Technical considerations

Rationale

Prokaryotic environmental sampling relies to a great extent on the ribosomal small subunit (SSU: 16S/18S) gene (Bengtsson et al. 2012). The SSU is typically alignable across orders and phyla, and most of the analysis and quality control tools that have been developed for the SSU rely on global alignments featuring the full query dataset. The ITS1 and ITS2 spacers of the fungal ITS region are however highly variable, and reliable ITS alignments can often not be built above the genus level in fungi. This means that most quality management software used for SSU sequences cannot be readily applied to fungal ITS sequences. Some few tools have been built specifically for the ITS region, and we have also found a handful of other, general-purpose tools to be quite useful for ITS purposes. Below we will discuss these tools for the user who seeks a greater degree of automation than that offered by the guidelines in the manuscript.

Sequence clustering

Sequence clustering is something of a research field in its own, and many disparate clustering algorithms with different features (with respect to, e.g., alignment strategy, distance measure, and linkage methods) are available; there is no single, optimal clustering tool for all situations. One trick used to single out deviant – as well as incorrectly purported - ITS sequences is nevertheless clustering. An on-line implementation of BLASTclust was recommended in the manuscript due to its very low learning threshold. Although BLASTclust will do a reasonable job at the task, it has a tendency to create overly inclusive clusters. The reason is that it does single-link clustering based on local alignment (Altschul et al. 1997). Thus, deviant sequences like chimeras can occasionally still be grouped into clusters with non-chimeras if they contain enough sequence data from one of the parent sequences. The UCLUST tool available through the USEARCH software suite (Edgar 2010) is a stringent clustering program that employs complete-link clustering based on global alignment, and we have yet to see it force a (badly) chimeric sequence into a cluster of non-chimeric ones. As a method to single out deviant sequences it is preferable to BLASTclust. It is however a technical software tool far from the paste-and-click type, and some scripting is often required to parse the output to the user's taste.

Hidden Markov models

While the ITS1 and ITS2 spacers are highly variable and thus difficult to work with from an

alignment-based point of view, the surrounding SSU and LSU genes – as well as the intercalary 5.8S - are more conserved. The variation in these genes can be represented in hidden Markov models (HMMs; Eddy 2011), and newly generated sequences can be run against these HMMs using, e.g., HMMER (http://hmmer.janelia.org) to examine which, if any, of these genes are present in the sequences. The Fungal ITS Extractor (Nilsson et al. 2010b) is built on this principle and allows the user to examine newly generated ITS sequences for coverage of the ITS region. The software will also indicate sequences for which none of these genes can be found, and it will extract the ITS1 and ITS2 from the query sequences. These extracted ITS1/ITS2 spacers make it easy for the user to subject only the variable components of the ITS region to sequence clustering and BLAST searches, a move that bypasses the potentially negative effects of query sequences of different degree of coverage and the presence of very conserved elements (e.g., the SSU or 5.8S) in the queries (cf. Kang et al. 2010). The Fungal ITS Reverse Complementary Checker (Nilsson et al. 2011b) similarly uses HMMs to examine ITS datasets for the presence of reverse complementary sequences and to reorient any such cases if found. Both these tools share a disadvantage: whereas the HMMs employed perform well with the Kingdom Fungi as a whole, they do not work well for fungi with very deviant ribosomal genes, notably *Cantharellus*, *Craterellus*, and *Tulasnella* (Moncalvo et al. 2006; Taylor and McCormick 2008). Complete revisions of these tools, coupled with a much extended set of HMMs, are meant to fill these gaps; a 2012 release of both tools is under way.

