Letter

# Moving beyond *de novo* clustering in fungal community ecology

## Introduction

High throughput sequencing (HTS) has rapidly become the *de facto* tool for characterizing microbial community structure in a wide variety of habitats (Caporaso *et al.*, 2011; Peay *et al.*, 2016; Truong *et al.*, 2017). Accompanying the expanding use of HTS to quantify microbial diversity is the need to delineate species, the ecological unit traditionally used to compare the richness and composition of communities across treatments, locations or habitats (Magurran, 2005). Due to the challenges in identifying microbial species using morphology or biology alone, designations are typically made by 'binning' DNA sequences that meet a similarity threshold into operational taxonomic units (OTUs; Blaxter *et al.*, 2005). Currently, the most widely employed approach for defining fungal OTUs is done according to similarities among sequences within the dataset (Supporting Information Fig. S1). Commonly referred to as *de novo* clustering (Bik *et al.*, 2012), this approach requires no input database as a reference, which is advantageous when characterizing communities with little *a priori* knowledge. Despite this benefit, the ecological insights gleaned from *de novo* clustering can be limited by the challenge of directly comparing OTU identity across different studies (Öpik *et al.*, 2014), and the coarse phylogenetic resolution of many taxonomic assignments (Halwachs *et al.*, 2017).

One alternative to *de novo* clustering is the closed reference approach, where OTUs are binned according to sequence similarity of those in a reference database. With this approach, both OTU clustering and taxonomic designations occur simultaneously. Although the use of closed reference clustering in fungal ecology has been scarce (Fig. S1), it has become increasingly common in the molecular characterization of arbuscular mycorrhizal (AM) fungal communities as well as in many bacterial 'microbiome' studies (Öpik *et al.*, 2014; Kelly *et al.*, 2016). The relatively low taxonomic and phylogenetic diversity of AM fungal communities (Stajich *et al.*, 2009; Redecker *et al.*, 2013; Davison *et al.*, 2015), combined with a curated database (Öpik *et al.*, 2010) and increasingly wide usage of the 18S rRNA gene for molecular characterization (Öpik *et al.*, 2014), may explain why AM fungal community ecologists (relative to other fungal ecologists) have readily embraced closed reference clustering. Notably, the closed reference clustering approach has contributed significant new ecological understanding to patterns of AM community assembly by tracking OTUs (referred to as VT, Öpik *et al.*, 2010) across studies with both contrasting habitat types and a wide variety of spatial scales (Davison *et al.*, 2015; García de León *et al.*, 2016). A second alternative to *de novo* clustering is an open reference approach, which first clusters sequences to a reference database, followed by *de novo* clustering of the remaining unmatched sequences. This hybrid approach can combine the advantages of the two aforementioned clustering approaches (Rideout *et al.*, 2014; He *et al.*, 2015), but its interpretation can be problematic if the OTU definitions between closed reference and *de novo* approaches differ. Although open reference clustering is the least commonly used in fungal community ecology analyses to date (Fig. S1), it has been employed in studies of both arbuscular mycorrhizal and ectomycorrhizal fungal communities (Dumbrell *et al.*, 2010; Jarvis *et al.*, 2015).

The increasingly widespread adoption of reference-based clustering in many microbial analyses raises the question: should fungal ecologists re-consider their default use of *de novo* clustering? In particular, it seems that reference-based clustering may represent an increasingly useful approach to fungal community analyses as databases such as UNITE (Kõljalg *et al.*, 2013) grow in size and a greater diversity of fungal habitats are molecularly characterized. Recent studies have suggested that reference-based clustering can increase OTU stability and taxonomic accuracy relative to *de novo* clustering (He *et al.*, 2015; Halwachs *et al.*, 2017), although how this clustering approach influences fungal community analyses across diverse habitats is currently unclear. To assess this gap in knowledge, we compared the relative performance of *de novo*, closed reference, and open reference clustering approaches on a mock community, as well as samples from four ecologically distinct habitats. These habitats varied in the degree to which fungal composition was captured by the UNITE database, providing an opportunity to investigate the importance of *a priori* habitat characterization on clustering approach performance.

