level hit. Lists of species and genera obtained by the three methods (454, DGGE, and fruit-body inventory) are given in Supplementary material.

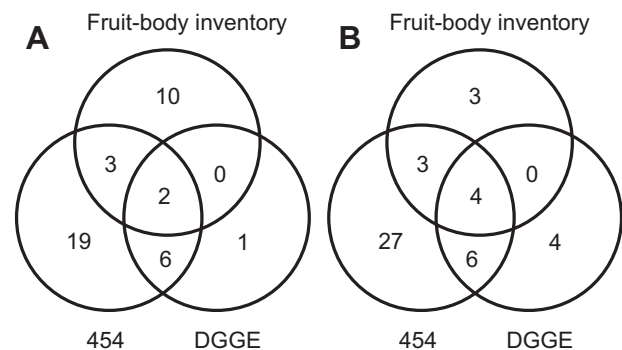


Fig 3 – The number of species (A) and genera (B) identified by 454 sequencing, DGGE analysis and fruit-body inventory. In case of the two molecular methods, only identifications with at least 90% confidence were included. Results from GenBank searches are not included.

Most of the reliably identified species, and especially those that were represented by a large number of sequences, belonged to the SAF database rather than the UNITE database (Fig 2). As our method for assessing the probability of correct identification can be applied only to a local database with fully identified reference sequences, we could not perform a systematic comparison with GenBank. However, we used the GenBank database to identify additional taxonomic groups present in the environmental samples. Out of the 26 158 sequences for which we did not obtain at least a 90% hit in terms of identity from the local libraries, 8 299 resulted in at least a 90% identity hit from GenBank (see [Supplementary material](#)). These represented 174 different GenBank entries, out of which 45% (27% of sequences) belonged to Ascomycota, 25% (7%) to Basidiomycota, 1% (0.04%) to Mucoromycotina, and 0.6% (0.01%) to Glomeromycota. Twenty-eight per cent (65%) of the GenBank hits represented unidentified fungal species.

Effect of sampling design

To assess the effect of the sampling design, we compared the DGGE-bands obtained separately for each of the 51 drilling points. In this analysis, we assumed that each band represented a single species without considering how reliable the species identification was. The number of species increased as a function of drilling points (Fig 4A), suggesting that roughly half of the species community is missed if less than five drilling points are used. Interestingly, the cumulative species accumulation curves were almost independent of whether the drilling points were distributed evenly or randomly within the log (Fig 4A).

As we barcoded the samples using a nested design prior to 454 sequencing, these data provide some clues to the effect of the sampling design. Accounting only for identifications that were estimated to be correct with at least 90% confidence, Fig 4B shows how well the species observed in subsamples represent the diversity observed in the larger samples. For log A, one species was found in the single sample, whereas the

combined sample contained five species. For the log B, eight species were found in the single sample, whereas the combined sample contained 17 species. All species of the single samples were also present in the combined samples, suggesting that the resolution of the 454 analysis was not compromised by mixing the 10 saw-dust samples before DNA extraction. The rectangles in Fig 4B compare the result for log D where we either mixed the samples at 3 m or 1 m intervals. The overlap between these two cases was 10 species, whereas six species were present only in the sample with 1 m intervals, and three species were present only in the samples taken at 3 m intervals. Finally, the triangles compare the species composition in the combined sample consisting of all four logs (including 19 species) compared to the prediction based on the species compositions obtained separately for each log (including all the 19 species and nine additional species). Given that the combined sample is a mixture of 51 saw-dust samples, it represents the predicted species frequencies surprisingly well, the missing species having a low abundance also in the individual samples.