Chimeras

Chimeric sequences have emerged as a serious problem in the wake of environmental sequencing. UNITE has a record of about 1,000 chimeric fungal ITS sequences, but since most of these are particularly striking examples of chimeras, they are likely just the tip of the iceberg. They stem from a total of 250 studies, most of which are of the environmental sequencing type, and many of which employed cloning. Users with such cloning and sequencing datasets should take advantage of available chimera control tools. UCHIME (Edgar et al. 2011) is an all-purpose chimera checker that we have found to work well for fungal ITS sequences. It has two main modes of operation: a *de novo* mode in which chimeras are detected based on the query sequences only, and a database mode where chimeras are inferred based on comparison to a user-provided inclusive, chimera-free reference database. One could argue that the only ITS dataset that is inclusive enough is INSD itself, which however is not a chimera-free dataset. The UNITE release of the fungal ITS sequences in

INSD/UNITE features the same sequences (plus some 4,000 UNITE core sequences), however with a total of 4,200 INSD sequences excluded for various quality-related reasons (including chimeras) and with a total of 13,500 taxonomic reannotations. This dataset can be downloaded as a FASTA file from http://unite.ut.ee/repository.php. An alternative fungal ITS chimera checker was released by Nilsson et al (2010a). It establishes potential chimeras through extracting the ITS1 and ITS2 from the queries; carrying out separate BLAST searches for each of these; and checking that both respective matches stem from the same fungal order. This tool cannot find chimeras between species of the same order, but in return it is not very sensitive to taxon sampling within orders and does not require that any of the parent sequences be present in the reference dataset for a chimera to be detected. A drawback is that it cannot readily be used to find chimeras that are already in INSD, since it uses this dataset as reference. MOTHUR (Schloss et al. 2009) is a software environment that offers several methods for chimera control, although these are to some extent oriented towards the ribosomal small subunit of prokaryotes.

454 pyrosequences

Neither INSD nor UNITE store full 454 pyrosequencing datasets as primary sequences. There are nevertheless 454 sequences in both these databases; some are released as representative sequences for OTUs whereas others were snuck in by their authors. This means that pyrosequences and regular database users will cross roads once in a while. In these situations it should be kept in mind that individual pyrosequences do not amount to much. In particular, individual sequences from datasets that were not carefully quality-controlled using software suites (e.g., AmpliconNoise (Quince et al. 2011) or Denoiser (Reeder and Knight 2010)) tailored specifically for denoising should always be handled with care. Simpler approaches to quality control of pyrosequences - such as only trimming sequences by their Phred scores (e.g., Schmieder and Edwards 2011) - should ideally be avoided in amplicon-based pyrosequencing studies, and any such sequences in the public databases should be treated with some degree of skepticism. A common practice in pyrosequencing-powered studies is to compute OTUs from the sequences, and then to let the computer pick a representative sequence (typically the most common sequence type) from each OTU (cf. Nilsson et al. 2011a). Such representative sequences are less likely to be compromised in terms of quality - particularly if the OTU holds say ten or more sequences - and their occasional use by INSD/UNITE visitors would seem legitimate to us. As many previous studies have noted, pyrosequencing singletons are highly likely to be compromised (e.g., Tedersoo et al. 2010).

The user should also keep in mind that the pyrosequencing technology struggles with homopolymer-rich regions (e.g., ...AAAAAAA...), and that pyrosequences that differ only in their homopolymeric regions may represent nothing but technical noise inherent to the technology (Huse et al. 2007; Balzer et al. 2011). CrunchClust (http://code.google.com/p/crunchclust/ ; Hartmann et al. 2012) is a clustering tool that does not erect OTUs based on differences in homopolymer regions only. It computes clusters through exact Needleman-Wunsch pairwise alignments (Needleman and Wunsch 1970) and the Levenshtein distance (Levenshtein 1966), forming – as we see it - a welcome step away from hard-coded percentage thresholds as arbiters of OTU inclusiveness. CROP (Hao et al. 2011) is another new clustering tool that offers promise in this regard.