Using dead wood, live wood, live leaf and forest soil samples, we quantified fungal species assignments, OTU richness and community composition from ITS1 amplicon libraries sequenced on the Illumina MiSeq platform. We compared two *de novo* clustering algorithms (CD-HIT and USEARCH; Li & Godzik, 2006; Edgar, 2010), two closed reference clustering algorithms (BLAST and NINJA-OPS; Altschul *et al.*, 1990; Al-Ghalith *et al.*, 2016), as well as two open reference clustering scenarios (NINJA/USEARCH; BLAST/CD-HIT) applying a 97% sequence similarity cutoff for OTU clustering as well as taxonomy assignments (Table S1). For the open reference clustering, sequences were first clustered by a closed reference algorithm (i.e. NINJA or BLAST); the remaining sequences that failed to cluster were then clustered by a *de novo* clustering approach (i.e. USEARCH or CD-HIT), and the OTU tables were combined (*sensu* Rideout *et al.*, 2014). The UNITE database (v.7.0) was used for reference-based clustering as well as for designating

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust

taxonomic assignments of *de novo* OTUs (Kõljalg *et al.*, 2013) via the BLASTN algorithm (Altschul *et al.*, 1990). See Methods S1 and S2 for further details on sample harvest, library preparation, sequence processing, and data accessibility.

## Results and Discussion

Taxonomic designations are an important component of fungal community analyses, as this information ultimately informs *how* communities vary across habitat types. To understand the influence of clustering approach on OTU taxonomic classifications, we analysed a mock community comprised of 25 fungal species, and encompassing a range of taxonomic and phylogenetic diversity (Nguyen *et al.*, 2015). Across clustering algorithms, mock OTU richness ranged between 23 and 29 OTUs (Fig. 1). USEARCH was the only algorithm to cluster the expected number of OTUs, although the closed reference clustering approach was the most precise in estimating mock richness (i.e. it had a narrower range of richness (23–26 OTUs) relative to *de novo* and open reference approaches, both ranging from 23 to 29 OTUs). Considering the taxonomy of the mock community, OTUs with expected species-level assignments ranged between 20 and 22 across clustering scenarios (Fig. 1), with the highest mock species recovery in USEARCH and open reference scenario BLAST/CD-HIT. At a slightly coarser taxonomic resolution, the number of mock OTUs with expected genus-level classifications increased slightly (21 and 23 OTUs across clustering scenarios), and was highest in NINJA and the open reference scenario NINJA/USEARCH (Fig. S2). Whilst the *de novo* and reference-based approaches performed relatively similarly across these metrics, the extent of OTU inflation in the mock community appeared to be larger in select *de novo* and open reference clustering scenarios. Collectively, these results suggest that there is general parity in OTU richness and taxonomic recovery between *de novo* and reference-based clustering approaches and that the choice of algorithm had a larger influence on mock community estimates than the clustering approach.

Across the live wood, dead wood, live leaf and soil samples, we compared OTU richness and OTU relative richness (i.e. the log-ratio of average OTU richness in one habitat relative to another) among the *de novo*, closed reference and open reference clustering approaches, following the removal of rare OTUs (i.e. OTUs with fewer than 10 sequences within a sample; see Lindahl *et al.*, 2013 and Oliver *et al.*, 2015). Total fungal OTU richness within each of the four habitats was significantly lower using the closed reference approaches (Table 1), where an average of 36% fewer OTUs were generated relative to *de novo* and open reference clustering methods. This substantial richness discrepancy is likely to be the result of two competing influences: sequences failing to cluster in the closed

| | | *De novo* | | Closed | | Open | |
|---|---|---|---|---|---|---|---|
| Total OTUs | (25) | 29 | 25 | 23 | 26 | 23 | 29 |
| Recovered mock species | (25) | 20 | 22 | 21 | 20 | 22 | 20 |
| Recovered mock species (genus) | (25) | 21 | 22 | 21 | 23 | 22 | 23 |

