Letters

Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns?

Progress in microbial community ecology is challenged by the fact that individuals often cannot be morphologically identified and counted, and the great majority of taxa are not phenotypically characterized because they lack recognizable traits and are not in culture. As a result, microbes are often distinguished from one another using molecular sequence data. For example, ≥ 97% sequence similarity in the 16S rRNA gene is commonly used to separate species in bacteria (Stackebrandt & Goebel, 1994; Schloss & Handelsman, 2006) although some argue that 99% is more appropriate (Acinas *et al.*, 2004). Such differences in universal thresholds may influence our ability to understand the ecology of sequence clusters, or operational taxonomic units (OTUs). If sequence similarities are too narrow (e.g. 99%), comparisons among communities are difficult, and if they are too broad (e.g. 90%), OTUs with different ecological roles and distribution patterns may be erroneously lumped together (discussed in Knights *et al.*, 2011a). Universal thresholds also do not consider differences in speciation and substitution rates among lineages and may therefore not capture equivalent units of diversity. Indeed, Koeppel & Wu (2013) showed that many bacterial OTUs are paraphyletic and span multiple ecological habitats, and Youngblut *et al.* (2013) demonstrated that ecological patterns may be optimally detected by using different OTU delineations among lineages.

Arbuscular mycorrhizal fungi (AMF) present similar challenges. AMF are asexual, obligate symbionts that exchange nutrients and other services for plant carbon (Smith & Read, 2008). Long-term asexual evolution has led to high genetic and ecological diversity within known AMF species (Munkvold *et al.*, 2004; Koch *et al.*, 2006; Rosendahl, 2008; Stockinger *et al.*, 2009) that traditionally have been identified based on spore features. Molecular techniques show that, like bacteria, many environmental sequences do not cluster with cultured species (Rosendahl, 2008), but there is currently no consensus on how to best organize sequences into biologically and ecologically meaningful taxonomic units (Öpik *et al.*, 2010). While some researchers use universal – albeit varying – OTU thresholds (Dumbrell *et al.*, 2011) or BLAST matches against published sequences (Lekberg *et al.*, 2007), others attempt to control for evolutionary differences by identifying OTUs as groups forming monophyletic clades (Rosendahl & Stukenbrock, 2004; Sykorova *et al.*, 2007). How these methods affect our understanding of AMF community patterns along environmental and spatial gradients is more or less unknown (but see Powell *et al.*, 2011). This is a pressing question given that high-throughput pipelines for post-sequencing analyses of next generation sequencing (NGS) data often utilize universal thresholds for OTU delineations (Caporaso *et al.*, 2010), as groupings based on evolutionary relationships are computationally expensive for large data sets.

Here we compare how similar two OTU delineation methods, one that considers evolutionary origins and another based on 97% sequence similarity, distribute AMF OTUs along environmental and spatial gradients. For the evolutionary approach, we manually combined sequence groups into OTUs that formed monophyletic clades (hereafter referred to as the monophyletic clade approach, or MCA). This resulted in a sequence variability within OTUs of 3–11.5%. For the 97% universal approach, we used a standard bioinformatics workflow developed using the open source Quantitative Insights into Microbial Ecology software package (QIIME, v.1.6.0, Caporaso *et al.*, 2010) with a 97% universal OTU threshold (i.e. a sequence variability within OTU of 3%) and the UCLUST clustering algorithm (hereafter referred to as the 97% approach). Both approaches are outlined in more detail in the Supporting Information Methods S1. Using the same data, quality filtering (QC) and removal of low abundance OTUs, we applied both methods to three published NGS (454-Titanium) AMF datasets of LSU rRNA gene sequences (all targeting the D2 region). The three datasets range in spatial scale from one Danish grassland (meters; Lekberg *et al.*, 2011; hereafter referred to as 'Site'; primers glo454-NDL22; 8912 sequences, post QC), 22 local plant communities in south-western Montana (tens of kilometers; Lekberg *et al.*, 2013; hereafter termed 'Local'; primers FLR3-FLR4; 302 527 sequences) and 19 sample sites located across regions in the western United States (hundreds of kilometers; Bunn *et al.*, 2014); termed 'Regional'; primers FLR3-FLR4; 28 711 sequences). The three datasets contained 37 independent variables (Table S1) that were used to assess if and how the two OTU delineation methods distributed AMF communities differently along spatial and environmental gradients. The correspondence between the two OTU delineation methods was assessed using four complimentary tests. Mantel tests were used to determine the correlation and significance of numerical variables for AMF community patterns, and Analysis of Similarities (ANOSIM) was used for categorical variables. BEST analysis constructed the optimal n-parameter model of all numerical variables at once, and Procrustes analyses were used to compare distributions of AMF communities in PCoA space (Caporaso *et al.*, 2012). Distance-based Redundancy Analysis (db-RDA) with forward selection was used to select a set of metadata variables that significantly explained non-overlapping portions of the AMF community variance (see 'ordistep' in the vegan package for R). All analyses were run in

© 2014 The Authors
New Phytologist © 2014 New Phytologist Trust

QIIME (Caporaso *et al.*, 2010) and are outlined in more detail in the Supporting Information.

