



Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities



Journal:	<i>New Phytologist</i>
Manuscript ID:	Draft
Manuscript Type:	L - Letter
Date Submitted by the Author:	n/a
Complete List of Authors:	Nilsson, Henrik; University of Gothenburg, Plant and Environmental Sciences Tedersoo, Leho; University of Tartu, Institute of Botany and Ecology Lindahl, Björn; Swedish University of Agricultural Sciences, Forest Mycology and Pathology Kjøller, Rasmus; University of Copenhagen, Biological Institute, Terrestrial Ecology Carlsen, Tor; University of Oslo, Department of Biology Quince, Christopher; University of Glasgow, School of Engineering Abarenkov, Kessy; University of Tartu, Institute of Botany and Ecology Pennanen, Taina; The Finnish Forest Research Institute, Vantaa Research Unit Stenlid, Jan; Swedish University of Agricultural Sciences, Forest Mycology and Pathology Bruns, Thomas; University of California Berkeley, Plant and Microbial Biology Larsson, Karl-Henrik; University of Oslo, The Mycological Herbarium of the Natural History Museum Koljalg, Urmas; University of Tartu, Institute of Botany and Ecology Kausrud, Håvard; University of Oslo, Department of Biology
Key Words:	next-generation sequencing, environmental sampling, fungal communities, standardization, fungal-plant interactions

SCHOLARONE™
Manuscripts

**Towards standardization of the description and publication of next-generation
sequencing datasets of fungal communities**

R. Henrik Nilsson^{1,2,*}, Leho Tedersoo^{2,3}, Björn Lindahl⁴, Rasmus Kjølner⁵, Tor Carlsen⁶,
Christopher Quince⁷, Kessy Abarenkov², Taina Pennanen⁸, Jan Stenlid⁴, Tom Bruns⁹, Karl-
Henrik Larsson¹⁰, Urmas Kõljalg^{2,3}, Håvard Kauserud⁶

¹ Department of Plant and Environmental Sciences, University of Gothenburg. Box 461, 405
30 Göteborg, Sweden

² Institute of Ecology and Earth Sciences, University of Tartu. 46 Vanemuise St. 51014 Tartu,
Estonia

³ Natural History Museum of Tartu University. 46 Vanemuise St. 51014 Tartu, Estonia

⁴ Department of Forest Mycology and Pathology, Swedish University of Agricultural
Sciences, Box 7026, 750 07 Uppsala

⁵ Biological Institute, Terrestrial Ecology, University of Copenhagen, Øster Farimagsgade
2D, DK-1353 Copenhagen, Denmark

⁶ Department of Biology, University of Oslo, P.O. Box 1066 Blindern, N-0316 Oslo, Norway

⁷ School of Engineering, University of Glasgow, Glasgow G12 8LT, United Kingdom

⁸ The Finnish Forest Research Institute, PL 18, FI-01301 Vantaa, Finland

⁹ Department of Plant and Microbial Biology, University of California, Berkeley, CA 94720,
USA

¹⁰ The Mycological Herbarium, Natural History Museum, University of Oslo, P.O. Box 1172,
Blindern, N-0318 Oslo, Norway

* author for correspondence: henrik.nilsson@dpes.gu.se, Tel. +46-31-7862623, Fax. +46-31-
7862560

