Phylogenetically Structured Differences in rRNA Gene Sequence Variation among Species of Arbuscular Mycorrhizal Fungi and Their Implications for Sequence Clustering

Geoffrey L. House, a Saliya Ekanayake, b Yang Ruan, b Ursel M. E. Schütte, a,c* Wittaya Kaonongbua, d Geoffrey Fox, b Yuzhen Ye, b James D. Bever a*

Department of Biology, Indiana University, Bloomington, Indiana, USA; School of Informatics and Computing, Indiana University, Bloomington, Indiana, USA; Integrated Program in the Environment, Indiana University, Bloomington, Indiana, USA; Department of Microbiology, Faculty of Science, King Mongkut’s University of Technology Thonburi, Bangkok, Thailand

ABSTRACT
Arbuscular mycorrhizal (AM) fungi form mutualisms with plant roots that increase plant growth and shape plant communities. Each AM fungal cell contains a large amount of genetic diversity, but it is unclear if this diversity varies across evolutionary lineages. We found that sequence variation in the nuclear large-subunit (LSU) rRNA gene from 29 isolates representing 21 AM fungal species generally assorted into genus- and species-level clades, with the exception of species of the genera Claroideoglomus and Entrophospora. However, there were significant differences in the levels of sequence variation across the phylogeny and between genera, indicating that it is an evolutionarily constrained trait in AM fungi. These consistent patterns of sequence variation across both phylogenetic and taxonomic groups pose challenges to interpreting operational taxonomic units (OTUs) as approximations of species-level groups of AM fungi. We demonstrate that the OTUs produced by five sequence clustering methods using 97% or equivalent sequence similarity thresholds failed to match the expected species of AM fungi, although OTUs from AbundantOTU, CD-HIT-OTU, and CROP corresponded better to species than did OTUs from mothur or UPARSE. 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. The OTU richness therefore will not reliably correspond to the AM fungal species richness in environmental samples. Conservatively, this error can overestimate species richness by 4-fold or underestimate richness by one-half, and the direction of this error will depend on the genera represented in the sample.

IMPORTANCE
Arbuscular mycorrhizal (AM) fungi form important mutualisms with the roots of most plant species. Individual AM fungi are genetically diverse, but it is unclear whether the level of this diversity differs among evolutionary lineages. We found that the amount of sequence variation in an rRNA gene that is commonly used to identify AM fungal species varied significantly between evolutionary groups that correspond to different genera, with the exception of two genera that are genetically indistinguishable from each other. When we clustered groups of similar sequences into operational taxonomic units (OTUs) using five different clustering methods, these patterns of sequence variation caused the number of OTUs to either over- or underestimate the actual number of AM fungal species, depending on the genus. Our results indicate that OTU-based inferences about AM fungal species composition from environmental sequences can be improved if they take these taxonomically structured patterns of sequence variation into account.

Sequences of rRNA genes have revolutionized our understanding of microbial diversity (1, 2) and the factors structuring microbial communities (3). When used in combination with high-throughput sequencing, rRNA gene sequences have provided critical insights into the composition and dynamics of microbial communities in a wide range of environments, from the deep ocean (4) to the microbiomes found within other organisms (5). To provide these insights, it is standard for bacterial sequence-based studies to cluster similar sequences into operational taxonomic units (OTUs) that are meant to approximate species-level groups. These OTUs are typically created by a clustering algorithm that uses a single sequence similarity threshold (usually 97%) to define OTU membership. This use of rRNA gene sequences as species barcodes also represents a powerful, culture-independent way to better understand the diversity of eukaryotic microbes. However, in these organisms, sequence variation cannot always be divided into species-level OTUs, because OTUs can either lump sequences from multiple species together or split sequences from the same species apart (6–9). It is currently unclear how well
OTUs correspond to species of arbuscular mycorrhizal (AM) fungi, an ecologically important but poorly described group.

AM fungi (phylum Glomeromycota) are a widely distributed group of fungi that form mutualisms with the roots of most terrestrial plant species (10) and can shape plant communities (11, 12). Because AM fungi are only known to reproduce asexually, species cannot be validated by using the biological species concept and are instead defined by the morphology of their spores formed in the soil. AM fungal species defined in this way can be functionally distinct (13), and the composition of AM fungal communities can change during ecological succession (14). However, in order to identify the AM fungal species participating in active mycorrhizal associations with plants, it is necessary to use DNA sequences from roots. In general, soil-dwelling fungi are genetically diverse (15), and AM fungi have an exceptionally large amount of rRNA gene sequence variation compared to other fungal groups (16, 17). However, in contrast to most other microbes, in AM fungi, this sequence variation can occur within a single multicellular cell (18, 19). While recent genome sequencing of isolates of the genus *Rhizoglomus* (20) has expanded our understanding of genetic variation in AM fungi beyond the rRNA genes (21–23), it remains largely unknown how either genome-wide or rRNA gene sequence variation may itself differ across evolutionary lineages. Previous studies that sampled a limited amount of rRNA gene sequence variation found that sequences from morphologically defined species generally formed clades on the gene tree (24–26), which is an essential prerequisite for clustering sequences into species-level OTUs. However, it is unclear whether this correspondence will remain robust to the sampling of additional sequence variation in AM fungi, as other work has called it into question for species of the genus *Claronoidogygromus* (27, 28).