We also applied a supervised learning approach (Breiman, 2001; Knights *et al.*, 2011b) to tests if and how the choice of universal OTU threshold (90–99%) influenced the ability to assign an AMF community to the correct environmental category. This approach requires a strong categorical predictor and we therefore used the Local dataset in which AMF communities differed significantly among plant community types (Lekberg *et al.*, 2013). Ninety percent of the sequence data was chosen at random to train the learning algorithm to recognize AMF assemblages associated with four plant community types (knapweed (*Centaurea stoebe*), cheatgrass (*Bromus tectorum*), spurge (*Euphorbia esula*) and mixed remnant native) and the remaining 10% of the data was classified based on the inferred function generated by the learning algorithm (Knights *et al.*, 2011b). In this scenario, the error rate expected by chance was 75% given that there were four plant community types (i.e. three in four designations would be incorrect due to chance alone). To obtain robust estimates of generalization error, 10-fold cross-validation was employed (Knights *et al.*, 2011b).

Compared with the MCA, the 97% approach drastically increased OTU numbers from 33 to 76 in the Site dataset, from 46 to 1083 in the Local dataset, and from 30 to 278 in the Regional dataset (Figs S1–S3). This increase may be due, in part, to sequencing errors (Dickie, 2010; Tedersoo *et al.*, 2010) because the recommended denoising step was not conducted (Reeder & Knight, 2010). An increase in OTU numbers was expected with the 97% approach, however, because many OTUs were manually combined in our MCA method, especially within the *Rhizophagus irregularis* and *Glomus microaggregatum* groups that had a within-OTUs sequence variability of up to 11.5%. This is very similar to the 11.8% intraspecific variation in the LSU gene region of *Rhizophagus intraradices* (FL isolate) reported by Stockinger *et al.* (2009). Based on this, the 97% approach may resolve smaller genetic differences within species, as has been shown for all fungi using the Internal Transcribed Spacer (ITS) region (Blaalid *et al.*, 2013). Indeed, the most abundant MCA OTU in the Site dataset (Rhizophagus P) clustered with *R. irregularis* (FR750199) in Krüger *et al.* (2012), whereas the 97% approach divided this monophyletic clade into three OTUs. Coincidentally, perhaps, the online database Maarj*AM* that targets the 18S gene region also separates this species into three distinct virtual taxa (Öpik *et al.*, 2010). Another unknown abundant OTU in the Site dataset, Rhizophagus D, was split into five OTUs with the 97% approach.

The finer-scale division of OTUs in the 97% approach increased (pairwise *t*-test $P<0.001$) β-diversity (γ/α) by an average 148% compared with MCA in the Local dataset. However, the increase in richness with the 97% approach did not appear to separate local ecotypes because the amount of variation explained by the spatial distribution of samples was not higher with the 97% approach compared with MCA (Table S1). Also, in spite of the dramatic increase in OTU numbers with the 97% approach and a change in *absolute* OTU richness, differences in *relative* richness among vegetation types observed in Lekberg *et al.* (2013) remained largely the same. If highly variable groups were over-represented in particular vegetation types, this correspondence may not have been observed.

OTU numbers aside, the distribution of AMF communities in ordination space, and their responses to environmental and spatial variables were very similar between the two OTU delineation approaches. More specifically, Mantel-*r* values, which assess the relationship between individual spatial and environmental variables and AMF community compositions, were highly correlated between the MCA and the 97% approach (Fig. 1). That is, important variables in the MCA (indicated by high Mantel-*r* and low *P*-values) were also important in the 97% approach (Table S1). The analyses of categorical variables were also in agreement, and the BEST analyses identified the same (Site and Local) or similar (Regional) predictor variables (Table S2). Both methods clustered the AMF communities according to vegetation type in the Local dataset (Fig. 2), and significant ($P<0.01$) Procrustes analyses (Caporaso *et al.*, 2012) were observed for all three studies (Table S3, Fig. S4).

