Sequence variation in the rRNA gene within isolates of arbuscular mycorrhizal fungi: Tests of phylogeny and clustering methodologies

Geoffrey L. House[a], Saliya Ekanayake[b], Yang Ruan[b,c], Ursel Schütte[a,d,e], Wittaya Kaonongbua[a,f], Geoffrey Fox[b], Yuzhen Ye[b], James D. Bever[a,g]

[a] Department of Biology, Indiana University, Bloomington, IN [b] School of Informatics and Computing, Indiana University, Bloomington, IN [c] Yelp, Inc. San Francisco, CA [d] School of Public and Environmental Affairs, Indiana University, Bloomington, IN [e] Department of Biology, University of Alaska, Fairbanks, AK [f] Department Of Microbiology, Faculty Of Science, King Mongkut's University of Technology Thonburi, Thailand [g] Department of Biology, University of Kansas, Lawrence, KS

Submitted to Proceedings of the National Academy of Sciences of the United States of America

Arbuscular mycorrhizal (AM) fungi form mutualisms with plant roots that increase plant growth and shape plant communities. A single AM fungal cell contains a large amount of genetic diversity but it is unclear how this diversity may vary across AM fungal lineages. To address this, we sequenced the nuclear ribosomal large subunit (LSU; 28S) gene from 21 species of phylogenetically diverse AM fungi. We then applied a novel multidimensional scaling (MDS) method and found that groups of similar sequences corresponded well with genus-level clades on the rRNA gene tree. Sequences from each species also generally formed monophyletic groups, with the exception of incomplete lineage sorting of sequence variants in the genera *Claroideoglomus* and *Entrophospora*. The level of sequence variation differed significantly between genera and also across the phylogeny. We used these patterns of sequence variation coupled with the MDS visualization to assess the accuracy of four different sequence clustering methods in delineating operational taxonomic units (OTUs) for species from different genera. The clustering methods AbundantOTU and CROP were consistently more accurate than mothur or UPARSE, although no clustering method gave OTUs that reliably approximated species-level groups. This lack of OTU-to-species correspondence resulted both from sequences of one species being split into multiple OTUs, and from sequences of multiple species being lumped into the same OTU. Using OTUs to identify putative AM fungal species from environmental samples will therefore result in biased richness estimates, and the direction of this bias will depend on the genera that are present in the sample.

Genetic polymorphism | Arbuscular mycorrhizal (AM) fungi | Incomplete lineage sorting | Sequence clustering | Multidimensional scaling

Introduction

Sequences of the phylogenetically-informative nuclear ribosomal RNA (rRNA) gene have revolutionized our understanding of microbial diversity. They have been instrumental both in the discovery of microbial groups, including the kingdom archaea (1), and also in clarifying their evolutionary relationships. When used in combination with high-throughput sequencing, rRNA gene sequences have provided critical insights into microbial communities and their dynamics in a wide range of environments, including marine microbes (2) and the microbiomes found within other organisms (3). rRNA gene sequences are a powerful, culture-independent way to better understand the genetic diversity of microbes; this is especially valuable in poorly described groups like arbuscular mycorrhizal (AM) fungi that form mutualisms with plant roots.

AM fungi (phylum Glomeromycota) are a widely distributed and ecologically important group of microbes that form mutualisms with the roots of most terrestrial plant species (4) and can shape plant communities in grassland ecosystems (5-7). Communities of soil-dwelling fungi are genetically and functionally diverse (8), and AM fungi have an exceptionally high amount of rRNA gene sequence variation compared to other fungal phyla (9). However, in contrast to most other microbes including bacteria and archaea, this sequence variation can occur within a single multi-nucleate cell of an AM fungal individual (10, 11).

The large amount of intra-organismal sequence variation in AM fungi presents challenges when attempting to use rRNA gene trees to better understand species relationships. Currently, morphology-based concepts of AM fungal species have generally been supported by rRNA gene trees (12-14), but these trees have been constructed using a limited amount of sequence variation. While the recent genome sequencing of isolates in the genus *Rhizophagus* has furthered our understanding of the range of genetic variation in AM fungi (15-17), it remains largely unknown how this genetic variation may itself vary between different evolutionary lineages. Justifiably, the most comprehensive phylogeny to-date (18) prioritized a taxonomically broad sampling at the cost of limiting the amount of sequence variation sampled within species. However, it is not currently known whether this phylogeny is robust to the inclusion of more of the inherent genetic variation that occurs within AM fungal species.

Determining how the amount of sequence variation may differ among species of AM fungi is important to more accurately

Significance

Arbuscular mycorrhizal (AM) fungi form important mutualisms with the roots of most plant species. Individual AM fungi are highly genetically diverse, but it is unclear whether the level of this diversity differs between evolutionary lineages. We find that sequence variation in AM fungi significantly varies between evolutionary lineages that correspond to genera, and that there is incomplete lineage sorting of this diversity among species in one genus. These consistent patterns of genetic diversity resulted in either an upward or downward bias in the correspondence between operational taxonomic units (OTUs) and species depending on the genus. Identifying putative AM fungal species from environmental samples using either OTUs or other sequence similarity-based methods should therefore take this taxonomically-structured genetic variability into account.