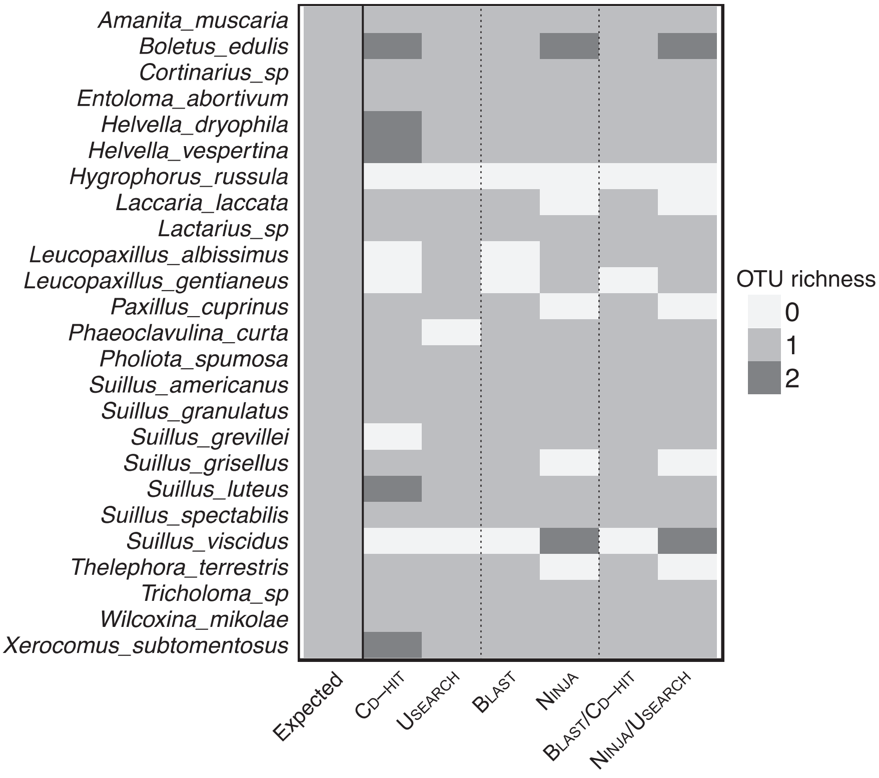

**Fig. 1** Mock community operational taxonomic unit (OTU) richness and taxonomic classifications calculated by six different *de novo*, closed reference, and open reference clustering approaches. Richness was compared to expected values based on the number of fungal species used to generate the mock community sample. Species and genus-level designations were quantified according to the number of mock species recovered at respective taxonomic levels. Multiple OTUs with same species name were counted only once; mismatched taxonomic IDs were not included in the visualization.

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust

**Table 1** Mean number of sequences clustered into operational taxonomic units (OTUs) (a) and average OTU richness (b) per environmental substrate and clustering algorithm

| | Approach | Algorithm | Live wood | Dead wood | Live leaf | Soil |
|---|---|---|---|---|---|---|
| (a) Sequences clustered (unrarefied samples) | *De novo* | Cd-hit | 5707 ± 4684 | 34 894 ± 19 680 | 72 393 ± 17 948 | 44 699 ± 22 539 |
| | | Usearch | 5737 ± 4691 | 34 956 ± 19 724 | 72 509 ± 17 934 | 45 132 ± 22 686 |
| | Closed reference | Blast | 1764 ± 2051 | 23 297 ± 17 247 | 64 580 ± 12 955 | 30 836 ± 15 412 |
| | | Ninja | 2651 ± 3725 | 23 906 ± 14 233 | 67 656 ± 15 582 | 32 187 ± 17 497 |
| | Open reference | Blast/Cd-hit | 5711 ± 4683 | 34 722 ± 19 869 | 72 403 ± 17 936 | 44 744 ± 22 544 |
| | | Ninja/Usearch | 5527 ± 4735 | 34 731 ± 20 170 | 70 930 ± 16 457 | 42 297 ± 22 116 |
| (b) OTU richness (rarefied samples) | *De novo* | Cd-hit | 53 ± 24.0$^{a}$ | 37 ± 9.1$^{a}$ | 61 ± 9.3$^{a}$ | 94 ± 13.6$^{a}$ |
| | | Usearch | 51 ± 23.7$^{a}$ | 31 ± 9.0$^{a}$ | 44 ± 8.3$^{bc}$ | 80 ± 13.6$^{a}$ |
| | Closed reference | Blast | 24 ± 10.6$^{b}$ | 20 ± 5.6$^{b}$ | 35 ± 5.1$^{c}$ | 61 ± 7.8$^{b}$ |
| | | Ninja | 25 ± 12.4$^{b}$ | 18 ± 4.7$^{b}$ | 38 ± 6.1 $^{cd}$ | 63 ± 8.1$^{b}$ |
| | Open reference | Blast/Cd-hit | 54 ± 23.7$^{a}$ | 36 ± 10.8$^{a}$ | 52 ± 7.7$^{ab}$ | 88 ± 13.1$^{a}$ |
| | | Ninja/Usearch | 53 ± 23.7$^{a}$ | 32 ± 9.8$^{a}$ | 50 ± 8.4$^{bc}$ | 82 ± 13.3$^{a}$ |