Is this correspondence to be expected given that the MCA often lumps OTUs generated by the 97% approach? Not necessarily, because one may have predicted that either (1) combining sequence types, and possibly ecotypes, in the MCA would have obscured environmental signals resulting in inflated *P*-values, or, on the contrary, (2) splitting sequence types within lineages, and possibly diluting ecologically relevant sequence types with the 97% approach would have reduced the power to identify significant environmental variables. We found neither and thus conclude that

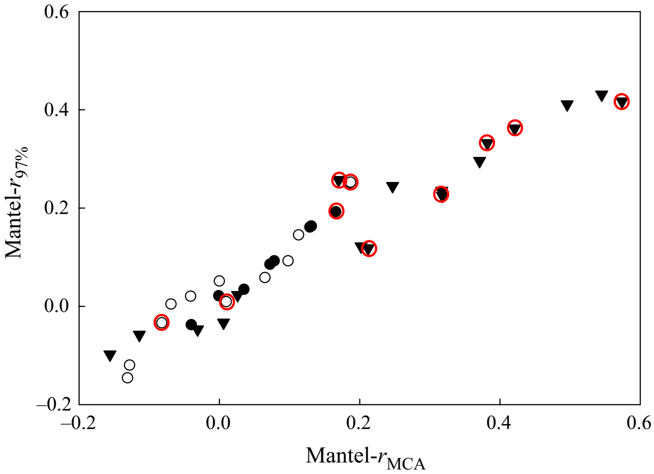

Fig. 1 Correlation of Mantel-*r* values obtained from statistical analyses with explanatory variables and arbuscular mycorrhizal fungal (AMF) communities generated from a monophyletic clade approach (MCA) or a universal (97%) threshold in all the three datasets (Site, closed circles; Local, open circles; Regional, triangles). The Mantel-*r* values were highly correlated ($R=0.96$, $P<0.001$) overall, as were separate correlations within datasets (Site: $R=0.99$, $P<0.001$; Local: $R=0.96$, $P<0.001$; Regional: $R=0.97$, $P<0.001$). The red circles indicate variables that explained a significant proportion of variation ($P<0.1$) in both operational taxonomic unit (OTU) delineation approaches using Distance-based Redundancy Analysis (db-RDA) and forward selection. The db-RDA eliminates potential issues with autocorrelations among explanatory variables and thus represents a very conservative, albeit still highly significant ($R=0.93$, $P<0.001$), measure of the relationship between the two OTU delineation approaches.

© 2014 The Authors
New Phytologist © 2014 New Phytologist Trust

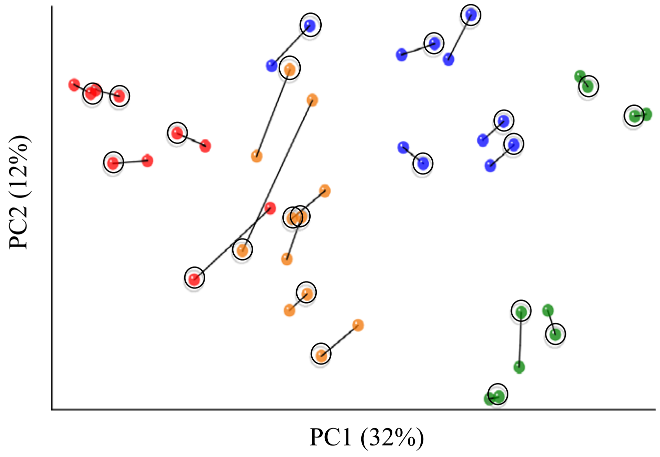

Fig. 2 Procrustes analysis of arbuscular mycorrhizal fungal (AMF) communities among cheatgrass (red), native (orange), knapweed (blue) and spurge (green) communities for the Local dataset. Duplicate points (connected by black lines) represent data processed through the monophyletic clade approach (MCA) and 97% (circled) operational taxonomic unit (OTU) picking pipelines. The lines between points highlight the Euclidean distance between points that represent the two different methods. Procrustes plots for the Site and Regional datasets are in Supporting Information Fig. S4.

AMF community patterns are robust across different OTU delineation methods (at least within the datasets and DNA target region tested here). We found support for this in the supervised learning, because even though OTU numbers increased exponentially with increasing similarity cut-offs, and ranged from 146 (90% threshold) to 2077 (99% threshold), the error rate was similar at 90 and 99% thresholds (Fig. S5). That is, the computer was equally good at assigning an AMF community to the correct aboveground plant community using AMF communities that had been delineated using 90% and 99% thresholds, which suggests that differences in AMF community patterns among plant community types in this dataset were deeply phylogenetically rooted. This was further supported by almost identical clustering of AMF communities among plant communities using the whole dataset and universal thresholds of 90, 97 and 98% (ANOSIM R = 0.60–0.62, P < 0.001).