Word count: 2226

Number of figures: 0

Number of tables: 0

Number of references: 20

Fungi play fundamental roles as agents of decomposition and nutrient recycling in most terrestrial ecosystems. The fact that fungi spend most of their life cycle below ground or within substrates has left the scientific community with a fragmentary understanding of fungal diversity, and a modest ~7% of the estimated 1.5 million extant species of fungi have been described (Kirk *et al.*, 2008). The poor correlation between the presence of fungal fruiting bodies or other macroscopic structures and the full diversity of the mycobiome at any sample site has shifted the focus in fungal ecology from fruiting bodies to molecular (DNA sequence) data, and nearly all recent attempts to characterize fungal communities are based on sequence data (Taylor, 2008). Such studies have hitherto been limited in sequence depth by the high cost and investment of effort associated with traditional Sanger sequencing of large numbers of samples, but recent methodological progress in the form of next-generation sequencing technologies (NGS; Shendure & Ji, 2008) offers a remedy to these problems. One of these NGS - massively parallel (“454”) pyrosequencing (Margulies *et al.*, 2005) - has the capacity to generate more than a million sequences of ~500 base-pairs (bp.) length in the course of a day, making it a groundbreaking tool for environmental sequencing of fungi.

For all the research venues opened by the NGS in general and pyrosequencing in particular, the technologies remain fairly complicated and may, in the absence of generally acknowledged standards, even prove counterproductive to fungal ecology and mycology at large. Various types of incompletely understood biases are introduced at different steps of the analyses (Quince *et al.*, 2009; Bellemain *et al.*, 2010; Tedersoo *et al.*, 2010), and these are often paid little attention to during the interpretation of the results. Approaches to delimitation of species or operational taxonomic units (OTUs) from molecular data differ widely among users, as do ideas on how to handle abundance data, taxonomic standards, and ecological classifications (Hibbett *et al.*, 2011). New web-based software has been developed specifically for sequence processing and analysis of fungal pyrosequencing data - *e.g.*, CLOTU (<http://www.bioportal.uio.no>), SCATA (<http://scata.mykopat.slu.se>), and PlutoF (<http://unite.ut.ee>) - and these resources approach sequence clustering and identification in different ways. Other areas where standards are lacking include what data and files to make available with the study in question and what estimates, statistics, and level of detail to report as part of the results. As a consequence, most NGS studies measure and quantify slightly to fundamentally different things and often do so in ways that are neither very clear to the reader nor directly amenable to independent repetition and verification.

Hibbett *et al.* (2011) provided an overview of the ten first 454-based studies of fungal communities. While each study represented a significant achievement, differences in

the nature and level of detail reported made precise comparison difficult. In more than half of the cases, email exchange with the corresponding authors was necessary for clarification, verification, and access to additional data. If allowed to continue unabated, the heterogeneous, non-standardized reporting of NGS data on fungal communities will prevent the detailed comparisons of communities and biological processes that the NGS technologies were hoped to enable. That is not in anyone's interest, and in this letter we introduce a set of core elements we feel every NGS-based study of fungal communities should report on. A standard for how environmental NGS data should be *generated and analyzed* is probably not warranted – or even desirable – at this stage, since many aspects of the processing and analysis of NGS data remain tentative and are likely to vary with the scope of the study. Our proposals are instead oriented towards how data and results should be *reported and made available to the scientific community*. We hope that they will be considered a lower bound on the level of detail necessary, although we recognize that it may not be possible to generate all of them for every NGS dataset.

Detailed comparisons and meta-analyses of studies are only possible if the underlying data are made available for download. This is not always the case, however, necessitating email exchange with authors regarding files that may no longer even exist (Whitlock *et al.*, 2010). To make the data available through the authors' personal web pages is similarly a makeshift solution that does not meet the criterion of long-term availability to the scientific community (Wren, 2008). We propose that all data relevant to the reanalysis and interpretation of fungal NGS studies should be deposited in central data archives. The European Nucleotide Archive (Leinonen *et al.*, 2011) is the recommended resource for storage of raw NGS data, including flowgram ("SFF") and unprocessed FASTA files in the case of pyrosequencing datasets. In addition, we propose that all the processed files underlying – and forming parts of - the results of any given NGS study should be deposited at the publisher's site as supplementary data along with the article presenting the study in question. Advances in data compression as implemented in, e.g., 7-Zip (<http://sourceforge.net/projects/sevenzips/>) will typically serve to reduce even large files to below the 15 MB-limit imposed by many journals, and as storage gets cheaper journals should expand this arbitrary limit.