Mean ± SD. For each substrate, OTU richness differences between algorithms were determined by one-way ANOVAs, followed by Tukey's HSD. Lowercase letters indicate statistical differences at $\alpha < 0.05$.

reference approaches due to an incomplete reference database as well as amplification and sequencing errors contributing to the clustering of additional OTUs in *de novo* datasets (Bik *et al.*, 2012; but see Aas *et al.*, 2017). Datasets incorporating *de novo* clustering may be more susceptible to these errors relative to closed reference datasets, as closed reference sequences must match reference sequences to be clustered into OTUs. Despite differences in total richness, relative richness estimates amongst habitats were consistent across the clustering approaches, as indicated by log richness ratios with overlapping 95% confidence intervals between soil and the three other fungal habitats, independent of specific clustering algorithm (Fig. S3). This latter result indicates that fungal richness discrepancies amongst clustering approaches were comparable across habitat types, despite large differences in fungal representation in the Unite database (e.g. well-characterized live leaves vs poorly-characterized live wood). Thus, it appears that clustering approach had minimal impact on the conclusions of richness that could be drawn when compared in a relative manner (i.e. between treatments). As such, we suggest that relative richness metrics in HTS datasets may actually be more ecologically meaningful than comparisons based on absolute richness (Hughes *et al.*, 2001; Haegeman *et al.*, 2013).

In addition to fungal OTU richness, we investigated whether clustering approach influenced β-diversity estimates of fungal community composition across the four habitats. We calculated Bray–Curtis distances among all samples (Bray & Curtis, 1957), following Hellinger transformations of the OTU abundances (Legendre & Gallagher, 2001) for each clustering algorithm. We found that each clustering scenario consistently differentiated fungal communities by habitat (i.e. live wood, dead wood, live leaves or soil), as indicated by four distinct groupings of fungal communities in Principal Coordinates Analysis (Fig. 2) as well as a significant habitat term in each PerMANOVA model (permutational multivariate ANOVA; $F_{3,39} = 8.9–11.1$; $R^2 = 0.41–0.42$; $P < 0.001$). Additionally, Bray–Curtis distance matrices were highly correlated among the *de novo*, closed reference and open reference approaches (Mantel test; $R = 0.97–0.99$), indicating that fungal community comparisons between samples were consistently similar (more precisely, dissimilar) and to the same extent across clustering approaches. Given the large divergence in community composition across these habitats, it is possible that the significant effects observed would not be present when similar analyses are made within habitats. To address this possibility, we repeated the same β-diversity estimates across the three clustering approaches within both the soil and dead wood habitats. Similar to the between-habitat results, we found relatively consistent sample distributions in ordination space for both habitat types (Figs S4, S5). Additionally, significant Mantel correlations between each pair of clustering algorithms for soil (Mantel $R = 0.75–0.92$; $P < 0.001$) and dead wood (Mantel $R = 0.92–0.98$; $P < 0.001$) indicate that the finer-scale compositional differences between samples within a habitat were preserved across clustering approaches. Taken together, the large consistencies in β-diversity across clustering approaches suggests that the current size of the Unite database is not a limitation for the analysis of fungal community composition across a diverse range of terrestrial habitat types.