Reserved for Publication Footnotes

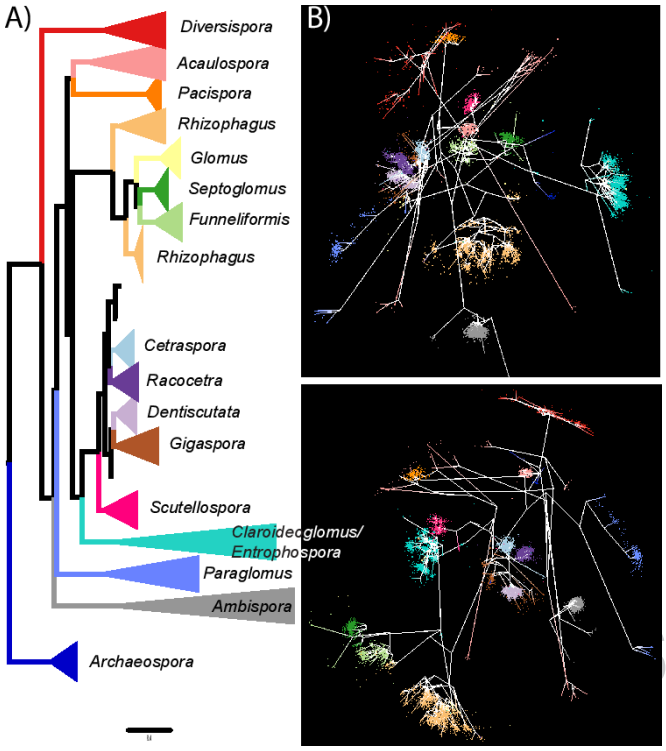

Fig. 1. Correspondence between the phylogenetic tree (left) and two different views of the MDS clustering (right) for sequences from 21 species of AM fungi colored by genus. Branches within genera on the phylogenetic tree are collapsed for clarity, but each sequence is represented as a point in the MDS visualization.

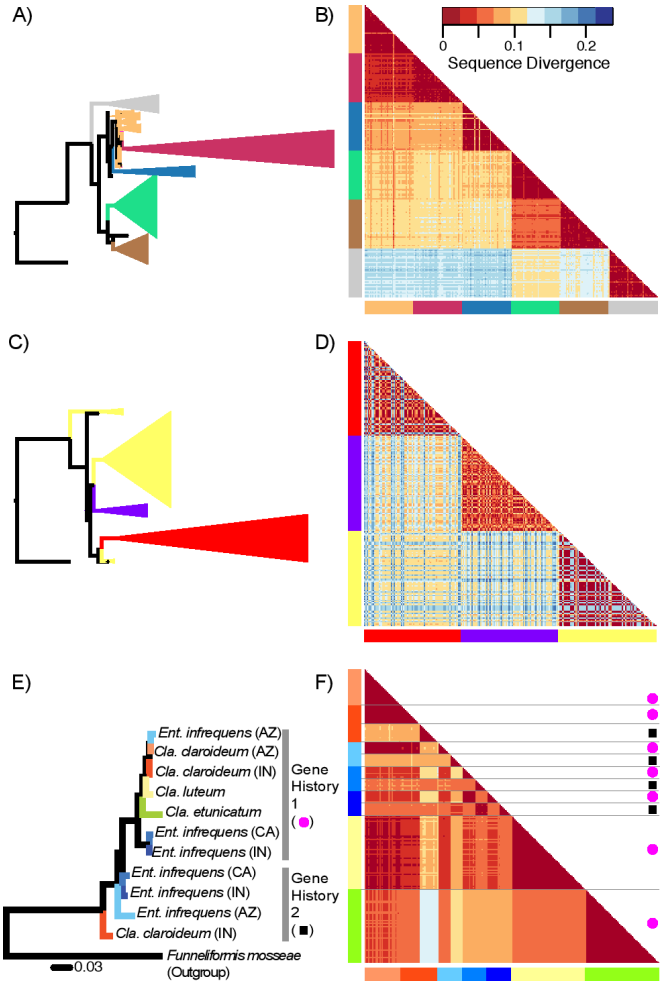

Fig. 2. Comparisons of sequence similarity between the gene tree and heatmaps of pairwise sequence similarity for species in the Gigasporaceae (top row), *Rhizophagus* (center row), and *Claroideoglomus/Entrophospora* (bottom row), with each species in the group being represented by the same color in both the gene tree and the heatmap. In the heatmap for *Claroideoglomus/Entrophospora*, sequences from gene history 1 are marked by pink circles and those from gene history 2 are marked by black squares.

determine species composition from environmental samples. Different species of AM fungi, as determined by the morphology of their resting spores, can be functionally distinct (19) and the composition of AM fungal communities can change during ecological succession (20). However, it is necessary to use DNA sequences from roots in order to identify the AM fungal species that have active mycorrhizal associations with plants. Analysis of these sequences typically relies on clustering similar sequences into operational taxonomic units (OTUs), and assumes that the sequence variation contained in a single species is approximated by each OTU (21-23).

For AM fungi the assumption that each OTU corresponds to a different species is largely untested, and may be incorrect given the large amount of intra-organismal sequence variation they contain. For studies of bacteria, including a 'mock community' of DNA from various strains with known OTU compositions in a sequencing run has allowed the evaluation of sequence clustering methods (24) and the assessment of the accuracy of OTU clustering for environmental samples (25). This 'mock community' approach has not been used for AM fungi. Rather, investigators have compared the utility of clustering programs for environmental samples by testing the correlation between specific OTUs and environmental variables (22, 26, 27). These studies, however, rely on the linked assumptions that environmental factors are correlated with functional traits of AM fungi that are detectable using rRNA gene sequences, and that those functional traits themselves are correlated with different AM fungal species, either of which may not be true. Moreover, the high level of intra-specific and intra-organismal sequence variation in AM fungi may violate an essential assumption of clustering sequences into OTUs - that all sequences from the same species form a monophyletic group on the gene tree. This assumption of sequence monophyly within species has not been tested in a comprehensive way across AM fungal groups, but recent work has called it into question, especially for the genus *Claroideoglomus* (28, 29).