The boldfaced bulleted subheadings presented below represent elements we feel all NGS studies of fungal communities should report on in a clear and comprehensive way. One or more examples are given for each subheading; they remain examples however and should not

necessarily be seen as recommendations of methodology or specific software packages. We have no opinion on the exact form in which the elements should be reported in a publication, and we envision the below as a checklist rather than as a mandatory table. The elements are grouped into logical units although some overlap is acknowledged.

Description of sample site and laboratory procedures

For the description of the sample site and sampling conditions, we propose that the MIMARKS/MIENS standard (Yilmaz *et al.*, 2010) should be followed. Insofar as the specifics of each sample differ, we argue that full metadata should be given for each sample. We advocate that the laboratory procedures should be described in comprehensive detail – including full primer sequence data, PCR enzyme specifics, and other points routinely left out by many authors – and we discourage the practice of referring to other articles instead of providing the corresponding information in writing.

Next generation sequencing platform

Most current NGS studies of fungal communities rely on pyrosequencing, but this is a situation that could change. The exact machine and protocol used should be detailed. *Example:* The samples were sequenced using a 454 GSFLX machine with the Titanium protocol.

Sequence data: filtering, denoising, and availability

*** Raw sequence data file.** *Example:* The 454 SFF file and the corresponding unprocessed FASTA and tag/barcode files were deposited in the European Nucleotide Archive as ERR00000X.

*** Filtering and trimming.** *Example:* Sequences with more than 2% ambiguity symbols were discarded. Leading and trailing, but not intercalary, ambiguity symbols were pruned in Flower 0.7 (<http://biohaskell.org/Applications/Flower>) prior to analysis. All sequences shorter than 250 bp. after the removal of barcodes, tags, and primers were discarded. All sequences longer than 450 bp. were trimmed down to 450 bp. from the 3' end.

*** Sequence denoising.** *Example:* AmpliconNoise 1.2 (Quince *et al.*, 2011) was used to denoise the entries; default settings were used.

*** Total number of discarded and retained sequences.** *Example:* Out of a total of 140,000 sequences, 18,234 were found to be of poor read quality, 15,213 too short, and 5,430 potentially chimeric, leaving 101,123 (72%) sequences for downstream analyses.

* **Sequencing depth.** *Examples:* Of the 101,123 sequences retained, 24,832 represented sample A, 31,323 sample B, 25,118 sample C, and 19,850 sample D. The number of sequences per sample ranged from 4,201 to 7,912 in the 15 samples.

* **FASTA files.** *Example:* All sequences passing the filtering step are available in the FASTA format as Supplementary Item X.

Sequence data: analysis and taxonomic assignment

Fully automated, high-quality species identification from fungal sequence data is presently not possible for any non-trivial assemblage of fungal lineages, leaving caution and taxonomic expertise as important elements in molecular identification of fungi. In particular we wish to discourage the use of single, static similarity thresholds for species demarcation as far as possible; these thresholds should be allowed to vary to better account for differences among fungal lineages. Any specific threshold value tailored for the ITS region will typically carry over poorly to other genetic markers, such as the ribosomal small subunit. The molecular identification procedure is underspecified in many NGS studies, making independent repetition difficult.

* **Details of the genetic marker used.** *Examples:* The full ITS1 of the nuclear ITS region as extracted using Nilsson *et al.* (2010) was used. The V2 and V3 regions – and their conserved intercalary segment - of the 18S was used (Hartmann *et al.*, 2010). Sequences covering less than 75% of the full length of the target region were discarded.

* **Type and specifics of sequence clustering.** *Examples:* Single-linkage clustering in BLASTclust 2.2.22 (Altschul *et al.*, 1997) was carried out at 97.0% similarity over $\geq 90.0\%$ of the length of the shortest sequence in the underlying pairwise alignments. Complete-linkage clustering at 98.5% similarity (global alignment) was performed in UCLUST 4.0 (Edgar, 2010) with the most abundant sequence types serving as cluster seeds.