Discussion

Emerging sequencing technologies, such as 454 pyrosequencing and parallelized implementations of Sanger sequencing, hold promise of generating hitherto unprecedented amounts of environmental sequence data. Consequently, automated tools for processing the raw data into useful summaries will become an absolute necessity. A key concern in the use of algorithmic approaches applied to sequence data and questions such as species identification relates to the reliability of the results. In this paper, we have developed a statistical framework which can be used to assess the reliability of BLAST search results for the purpose of species identification. The methods developed here serve as a simple starting point, and they can be refined and extended in various directions. To start with, instead of accounting for indices of sequence similarity as reported by BLAST, one could

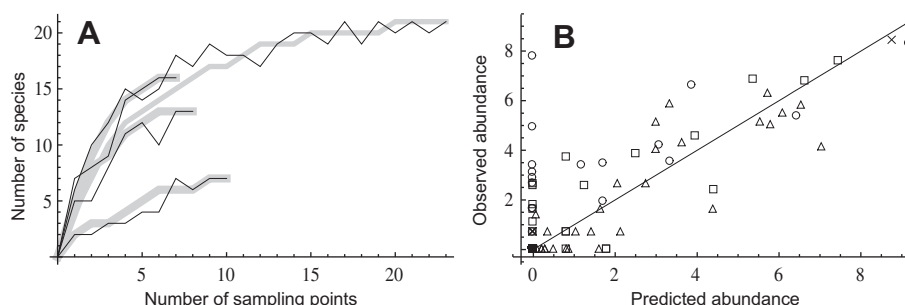


Fig 4 – Consistency of the data among samples. (A), the cumulative number of distinct DGGE-bands as a function of drilling points per log. The black lines are based on placing the drilling points as evenly as possible within the logs, and the grey lines show the expectation if the drilling points are selected randomly (median of 1 000 randomizations). **(B),** the observed number of sequences in four pooled 454 samples versus the predicted number based on subsamples, accounting only for species that were correctly identified with at least 90% probability. The crosses correspond to sample 3 (all drilling spots of log A mixed) predicted by sample 1 (a single drilling spot of log A). The circles correspond to sample 4 predicted by sample 2, rectangles to sample 7 predicted by sample 6, and triangles to sample 8 predicted by samples 3, 4, 5 and 7. Abundance is measured as log (number of sequences + 1), and the line shows identity.

integrate the probabilistic approach with more sophisticated algorithms targeted specifically for taxonomic classification, such as those used in MEGAN and CARMA. Further, we treated all species equally, except the three genera within which species-level classification was known to be partly tentative. Analogous to taxon-specific thresholds (Menkis *et al.* 2005), it would be possible to include taxon-specific information about levels of sequence similarity within and among species, e.g., using ordination techniques (Linton *et al.* 2007). Finally, we emphasize the need to combine automated and manual species identification, the automated algorithm singling out the sequences that may require further scrutiny by the user.

Results obtained here and by earlier studies (Nilsson *et al.* 2009; Ryberg *et al.* 2009) point out two reasons why species identifications from sequence data may fail. First, the levels of within and between species variation in the ITS1-region vary among different groups of wood-inhabiting fungi (Kausserud & Schumacher 2003; Schmidt & Moreth 2003; Högberg & Land 2004). Cases for which the ITS1-region is not sufficiently variable among species lead to a low *gap* value, indicating that there is another species with almost equally as high *identity* score such hits are located in the lower-right corners of the panels of Figs 1 and 2. To improve the reliability of these identifications, longer sequences or sequence information from other regions of the genome is required. Second, the focal species may simply be missing from the reference database, with the BLAST results being characterized by a low *identity* score (the hits shown in the lower-left corners of the panels of Figs 1 and 2). Improvement here can obviously be done only by extending the coverage of the reference database with respect to taxonomic and ecological sampling (cf. Ryberg *et al.* 2009).