While patterns of fungal community α- and β-diversity among varying habitats were remarkably similar regardless of clustering approach, one notable limitation for the closed reference clustering approach that we encountered was the low number of clustered sequences in poorly characterized habitats. Among the habitats sampled, the mean number of sequences clustered using closed reference approaches was 62% lower in live wood compared to 5% lower in live leaves, relative to *de novo* clustering (Table 1). When sampling effort was standardized across habitat types (see McMurdie & Holmes, 2014 for potential limitations), the low proportion of clustered sequences in live wood along with the low initial ITS sequence recovery for this habitat type substantially limited the number of sequences that could be included in the analysis. Specifically, we rarefied each sample to 1000 sequences to retain as many live wood samples as possible, removing any sample that did not meet this threshold across all clustering algorithms (Table S2). If the *de novo* clustering approach alone was considered, we could have increased our rarefaction threshold four-fold while maintaining the same sample count in live wood. Yet, of the sequences that

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust

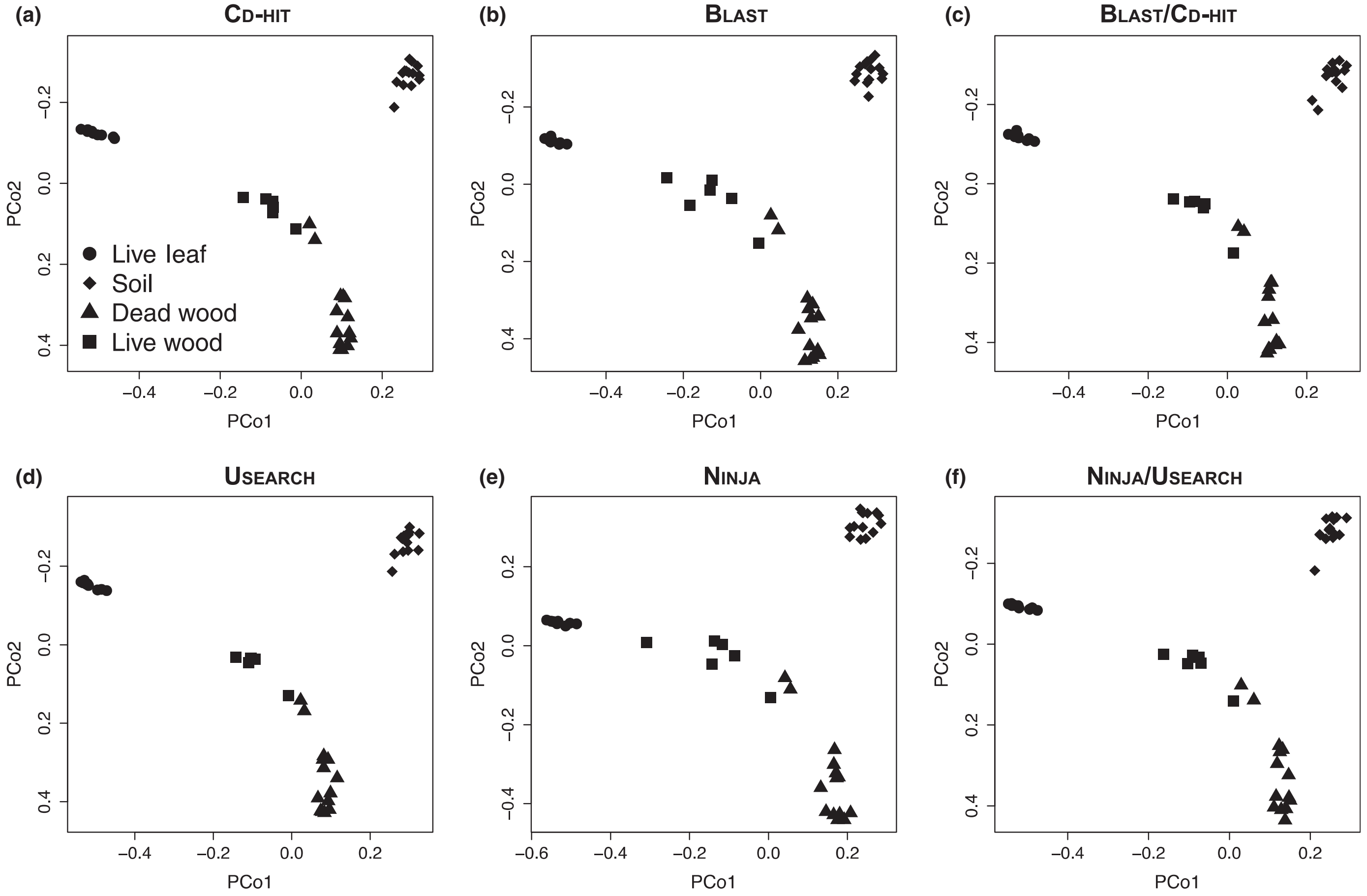

**Fig. 2** Principal coordinates (PCo) analysis of fungal β-diversity between live leaf, soil, dead wood, and live wood habitats, when operational taxonomic units (OTUs) were clustered by six different (a, d) *de novo*, (b, e) closed reference and (c, f) open reference clustering approaches. β-diversity was calculated by the Bray-Curtis dissimilarity metric. OTU abundances were Hellinger transformed before dissimilarity calculation.

failed to cluster under the Blast and Ninja algorithms, the majority (0.69 and 0.82, respectively) were not identified at the genus level following *de novo* clustering and subsequent taxonomic assignment (Fig. S6), validating the expectation that closed reference results recovered nearly as many genera as *de novo* and open reference clustering approaches. Together, these results suggest that for particularly poorly characterized systems, such as live wood, or potentially aquatic and marine systems (Peay *et al.*, 2016), either *de novo* or open reference clustering may still be preferable to the closed reference approach. Nonetheless, excluding sequences with no taxonomic assignment during closed reference clustering did not impact richness comparisons in even the least well-characterized habitat, although this may not necessarily be the case across other habitat comparisons.