References

Altschul SF, Madden TL, Schäffer 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.
doi: 10.1093/nar/25.17.3389

Balzer S, Malde K, Jonassen I (2011) Systematic exploration of error sources in pyrosequencing flowgram data. Bioinformatics 27: i304-i309.
doi: 10.1093/bioinformatics/btr251

Bengtsson J, Hartmann M, Unterseher M et al. (2012) Megraft: a software package to graft ribosomal small subunit (16S/18S) fragments onto full-length sequences for accurate species richness and sequencing depth analysis in pyrosequencing-length metagenomes and similar environmental datasets. Research in Microbiology (in press).
doi: 10.1016/j.resmic.2012.07.001

Eddy SR (2011) Accelerated profile HMM searches. PLoS Computational Biology 7: e1002195.
doi:10.1371/journal.pcbi.1002195

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26: 2460-2461.

doi: 10.1093/bioinformatics/btq461

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27: 2194-2200.
doi: 10.1093/bioinformatics/btr381

Hao X, Jiang R, Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. Bioinformatics 27(5): 611-618.
doi: 10.1093/bioinformatics/btq725

Hartmann M, Howes CG, VanInsberghe D et al. (2012) Significant and persistent impact of timber harvesting on soil microbial communities in northern coniferous forests (in press).

Huse SM, Hiber JA, Morrison HG, Sogin ML, Welch DM (2007) Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biology 8: R143.
doi:10.1186/gb-2007-8-7-r143

Kang S, Mansfield MA, Park B et al. (2010) The promise and pitfalls of sequence-based identification of plant pathogenic fungi and oomycetes. Phytopathology 100: 732-737.
doi: 10.1094/PHYTO-100-8-0732

Levenshtein V (1966) Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady 10: 707-710.

Moncalvo J-M, Nilsson RH, Koster B et al. (2006) The cantharelloid clade: dealing with incongruent gene trees and phylogenetic reconstruction methods. Mycologia 98: 937-948.

Needleman SB, Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48: 443–453.
doi:10.1016/0022-2836(70)90057-4

Nilsson RH, Abarenkov K, Veldre V, Nylinder S, De Wit P, Brosche S, Alfredsson JF, Ryberg M, Kristiansson E (2010a) An open source chimera checker for the fungal ITS region. Molecular Ecology Resources 10: 1076-1081.

doi: 10.1111/j.1755-0998.2010.02850.x

Nilsson RH, Veldre V, Hartmann M et al. (2010b) An open source software package for automated extraction of ITS1 and ITS2 from fungal ITS sequences for use in high-throughput community assays and molecular ecology. Fungal Ecology 3: 284-287.
doi: 10.1016/j.funeco.2010.05.002

Nilsson RH, Tedersoo L, Lindahl BD et al. (2011a) Towards standardization of the description and publication of next-generation sequencing datasets of fungal communities. New Phytologist 191(2): 314-318.
doi: 10.1111/j.1469-8137.2011.03755.x

Nilsson RH, Veldre V, Wang Z et al. (2011b) A note on the incidence of reverse complementary fungal ITS sequences in the public sequence databases and a software tool for their detection and reorientation. Mycoscience 52: 278-282.
doi: 10.1007/s10267-010-0086-z

Reeder J, Knight R (2010) Rapid denoising of pyrosequencing amplicon data: exploiting the rank-abundance distribution. Nature Methods 7: 668-669.
doi:10.1038/nmeth0910-668b

Schloss PD, Westcott SL, Ryabin T et al. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75(23): 7537-7541.
doi: 10.1128/AEM.01541-09

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27: 863-864.
doi: 10.1093/bioinformatics/btr026

Taylor DL, McCormick MK (2008) Internal transcribed spacer primers and sequences for improved characterization of basidiomycetous orchid mycorrhizas. New Phytologist 177: 1020–1033.
doi: 10.1111/j.1469-8137.2007.02320.x

Tedersoo L, Abarenkov K, Nilsson RH et al. (2011) Tidying up International Nucleotide Sequence Databases: ecological, geographical, and sequence quality annotation of ITS sequences of mycorrhizal fungi. PLoS ONE 6: e24940.
doi: 10.1371/journal.pone.0024940

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12: 38.
doi: 10.1186/1471-2105-12-38