* **Sequence data used for taxonomic annotation.** *Example:* The most frequent sequence type in each cluster was used for the BLAST searches, and the corresponding FASTA file is provided as Supplementary Item X.

* **Specification of the taxonomic reference database.** *Examples:* All fully identified entries in INSD (Benson *et al.*, 2011) and UNITE (Abarenkov *et al.*, 2010) as of December 2010 were used as reference sequences.

* **Specification of the taxonomic annotation procedure.** *Example:* BLAST 2.2.22 was used. $\geq 97\%$ similarity across the entire length of the pairwise alignment was taken to indicate

conspicuity; however, $\geq 99\%$ was required in the cases of *Cortinarius*, *Aspergillus*, and *Penicillium*. If only one reference sequence was available for some given species, or if the taxonomic annotation was contradicted by another, equally close reference sequence, an asterisk and a question mark, respectively, was added to the taxonomic annotation of the sequence. $\geq 90\%$ similarity was taken to approximate the genus level (e.g., *Hydnum* sp.) and $\geq 60\%$ similarity the ordinal level (e.g., *Boletales* sp.). Only sequences determined at least to ordinal level were used for the phylum-level comparison of Fig. X. Sequences not having a $\geq 60\%$ BLAST match to any reference sequence were considered potentially compromised; were marked as such; and were excluded from the ecological statistics but not from the list of OTUs recovered.

*** Handling of singletons and OTUs with few reads.** *Examples:* All singletons were discarded. All OTUs with fewer than 5 reads were excluded from further analysis. Only singletons at least 90% identical to the reference sequence of a species not otherwise recovered in the study were kept.

Post-clustering/taxonomic results

*** Count of OTUs recovered.** *Example:* A total of 242 non-singleton OTUs were recovered. Forty-two were unique to single samples, and the rest were shared by two or more samples (Supplementary Item X).

*** List of OTUs recovered.** *Example:* A complete annotated list of the abundance of all fungal OTUs recovered in each sample is provided in the QIIME format (http://qiime.sourceforge.net/documentation/file_formats.html) as Supplementary Item X.

*** Taxonomic affiliations.** The dynamic nature of fungal taxonomy, coupled with the many pitfalls of molecular identification, suggest that any taxonomic affiliations derived should be kept separate from the name (identifier) of the OTUs. *Example:* The OTUs were identified to species, genus, or order as applicable, and the taxonomic affiliations are provided as a separate file (Supplementary Item X).

*** Proportion of fully identified OTUs.** *Example:* We tentatively identified 72 (30%) of the 242 OTUs to species level. Of the remaining 170 OTUs, we tentatively identified 60 to genus and 88 to order level.

*** Phylum-level distribution.** *Example:* 97% of the OTUs were of fungal origin. Of these, 58% of were basidiomycetes; 22% were ascomycetes; 12% were glomeromycetes; and 8% were found to belong to other fungal lineages.

Discretionary consideration

Molecular identification of fungi is fraught with methodological complications, but above all it is severely hampered by the lack of reference sequences for much of the extant diversity of fungi. It would seem appropriate to plan each NGS-study of environmental samples so that taxonomic expertise is accounted for among, or available to, the authors of the study. If also some part of the budget could be allocated to generating reference sequences from fruiting bodies relevant for the ecosystem and geographical region under study, each new NGS study would contribute to the reference sequence databases and ultimately the possibilities to disentangle the diversity discovered (Brock *et al.*, 2009).

Conclusion

Each NGS study represents a considerable investment in terms of time and money, and for that investment to be of maximum use to the broader scientific community, the data generated and results obtained should be presented and made available in a comprehensive, transparent way. We believe the elements discussed above are a significant contribution to the development of such a standard, and their specification is unlikely to take more than a few hours. The NGS is the most exciting development in fungal ecology for many years, and correctly employed it will enable great strides to be made towards a much deeper understanding of fungi and their trophic roles in ecosystems.