Judging from the present results, a relatively dense grid of sampling points may be needed to cover the majority of the fungal species in a log (Fig 4A). Large sample size leads to high cost and high workload, making it important to optimize the study design. Contrary to fingerprinting methods, combining even 10–20 drilling samples prior to 454 sequencing did not seem to compromise the observed species frequencies (Fig 4B). Thus, high-throughput sequencing has an advantage in terms of reduced amount of samples required for DNA release and purification. However, it is clear that pooling too many samples will eventually reduce the resolution of the method. This is either because of imperfect mixing as saw-dust, or due to the randomness associated with the finite number of sequences obtained by PCR or 454 sequencing. A further limitation of any approach involving PCR is that in complex samples all templates are not amplified with identical efficiency (Polz & Cavanaugh 1998; Acinas *et al.* 2004).

Much of the emphasis in metagenomics has been in shotgun sequencing of environmental samples, targeting, for example, gene functions and metabolic capacities in microbial communities (Tringe & Rubin 2005; Ward 2006). Modern sequencing methods provide efficient tools also for the much simpler task of identifying the species that are present in a given environmental sample. While simply generating a list of the species present in the sample may seldom be the ultimate goal of any scientific study, such data nevertheless form the pillar of many kinds of studies, such as analyses of species-specific habitat requirements, models of single-

species and community dynamics, and the assessment of the functional roles of species within an ecosystem. The framework developed here facilitates the use of molecular methods in population and community ecology by assessing the reliability of identifications using probabilities, the natural unit for measuring uncertainty. This is not only of conceptual value, but also makes it practical to transfer the unavoidable uncertainty in species identifications to further analyses. If these uncertainties are used as a part of the input data for, e.g., Bayesian models of population dynamics, they can be naturally propagated throughout the sequence of analyses. When considering the dramatic improvements in sequencing technology during the last few years, there can be little doubt that we are well on our way to crossing the border into a new era in fungal inventories and in how fungi are studied in their natural environments.

Acknowledgements

We thank Dr Jenni Hultman and especially Dr Henrik Nilsson for valuable comments. This study was supported by the Academy of Finland (Grant no. 124242 to O.O. and Grant no. 121630 to R.M.) and the European Research Council (ERC Starting Grant no. 205905 to O.O.).

Supplementary material

Supplementary data associated with this article can be found in the online version, at doi:10.1016/j.funeco.2010.01.001

REFERENCES

- Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF, 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.
- Allmér J, Vasiliauskas R, Ihrmark K, Stenlid J, Dahlberg A, 2006. Wood-inhabiting fungal communities in woody debris of Norway spruce (*Picea abies* (L.) Karst.), as reflected by sporocarps, mycelial isolations and T-RFLP identification. *Fems Microbiology Ecology* **55**: 57–67.
- Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389–3402.
- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW, 2009. GenBank. *Nucleic Acids Research* **37**: D26–D31.
- Berglund H, Edman M, Ericson L, 2005. Temporal variation of wood-fungi diversity in boreal old-growth forests: implications for monitoring. *Ecological Applications* **15**: 970–982.
- Buée M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F, 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184**: 449–456.
- Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer ELL, Bateman A, 2008. The Pfam protein families database. *Nucleic Acids Research* **36**: D281–D288.
- Froslev TG, Matheny PB, Hibbett DS, 2005. Lower level relationships in the mushroom genus *Cortinarius* (basidiomycota, agaricales):

