Letter

Parsing ecological signal from noise in next generation amplicon sequencing

Introduction

It is clear that the use of next generation sequencing (NGS) applied to environmental DNA is changing the way researchers conduct experiments and significantly deepening our understanding of microbial communities around the globe (Amend *et al.*, 2010; Caporaso *et al.*, 2011; Bik *et al.*, 2012; Bates *et al.*, 2013). The lower per unit cost and sheer number of sequences relative to traditional methods provide tremendous advantages in characterizing the richness and composition of highly diverse microbial systems (Bokulich *et al.*, 2013). In a recent volume of *New Phytologist*, Lindahl *et al.* (2013) presented an excellent introduction into high-throughput sequencing of amplified gene markers for fungi, and broadly discussed field sampling and handling, DNA extraction, markers, primers, amplicon library construction, sequencing platform, bioinformatic analyses and data interpretation. We applaud their overview as an important general guide, but we have found that there are significant additional issues regarding NGS that have not been well articulated in the literature, especially when applied to fungi. Below we highlight a series of platform-independent recommendations based on our recent experiences with NGS, which we think are critical for maximizing the signal: noise ratio in molecular ecological analyses.

Controls, both negative and positive

The inclusion of both negative and positive controls is indispensable in NGS-based studies due to the greater detection level than traditional sequencing (i.e. sequences can be readily detected in controls with NGS methods even in the absence of positive PCR bands). These controls are essential at multiple steps during the experimental process, for example, in the field, in laboratory settings where samples are processed, during DNA extraction, as well as before and after PCR. To be useful, the controls must be treated identically to other samples from initial processing through library preparation. As an example, we recently conducted a field-based study using the Illumina MiSeq platform to amplify the ITS1 region of soil fungi, which included a series of negative controls. Of the total sequence pool generated, we detected 0.01% from soil sieve controls (3.17% total OTUs (operational taxonomic unit (s))), 0.0001% (0.2% total OTUs) from DNA extraction controls, and 0.001% sequences (0.67% total OTUs) from PCR controls. Together, these controls accounted for 0.01% of total sequences (3.8% of total OTUs).

While detection of fungal taxa in negative controls is key to determining which fungal taxa should be included in subsequent ecological analyses, there is currently no consensus on how to handle these sequences. One approach would be to simply delete any OTUs that appeared in negative controls across all samples (e.g. Vik *et al.*, 2013). However, in our study, this would have deleted many of the most abundant OTUs in the experimental samples. It seems highly likely that those abundant OTUs were in fact present in the field because (1) many had been previously encountered in soil and (2) their abundance in the controls was multiple orders of magnitude lower. To avoid eliminating OTUs that appeared to be ecologically valid, we addressed this issue by subtracting the number of sequences of each OTU present in the negative controls from the sequence abundance of that OTU in the experimental samples (essentially, after subtraction, the negative control samples will contain zero sequences, and other samples will have reduced abundances). In our dataset, this approach eliminated only two low abundance OTUs (each had < 40 total sequences) instead of 56 OTUs had we used the deletion approach. While we recognize that the inclusion of two low abundance OTUs would have negligible effects on the large conclusions drawn, the alternative deletion approach would have most likely created ecologically spurious results. Although this subtraction approach may not be best for all studies, we strongly recommend that researchers sequence all negative controls and explicitly report those results as well as how sequences in controls are processed before ecological analyses. Further, if this subtraction method is applied, we advocate adding as equivalent volumes of the negative control and experimental samples as possible to avoid sequence inflation biases.

We also suggest that researchers consider additional types of negative controls in experimental designs. In the aforementioned study, we were particularly interested in characterizing the richness of active ectomycorrhizal (ECM) fungal communities, so we sampled soils from monodominant plots containing either ECM or arbuscular mycorrhizal (AM) hosts. Because ECM hosts are absent in the AM host plots, we considered those samples to be a negative field control for ECM fungal taxa that were present as spores but not mycelium. We found that while ECM fungal taxa represented only 1.6% of the sequences from the AM host plots, they included nearly 60% of all the ECM fungal OTUs present in the ECM host plots. Subtracting the abundance of each ECM fungal OTU in AM plots from the abundance of the same ECM fungal OTU in ECM plots had the greatest impact on low abundance ECM fungal OTUs and reduced the total number of OTUs in our host samples by 10%. However, this extra type of negative control made us more confident that the ECM fungal taxa included in our final dataset were most likely to be present as mycelium rather than spores. This

kind of subtraction method works for DNA-based samples, but an RNA-based approach is another alternative for eliminating fungal OTUs present only as spores (Van Der Linde & Haller, 2013).