Here we first evaluated the range of sequence variation within isolates of AM fungi distributed across 21 morphologically defined species by pyrosequencing the nuclear large subunit (LSU; 28S) rRNA gene from resting spores. This taxonomically broad sampling allowed us to determine the extent to which the rRNA gene tree based on few sequences (18) is representative of the range of genetic diversity present in individual species of AM fungi. We used a novel multidimensional scaling (MDS) method that we developed (30, 31) to visualize the sequence variation in the dataset and to compare it with evolutionary relationships inferred by the gene tree. We then tested whether there are taxonomic or phylogenetic patterns in the distribution of sequence variation across AM fungi. Finally, we used this range of variation together with the MDS visualization method to evaluate how accurately the OTUs generated by four different sequence clustering methods match the known species composition. These methods represented three distinct types of clustering algorithms: 1) Greedy (or top-down: AbundantOTU (32) and UPARSE (33), and 2) Hierarchical (or bottom-up: mothur (34)), algorithms that

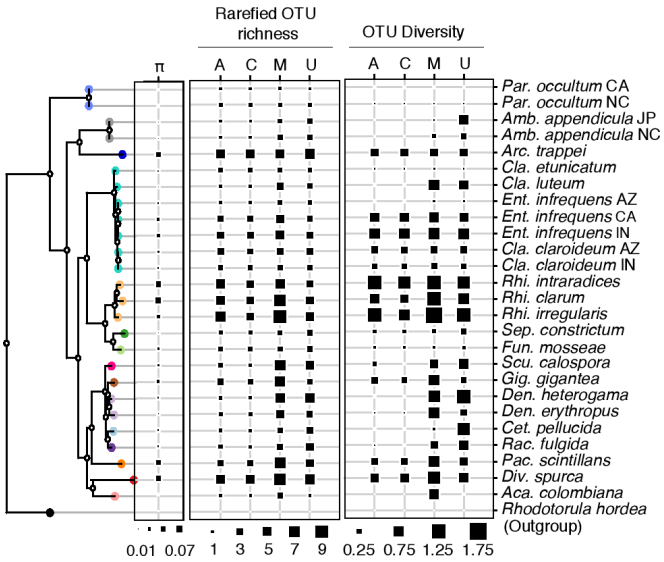

Fig. 3. Phylogenetic differences in nucleotide diversity (π; left) as well as both rarefied OTU richness (middle) and OTU diversity (right) for each of the four clustering methods: AbundantOTU (A), CROP (C), mothur (M), and UPARSE (U) for all species and geographic isolates. The rooted phylogeny was made using representative extended sequences (~675 bp) from each species. The leaves are colored by genus to match Fig. 1 and genera are abbreviated as follows: *Par*: *Paraglomus*, *Amb*: *Ambispora*, *Arc*: *Archaeospora*, *Cla*: *Claroideoglomus*, *Ent*: *Entrophospora*, *Rhi*: *Rhizophagus*, *Sep*: *Septoglomus*, *Fun*: *Funneliformis*, *Scu*: *Scutellospora*, *Gig*: *Gigaspora*, *Den*: *Dentiscutata*, *Cet*: *Cetraspora*, *Rac*: *Racocetra*, *Pac*: *Pacispora*, *Div*: *Diversispora*,and *Aca*: *Acaulospora*.

require a defined sequence similarity threshold, as well as 3) a Bayesian clustering algorithm (CROP (35)) that does not.

Methods

Sequences were primarily obtained from resting spores of 21 species of AM fungi from soil-based cultures. Four species were represented by multiple geographic isolates, and eight species had multiple independent DNA extractions. Spore surfaces were cleaned before crushing, and spore contents served as DNA template to amplify the phylogenetically informative D2 region of the LSU rRNA gene, which was then pyrosequenced (454, Roche). Sequences were subjected to stringent quality screening and chimeric sequences were removed using UCHIME (36). These sequences were then supplemented by sequences from GenBank and from the most comprehensive AM fungal phylogeny to-date (18) to add taxonomic breadth for phylogeny construction. We followed the consensus names for AM fungal genera proposed by (14), except for *Claroideoglomus* and *Entrophospora*, which we grouped together (see results and discussion). See the supplemental methods and Table S1 for more detail about accessions and sequencing methods.

Sequence variation was directly visualized in three-dimensions using a novel MDS visualization algorithm (30), with each sequence represented by an individual point. To compare patterns of sequence similarity with evolutionary relationships, we interpolated a maximum likelihood phylogenetic tree constructed using RAxML (37) into the MDS visualization using a neighbor-joining algorithm we developed previously (31). To delineate OTUs among the sequences, we clustered the dataset using AbundantOTU, mothur, and UPARSE with 97% sequence similarity thresholds using default settings, and using CROP with parameter values meant to approximate a 97% similarity threshold. To visually compare the extents of the resulting OTUs, we color-coded the point representing each sequence in the MDS visualization according to its OTU membership. We used the adjusted Rand index to quantify how accurately the OTUs produced by each clustering method compared to the known species composition.