Understanding how the amount of within-species sequence variation may differ across the AM fungal phylogeny is important to more confidently determine species composition from environmental samples. For studies of bacteria, the inclusion of a mixture of DNA from specific strains with known OTU compositions is an essential prerequisite for clustering sequences into species-level OTUs. However, in contrast to most other microbes, in AM fungi, this sequence variation may itself differ across evolutionary lineages. Previous studies that sampled a limited amount of rRNA gene sequence variation found that sequences from morphologically defined species generally formed clades on the gene tree (24–26), which is an essential prerequisite for clustering sequences into species-level OTUs. However, it is unclear whether this correspondence will remain robust to the sampling of additional sequence variation in AM fungi, as other work has called it into question for species of the genus *Claronoidogygromus* (27, 28).

The resulting sequences were then quality screened before clustering into OTUs. The clustering method CD-HIT-OTU (38) employs its own sequence quality control and chimera sequence removal pipeline that does not rely on base quality scores, and so the full data set of all raw test sequences (188,713) with the PCR primer sequences still attached and all reference sequences was used directly with CD-HIT-OTU. After quality control screening, 51,135 sequences remained and were automatically truncated to 350 bp before being quality filtered by USEARCH (39). The purified PCR products were then pyrosequenced (454 Life Sciences, Branford, CT).

The resulting sequences were then quality screened before clustering into OTUs. The clustering method CD-HIT-OTU (38) employs its own sequence quality control and chimera sequence removal pipeline that does not rely on base quality scores, and so the full data set of all raw test sequences (188,713) with the PCR primer sequences still attached and all reference sequences was used directly with CD-HIT-OTU. After quality control screening, 51,135 sequences remained and were automatically used for clustering. For the AbundantOTU (39), CROP (40), mothur (41), and UPARSE (42) clustering methods, a common pipeline for stringent sequence quality control was applied to all raw test sequences. This pipeline removed sequences with any undetermined bases (N) or any insertions, any deletions, or more than one substitution in the PCR primer site before trimming of the primer site from the sequences. Sequences were then truncated to 350 bp before being quality filtered by using USEARCH (43), allowing a maximum of 1 expected error per sequence. The remaining sequences were dereplicated to remove identical sequences, which was necessary before chimeric sequences could be removed by using UCHIME de novo (44). We then used the multidimensional scaling (MDS) algorithm described below to visually identify a small number of sequences (253 sequences when dereplicated and 1,238 sequences when rereplicated) that closely grouped with sequences from phylogenetically distant species instead of their expected species affiliation. These sequences were removed from the data set, and the number of affected sequences from each barcode and the corresponding isolate names are listed in Table S1 in the supplemental material. After quality control, 9,094 unique test sequences remained. These sequences were added to the dereplicated reference sequences to form the data set for use with the MDS visualization as well as CROP and UPARSE (10,081 sequences); for clustering with AbundantOTU and mothur, the unique test
sequences were rereplicated to reconstruct the abundance of each unique, nonchimeric sequence and were then added to the reference sequences, yielding 51,543 sequences. The number of sequences passing the CD-HIT-OTU quality screening pipeline was within 1% of the number passing the common quality screening pipeline for the other clustering methods, and the distribution of sequences among barcodes was highly linearly correlated between the two quality screening pipelines ($r = 0.93$), suggesting that the two pipelines are comparable overall and that all clustering results can be compared. The reference sequences extended the phylogenetic coverage of the data set and allowed independent validation of species attributions in phylogenies; however, only the test sequences were used for all analyses involving metrics of sequence variation.

**Clustering of sequences into OTUs.** The five different clustering methods that we tested were selected to represent a range of clustering approaches based on three distinct types of underlying clustering algorithms: (i) greedy (or top-down), with AbundantOTU v.0.93b (39), CD-HIT-OTU v.0.0.2 (38), and UPARSE v.8.1.1 (42); (ii) hierarchical (or bottom-up), with mothur v.1.34.0 (41); and (iii) Bayesian, with CROP v.1.33 (40). We used a 97% sequence similarity as the required threshold for the greedy and hierarchical algorithms and ran them with default settings, except that we allowed CD-HIT-OTU to find the consensus PCR primer sequence as the first 21 bp of each sequence. For CROP, similarity levels approximated 97% (see the supplemental material for clustering commands). For each clustering method, we calculated the number of OTUs represented per isolate (OTU richness) and the Shannon diversity index of sequence distribution among OTUs for each isolate (OTU diversity). To correct for uneven numbers of sequences between isolates, we rarefied the OTU richness value for each isolate to the minimum number of sequences per isolate (Diversispora spura, with 85 sequences for AbundantOTU, CROP, mothur, and UPARSE and 93 sequences for CD-HIT-OTU) by using the vegan package in R (v.3.1.2; R Core Team, Vienna, Austria).