## Conclusions

Despite the prevalence of *de novo* clustering in fungal community analysis, our results suggest that reference-based clustering approaches have a similar potential to accurately characterize the relative richness and β-diversity of fungal communities across a range of environments. Coupled with the fact that closed reference approaches consistently estimated the expected richness and taxonomy of the mock community in our analysis, we believe that non-AM fungal ecologists should reconsider their default *de novo* clustering approach, especially in well-characterized systems (e.g. Ovaskainen *et al.,* 2010). We recognize that reference-based clustering may not always be the best option, as clustering and taxonomic assignment depend on many factors, including the specific algorithm employed, the evolutionary properties of the sequences being analyzed (Kopylova *et al.*, 2016; Westcott & Schloss, 2015), and the fact that only a small fraction of the estimated 1.5–6 million fungal species on Earth are currently included in public sequence repositories (Halwachs *et al.*, 2017; Tedersoo *et al.,* 2017). However, the growing use of standardized barcoding regions (Schoch *et al.,* 2012), coupled with the increasing number of well-curated reference databases, suggests that the performance of reference-based clustering will continue to improve. Regardless of clustering approach, taxonomic identity appears to be sensitive to the specific algorithm used, which highlights the ongoing challenges associated with comparing taxonomies across studies as well as clustering approaches. Despite the consistent clustering and taxonomic assignment thresholds we used, individual algorithms have variations (Methods S2) that can impact results and interpretations. Recently, new methods have been developed that avoid clustering approaches altogether (e.g.

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust

DADA2; Callahan *et al.,* 2016), which circumvent the need for arbitrary thresholds to the increase accuracy and reproducibility of fungal diversity estimates. Collectively, we believe the results presented in this study, combined with the recent examples from the AM fungal ecology literature (e.g. Davison *et al.*, 2015), suggest that the fungal research community could benefit from more frequent use of reference-based clustering approaches moving forward.

## Acknowledgements

Authors would like to thank Jonathan Schilling and Scott Bates for help with sample collection. Financial support was provided by a USDA Hatch grant (MN-71-031) to P.G.K.

## Author contributions

P.G.K., L.C.C., Z.S., G.A.A-G. and D.K. conceived the idea and designed methodology; L.C.C., P.G.K., Z.S. and G.A.A-G. collected samples and generated data; L.C.C. led the analysis and writing of the manuscript with substantial contributions from P.G.K. and Z.S. All authors contributed critically to the manuscript and gave final approval for publication.