For each 454 barcode, which represents sequences from different species, geographic isolates, or replicate DNA extractions, we calculated: per-site nucleotide diversity (π), and as well as both the number of OTUs in which sequences from each barcode appeared, standardized by sequence number (rarefied OTU richness), and the Shannon diversity index of the number of sequences from each barcode contained in each OTU (OTU diversity) repeated for each clustering method. We used mixed models to determine how these three metrics varied across taxonomic groups, and tested for phylogenetic signal in them by adapting a linear regression-based approach developed to estimate the heritability using genetic markers (38) (See supplemental methods).

Results

rRNA gene tree closely corresponds to the MDS visualization

The gene tree from a representative set of consensus sequences generally assigned sequences from each genus to a single monophyletic group with the exception of *Rhizophagus* (Fig. 1A). Genus-level clades on the gene tree closely matched the MDS visualization of variation among all sequences due to the tree topology connecting clusters of sequences from the same genus with branches derived from relatively shallow nodes close to the leaves (Fig. 1B and C). In both the MDS visualization and the gene tree, sequences from *Entrophospora infrequens* and those from *Claroideoglomus* species were indistinguishable from each other, and we therefore consider them to be in the same genus.

Test of monophyly of rRNA sequences within species

For the Gigasporaceae (genera *Scutellospora, Racocetra, Gigaspora, Cetraspora,* and *Dentiscutata*), sequences from each species consistently formed monophyletic groups (Fig. 2A). This pattern of monophyly was also evident in the heatmap of pairwise sequence divergence between randomly selected sequences that was calculated independently of the phylogeny: sequences from the same species (triangular tiles on the diagonal), were most similar to each other, followed by sequences from closely related species (square tiles near the diagonal) (Fig. 2B). This pattern was also apparent for species in the genus *Rhizophagus* (family Glomeraceae) despite the larger amount of intraspecific sequence variation in this group indicated by less consistent blocks of color in the heatmap (Fig. 2C-D).

In contrast, sequences from species in *Claroideoglomus* and *Entrophospora* did not always form monophyletic groups on the gene tree (Fig. 2E). All sequences from *C. etunicatum* (green) and *C. luteum* (yellow), as well as a subset of sequences from *C. claroideum* (light orange: Arizona isolate; dark orange: Indiana isolate) and *E. infrequens* (light blue: Arizona isolate; medium blue: California isolate; dark blue: Indiana isolate) that group most centrally in the MDS visualization (Fig. S1) form a clade on the gene tree (referred to here as 'Gene history 1'; Fig. 2E top). Sequences from *C. claroideum* (excluding the Arizona isolate) and *E. infrequens* that are more peripheral in the MDS visualization (Fig. S1) also branch basally on the gene tree (referred to here as 'Gene history 2'; Fig. 2E bottom). The checkerboard pattern in the heatmap of pairwise sequence divergence for these two species indicates sequences from the same gene history are more similar to each other than to sequences from the alternate gene history regardless of the geographic isolate in which they are found (Fig. 2F). In this group there is also more sequence homogeneity within species, geographic isolate, or gene history compared to within species in the Gigasporaceae or *Rhizophagus* as indicated by more solid blocks of color in the heatmap (Fig. 2B, D, and F). Interestingly, sequences from *E. infrequens* consistently shared a single nucleotide polymorphism (SNP) compared to *C. claroideum* across all gene histories and geographic isolates.

Distribution of rRNA sequence variation across the phylogeny

There were substantial differences in nucleotide diversity (π) for species across the AM fungal phylogeny (Fig. 3), and those differences were correlated with observed rarefied OTU richness (number of OTUs) and OTU diversity (equality of sequence numbers among OTUs) (Fig. S2). The phylogenetic patterns in OTU characteristics were usually consistent regardless of clustering method, although AbundantOTU and CROP generally gave significantly lower values of both rarefied OTU richness ($F_{3,204} = 10.61$, $p < 0.0001$) and OTU diversity ($F_{3,204} = 8.03$, $p < 0.0001$) per species compared to mothur or UPARSE, except for the comparison between AbundantOTU and UPARSE for OTU richness (Tukey's HSD $p = 0.21$).

A large and statistically significant amount of the variance in both nucleotide diversity and rarified OTU richness was explained by genus regardless of clustering method used (Table 1).

Table 1. Amount of variance in nucleotide diversity (π) and rarefied OTU richness explained by genus and species, as well as the variance explained by all phylogenetic relationships across each of the four clustering methods. Total variance is the sum of s^2_{Genus}, $s^2_{Species}$, and the variance explained by geographic isolate within species ($s^2_{Isolate}$; not shown). The percentage of s^2_{Total} explained by genus and species is noted, as is the phylogenetic heritability, h^2. Values in bold are significant at the α<=0.05 level.

		Total	Genus		Species		Phylogeny	
		s^2	s^2	%	s^2	%	s^2	h^2
π		**4.19×10^{-4} (1.73×10^{-4})**	**3.36×10^{-4} (1.67×10^{-4})**	80	0.0	0	2.76×10^{-6}	0.07
OTU richness	AbundantOTU	**2.199 (0.882)**	**1.922 (0.863)**	87	0.0	0	**0.312**	0.14
	CROP	**1.714 (0.711)**	**1.493 (0.697)**	87	0.0	0	**0.226**	0.13
	mothur	**5.551 (2.436)**	**4.362 (2.121)**	79	0.057 (0.874)	1	**0.600**	0.11
	UPARSE	**1.983 (0.867)**	**1.506 (0.821)**	76	0.0	0	**0.396**	0.20

However genus only explained a significant amount of the variance in OTU diversity when the sequences were clustered with AbundantOTU or mothur (Table S1). A significant though small proportion of the total variance in OTU richness and diversity was explained by phylogeny regardless of clustering method, but this was not the case for nucleotide diversity (Table 1 and S1).