A good positive control is a 'mock community' – a known set of organisms from quantified amounts of DNA (Amend *et al.*, 2010; Huse *et al.*, 2010; Caporaso *et al.*, 2011; Ihrmark *et al.*, 2012; Bokulich *et al.*, 2013; Egge *et al.*, 2013). There are two primary ways to set up a mock community; one is to have the same amount of DNA for all organisms and the other is to vary them as if they were found in a natural community with abundant and less abundant taxa represented. Given that gene copy number can vary significantly across fungal taxa (Debaud *et al.*, 1999), the mock community could also be built from individual PCR products of each taxon to account for this variation. Regardless of the specific method chosen, the mock community should best cover taxonomic breadth, with some groups having more closely related species. Having breadth will be more representative of the taxonomic variation typically present in community-level studies, while having more closely related species will allow the user to adjust the clustering threshold to recover the particular number of taxa in the mock community as a proxy for the total community. The latter is important because the clustering process may obscure taxonomic richness patterns (Yamamoto & Bibby, 2014). A mock community will also serve to inform the quality of the sequencing run (i.e. helps address run-to-run variation) and the processing steps necessary to retain the most data (i.e. addresses sequencing data quality).

As an example, we created a modular mock community consisting of equimolar genomic DNA aliquots from 27 species of fungi in the Ascomycota and Basidiomycota (Table 1). We used tissue from recently dried mushrooms, although the best option would be to use pure cultures to avoid contaminants. We treated the mock community the same as all other experimental samples and controls during PCR, library preparation and sequencing of the field study outlined earlier (see Smith & Peay (2014) for details about primers and PCR conditions). We began our analysis by first carefully examining the mock community. After bioinformatics processing, we found that we were only able to recover a maximum of 25 out of the 27 species. Two species failed to sequence completely, whereas others were eliminated depending on the data filtering parameters (Table 1). Interestingly, the species that failed to sequence were those that had unusually long ITS1 sequences such as *Leccinum* and *Cantharellus* (see Ihrmark *et al.* (2012) for data showing relationships between amplicon length and amplification). Furthermore, *Cantharellus* have been documented to amplify poorly with ITS primers (Buyck & Hofstetter, 2011), which could also account for the failure to sequence. Equivalent results were produced in a second independent Illumina run, suggesting the aforementioned patterns are not likely to be artifactual. Overall, the presence of the mock community allowed us to choose the appropriate sequence quality filtering method as well as calibrate our clustering threshold (i.e. 95%, 97%, 97.5%, etc.) to best recover the actual number of species in the total dataset. As such, we consider the inclusion of this kind of positive control to be just as essential to any NGS run as negative controls.

Table 1 Comparison of sequence abundances and operational taxomomic unit (OTU) counts among forward single direction sequences (ITS1F), reverse direction sequences (ITS2-barcode), and paired sequences in a mock community of 27 Basidiomycota and Ascomycota species

Mock community taxa	Single direction forward sequences	Single direction reverse sequences	Paired sequences
Hygrophorus russula	10717	7404	7696
Cortinarius sp.	886	615	631
Amanita muscaria	883	513	531
Tricholoma sp.	562	340	407
Entoloma abortivum	557	295	371
Xerocomus subtomentosus	485	337	347
Pholiota spumosa	451	311	291
Suillus laricinus	312	131	198
Thelephora terrestris	311	199	197
Suillus granulatus	227	122	130
Suillus americanus	221	96	125
Suillus grevillei	210	94	123
Helvella vespertina	201	0	6
Suillus luteus	152	77	104
Leucopaxillus gentianeus	108	66	70
Lactarius sp.	103	47	58
Laccaria laccata	103	54	67
Paxillus cuprinus	70	48	43
Helvella dryophila	66	0	5
Suillus spectabilis	47	12	29
Boletus edulis	45	28	32
Leucopaxillus albissimus	38	22	21
Phaeoclavulina curta	30	0	0
Suillus grisellus	23	0	0
Wilcoxina mikolae	13	0	7
Cantharellus sp.	0	0	0
Leccinum sp.	0	0	0
Total sequences	16821	10811	11489
OTU count	25	20	23