**Lauren C. Cline[1], Zewei Song[2], Gabriel A. Al-Ghalith[3], Dan Knights[3,4] and Peter G. Kennedy[1,5]***

[1]Department of Plant and Microbial Biology, University of Minnesota, St Paul, MN 55108, USA;
[2]Department of Plant Pathology, University of Minnesota, St Paul, MN 55108, USA;
[3]Biomedical Informatics and Computational Biology, University of Minnesota, Minneapolis, MN 55455, USA;
[4]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN 55455, USA;
[5]Department of Ecology, Evolution, and Behavior, University of Minnesota, St Paul, MN 55108, USA
(*Author for correspondence: tel +1 612 624 8519; email kennedyp@umn.edu)

## References

**Aas AB, Davey ML, Kauserud H. 2017.** ITS all right mama: investigating the formation of chimeric sequences in the ITS2 region by DNA metabarcoding analyses of fungal mock communities of different complexities. *Molecular Ecology Resources* 17: 730–741.

**Al-Ghalith GA, Montassier E, Ward HN, Knights D. 2016.** NINJA-OPS: fast accurate marker gene alignment using concatenated ribosomes. *PLoS Computational Biology* 12: e1004658.

**Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990.** Basic local alignment search tool. *Journal of Molecular Biology* 215: 403–410.

**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: 234–244.

**Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E. 2005.** Defining operational taxonomic units using DNA barcode data. *Philosophical transactions of the Royal Society of London Series B, Biological Sciences* 360: 1935–1943.

**Bray JR, Curtis JT. 1957.** An ordination of upland forest communities of southern Wisconsin. *Ecological Monographs* 27: 325–349.

**Callahan BJ, Mcmurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016.** DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13: 581–583.

**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.

**Davison J, Moora M, Öpik M, Adholeya A, Ainsaar L, Bâ A, Burla S, Diedhiou AG, Hiiesalu I, Jairus T *et al.* 2015.** Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* 127: 970–973.

**Dumbrell AJ, Nelson M, Helgason T, Dytham C, Fitter AH. 2010.** Relative roles of niche and neutral processes in structuring a soil microbial community. *ISME Journal* 4: 337–345.

**Edgar RC. 2010.** Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26: 2460–2461.

**García de León DG, Moora M, Öpik M, Neuenkamp L, Gerz M. 2016.** Symbiont dynamics during ecosystem succession : co-occurring plant and arbuscular mycorrhizal fungal communities. *FEMS Microbiology Ecology* **92**: fiw097.

**Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. 2013.** Robust estimation of microbial diversity in theory and in practice. *ISME Journal* 7: 1092–1101.

**Halwachs B, Madhusudhan N, Krause R, Nilsson RH, Moissl-Eichinger C, Högenauer C, Thallinger GG, Berry D. 2017.** Critical issues in mycobiota analysis. *Frontiers in Microbiology* 8: 180.

**He Y, Caporaso JG, Jiang X-T, Sheng H-F, Huse SM, Rideout JR, Edgar RC, Kopylova E, Walters WA, Knight R *et al.* 2015.** Stability of operational taxonomic units: an important but neglected property for analyzing microbial diversity. *Microbiome* 3: 20.

**Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJM. 2001.** Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* 67: 4399–4406.

**Jarvis SG, Woodward S, Taylor AFS. 2015.** Strong altitudinal partitioning in the distributions of ectomycorrhizal fungi along a short (300 m) elevation gradient. *New Phytologist* 206: 1145–1155.

**Kelly BJ, Imai I, Bittinger K, Laughlin A, Fuchs BD, Bushman FD, Collman RG. 2016.** Composition and dynamics of the respiratory tract microbiome in intubated patients. *Microbiome* 4: 7.

**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.

**Kopylova E, Navas-Molina JA, Mercier C, Xu ZZ, Mahé F, He Y, Zhou H-W, Rognes T, Caporaso JG, Knight R. 2016.** Open-source sequence clustering methods improve the state of the art. *mSystems* 1: e00003–e00015.

**Legendre P, Gallagher E. 2001.** Ecologically meaningful transformations for ordination of species data. *Oecologia* 129: 271–280.

**Li W, Godzik A. 2006.** Cd-hit : a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22: 1658–1659.

**Lindahl BD, Nilsson RH, Tedersoo L, Abarenkov K, Carlsen T, Kjøller 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.

**Magurran AE. 2005.** Species abundance distributions: pattern or process? *Functional Ecology* 19: 177–181.

**McMurdie PJ, Holmes S. 2014.** Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Computational Biology* 10: e1003531.