Comparison of clustering methods

Although sequences from each species in the Gigasporaceae and *Rhizophagus* generally grouped separately in the MDS visualization (Fig. 4A and F), the increased intraspecific variation in *Rhizophagus* (Fig. 2D) is evident in the wider spread of sequences from each species. For these groups, none of the four sequence clustering methods consistently gave a single OTU for each AM fungal species (Fig. 4). We observed both ways in which OTUs can fail to correspond to species-level groups of sequences: 1) Sequences from a single species were split into multiple OTUs, particularly for the three *Rhizophagus* species (Fig. 4 F-J), and 2) sequences from two different species were combined into a single OTU, especially for sequences from *D. erythropus* and *D. heterogama* in the Gigasporaceae, colored orange and red respectively (Fig. 4A) that were assigned to the same OTU by AbundantOTU and CROP (Fig. 4B and C). OTUs that lumped sequences from different species together also occurred in the *Claroideoglomus*/*Entrophospora* group (Fig. S1) as expected given that rRNA sequences in these species are not monophyletic (Fig. 2E).

Generally, we found closer visual correspondence between OTU delineations (Fig. 4B-E and G-J) and species boundaries (Fig. 4A and F) in the Gigasporaceae than in *Rhizophagus*, and this was corroborated by the adjusted Rand index of clustering fit for each clustering method being substantially higher for the Gigasporaceae than for *Rhizophagus* (Fig. 4). AbundantOTU and CROP each gave similar correspondence between OTUs and species boundaries, and both clustering methods were more accurate compared to either mothur or UPARSE for both AM fungal groups (Fig. 4).

Discussion

Sequence variation in AM fungi is generally monophyletic

The novel use of MDS to visualize sequence differences in large datasets from multiple species allows visual comparisons to be made between the range of sequence variation and both projected phylogenies and OTU delineations determined using various sequence clustering methods. Our gene tree, made using a large amount of the sequence variation contained in the 21 species we surveyed (Fig. 1A), generally agreed with the current consensus genus-level phylogeny of AM fungi (14) and the most current AM fungal rRNA gene tree (18). Furthermore, the overall concordance between our AM fungal rRNA gene tree and the MDS visualization of all sequences (Fig. 1B) is consistent with the fact that sequence variation within both species and genera is generally monophyletic regardless of the magnitude of that variation (Fig. 2A-F). The notable exception to this general agreement between the MDS visualization and the phylogeny is the *Claroideoglomus*/*Entrophospora* clade.

Claroideoglomus and Entrophospora form one group that contains incomplete lineage sorting of rRNA gene sequence variation

Sequences from species in both *Claroideoglomus* and *Entrophospora* did not correspond to separate clades on the gene tree, and because of this we identify both genera as a single group. Species within this group are morphologically well defined. For example, asexually produced spores of *E. infrequens* have a different developmental sequence and a unique wall structure compared to spores from *Claroideoglomus* species. However, species in this group show little genetic differentiation over the portion of the LSU rRNA gene used here (Fig. 2E and F), and there is less background variation within species compared to the Gigasporaceae or *Rhizophagus* (Fig. 2B,D, and F). Previous studies have also documented a lack of genetic distinction between *Claroideoglomus* species (28, 29), but have not included *E. infrequens* in their analysis, nor has the most comprehensive phylogeny of AM fungi published to-date (18). Within *Claroideoglomus* species, previous work has also demonstrated the presence of two evolutionary lineages in rRNA gene sequences ('L' and 'S' variants) (29). The PCR primer set used in this study only amplified the 'L' sequence variants due to primer site mismatches with 'S' variants. Most 'L' variant sequences from (29) represent 'Gene history 1', but 9% (17 of 195) of sequences represent 'Gene history 2', including five sequences from *C. luteum*, suggesting this species is subject to the same incomplete lineage sorting of sequence variants that we observed for *E. infrequens* and *C. claroideum* (Fig. 2E and S1). The stronger similarity of sequences from each gene history in *E. infrequens* and *C. claroideum* compared to sequences from the same geographic isolate (Fig. 2F) suggests that this pattern of incomplete lineage sorting is due to the stable coexistence of both sequence types in these species. Despite this, *E. infrequens* is likely to have a single origin as indicated by a SNP that differs from all *Claroideoglomus* species in our dataset and all but six sequences of *C. etunicatum* from (29), and this allows its sequences to be differentiated from those of other species in the group.