**Visualization of sequence variation.** Sequence variation in the data set was visualized by using a novel MDS algorithm (45) that represents each sequence as a point in three-dimensional space, with the location of each point being determined by the pairwise differences between that sequence and all other sequences in the data set. To compare the MDS visualization of sequence variation with the inferred evolutionary relationships between sequences, we interpolated the gene tree into the visualization using a neighbor-joining-based algorithm that we developed previously (46). For clarity, not all sequences were included in this gene tree; instead, the data set was preclustered by using AbundantOTU (39) with a 99% sequence similarity threshold. The resulting sequences still represented all genera in the initial data set. We aligned the sequences with MUSCLE v.3.8.3 (47) and created an unrooted maximum likelihood gene tree using RAxML v.8.0.0 (48) with the generalized time-reversible (GTR) gamma model of nucleotide substitution and 1,000 rapid bootstrap replicates to determine statistical confidence in the tree topology; all subsequent multiple-sequence alignments and gene trees were constructed in the same way, except that the gene trees were rooted by including an outgroup sequence.

To evaluate the sequence similarity between isolates from three clades that contained different amounts of sequence variation, we constructed both a rooted gene tree and a heat map of pairwise sequence divergence for each clade. For the gene trees, all aligned sequences were used from the clades corresponding to the Gigasporaceae and Rhizoglomus, and for the clad corresponding to the genera Claroideoglomus and Entrophospora, the gene tree was made by using representative sequences for each species and each evolutionary history (sequence group) (see Results) that were created by clustering sequences using AbundantOTU (39) with a 97% sequence similarity threshold. Each of the three gene trees was rooted by using a sequence from the most closely related clade in the data set (Fig. 1). Clades on each gene tree that represented a single species were collapsed for clarity and colored by species. For each of the three clades, heat maps of pairwise sequence divergence were made from a random subsample of 240 sequences, split evenly among all species, geographic isolates, and both sequence groups for Claroideoglomus and Entrophospora. Pairwise genetic distances for the aligned sequences were calculated with MEGA 7 (49), using the Kimura 2-parameter model of nucleotide substitution (50) with gamma rate distribution, which was the best-fitting substitution model overall for the three clades, as determined by the substitution model-fitting function in MEGA 7; the use of different substitution models gave qualitatively identical results.

**Evaluation of the phylogenetic signal in patterns of sequence variation.** To estimate the amount of sequence variation contained in each barcode, which represents sequences from different species, geographic isolates, or replicate DNA extractions, we calculated the mean per-site nucleotide diversity (denoted $\pi$). We did this by first aligning all sequences for each barcode and then taking 10 independent, random subsamples from that barcode, with each subsample comprising 85 aligned sequences (the minimum number of sequences per isolate). For each of these subsamples, we calculated the value of $\pi$ averaged across all sites in the alignment that were represented by two or more sequences by using an extension of the single-site method reported previously (51):

$$\pi = \left(1/B \times \sum_{i=1}^{B} \left[ F_{A} + F_{T} + F_{C} + F_{G} \right] - 1 \right) \times \left[ 1 - (F_{A} + F_{T} + F_{C} + F_{G}) \right]$$

where $B$ is the number of positions (columns) with two or more aligned bases and therefore the possibility of a single nucleotide polymorphism (SNP); $x$ is the number of positions in the alignment where a SNP occurs; $n_{i}$ is the number of different sequences represented in the alignment at each position with a SNP; and $F_{A}, F_{T}, F_{C},$ and $F_{G}$ are the frequencies of nucleotides A, T, C, and G, respectively, in the alignment at each position with a SNP. For each barcode, the $\pi$ values calculated from all 10 subsamples were then averaged and were used for all analyses.

![FIG 1 Correspondence between the rRNA gene tree (left) and two different views of the multidimensional scaling (MDS) visualization (right) for sequences from 21 species of AM fungi colored by genus. Branches within genera on the gene tree are collapsed for clarity, with the bootstrap value for each of these genus-level clades being noted on its branch; in contrast, each sequence is represented as a point in the MDS visualization.](image-url)
To determine how nucleotide diversity (π), rarefied OTU richness, and OTU diversity may vary across different taxonomic groups, we tested the magnitude and statistical significance of variance components corresponding to AM fungal genera, species, and isolates using mixed models with the MIXED procedure in SAS (v.9.4; SAS Institute, Inc., Cary, NC). We followed the consensus AM fungal genus names proposed previously (52), except for *Rhizoglosmus* instead of *Rhizopagus* (20) and *Clarodeoglomus* and *Entrophosporoa*, which we considered the same genus for statistical tests (see Results and Discussion). Total variance was calculated as the sum of the variances explained by differences among isolates, species, and genera and was considered to be significantly different from zero if at least one of those variance components was significantly different from zero. For rarefied OTU richness and OTU diversity, we repeated this analysis for each of the five clustering methods.