Overall, our results agree remarkably well with findings by Powell *et al.* (2011), which showed that an equivalent amount of variation was accounted for using a universal OTU delineation approach (97%) and one based on evolutionary processes (general mixed Yule-coalescent or GMYC model). Their comparison used Sanger sequencing datasets that targeted the more conservative 18S gene region (Stockinger *et al.*, 2010), which suggests that our findings may extend beyond the three datasets included here. Powell *et al.* (2011) argued for the use of the GMYC model because the same amount of variation was explained by fewer OTUs. The known difference in lineage age and divergence among known AMF species also favors an evolutionary approach because OTUs can be more easily aligned with known species (Krüger *et al.*, 2012) or virtual taxa (Öpik *et al.*, 2010). Niche-based methods using variable thresholds are being developed for 16S gene sequence data (Koeppel & Wu, 2013), but they are currently only computationally feasible for smaller-sized datasets (< 10 000 sequences). Until these approaches become readily available, our results indicate that even though comparisons of richness are difficult, AMF community patterns generated using universal thresholds do not differ from those grounded in evolutionary theory. While a more unified approach in AMF community ecology is desirable, cross-validations against expert-curated databases (Öpik *et al.*, 2010) will allow for comparisons among studies regardless of OTU delineation approach, and will forward our understanding of AMF ecology and biogeography.

Acknowledgements

The authors thank Roger Koide, Philip Ramsey, Beau Larkin, Dan Mummey and four anonymous reviewers for their constructive comments on the manuscript, Rebecca Bunn for providing comments and allowing the authors to use the Regional dataset, Yvette Ortega for inputs on statistics and MPG Operations for funding. S.M.G. is supported, in part, by an EPA STAR Graduate Fellowship.

Ylva Lekberg[1,2]*, Sean M. Gibbons[1,3,4] and Søren Rosendahl[5]

[1]MPG Ranch, Missoula, MT 59801, USA;
[2]Department for Ecosystem and Conservation Sciences, University of Montana, Missoula, MT 59812, USA;
[3]Graduate Program in Biophysical Sciences, University of Chicago, Chicago, IL 60637, USA;
[4]Argonne National Laboratory, Institute for Genomics and Systems Biology, 9700 S. Cass Ave., Lemont, IL 60439, USA;
[5]Department of Biology, University of Copenhagen, Universitetsparken 15, DK-2100, Copenhagen, Denmark
(*Author for correspondence: tel +1 406 396 6159; email ylekberg@mpgranch.com)

References

Acinas SG, Klepac-Ceraj V, Hunt DE, Pharino C, Ceraj I, Distel DL, Polz MF. 2004. Fine-scale phylogenetic architecture of a complex bacterial community. *Nature* **430**: 551–554.

Blaalid R, Kumar S, Nilsson RH, Abarenkov K, Kirk PM, Kauserud H. 2013. ITS1 versus ITS2 as DNA metabarcodes for fungi. *Molecular Ecology Resources* **13**: 218–224.

Breiman L. 2001. Random forests. *Machine Learning* **45**: 5–32.

Bunn RA, Lekberg Y, Gallagher C, Rosendahl S, Ramsey PW. 2014. Grassland invaders and their mycorrhizal symbionts: a study across climate and invasion gradients. *Ecology and Evolution.* doi: 10.1002/ece3.917.

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 CH, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M *et al.* 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *The ISME Journal* **6**: 1621–1624.

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

Dumbrell AJ, Ashton PD, Aziz N, Feng G, Nelson M, Dytham C, Fitter AH, Helgason T. 2011. Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytologist* **190**: 794–804.

© 2014 The Authors
New Phytologist © 2014 New Phytologist Trust

Knights D, Costello EK, Knight R. 2011b. Supervised classification of human microbiota. *FEMS Microbiology Reviews* **35**: 343–359.

Knights D, Parfrey Wegener L, Zaneveld J, Lozupone C, Knight R. 2011a. Human-associated microbial signatures: examining their predictive value. *Cell Host & Microbe* **10**: 292–296.

Koch AM, Croll D, Sanders IR. 2006. Genetic variability in populations of arbuscular mycorrhizal fungi causes variation in plant growth. *Ecology Letters* **9**: 103–110.

Koeppel AF, Wu M. 2013. Surprisingly extensive mixed phylogenetic and ecological signals among bacterial Operational Taxonomic Units. *Nucleic Acids Research* **41**: 5175–5188.

Krüger M, Krüger C, Walker C, Stockinger H, Schüßler A. 2012. Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizaal fungi from phylum to species level. *New Phytologist* **193**: 970–984.