Sequence variation differs across genera and the phylogeny

We find strong evidence that the amount of genetic variation in AM fungi varies at the genus level, as indicated by patterns of nucleotide diversity, OTU richness, and OTU evenness (Fig. 3, Table 1). The fact that genus account for more variation in nucleotide diversity and both of the OTU metrics than the full structure of the phylogeny is perhaps surprising, but can be clearly illustrated with the genus *Rhizophagus*. Species in *Rhizophagus* have high nucleotide diversity, OTU richness, and OTU diversity (Fig.s 2, 3), while species from the sister clade of *Septoglomus* and *Funneliformis* have markedly lower variation in all three

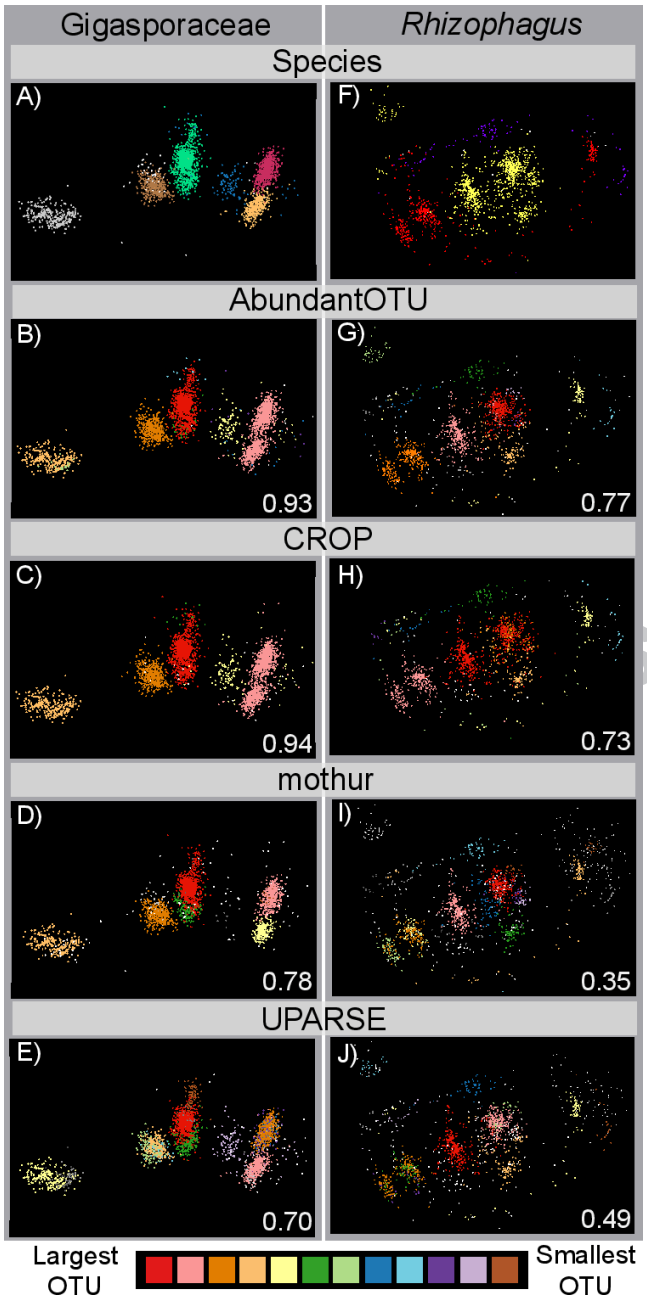

Fig. 4. MDS visualizations of OTU clusters for: left column – species in the family Gigasporaceae and right column – species in the genus *Rhizophagus*; for both columns the sequences (points) in the top row are colored by the species they were isolated from (the same colors are used as in Fig. 2), and the remaining rows (top to bottom) show the same sequences colored by OTU (with >10 sequences) for each of the four clustering methods: AbundantOTU (second row), CROP (third row), mothur (fourth row), and UPARSE (fifth row). The adjusted Rand index values that in the bottom right-hand corner of each OTU panel quantify the fit of the OTU delineations compared to the known species attribution for each sequence. Larger adjusted Rand index values indicate that sequences originating from the same AM fungal species were more accurately grouped together into the same OTU: all four clustering methods had greater accuracy for the Gigasporaceae with its smaller amount of intraspecific variation compared to *Rhizophagus*, but AbundantOTU and CROP were consistently more accurate than mothur or UPARSE for either group.

metrics (Fig. 3). The lower predictive power of the full phylogeny compared to genus indicates that sequence variation is a trait that evolves quickly relative to the deep history represented by the AM fungal phylogeny.

Comparison of Clustering Methodologies

The relationship between rarefied OTU richness and actual species richness varied across genera. Given the generally large amount of intraspecific sequence variation present among AM fungal species (10, 11) (Fig. 2B and D), we expected the clustering methods to generate multiple OTUs for each biological species. This was generally the pattern we found, with sequences from the same species placed into multiple OTUs, especially for *Rhizophagus* with its increased amount of intraspecific variation (Fig. 2B and Fig. 4G-J). Compared to species of *Rhizophagus*, species in the Gigasporaceae contained less intraspecific variation (Fig. 2B), and their OTUs better matched species-level groups, as indicated by a combination of lower OTU richness (Fig. 3A), lower OTU diversity (Fig. 3B), and higher adjusted Rand index values (Fig. 4B-E).

However for genera with both little intra- and interspecific variation, sequences from each species were not always assigned to separate OTUs. Although sequences from *D. erythropus* and *D. heterogama* (colored orange and red, respectively in Fig. 2 and 4A) formed monophyletic groups (Fig. 2B and C), sequences from both species were lumped into the same OTU by AbundantOTU and CROP, the two clustering methods that were otherwise the most accurate (Fig. 4B and C). Sequences from separate species in the *Claroideoglomus*/*Entrophospora* group were also lumped together into the same OTU (Fig. S1) due to a combination of low interspecific sequence variation (Fig. 2F) and incomplete lineage sorting (Fig. 2E), with sequences from each evolutionary group (gene history) clustered together regardless of species identity (Fig. S1). An essential assumption of all sequence clustering methods is that sequences from each species form a monophyletic group on the gene tree, and therefore no clustering method will be able to accurately create species-level OTUs in the presence of incomplete lineage sorting. We demonstrate that the number of OTUs cannot reliably estimate the number of AM fungal species contained in any given sample because not all sequences from known species of AM fungi were consistently grouped together into the same OTU. Furthermore, the direction of this bias in OTU-based estimates of species richness was not consistent across AM fungal genera.