- a comparison of RPB1, RPB2, and ITS phylogenies. *Molecular Phylogenetics and Evolution* 37: 602–618.
- Gärdenfors U (ed), 2005. *The 2005 Red-List of Swedish Species*. ArtDataBanken, SLU, Uppsala.
- 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.
- Geml J, Laursen GA, O'Neill K, Nusbaum HC, Taylor DL, 2006. Beringian origins and cryptic speciation events in the fly agaric (*Amanita muscaria*). *Molecular Ecology* 15: 225–239.
- Gyllenberg M, Koski T, 2001. Probabilistic models for bacterial taxonomy. *International Statistical Review* 69: 249–276.
- Hall TA, 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41: 95–98.
- Halme P, Kotiaho JS, Ylisirnio AL, Hottola J, Junninen K, Kouki J, Lindgren M, Monkkonen M, Penttilä R, Renvall P, Siitonen J, Simila M, 2009. Perennial polypores as indicators of annual and red-listed polypores. *Ecological Indicators* 9: 256–266.
- Nordic macromycetes. In: Hansen L, Knudsen H (eds), *Heterobasidioid, Aphyllophoroid and Gastromycetoid Basidiomycetes*, vol. 3. Nordsvamp, Copenhagen.
- Hjortstam K, Larsson K-H, Ryvarde L, 1987. *Introduction and keys*. In: *Corticaceae of North Europe*, vol. 1. Fungiflora, Oslo.
- Högberg N, Land CJ, 2004. Identification of *Serpula lacrymans* and other decay fungi in construction timber by sequencing of ribosomal DNA – a practical approach. *Holzforschung* 58: 199–204.
- Hottola J, Siitonen J, 2008. Significance of woodland key habitats for polypore diversity and red-listed species in boreal forests. *Biodiversity and Conservation* 17: 2559–2577.
- Hughes KW, Petersen RH, Lickey EB, 2009. Using heterozygosity to estimate a percentage DNA sequence similarity for environmental species' delimitation across basidiomycete fungi. *New Phytologist* 182: 795–798.
- Huson DH, Auch AF, Qi J, Schuster SC, 2007. MEGAN analysis of metagenomic data. *Genome Research* 17: 377–386.
- Jasalavich CA, Ostrofsky A, Jellison J, 2000. Detection and identification of decay fungi in spruce wood by restriction fragment length polymorphism analysis of amplified genes encoding rRNA. *Applied and Environmental Microbiology* 66: 4725–4734.
- Johannesson H, Stenlid J, 1999. Molecular identification of wood-inhabiting fungi in an unmanaged *Picea abies* forest in Sweden. *Forest Ecology and Management* 115: 203–211.
- Jumpponen A, Jones KL, 2009. Massively parallel 454 sequencing indicates hyperdiverse fungal communities in temperate *Quercus macrocarpa* phyllosphere. *New Phytologist* 184: 438–448.
- Karen O, Hogberg N, Dahlberg A, Jonsson L, Nylund JE, 1997. Inter- and intraspecific variation in the ITS region of rDNA of ectomycorrhizal fungi in Fennoscandia as detected by endonuclease analysis. *New Phytologist* 136: 313–325.
- Kauserud H, Schumacher T, 2003. Ribosomal DNA variation, recombination and inheritance in the basidiomycete *Trichaptum abietinum*: implications for reticulate evolution. *Heredity* 91: 163–172.
- Knudsen H, Vesterholt J (eds), 2008. *Funga Nordica: Agaricoid, Boletoid and Cyphelloid Genera*. Nordsvamp, Copenhagen.
- Köljalg U, Larsson KH, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM, 2005. UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist* 166: 1063–1068.
- Korkkama-Rajala T, Mueller MM, Pennanen T, 2008. Decomposition and fungi of needle litter from slow- and fast-growing Norway spruce (*Picea abies*) clones. *Microbial Ecology* 56: 76–89.
- Kotiranta H, Saarenoksa R, Kytövuori I, 2009. Aphyllophoroid fungi of Finland. A check-list with ecology, distribution, and threat categories. *Norrinia* 19: 1–223.
- Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J, 2008. Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research* 36: 2230–2239.