The mock community was built by mixing equimolar amounts of DNA from each species and then processed in the same way as all other samples. All three datasets were treated identically in the bioinformatics quality filtering and OTU clustering using the multi-step full-linkage OTU clustering approach implemented in QIIME.

To pair or not to pair?

The use of paired end sequencing is becoming more popular amongst microbial ecologists because there is the potential to increase sequence quality since two quality scores inform each base (Masella *et al.*, 2012), and pairing algorithms are now common in NGS pipelines such as MOTHUR (Schloss *et al.*, 2009), QIIME (Caporaso *et al.*, 2010) and UPARSE (Edgar, 2013). In our recent experience, however, we have found that there can be tradeoffs to pairing because some taxa fail to pair successfully. To illustrate this point, we compared the results of single vs. paired direction sequences in our mock community using the QIIME analysis pipeline (Table 1). While we were able to recover 25 OTUs from single forward direction sequences, we were only able to recover 23 OTUs from paired direction sequences. Similarly, we were only able to recover 20 OTUs from the single reverse direction dataset,

which contained 30% fewer sequences than the single forward direction dataset (also reported by Caporaso *et al.*, 2011; Smith & Peay, 2014).

It appears that the failure to pair was largely due to the poor quality and quantity of the reverse direction sequences (we retained only high quality reads for the analyses presented in Table 1). This could be particularly seen for Ascomycota species (Table 1), which were well represented by single forward direction reads, but due to the poor quality of the reverse direction reads, a much smaller number of sequences were recovered when paired. Some Basidiomycota species (e.g. *Phaeoclavulina curta* and *Suillus grisellus*) also failed to pair even as we relaxed pairing parameters. In addition, the process of pairing of sequences can produce ambiguous bases (N) due to call conflicts. Since sequences containing Ns are typically filtered out, this reduced the total number of sequences in the paired dataset by 31% compared to the single forward direction dataset. In this example, we found that using single highest quality read direction (in this case just the forward direction reads) provided a more accurate picture of the underlying community than applying a paired sequence approach. We readily acknowledge that different sequencing runs and analysis pipelines can produce different results, so the approach taken in this example may not be best for all datasets. Instead, we stress the only accurate way to tell whether to pair or not pair sequence reads from individual datasets is to analyze them separately, using the mock community as an initial guide.

Ensuring accurate OTU clustering, filtering and identification

Being able to produce accurate OTUs from a dataset is a primary goal of all NGS-based studies and many OTU clustering strategies have been implemented in various bioinformatics pipelines: complete-linkage (furthest neighbor), average-linkage (average neighbor) and single-linkage (nearest-neighbor). The advantages and disadvantages to each of these strategies have been discussed elsewhere (Huse *et al.*, 2010; Lindahl *et al.*, 2013). We note here that combining different linkage strategies (also known as a chain-picking method) may sometimes recover a more accurate number of OTUs than using just a single strategy. For example, in our mock community, we began by using USEARCH, which recovered 29 OTUs, five of which belonged to the same OTU (*Hygrophorus russula*) but did not cluster. We then applied the UCLUST algorithm to the OTUs from the previous USEARCH step, which collapsed the five duplicate *H. russula* OTUs into one, thereby recovering the 25 total OTUs mentioned earlier. Although this approach worked well in this example, different clustering strategies or combinations of strategies will be sensitive to scaling issues (i.e. decreased accuracy of OTU clustering and exponentially increased computational time for much larger datasets). This complicates the ability to use a 'one-size-fits-all' clustering strategy, but the presence of a mock community will help to calibrate parameters to ensure the most accurate OTU clustering possible.