While no clustering method created OTUs that reliably accommodated the genus-specific differences in sequence variation, AbundantOTU and CROP performed nearly identically and were substantially more accurate than either mothur or UPARSE both across the AM fungal phylogeny (Fig. 3) as well as across a range of intraspecific sequence variation in the Gigasporaceae and *Rhizophagus* (Fig. 4). With the exception of *D. erythropus* and *D. heterogama* discussed above, OTUs from AbundantOTU and CROP also better matched the range of sequence variation within each species compared to those from mothur and UPARSE as determined from visual comparisons and their larger adjusted Rand index values (Fig. 4). CROP was the only method tested here that uses Bayesian inference to delineate OTUs instead of a fixed sequence similarity cutoff (35) and it is therefore surprising that its results were largely indistinguishable from those of AbundantOTU using a 97% sequence similarity threshold.

Implications for environmental sequencing

Interpretations of environmental sequence data should be guided by knowledge of how AM fungal sequence variation is distributed across taxonomic groups, and how different clustering methods delineate OTUs based on that variation. Several evaluations and guidelines for determining AM fungal community composition from environmental samples have recently been published (39-41), but they do not accommodate systematic differences in the amount of intraspecific sequence variation across the sampled taxonomic groups. Caution should also be exer-

cised when assigning taxonomic attributions to environmental sequences using phylogenetic differentiation such as the GMYC (26) and PTP (42) methods, or using entities like the virtual taxa that are represented by voucher sequences in the MaarjAM database (21, 27, 43) on the basis of phylogenetic monophyly and high sequence similarity (26). For example, these methods would consistently under-represent the diversity of groups like the Gigasporaceae that has both low intra- and interspecific sequence variation (Fig. 2B), and *Claroideoglomus/Entrophospora* that has incomplete lineage sorting of sequence variation (Fig. 2E-F). Indeed for recent estimates of global AM fungal endemism, when taxa were defined instead using OTUs that represented more of the inherent sequence variation compared to phylogenetically-defined virtual taxa, the estimated endemism increased by an order of magnitude (43). Conversely, for AM fungal groups with a large amount of intraspecific sequence variation, like species in *Rhizophagus*, we show that all four of the clustering methods tested split the sequence variation from a single species into multiple OTUs (Fig. 4G-J). This phenomenon may partly underlie field observations of phylogenetic aggregation in AM fungal communities over short distances (meters) (22) as this aggregation may represent sequences from a single organism that were assigned to several OTUs instead of the presence of multiple, functionally similar species.

Conclusions

Identifying putative AM fungal species in environmental samples typically uses rRNA gene sequences. We find that the level of sequence variation in the rRNA gene consistently differs across AM fungal genera, and that this sequence variation is also subject to incomplete lineage sorting in the *Claroideoglomus/Entrophospora* group. The lack of consistency in the level of rRNA sequence variation that occurs across both taxonomic groups and the full phylogeny presents genuine problems in using OTU richness to estimate species richness for AM fungi. The MDS visualization we demonstrate here can assist as a diagnostic tool to identify groups that may be especially affected by differences in rRNA sequence variation, but no current method of sequence-based species identification is able to overcome this problem.

Acknowledgements:.

We thank Joe Morton of INVAM for the maintenance of and access to AM fungal cultures used in this study, Ed Kim for assistance with spore isolation and sample preparation, and Anna Rosling for initial exploration of the sequence dataset. Funding was provided by NSF grants DEB-0616891 and DEB-0919434, and SERDP RC-2330 to J.D.B. This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute, and in part by the Indiana METACyt Initiative.