**Nguyen N, Smith D, Peay K, Kennedy P. 2015.** Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* 4: 1389–1393.

**Oliver AK, Brown SP, Callaham MA, Jumpponen A. 2015.** Polymerase matters: non-proofreading enzymes inflate fungal community richness estimates by up to 15%. *Fungal Ecology* 15: 86–89.

**Öpik M, Davison J, Moora M, Zobel M. 2014.** DNA-based detection and identification of Glomeromycota : the virtual taxonomy of environmental sequences. *Botany-Botanique* 92: 135–147.

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust

**Öpik M, Vanatoa A, Vanatoa E, Moora M, Davison J, Kalwij JM, Reier Ü, Zobel M. 2010.** The online database Maarj*AM* reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* **188**: 223–241.

**Ovaskainen O, Nokso-koivisto J, Hottola J, Rajala T, Pennanen T, Ali-Kovero H, Miettinen O, Oinonen P, Auvinen P, Paulin L *et al.* 2010.** Identifying wood-inhabiting fungi with 454 sequencing – What is the probability that BLAST gives the correct species? *Fungal Ecology* **3**: 274–283.

**Peay KG, Kennedy PG, Talbot JM. 2016.** Dimensions of biodiversity in the earth mycobiome. *Nature Reviews Microbiology* **14**: 434–447.

**Redecker D, Schüßler A, Stockinger H, Stürmer SL, Morton JB, Walker C. 2013.** An evidence-based consensus for the classification of arbuscular mycorrhizal fungi (Glomeromycota). *Mycorrhiza* **23**: 515–531.

**Rideout JR, He Y, Navas-Molina JA, Walters WA, Ursell LK, Gibbons SM, Chase J, McDonald D, Gonzalez A, Robbins-Pianka A *et al.* 2014.** Subsampled open-reference clustering creates consistent, comprehensive OTU definitions and scales to billions of sequences. *PeerJ* **2**: e545.

**Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, Levesque CA, Chen W. 2012.** Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences, USA* **109**: 6241–6246.

**Stajich JE, Berbee ML, Blackwell M, Hibbett DS, James TY, Spatafora JW, Taylor JW. 2009.** The fungi. *Current Biology* **19**: R840–R845.

**Tedersoo L, Bahram M, Puusepp R, Nilsson RH, James TY. 2017.** Novel soil-inhabiting clades fill gaps in the fungal tree of life. *Microbiome* **5**: 42.

**Truong C, Mujic AB, Healy R, Kuhar F, Furci G, Torres D, Niskanen T, Sandoval-Leiva PA, Fernández N, Escobar JM *et al.*** 2017. How to know the fungi: combining field inventories and DNA-barcoding to document fungal diversity. *New Phytologist.* **214**: 913–919.

**Westcott SL, Schloss PD. 2015.** De novo clustering methods outperform reference-based methods for assigning 16S rRNA gene sequences to operational taxonomic units. *PeerJ* **3**: e1487.

## Supporting Information

Additional Supporting Information may be found online in the Supporting Information tab for this article:

**Fig. S1** Number of times *de novo,* closed reference and open reference clustering approaches were used in a Web of Science literature survey of 141 fungal molecular studies.

**Fig. S2** Genus-level classifications of mock community across clustering scenarios.

**Fig. S3** Relative OTU richness between habitats.

**Fig. S4** Principal coordinates analysis of fungal β-diversity between soil samples.

**Fig. S5** Principal coordinates analysis of fungal β-diversity between dead wood samples.

**Fig. S6** Genus-level taxonomic identifications of *de novo* OTUs that failed to cluster under closed reference algorithms.

**Table S1** Requirements for OTU clustering and taxonomic assignment under each clustering scenario

**Table S2** Number of samples by substrate in which more than 1000 sequences clustered to OTUs across clustering algorithms

**Methods S1** The full method description for specimen harvest, DNA extraction and amplification, and statistical analysis.

**Methods S2** Commands for sequence processing.

Please note: Wiley Blackwell are not responsible for the content or functionality of any Supporting Information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

**Key words:** closed reference, community analysis, *de novo*, diversity, fungi, open reference, operational taxonomic unit (OTU) clustering.

© 2017 The Authors
*New Phytologist* © 2017 New Phytologist Trust