A result we feel is particularly important to highlight is that some OTUs generated by any OTU clustering strategy may not be biologically valid, even if they match with very high identity to a sequence in a database. In our mock community dataset, we found that 3.3% of the OTUs had good identity matches (95–100%) but only very short length matches to a database sequence (e.g. the query sequence matched only 6.5–36% of subject sequence). In comparing the unmatched part of the query sequence using BLAST, we found that it often either did not match to anything fungal or to any sequences in GenBank. As such, we think these sequences are likely chimeric artifacts, but since the UNITE SH database (Kõljalg *et al.* 2013) that we used to inform the chimera checking analyses only contains fungal sequences, these nonfungal regions were not detected (a database containing ITS sequences of all eukaryotic organisms would be more appropriate for chimera checking). Based on these results, we suggest it is important to explicitly consider both BLAST match length and identity of matched length for proper OTU filtering. One example addressing this problem was recently presented by Branco *et al.* (2013), who used a 95% match length-filtering step to remove these kinds of OTUs. In our own analyses, we have found that 85% match length seems to represent a reasonable balance between retaining good vs. spurious OTUs. Alternatively, an e-value or bit-score cutoff could be used to remove short sequences that have good identity matches.

The low OTU abundance dilemma

Popular analysis pipelines for next generation sequencing such as UPARSE (Edgar, 2013) suggest removing OTUs represented by a single sequence (i.e. singletons) because of the likely chance for errors when sequencing so deeply (see also Tedersoo *et al.* (2010) and Dickie (2010) for issues regarding retention or removal of singleton taxa). Importantly, we note certain OTU clustering strategies or combinations of strategies can also produce small numbers of false OTUs represented by *more* than one sequence. As such, the elimination of just singletons may not be sufficient data quality control. Here, again, the mock community can be used to determine how many sequences should be removed to best recover the actual community (Table 1). In our mock community example, one of the OTU clustering methods we used ('subsampled open reference OTU picking' in QIIME) produced two OTUs that were not actually present in the mock community. These extra OTUs had fewer than three sequences whereas the rest of the mock community OTUs had > 10 sequences. Thus, if we were to use this particular method of OTU clustering, we would remove any OTUs that had three or less sequences across the whole experimental dataset. While it did not happen in our example, the fungal OTUs in the positive control could potentially appear in low abundance (i.e. singletons) in other samples due to primer contamination or tag switching (see below). If this were to happen, these low abundance OTUs should of course also be removed from across all the experimental samples.

Incidence- or abundance-based results?

A number of molecular-based ecological datasets have shown that sequence abundance does not necessarily correlate well with tissue abundance across different species (Manter & Vivanco, 2007; Liti *et al.*, 2009; Amend *et al.*, 2010; Avis *et al.*, 2010; Egge *et al.*, 2013; Weber & Pawlowski, 2013). This was also reflected in our mock

community where despite combining equal amounts of DNA of all 27 species, we found that one OTU appeared three orders of magnitude higher in total sequence number than eight of the OTUs and two orders of magnitude > 16 of the OTUs, much like a rank-abundance curve of any natural community (Table 1). While we think that this issue (which could be due to factors such as unequal gene copy number in fungal genomes or taxon-specific PCR bias) is important to keep in mind, we suggest that analyses based on sequence abundance data often have ecological relevance. For example, the most abundant ECM fungal species on root tips often have the highest number of sequences in NGS datasets (Tedersoo *et al.*, 2010; P. Kennedy *et al.*, unpublished data; N. H. Nguyen *et al.*, unpublished data). Similarly, in mock communities containing different concentrations of known species across different samples, Amend *et al.* (2010) showed that sequence abundances generally scaled well with relative DNA concentration within but not between species (i.e. 'semi-quantitative'). Further, Smith & Peay (2014) compared results based on incidence- (i.e. presence/absence) and abundance-based data and showed that only using the former led to artificially high estimates of β-diversity when re-sequencing the same DNA extract. Based on these combined examples, we suggest conducting ecological analyses of fungal communities using both incidence- and abundance-based sequence data is a better approach than using only one or the other, as the results from one data type can help to inform the other. If using only abundance-based data is preferred due to concerns about the amount of information lost when transforming to incidence data, we remind researchers that a variety of data transformations (i.e. log, square-root) can be used to down weight the importance of more abundant OTUs before ecological analyses.