1. Woese CR & Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences* 74(11):5088-5090.
2. Sogin ML, *et al.* (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences* 103(32):12115-12120.
3. Ley RE, *et al.* (2008) Evolution of mammals and their gut microbes. *Science* 320(5883):1647-1651.
4. Smith SE & Read DJ (2008) *Mycorrhizal symbiosis* (Academic press).
5. Hartnett DC & Wilson GWT (1999) Mycorrhizae influence plant community structure and diversity in tallgrass prairie. *Ecology* 80(4):1187-1195.
6. van der Heijden MGA, *et al.* (1998) Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* 396(6706):69-72.
7. Vogelsang KM, Reynolds HL, & Bever JD (2006) Mycorrhizal fungal identity and richness determine the diversity and productivity of a tallgrass prairie system. *New Phytologist* 172(3):554-562.
8. Tedersoo L, *et al.* (2014) Global diversity and geography of soil fungi. *Science* 346(6213).
9. Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, & Larsson K-H (2008) Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics* 4:193-201.
10. Sanders IR, Alt M, Groppe K, Boller T, & Wiemken A (1995) Identification of ribosomal DNA polymorphisms among and within spores of the Glomales: application to studies on the genetic diversity of arbuscular mycorrhizal fungal communities. *New Phytologist* 130(3):419-427.
11. Clapp JP, Fitter AH, & Young JPW (1999) Ribosomal small subunit sequence variation within spores of an arbuscular mycorrhizal fungus, *Scutellospora* sp. *Molecular Ecology* 8(6):915.
12. Morton JB (2009) Reconciliation of conflicting phenotypic and rRNA gene phylogenies of fungi in Glomeromycota based on underlying patterns and processes. *Mycorrhizas - Functional processes and ecological impact*, eds Azcón-Aguilar C, Barea JM, Gianinazzi S, & Gianinazzi-Pearson V (Springer, Berlin).
13. Kaonongbua W, Morton JB, & Bever JD (2010) Taxonomic revision transferring species in *Kuklospora* to *Acaulospora* (Glomeromycota) and a description of *Acaulospora colliculosa* sp. nov. from field collected spores. *Mycologia* 102(6):1497-1509.
14. Redecker D, *et al.* (2013) An evidence-based consensus for the classification of arbuscular mycorrhizal fungi (Glomeromycota). *Mycorrhiza* 23(7):515-531.
15. Tisserant E, *et al.* (2013) Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences* 110(50):20117-20122.
16. Boon E, Halary S, Bapteste E, & Hijri M (2015) Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm. *Genome Biology and Evolution* 7:505-521.
17. Lin K, *et al.* (2014) Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genet* 10(1):e1004078.
18. Krüger M, Krüger C, Walker C, Stockinger H, & Schüßler A (2012) Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. *New Phytologist* 193(4):970-984.
19. Bever JD, Morton JB, Antonovics J, & Schultz PA (1996) Host-dependent sporulation and species diversity of arbuscular mycorrhizal fungi in a mown grassland. *Journal of Ecology* 84(1):71-82.
20. Johnson NC, Zak DR, Tilman D, & Pfleger FL (1991) Dynamics of vesicular-arbuscular mycorrhizae during old field succession. *Oecologia* 86(3):349-358.
21. Öpik M, Davison J, Moora M, & Zobel M (2013) DNA-based detection and identification of Glomeromycota: the virtual taxonomy of environmental sequences. *Botany* 92(2):135-147.
22. Horn S, Caruso T, Verbruggen E, Rillig MC, & Hempel S (2014) Arbuscular mycorrhizal fungal communities are phylogenetically clustered at small scales. *ISME J* 8(11):2231-2242.
23. Cheeke TE, *et al.* (2015) Spatial soil heterogeneity has a greater effect on symbiotic arbuscular mycorrhizal fungal communities and plant growth than genetic modification with *Bacillus thuringiensis* toxin genes. *Molecular Ecology* 24(10):2580-2593.
24. Huse SM, Welch DM, Morrison HG, & Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* 12(7):1889-1898.
25. Bokulich NA, *et al.* (2013) Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. *Nat Meth* 10(1):57-59.
26. Powell JR, Monaghan MT, Öpik M, & Rillig MC (2011) Evolutionary criteria outperform operational approaches in producing ecologically relevant fungal species inventories. *Molecular Ecology* 20(3):655-666.
27. Öpik M, *et al.* (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* 188(1):223-241.
28. den Bakker HC, VanKuren NW, Morton JB, & Pawlowska TE (2010) Clonality and recombination in the life history of an asexual arbuscular mycorrhizal fungus. *Molecular Biology and Evolution* 27(11):2474-2486.
29. VanKuren NW, den Bakker HC, Morton JB, & Pawlowska TE (2013) Ribosomal RNA gene diversity, effective population size, and evolutionary longevity in asexual Glomeromycota. *Evolution* 67(1):207-224.
30. Ruan Y, *et al.* (2012) DACIDR: deterministic annealed clustering with interpolative dimension reduction using a large collection of 16S rRNA sequences. in *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pp 329-336.
31. Ruan Y, *et al.* (2014) Integration of Clustering and Multidimensional Scaling to Determine Phylogenetic Trees as Spherical Phylograms Visualized in 3 Dimensions. in *IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, pp 720-729.
32. Ye Y (2010) Identification and quantification of abundant species from pyrosequences of 16S rRNA by consensus alignment. in *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp 153-157.
33. Edgar RC (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Meth* 10(10):996-998.
34. Schloss PD, *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.
35. Hao X, Jiang R, & Chen T (2011) Clustering 16S rRNA for OTU prediction: a method of unsupervised Bayesian clustering. *Bioinformatics* 27(5):611-618.
36. Edgar RC, Haas BJ, Clemente JC, Quince C, & Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* 27(16):2194-2200.
37. Stamatakis A (2014) RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*.
38. Ritland K (2000) Marker-inferred relatedness as a tool for detecting heritability in nature. *Molecular Ecology* 9(9):1195-1204.
39. Lekberg Y, Gibbons SM, & Rosendahl S (2014) Will different OTU delineation methods change interpretation of arbuscular mycorrhizal fungal community patterns? *New Phytologist* 202(4):1101-1104.
40. Lindahl BD, *et al.* (2013) Fungal community analysis by high-throughput sequencing of amplified markers – a user's guide. *New Phytologist* 199(1):288-299.
41. Hart MM, *et al.* (2015) Navigating the labyrinth: a guide to sequence-based, community ecology of arbuscular mycorrhizal fungi. *New Phytologist* 207(1):235-247.
42. Zhang J, Kapli P, Pavlidis P, & Stamatakis A (2013) A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* 29(22):2869-2876.
43. Davison J, *et al.* (2015) Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* 349(6251):970-973.

Please review all the figures in this paginated PDF and check if the figure size is appropriate to allow reading of the text in the figure.

If readability needs to be improved then resize the figure again in 'Figure sizing' interface of Article Sizing Tool.