- Kulhankova A, Beguiristain T, Moukoumi J, Berthelin J, Ranger J, 2006. Spatial and temporal diversity of wood decomposer communities in different forest stands, determined by ITS rDNA targeted TGGE. *Annals of Forest Science* 63: 547–556.
- Kwasna H, Bateman GL, Ward E, 2008. Determining species diversity of microfungal communities in forest tree roots by pure-culture isolation and DNA sequencing. *Applied Soil Ecology* 40: 44–56.
- Linton CJ, Borman AM, Cheung G, Holmes AD, Szekeley A, Palmer MD, Bridge PD, Campbell CK, Johnson EM, 2007. Molecular identification of unusual pathogenic yeast isolates by large ribosomal subunit gene sequencing: 2 years of experience at the United Kingdom Mycology Reference Laboratory. *Journal of Clinical Microbiology* 45: 1152–1158.
- Markowitz VM, Ivanova N, Palaniappan K, Szeto E, Korzeniewski F, Lykidis A, Anderson I, Mavrommatis K, Kunin V, Martin HG, Dubchak I, Hugenholtz P, Kyrpides NC, 2006. An experimental metagenome data management and analysis system. *Bioinformatics* 22: E359–E367.
- Menkis A, Vasiliauskas R, Taylor AFS, Stenlid J, Finlay R, 2005. Fungal communities in mycorrhizal roots of conifer seedlings in forest nurseries under different cultivation systems, assessed by morphotyping, direct sequencing and mycelial isolation. *Mycorrhiza* 16: 33–41.
- Niemelä T, 2005. *Polypores, Lignicolous Fungi*. Helsinki University Press, Helsinki.
- Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson KH, 2008. Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics*: 193–201.
- Nilsson RH, Ryberg M, Abarenkov K, Sjökvist E, Kristiansson E, 2009. The ITS region as a target for characterization of fungal communities using emerging sequencing technologies. *Fems Microbiology Letters* 296: 97–101.
- Peay KG, Kennedy PG, Bruns TD, 2008. Fungal community ecology: a hybrid beast with a molecular master. *Bioscience* 58: 799–810.
- Polz MF, Cavanaugh CM, 1998. Bias in template-to-product ratios in multitemplate PCR. *Applied and Environmental Microbiology* 64: 3724–3730.
- Porter TM, Skillman JE, Moncalvo JM, 2008. Fruiting body and soil rDNA sampling detects complementary assemblage of *Agaricomycotina* (basidiomycota, fungi) in a hemlock-dominated forest plot in southern Ontario. *Molecular Ecology* 17: 3037–3050.
- Raes J, Foerstner KU, Bork P, 2007. Get the most out of your metagenome: computational analysis of environmental sequence data. *Current Opinion in Microbiology* 10: 490–498.
- Rassi P, Alanen A, Kanerva T, Mannerkoski I (eds), 2001. *Suomen lajien uhanalaisuus 2000*. Ministry of the Environment and Finnish Environment Institute, Helsinki.
- Ryberg M, Kristiansson E, Sjökvist E, Nilsson RH, 2009. An outlook on the fungal internal transcribed spacer sequences in GenBank and the introduction of a web-based tool for the exploration of fungal diversity. *New Phytologist* 181: 471–477.
- Ryberg M, Nilsson RH, Kristiansson E, Topel M, Jacobsson S, Larsson E, 2008. Mining metadata from unidentified ITS

- sequences in GenBank: a case study in *Inocybe* (basidiomycota). *BMC Evolutionary Biology* **8**: 50.
- Schmidt O, Moreth U, 2003. Molecular identity of species and isolates of internal pore fungi *Antrrodia* spp. and *Oligoporus placenta*. *Holzforschung* **57**: 120–126.
- Tringe SG, Rubin EM, 2005. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* **6**: 805–814.
- Vainio EJ, Hantula J, 2000. Direct analysis of wood-inhabiting fungi using denaturing gradient gel electrophoresis of amplified ribosomal DNA. *Mycological Research* **104**: 927–936.
- Ward N, 2006. New directions and interactions in metagenomics research. *Fems Microbiology Ecology* **55**: 331–338.
- White TJ, Bruns TD, Lee S, Taylor J, 1990. Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfaud DH, Sninsky JJ, White TJ (eds), *PCR Protocols: a Guide to Methods and Applications*. Academic Press, San Diego, USA, pp. 315–322.