Overlooked contamination sources

Primer cross-contamination in a multiplexed library is a serious issue and should be discussed openly. In a sequencing library that has primer cross-contamination, a small number of sequences can be erroneously assigned to a different sample, potentially skewing the ecological interpretation of the data. Primer cross-contamination could happen at any stage, from oligonucleotide manufacturing to PCR. Primers maintained in plates, be they from the manufacturer or aliquots, have a greater chance of being contaminated due to repeated opening and closing of the sealing mats/film. We emphasize the importance of asking explicit questions about the chance of cross-contamination and purification costs for primers from the manufacturer before ordering. In addition, we suggest that primers be ordered in individual tubes and aliquots be made in tubes instead of plates to minimize the possibility of cross-contamination. Tag-switching (*sensu* Carlsen *et al.*, 2012) (where primer barcode tags from one sample may jump onto another sample during PCR) is a related issue, which can be accounted for by using primers tagged on both ends. Unfortunately, this would double the cost of primers, so may not be practical for the majority of researchers.

Another issue is the accidental inclusion of previously amplified DNA from one project in the post-PCR sample processing of a different project. Fortunately, for a laboratory where NGS is used for the first time, there is no chance for this kind of contamination. It will, however, become immediately relevant in laboratories that have built multiple libraries from the same primer sets (P. Kennedy *et al.*, unpublished data). While specifically accounting for post-PCR contamination is difficult (because controls at these steps cannot be easily parsed bioinformatically due to the absence of barcodes), careful additional laboratory hygiene (e.g. doing all post-PCR reactions with pipette tips with barriers, wiping down pipettors regularly with nuclease solutions, using a flow hood with frequent ultraviolet (UV) radiation sterilization) will help reduce this possibility.

Conclusion

Although the examples provided here are specific to fungal molecular ecology, we think that it could be broadly applied to other study systems using amplified markers. Specifically, we think researchers need to pay even closer attention to controls in NGS analyses in order to reduce as much noise from their datasets as possible. We stress that positive and negative controls give different important information about NGS-based data and that, for each dataset, the inclusion and independent examination of both is key to determining the best criteria used to generate the OTU tables used in ecological analyses. We recognize that more experienced researchers have already begun grappling with these issues, but we hope this letter will complement the 'User's Guide' by Lindahl *et al.* (2013) for researchers just starting to sequence fungi in ecological studies.

Acknowledgements

The authors thank Y. Lekberg, Leho Tedersoo and two anonymous reviewers for constructive comments on previous drafts of this manuscript.

Nhu H. Nguyen[1]*, Dylan Smith[2], Kabir Peay[2] and Peter Kennedy[1]

[1]Department of Plant Biology, University of Minnesota, Twin Cities, St Paul, MN 55108, USA
[2]Department of Biology, Stanford University, Palo Alto, CA 94305, USA
(*Author for correspondence: tel +1 612 624 8519; email nhnguyen@umn.edu)

References

Amend AS, Seifert KA, Bruns TD. 2010. Quantifying microbial communities with 454 pyrosequencing: does sequence abundance count? *Molecular Ecology* **19**: 5555–5565.

Avis PG, Branco S, Tang Y, Mueller G. 2010. Pooled samples bias fungal community descriptions. *Molecular Ecology Resources* **10**: 135–141.

Bates ST, Clemente JC, Flores GE, Walters WA, Parfrey LW, Knight R, Fierer N. 2013. Global biogeography of highly diverse protistan communities in soil. *The ISME Journal* **7**: 652–659.

Bik HM, Porazinska DL, Creer S, Caporaso JG, Knight R, Thomas WK. 2012. Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology & Evolution* **27**: 233–243.

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nature Methods* **10**: 57–59.

Branco S, Bruns TD, Singleton I. 2013. Fungi at a small scale: spatial zonation of fungal assemblages around single trees. *PLoS ONE* **8**: 1–10.