To test for phylogenetic signal in the variation of nucleotide diversity, rarefied OTU richness, and OTU diversity, we adapted an approach developed previously to estimate heritability using genetic markers (53). Phylogenetic heritability is estimated as the slope of the regression line between the pairwise phylogenetic distance (predictor variable) and the pairwise cross product of sequence variation (response variable). The pairwise cross product of sequence variation is calculated as

\[Z_{ij} = \frac{(y_i - \mu) \times (y_j - \mu)}{V}\]  

(2)

where \(y_i\) and \(y_j\) are the trait (nucleotide diversity, rarefied OTU richness, or OTU diversity) values for each of the samples (sequences from each barcode) in the pairwise comparison, \(\mu\) is the mean trait value for all the samples, and \(V\) is the unbiased variance of the trait value for all of the samples.

Pairwise phylogenetic distances were calculated by using the cophenetic function in the ape package of R on a rooted phylogeny of all species in the test data set using an alignment of extended sequences (~675 bp) that were either collected from data reported previously (25) (19 species), obtained from GenBank (*Cetraspora pellucida* and the outgroup *Rhodotula hordae*), or sequenced by us (*Entrophosporoa infrequens*). These extended sequences were used to obtain better resolution for internal phylogenetic nodes and therefore more accurate phylogenetic distances between genera than would be possible by using our shorter test sequences. Geographic isolates or replicate DNA extractions of the same species were added to the phylogenetic distance matrix as entries with a value of zero.

Phylogenetic heritability is estimated by using the following regression:

\[Z_{ij} = 2r_{ij}h^2 + r_{ij} + e_{ij}\]  

(3)

where \(r_{ij}\) is the pairwise phylogenetic distance for each of the samples in the pairwise comparison, \(h^2\) is the phylogenetic heritability, \(r_{ij}\) is the pairwise correlation of each of the samples due to environmental factors, and \(e_{ij}\) is measurement error. Because \(r_{ij}\) and \(e_{ij}\) are assumed to be independent of \(r_{ij}\), phylogenetic heritability can be calculated as follows:

\[h^2 = Z_{ij}/2r_{ij}\]  

(4)

We then estimated the amount of variance explained by phylogenetic relationships by multiplying the total variance (calculated above) with the phylogenetic heritability from equation 4.

**Assessing the match between OTUs and known AM fungal species.** We visualized the OTUs from each of the four sequence clustering methods by color-coding points (sequences) in the MDS visualization according to their OTU membership for OTUs containing at least 10 sequences, and OTUs were colored according to their size (the number of sequences that they contained). We evaluated how well the OTUs produced by each clustering method matched the known species composition by calculating the adjusted Rand index (54, 55) using the phyclust package in R with the null model assumption that all sequences from each species would be assigned to the same OTU. Rand index values can range from 0 to 1, with a value of 1 indicating a perfect match between OTUs and species.

**Accession number(s).** The raw 454 sequences from this study have been deposited in the NCBI Sequence Read Archive (SRA) under accession number SRP067281.

**RESULTS**

The rRNA gene tree closely corresponds to the MDS visualization. The gene tree from a representative set of sequences generally assigned sequences from each genus to a single clade, with the exception of *Rhizoglosmus* (Fig. 1A). Within genera, groups of similar sequences in the MDS visualization were connected by branches of the gene tree that had relatively shallow nodes, suggesting that the visualization can help identify phylogenetically meaningful variation in large sequence data sets (Fig. 1B). In both the MDS visualization and the gene tree, sequences from isolates of *Entrophosporoa infrequens* and isolates of *Clarodeoglomus* species were indistinguishable, and therefore we considered them to be in the same group.

**Variation in rRNA sequences is generally assorbed into species-level clades.** For the Gigasporaceae (genera *Cetraspora*, *Denticutata*, *Gigaspora*, *Racocetra*, and *Scutellospora*) and the genus *Rhizoglosmus* (family Glomeraceae), sequences from isolates of each species generally formed a clade (Fig. 2A and C), consistent with recombination of rRNA gene sequences within species. The two main exceptions were the majority of sequences from *Denticutata erythropus*, which did not form a clade with bootstrap support and were paraphyletic relative to *D. heterogama*, and sequences from *Rhizoglosmus irregularae*, which formed groups that generally had weak bootstrap support and that were paraphyletic relative to the other *Rhizoglosmus* species. However, the heat maps of pairwise sequence divergence, which were calculated independently of the gene trees, emphasize a pattern of sequence similarity within species. In the Gigasporaceae, sequences from isolates of the same species or from different geographic isolates of *Racocetra fulgida* were most similar to each other, followed by sequences from isolates of closely related species (Fig. 2B). Sequences from isolates of *Rhizoglosmus* species were also distinct from one another despite the larger amount of within-species sequence variation in this group, as indicated by the less consistent blocks of color in the heat map, including the geographic isolates of *Rhizoglosmus claram* that are so differentiated that they appear to be separate species (Fig. 2D).