Lekberg Y, Gibbons SM, Rosendahl S, Ramsey PW. 2013. Severe plant invasions can increase mycorrhizal fungal abundance and diversity. *The ISME Journal* **7**: 1424–1433.

Lekberg Y, Koide RT, Rohr JR, Aldrich-Wolfe L, Morton JB. 2007. Role of niche restrictions and dispersal in the composition of arbuscular mycorrhizal fungal communities. *Journal of Ecology* **95**: 95–105.

Lekberg Y, Schnoor T, Kjøller R, Gibbons SM, Hansen LH, Al-Soud WA, Sørensen SJ, Rosendahl S. 2011. 454-sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. *Journal of Ecology* **100**: 151–160.

Munkvold L, Kjøller R, Vestberg M, Rosendahl S, Jakobsen I. 2004. High functional diversity within species of arbuscular mycorrhizal fungi. *New Phytologist* **164**: 357–364.

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.

Powell JR, Monaghan MT, Öpik M, Rillig MC. 2011. Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology* **20**: 655–666.

Reeder J, Knight R. 2010. Rapidly denoising pyrosequencing amplicon reads by exploiting rank-abundance distributions. *Nature Methods* **7**: 668–669.

Rosendahl S. 2008. Communities, populations and individuals of arbuscular mycorrhizal fungi. *New Phytologist* **178**: 253–266.

Rosendahl S, Stukenbrock EH. 2004. Community structure of arbuscular mycorrhizal fungi in undisturbed vegetation revealed by analyses of LSU rDNA sequences. *Molecular Ecology* **13**: 3179–3186.

Schloss PD, Handelsman J. 2006. Toward a census of bacteria in soil. *PLoS Computational Biology* **2**: 786–793.

Smith SE, Read DJ. 2008. *Mycorrhizal symbiosis*. Cambridge, UK: Academic Press.

Stackebrandt E, Goebel BM. 1994. Taxonomic note: a place for DNA–DNA reassociation and 16S rDNA sequence analysis in the present species definition in bacteriology. *International Journal of Systematic Bacteriology* **44**: 846–849.

Stockinger H, Krüger M, Schüßler A. 2010. DNA barcoding of arbuscular mycorrhizal fungi. *New Phytologist* **187**: 461–474.

Stockinger H, Walker C, Schüßler A. 2009. '*Glomus intraradices* DAOM197198', a model fungus in arbuscular mycorrhiza research, is not *Glomus intraradices*. *New Phytologist* **183**: 1176–1187.

Sykorova Z, Ineichen K, Wiemken A, Redecker D. 2007. The cultivation bias: different communities of arbuscular mycorrhizal fungi detected in roots from the field, from bait plants transplanted to the field, and from a greenhouse trap experiment. *Mycorrhiza* **18**: 1–14.

Tedersoo L, Nilsson RH, Abarenkov K, Jairus T, Sadam A, Saar I, Bahram M, Bechem E, Chuyong G, Kõljalg U. 2010. 454 Pyrosequencing and Sanger sequencing of tropical mycorrhizal fungi provide similar results but reveal substantial methodological biases. *New Phytologist* **188**: 291–301.

Youngblut N, Shae A, Read JS, McMahon KD, Whitaker RJ. 2013. Lineage-specific responses of microbial communities to environmental change. *Applied Environmental Microbiology* **79**: 39–47.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 NeighborNet split network based on MCA or 97% universal threshold using the Site dataset.

Fig. S2 NeighborNet split network based on MCA or 97% universal threshold using the Local dataset.

Fig. S3 NeighborNet split network based on MCA or 97% universal threshold using the Regional dataset.

Fig. S4 Procrustes analysis of the Site and Regional datasets.

Fig. S5 Supervised learning results of OTU numbers and error rates for the Local dataset.

Table S1 Mantel-*r* and *P*-values using the 97% universal threshold and the monophyletic clade approach (MCA) on the three datasets

Table S2 ANOSIM and BEST analyses for the three datasets using the monophyletic clade approach (MCA) and the 97% universal threshold to delineate OTUs

Table S3 Correlations beta-diversity distance matrices and Procrustes analyses between the monophyletic clade approach (MCA) and the 97% approach

Methods S1 This outlines the sequence processing and analysis associated with the MCA and 97% OTU delineation methods.

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: 454-Titanium sequencing, arbuscular mycorrhizal fungi (AMF), monophyletic clade, operational taxonomic unit (OTU) delineation, QIIME, universal threshold.

© 2014 The Authors
New Phytologist © 2014 New Phytologist Trust