Buyck B, Hofstetter V. 2011. The contribution of tef-1 sequences to species delimitation in the *Cantharellus cibarius* complex in the southeastern USA. *Fungal Diversity* **49**: 35–46.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI *et al.* 2010. QIIME allows analysis of high-throughput community sequencing data. *Nature Methods* **7**: 335–336.

Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences, USA* **108**: 4516–4522.

Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. 2012. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? *Fungal Ecology* **5**: 747–749.

Debaud JC, Marmiesse R, Gay G 1999. Intraspecific genetic variation and populations of ectomycorrhizal fungi. In: Varma A, Hock B, eds. *Mycorrhiza*. Berlin, Germany: Spinger-Verlag, 75–110.

Dickie I. 2010. Insidious effects of sequencing errors on perceived diversity of molecular surveys. *New Phytologist* **188**: 916–918.

Edgar RC. 2013. UPARSE: highly accurate OTU sequences from microbial amplicon sequences. *Nature Methods* **10**: 1–3.

Egge E, Bittner L, Andersen T, Audic S, de Vargas C, Edvardsen B. 2013. 454 Pyrosequencing to describe microbial eukaryotic community composition, diversity and relative abundance: a test for marine haptophytes. *PLoS ONE* **8**: e74371.

Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12**: 1889–98.

Ihrmark K, Bödeker ITM, Cruz-Martinez K, Friberg H, Kubartova A, Schenck J, Strid Y, Stenlid J, Brandström-Durling M, Clemmensen KE *et al.* 2012. New primers to amplify the fungal ITS2 region – evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiology Ecology* **82**: 666–677.

Kõljalg U, Nilsson RH, Abarenkov K, Tedersoo L, Taylor AFS, Bahram M, Bates ST, Bruns TD, Bengtsson-Palme J, Callaghan TM *et al.* 2013. Towards a unified paradigm for sequence-based identification of fungi. *Molecular Ecology* **22**: 5271–5277.

Lindahl DB, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjoller R, Kõljalg U, Pennanen T, Rosendahl S, Stenlid J *et al.* 2013. Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist* **199**: 288–299.

Liti G, Carter DM, Moses AM, Warringer J, Parts L, James SA, Davey RP, Roberts IN, Burt A, Koufopanou V *et al.* 2009. Population genomics of domestic and wild yeasts. *Nature* **458**: 337–341.

Manter DK, Vivanco JM. 2007. Use of the ITS primers, ITS1F and ITS4, to characterize fungal abundance and diversity in mixed-template samples by qPCR and length heterogeneity analysis. *Journal of Microbiological Methods* **71**: 7–14.

Masella AP, Bartram AK, Truszkowski JM, Brown DG, Neufeld JD. 2012. PANDAseq: paired-end assembler for illumina sequences. *BMC Bioinformatics* **13**: 31.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ *et al.* 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**: 7537–7541.

Smith DP, Peay KG. 2014. Sequence depth, not PCR replication, improves ecological inference from next generation DNA sequencing. *PLoS ONE* **9**: e90234.

Tedersoo L, May TW, Smith ME. 2010. Ectomycorrhizal lifestyle in fungi: global diversity, distribution, and evolution of phylogenetic lineages. *Mycorrhiza* **20**: 217–263.

Van Der Linde S, Haller S. 2013. Obtaining a spore free fungal community composition. *Fungal Ecology* **6**: 522–526.

Vik U, Logares R, Blaadid R, Halvorsen R, Carlsen T, Bakke I, Kolsto A-B, Okstad OA, Kauserud H. 2013. Different bacterial communities in ectomycorrhizae and surrounding soil. *Scientific Reports* **3**: 3471.

Weber A-TA, Pawlowski J. 2013. Can abundance of protists be inferred from sequence data: a case study of foraminifera. *PLoS ONE* **8**: e56739.

Yamamoto N, Bibby K. 2014. Clustering of fungal community internal transcribed spacer sequence data obscures taxonomic diversity. *Environmental Microbiology*. doi: 10.1111/1462-2920.12390.

Key words: control, ecology, fungi, mock community, next generation sequencing (NGS).