In contrast, sequences from isolates of each *Clarodeoglomus-Entrophosporoa* species generally did not form a clade but instead assorted into two groups with different evolutionary histories (Fig. 2E). All sequences from *Clarodeoglomus etunicatum* and *C. luteum* as well as a subset of sequences from *C. claroideum* and *E. infrequens* that aggregated in the MDS visualization (see Fig. S1 in the supplemental material) are referred to here as “group A,” but they did not form a clade (Fig. 2E, top). Rather, these sequences were paraphyletic relative to the clade representing other sequences from *C. claroideum* Indiana isolate 2 and all three geographic isolates of *E. infrequens* (referred to here as “group B”) (Fig. 2E, bottom) that are peripheral in the MDS visualization (see Fig. S1 in the supplemental material). The checkerboard pattern in the heat map of pairwise sequence divergence for isolates of these two species indicates that sequence similarity is determined more by sequence group than by species (Fig. 2F).

**Distribution of rRNA sequence variation across the phylogeny.** Nucleotide diversity (π), rarefied OTU richness, and OTU diversity differed substantially for isolates across the AM fungal
phylogeny but remained relatively constant between geographic isolates of the same species (Fig. 3). These phylogenetic patterns of the OTU-based metrics were usually consistent regardless of the clustering method, although all clustering methods had significantly lower values of rarefied OTU richness per isolate than those determined by mothur (\(F_{L,140} = 8.52; P < 0.001\)), and AbundantOTU, CD-HIT-OTU, and CROP had significantly lower values of OTU diversity (\(F_{L,140} = 5.50; P < 0.001\)) per isolate than those determined by mothur or UPARSE, except for the comparison between AbundantOTU and UPARSE (\(P = 0.13\) by Tukey’s honestly significant difference [HSD] test). Rarefied OTU richness for all clustering methods consistently overestimated true species richness across the phylogeny, by up to a factor of 4 for CD-HIT-OTU and a factor of 9 for mothur, with the other
clustering methods having intermediate levels of overestimation (Fig. 3).

Genera explained a large and statistically significant amount of variance in rarefied OTU richness for all methods except CD-HIT-OTU (\( \rho = 0.10 \)) and in OTU diversity for all methods except UPARSE (\( \rho = 0.12 \)) and CD-HIT-OTU (\( \rho = 0.10 \) (Table 1). Genera also explained a large although not significant amount of the variance in nucleotide diversity (\( \rho = 0.07 \)) (Table 1). In contrast, different isolates accounted for a relatively small and nonsignificant amount of variance in both nucleotide diversity and rarefied OTU richness, although they accounted for a significant and much larger amount of the variance in OTU diversity for all methods except UPARSE (\( \rho = 0.07 \)) (Table 1). A significant although small amount of the total variance in both OTU richness and diversity was explained by the full phylogeny regardless of the clustering method, but this was not the case for nucleotide diversity (\( \rho = 0.13 \) (Table 1).

No clustering method produced species-level OTUs for AM fungi. In the Gigasporaceae and in Rhizoglomus, none of the four sequence clustering methods that used the common sequence processing pipeline consistently gave a single OTU for isolates of each AM fungal species (Fig. 4). We observed both ways in which OTUs fail to correspond to species-level groups of sequences: (i) sequences from a single species were split into multiple OTUs, particularly for the three Rhizoglomus species (Fig. 4F to J), and (ii) sequences from different species were lumped into a single OTU, especially for sequences from D. erythropus and D. heterogama in the Gigasporaceae (Fig. 4A), which were assigned to the same OTU by AbundantOTU and CROP (Fig. 4B and C) and also by CD-HIT-OTU (results not shown). OTUs that lumped sequences from different species also occurred for Claroideoglomus-Entrophospora (see Fig. S1 in the supplemental material), as expected given that the rRNA sequence variation does not assort into species-level clades in this group (Fig. 2E). Overall, OTU delineations were visually more closely matched to species of the Gigasporaceae (Fig. 4A to E) than Rhizoglomus (Fig. 4F to J). This was corroborated by the substantially higher adjusted Rand index values, which quantify the match between OTUs and species, for all clustering methods for the Gigasporaceae (CD-HIT-OTU value, 0.95) (Fig. 4B to E) than those for Rhizoglomus (CD-HIT-OTU value, 0.87) (Fig. 4G to J). For both AM fungal groups, AbundantOTU, CROP, and CD-HIT-OTU produced OTUs that better matched species than did OTUs from either mother or UPARSE.