References

- Abarenkov K, Nilsson RH, Larsson K-H, Alexander IJ, Eberhardt U, Erland S, Høiland K, Kjoller R, Larsson E, Pennanen T et al. 2010.** The UNITE database for molecular identification of fungi - recent updates and future perspectives. *New Phytologist* 186: 281-285.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, 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.
- Bellemain E, Carlsen T, Brochmann C, Coissac E, Taberlet P, Kauserud H. 2010.** ITS as an environmental DNA barcode for fungi: an *in silico* approach reveals potential PCR biases. *BMC Microbiology* 10: 189.

- Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. 2011. GenBank. *Nucleic Acids Research* 39: D32-D37
- Brock PM, Döring H, Bidartondo MI. 2009. How to know unknown fungi: the role of a herbarium. *New Phytologist* 181: 719-724.
- Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460-2461.
- Hartmann M, Howes CG, Abarenkov K, Mohn WW, Nilsson RH. 2010. V-Xtractor: An open-source, high-throughput software tool to identify and extract hypervariable regions of small subunit (16S/18S) ribosomal RNA gene sequences. *Journal of Microbiological Methods* 83: 250-253.
- Hibbett DS, Ohman A, Glotzer D, Nuhn M, Kirk P, Nilsson RH. 2011. Progress in molecular and morphological taxon discovery in *Fungi* and options for formal classification of environmental sequences. *Fungal Biology Reviews*, in press. doi:10.1016/j.fbr.2011.01.001
- Kirk PM, Cannon PF, Minter DW, Stalpers JA. 2008. *Dictionary of the Fungi*, 10th Edition. Wallingford, UK: CABI Publishing.
- Leinonen R, Akhtar R, Birney E, Bower L, Cerdano-Tárraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R et al. 2011. The European Nucleotide Archive. *Nucleic Acids Research* 39: D28-D31.
- Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen Y-J, Chen Z et al. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437: 376-380.
- Nilsson RH, Veldre V, Hartmann M, Unterseher M, Amend A, Bergsten J, Kristiansson E, Ryberg M, Jumpponen A, Abarenkov K. 2010. An open source software package for

- 267 automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput
 268 community assays and molecular ecology. *Fungal Ecology* 3: 284-287.
 269
- 270 **Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT.**
 271 **2009.** Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature*
 272 *Methods* 6: 639-641.
 273
- 274 **Quince C, Lanzén A, Davenport RJ, Turnbaugh PJ. 2011.** Removing noise from
 275 pyrosequenced amplicons. *BMC Bioinformatics* 12: 38.
 276
- 277 **Shendure J, Ji H. 2008.** Next-generation DNA sequencing. *Nature Biotechnology* 26: 1135-
 278 1145.
 279
- 280 **Taylor AFS. 2008.** Recent advances in our understanding of fungal ecology. *Coolia* 52: 197-
 281 212.
 282
- 283 **Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem**
 284 **E, Chuyong G, Kõljalg U. 2010.** 454 Pyrosequencing and Sanger sequencing of tropical
 285 mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New*
 286 *Phytologist* 188: 291-301.
 287
- 288 **Whitlock MC, McPeck MA, Rausher MD, Rieseberg L, Moore AJ. 2010.** Data archiving.
 289 *American Naturalist* 175: 145-146.
 290
- 291 **Wren JD. 2008.** URL decay in MEDLINE—a 4-year follow-up study. *Bioinformatics* 24:
 292 1381-1385.
 293
- 294 **Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA,**
 295 **Karsch-Mizrachi I, Johnston A, Cochrane G et al. 2010.** The “Minimum Information
 296 about an ENvironmental Sequence” (MIENS) specification, version 2. *Nature Precedings*
 297 doi:10.1038/npre.2010.5252.2
 298
 299

300 **Key words:** next-generation sequencing, environmental sampling, fungal communities,
301 standardization

For Peer Review