**DISCUSSION**

Sequence variation in genera and species is generally assorted into clades. With the exception of the genus Diversispora, the gene tree made by using the large range of sequence variation contained in isolates of the 21 species surveyed (Fig. 1A) generally agreed

**TABLE 1** Total variance in nucleotide diversity, rarefied OTU richness, and OTU diversity

<table>
<thead>
<tr>
<th>Metric</th>
<th>OTU clustering method</th>
<th>Total</th>
<th>Genus</th>
<th>Species</th>
<th>Isolate</th>
<th>Phylogeny</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>( \pi )</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>( \rho )</td>
<td>( h^2 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td>( s^2 ) (SE)</td>
<td>( s^2 ) (SE)</td>
<td>%</td>
<td>( s^2 ) (SE)</td>
<td>%</td>
</tr>
<tr>
<td>OTU richness</td>
<td>AbundantOTU</td>
<td>2.18 (0.88)</td>
<td>1.98 (0.86)</td>
<td>91</td>
<td>0.0</td>
<td>0.20 (0.16)</td>
</tr>
<tr>
<td></td>
<td>CROP</td>
<td>1.75 (0.72)</td>
<td>1.55 (0.70)</td>
<td>89</td>
<td>0.0</td>
<td>0.20 (0.12)</td>
</tr>
<tr>
<td></td>
<td>mothur</td>
<td>5.70 (2.39)</td>
<td>4.40 (2.18)</td>
<td>77</td>
<td>0.57 (0.09)</td>
<td>0.73 (0.49)</td>
</tr>
<tr>
<td></td>
<td>UPARSE</td>
<td>2.03 (1.16)</td>
<td>1.49 (0.89)</td>
<td>74</td>
<td>0.18 (0.65)</td>
<td>0.35 (0.37)</td>
</tr>
<tr>
<td></td>
<td>CD-HIT-OTU</td>
<td>0.51 (0.32)</td>
<td>0.26 (0.21)</td>
<td>51</td>
<td>0.02 (0.17)</td>
<td>0.23 (0.18)</td>
</tr>
<tr>
<td>OTU diversity</td>
<td>AbundantOTU</td>
<td>0.11 (0.05)</td>
<td>0.06 (0.03)</td>
<td>50</td>
<td>0.01 (0.03)</td>
<td>0.04 (0.02)</td>
</tr>
<tr>
<td></td>
<td>CROP</td>
<td>0.09 (0.04)</td>
<td>0.04 (0.03)</td>
<td>48</td>
<td>0.01 (0.02)</td>
<td>0.04 (0.02)</td>
</tr>
<tr>
<td></td>
<td>mothur</td>
<td>0.09 (0.07)</td>
<td>0.11 (0.06)</td>
<td>61</td>
<td>0.00</td>
<td>0.07 (0.03)</td>
</tr>
<tr>
<td></td>
<td>UPARSE</td>
<td>0.14 (0.08)</td>
<td>0.05 (0.04)</td>
<td>36</td>
<td>0.03 (0.05)</td>
<td>0.06 (0.04)</td>
</tr>
<tr>
<td></td>
<td>CD-HIT-OTU</td>
<td>0.09 (0.04)</td>
<td>0.03 (0.02)</td>
<td>33</td>
<td>0.02 (0.03)</td>
<td>0.04 (0.02)</td>
</tr>
</tbody>
</table>

*a* Shown are data for the total variance (\( s^2 \)) in nucleotide diversity (\( \pi \)), rarefied OTU richness, and OTU diversity, including the amount of variance and percentage of the total variance explained by genus, species, and isolate as well as the amount of variance explained by phylogeny and phylogenetic heritability (\( h^2 \)). Values in boldface type denote significant differences in variance across the taxonomic group or across the phylogeny at the level of an \( \alpha \) of 0.05.
for both groups, AbundantOTU and CROP consistently generated OTUs that with its smaller amount of within-isolate variation, than for methods had closer OTU-to-species correspondence for the Gigasporaceae, and the sequences originating from each AM fungal species. All four clustering adjusted Rand index values indicate closer correspondence between each OTU containing the number of sequences that they contain. The adjusted Rand index value in the top right corner of each panel for the clustering methods quantifies the fit of the top four rows (top to bottom) show the same sequences colored by OTU for each of the following four clustering methods that used the common sequence processing pipeline: AbundantOTU (second row), CROP (third row), mothur (fourth row), and UPARSE (fifth row). Only OTUs containing >10 sequences are shown, and OTUs are colored according to the number of sequences that they contain. The adjusted Rand index value in the top right corner of each panel for the clustering methods quantifies the fit of the OTU delineations compared to the known species attribution. Higher adjusted Rand index values indicate closer correspondence between each OTU and the sequences originating from each AM fungal species. All four clustering methods had closer OTU-to-species correspondence for the Gigasporaceae, with its smaller amount of within-isolate variation, than for Rhizoglomus, but for both groups, AbundantOTU and CROP consistently generated OTUs that better matched species than OTUs generated from mothur or UPARSE. Interactive three-dimensional versions of these MDS visualizations are available at https://spidal-gw.dsc.soic.indiana.edu/public/groupdashboard/AM%20fungal%20clustering%20AEM.

that the correspondence between morphologically defined taxa and clades on the AM fungal gene tree is robust to differences in the magnitude of sequence variation (Fig. 2A to D). The notable exception to this general agreement between morphologically and phylogenetically defined taxa is the Claroideoglomus-Entrophospora clade.

Claroideoglomus and Entrophospora form one group with gene tree discordance. Sequences from isolates of Claroideoglomus and Entrophospora species did not correspond to separate clades on the gene tree, and because of this, we identify both genera as a single group. Species in this group are morphologically well defined. For example, spores of E. infrequens develop differently and have a unique wall structure compared to spores of Claroideoglomus species (see Fig. S2 in the supplemental material). However, isolates of both species show little genetic differentiation over the portion of the LSU rDNA gene used here (Fig. 2E and F) and have less variation within species than the Gigasporaceae or Rhizoglomus (Fig. 2B, D, and F). Previous studies also documented a lack of genetic distinction between Claroideoglomus species (27, 28) but did not include E. infrequens in their analyses, nor has the most comprehensive AM fungal rRNA gene tree (25). In Claroideoglomus species, previous work demonstrated the presence of two evolutionary lineages in LSU rRNA gene sequences (“L” and “S” variants) (28). The PCR primer set used here amplified only L sequence variants due to primer site mismatches with S variants. Most L variant sequences reported previously (28) correspond to group A from this study, but 9% (17 of 195) of sequences correspond to group B, including 5 sequences from C. luteum, suggesting that it has the same gene tree discordance that we observed for E. infrequens and C. claroideum (Fig. 2E and F). Because this discordance is shared among species, the sequence variation that underlies it was likely present in a common ancestor of the group and may be due to introgression or to incomplete lineage sorting of standing sequence variation.

Sequence variation differs across clades and the full phylogeny. We find strong evidence that the amount of genetic variation in AM fungi varies across clades that correspond to genera, as indicated by patterns of nucleotide diversity, rarefied OTU richness, and OTU diversity (Fig. 3 and Table 1). Although these genus-level differences were not statistically significant for CD-HIT-OTU (Table 1), this was likely due to its good overall performance in giving consistently low rarefied OTU richness estimates across nearly all species and geographic isolates (Fig. 3). These patterns were also consistent regardless of how many spores were sampled. This indicates that the same amount of sequence variation that occurs in multiple spores is also contained within a single spore and provides evidence of phylogenetic heritability of intracellular variation. In addition, the small overall amount of variance explained by isolate for nucleotide diversity and rarefied OTU richness is indicative of the presence of relatively little variation between replicate DNA extractions of the same culture, although this variance was larger for OTU diversity (Table 1).

The fact that genus-level clades account for more of the variance in all three metrics of sequence variation than the full phylogeny (Table 1) is perhaps surprising but can be clearly illustrated with the genus Rhizoglomus. Species of Rhizoglomus have markedly higher values for each of the three metrics than do species of the sister genera Septoglomus and Funneliformis (Fig. 3). The lower predictive power of the full phylogeny than of genus-level clades indicates that sequence variation is a trait that evolves quickly
Bidirectional error in linking OTUs to AM fungal species. None of the sequence clustering methods tested here were effective at consistently creating species-level OTUs, and therefore OTU richness cannot reliably estimate AM fungal species richness. The internal transcribed spacer (ITS) region of the rRNA genes has been proposed as a universal sequence barcode for fungal species, but it has limited use for AM fungi due to the large amount of ITS sequence variation in this group, and either the small-subunit (SSU) or LSU rRNA gene is typically used instead (17, 57). However, even within LSU sequences, we found enough variation across genus-level clades of AM fungi to cause bidirectional errors in the correspondence between OTUs and species.

For groups like *Rhizoglomus* with a large amount of within-isolate variation (Fig. 2D), sequences from the same species were commonly split into multiple OTUs (Fig. 4G to J), a finding similar to the findings of another recent study (34). In contrast, for genera with little between-isolate variation, sequences from isolates of different species could be lumped into the same OTU. For example, although sequences from isolates of *D. eurythrops* and *D. heterogama* (Fig. 2 and 4A, red and orange, respectively) showed consistent differences (Fig. 2A and B), they were assigned to the same OTU by AbundantOTU and CROP (Fig. 4B and C) and also by CD−HIT−OTU (results not shown), the three clustering methods that otherwise gave the closest matches between OTUs and species. Sequences from isolates of different *Claroideoglomus-Entrophospora* species were also lumped together into the same OTU (see Fig. S1 in the supplemental material). An essential assumption of all sequence clustering methods is that sequences from each species form a clade on the gene tree, and therefore, no clustering method can create species-level OTUs for groups like this that have gene tree discordance. A similar bidirectional error in the correspondence between OTUs and species occurs for other groups of fungi (6) and eukaryotic microbes (7, 8), suggesting that it is a shared pattern across a range of genetic systems that is determined by the evolutionary history of the group (9).

No clustering method created OTUs that accommodated the range of sequence variation that occurs across the AM fungal phylogeny (Fig. 3). However, AbundantOTU, CROP, and CD−HIT−OTU performed comparably and gave OTUs better matched to species than either mothur or UPARSE, both across the phylogeny (Fig. 3) and across a range of sequence variation in the Gigasporaceae and *Rhizoglomus* (see Results for adjusted Rand index values for CD−HIT−OTU) (Fig. 4). Bayesian inference, as used by CROP, potentially has the flexibility to accommodate differences in sequence variation when generating OTUs (40), so it is surprising that AbundantOTU and CD−HIT−OTU, with their fixed similarity threshold, gave clustering results nearly identical to those of CROP.

Implications for environmental sequencing. Determining AM fungal community composition from environmental sequence data should be guided by knowledge of how sequence variation is distributed across the phylogeny and also how different clustering methods delineate OTUs based on that variation. Several evaluations and guidelines for characterizing AM fungal communities in environmental samples have recently been reported (58–60), but they do not accommodate the systematic differences in within-species sequence variation that occurs in isolates across the phylogeny (Fig. 3). Caution should also be exercised when assigning taxonomic attributions to environmental sequences using phylogenetic differentiation, such as the generalized mixed Yule coalescent (GMYC) (33) and Poisson tree processes (PTP) (61) methods, or using entities like the virtual taxa that are represented by voucher sequences in the MaariAM database on the basis of both phylogenetic monophyly and high sequence similarity (32, 57, 62). For example, these methods would consistently underrepresent the diversity in groups with a small amount of between-species sequence variation, like some species in the Gigasporaceae (Fig. 2A and B), and in groups that have gene tree discordance, like the *Claroideoglomus-Entrophospora* clade (Fig. 2E and F). For the isolates considered here, the MDS visualization demonstrates that OTU richness can underestimate species richness by one-half (Fig. 4A to C), although more extensive sequence sampling within these groups is necessary to better understand the magnitude of this effect. Conversely, for AM fungal groups with a large amount of within-species sequence variation, such as species of *Rhizoglomus*, OTU richness overestimated actual species richness by at least 4-fold across all clustering methods, which was the most of any AM fungal genus (Fig. 3). This phenomenon may partly underlie field observations of phylogenetic aggregation in AM fungal communities over short distances (meters) (31), 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. Finally, recent evidence documenting variation in the level of heterokaryosis that occurs among isolates of the same AM fungal species (63) may also affect the range of sequence variation observed for AM fungi.

Identification of putative AM fungal species in environmental samples typically is done by using rRNA gene sequences. We find that the level of sequence variation in the LSU rRNA gene consistently differs across clades of AM fungi that correspond to genera and that this sequence variation does not correspond to morphologically defined species of *Claroideoglomus-Entrophospora*. The different levels of rRNA sequence variation that occur across both genus-level clades and the full phylogeny present genuine problems in the use of OTU richness to estimate species richness for AM fungi. The MDS visualization that 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, and we suggest that interpretations of AM fungal OTU composition from environmental sequences should be made in the context of these limitations.

ACKNOWLEDGMENTS

We thank J. Morton and INVAM for access to AM fungal cultures, E. Kim for laboratory assistance, A. Rosling for initial data exploration, M. Hahn for insightful comments that substantially improved the manuscript, and two anonymous reviewers and S. Hempel for their constructive comments that also improved the manuscript.

Computing resources were provided by the Digital Science Center at Indiana University and in part by Lilly Endowment, Inc., through the Indiana University Pervasive Technology Institute and the Indiana METACy Initiative.

Wittaya Kaonongbua and James D. Bever designed the research; Wittaya Kaonongbua performed the research; Yang Ruan, Geoffrey Fox, and Yuzhen Ye contributed new analytic methods and tools; Geoffrey L. House, Saliya Ekanayake, Ursel M. E. Schütte, and James D. Bever ana-
FUNDING INFORMATION
This work, including the efforts of Geoffrey L. House, Ursel M. E. Schütte, and James D. Bever, was funded by the National Science Foundation (NSF) (DEB-0616891). This work, including the efforts of Saliya Ekanayake and Geoffrey Fox, was funded by the National Science Foundation (NSF) (DEB-0919434). This work, including the efforts of Wittaya Kaonongbua and James D. Bever, was funded by the Strategic Environmental Research and Development Program (SERDP) (RC-2330). This work, including the efforts of Geoffrey L. House, Ursel M. E. Schütte, van der Heijden MGA, Klironomos JN, Ursic M, Moutoglis P, Streitwolf-Engel R, Boller T, Wiemken A, Sanders IR, Alt M, Groppke B, Tikhomirova I, Li H, Foo S, Yuan J, Yang Y, Zhou H, Zhang Z, Alvernas J, Glickman J, Hsu J, and Li L, was funded by the National Science Foundation (NSF) (DEB-1443054).

REFERENCES
Trouvelot S, van Tuinen D, Hijri M, Gianinazzi-Pearson V.

Powell JR, Monaghan MT, O¨ pik M, Rillig MC.

Edgar RC.

Katoh K, Frith MC.


